YAYASAN PRIMA AGUS TEKNIK

TEORI EKONOMI BERBASIS BIG DATA

Dr. Agus Wibowo, M.Kom, M.Si, MM

TEORI EKONOMI BERBASIS BIG DATA

Penulis :
Dr. Agus Wibowo, M.Kom., M.Si., MM.

ISBN : 9 786238 120765

Editor :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom.

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom.

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : info@stekom.ac.id

KATA PENGANTAR

Puji syukur penulis ucapkan kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan buku ajar yang berjudul ***"Teori Ekonomi Berbasis Big Data"*** dengan baik. Manfaat utama dari teori ekonomi berbasis big data adalah kemampuannya untuk memberikan wawasan yang lebih akurat dan real-time terkait dengan dinamika ekonomi. Analisis data yang canggih dapat mengidentifikasi pola-pola tersembunyi, trend, dan perubahan perilaku konsumen dengan lebih cepat. Hal ini memungkinkan pengambil keputusan ekonomi, seperti perusahaan dan pemerintah, untuk merespons secara lebih adaptif terhadap perubahan pasar dan menciptakan kebijakan yang lebih efektif.

"Teori Ekonomi Berbasis Big Data" merupakan pendekatan yang menggabungkan prinsip-prinsip ekonomi tradisional dengan analisis data yang massif dan kompleks. Dalam konteks ini, big data merujuk pada kumpulan data besar dan heterogen yang diperoleh dari berbagai sumber, termasuk transaksi online, media sosial, dan sensor-sensor yang terhubung dengan Internet of Things (IoT). Teori ini mengakui bahwa data yang sangat besar dan bervariasi dapat memberikan wawasan mendalam terkait perilaku ekonomi, keputusan konsumen, dan dinamika pasar yang tidak dapat dipahami sepenuhnya oleh teori ekonomi konvensional.

Keuntungan lainnya adalah kemampuan teori ekonomi berbasis big data untuk memodelkan keterkaitan yang kompleks di antara berbagai variabel ekonomi. Dengan memanfaatkan teknik-teknik analisis prediktif dan machine learning, teori ini dapat memahami hubungan yang tidak linear dan dinamika yang kompleks di pasar. Hal ini memungkinkan para pengambil keputusan untuk mengoptimalkan strategi mereka dan meningkatkan efisiensi operasional.

Selain itu, teori ekonomi berbasis big data dapat menghasilkan solusi inovatif untuk masalah-masalah ekonomi dan sosial. Melalui pemahaman mendalam terhadap perilaku konsumen dan dinamika pasar, teori ini dapat memberikan dasar untuk pengembangan produk baru, perencanaan pemasaran yang lebih efektif, dan strategi bisnis yang lebih adaptif. Dengan demikian, teori ekonomi berbasis big data bukan hanya alat prediktif, tetapi juga katalisator untuk inovasi dan pertumbuhan ekonomi yang berkelanjutan.

Buku ini dibagi menjadi empat bagian: Bagian I *"Peluang Big Data"* mengeksplorasi potensi nilai big data dengan fokus khusus pada konteks Eropa. Bab 1 memaparkan potensi nilai dari big data dan mengkaji dimensi hukum, bisnis, dan sosial yang perlu ditangani untuk memenuhi janjinya. Selanjutnya, Bab. 2 secara singkat memperkenalkan proyek besar Komisi Eropa dan kewenangannya untuk menetapkan peta jalan penelitian big data untuk Horizon 2020 guna mendukung dan mendorong penelitian dan inovasi di Area Penelitian Eropa.

Bagian II *"Rantai Nilai Big Data"* merinci keseluruhan siklus hidup big data dari sudut pandang teknis, mulai dari akuisisi data, analisis, kurasi, dan penyimpanan hingga penggunaan dan eksploitasi data. Bab 3 memperkenalkan konsep inti rantai nilai big data. Lima bab

berikutnya memerinci setiap tahapan rantai nilai data, termasuk ringkasan mutakhir, kasus penggunaan yang muncul, dan pertanyaan penelitian terbuka yang penting. Bab 4 memberikan cakupan komprehensif tentang akuisisi data besar, yang merupakan proses pengumpulan, pemfilteran, dan pembersihan data sebelum dimasukkan ke dalam gudang data atau solusi penyimpanan lainnya untuk diproses lebih lanjut. Pada Bab 5 membahas analisis big data yang berfokus pada transformasi data mentah yang diperoleh menjadi sumber daya yang koheren dan dapat digunakan, sesuai untuk analisis guna mendukung pengambilan keputusan dan skenario penggunaan spesifik domain. Dan bab 6 menyelidiki bagaimana lanskap big data yang muncul mendefinisikan persyaratan baru untuk infrastruktur kurasi data dan bagaimana infrastruktur kurasi big data berevolusi untuk memenuhi tantangan-tantangan ini. Selanjutnya bab 7 memberikan gambaran singkat tentang sistem penyimpanan data besar yang mampu menangani data berkecepatan tinggi, bervolume tinggi, dan beragam. Akhirnya, Bab. 8 mengkaji tujuan bisnis yang memerlukan akses ke data dan analisis serta integrasinya ke dalam pengambilan keputusan bisnis di berbagai sektor.

Bagian III *"Penggunaan dan Eksploitasi Big Data"* menggambarkan kemungkinan penciptaan nilai dari penerapan big data di berbagai sektor, termasuk industri, layanan kesehatan, keuangan, energi, media, dan layanan publik. Dalam bab 9 memberikan latar belakang konseptual dan tinjauan umum mengenai inovasi berbasis data besar di masyarakat, menyoroti faktor-faktor dan tantangan yang terkait dengan penyebaran, penyerapan, dan keberlanjutan inisiatif berbasis data besar yang memadai. Bab-bab selanjutnya menjelaskan kecanggihan big data di berbagai sektor, mengkaji faktor-faktor pendukung, kebutuhan industri, dan skenario penerapan, serta menyaring analisisnya menjadi serangkaian persyaratan komprehensif di seluruh rantai nilai big data. Bab 10 merinci beragam peluang teknologi big data untuk meningkatkan pemberian layanan kesehatan secara keseluruhan. Dan bab 11 menyelidiki potensi nilai yang dapat diperoleh dari big data oleh organisasi pemerintah dengan meningkatkan produktivitas di lingkungan dengan keterbatasan anggaran yang signifikan. Bab 12 mengeksplorasi berbagai keuntungan big data bagi lembaga keuangan. Selanjutnya bab 13 mengkaji teknologi big data khusus domain yang diperlukan untuk sistem energi dan transportasi cyber-fisik, yang fokusnya perlu beralih dari sekadar big data ke teknologi smart data. Bab 14 membahas sektor media dan hiburan yang dalam banyak hal merupakan pengguna awal teknologi big data karena memungkinkan mereka mendorong transformasi digital, memanfaatkan secara lebih penuh tidak hanya data yang sudah tersedia namun juga sumber data baru baik dari dalam maupun luar organisasi.

Terakhir, Bagian IV *"Peta Jalan untuk Riset Big Data"* mengidentifikasi dan memprioritaskan persyaratan lintas sektoral untuk riset big data dan menguraikan isu-isu teknologi, ekonomi, politik, dan sosial yang paling mendesak dan menantang untuk big data di Eropa. Bab 15 memerinci proses yang digunakan untuk menggabungkan persyaratan big data dari berbagai sektor ke dalam satu set persyaratan lintas sektor yang diprioritaskan yang digunakan untuk menentukan peta jalan kebijakan teknologi, bisnis, dan masyarakat serta rekomendasi tindakan. Bab 16 yang merupakan bab terakhir dalam buku ini akan menjelaskan peta jalan di bidang teknologi, bisnis, kebijakan, dan masyarakat. Bab ini memperkenalkan Big

Data Value Association (BDVA) dan kontrak Big Data Value Public Private Partnership (BDV cPPP) yang memberikan kerangka kerja bagi kepemimpinan industri, investasi, dan komitmen pihak swasta dan publik untuk membangun ekonomi berbasis data di seluruh Eropa. Demikian buku ajar ini kami buat, dengan harapan agar pembaca dapat memahami informasi dan juga mendapatkan wawasan mengenai teori ekonomi berbasis big data.

Semarang, Desember 2023
Penulis

Dr. Agus Wibowo M.Kom , M.Si, MM

DAFTAR ISI

BAB 1
PELUANG NILAI BIG DATA

1.1 PENDAHULUAN

Volume data tumbuh secara eksponensial, dan diperkirakan pada tahun 2020 akan terdapat lebih dari 16 zettabytes (16 Triliun GB) data yang berguna (Turner dkk. 2014). Kita berada di ambang era di mana setiap perangkat online, di mana sensor ada di mana-mana di dunia yang menghasilkan aliran data secara terus-menerus, di mana volume data yang ditawarkan dan dikonsumsi di Internet akan meningkat berkali-kali lipat, di mana Internet Segala sesuatunya akan menghasilkan sidik jari digital dunia kita.

Big Data adalah bidang baru di mana teknologi inovatif menawarkan cara-cara baru untuk mengekstraksi nilai dari banyaknya informasi baru. Kemampuan untuk mengelola informasi dan mengekstrak pengetahuan secara efektif kini dipandang sebagai keunggulan kompetitif utama. Banyak organisasi membangun bisnis inti mereka berdasarkan kemampuan mereka mengumpulkan dan menganalisis informasi untuk mengekstrak pengetahuan dan wawasan bisnis. Adopsi teknologi Big Data dalam sektor industri bukanlah suatu kemewahan namun merupakan kebutuhan penting bagi sebagian besar organisasi untuk bertahan hidup dan mendapatkan keunggulan kompetitif.

Bab ini mengeksplorasi potensi nilai dari Big Data dengan fokus khusus pada konteks Eropa dan mengidentifikasi potensi transformasi positif dari Big Data dalam sejumlah sektor utama. Pertemuan ini membahas perlunya strategi yang jelas untuk meningkatkan daya saing industri Eropa guna mendorong inovasi dan daya saing. Terakhir, bab ini menjelaskan dimensi-dimensi utama, termasuk keterampilan, hukum, bisnis, dan sosial, yang perlu ditangani dalam Ekosistem Big Data Eropa.

1.2 MEMANFAATKAN BIG DATA

Dampak Big Data tidak hanya berdampak pada dunia komersial; dalam komunitas ilmiah, ledakan data yang tersedia menghasilkan apa yang disebut ilmu data, sebuah pendekatan baru yang intensif data terhadap penemuan ilmiah. Kemampuan teleskop atau akselerator partikel untuk menghasilkan beberapa petabyte data per hari menimbulkan masalah berbeda dalam hal penyimpanan dan pemrosesan. Para ilmuwan tidak memiliki solusi siap pakai yang siap menganalisis dan membandingkan dengan tepat kumpulan data yang tersebar dan sangat besar. Untuk mewujudkan visi ini diperlukan teknologi Big Data yang inovatif untuk pengelolaan, pemrosesan, analisis, penemuan, dan penggunaan data.

Data telah menjadi faktor produksi baru, sama seperti aset fisik dan sumber daya manusia. Memiliki basis teknologi dan struktur organisasi yang tepat untuk mengeksploitasi data sangatlah penting. Eropa harus memanfaatkan potensi Big Data untuk menciptakan nilai bagi masyarakat, warga negara, dan dunia usaha. Namun, dari sudut pandang adopsi industri, eropa tertinggal dibandingkan amerika serikat dalam hal teknologi Big Data dan tidak

memanfaatkan potensi manfaat Big Data di seluruh sektor industrinya. Diperlukan strategi yang jelas untuk meningkatkan daya saing industri eropa melalui Big Data. Meskipun perusahaan-perusahaan yang berbasis di AS dikenal luas atas karya mereka di bidang Big Data, sangat sedikit organisasi di eropa yang terkenal atas karya mereka di bidang Big Data. Hal ini membuat eropa bergantung pada teknologi yang datang dari luar dan mungkin menghalangi pemangku kepentingan di eropa untuk mengambil keuntungan penuh dari teknologi Big Data. Menjadi kompetitif dalam teknologi dan solusi Big Data akan memberikan eropa sumber daya saing baru dan potensi untuk mengembangkan industri terkait data baru yang akan menghasilkan lapangan kerja baru.

Untuk mengatasi permasalahan yang ada saat ini memerlukan pendekatan holistik, dimana kegiatan teknis bekerja sama dengan aspek bisnis, kebijakan, dan masyarakat. Eropa perlu menentukan tindakan yang mendukung penerapan dan adopsi teknologi yang lebih cepat dalam kasus nyata. Dukungan diperlukan tidak hanya untuk membangun teknologi tetapi juga untuk menumbuhkan ekosistem yang memungkinkan inovasi. Ada banyak tantangan teknis yang memerlukan penelitian lebih lanjut, namun upaya ini harus disertai dengan pemahaman berkelanjutan tentang bagaimana teknologi Big Data mendukung tantangan bisnis dan masyarakat. Bagaimana inovasi berbasis data dapat diintegrasikan ke dalam proses, nilai budaya, dan strategi bisnis organisasi? eropa memiliki rekam jejak dalam upaya penelitian bersama, serta kekuatan dalam menyatukan kebijakan atau menghilangkan hambatan adopsi. Ada peluang untuk memanfaatkan hal ini dan kekuatan eropa lainnya untuk mewujudkan visi di mana data besar berkontribusi menjadikan eropa sebagai negara dengan perekonomian paling kompetitif di dunia pada tahun 2020.

1.3 VISI BIG DATA

Sektor Teknologi Informasi dan Komunikasi (TIK) bertanggung jawab langsung atas 5% PDB eropa, dengan nilai pasar sebesar 660 miliar per tahun, hal ini juga memberikan kontribusi yang signifikan terhadap pertumbuhan produktivitas secara keseluruhan (20 % langsung dari sektor ICT dan 30 % dari investasi ICT). Solusi Big Data dapat berkontribusi untuk meningkatkan daya saing eropa dengan menghadirkan alat, aplikasi, dan layanan yang bernilai tambah. Salah satu perkiraan untuk tahun 2020 menyebutkan potensi data besar dan terbuka untuk meningkatkan PDB Eropa sebesar 1,9%, setara dengan pertumbuhan ekonomi satu tahun penuh di UE (Buchholtz dkk. 2014). International Data Corporation (IDC) memperkirakan bahwa pasar teknologi dan layanan Big Data akan tumbuh pada tingkat pertumbuhan tahunan gabungan (CAGR) sebesar 27% menjadi Rp.32,4 miliar hingga tahun 2017.

Komisi eropa pada bulan maret 2010 meluncurkan strategi eropa 2020 (Komisi Eropa 2010) untuk keluar dari krisis dan mempersiapkan perekonomian UE menghadapi tantangan berikutnya dalam hal produktivitas, ekonomi, dan kohesi sosial. Agenda digital untuk eropa adalah salah satu dari tujuh inisiatif unggulan strategi eropa 2020, hal ini mendefinisikan peran pendukung utama yang harus dimainkan oleh penggunaan TIK jika Eropa ingin berhasil mencapai ambisinya pada tahun 2020. Pentingnya Big Data diakui dengan memasukkan topik

tertentu dalam agenda digital untuk mendapatkan manfaat maksimal dari data yang ada dan khususnya kebutuhan untuk membuka sumber daya data publik untuk digunakan kembali. Seperti yang dinyatakan oleh Komisaris Uni Eropa Kroes, *"Big Data adalah Minyak baru"* yang dapat dikelola, dimanipulasi, dan digunakan dengan cara yang belum pernah ada sebelumnya berkat alat digital berkinerja tinggi, menjadikan Big Data sebagai bahan bakar inovasi.

Transformasi Sektor Industri

Potensi Big Data diperkirakan akan berdampak pada semua sektor, mulai dari layanan kesehatan hingga media, dari energi hingga ritel. Potensi transformasi positif telah teridentifikasi di sejumlah sektor utama.

a) **Layanan Kesehatan**: Pada awal abad ke-21, eropa merupakan negara dengan masyarakat menua yang sangat menuntut infrastruktur layanan kesehatan. Ada kebutuhan mendesak untuk meningkatkan efisiensi sistem layanan kesehatan saat ini agar lebih berkelanjutan. Penerapan Big Data memiliki potensi yang signifikan di sektor ini dengan perkiraan penghematan pengeluaran sebesar Rp.90 miliar dari anggaran layanan kesehatan nasional di UE (Manyika dkk. 2011). Penerapan klinis dari data besar berkisar dari penelitian efektivitas komparatif di mana efektivitas intervensi klinis dan finansial dibandingkan dengan sistem pendukung keputusan klinis generasi berikutnya yang menggunakan kumpulan data kesehatan heterogen yang komprehensif serta analisis operasi klinis yang canggih. Aplikasi litbang layanan kesehatan mencakup pemodelan prediktif, alat statistik, dan algoritme untuk meningkatkan desain uji klinis, pengobatan yang dipersonalisasi, dan analisis pola penyakit.

b) **Sektor Publik**: Sektor publik di eropa menyumbang hampir setengah dari PDB dan dapat memperoleh manfaat yang signifikan dari Big Data untuk mendapatkan efisiensi dalam proses administrasi. Big Data dapat mengurangi biaya kegiatan administratif sebesar 15–20%, sehingga menciptakan nilai baru yang setara dengan Rp.150 miliar hingga Rp.300 miliar (OECD 2013). Potensi manfaat di sektor publik mencakup peningkatan transparansi melalui pemerintahan terbuka dan data terbuka, peningkatan pengadaan publik, peningkatan alokasi pendanaan ke dalam program-program, layanan berkualitas lebih tinggi, peningkatan akuntabilitas sektor publik, dan masyarakat yang lebih berpengetahuan. Yang terutama pada masa depan adalah definisi kebijakan untuk berbagi data di seluruh lembaga pemerintah dan untuk memberi informasi kepada masyarakat tentang trade-off antara risiko privasi dan keamanan dari berbagi data dan manfaat yang bisa mereka peroleh. Big Data juga akan mengubah hubungan antara warga negara dan pemerintah dengan memberdayakan masyarakat untuk memahami isu-isu politik dan sosial dengan cara-cara baru yang transparan, memungkinkan mereka untuk terlibat dengan isu-isu lokal, regional, nasional, dan global melalui partisipasi.

c) **Keuangan dan Asuransi**: Ada sejumlah cara bagi perusahaan jasa keuangan untuk mencapai keuntungan bisnis dengan menambang dan menganalisis data. Hal ini mencakup peningkatan layanan pelanggan ritel, deteksi penipuan, dan peningkatan efisiensi operasional. Big Data dapat digunakan untuk mengidentifikasi eksposur

secara real-time di berbagai instrumen keuangan canggih seperti derivatif. Analisis prediktif terhadap data internal dan eksternal menghasilkan pengelolaan yang lebih baik dan proaktif terhadap berbagai masalah mulai dari risiko kredit dan operasional (misalnya risiko penipuan dan reputasi) hingga loyalitas pelanggan dan profitabilitas. Tantangan bagi sektor keuangan adalah bagaimana memanfaatkan keluasan dan kedalaman data yang tersedia untuk memenuhi tuntutan regulator sekaligus menyediakan layanan yang dipersonalisasi bagi pelanggan mereka.

d) **Telekomunikasi, Media, dan Hiburan**: Teknik analisis dan visualisasi Big Data dapat memungkinkan penemuan dan penyampaian konten media secara efektif sehingga memungkinkan pengguna berinteraksi secara dinamis dengan media dan konten baru di berbagai platform. Domain data lokasi pribadi menawarkan potensi penciptaan nilai baru melalui aplikasi, termasuk pengiriman konten berbasis lokasi untuk individu, perutean konten yang dipersonalisasi secara cerdas, telematika otomotif, layanan berbasis lokasi seluler, dan iklan bertarget geografis.

e) **Ritel**: Peluang besar untuk menggunakan teknologi Big Data terletak pada interaksi antara pengecer dan konsumen. Data memainkan peran yang semakin besar ketika konsumen mencari, meneliti, membandingkan, membeli, dan mendapatkan dukungan secara online dan produk yang dijual oleh pengecer semakin banyak menghasilkan jejak data mereka sendiri. Big Data dapat meningkatkan produktivitas dan efisiensi sehingga menghasilkan potensi peningkatan margin operasi pengecer sebesar 60%. Big Data dapat berdampak pada ritel di berbagai bidang seperti pemasaran: penjualan silang, pemasaran berbasis lokasi, analisis perilaku di dalam toko, segmentasi mikro pelanggan, analisis sentimen pelanggan, peningkatan pengalaman konsumen multi-saluran; merchandising: optimalisasi bermacam-macam, optimalisasi harga, optimalisasi penempatan dan desain; operasi: transparansi kinerja, optimalisasi input tenaga kerja; Jaringan pasokan: manajemen inventaris, optimalisasi distribusi dan logistik, menginformasikan negosiasi pemasok; model bisnis baru: layanan perbandingan harga, pasar berbasis web.

f) **Manufaktur**: Sektor manufaktur merupakan pengguna awal TI untuk merancang, membangun, dan mendistribusikan produk. Pabrik pintar generasi berikutnya dengan mesin cerdas dan berjaringan (yaitu Internet of Things, Industri 4.0) akan mengalami peningkatan efisiensi lebih lanjut dalam desain, produksi, dan kualitas produk. Big Data akan memungkinkan pemenuhan kebutuhan pelanggan melalui produk yang tepat sasaran dan distribusi yang efektif. Selain peningkatan efisiensi dan pemeliharaan prediktif, data besar akan memungkinkan model bisnis yang sepenuhnya baru di bidang produksi massal produk-produk individual.

g) **Energi dan Transportasi**: Big Data akan membuka peluang baru dan cara-cara inovatif untuk memantau dan mengendalikan jaringan transportasi dan logistik menggunakan berbagai sumber data dan Internet of Things. Potensi Big Data di sektor transportasi diperkirakan mencapai Rp. 500 Triliun di seluruh dunia dalam bentuk penghematan waktu dan bahan bakar, dengan penghindaran emisi CO2 sebesar 380 megaton (OECD

2013). Digitalisasi sistem energi memungkinkan perolehan data real-time dan beresolusi tinggi melalui smart meter yang dapat dimanfaatkan dalam analisis canggih untuk meningkatkan tingkat efisiensi dalam sisi permintaan dan pasokan jaringan energi. Bangunan pintar dan kota pintar akan menjadi pendorong utama peningkatan efisiensi di sektor energi. Teknologi Big Data di sektor utilitas mempunyai potensi mengurangi emisi CO2 lebih dari 2 gigaton, setara dengan Rp. 79 Miliyar (OECD 2013).

Ekosistem data yang sukses akan menyatukan pemilik data, perusahaan analisis data, profesional data yang terampil, penyedia layanan cloud, perusahaan dari industri pengguna, pemodal ventura, wirausahawan, lembaga penelitian, dan universitas. Ekosistem data yang sukses, yang merupakan ciri utama ekonomi berbasis data, akan membuat para pemangku kepentingan berinteraksi secara lancar dalam pasar tunggal digital, sehingga menghasilkan peluang bisnis, akses yang lebih mudah terhadap pengetahuan, dan permodalan (Komisi Eropa 2014). "Komisi dapat berkontribusi dalam hal ini dengan menyatukan para pemain terkait dan mengarahkan sumber daya keuangan yang tersedia untuk memfasilitasi kolaborasi di antara berbagai pemangku kepentingan dalam ekonomi data Eropa" (DG Connect 2013).

Big Data menawarkan potensi nilai luar biasa yang belum dimanfaatkan bagi banyak sektor namun, tidak ada ekosistem data yang terhubung di Eropa. Seperti yang dijelaskan oleh Komisaris Kroes, *"Fragmentasi ini menyangkut sektor, bahasa, serta perbedaan hukum dan praktik kebijakan antar negara UE"* (Komisi Eropa 2013; Kroes 2013). Selama Konferensi ICT 2013, Komisaris Kroes menyerukan kemitraan publik-swasta eropa dalam Big Data untuk menciptakan ekosistem data eropa yang berhubungan untuk mendorong penelitian dan inovasi seputar data, serta penyerapan data lintas sektor, lintas bahasa, serta layanan dan produk data lintas batas. Ia juga mencatat perlunya memastikan privasi *"Menguasai Big Data berarti menguasai privasi juga"* (Kroes 2013). Agar hal ini dapat terwujud, diperlukan pendekatan interdisipliner untuk menciptakan lingkungan bisnis yang optimal bagi Big Data yang akan mempercepat adopsi di Eropa.

1.4 EKOSISTEM INOVASI BIG DATA

Untuk mendorong inovasi dan daya saing, Eropa perlu mendorong pengembangan dan adopsi teknologi Big Data, kasus penggunaan yang memberikan nilai tambah, dan model bisnis yang berkelanjutan. Meskipun tidak ada ekosistem data yang berhubungan di tingkat Eropa (DG Connect 2013), manfaat dari berbagi dan menghubungkan data antar domain dan industri menjadi jelas. Pendekatan ekosistem memungkinkan organisasi untuk menciptakan nilai baru yang tidak dapat dicapai oleh satu organisasi pun (Adner 2006). Ekosistem Big Data Eropa merupakan faktor penting dalam komersialisasi dan komoditisasi layanan, produk, dan platform Big Data. Dalam ekosistem bisnis yang sehat, perusahaan dapat bekerja sama dalam jaringan bisnis yang kompleks di mana mereka dapat dengan mudah bertukar dan berbagi sumber daya penting. Jika Ekosistem Big Data ingin muncul di Eropa, penting bagi berbagai aktor dalam ekosistem tersebut untuk mendefinisikan visi bersama dan bersama-sama mengidentifikasi kesenjangan dalam lanskap data saat ini. Ekosistem Big Data yang sukses akan membuat semua pemangku kepentingan berinteraksi secara lancar dalam pasar tunggal

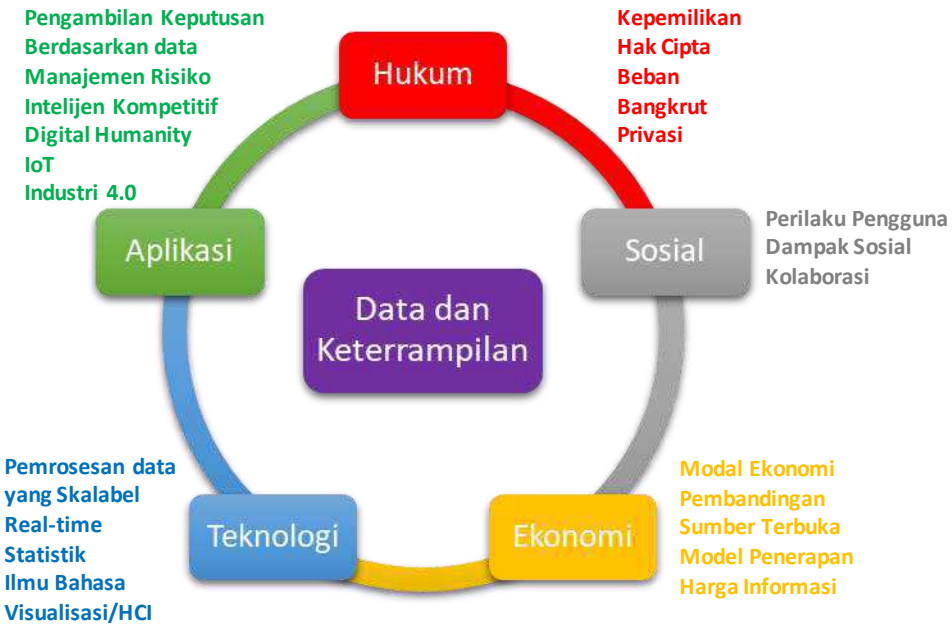
digital, yang mengarah pada peluang bisnis, akses lebih mudah terhadap pengetahuan, dan modal

Dimensi Ekosistem Big Data Eropa

Penggunaan dan pemahaman Big Data sebagai aset ekonomi secara efisien membawa potensi besar bagi perekonomian dan masyarakat UE. Tantangan dalam membangun ekosistem Big Data di eropa telah didefinisikan ke dalam serangkaian dimensi utama, seperti yang diilustrasikan pada Gambar 1.1. Eropa harus mengatasi berbagai tantangan ini untuk mendorong pengembangan ekosistem Big Data.

1) **Data**: Ketersediaan dan akses terhadap data akan menjadi dasar ekosistem yang berpusat pada data. Ekosistem data yang sehat akan terdiri dari beragam tipe data yang berbeda: data terstruktur, tidak terstruktur, multibahasa, dihasilkan mesin dan sensor, data statis, dan real-time. Data dalam ekosistem harus berasal dari berbagai sektor, termasuk layanan kesehatan, energi, ritel, dan dari sumber publik dan swasta. Nilai dapat dihasilkan dengan berbagai cara, dengan memperoleh data, menggabungkan data dari berbagai sumber dan lintas sektor, menyediakan akses latensi rendah, meningkatkan kualitas data, memastikan integritas data, memperkaya data, mengekstraksi wawasan, dan menjaga privasi.
2) **Keterampilan**: Tantangan penting bagi Eropa adalah memastikan ketersediaan pekerja terampil di ekosistem data. Ekosistem yang aktif akan membutuhkan ilmuwan dan insinyur data yang memiliki keahlian di bidang analitik, statistik, pembelajaran mesin, penambangan data, dan manajemen data. Pakar teknis perlu dipadukan dengan pakar bisnis yang paham data dengan pengetahuan domain yang kuat dan kemampuan menerapkan pengetahuan data mereka dalam organisasi untuk menciptakan nilai.
3) **Hukum**: Lingkungan peraturan yang sesuai diperlukan untuk memfasilitasi pengembangan pasar data besar di eropa. Kejelasan hukum diperlukan mengenai isu-isu seperti kepemilikan data, penggunaan, perlindungan, privasi, keamanan, tanggung jawab, kejahatan dunia maya, hak kekayaan intelektual, dan implikasi dari kebangkrutan dan kebangkrutan.
4) **Teknis**: Tantangan teknis utama yang perlu diatasi termasuk akuisisi data berskala besar dan heterogen, penyimpanan data yang efisien, pemrosesan data dan analisis data real-time yang masif, kurasi data, pengambilan dan visualisasi data tingkat lanjut, antarmuka pengguna yang intuitif, interoperabilitas dan menghubungkan data , informasi, dan konten. Semua topik ini perlu dikembangkan untuk mempertahankan atau mengembangkan keunggulan kompetitif.
5) **Penerapan**: Big Data berpotensi mentransformasi banyak sektor dan domain termasuk sektor kesehatan, sektor publik, keuangan, energi, dan transportasi. Aplikasi dan solusi inovatif yang berbasis nilai harus dikembangkan, divalidasi, dan diterapkan dalam ekosistem Big Data jika eropa ingin menjadi pemimpin dunia.
6) **Bisnis**: Ekosistem Big Data dapat mendukung transformasi sektor bisnis yang ada dan pengembangan start-up baru dengan model bisnis inovatif untuk mendorong pertumbuhan lapangan kerja dan aktivitas ekonomi.

7) **Sosial**: Penting untuk meningkatkan kesadaran akan manfaat Big Data bagi dunia usaha, sektor publik, dan masyarakat. Big Data akan memberikan solusi bagi tantangan-tantangan sosial yang besar di eropa, seperti peningkatan efisiensi dalam layanan kesehatan, peningkatan kelayakan hidup di kota-kota, peningkatan transparansi dalam pemerintahan, dan peningkatan keberlanjutan.



Gambar 1.1 Dimensi Ekosistem Nilai Big Data

Ringkasan

Big Data adalah salah satu aset ekonomi utama masa depan. Menguasai potensi teknologi Big Data dan memahami potensinya untuk mentransformasi sektor industri akan meningkatkan daya saing perusahaan-perusahaan eropa dan menghasilkan pertumbuhan ekonomi dan lapangan kerja. Eropa memerlukan strategi yang jelas untuk meningkatkan daya saing industri eropa guna mendorong inovasi. Eropa perlu mendorong pengembangan dan adopsi teknologi Big Data secara luas, penggunaan nilai tambah, dan model bisnis berkelanjutan melalui ekosistem Big Data. Investasi strategis diperlukan baik oleh sektor publik maupun swasta untuk memungkinkan Eropa menjadi pemimpin dalam ekonomi digital berbasis data global dan untuk memperoleh manfaat yang ditawarkan dengan terciptanya ekosistem Big Data eropa.

BAB 2
PROYEK BESAR

2.1 PENDAHULUAN

Proyek Big Data Public Private Forum (BIG) adalah tindakan koordinasi dan dukungan UE untuk menyediakan peta jalan bagi Big Data di eropa. Proyek BIG berupaya untuk mendefinisikan dan menerapkan strategi Big Data yang jelas yang menangani aktivitas-aktivitas penting yang diperlukan dalam penelitian dan inovasi, adopsi teknologi, dan dukungan yang diperlukan dari Komisi Eropa yang diperlukan untuk keberhasilan penerapan ekonomi Big Data. Sebagai bagian dari strategi ini, hasil proyek digunakan sebagai masukan untuk Horizon 2020.

Teknologi penelitian dasar dan aplikasi sektoral yang inovatif dianalisis dan dinilai dalam proyek BIG untuk menciptakan peta jalan teknologi dan strategi sehingga komunitas bisnis dan operasional memahami potensi teknologi Big Data dan mampu menerapkan strategi dan teknologi yang tepat untuk keuntungan komersial.

Bab ini memberikan gambaran umum proyek besar yang merinci misi dan tujuan strategis proyek. Bab ini menjelaskan mitra dalam konsorsium dan keseluruhan struktur pekerjaan proyek. Metodologi tiga tahap yang digunakan dalam proyek ini dijelaskan, termasuk rincian teknik yang digunakan dalam kelompok kerja teknis, bentuk sektoral, dan kegiatan pemetaan jalan. Terakhir, peran proyek dalam membentuk Kemitraan Pemerintah Swasta dan Asosiasi Nilai Big Data yang bersifat kontrak Horizon 2020 Big Data Value dibahas.

2.2 MISI PROYEK

Untuk mewujudkan visi masyarakat berbasis data pada tahun 2020, eropa harus mempersiapkan ekosistem yang tepat seputar Big Data. Organisasi publik dan swasta perlu memiliki infrastruktur dan teknologi yang diperlukan untuk menghadapi kompleksitas data besar, namun juga harus mampu menggunakan data untuk memaksimalkan daya saing mereka dan memberikan nilai bisnis.

Membangun komunitas industri seputar Big Data di eropa merupakan prioritas utama proyek BIG, bersamaan dengan menyiapkan infrastruktur kolaborasi dan diseminasi yang diperlukan untuk menghubungkan pemasok teknologi, integrator, dan organisasi pengguna terkemuka. Proyek BIG berupaya untuk mendefinisikan dan menerapkan strategi yang mencakup penelitian dan inovasi, serta adopsi teknologi. Pembentukan komunitas bersama dengan sumber daya yang memadai untuk bekerja di semua tingkatan (teknis, bisnis, politik, dll.), merupakan dasar bagi strategi jangka panjang eropa. Yakin bahwa diperlukan reaksi yang kuat, BIG mendefinisikan misinya sebagai berikut:

Misi BIG adalah membangun ekosistem yang akan menyatukan semua pemangku kepentingan terkait yang diperlukan untuk mewujudkan masyarakat berbasis data pada tahun 2020. Ekosistem ini akan memastikan bahwa Eropa memainkan peran utama dalam mendefinisikan konteks baru dengan membangun infrastruktur yang diperlukan. dan teknologi, menghasilkan ruang inovasi yang sesuai di mana semua organisasi mendapat

manfaat dari data, dan menyediakan kerangka kerja pan-Eropa untuk secara hubungan mengatasi hambatan kebijakan, peraturan, hukum, dan keamanan. Misi besar dipecah menjadi sejumlah tujuan strategis khusus untuk proyek tersebut.

2.3 TUJUAN STRATEGIS

Pada bulan September 2012, proyek ini mengidentifikasi serangkaian tujuan strategis untuk memastikan misinya tercapai. Tujuan spesifiknya adalah:

a) BIG akan membentuk inisiatif berbasis industri seputar manajemen informasi cerdas dan Big Data untuk berkontribusi terhadap daya saing UE dan menempatkannya dalam Horizon 2020: Kepemimpinan industri akan memandu tindakan menuju manfaat bisnis yang nyata, namun akan dilengkapi dengan pandangan dari akademisi dan organisasi penelitian, yang juga akan mengambil bagian dalam upaya ini. Proyek ini akan mengambil pendekatan jangka panjang untuk mewakili pandangan dan kepentingan para pemangku kepentingan IIM, dengan fokus khusus pada Big Data karena relevansinya dalam konteks saat ini dan masa depan. Keputusan seperti menetapkannya sebagai badan hukum akan dipertimbangkan, dan potensi merger dengan asosiasi terkait di tingkat UE juga akan dipertimbangkan demi keberlanjutan dan dampak.

b) BIG akan menguraikan peta jalan terpadu yang mempertimbangkan aspek teknis, bisnis, kebijakan, dan masyarakat, dengan fokus tidak hanya pada masalah teknis semata, namun juga menetapkan prioritas berdasarkan dampak yang diharapkan. Konsorsium BIG akan melibatkan keahlian yang diperlukan untuk memastikan kontribusi tidak hanya dari mitra proyek, namun juga dari komunitas yang lebih luas yang terdiri dari para ahli di bidang teknis yang relevan serta para ahli di sektor atau bidang aplikasi di mana penggunaan teknologi ini diharapkan menghasilkan dampak tinggi.

c) BIG akan memastikan bahwa bidang penelitian teknis yang dipilih oleh proyek ini mencakup kebutuhan yang diungkapkan oleh industri dalam domain aplikasi yang berbeda: Agar hal ini dapat terwujud, diperlukan pemahaman yang tajam tentang bagaimana Big Data dapat diterapkan dalam sektor industri. Pemahaman ini perlu disampaikan kepada para ahli di bidangnya untuk menentukan jalur yang jelas dalam penerapan teknologi di setiap sektor yang dipilih.

d) BIG akan mendorong adopsi teknologi Big Data gelombang awal: Daripada hanya mengadopsi pendekatan futuristik, BIG akan menggunakan teknologi yang sudah ada sebagai titik awal. Tujuannya adalah untuk mencapai pemahaman yang jelas tentang tingkat kematangan berbagai solusi teknis serta kelayakan penerapannya. Hal ini akan menjadi informasi berharga sehubungan dengan perkembangan terkini dan akan digunakan sebagai masukan untuk penjabaran peta jalan sektoral dan terpadu.

e) BIG akan mendefinisikan dan mendorong tindakan yang berhubungan dengan kebijakan dan peraturan, termasuk aspek-aspek seperti keamanan data, kekayaan intelektual, privasi, tanggung jawab, dan akses data. BIG akan berkontribusi pada

seluruh ekosistem terkait implementasi Big Data tanpa membatasi aktivitasnya hanya pada masalah teknis saja.

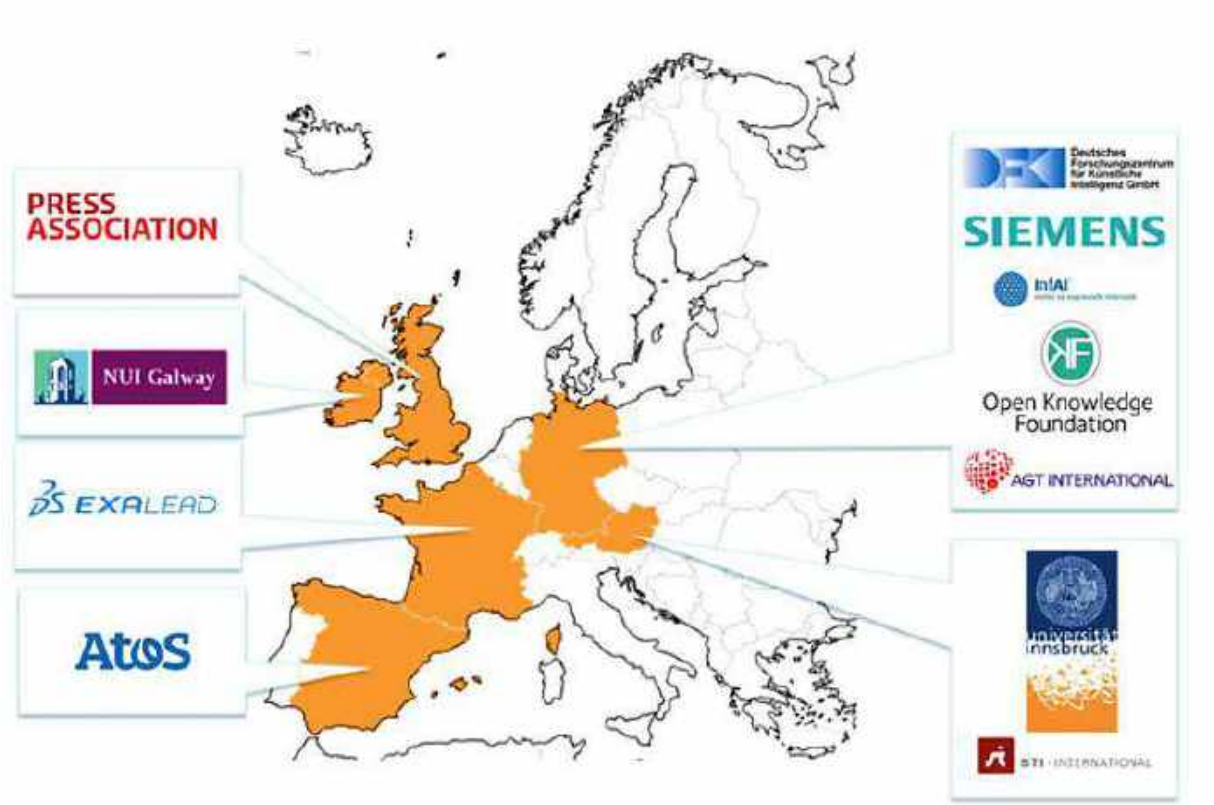
f) BIG akan melaksanakan aksi sosialisasi yang menyasar berbagai pemangku kepentingan dan pemain dalam Jaringan nilai: Aksi sosialisasi akan disesuaikan dengan komunitas yang berbeda (misalnya pakar teknis, ilmuwan data, manajer teknis, manajer bisnis, dan eksekutif di Perusahaan Multinasional (MNC)) dan Usaha Kecil dan Menengah (UKM)). BIG menangani semua komunitas yang relevan dengan strategi yang ambisius termasuk kehadiran di media massa, konferensi yang relevan, penyelenggaraan lokakarya dan acara, dan maksimalisasi penggunaan saluran web.

g) Semua ini tidak akan mungkin terjadi tanpa penyediaan infrastruktur kolaborasi yang tepat. Kolaborasi antar proyek, serta banyak diskusi antara seluruh pemangku kepentingan dan pelaku terkait dalam Jaringan nilai, termasuk organisasi industri besar di lanskap UE, akan dilakukan. Mengingat hal ini BIG menyiapkan dan memelihara infrastruktur pendukung yang memungkinkan kolaborasi, berbagi informasi, dan penyesuaian tindakan terhadap audiens sasaran yang berbeda.

2.4 KONSORSIUM

Para peserta konsorsium BIG (diilustrasikan pada Gambar 2.1), dipilih dengan cermat untuk memasukkan pemain-pemain kunci dengan keterampilan yang saling melengkapi baik di industri maupun akademisi. Masing-masing mitra proyek memiliki pengalaman dalam proyek-proyek eropa yang modern dan koneksi yang signifikan dengan pemangku kepentingan utama di pasar data besar. Mitra akademis yang menggunakan pengetahuan ahli mereka di lapangan memimpin penyelidikan teknis teknologi data besar. Mitra industri memiliki pengetahuan yang baik tentang produk dan layanan manajemen data berskala besar serta penerapannya dalam berbagai sektor industri. Mitra konsorsium BIG adalah:

- **Industri**: Atos, Press Association (PA), Siemens, AGT International, Exalead, dan Open Knowledge Foundation (OKF)
- **Akademisi**: Universitas Innsbruck (UIBK), Universitas Nasional Irlandia Galway (NUIG), Universitas Leipzig, Pusat Penelitian Kecerdasan Buatan Jerman (DFKI), dan STI Internasional



Gambar 2.1 Anggota konsorsium proyek BESAR

2.5 KETERLIBATAN PEMANGKU KEPENTINGAN

Hal yang penting bagi keberhasilan upaya pemupukan silang dan pemetaan jalan yang luas adalah keterlibatan sebagian besar masyarakat dan industri, tidak hanya dari sudut pandang penyediaan teknologi tetapi juga adopsi teknologi. Proyek ini mengambil pendekatan inklusif terhadap keterlibatan pemangku kepentingan dan secara aktif meminta masukan dari masyarakat luas yang terdiri dari para ahli di bidang teknis serta para ahli di sektor bisnis. Filosofi terbuka diterapkan pada semua dokumen yang dihasilkan oleh proyek, yang dipublikasikan kepada komunitas luas untuk kontribusi aktif dan validasi konten. Proyek ini mengadakan lokakarya pemangku kepentingan untuk melibatkan masyarakat dalam proyek tersebut. Lokakarya pertama diadakan pada pean Data Forum (EDF) 2013 di Dublin untuk mengumumkan proyek tersebut kepada masyarakat dan mengumpulkan peserta. Lokakarya kedua berlangsung di EDF 2014 di Athena untuk menyajikan hasil sementara proyek untuk mendapatkan masukan dan validasi lebih lanjut dengan para pemangku kepentingan. Selama proyek berlangsung, sejumlah lokakarya khusus sektoral yang dihadiri banyak orang diadakan untuk mengumpulkan kebutuhan dan memvalidasi temuan. Di akhir proyek, lokakarya akhir diadakan untuk mempresentasikan hasil proyek pada bulan Oktober 2014 di Heidelberg.

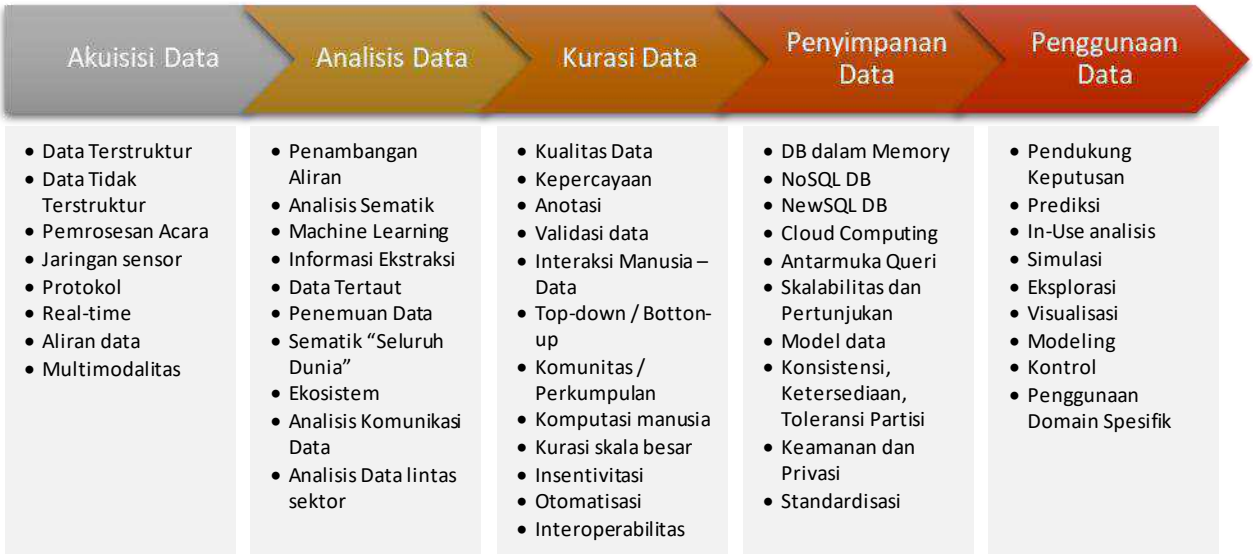
2.6 STRUKTUR PROYEK

Pekerjaan proyek BIG dibagi menjadi beberapa kelompok yang berfokus pada sektor industri dan bidang teknis. Struktur proyek terdiri dari forum sektoral dan kelompok kerja teknis. Forum sektoral mengkaji bagaimana teknologi Big Data dapat memungkinkan inovasi

dan transformasi bisnis di berbagai sektor. Forum sektoral dipimpin oleh mitra industri proyek. Tujuan mereka adalah mengumpulkan kebutuhan Big Data dari sektor industri vertikal, termasuk kesehatan, sektor publik, keuangan, asuransi, telekomunikasi, media, hiburan, manufaktur, ritel, energi, dan transportasi (lihat Gambar 2.2).

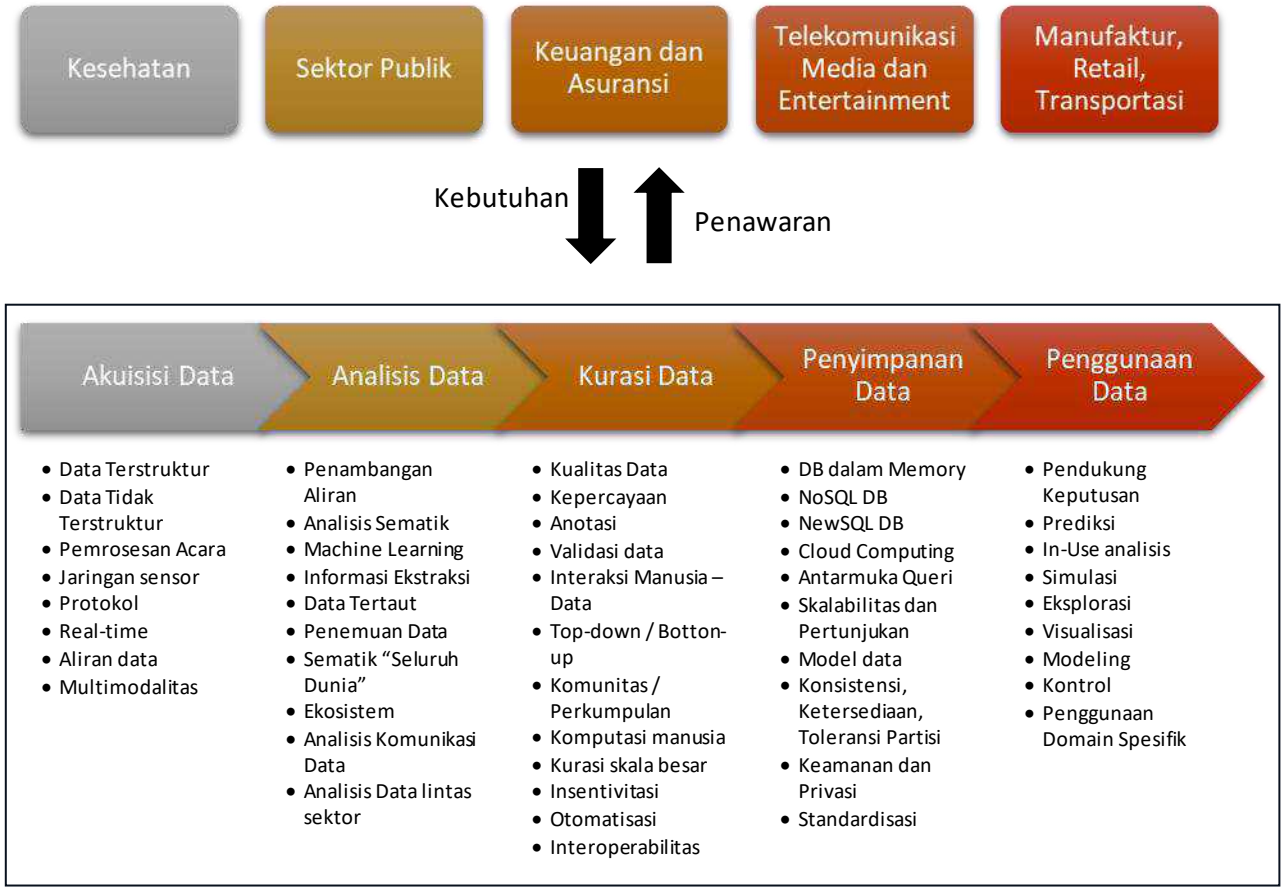
Gambar 2.2 Forum sektoral dalam proyek BIG

Kelompok kerja teknis berfokus pada teknologi Big Data untuk setiap aktivitas dalam Jaringan nilai data untuk memeriksa kemampuan, tingkat kematangan, kejelasan, pemahaman, dan kesesuaian penerapannya. Kelompok teknis (lihat Gambar 2.3) dipimpin oleh mitra akademis di BIG dan mengkaji tren teknologi dan penelitian yang muncul untuk mengatasi Big Data.



Gambar 2.3 Kelompok kerja teknis dalam proyek BIG

Seperti diilustrasikan pada Gambar 2.4, kebutuhan yang diidentifikasi oleh forum sektor digunakan untuk memahami kematangan dan kesenjangan dalam kemampuan yang ditawarkan oleh teknologi Big Data saat ini. Analisis ini memberikan gambaran jelas mengenai keterbatasan dan harapan terkait penerapan teknologi Big Data. Hasil analisis digunakan untuk menghasilkan serangkaian peta jalan yang mencerminkan konsensus yang menentukan prioritas dan tindakan yang diperlukan untuk Big Data di setiap sektor.

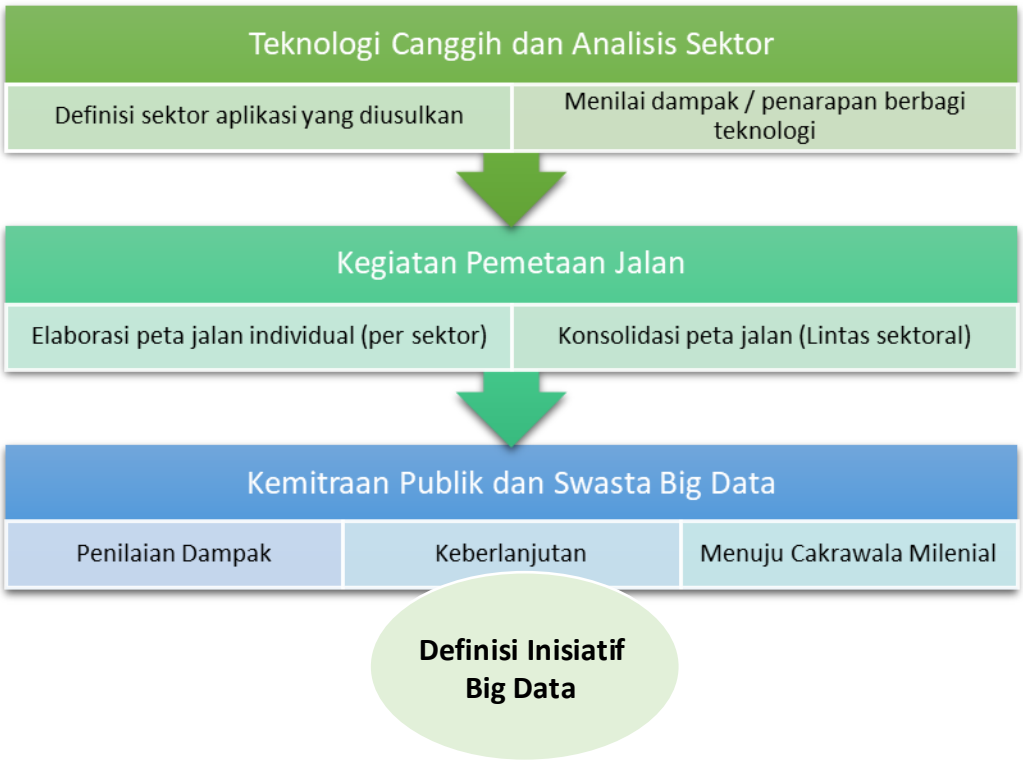


Gambar 2.4 Struktur proyek BESAR

2.7 METODOLOGI

Dari sudut pandang operasional, BIG mendefinisikan serangkaian kegiatan berdasarkan pendekatan tiga fase seperti yang diilustrasikan pada Gambar 2.5. Ketiga fase tersebut adalah:

1. Teknologi canggih dan analisis sektoral
2. Kegiatan pemetaan jalan
3. Kemitraan publik-swasta Big Data



Gambar 2.5 Metodologi tiga fase BIG

Teknologi Terkini dan Analisis Sektor

Pada tahap pertama proyek, forum sektoral dan kelompok kerja teknis melakukan penyelidikan paralel untuk mengidentifikasi:

- Kebutuhan dan persyaratan sektoral yang dikumpulkan dari berbagai pemangku kepentingan
- Kecanggihan teknologi Big Data serta identifikasi tantangan penelitian

Sebagai bagian dari investigasi, sektor aplikasi mengungkapkan kebutuhan mereka sehubungan dengan teknologi serta kemungkinan keterbatasan dan harapan terkait penerapannya saat ini dan di masa depan.

Berdasarkan hasil investigasi tersebut dilakukan gap analysis antara kemampuan teknologi apa yang sudah siap, dengan ekspektasi sektoral mengenai kemampuan teknologi apa yang dibutuhkan saat ini serta kebutuhan di masa depan. Analisis ini menghasilkan serangkaian peta jalan sektoral yang mencerminkan konsensus dan menentukan prioritas dan tindakan untuk memandu langkah lebih lanjut dalam penelitian Big Data.

Kelompok Kerja Teknis

Tujuan dari kelompok kerja teknis ini adalah untuk menyelidiki kecanggihan teknologi Big Data untuk menentukan tingkat kematangan, kejelasan, pemahaman, dan kesesuaian untuk implementasi. Untuk memungkinkan dilakukannya investigasi ekstensif dan pemetaan perkembangan secara rinci, kelompok kerja teknis menerapkan kombinasi pendekatan top-down dan bottom-up, dengan fokus pada pendekatan bottom-up. Pendekatan kelompok kerja didasarkan pada pendekatan 4 langkah:

1. Penelitian Literatur
2. Wawancara Ahli Materi Pelajaran
3. Lokakarya Pemangku Kepentingan
4. Survei Teknis.

Pada langkah pertama setiap kelompok kerja teknis melakukan tinjauan literatur sistematis berdasarkan kegiatan berikut:

1. Identifikasi jenis dan sumber informasi yang relevan
2. Analisis informasi penting di setiap sumber
3. Identifikasi topik-topik utama untuk setiap kelompok kerja teknis
4. Identifikasi ahli pokok bahasan utama untuk setiap topik sebagai calon wawancara potensial.
5. Menyatukan pesan utama dari setiap sumber data ke dalam deskripsi canggih untuk setiap topik yang teridentifikasi

Para ahli dalam konsorsium menguraikan titik awal awal untuk setiap bidang teknis, dan topiknya diperluas melalui penelusuran literatur dan wawancara ahli pokok bahasan.

Jenis sumber data berikut ini digunakan: makalah ilmiah yang diterbitkan dalam lokakarya, simposium, konferensi, jurnal dan majalah, kertas putih perusahaan, situs web vendor teknologi, proyek sumber terbuka, majalah online, data analis, blog web, sumber online lainnya, dan wawancara yang dilakukan oleh konsorsium BIG. Kelompok-kelompok tersebut berfokus pada sumber-sumber yang menyebutkan teknologi konkrit dan menganalisisnya sehubungan dengan nilai dan manfaatnya.

Langkah sintesis membandingkan pesan-pesan utama dan mengekstraksi pandangan-pandangan yang disepakati yang kemudian dirangkum dalam kertas putih teknis. Topik-topik diprioritaskan berdasarkan sejauh mana topik-topik tersebut mampu menjawab kebutuhan dunia usaha sebagaimana diidentifikasi oleh kelompok kerja forum sektoral.

Survei literatur dilengkapi dengan serangkaian wawancara dengan para ahli di bidang topik yang relevan. Wawancara ahli materi pokok adalah teknik yang cocok untuk pengumpulan data dan khususnya untuk penelitian eksplorasi karena memungkinkan diskusi luas yang menjelaskan faktor-faktor penting. Informasi yang dikumpulkan kemungkinan besar lebih akurat dibandingkan informasi yang dikumpulkan dengan metode lain karena pewawancara dapat menghindari jawaban yang tidak akurat atau tidak lengkap dengan menjelaskan pertanyaan kepada orang yang diwawancarai (Oppenheim 1992).

Wawancara mengikuti protokol semi-terstruktur. Topik wawancara mencakup berbagai aspek Big Data, dengan fokus pada:

1. Tujuan teknologi Big Data
2. Penerima manfaat teknologi Big Data
3. Pendorong dan hambatan bagi teknologi Big Data
4. Teknologi dan standar untuk teknologi Big Data

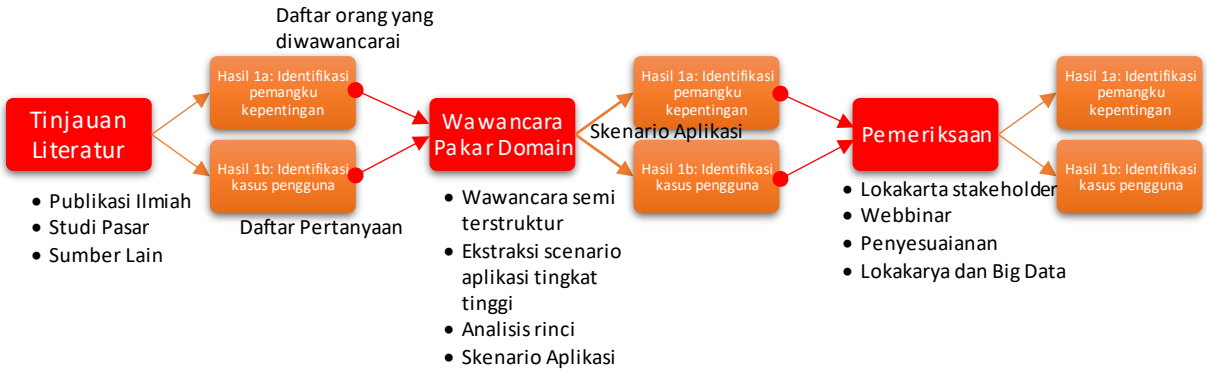
Sekelompok orang yang diwawancarai diidentifikasi dari survei literatur, kontak dalam konsorsium, dan penelusuran yang lebih luas pada ekosistem Big Data. Orang yang diwawancarai dipilih untuk mewakili berbagai pemangku kepentingan dalam ekosistem Big

Data. Pemilihan narasumber meliputi (1) penyedia teknologi Big Data yang sudah mapan (biasanya MNC), (2) pemain sektoral inovatif yang berhasil memanfaatkan Big Data, (3) UKM baru dan yang sedang berkembang di bidang Big Data, dan (4) otoritas akademis terkemuka dunia di bidang teknis yang terkait dengan Jaringan Nilai Big Data.

Forum Sektoral

Tujuan keseluruhan dari forum sektoral ini adalah untuk memperoleh pemahaman mendalam tentang bagaimana teknologi Big Data dapat digunakan di berbagai sektor industri, seperti layanan kesehatan, publik, keuangan dan asuransi, serta media.

Untuk mengidentifikasi kebutuhan pengguna dan kebutuhan industri di setiap domain, forum sektoral mengikuti metodologi penelitian yang mencakup tiga langkah berikut seperti yang diilustrasikan pada Gambar 2.6. Untuk setiap sektor industri, langkah-langkah tersebut dilakukan secara terpisah. Namun, jika sektor-sektor tersebut terkait (seperti energi dan transportasi), hasilnya akan digabungkan untuk sektor-sektor tersebut guna menyoroti perbedaan dan persamaan.



Gambar 2.6 Tiga langkah metodologi penelitian forum sektoral

Tujuan dari langkah pertama adalah untuk mengidentifikasi pemangku kepentingan dan kasus penggunaan aplikasi Big Data di berbagai sektor. Oleh karena itu, survei dilakukan termasuk tinjauan ilmiah, studi pasar, dan sumber Internet lainnya. Pengetahuan ini memungkinkan forum sektoral untuk mengidentifikasi dan memilih mitra wawancara potensial dan memandu pengembangan kuesioner untuk wawancara ahli di bidangnya.
Kuesioner terdiri dari hingga 12 pertanyaan yang dikelompokkan menjadi tiga bagian:

a. Penyelidikan langsung terhadap kebutuhan spesifik pengguna
b. Evaluasi tidak langsung terhadap kebutuhan pengguna dengan mendiskusikan relevansi kasus penggunaan yang diidentifikasi pada Langkah 1 serta aplikasi Big Data lainnya yang mereka ketahui
c. Meninjau kendala-kendala yang perlu diatasi untuk mendorong penerapan aplikasi Big Data di setiap sektor

Pada langkah kedua, wawancara semi terstruktur dilakukan dengan menggunakan kuesioner yang dikembangkan. Setidaknya satu perwakilan dari setiap kelompok pemangku kepentingan yang diidentifikasi pada Langkah 1 diwawancarai. Untuk memperoleh kebutuhan pengguna

dari materi yang dikumpulkan, kasus penggunaan yang paling relevan dan sering disebutkan dikumpulkan ke dalam skenario aplikasi tingkat tinggi. Strategi pengumpulan dan analisis data terinspirasi oleh pendekatan triangulasi (Flick 2004). Meninjau dan menilai secara kuantitatif skenario aplikasi tingkat tinggi menghasilkan analisis kebutuhan pengguna yang andal. Pemeriksaan terhadap kemungkinan kendala penerapan Big Data membantu mengidentifikasi persyaratan relevan yang perlu ditangani.

Langkah ketiga melibatkan pemeriksaan silang dan validasi hasil awal dari dua langkah pertama dengan melibatkan pemangku kepentingan domain tersebut. Beberapa sektor mengadakan lokakarya dan webinar khusus dengan pemangku kepentingan industri untuk mendiskusikan dan meninjau hasilnya. Hasil lokakarya dipelajari dan diintegrasikan bila diperlukan.

Pemetaan Jalan Lintas Sektoral

Perbandingan antar sektor memungkinkan identifikasi persamaan dan perbedaan di berbagai tingkat, termasuk teknis, kebijakan, bisnis, dan peraturan. Analisis ini digunakan untuk menentukan peta jalan lintas sektoral terpadu yang memberikan pandangan holistik yang berhubungan mengenai domain Big Data. Peta jalan Big Data lintas sektoral ditentukan menggunakan tiga langkah berikut:

1. Konsolidasi untuk membangun pemahaman umum tentang persyaratan serta deskripsi teknologi dan istilah yang digunakan di seluruh domain
2. Pemetaan untuk mengidentifikasi teknologi apa pun yang diperlukan untuk memenuhi kebutuhan lintas sektor yang teridentifikasi
3. Penyesuaianan temporal untuk menyoroti teknologi mana yang perlu tersedia dan pada titik waktu tertentu dengan memasukkan perkiraan tingkat adopsi oleh pemangku kepentingan yang terlibat

Sisa bagian ini menjelaskan masing-masing langkah ini secara lebih rinci.

Konsolidasi

Penyesuaianan antar kelompok kerja teknis, dan antara kelompok kerja teknis dan forum sektoral, merupakan hal yang penting dan difasilitasi melalui pertukaran awal rancangan, pertemuan tatap muka, dan pengumpulan persyaratan yang terkonsolidasi melalui SF. Untuk menyelaraskan pelabelan persyaratan yang spesifik pada sektor tertentu, deskripsi gabungan telah dibuat. Dalam melakukan hal ini, masing-masing sektor menyediakan persyaratannya dengan kebutuhan pengguna yang terkait. Dalam pertemuan khusus, persyaratan serupa dan terkait dikumpulkan dan kemudian digabungkan, diselaraskan, atau direstrukturisasi. Dengan demikian, daftar awal dari 13 persyaratan tingkat tinggi dan 28 persyaratan sub-tingkat dapat dikurangi menjadi 8 persyaratan tingkat tinggi dan 25 persyaratan sub-tingkat. Singkatnya fase konsolidasi mengurangi jumlah persyaratan sebesar 20%.

Pemetaan

Untuk memetakan teknologi ke persyaratan, kelompok kerja teknis menunjukkan teknologi mana yang dapat digunakan untuk memenuhi persyaratan gabungan. Selain

memberikan pemetaan antara persyaratan dan teknologi, kelompok kerja teknis juga menunjukkan tantangan penelitian terkait.

Dalam lokakarya 1 hari, pemetaan awal teknologi dan persyaratan dikonsolidasikan dalam dua langkah. Pertama, kemampuan teknologi yang ditunjukkan dianalisis secara lebih rinci dengan menjelaskan bagaimana aspek spesifik sektor dari setiap kebutuhan lintas sektor dapat ditangani. Kedua, untuk setiap persyaratan lintas sektor, diselidiki apakah teknologi dari berbagai kelompok kerja teknis perlu digabungkan untuk memenuhi seluruh cakupan persyaratan. Di akhir diskusi, teknologi apa pun yang diminta oleh setidaknya dua sektor dimasukkan ke dalam peta jalan lintas sektor.

Penyesuaianan Temporal

Setelah mengidentifikasi teknologi-teknologi utama, penyesuaian temporalnya perlu ditentukan. Hal ini dicapai dengan menjawab dua pertanyaan:

- Berapa lama waktu pengembangan teknologi tersebut?
- Kapan pemangku kepentingan yang terlibat akan mengadopsi teknologi ini?

Waktu pengembangan untuk setiap teknologi menunjukkan berapa banyak waktu yang dibutuhkan untuk menyelesaikan tantangan penelitian terkait. Kerangka waktu ini bergantung pada kompleksitas teknis dari tantangan tersebut serta sejauh mana perluasan yang spesifik pada sektor tertentu diperlukan. Untuk menentukan tingkat adopsi teknologi Big Data (atau kasus penggunaan terkait) diperlukan persyaratan non-teknis seperti ketersediaan kasus bisnis, struktur insentif yang sesuai, kerangka hukum, potensi manfaat, serta total biaya untuk semua teknologi Big Data. pemangku kepentingan yang terlibat (Adner 2012) dipertimbangkan.

2.8 KEMITRAAN PEMERINTAH SWASTA BIG DATA

Forum Publik Swasta Big Data, demikian sebutan awalnya, dimaksudkan untuk menciptakan jalan menuju implementasi peta jalan. Jalur ini memerlukan dua elemen utama: (1) mekanisme untuk memasukkan isi peta jalan ke dalam agenda nyata yang didukung oleh sumber daya yang diperlukan (investasi ekonomi dari pemangku kepentingan publik dan swasta) dan (2) masyarakat yang tertarik dengan topik tersebut dan berkomitmen untuk mewujudkannya. investasi dan berkolaborasi dalam implementasi agenda.

Konsorsium BIG yakin bahwa untuk mencapai hasil ini diperlukan kesadaran dan komitmen yang luas di luar proyek. BIG mengambil langkah-langkah yang diperlukan untuk menghubungi pemain-pemain besar dan bekerja sama dengan platform teknologi eropa NESSI untuk bersama-sama berupaya mewujudkan upaya ini. Kolaborasi ini dimulai pada musim panas tahun 2013 dan memungkinkan mitra BIG untuk membangun koneksi tingkat tinggi yang diperlukan baik di tingkat industri maupun politik. Kolaborasi ini membuahkan hasil sebagai berikut:

1 Agenda Riset & Inovasi Strategis (SRIA) mengenai nilai Big Data yang awalnya didasari oleh makalah teknis dan peta jalan BIG dan telah diperluas dengan masukan dari konsultasi publik yang mencakup ratusan pemangku kepentingan tambahan yang mewakili sisi penawaran dan permintaan .

2 Proposal cPPP (kontrak KPS) sebagai langkah formal untuk membentuk KPS mengenai nilai Big Data. Proposal cPPP dibangun berdasarkan SRIA dengan menambahkan elemen konten tambahan seperti instrumen potensial yang dapat digunakan untuk implementasi agenda.
3 Pembentukan komunitas perwakilan pemangku kepentingan yang telah mendukung SRIA dan menyatakan minat dan komitmen untuk terlibat dalam cPPP. Identifikasi kelompok inti yang dipimpin oleh industri yang siap berkomitmen terhadap tujuan cPPP dengan kemauan untuk menginvestasikan uang dan waktu.
4 Pembentukan badan hukum yang berbasis di Belgia: sebuah organisasi nirlaba bernama *Big Data Value Association* (BDVA) untuk mewakili pihak swasta cPPP. BDVA mempunyai 24 anggota pendiri, termasuk banyak mitra proyek BIG.
5 Dan terakhir, penandatanganan cPPP nilai Big Data antara BDVA dan komisi eropa. CPPP ditandatangani oleh Wakil Presiden Neelie Kroes, yang saat itu menjabat sebagai komisaris agenda digital UE, dan Jan Sundelin, presiden Big Data Value Association (BDVA), pada 13 Oktober 2014 di Brussels. CPPP BDV memberikan kerangka kerja yang menjamin kepemimpinan industri, investasi, dan komitmen pihak swasta dan publik untuk membangun ekonomi berbasis data di seluruh eropa, menguasai pembangkitan nilai dari Big Data dan menciptakan keunggulan kompetitif yang signifikan bagi eropa. industri yang akan meningkatkan pertumbuhan ekonomi dan lapangan kerja.

Ringkasan

Proyek Big Data Public Private Forum (BIG) adalah tindakan koordinasi dan dukungan UE untuk menyediakan peta jalan bagi Big Data di eropa. Proyek BIG berupaya untuk mendefinisikan dan menerapkan strategi Big Data yang jelas yang menangani aktivitas-aktivitas penting yang diperlukan dalam penelitian dan inovasi, adopsi teknologi, dan dukungan yang diperlukan dari Komisi Eropa yang diperlukan untuk keberhasilan penerapan ekonomi Big Data.

Proyek BIG menggunakan metodologi tiga fase dengan kelompok kerja teknis yang mengkaji teknologi dasar, forum sektoral yang mengkaji aplikasi sektoral yang inovatif, dan kegiatan pemetaan jalan untuk menciptakan peta jalan teknologi dan strategi sehingga komunitas bisnis dan operasional memahami potensi teknologi Big Data dan dimungkinkan untuk menerapkan strategi dan teknologi yang tepat untuk keuntungan komersial. Proyek ini merupakan kontributor utama pembentukan Horizon 2020, Big Data Value Association kontrak Public Private Partnership (cPPP) dan Big Data Value Association.

BAB 3
DEFINISI, KONSEP, DAN PENDEKATAN TEORETIS BIG DATA

3.1 PENDAHULUAN

Munculnya gelombang baru data dari berbagai sumber, seperti Internet of Things, Sensor Networks, Open Data on the Web, data dari aplikasi seluler, data jaringan sosial, serta pertumbuhan alami kumpulan data di dalam organisasi, menciptakan permintaan akan strategi pengelolaan data baru yang dapat mengatasi skala lingkungan data yang baru ini. Big Data adalah bidang yang sedang berkembang di mana teknologi inovatif menawarkan cara-cara baru untuk menggunakan kembali dan mengekstraksi nilai dari informasi. Kemampuan untuk mengelola informasi dan mengekstrak pengetahuan secara efektif kini dipandang sebagai keunggulan kompetitif utama, dan banyak organisasi membangun bisnis inti mereka berdasarkan kemampuan mereka mengumpulkan dan menganalisis informasi untuk mengekstrak pengetahuan dan wawasan bisnis. Adopsi teknologi Big Data dalam sektor industri bukanlah suatu kemewahan namun merupakan kebutuhan penting bagi sebagian besar organisasi untuk mendapatkan keunggulan kompetitif.

Bab ini membahas definisi dan konsep terkait Big Data. Bab ini dimulai dengan mengeksplorasi berbagai definisi *"Big Data"* yang muncul selama beberapa tahun terakhir untuk memberi label pada data dengan atribut berbeda. Jaringan nilai Big Data diperkenalkan untuk menggambarkan aliran informasi dalam sistem Big Data sebagai serangkaian langkah yang diperlukan untuk menghasilkan nilai dan wawasan berguna dari data. Bab ini mengeksplorasi konsep ekosistem, asal usulnya dari komunitas bisnis, dan bagaimana konsep tersebut dapat diperluas ke konteks Big Data. Pemangku kepentingan utama dalam ekosistem Big Data diidentifikasi beserta tantangan-tantangan yang perlu diatasi untuk mewujudkan ekosistem Big Data di eropa.

3.2 DEFINISI BIG DATA

Selama beberapa tahun terakhir, istilah *"Big Data"* digunakan oleh berbagai pemain besar untuk memberi label data dengan atribut berbeda. Beberapa definisi Big Data telah diajukan selama dekade terakhir, lihat Tabel 3.1. Definisi pertama, oleh Doug Laney dari META Group (kemudian diakuisisi oleh Gartner), mendefinisikan Big Data menggunakan perspektif tiga dimensi: *"Big Data adalah aset informasi bervolume tinggi, berkecepatan tinggi, dan/atau sangat beragam yang memerlukan bentuk-bentuk baru. pemrosesan untuk memungkinkan peningkatan pengambilan keputusan, penemuan wawasan, dan optimalisasi proses"* (Laney 2001).

Loukides (2010) mendefinisikan Big Data sebagai *"ketika ukuran data itu sendiri menjadi bagian dari masalah dan teknik tradisional dalam menangani data sudah tidak ada lagi"*. Jacobs (2009) menggambarkan Big Data sebagai *"data yang ukurannya memaksa kita untuk melihat lebih jauh dari metode yang lazim digunakan pada saat itu".*

Big Data menyatukan serangkaian tantangan manajemen data untuk bekerja dengan data dalam skala ukuran dan kompleksitas baru. Banyak dari tantangan-tantangan ini bukanlah hal baru. Namun yang baru adalah tantangan yang ditimbulkan oleh karakteristik spesifik Big Data terkait **3V** :

- **Volume (jumlah data)**: menangani data berskala besar dalam pemrosesan data (misalnya Jaringan Pasokan Global, Analisis Keuangan Global, Large Hadron Collider).
- **Velocity (kecepatan data)**: berhubungan dengan aliran data real-time yang masuk dengan frekuensi tinggi (misalnya Sensor, Lingkungan Pervasif, Perdagangan Elektronik, Internet of Things).
- **Variasi (rentang tipe/sumber data)**: menangani data menggunakan format sintaksis yang berbeda (misalnya Spreadsheet, XML, DBMS), skema, dan makna (misalnya Integrasi Data Perusahaan).

Vs Big Data menantang dasar-dasar pendekatan teknis yang ada dan memerlukan bentuk pemrosesan data baru untuk memungkinkan peningkatan pengambilan keputusan, penemuan wawasan, dan optimalisasi proses. Seiring dengan semakin matangnya bidang Big Data, V lainnya telah ditambahkan seperti Veracity (mendokumentasikan kualitas dan ketidakpastian), Value, dll. Nilai Big Data dapat digambarkan dalam konteks dinamika organisasi berbasis pengetahuan (Choo 1996), dimana proses pengambilan keputusan dan tindakan organisasi bergantung pada proses pengambilan akal dan penciptaan pengetahuan.

Tabel 3.1 Definisi Big Data

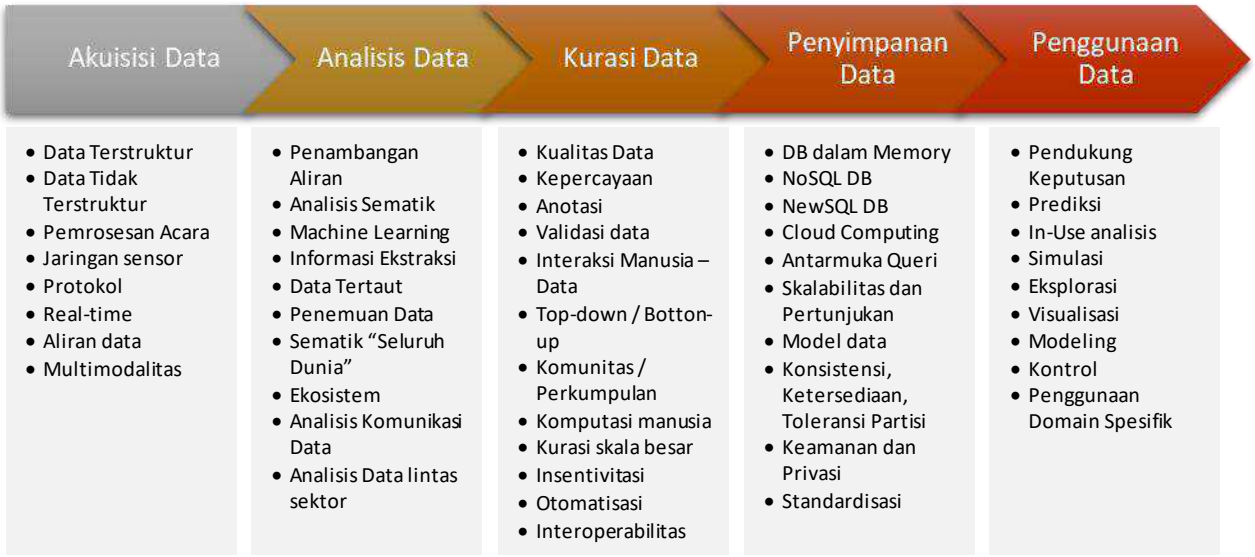
Definisi Data Besar	Sumber
"Big Data adalah aset informasi bervolume tinggi, berkecepatan tinggi, dan/atau beragam yang memerlukan bentuk pemrosesan baru untuk memungkinkan pengambilan keputusan yang lebih baik, penemuan wawasan, dan optimalisasi proses"	Laney (2001), Manyika et al. (2011)
"Ketika ukuran data itu sendiri menjadi bagian dari masalah dan teknik tradisional dalam menangani data menjadi sia-sia"	Loukides (2010)
Big Data adalah "data yang ukurannya memaksa kita untuk melihat melampaui metode yang sudah terbukti dan lazim pada saat itu"	Jacobs (2009)
"Teknologi Big Data [adalah] teknologi dan arsitektur generasi baru yang dirancang untuk mengekstraksi nilai secara ekonomis dari volume yang sangat besar dari beragam data dengan memungkinkan penangkapan, penemuan, dan/atau analisis berkecepatan tinggi"	IDC (2011)
"Istilah untuk kumpulan kumpulan data yang begitu besar dan kompleks sehingga menjadi sulit untuk diproses menggunakan alat manajemen basis data yang ada atau aplikasi pemrosesan data tradisional"	Wikipedia (2014)

"Kumpulan kumpulan data yang besar dan kompleks yang hanya dapat diproses dengan susah payah menggunakan alat manajemen basis data yang tersedia"	Mike 2.0 (2014)
"Big Data adalah istilah yang mencakup penggunaan teknik untuk menangkap, memproses, menganalisis, dan memvisualisasikan kumpulan data yang berpotensi berukuran besar dalam jangka waktu yang wajar dan tidak dapat diakses oleh teknologi TI standar." Lebih jauh lagi, platform, alat, dan perangkat lunak yang digunakan untuk tujuan ini secara kolektif disebut "teknologi Big Data"	NESSI (2012)
"Data besar dapat berarti volume yang besar, kecepatan yang besar, atau variasi yang besar"	Stonebraker (2012)

3.3 JARINGAN NILAI BIG DATA

Dalam bidang manajemen bisnis, jaringan nilai telah digunakan sebagai alat pendukung keputusan untuk memodelkan Jaringan aktivitas yang dilakukan organisasi untuk memberikan produk atau layanan yang berharga ke pasar (Porter 1985). Jaringan nilai mengkategorikan aktivitas umum yang memberi nilai tambah pada suatu organisasi sehingga aktivitas tersebut dapat dipahami dan dioptimalkan. Jaringan nilai terdiri dari serangkaian subsistem yang masing-masing memiliki masukan, proses transformasi, dan keluaran. Rayport dan Sviokla (1995) adalah salah satu orang pertama yang menerapkan metafora Jaringan nilai pada sistem informasi dalam pekerjaan mereka mengenai jaringan nilai virtual, sebagai alat analisis, Jaringan nilai dapat diterapkan pada arus informasi untuk memahami penciptaan nilai teknologi data. Dalam jaringan nilai data, arus informasi digambarkan sebagai serangkaian langkah yang diperlukan untuk menghasilkan nilai dan wawasan berguna dari data. Komisi Eropa melihat Jaringan nilai data sebagai "pusat ekonomi pengetahuan masa depan, yang membawa peluang perkembangan digital ke sektor-sektor yang lebih tradisional (misalnya transportasi, jasa keuangan, kesehatan, manufaktur, ritel)".

Jaringan nilai Big Data (Curry et al. 2014), seperti yang diilustrasikan pada Gambar 3.1, dapat digunakan untuk memodelkan aktivitas tingkat tinggi yang membentuk sistem informasi. Jaringan Nilai Big Data mengidentifikasi aktivitas-aktivitas utama tingkat tinggi berikut ini:



Gambar 3.1 Jaringan Nilai Big Data seperti yang dijelaskan di dalamnya

Akuisisi Data adalah proses pengumpulan, penyaringan, dan pembersihan data sebelum dimasukkan ke dalam gudang data atau solusi penyimpanan lainnya di mana analisis data dapat dilakukan. Akuisisi data adalah salah satu tantangan Big Data dalam hal kebutuhan infrastruktur. Infrastruktur yang diperlukan untuk mendukung akuisisi data besar harus memberikan latensi yang rendah dan dapat diprediksi baik dalam menangkap data maupun dalam menjalankan kueri; mampu menangani volume transaksi yang sangat tinggi, seringkali dalam lingkungan terdistribusi; dan mendukung struktur data yang fleksibel dan dinamis. Akuisisi data dirinci lebih lanjut dalam bab ini.

Analisis data berkaitan dengan membuat data mentah yang diperoleh dapat digunakan dalam pengambilan keputusan serta penggunaan khusus domain. Analisis data melibatkan eksplorasi, transformasi, dan pemodelan data dengan tujuan menyoroti data yang relevan, mensintesis dan mengekstrak informasi tersembunyi yang berguna dengan potensi tinggi dari sudut pandang bisnis. Area terkait mencakup penambangan data, intelijen bisnis, dan pembelajaran mesin.

Kurasi Data adalah pengelolaan data secara aktif selama siklus hidupnya untuk memastikan data tersebut memenuhi persyaratan kualitas data yang diperlukan untuk penggunaannya yang efektif. Proses kurasi data dapat dikategorikan ke dalam aktivitas berbeda seperti pembuatan konten, seleksi, klasifikasi, transformasi, validasi, dan pelestarian. Kurasi data dilakukan oleh kurator ahli yang bertanggung jawab meningkatkan aksesibilitas dan kualitas data. Kurator data (juga dikenal sebagai kurator ilmiah, atau anotator data) mempunyai tanggung jawab untuk memastikan bahwa data dapat dipercaya, dapat ditemukan, dapat diakses, dapat digunakan kembali, dan sesuai dengan tujuannya. Tren utama dalam kurasi Big Data adalah dengan menggunakan pendekatan komunitas dan crowdsourcing.

Penyimpanan Data adalah persistensi dan pengelolaan data dengan cara yang dapat diskalakan yang memenuhi kebutuhan aplikasi yang memerlukan akses cepat ke data. *Sistem Manajemen Basis Data Relasional* (RDBMS) telah menjadi solusi utama dan hampir unik dalam

paradigma penyimpanan selama hampir 40 tahun. Namun, properti ACID *(Atomicity, Consistency, Isolation, dan Durability)* yang menjamin transaksi database kurang fleksibel dalam hal perubahan skema dan kinerja serta toleransi kesalahan ketika volume dan kompleksitas data bertambah, sehingga tidak cocok untuk skenario Big Data. Teknologi NoSQL telah dirancang dengan tujuan skalabilitas dan menghadirkan berbagai solusi berdasarkan model data alternatif.

Penggunaan data mencakup aktivitas bisnis berbasis data yang memerlukan akses ke data, analisisnya, dan alat yang diperlukan untuk mengintegrasikan analisis data dalam aktivitas bisnis. Penggunaan data dalam pengambilan keputusan bisnis dapat meningkatkan daya saing melalui pengurangan biaya, peningkatan nilai tambah, atau parameter lainnya yang dapat diukur berdasarkan kriteria kinerja yang ada.

3.4 EKOSISTEM

Istilah ekosistem diciptakan oleh Tansley pada tahun 1935 untuk mengidentifikasi unit ekologi dasar yang terdiri dari lingkungan dan organisme yang menggunakannya. Dalam konteks bisnis, James F. Moore (1993, 1996, 2006) mengeksploitasi metafora biologis dan menggunakan istilah tersebut untuk menggambarkan lingkungan bisnis. Moore mendefinisikan ekosistem bisnis sebagai *"komunitas ekonomi yang didukung oleh landasan interaksi organisasi dan individu"* (Moore 1996). Sebuah strategi yang melibatkan sebuah perusahaan yang berusaha untuk sukses sendirian telah terbukti terbatas dalam hal kapasitasnya untuk menciptakan produk atau jasa yang bernilai. Sangat penting bagi dunia usaha untuk berkolaborasi satu sama lain agar dapat bertahan dalam ekosistem bisnis (Moore 1993; Gossain dan Kandiah 1998). Ekosistem memungkinkan perusahaan menciptakan nilai baru yang tidak dapat dicapai oleh perusahaan sendiri. Dalam ekosistem bisnis yang sehat, perusahaan dapat bekerja sama dalam jaringan bisnis yang kompleks di mana mereka dapat dengan mudah bertukar dan berbagi sumber daya penting.

Studi tentang Ekosistem Bisnis adalah bidang penelitian aktif di mana para peneliti menyelidiki banyak aspek metafora ekosistem bisnis untuk mengeksplorasi aspek-aspek seperti komunitas, kerjasama, saling ketergantungan, ko-evolusi, fungsi ekosistem, dan batas-batas lingkungan bisnis. Koening (2012) memberikan tipologi sederhana Ekosistem Bisnis berdasarkan tingkat kendali sumber daya utama dan jenis saling ketergantungan anggota. Jenis ekosistem bisnis mencakup sistem pasokan (misalnya Nike), platform (Apple iTunes), komunitas takdir (misalnya Sematech di industri semikonduktor), dan komunitas yang berkembang.

Ekosistem Big Data

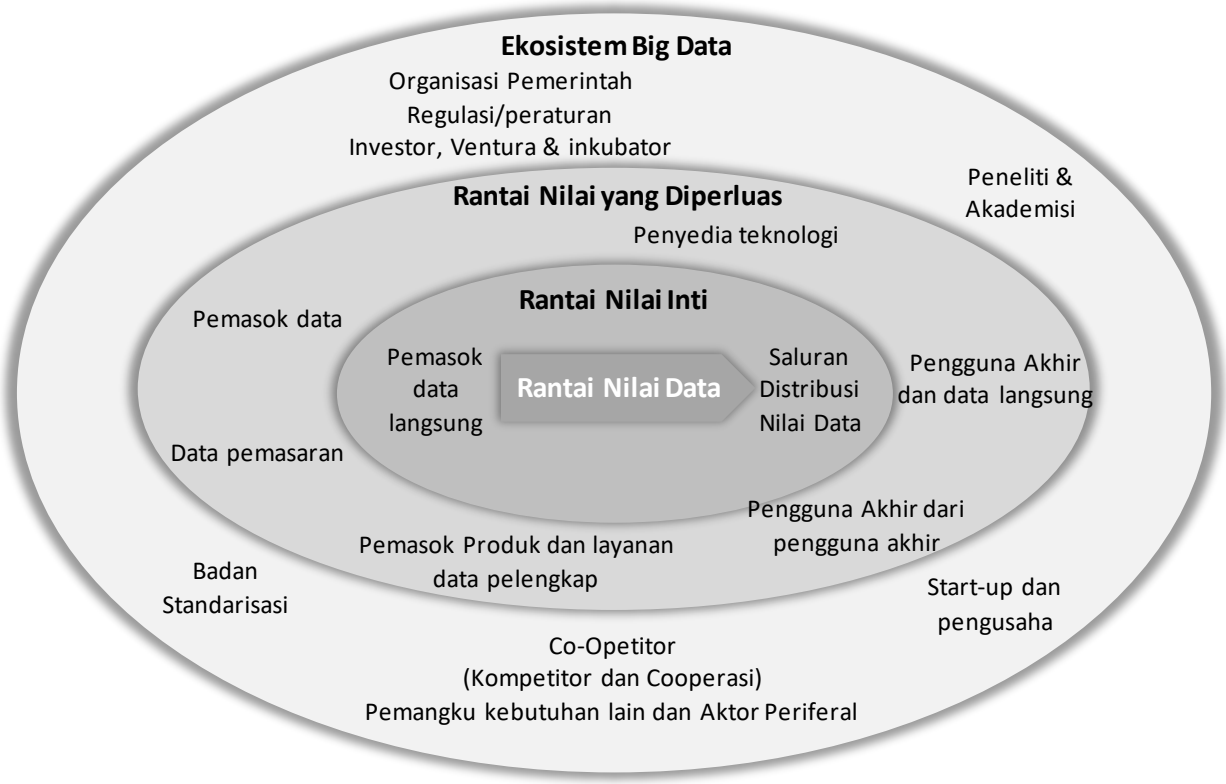
Dalam ekosistem alami, organisme cerdas mengendalikan energinya. Dalam ekosistem bisnis, perusahaan yang cerdas mengelola informasi dan arusnya. Dalam kaitannya dengan data, metafora ekosistem berguna untuk menggambarkan lingkungan data yang didukung oleh komunitas organisasi dan individu yang saling berinteraksi. Ekosistem Big Data dapat terbentuk dengan berbagai cara di dalam organisasi, platform teknologi komunitas, atau di dalam atau lintas sektor. Ekosistem Big Data ada di banyak sektor industri di mana sejumlah

besar data berpindah antar aktor dalam Jaringan pasokan informasi yang kompleks. Sektor-sektor dengan ekosistem data yang sudah mapan dan baru mencakup layanan kesehatan, keuangan, logistik, media, manufaktur, dan farmasi. Selain data itu sendiri, Ekosistem Big Data juga dapat didukung oleh platform pengelolaan data, infrastruktur data (misalnya berbagai proyek sumber terbuka Apache), dan layanan data.

Ekosistem Big Data Eropa

Meskipun tidak ada ekosistem data yang berhubungan di tingkat Eropa (DG Connect 2013), manfaat dari berbagi dan menghubungkan data antar domain dan sektor industri menjadi jelas. Inisiatif seperti kota pintar menunjukkan bagaimana berbagai sektor (misalnya energi dan transportasi) dapat berkolaborasi untuk memaksimalkan potensi optimalisasi dan pengembalian nilai. Perpaduan antar pemangku kepentingan dan kumpulan data dari berbagai sektor merupakan elemen kunci untuk memajukan perekonomian Big Data di Eropa.

Ekosistem bisnis Big Data di Eropa merupakan faktor penting dalam komersialisasi dan komoditisasi layanan, produk, dan platform Big Data. Ekosistem Big Data yang sukses akan membuat semua pemangku kepentingan berinteraksi secara lancar dalam pasar tunggal digital, yang mengarah pada peluang bisnis, akses yang lebih mudah terhadap pengetahuan dan modal.



Gambar 3.2 Tingkat Mikro, Meso, dan Makro dari Ekosistem Big Data [diadaptasi dari Moore (1996)]

Ekosistem data yang berfungsi dengan baik harus mempertemukan para pemangku kepentingan utama dengan manfaat yang jelas bagi semua pihak. Aktor-aktor kunci dalam ekosistem Big Data, seperti yang diilustrasikan pada Gambar 3.2, adalah:

1. **Pemasok Data**: Orang atau organisasi (Usaha besar dan kecil dan menengah (UKM)), yang membuat, mengumpulkan, menggabungkan, dan mengubah data dari sumber publik dan swasta
2. **Penyedia Teknologi**: Biasanya organisasi (Besar dan UKM) sebagai penyedia alat, platform, layanan, dan pengetahuan untuk pengelolaan data
3. **Pengguna Akhir Data**: Orang atau organisasi dari berbagai sektor industri (swasta dan publik) yang memanfaatkan teknologi dan layanan data besar untuk keuntungan mereka.
4. **Pasar Data**: Orang atau organisasi yang menampung data dari penerbit dan menawarkannya kepada konsumen/pengguna akhir.
5. **Startup dan Wirausahawan**: Mengembangkan teknologi, produk, dan layanan berbasis data yang inovatif.
6. **Peneliti dan Akademisi**: Selidiki algoritme, teknologi, metodologi, model bisnis, dan aspek kemasyarakatan baru yang diperlukan untuk memajukan Big Data.
7. **Regulator untuk privasi data dan masalah hukum**.
8. **Badan Standardisasi**: Menetapkan standar teknologi (baik resmi maupun de facto) untuk mendorong adopsi teknologi Big Data secara global.
9. **Investor, Kapitalis Ventura, dan Inkubator**: Orang atau organisasi yang menyediakan sumber daya dan layanan untuk mengembangkan potensi komersial ekosistem.

Menuju Ekosistem Big Data

Untuk mewujudkan ekosistem data yang luas di Eropa, diperlukan sejumlah tantangan teknis yang harus diatasi terkait dengan biaya dan kompleksitas penerbitan dan pemanfaatan data. Ekosistem saat ini menghadapi sejumlah masalah seperti penemuan data, kurasi, penautan, sinkronisasi, distribusi, pemodelan bisnis, serta penjualan dan pemasaran. Sejumlah tantangan sosial dan lingkungan hidup perlu diatasi untuk membangun ekosistem Big Data yang efektif; ini termasuk namun tidak terbatas pada:

1. Memahami nilai dan kontribusi teknologi Big Data
2. Menentukan nilai data
3. Identifikasi model bisnis yang akan mendukung ekosistem berbasis data
4. Memungkinkan pengusaha dan pemodal ventura mengakses ekosistem dengan mudah
5. Pelestarian privasi dan keamanan bagi seluruh aktor dalam ekosistem
6. Mengurangi fragmentasi bahasa, hak kekayaan intelektual, hukum, dan praktik kebijakan antar negara UE

Ringkasan

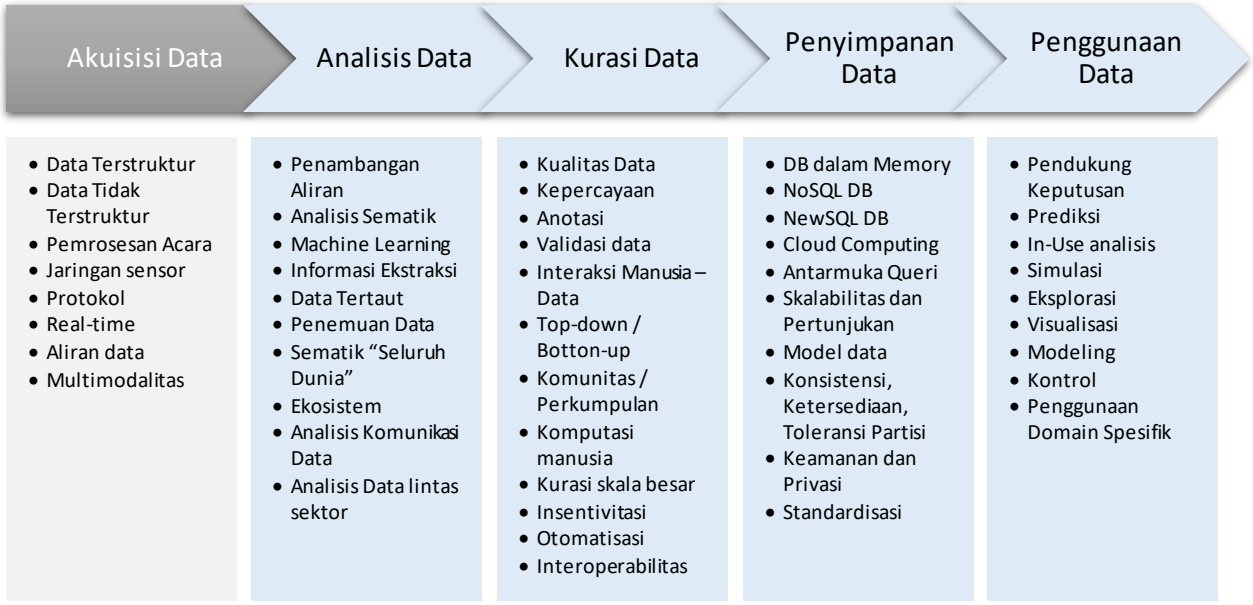
Big Data adalah bidang baru di mana teknologi inovatif menawarkan cara-cara baru untuk mengambil manfaat dari banyaknya informasi yang tersedia. Seperti halnya bidang baru lainnya, istilah dan konsep dapat menimbulkan interpretasi yang berbeda-beda. Domain Big

Data juga demikian. Perbedaan definisi “Big Data” yang muncul menunjukkan keragaman dan penggunaan istilah tersebut untuk memberi label data dengan atribut yang berbeda. Dua alat dari komunitas bisnis, Jaringan Nilai dan Ekosistem Bisnis, dapat digunakan untuk memodelkan sistem Big Data dan lingkungan bisnis Big Data. Jaringan Nilai Big Data dapat menggambarkan aliran informasi dalam sistem Big Data sebagai serangkaian langkah yang diperlukan untuk menghasilkan nilai dan wawasan berguna dari data.

BAB 4
AKUISISI DATA BESAR

4.1 PENDAHULUAN

Selama beberapa tahun terakhir, istilah Big Data digunakan oleh berbagai pemain besar untuk memberi label data dengan atribut berbeda. Selain itu, arsitektur memproses data yang berbeda untuk Big Data telah diusulkan untuk mengatasi perbedaan karakteristik Big Data. Secara keseluruhan, akuisisi data dipahami sebagai proses pengumpulan, pemfilteran, dan pembersihan data sebelum data dimasukkan ke dalam gudang data atau solusi penyimpanan lainnya.



Gambar 4.1 Akuisisi data dalam Jaringan nilai Big Data

Posisi akuisisi Big Data dalam keseluruhan Jaringan nilai Big Data dapat dilihat pada Gambar 4.1. Akuisisi Big Data umumnya diatur oleh empat faktor V: volume, kecepatan, variasi, dan nilai. Sebagian besar skenario akuisisi data mengasumsikan data bervolume tinggi, berkecepatan tinggi, bervariasi tinggi, namun bernilai rendah, sehingga penting untuk memiliki algoritma pengumpulan, pemfilteran, dan pembersihan yang dapat beradaptasi dan efisien waktu untuk memastikan bahwa hanya fragmen bernilai tinggi dari data yang diambil. data sebenarnya diproses oleh analisis data-gudang. Namun, bagi beberapa organisasi, sebagian besar data berpotensi bernilai tinggi karena penting untuk merekrut pelanggan baru. Untuk organisasi seperti itu, analisis, klasifikasi, dan pengemasan data pada volume data yang sangat tinggi memainkan peran paling sentral setelah akuisisi data.

Tujuan dari bab ini ada tiga: Pertama, bertujuan untuk mengidentifikasi persyaratan umum saat ini untuk akuisisi data dengan menyajikan kerangka kerja dan protokol terbuka yang canggih untuk akuisisi data besar bagi perusahaan. Tujuan kedua kami adalah mengungkap pendekatan yang saat ini digunakan untuk akuisisi data di berbagai sektor.

Terakhir, bab ini membahas bagaimana persyaratan akuisisi data dipenuhi dengan pendekatan yang ada saat ini serta kemungkinan pengembangan di masa depan dalam bidang yang sama.

4.2 WAWASAN PENTING UNTUK AKUISISI BIG DATA

Untuk mendapatkan pemahaman yang lebih baik tentang akuisisi data, bab ini pertama-tama akan melihat perbedaan arsitektur Big Data yang dimiliki Oracle, Vivisimo, dan IBM. Hal ini akan mengintegrasikan proses akuisisi dalam jalur pemrosesan data besar. Jalur pemrosesan data besar telah diabstraksikan dalam berbagai cara pada penelitian sebelumnya. Oracle (2012) mengandalkan pendekatan tiga langkah untuk pemrosesan data. Pada langkah pertama, konten dari sumber data yang berbeda diambil dan disimpan dalam solusi penyimpanan terukur seperti database NoSQL atau Sistem File Terdistribusi Hadoop (HDFS). Data yang disimpan selanjutnya diproses dengan terlebih dahulu ditata ulang dan disimpan dalam perangkat lunak analisis data besar berkemampuan SQL dan terakhir dianalisis dengan menggunakan algoritma analisis data besar.

Velocity (Vivisimo 2012) mengandalkan pandangan berbeda mengenai data besar. Di sini, pendekatannya lebih berorientasi pada pencarian. Komponen utama arsitekturnya adalah lapisan konektor, di mana berbagai sumber data dapat ditangani. Konten sumber data ini dikumpulkan secara paralel, dikonversi, dan akhirnya ditambahkan ke indeks, yang membangun dasar untuk analisis data, intelijen bisnis, dan semua aplikasi berbasis data lainnya. Pemain besar lainnya seperti IBM mengandalkan arsitektur yang mirip dengan Oracle (IBM 2013).

Di berbagai arsitektur pemrosesan data besar, inti dari akuisisi data adalah pengumpulan data dari sumber informasi terdistribusi dengan tujuan menyimpannya dalam penyimpanan data yang skalabel dan berkemampuan data besar. Untuk mencapai tujuan ini, diperlukan tiga komponen utama:

1. Protokol yang memungkinkan pengumpulan informasi untuk sumber data terdistribusi jenis apa pun (tidak terstruktur, semi terstruktur, terstruktur)
2. Kerangka kerja dimana data dikumpulkan dari sumber terdistribusi dengan menggunakan protokol yang berbeda
3. Teknologi yang memungkinkan penyimpanan data yang diambil oleh kerangka kerja secara terus-menerus

4.3 DAMPAK SOSIAL DAN EKONOMI DARI AKUISISI BIG DATA

Selama beberapa tahun terakhir, jumlah data yang dihasilkan secara stabil telah meningkat. Sembilan puluh persen data di dunia saat ini dihasilkan selama 2 tahun terakhir. Sumber dan sifat data ini beragam. Mulai dari data yang dikumpulkan oleh sensor hingga data yang menggambarkan transaksi (online). Jumlahnya yang terus meningkat diproduksi di media sosial dan melalui perangkat seluler. Jenis data (terstruktur atau tidak terstruktur) dan semantiknya juga beragam. Namun, semua data ini harus dikumpulkan untuk membantu menjawab pertanyaan bisnis dan membentuk gambaran pasar yang luas.

Bagi bisnis, tren ini mempunyai beberapa peluang dan tantangan baik dalam menciptakan model bisnis baru maupun meningkatkan operasi yang ada, sehingga menghasilkan keuntungan pasar. Alat dan metode untuk menangani data besar yang didorong oleh empat V dapat digunakan untuk meningkatkan periklanan spesifik pengguna atau riset pasar secara umum. Misalnya, sistem pengukuran cerdas sedang diuji di sektor energi. Selain itu, jika digabungkan dengan sistem penagihan baru, sistem ini juga dapat bermanfaat di sektor lain seperti telekomunikasi dan transportasi.

Big Data telah mempengaruhi banyak bisnis dan berpotensi berdampak pada semua sektor bisnis. Meskipun ada beberapa tantangan teknis, dampaknya terhadap manajemen dan pengambilan keputusan dan bahkan budaya perusahaan juga tidak kalah besarnya (McAfee dan Brynjolfsson 2012).

Namun masih ada beberapa batasan. Yaitu masalah privasi dan keamanan yang perlu ditangani oleh sistem dan teknologi ini. Banyak sistem telah menghasilkan dan mengumpulkan data dalam jumlah besar, namun hanya sebagian kecil yang digunakan secara aktif dalam proses bisnis. Selain itu, banyak dari sistem ini tidak memiliki persyaratan waktu nyata.

4.4 AKUISISI BIG DATA

Sebagian besar akuisisi data besar dilakukan dalam paradigma antrian pesan, terkadang juga disebut paradigma streaming, paradigma terbitkan/berlangganan (Carzaniga dkk. 2000), atau paradigma pemrosesan peristiwa (Cugola dan Margara 2012; Luckham 2002). Di sini, asumsi dasarnya adalah bahwa berbagai sumber data yang mudah menguap menghasilkan informasi yang perlu ditangkap, disimpan, dan dianalisis oleh platform pemrosesan data besar. Informasi baru yang dihasilkan oleh sumber data diteruskan ke penyimpanan data melalui kerangka akuisisi data yang menerapkan protokol yang telah ditentukan sebelumnya. Bagian ini menjelaskan dua teknologi inti untuk memperoleh data besar.

Protokol

Beberapa organisasi yang mengandalkan pemrosesan Big Data secara internal telah merancang protokol khusus perusahaan yang sebagian besar belum dirilis secara publik sehingga tidak dapat dijelaskan dalam bab ini. Bagian ini menyajikan protokol terbuka yang umum digunakan untuk akuisisi data.

AMQP

Alasan pengembangan *Advanced Message Queuing Protocol* (AMQP) adalah kebutuhan akan protokol terbuka yang dapat memenuhi kebutuhan perusahaan besar terkait dengan akuisisi data. Untuk mencapai tujuan ini, 23 perusahaan menyusun serangkaian persyaratan protokol akuisisi data. AMQP (Advanced Message Queuing Protocol) yang dihasilkan menjadi standar OASIS pada bulan Oktober 2012. Alasan di balik AMQP (Bank of America et al. 2011) adalah untuk menyediakan protokol dengan karakteristik berikut:

- **Ubiquity**: Properti AMQP ini mengacu pada kemampuannya untuk digunakan di berbagai industri baik dalam arsitektur akuisisi data saat ini dan di masa depan. Keberadaan AMQP dicapai dengan membuatnya mudah diperluas dan diterapkan.

Banyaknya framework yang mengimplementasikannya, termasuk SwiftMQ, Microsoft Windows Azure Service Bus, Apache Qpid, dan Apache ActiveMQ, mencerminkan betapa mudahnya implementasi protokol ini.

- **Keselamatan**: Properti keselamatan diterapkan pada dua dimensi yang berbeda. Pertama, protokol ini memungkinkan integrasi enkripsi pesan untuk memastikan bahwa pesan yang disadap pun tidak dapat didekodekan dengan mudah. Dengan demikian, ini dapat digunakan untuk mentransfer informasi penting bisnis. Protokol ini kuat terhadap suntikan spam, membuat broker AMQP sulit untuk diserang. Kedua, AMQP menjamin ketahanan pesan, artinya memungkinkan pesan untuk ditransfer bahkan ketika pengirim dan penerima tidak online pada waktu yang sama.
- **Fidelity**: Karakteristik ketiga ini berkaitan dengan integritas pesan. AMQP mencakup sarana untuk memastikan bahwa pengirim dapat mengekspresikan semantik pesan dan dengan demikian memungkinkan penerima memahami apa yang diterimanya. Protokol ini mengimplementasikan semantik kegagalan yang andal yang memungkinkan sistem mendeteksi kesalahan mulai dari pembuatan pesan di pihak pengirim hingga penyimpanan informasi oleh penerima.
- **Penerapan**: Tujuan di balik properti ini adalah untuk memastikan bahwa klien dan broker AMQP dapat berkomunikasi dengan menggunakan beberapa protokol lapisan model Open Systems Interconnection (OSI) seperti Transmission Control Protocol (TCP), User Datagram Protocol (UDP), dan juga Protokol Transmisi Kontrol Aliran (SCTP). Dengan cara ini, AMQP dapat diterapkan di banyak skenario dan industri di mana tidak semua protokol lapisan model OSI diperlukan dan digunakan. Selain itu, protokol ini dirancang untuk mendukung pola pesan yang berbeda termasuk pesan langsung, permintaan/balas, publikasi/berlangganan, dll.
- **Interoperabilitas**: Protokol dirancang agar independen terhadap implementasi dan vendor tertentu. Dengan demikian, klien dan broker dengan implementasi, arsitektur, dan kepemilikan yang sepenuhnya independen dapat berinteraksi melalui AMQP. Seperti disebutkan di atas, beberapa kerangka kerja dari berbagai organisasi kini menerapkan protokol ini.
- **Keterkelolaan**: Salah satu perhatian utama selama spesifikasi AMQP adalah memastikan bahwa kerangka kerja yang menerapkan AMQP dapat diperluas dengan mudah. Hal ini dicapai dengan memastikan bahwa AMQP adalah protokol kabel yang toleran terhadap kesalahan dan lossless yang melaluinya semua jenis informasi (misalnya XML, audio, video) dapat ditransfer.

Untuk mengimplementasikan persyaratan ini, AMQP bergantung pada sistem tipe dan empat lapisan berbeda: lapisan transport, lapisan pesan, lapisan transaksi, dan lapisan keamanan. Sistem tipe didasarkan pada tipe primitif dari database (bilangan bulat, string, simbol, dll.), tipe yang dijelaskan seperti yang diketahui dari pemrograman, dan nilai deskriptor yang dapat diperluas oleh pengguna protokol. Selain itu, AMQP memungkinkan penggunaan pengkodean untuk menyimpan simbol dan nilai serta definisi tipe gabungan yang terdiri dari kombinasi beberapa tipe primer.

Lapisan transport menentukan bagaimana pesan AMQP diproses. Jaringan AMQP terdiri dari node yang terhubung melalui tautan. Pesan dapat berasal dari (pengirim), diteruskan oleh (relay), atau dikonsumsi oleh node (penerima). Pesan hanya diperbolehkan melintasi suatu tautan jika tautan tersebut mematuhi kriteria yang ditentukan oleh sumber pesan. Lapisan transport mendukung beberapa jenis pertukaran rute termasuk penyebaran pesan dan pertukaran topik. Lapisan pesan AMQP menjelaskan struktur pesan yang valid. Pesan telanjang adalah pesan yang dikirimkan oleh pengirim ke jaringan AMQP.

Lapisan transaksi memungkinkan hasil terkoordinasi dari transfer independen. Ide dasar di balik arsitektur pendekatan pesan transaksional yang diikuti oleh lapisan terletak pada pengirim pesan yang bertindak sebagai pengontrol, sedangkan penerima bertindak sebagai sumber daya ketika pesan ditransfer sebagaimana ditentukan oleh pengontrol. Dengan cara ini, pemrosesan pesan yang terdesentralisasi dan terukur dapat dicapai. Lapisan AMQP terakhir adalah lapisan keamanan, yang memungkinkan definisi cara untuk mengenkripsi konten pesan AMQP. Protokol untuk mencapai tujuan ini seharusnya ditentukan secara eksternal dari AMQP itu sendiri. Protokol yang dapat digunakan untuk tujuan ini mencakup keamanan lapisan transport (TSL) dan lapisan otentikasi dan keamanan sederhana (SASL).

Karena penerapannya di beberapa industri dan fleksibilitasnya yang tinggi, AMQP kemungkinan besar akan menjadi pendekatan standar untuk pemrosesan pesan di industri yang tidak mampu menerapkan protokol khusus mereka sendiri. Dengan hadirnya industri *data-as-a-service*, industri ini juga menjanjikan solusi tepat untuk mengimplementasikan layanan seputar aliran data. Salah satu broker AMQP yang paling umum digunakan adalah RabbitMQ, yang popularitasnya sebagian besar disebabkan oleh fakta bahwa ia mengimplementasikan beberapa protokol pengiriman pesan termasuk JMS.

Layanan Pesan Java

Java Message Service (JMS) API disertakan dalam Java 2 Enterprise Edition pada tanggal 18 Maret 2002, setelah Java Community Process dalam versi finalnya 1.1 meratifikasinya sebagai standar.

Menurut spesifikasi 1.1 JMS “menyediakan cara umum bagi program Java untuk membuat, mengirim, menerima dan membaca pesan sistem pesan perusahaan”. Alat administratif memungkinkan seseorang untuk menetapkan tujuan dan menghubungkan pabrik ke dalam namespace *Java Naming and Directory Interface* (JNDI). Klien JMS kemudian dapat menggunakan injeksi sumber daya untuk mengakses objek yang dikelola di namespace dan kemudian membuat koneksi logis ke objek yang sama melalui penyedia JMS.

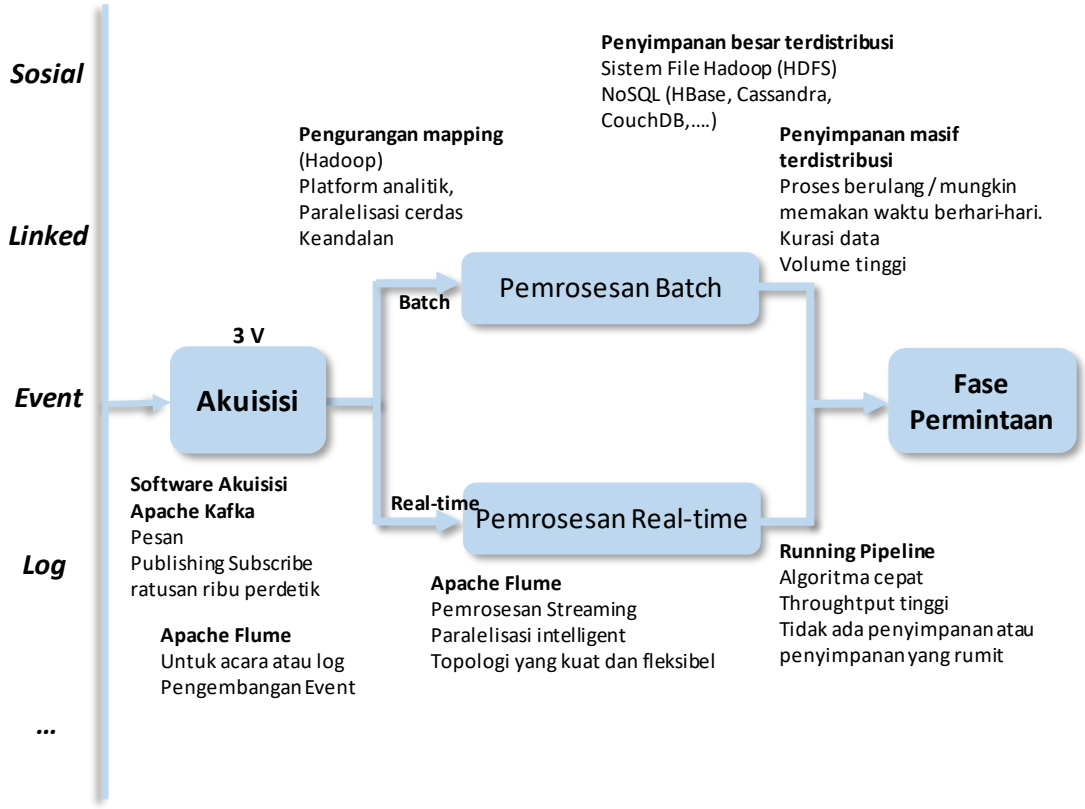
JNDI dalam hal ini berfungsi sebagai moderator antara klien berbeda yang ingin bertukar pesan. Perhatikan bahwa istilah "klien" digunakan di sini (seperti halnya spesifikasi) untuk menunjukkan pengirim dan penerima pesan, karena JMS pada awalnya dirancang untuk bertukar pesan secara *peer-to-peer*. Saat ini, JMS menawarkan dua model pengiriman pesan: *point-to-point* dan *penerbit-pelanggan*, dimana yang terakhir adalah koneksi satu-ke-banyak.

AMQP kompatibel dengan JMS, yang merupakan standar de facto untuk penyampaian pesan di dunia Java. Meskipun AMQP didefinisikan pada tingkat format (yaitu aliran byte oktet), JMS distandarisasi pada tingkat API dan oleh karena itu tidak mudah untuk diterapkan dalam bahasa pemrograman lain (seperti yang disarankan oleh "J" dalam "JMS"). JMS juga tidak menyediakan fungsionalitas untuk penyeimbangan beban/toleransi kesalahan, pemberitahuan kesalahan/penasihat, administrasi layanan, keamanan, protokol kabel, atau penyimpanan jenis pesan (akses database). Namun, keuntungan besar dari AMQP adalah kemandirian implementasi bahasa pemrograman yang menghindari penguncian vendor dan kompatibilitas platform.

Alat Perangkat Lunak

Sehubungan dengan perangkat lunak untuk akuisisi data, banyak di antaranya yang terkenal dan banyak kasus penggunaan tersedia di seluruh web sehingga layak untuk melakukan pendekatan pertama terhadap perangkat tersebut. Meskipun demikian, penggunaan setiap alat yang benar memerlukan pengetahuan mendalam tentang cara kerja internal dan implementasi perangkat lunak. Paradigma akuisisi data yang berbeda-beda muncul tergantung pada cakupan alat yang menjadi fokus alat ini. Diagram arsitektur pada Gambar 4.2 menunjukkan gambaran keseluruhan alur kerja Big Data lengkap yang menyoroti bagian akuisisi data.

Di sisa bagian ini, alat-alat ini dan alat-alat lain yang berkaitan dengan akuisisi data dijelaskan secara rinci.



Gambar 4.2 Alur Kerja Data Besar

Storm

Storm adalah kerangka kerja sumber terbuka untuk komputasi real-time terdistribusi yang kuat pada aliran data. Ini dimulai sebagai proyek sumber terbuka dan sekarang memiliki komunitas yang besar dan aktif. Storm mendukung berbagai bahasa pemrograman dan fasilitas penyimpanan (database relasional, penyimpanan NoSQL, dll.). Salah satu keuntungan utama Storm adalah dapat digunakan dalam banyak skenario pengumpulan data termasuk pemrosesan aliran dan RPC terdistribusi untuk menyelesaikan fungsi komputasi intensif saat itu juga, dan aplikasi komputasi berkelanjutan (Gabriel 2012). Banyak perusahaan dan aplikasi menggunakan Storm untuk mendukung berbagai sistem produksi yang memproses data, termasuk Groupon, The Weather Channel, fullcontact.com, dan Twitter.

Jaringan logis Storm terdiri dari tiga jenis node: node master yang disebut Nimbus, satu set node Zookeeper perantara, dan satu set node Supervisor.

a. **Nimbus**: setara dengan JobTracker Hadoop: ia mengunggah komputasi untuk dieksekusi, mendistribusikan kode ke seluruh cluster, dan memantau komputasi.
b. **Penjaga Kebun Binatang**: menangani koordinasi cluster secara lengkap. Lapisan organisasi cluster ini didasarkan pada proyek Apache ZooKeeper.
c. **Daemon Supervisor**: memunculkan node pekerja; ini sebanding dengan TaskTracker Hadoop. Di sinilah sebagian besar pekerjaan pengembang aplikasi dilakukan. Node pekerja berkomunikasi dengan Nimbus melalui Zookeepers untuk menentukan apa yang akan dijalankan pada mesin, memulai dan menghentikan pekerja.

Komputasi disebut topologi di Storm. Setelah diterapkan, topologi berjalan tanpa batas. Ada empat konsep dan lapisan abstraksi dalam Storm:

a. **Streams**: rangkaian tupel tak terbatas, yang diberi nama daftar nilai. Nilai dapat berupa objek sembarang yang mengimplementasikan antarmuka serialisasi.
b. **Spouts**: adalah sumber aliran dalam komputasi, mis. pembaca untuk sumber data seperti Twitter Streaming API.
c. **Baut**: memproses sejumlah aliran masukan dan menghasilkan sejumlah aliran keluaran. Di sinilah sebagian besar logika aplikasi berjalan.
d. **Topologi**: adalah abstraksi tingkat atas dari Storm. Pada dasarnya, topologi adalah jaringan cerat dan baut yang dihubungkan oleh tepian. Setiap tepinya adalah baut yang mengikuti aliran cerat atau baut lain.

Cerat dan baut merupakan node tanpa kewarganegaraan dan secara inheren paralel, menjalankan banyak tugas di seluruh cluster. Dari sudut pandang fisik, pekerja adalah proses Java Virtual Machine (JVM) dengan sejumlah tugas yang berjalan di dalamnya. Baik cerat maupun baut didistribusikan ke sejumlah tugas dan pekerja. Storm mendukung sejumlah pendekatan pengelompokan aliran mulai dari pengelompokan acak hingga tugas, hingga pengelompokan bidang, di mana tupel dikelompokkan berdasarkan bidang tertentu untuk tugas yang sama.

Storm menggunakan model tarik; setiap baut menarik peristiwa dari sumbernya. Tupel melintasi seluruh jaringan dalam jangka waktu tertentu atau dianggap gagal. Oleh karena itu,

dalam hal pemulihan, spouts bertanggung jawab untuk menjaga tupel tetap siap untuk diputar ulang.

S4

S4 (sistem streaming yang dapat diskalakan) adalah platform terdistribusi dan bertujuan umum untuk mengembangkan aplikasi yang memproses aliran data. Dimulai pada tahun 2008 oleh Yahoo! Inc., sejak 2011 merupakan proyek Apache Incubator. S4 dirancang untuk bekerja pada perangkat keras komoditas, menghindari kemacetan I/O dengan mengandalkan pendekatan all-in-memory (Neumeyer 2011).

Secara umum, kejadian data penting dialihkan ke elemen pemrosesan (PE). PP menerima peristiwa dan memancarkan peristiwa yang dihasilkan dan/atau memublikasikan hasilnya. Mesin S4 terinspirasi oleh model MapReduce dan menyerupai model Aktor (semantik enkapsulasi dan transparansi lokasi). Antara lain menyediakan antarmuka pemrograman sederhana untuk memproses aliran data dalam arsitektur terdesentralisasi, simetris, dan pluggable.

Aliran di S4 adalah urutan elemen (peristiwa) dari kunci dan atribut bernilai tupel. Unit komputasi dasar PE diidentifikasi oleh empat komponen berikut: (1) fungsinya yang disediakan oleh kelas PE dan konfigurasi terkait, (2) jenis peristiwa yang dikonsumsi, (3) atribut yang dikunci dalam peristiwa ini, dan (4) nilai atribut kunci dari peristiwa yang memakan waktu. PE dibuat instance-nya oleh platform untuk setiap nilai atribut kunci. PE tanpa kunci adalah kelas PE khusus tanpa atribut dan nilai yang dikunci. PE ini menggunakan semua peristiwa dari jenis yang sesuai dan biasanya berada di lapisan input klaster S4. Ada sejumlah besar PE standar yang tersedia untuk sejumlah tugas umum seperti agregat dan penggabungan. Host logis dari PE adalah node pemrosesan (PN). PN mendengarkan kejadian, menjalankan operasi untuk kejadian yang masuk, dan mengirimkan kejadian dengan bantuan lapisan komunikasi.

S4 merutekan setiap peristiwa ke PN berdasarkan fungsi hash pada semua nilai yang diketahui dari atribut yang dikunci dalam peristiwa tersebut. Ada jenis objek PE khusus lainnya: prototipe PE. Hal ini diidentifikasi oleh tiga komponen pertama. Objek-objek ini dikonfigurasikan pada saat inisialisasi dan untuk nilai apa pun objek tersebut dapat mengkloning dirinya sendiri untuk membuat PE yang sepenuhnya memenuhi syarat. Peristiwa kloning ini dipicu oleh PN untuk setiap nilai unik dari atribut yang dikunci. Aplikasi S4 adalah grafik yang terdiri dari prototipe PE dan aliran yang menghasilkan, menggunakan, dan mengirimkan pesan, sedangkan instance PE adalah klon dari prototipe terkait yang berisi status dan dikaitkan dengan kunci unik (Neumeyer et al. 2011).

Sebagai konsekuensi dari desain ini, S4 menjamin bahwa semua peristiwa dengan nilai spesifik dari atribut yang dikunci tiba di PN yang sesuai dan di dalamnya dialihkan ke contoh PE tertentu (Bradic 2011). Status PE saat ini tidak dapat diakses oleh PE lainnya. S4 didasarkan pada model dorong: peristiwa dialihkan ke PE berikutnya secepat mungkin. Oleh karena itu, jika buffer penerima terisi, acara mungkin akan dibatalkan. Melalui pos pemeriksaan lossy, S4 menyediakan pemulihan status. Jika terjadi kerusakan node, node baru akan mengambil alih tugasnya dari snapshot terbaru. Lapisan komunikasi didasarkan pada proyek Apache

ZooKeeper. Ia mengelola cluster dan menyediakan penanganan failover ke node stand-by. PE dibangun di Java menggunakan API yang cukup sederhana dan dirangkai ke dalam aplikasi menggunakan kerangka Spring.

Kafka

Kafka adalah sistem pesan terbitkan-langganan terdistribusi yang dirancang untuk mendukung pesan persisten dengan throughput tinggi. Kafka bertujuan untuk menyatukan pemrosesan offline dan online dengan menyediakan mekanisme pemuatan paralel ke dalam Hadoop serta kemampuan untuk mempartisi konsumsi real time pada sekelompok mesin. Penggunaannya untuk pemrosesan aliran aktivitas membuat Kafka sebanding dengan Apache Flume, meskipun arsitektur dan primitifnya sangat berbeda dan menjadikan Kafka lebih sebanding dengan sistem pesan tradisional. Kafka awalnya dikembangkan di LinkedIn untuk melacak sejumlah besar peristiwa aktivitas yang dihasilkan oleh situs web. Peristiwa aktivitas ini sangat penting untuk memantau keterlibatan pengguna serta meningkatkan relevansi dalam produk berbasis data mereka. Diagram sebelumnya memberikan gambaran sederhana tentang topologi penerapan di LinkedIn.

Perhatikan bahwa satu klaster Kafka menangani semua data aktivitas dari sumber berbeda. Ini menyediakan saluran data tunggal untuk konsumen online dan offline. Tingkat ini bertindak sebagai penyangga antara aktivitas langsung dan pemrosesan asinkron. Kafka juga dapat digunakan untuk mereplikasi semua data ke pusat data berbeda untuk konsumsi offline. Kafka dapat digunakan untuk memberi makan Hadoop untuk analisis offline, serta cara untuk melacak metrik operasional internal yang memberi makan grafik secara real-time. Dalam konteks ini, penggunaan yang sangat tepat untuk Kafka dan mekanisme terbitkan-langganannya adalah memproses data aliran terkait, mulai dari melacak tindakan pengguna di situs web berskala besar hingga tugas relevansi dan pemeringkatan.

Di Kafka, setiap aliran disebut “topik”. Topik dipartisi untuk tujuan penskalaan. Produsen pesan menyediakan kunci yang digunakan untuk menentukan partisi tujuan pengiriman pesan. Dengan demikian, semua pesan yang dipartisi dengan kunci yang sama dijamin berada di partisi topik yang sama. Broker Kafka menangani beberapa partisi dan menerima serta menyimpan pesan yang dikirim oleh produsen.

Konsumen Kafka membaca suatu topik dengan mendapatkan pesan dari semua partisi topik tersebut. Jika konsumen ingin membaca semua pesan dengan kunci tertentu (misalnya ID pengguna jika ada klik pada situs web), ia hanya perlu membaca pesan dari partisi tempat kunci tersebut berada, bukan topik lengkapnya. Selain itu, dimungkinkan untuk mereferensikan titik mana pun dalam file log broker menggunakan offset. Offset ini menentukan di mana konsumen berada dalam pasangan topik atau partisi tertentu. Offset bertambah setelah konsumen membaca pasangan topik atau partisi.

Kafka memberikan jaminan pengiriman pesan setidaknya sekali dan ketersediaan partisi yang tinggi. Untuk menyimpan pesan dalam cache, Kafka mengandalkan sistem file, sedangkan semua data langsung ditulis ke log persisten tanpa harus dipindahkan ke disk. Secara kombinasi, protokol ini dibangun berdasarkan abstraksi kumpulan pesan, yang mengelompokkan pesan-pesan menjadi satu. Dengan demikian, ini meminimalkan overhead

jaringan dan operasi disk berurutan. Baik konsumen maupun produsen berbagi format pesan yang sama.

Saluran

Flume adalah layanan untuk mengumpulkan dan memindahkan data log dalam jumlah besar secara efisien. Ini memiliki arsitektur sederhana dan fleksibel berdasarkan aliran data streaming. Ini kuat dan toleran terhadap kesalahan dengan mekanisme keandalan yang dapat disesuaikan serta banyak mekanisme failover dan pemulihan. Ini menggunakan model data sederhana yang dapat diperluas yang memungkinkan aplikasi analitik online. Sistem ini dirancang dengan empat tujuan utama berikut: keandalan, skalabilitas, pengelolaan, dan ekstensibilitas.

Tujuan Flume adalah menyediakan sistem yang terdistribusi, andal, dan tersedia untuk mengumpulkan, menggabungkan, dan memindahkan data log dalam jumlah besar secara efisien dari berbagai sumber ke penyimpanan data terpusat. Arsitektur Flume NG didasarkan pada beberapa konsep yang bersama-sama membantu mencapai tujuan ini:

- **Peristiwa**: payload byte dengan header string opsional yang mewakili unit data yang dapat diangkut Flume dari titik asal ke tujuan akhirnya.
- **Aliran**: pergerakan peristiwa dari titik asal ke tujuan akhirnya dianggap sebagai aliran data, atau sekadar aliran.
- **Klien**: implementasi antarmuka yang beroperasi pada titik asal peristiwa dan mengirimkannya ke agen Flume.
- **Agen**: sebuah proses independen yang menampung komponen flume seperti sumber, saluran, dan sink, sehingga memiliki kemampuan untuk menerima, menyimpan, dan meneruskan kejadian ke tujuan hop berikutnya.
- **Sumber**: implementasi antarmuka yang dapat menggunakan peristiwa yang dikirimkan melalui mekanisme tertentu.
- **Saluran**: penyimpanan sementara acara, di mana acara dikirimkan ke saluran melalui sumber yang beroperasi di dalam agen. Peristiwa yang dimasukkan ke dalam saluran akan tetap berada di saluran tersebut sampai wastafel memindahkannya untuk diangkut lebih lanjut.
- **Sink**: implementasi antarmuka yang dapat menghapus peristiwa dari saluran dan mengirimkannya ke agen berikutnya dalam aliran, atau ke tujuan akhir peristiwa.

Konsep-konsep ini membantu menyederhanakan arsitektur, implementasi, konfigurasi, dan penerapan Flume.

Aliran di Flume NG dimulai dari klien. Klien mengirimkan acara tersebut ke tujuan hop berikutnya. Tujuan ini adalah agen. Lebih tepatnya, tujuan adalah sumber yang beroperasi di dalam agen. Sumber yang menerima peristiwa ini kemudian akan mengirimkannya ke satu atau lebih saluran. Saluran yang menerima kejadian tersebut dikuras oleh satu atau lebih sink yang beroperasi dalam agen yang sama. Jika sink adalah sink biasa, maka event tersebut akan diteruskan ke tujuan hop berikutnya, yaitu agen lain. Jika yang terjadi adalah terminal sink, maka event tersebut akan diteruskan ke tujuan akhirnya. Saluran memungkinkan pemisahan

sumber dari sink menggunakan model pertukaran data produsen-konsumen yang sudah dikenal. Hal ini memungkinkan sumber dan sink memiliki kinerja dan karakteristik runtime yang berbeda namun tetap dapat menggunakan sumber daya fisik yang tersedia untuk sistem secara efektif.

Kasus penggunaan utama Flume adalah sebagai sistem pencatatan yang mengumpulkan sekumpulan file log pada setiap mesin dalam sebuah cluster dan menggabungkannya ke penyimpanan persisten terpusat seperti Sistem File Terdistribusi Hadoop (HDFS). Selain itu, Flume dapat digunakan sebagai pengelola peristiwa HTTP yang menangani berbagai jenis permintaan dan mengarahkan masing-masing permintaan ke penyimpanan data tertentu selama proses akuisisi data, seperti database NoSQL seperti HBase.

Oleh karena itu, Apache Flume bukanlah sistem akuisisi data murni namun bertindak sebagai pelengkap dengan mengelola berbagai tipe data berbeda yang diperoleh dan mentransformasikannya ke penyimpanan atau repositori data tertentu.

Hadoop

Apache Hadoop adalah proyek sumber terbuka yang mengembangkan kerangka kerja untuk komputasi yang andal, terukur, dan terdistribusi pada data besar menggunakan kelompok perangkat keras komoditas. Itu berasal dari MapReduce Google dan Google File System (GFS) dan ditulis dalam JAVA. Ini digunakan dan didukung oleh komunitas besar dan digunakan dalam lingkungan produksi dan penelitian oleh banyak organisasi, terutama: Facebook, a9.com, AOL, Baidu, IBM, Imageshack, dan Yahoo. Proyek Hadoop terdiri dari empat modul:

a. **Hadoop Common**: untuk utilitas umum yang digunakan di seluruh Hadoop.
b. **Hadoop Distributed File System (HDFS)**: sistem file yang sangat tersedia dan efisien.
c. **Hadoop YARN (Yet Another Resource Negotiator)**: kerangka kerja untuk penjadwalan pekerjaan dan manajemen cluster.
d. **Hadoop MapReduce**: sistem untuk memproses data dalam jumlah besar secara paralel.

Cluster Hadoop dirancang berdasarkan prinsip master-slave. Master adalah nama node. Itu melacak metadata tentang distribusi file. File besar biasanya dipecah menjadi beberapa bagian berukuran 128 MB. Bagian-bagian ini disalin tiga kali dan replikanya didistribusikan melalui cluster node data (simpul budak). Jika terjadi kegagalan node, informasinya tidak hilang; node nama dapat mengalokasikan data lagi. Untuk memantau cluster, setiap node budak secara teratur mengirimkan detak jantung ke node nama. Jika seorang budak tidak dikenali dalam jangka waktu tertentu, maka ia dianggap mati. Karena node master merupakan satu titik kegagalan, node tersebut biasanya dijalankan pada perangkat keras yang sangat andal. Dan, sebagai tindakan pencegahan, node nama sekunder dapat melacak perubahan dalam meta data dengan bantuannya dimungkinkan untuk membangun kembali fungsionalitas node nama dan dengan demikian memastikan fungsionalitas cluster.

YARN adalah penjadwal cluster Hadoop. Ini mengalokasikan sejumlah kontainer (yang merupakan proses penting) dalam sekelompok mesin dan mengeksekusi perintah sewenang-wenang pada mereka. YARN terdiri dari tiga bagian utama: ResourceManager, NodeManager, dan ApplicationMaster. Dalam sebuah cluster, setiap mesin menjalankan NodeManager, yang bertanggung jawab untuk menjalankan proses pada mesin lokal. Resource-Manager memberi tahu NodeManager apa yang harus dijalankan, dan Aplikasi memberi tahu ResourceManager kapan harus menjalankan sesuatu di cluster.

Data diproses sesuai dengan paradigma MapReduce. MapReduce adalah kerangka kerja untuk komputasi terdistribusi paralel. Karena pemrosesan penyimpanan data bekerja dalam mode master-slave, tugas komputasi disebut pekerjaan dan didistribusikan oleh pelacak pekerjaan. Alih-alih memindahkan data ke penghitungan, Hadoop memindahkan penghitungan ke data. Pelacak pekerjaan berfungsi sebagai master yang mendistribusikan dan mengelola pekerjaan di cluster. Pelacak tugas melakukan pekerjaan sebenarnya pada pekerjaan. Biasanya setiap node cluster menjalankan instance pelacak tugas dan node data. Kerangka kerja MapReduce memudahkan pemrograman program paralel yang sangat terdistribusi. Seorang pemrogram dapat fokus pada penulisan fungsi map dan pengurangan yang lebih sederhana untuk menangani tugas yang ada, sementara infrastruktur MapReduce menangani pengoperasian dan pengelolaan tugas di cluster.

Dalam orbit proyek Hadoop sejumlah proyek terkait telah bermunculan. Proyek Apache Pig misalnya dibangun di atas Hadoop dan menyederhanakan penulisan dan pemeliharaan implementasi Hadoop. Hadoop sangat efisien untuk pemrosesan batch. Proyek Apache HBase bertujuan untuk menyediakan akses real-time ke data besar.

4.5 TREN DAN SYARAT UNTUK AKUISISI BIG DATA

Peralatan akuisisi data besar harus menangani akuisisi data berkecepatan tinggi, beragam, dan real-time. Oleh karena itu, peralatan untuk akuisisi data harus memastikan throughput yang sangat tinggi. Artinya, data dapat berasal dari berbagai sumber (jaringan sosial, sensor, penambangan web, log, dll.) dengan struktur berbeda, atau tidak terstruktur (teks, video, gambar, dan file media) dan dalam kecepatan yang sangat tinggi (puluhan atau ratusan ribu kejadian per detik). Oleh karena itu, tantangan utama dalam memperoleh Big Data adalah menyediakan kerangka kerja dan alat yang memastikan hasil yang diperlukan untuk masalah yang dihadapi tanpa kehilangan data apa pun dalam prosesnya.

Dalam konteks ini, tantangan yang muncul dalam akuisisi Big Data adalah sebagai berikut:

a. Akuisisi data sering kali dimulai dengan alat yang menyediakan semacam masukan data ke sistem, seperti jaringan sosial dan algoritme penambangan web, perangkat lunak akuisisi data sensor, log yang dimasukkan secara berkala, dll. Biasanya proses akuisisi data dimulai dengan satu atau beberapa data. beberapa titik akhir dari mana data berasal. Titik akhir ini dapat memiliki tampilan teknis yang berbeda, seperti pengimpor log, algoritme berbasis Storm, atau bahkan akuisisi data mungkin menawarkan API ke dunia luar untuk memasukkan data, dengan menggunakan

layanan RESTful atau API terprogram lainnya. Oleh karena itu, setiap solusi teknis yang bertujuan untuk memperoleh data dari berbagai sumber harus mampu menangani berbagai macam implementasi yang berbeda.

b. Untuk menyediakan mekanisme untuk menghubungkan akuisisi data dengan data sebelum dan sesudah pemrosesan (analisis) dan penyimpanan, baik dalam lapisan historis maupun real-time. Untuk melakukan hal ini, alat pemrosesan batch dan real-time (yaitu Storm dan Hadoop) harus dapat dihubungi oleh alat akuisisi data. Hal ini diterapkan dengan cara yang berbeda. Misalnya Apache Kafka menggunakan mekanisme terbitkan-berlangganan di mana Hadoop dan Storm dapat berlangganan, sehingga pesan yang diterima akan tersedia bagi mereka. Apache Flume di sisi lain mengikuti pendekatan yang berbeda, menyimpan data dalam penyimpanan nilai kunci NoSQL untuk memastikan kecepatan, dan mengirimkan data ke satu atau beberapa penerima (yaitu Hadoop dan Storm). Ada garis tipis berwarna merah antara akuisisi, penyimpanan, dan analisis data dalam proses ini, karena akuisisi data biasanya diakhiri dengan menyimpan data mentah dalam kumpulan data master yang sesuai, dan menghubungkannya dengan saluran analitik (terutama untuk real-time, tetapi juga batch pengolahan).
c. Menghasilkan model terstruktur atau semi-terstruktur yang valid untuk analisis data, agar dapat melakukan pra-pemrosesan data yang diperoleh secara efektif, khususnya data tidak terstruktur. Batasan antara perolehan dan analisis data menjadi kabur pada tahap pra-pemrosesan. Beberapa orang mungkin berargumentasi bahwa pra-pemrosesan adalah bagian dari pemrosesan, dan juga bagian dari analisis data, sementara yang lain percaya bahwa perolehan data tidak berakhir pada pengumpulan sebenarnya, namun juga dengan pembersihan data dan menyediakan serangkaian hubungan dan metadata minimal. dia. Pembersihan data biasanya memerlukan beberapa langkah, seperti penghapusan boilerplate (yaitu menghapus header HTML dalam akuisisi penambangan web), deteksi bahasa dan pengenalan entitas bernama (untuk sumber daya tekstual), dan menyediakan metadata tambahan seperti stempel waktu, informasi asal (satu lagi yang tumpang tindih). dengan kurasi data), dll.
d. Akuisisi media (gambar, video) merupakan tantangan yang signifikan, namun tantangan yang lebih besar lagi adalah melakukan analisis dan penyimpanan video dan gambar.
e. Keragaman data memerlukan pemrosesan semantik dalam data agar dapat menggabungkan data dari berbagai sumber dengan benar dan efektif saat pemrosesan. Penelitian mengenai pemrosesan peristiwa semantik seperti perkiraan semantik (Hasan dan Curry 2014a), pemrosesan peristiwa tematik (Hasan dan Curry 2014b), dan penandaan thingsonomy (Hasan dan Curry 2015) merupakan pendekatan yang muncul di bidang ini, dalam konteks ini.
f. Untuk melakukan pasca dan pra-pemrosesan data yang diperoleh, teknologi terkini menyediakan seperangkat alat dan kerangka kerja sumber terbuka dan komersial. Oleh karena itu, tujuan utama ketika menentukan strategi akuisisi data yang benar

adalah untuk memahami kebutuhan sistem dalam hal volume, variasi, dan kecepatan data, serta mengambil keputusan yang tepat tentang alat mana yang terbaik untuk memastikan akuisisi dan throughput yang diinginkan.

4.6 STUDI KASUS SEKTOR UNTUK AKUISISI BIG DATA

Bagian ini menganalisis penggunaan teknologi akuisisi data besar di sejumlah sektor.

Bidang Kesehatan

Dalam sektor kesehatan, teknologi Big Data bertujuan untuk membangun pendekatan holistik di mana data klinis, keuangan, dan administratif serta data perilaku pasien, data populasi, data perangkat medis, dan data kesehatan terkait lainnya digabungkan dan digunakan untuk retrospektif, data nyata. waktu, dan analisis prediktif.

Untuk membangun landasan bagi keberhasilan implementasi aplikasi kesehatan Big Data, tantangan digitalisasi dan akuisisi data (yaitu menempatkan data kesehatan dalam bentuk yang sesuai sebagai masukan untuk solusi analitik) perlu diatasi. Saat ini, sejumlah besar data kesehatan disimpan dalam silo data dan pertukaran data hanya dapat dilakukan melalui Pindai, Faks, atau email. Karena antarmuka yang tidak fleksibel dan standar yang hilang, pengumpulan data kesehatan bergantung pada solusi individual dengan biaya tinggi.

Di rumah sakit, data pasien disimpan dalam sistem CIS (sistem informasi klinis) atau EHR (catatan kesehatan elektronik). Namun, departemen klinis yang berbeda mungkin menggunakan sistem yang berbeda, seperti RIS (sistem informasi radiologi), LIS (sistem informasi laboratorium), atau PACS (sistem pengarsipan gambar dan komunikasi) untuk menyimpan datanya. Tidak ada model data standar atau sistem EHR. Mekanisme yang ada untuk integrasi data adalah adaptasi dari solusi gudang data standar dari penyedia TI horizontal seperti Oracle Healthcare Data Model, Healthcare Logical Data Model Teradata, IBM Healthcare Provider Data Model, atau solusi baru seperti platform i2b2. Meskipun tiga kriteria pertama terutama digunakan untuk menghasilkan tolok ukur mengenai kinerja organisasi rumah sakit secara keseluruhan, platform i2b2 membangun gudang data yang memungkinkan integrasi data dari berbagai departemen klinis untuk mendukung tugas mengidentifikasi kelompok pasien. Dengan melakukan hal ini, data terstruktur seperti diagnosis dan nilai laboratorium dipetakan ke sistem pengkodean standar. Namun, data tidak terstruktur tidak diberi label lebih lanjut dengan informasi semantik. Selain fungsi utamanya untuk identifikasi kelompok pasien, sarang i2b2 menawarkan beberapa modul tambahan. Selain modul khusus untuk tugas impor, ekspor, dan visualisasi data, juga tersedia modul untuk membuat dan menggunakan semantik tambahan. Misalnya, alat pemrosesan bahasa alami (NLP) menawarkan sarana untuk mengekstraksi konsep dari istilah-istilah tertentu dan menghubungkannya dengan pengetahuan terstruktur.

Saat ini, data dapat dipertukarkan dengan menggunakan format pertukaran seperti HL7. Namun, karena alasan non-teknis seperti privasi, data kesehatan biasanya tidak dibagikan ke seluruh organisasi (fenomena silo organisasi). Informasi tentang diagnosis, prosedur, nilai laboratorium, demografi, pengobatan, penyedia layanan, secara umum disediakan dalam format terstruktur, namun tidak secara otomatis dikumpulkan dalam cara

yang terstandarisasi. Misalnya, departemen lab menggunakan sistem pengkodeannya sendiri untuk nilai lab tanpa pemetaan eksplisit ke standar LOINC *(Logical Observation Identifiers Names and Codes)*. Selain itu, departemen klinis yang berbeda sering kali menggunakan templat laporan yang berbeda namun disesuaikan tanpa menentukan semantik umum. Kedua skenario tersebut menyebabkan kesulitan dalam perolehan data dan integrasi yang diakibatkannya.

Mengenai data tidak terstruktur seperti teks dan gambar, standar untuk mendeskripsikan informasi meta tingkat tinggi hanya dikumpulkan sebagian. Dalam domain pencitraan, tersedia standar DICOM *(Digital Imaging and Communications in Medicine)* untuk menentukan metadata gambar. Namun, untuk menggambarkan meta-informasi laporan klinis atau studi klinis, standar umum (yang disepakati) tidak ada. Sejauh pengetahuan kami, untuk representasi informasi konten data tidak terstruktur seperti gambar, teks, atau data genomik, tidak ada standar yang tersedia. Upaya awal untuk mengubah situasi ini adalah inisiatif seperti inisiatif pelaporan terstruktur oleh RSNA atau anotasi semantik menggunakan kosakata standar. Misalnya, Medical Subject Headings (MeSH) adalah tesaurus kosakata terkontrol dari Perpustakaan Kedokteran Nasional AS untuk menangkap topik teks dalam domain medis dan biologi. Ada juga beberapa terjemahan ke bahasa lain.

Karena setiap vendor EHR menyediakan model datanya sendiri, tidak ada model data standar untuk penggunaan sistem pengkodean untuk mewakili konten laporan klinis. Dalam hal cara yang mendasari representasi data, sistem EHR yang ada saat ini lebih mengandalkan representasi data kesehatan yang berpusat pada kasus (case-centric) dan bukan pada representasi data kesehatan yang berpusat pada pasien. Hal ini menghambat perolehan dan integrasi data kesehatan jangka panjang.

Diperlukan alat pelaporan terstruktur yang mudah digunakan dan tidak menimbulkan pekerjaan ekstra bagi dokter, yaitu sistem ini harus diintegrasikan secara mulus ke dalam alur kerja klinis. Selain itu, informasi konteks yang tersedia harus digunakan untuk membantu dokter. Mengingat alat pelaporan terstruktur diimplementasikan sebagai alat yang mudah digunakan, alat tersebut dapat diterima oleh dokter sehingga sebagian besar dokumentasi klinis dilakukan dalam bentuk semi-terstruktur dan kualitas serta kuantitas anotasi semantik meningkat.

Dari sudut pandang organisasi, penyimpanan, pemrosesan, akses, dan perlindungan data besar harus diatur di beberapa tingkat berbeda: tingkat kelembagaan, regional, nasional, dan internasional. Ada kebutuhan untuk mendefinisikan siapa yang mengizinkan proses tertentu, siapa yang mengubah proses, dan siapa yang mengimplementasikan perubahan proses. Oleh karena itu, kerangka atau pedoman hukum yang tepat dan konsisten (misalnya. ISO/IEC 27000) untuk keempat level diperlukan.

IHE (mengintegrasikan perusahaan layanan kesehatan) memungkinkan akses plug-and-play dan aman ke informasi kesehatan kapan pun dan di mana pun diperlukan. Ini menyediakan spesifikasi, alat, dan layanan yang berbeda. IHE juga mempromosikan penggunaan standar yang sudah mapan dan diterima secara internasional (misalnya Pencitraan Digital dan Komunikasi dalam Kedokteran, Tingkat Kesehatan 7). Data farmasi dan

penelitian dan pengembangan yang mencakup uji klinis, studi klinis, data populasi dan penyakit, dll. biasanya dimiliki oleh perusahaan farmasi, laboratorium penelitian/akademisi, atau pemerintah. Saat ini, banyak upaya manual dilakukan untuk mengumpulkan semua kumpulan data untuk melakukan studi klinis dan analisis terkait. Upaya manual untuk mengumpulkan data cukup tinggi.

Manufaktur, Ritel, dan Transportasi

Akuisisi Big Data dalam konteks sektor ritel, transportasi, dan manufaktur menjadi semakin penting. Ketika biaya pemrosesan data menurun dan kapasitas penyimpanan meningkat, data kini dapat dikumpulkan secara terus-menerus. Perusahaan manufaktur serta pengecer dapat memantau saluran seperti Facebook, Twitter, atau berita untuk mengetahui adanya penyebutan dan menganalisis data ini (misalnya analisis sentimen pelanggan). Pengecer di web juga mengumpulkan data dalam jumlah besar dengan menyimpan file log dan menggabungkan informasi tersebut dengan sumber data lain seperti data penjualan untuk menganalisis dan memprediksi perilaku pelanggan. Di bidang manufaktur, semua perangkat yang berpartisipasi saat ini saling terhubung (misalnya sensor, RFID), sehingga informasi penting dikumpulkan secara konstan untuk memprediksi komponen yang rusak pada tahap awal.

Ketiga sektor tersebut memiliki kesamaan yaitu data berasal dari sumber yang sangat heterogen (misalnya file log, data dari media sosial yang perlu diekstraksi melalui API milik sendiri, data dari sensor, dll.). Data masuk dengan kecepatan yang sangat tinggi, sehingga memerlukan pemilihan teknologi yang tepat untuk ekstraksi (misalnya MapReduce). Tantangannya mungkin juga mencakup integrasi data. Misalnya, nama produk yang digunakan oleh pelanggan di platform media sosial harus dicocokkan dengan ID yang digunakan untuk halaman produk di web dan kemudian dicocokkan dengan ID internal yang digunakan dalam sistem Enterprise Resource Planning (ERP). Alat yang digunakan untuk akuisisi data di ritel dapat dikelompokkan berdasarkan dua jenis data yang biasanya dikumpulkan di ritel:

- Data penjualan dari departemen akuntansi dan pengendalian
- Data dari departemen pemasaran

Monitor saluran data dinamit, yang baru-baru ini dibeli oleh Market Track LLC, memberikan solusi untuk mengumpulkan informasi tentang harga produk di lebih dari Rp.1 miliar halaman beli di lebih dari 4000 pengecer global secara real time, dan dengan demikian memungkinkan untuk mempelajari dampak promosi investasi, memantau harga, dan melacak sentimen konsumen terhadap merek dan produk.

Meningkatnya penggunaan media sosial tidak hanya memberdayakan konsumen untuk dengan mudah membandingkan layanan dan produk baik dari segi harga dan kualitas, namun juga memungkinkan pengecer untuk mengumpulkan, mengelola, dan menganalisis volume dan kecepatan data yang besar, memberikan peluang besar bagi industri ritel. . Untuk mendapatkan keunggulan kompetitif, informasi real-time sangat penting untuk prediksi akurat dan model optimasi. Dari perspektif akuisisi data, diperlukan sarana untuk komputasi aliran data, yang dapat mengatasi tantangan Vs data.

Agar dapat memberikan manfaat bagi sektor transportasi (khususnya transportasi perkotaan multimoda), alat yang mendukung akuisisi data besar harus menyelesaikan dua tugas utama (DHL 2013; Davenport 2013). Pertama, mereka harus menangani data pribadi dalam jumlah besar (misalnya informasi lokasi) dan menangani masalah privasi terkait. Kedua, mereka harus mengintegrasikan data dari berbagai penyedia layanan, termasuk sensor yang tersebar secara geografis yaitu Internet of Things (IoT) dan sumber data terbuka.

Berbagai pemain mendapatkan manfaat dari Big Data di sektor transportasi. Pemerintah dan lembaga publik menggunakan semakin banyak data untuk pengendalian lalu lintas, perencanaan rute, dan manajemen transportasi. Sektor swasta memanfaatkan semakin banyak waktu untuk perencanaan rute dan pengelolaan pendapatan untuk mendapatkan keunggulan kompetitif, menghemat waktu, dan meningkatkan efisiensi bahan bakar. Individu semakin banyak menggunakan data melalui situs web, aplikasi perangkat seluler, dan informasi GPS untuk perencanaan rute guna meningkatkan efisiensi dan menghemat waktu perjalanan.

Di sektor manufaktur, alat akuisisi data terutama perlu memproses data sensor dalam jumlah besar. Alat-alat tersebut perlu menangani data sensor yang mungkin tidak kompatibel dengan data sensor lainnya sehingga tantangan integrasi data perlu diatasi, terutama ketika data sensor dikirimkan melalui beberapa perusahaan dalam satu Jaringan nilai. Kategori alat lainnya perlu mengatasi masalah pengintegrasian data yang dihasilkan oleh sensor dalam lingkungan produksi dengan data dari, misalnya. Sistem ERP dalam perusahaan. Hal ini paling baik dicapai ketika alat menghasilkan dan menggunakan format metadata standar.

Pemerintah, Masyarakat, Nirlaba

Mengintegrasikan dan menganalisis data dalam jumlah besar memainkan peran yang semakin penting dalam masyarakat saat ini. Namun seringkali, penemuan dan wawasan baru hanya dapat dicapai dengan mengintegrasikan informasi dari sumber yang tersebar. Meskipun ada kemajuan baru-baru ini dalam penerbitan data terstruktur di web (seperti penggunaan RDF dalam atribut (RDFa) dan inisiatif skema.org), pertanyaan yang muncul adalah bagaimana kumpulan data yang lebih besar dapat dipublikasikan dengan cara yang membuatnya mudah ditemukan dan juga memfasilitasi integrasi. sebagai analisis.

Salah satu pendekatan untuk mengatasi masalah ini adalah portal data, yang memungkinkan organisasi mengunggah dan mendeskripsikan kumpulan data menggunakan skema metadata yang komprehensif. Mirip dengan perpustakaan digital, jaringan portal data tersebut dapat mendukung deskripsi, pengarsipan, dan penemuan kumpulan data di web. Baru-baru ini, terjadi pertumbuhan pesat dalam katalog data yang tersedia di web. Registri katalog data datacatalogs.org mencantumkan 314 katalog data di seluruh dunia. Contoh meningkatnya popularitas katalog data adalah portal Open Government Data, portal data organisasi internasional dan LSM, serta portal data ilmiah. Di sektor publik dan pemerintahan, beberapa katalog dan pusat data dapat digunakan untuk menemukan metadata atau setidaknya untuk menemukan lokasi (tautan) ke file media yang menarik seperti publicdata.eu.

Sektor publik berpusat pada aktivitas warga negara. Akuisisi data di sektor publik meliputi pengumpulan pajak, statistik kejahatan, data polusi air dan udara, laporan cuaca, konsumsi energi, regulasi bisnis internet: game online, kasino online, perlindungan kekayaan intelektual, dan lain-lain.

Inisiatif data terbuka yang dilakukan pemerintah (data.gov, data.gov.uk untuk data publik terbuka, atau govdata.de) adalah contoh terkini dari semakin pentingnya data publik dan nirlaba. Ada inisiatif serupa di banyak negara. Sebagian besar data yang dikumpulkan oleh lembaga publik dan pemerintah negara-negara tersebut pada prinsipnya tersedia untuk digunakan kembali. Panduan W3C mengenai keterbukaan data pemerintah (Bennett dan Harvey 2009) menyarankan bahwa data harus dipublikasikan segera setelah tersedia dalam format mentah aslinya, kemudian disempurnakan dengan semantik dan metadata. Namun, dalam banyak kasus, pemerintah kesulitan mempublikasikan data tertentu, karena faktanya data tersebut harus bersifat non-pribadi dan tidak sensitif serta mematuhi peraturan privasi dan perlindungan data. Banyak sektor dan pelaku yang dapat memperoleh manfaat dari data publik ini.

Berikut ini disajikan beberapa studi kasus penerapan teknologi Big Data di berbagai bidang sektor publik.

Daerah Pemungutan Pajak

Salah satu bidang utama solusi Big Data adalah pemulihan pendapatan pajak sebesar jutaan per tahun. Tantangan bagi penerapan semacam ini adalah untuk mengembangkan resolusi identitas yang cepat dan akurat serta kemampuan mencocokkan untuk departemen pajak negara bagian yang memiliki anggaran terbatas dan staf terbatas untuk menentukan di mana harus mengerahkan sumber daya audit yang langka dan meningkatkan efisiensi pengumpulan pajak. Sorotan implementasi utama adalah:

a. Mengidentifikasi dengan cepat kecocokan yang tepat dan mirip
b. Aktifkan de-duplikasi dari kesalahan entri data
c. Throughput dan skalabilitas yang tinggi menangani pertumbuhan volume data
d. Mengakomodasi perubahan format file dengan cepat dan mudah, serta penambahan sumber data baru

Salah satu solusi didasarkan pada perangkat lunak yang dikembangkan oleh perusahaan Pervasive Software: mesin Pervasive DataRush, Pervasive DataMatcher, dan Pervasive Data Integrator. DataRush Pervasif menyediakan konstruksi sederhana untuk:

a. Membuat unit-unit kerja (proses) yang masing-masing dapat dibuat paralel.
b. Menyatukan proses-proses dalam grafik aliran data (rakitan), namun kemudian memungkinkan penggunaan kembali rakitan kompleks sebagai operator sederhana dalam aplikasi lain.
c. Lebih mengikat operator ke dalam aplikasi aliran data baru yang lebih luas.
d. Jalankan kompiler yang dapat melintasi semua sub-rakitan sambil menjalankan penyesuai untuk secara otomatis menentukan strategi eksekusi paralel berdasarkan sumber daya yang ada saat itu dan/atau heuristik yang lebih kompleks (hal ini akan semakin membaik seiring berjalannya waktu).

Hal ini dicapai dengan menggunakan teknik seperti pencocokan fuzzy, penautan rekaman, dan kemampuan untuk mencocokkan kombinasi bidang apa pun dalam kumpulan data. Teknik utama lainnya mencakup integrasi data dan proses Ekstrak, Transformasi, Muat (ETL) yang menyimpan dan menyimpan semua metadata desain dalam repositori desain terbuka berbasis XML untuk memudahkan pertukaran dan penggunaan kembali metadata. Hal ini memungkinkan implementasi dan penerapan yang cepat serta mengurangi biaya seluruh proses integrasi.

Konsumsi Energi

Sebuah artikel melaporkan permasalahan dalam regulasi konsumsi energi. Persoalan utamanya adalah ketika energi dialirkan ke jaringan distribusi, maka energi tersebut harus digunakan pada saat itu juga. Penyedia energi sedang bereksperimen dengan perangkat penyimpanan untuk membantu mengatasi masalah ini, namun perangkat tersebut masih baru dan mahal. Oleh karena itu masalah ini diatasi dengan perangkat pengukuran pintar.

Saat mengumpulkan data dari perangkat pengukuran pintar, tantangan pertama adalah menyimpan data dalam jumlah besar. Misalnya, dengan asumsi 1 juta perangkat pengumpulan mengambil 5 kB data per pengumpulan tunggal, potensi pertumbuhan volume data dalam satu tahun bisa mencapai 2.920 TB.

Tantangan selanjutnya adalah menganalisis volume data yang sangat besar ini, melakukan referensi silang data tersebut dengan informasi pelanggan, distribusi jaringan, dan informasi kapasitas berdasarkan segmen, informasi cuaca lokal, dan data biaya pasar spot energi. Memanfaatkan data ini akan memungkinkan perusahaan utilitas untuk lebih memahami struktur biaya dan pilihan strategis dalam jaringan mereka, yang dapat mencakup:

a. Menambah kapasitas pembangkit dibandingkan membeli energi di luar pasar (misalnya energi terbarukan seperti tenaga angin, tenaga surya, mobil listrik di luar jam sibuk)
b. Berinvestasi pada perangkat penyimpanan energi dalam jaringan untuk mengimbangi penggunaan puncak dan mengurangi pembelian dan biaya spot
c. Memberikan insentif kepada konsumen individu, atau kelompok konsumen, untuk mengubah perilaku konsumsi energi

Salah satu pendekatan dari perusahaan Lavastorm adalah proyek yang mengeksplorasi masalah analitik dengan perusahaan inovatif seperti FalbygdensEnergi AB (FEAB) dan Sweco. Untuk menjawab pertanyaan-pertanyaan kunci, Platform Analitik Lavastorm digunakan. Lavastorm Analytics Engine adalah solusi analisis bisnis mandiri yang memberdayakan analis untuk memperoleh, mengubah, menganalisis, dan memvisualisasikan data dengan cepat, serta berbagi wawasan penting dan jawaban tepercaya atas pertanyaan bisnis dengan manajer dan eksekutif non-teknis. Mesin ini menawarkan serangkaian kemampuan analitik terintegrasi yang memungkinkan analis untuk secara mandiri mengeksplorasi data perusahaan dari berbagai sumber data, membuat dan berbagi model analitik tepercaya, menghasilkan perkiraan akurat, dan mengungkap wawasan yang sebelumnya tersembunyi dalam satu lingkungan tunggal yang sangat visual dan terukur.

Media dan Hiburan

Media dan hiburan berpusat pada pengetahuan yang disertakan dalam file media. Dengan pertumbuhan signifikan file media dan metadata terkait, akibat evolusi Internet dan web sosial, perolehan data di sektor ini telah menjadi tantangan besar.

Menurut laporan Quantum, mengelola dan berbagi konten dapat menjadi sebuah tantangan, terutama bagi industri media dan hiburan. Dengan kebutuhan untuk mengakses rekaman video, file audio, gambar beresolusi tinggi, dan konten lainnya, diperlukan solusi berbagi data yang andal dan efektif.

Alat yang umum digunakan di sektor media dan hiburan meliputi:

- Sistem file khusus yang digunakan sebagai alternatif berkinerja tinggi dibandingkan NAS dan jaringan berbagi
- Teknologi pengarsipan khusus yang memungkinkan pembuatan arsip digital yang mengurangi biaya dan melindungi konten
- Klien khusus yang memungkinkan aplikasi berbasis LAN dan aplikasi berbasis SAN untuk berbagi kumpulan konten tunggal
- Berbagai solusi penyimpanan khusus (untuk berbagi file berkinerja tinggi, penyimpanan near-line yang hemat biaya, retensi data offline, untuk penyimpanan primer berkecepatan tinggi)

Layanan digital on-demand telah secara radikal mengubah pentingnya jadwal bagi konsumen dan lembaga penyiaran. Perusahaan media terbesar telah banyak berinvestasi pada infrastruktur teknis untuk mendukung penyimpanan dan streaming konten. Misalnya, jumlah situs pengunduhan dan streaming musik legal, serta layanan radio Internet, telah meningkat pesat dalam beberapa tahun terakhir—konsumen memiliki pilihan yang hampir membingungkan tergantung pada genre musik, opsi berlangganan, perangkat, manajemen hak digital. (DRM) yang mereka sukai. Lebih dari 391 juta lagu terjual di Eropa pada tahun 2012, dan 75 juta lagu diputar di stasiun radio online.

Menurut stat, telah terjadi peningkatan besar dalam akses rumah tangga terhadap broadband sejak tahun 2006. Di seluruh "EU27" (negara anggota UE dan enam negara lain di wilayah geografis Eropa) penetrasi broadband berada pada kisaran 30% pada tahun 2006 namun mencapai 72% pada tahun 2012. Bagi rumah tangga dengan broadband berkecepatan tinggi, streaming media adalah cara yang sangat menarik dalam mengonsumsi konten. Demikian pula, kecepatan unggah yang lebih cepat berarti orang dapat membuat video mereka sendiri untuk platform media sosial.

Telah terjadi pergeseran besar dari media arus utama yang bersifat massal dan anonim, menuju pengalaman yang dipersonalisasi dan berdasarkan permintaan. Pengalaman konsumen bersama dalam skala besar seperti acara olahraga besar, reality show, dan sinetron kini menjadi populer. Konsumen berharap dapat menonton atau mendengarkan apa pun yang mereka inginkan, kapan pun mereka mau.

Layanan streaming memberikan kendali kepada pengguna yang memilih kapan akan menonton acara, konten web, atau musik favorit mereka. Perusahaan media terbesar telah banyak berinvestasi pada infrastruktur teknis untuk mendukung penyimpanan dan streaming

konten. Perusahaan media menyimpan sejumlah besar data pribadi, baik data pelanggan, pemasok, konten, atau karyawan mereka sendiri. Perusahaan mempunyai tanggung jawab tidak hanya terhadap diri mereka sendiri sebagai pengontrol data, namun juga penyedia layanan cloud (pemroses data). Banyak organisasi media besar dan kecil telah mengalami pelanggaran data yang parah dua di antara korban paling terkenal adalah Sony dan LinkedIn. Mereka tidak hanya menanggung biaya untuk memperbaiki pelanggaran data, namun juga denda dari badan perlindungan data seperti Kantor Komisaris Informasi (ICO) di Inggris.

Keuangan dan Asuransi

Mengintegrasikan data dalam jumlah besar dengan sistem intelijen bisnis untuk analisis memainkan peran penting dalam sektor keuangan dan asuransi. Beberapa bidang utama untuk memperoleh data di sektor ini adalah pasar valuta asing, investasi, perbankan, profil pelanggan, dan perilaku.

Menurut McKinsey Global Institute Analysis, "Layanan Keuangan mempunyai manfaat terbesar dari Big Data". Untuk kemudahan menangkap dan menilai potensi, "pemain keuangan mendapatkan nilai tertinggi untuk peluang penciptaan nilai". Bank dapat menambah nilai dengan meningkatkan sejumlah produk, misalnya menyesuaikan UX, meningkatkan penargetan, mengadaptasi model bisnis, mengurangi kerugian portofolio dan biaya modal, efisiensi kantor, dan proposisi nilai baru. Beberapa data keuangan yang tersedia untuk umum disediakan oleh lembaga statistik internasional seperti stat, Bank Dunia, Bank Sentral Eropa, Dana Moneter Internasional, Perusahaan Keuangan Internasional, Organisasi untuk Kerja Sama dan Pembangunan Ekonomi. Meskipun sumber data ini tidak terlalu sensitif terhadap waktu dibandingkan dengan pasar bursa, sumber data ini memberikan data pelengkap yang berharga.

Deteksi penipuan adalah topik penting dalam keuangan. Menurut Studi Penipuan Global tahun 2014, sebuah organisasi pada umumnya kehilangan sekitar 5% pendapatannya setiap tahun karena penipuan. Sektor perbankan dan jasa keuangan mempunyai banyak penipuan. Sekitar 30 % skema penipuan terdeteksi melalui tip off dan hingga 10 % secara tidak sengaja, namun hanya hingga 1 % melalui pengendalian TI (ACFE 2014). Metode deteksi penipuan yang lebih baik dan lebih baik mengandalkan analisis data besar secara real-time. Untuk metode deteksi penipuan yang lebih akurat dan tidak terlalu mengganggu, bank dan lembaga jasa keuangan semakin banyak menggunakan algoritme yang mengandalkan data transaksi real-time. Teknologi ini memanfaatkan data dalam jumlah besar yang dihasilkan dengan kecepatan tinggi dan dari sumber hibrid. Seringkali, data dari sumber seluler dan data sosial seperti informasi geografis digunakan untuk prediksi dan deteksi. Dengan menggunakan algoritma pembelajaran mesin, sistem modern mampu mendeteksi penipuan dengan lebih andal dan lebih cepat. Namun ada batasan untuk sistem seperti itu. Karena layanan keuangan beroperasi dalam lingkungan peraturan, penggunaan data pelanggan tunduk pada undang-undang dan peraturan privasi.

Kesimpulan

Akuisisi data merupakan proses yang penting dan memungkinkan alat-alat berikutnya dalam Jaringan nilai data melakukan tugasnya dengan benar (misalnya alat analisis data).

Kecanggihan alat akuisisi data menunjukkan bahwa ada banyak alat dan protokol, termasuk solusi sumber terbuka yang mendukung proses akuisisi data. Banyak dari alat ini telah dikembangkan dan dioperasikan dalam lingkungan produksi atau pemain besar seperti Facebook atau Amazon.

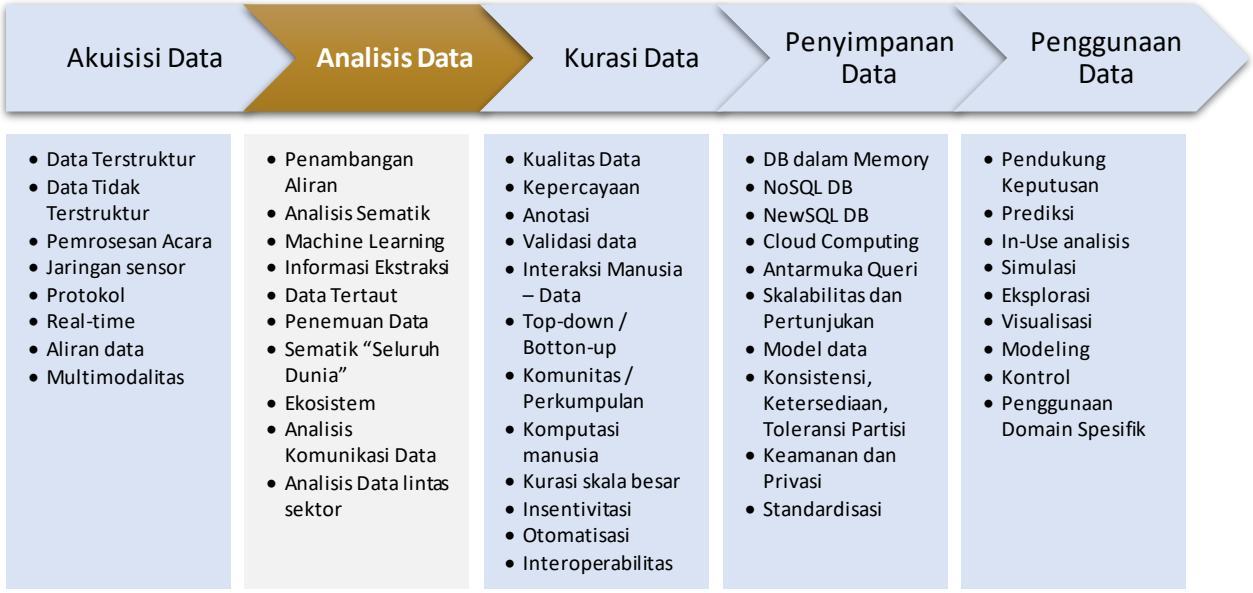
Meskipun demikian, terdapat banyak tantangan terbuka agar berhasil menerapkan solusi Big Data yang efektif untuk akuisisi data di berbagai sektor (lihat bagian “Persyaratan Masa Depan dan Tren yang Muncul untuk Akuisisi Big Data”). Masalah utamanya adalah menghasilkan solusi yang kuat dan memiliki skalabilitas tinggi untuk saat ini dan meneliti sistem generasi berikutnya untuk memenuhi kebutuhan industri yang terus meningkat.

BAB 5
ANALISIS DATA BESAR

5.1 PENDAHULUAN

Data hadir dalam berbagai bentuk dan satu dimensi untuk mempertimbangkan dan membandingkan format data yang berbeda adalah jumlah struktur yang terkandung di dalamnya. Semakin banyak struktur yang dimiliki kumpulan data, semakin besar kemungkinannya untuk diproses oleh mesin. Pada tingkat ekstrim, representasi semantik akan memungkinkan penalaran mesin. Analisis data besar adalah sub-bidang data besar yang berkaitan dengan penambahan struktur pada data untuk mendukung pengambilan keputusan serta mendukung skenario penggunaan spesifik domain. Bab ini menguraikan wawasan utama, kecanggihan, tren yang muncul, kebutuhan masa depan, dan studi kasus sektoral untuk analisis data.

Posisi analisis Big Data dalam keseluruhan jaringan nilai Big Data dapat dilihat pada gambar 5.1. Data yang mungkin terstruktur atau tidak dan biasanya terdiri dari berbagai format berbeda diubah agar siap untuk kurasi data, penyimpanan data, dan penggunaan data. Itu sebabnya tanpa analisis Big Data, sebagian besar data yang diperoleh tidak akan berguna.



Gambar 5.1 Analisis Data Dalam Jaringan Nilai Big Data

Analisis menemukan bahwa teknik umum berikut berguna saat ini atau akan berguna dalam jangka pendek hingga menengah: penalaran (termasuk penalaran aliran), pemrosesan semantik, penambangan data, pembelajaran mesin, ekstraksi informasi, dan penemuan data.

Wilayah umum ini bukanlah hal baru. Namun yang baru adalah tantangan yang ditimbulkan oleh karakteristik spesifik Big Data terkait tiga V:

1. **Volume**— menempatkan skalabilitas sebagai inti dari seluruh pemrosesan. Diperlukan penalaran skala besar, pemrosesan semantik, penambangan data, pembelajaran mesin, dan ekstraksi informasi.
2. **Velocity**— tantangan ini mengakibatkan munculnya area pemrosesan data aliran, penalaran aliran, dan penambangan data aliran untuk mengatasi tingginya volume data mentah yang masuk.
3. **Variasi**— dapat berupa format sintaksis yang berbeda (misalnya spreadsheet vs. csv) atau skema data yang berbeda atau makna berbeda yang melekat pada bentuk sintaksis yang sama (misalnya Paris sebagai kota atau orang). Teknik semantik, terutama yang berkaitan dengan Data Tertaut, telah terbukti paling berhasil diterapkan sejauh ini meskipun masalah skalabilitas masih harus diatasi.

5.2 WAWASAN UTAMA UNTUK ANALISIS BIG DATA

Wawancara dengan berbagai pemangku kepentingan terkait analisis Big Data telah mengidentifikasi wawasan utama berikut ini. Daftar lengkap orang yang diwawancarai disajikan pada Tabel 5.1. Teknologi lama diterapkan dalam konteks baru teknologi individu dan kombinasi teknologi lama diterapkan dalam konteks Big Data. Perbedaannya terletak pada skala (volume) dan besarnya heterogenitas yang ditemui (variasi). Secara khusus, dalam konteks web, fokusnya terlihat pada kumpulan data besar berbasis semantik seperti Freebase dan pada ekstraksi data berkualitas tinggi dari web. Selain skala, ada hal baru dalam kenyataan bahwa teknologi-teknologi ini hadir bersamaan.

Tabel 5.1 Analisis Big Data Narasumber

NO.	NAMA DEPAN	NAMA KELUARGA	ORGANISASI	PERAN/POSISI
1	S€OREN	AUER	LEIPZIG	PROFESOR
2	RICARDO	BAEZA- YATES	YAHOO!	WAKIL PRESIDEN PENELITIAN
3	FRANC¸OIS	BANCILHON	DATA PUBLICA	CEO
4	RICHARD	BENJAMINS	TELEFONCICA	DIREKTUR BIZ INTEL
5	HJALMAR	GISLASON	DATAMARKET.COM	PENDIRI
6	ALON	HALVEY	GOOGLE	ILMUWAN RISET
7	USMAN	HAQUE	COSM (PACHUBE)	DIREKTUR DIVISI PROYEK PERKOTAAN
8	STEVE	HARRIS	GARLIK/EXPERIAN	CTO
9	JIM	HENDLER	RPI	PROFESOR
10	ALEK	KOŁCZ	TWITTER	ILMUWAN DATA
11	PRASANNA	LAL DAS	WORLD BANK	SNR PROG. PETUGAS, KEPALA PROGRAM DATA KEUANGAN TERBUKA
12	PETER	MIKA	YAHOO!	PENELITI
13	ANDREAS	RIBBROCK	TERADATA GMBH	PEMIMPIN TIM ANALISIS BIG DATA DAN ARSITEK SENIOR

14	JENI	TENNISON	OPEN DATA INSTITUTE	DIREKTUR TEKNIS
15	BILL	THOMPSON	BBC	KEPALA PENGEMBANGAN MITRA
16	ANDRAZˇ	TORI	ZEMANTA	PEMILIK DAN CTO
17	FRANK	VAN HARMELEN	AMSTERDAM	PROFESOR
18	MARCO	VICECONTI	UNIVERSITY OF SHEFFIELD AND THE VPH INSTITUTE	PROFESOR DAN DIREKTUR
19	JIM	WEBBER	NEO	KEPALA ILMUWAN

Penambangan Data Aliran Ini diperlukan untuk menangani aliran data dalam jumlah besar yang berasal dari jaringan sensor atau aktivitas online dari sejumlah besar pengguna. Kemampuan ini akan memungkinkan organisasi untuk memberikan personalisasi yang sangat adaptif dan akurat. Penemuan data yang menarik pertanyaan yang sering diajukan oleh pengguna dan pengembang adalah: Di mana kita bisa mendapatkan data tentang X? Dimana kita bisa mendapatkan informasi tentang Y? Sulit untuk menemukan data dan data yang ditemukan sering kali kedaluwarsa dan tidak dalam format yang benar. Crawler diperlukan untuk menemukan kumpulan data besar, metadata untuk data besar, hubungan bermakna antara kumpulan data terkait, dan mekanisme pemeringkatan kumpulan data yang berfungsi sama baiknya dengan Page Rank untuk dokumen web.

Berurusan dengan Data yang Sangat Luas dan Sangat Spesifik Hal yang paling menarik dalam ekstraksi informasi dari web adalah bahwa web adalah tentang segala hal sehingga cakupannya luas. Pra-web fokusnya adalah pada domain tertentu ketika membangun database dan basis pengetahuan. Hal ini tidak lagi dapat dilakukan dalam konteks web. Seluruh gagasan mengkonseptualisasikan domain telah diubah: Sekarang domain adalah segalanya di dunia. Sisi positifnya, manfaatnya adalah Anda mendapatkan banyak keluasan, dan tantangan penelitiannya adalah bagaimana seseorang dapat mendalami suatu domain sambil mempertahankan konteks yang luas.

Kesederhanaan menghasilkan adopsi Hadoop berhasil karena merupakan alat termudah untuk digunakan oleh pengembang, mengubah permainan di bidang data besar. Hal ini tidak berhasil karena merupakan yang terbaik namun karena merupakan yang termudah untuk digunakan (bersama dengan HIVE). Hadoop berhasil menyeimbangkan penanganan kompleksitas (pemrosesan data besar) dan kesederhanaan bagi pengembang. Sebaliknya, teknologi semantik seringkali sulit digunakan. Hjalmar Gislason, salah satu narasumber kami menganjurkan perlunya “demokratisasi teknologi semantik”.

Ekosistem yang Dibangun di Sekitar Kumpulan Alat Memiliki Dampak yang Signifikan Hal ini sering kali didorong oleh perusahaan besar di mana teknologi diciptakan untuk memecahkan masalah internal dan kemudian diberikan begitu saja. Apache Cassandra3 adalah contoh yang awalnya dikembangkan oleh Facebook untuk mendukung fitur pencarian kotak masuk mereka hingga tahun 2010. Ekosistem di sekitar Hadoop mungkin yang paling terkenal.

Komunitas dan Big Data Akan Dilibatkan dalam Hubungan Baru dan Menarik Komunitas akan dilibatkan dengan Big Data di seluruh tahapan Jaringan nilai dan dalam berbagai cara. Secara khusus, masyarakat akan dilibatkan secara mendalam dalam pengumpulan data, meningkatkan keakuratan data, dan penggunaan data. Big Data juga akan meningkatkan keterlibatan komunitas dalam masyarakat secara umum.

Penggunaan Big Data lintas sektoral akan membuka peluang bisnis baru bagian ritel mengenai kebutuhan masa depan dan tren yang sedang berkembang menjelaskan contohnya. O2 UK bersama dengan Telef´onica digital baru-baru ini meluncurkan layanan yang memetakan dan menggunakan kembali data seluler untuk industri ritel. Layanan ini memungkinkan pengecer untuk merencanakan lokasi gerai ritel berdasarkan pergerakan harian calon pelanggan. Layanan ini menyoroti pentingnya data besar internal (dalam hal ini catatan seluler) yang kemudian digabungkan dengan sumber data eksternal (data geografis dan preferensi) untuk menghasilkan jenis bisnis baru. Secara umum, pengumpulan data lintas organisasi dan sektor akan meningkatkan daya saing industri Eropa. Tantangan terbesar bagi sebagian besar industri saat ini adalah menggabungkan teknologi Big Data dalam proses dan infrastruktur mereka. Banyak perusahaan mengidentifikasi kebutuhan untuk melakukan analisis data besar, namun tidak memiliki sumber daya untuk menyiapkan infrastruktur untuk menganalisis dan memelihara saluran analisis (Benjamins). Meningkatkan kesederhanaan teknologi akan membantu tingkat adopsi. Selain itu, sejumlah besar pengetahuan domain harus dibangun dalam setiap industri tentang bagaimana data dapat digunakan; apa yang berharga untuk diekstraksi dan keluaran apa yang dapat digunakan dalam operasi sehari-hari.

Biaya penerapan analisis Big Data merupakan hambatan bisnis dalam penerapan teknologi Big Data. Anonimitas, privasi, dan perlindungan data merupakan persyaratan lintas sektoral yang disoroti untuk teknologi Big Data. Informasi tambahan dapat ditemukan dalam analisis akhir mengenai kebutuhan sektor ini.

5.3 ANALISIS BIG DATA TERCANGGIH

Industri saat ini menerapkan pembelajaran mesin berskala besar dan algoritme lain untuk menganalisis kumpulan data berukuran besar, dikombinasikan dengan pemrosesan peristiwa kompleks dan pemrosesan aliran untuk analisis waktu nyata. Ditemukan juga bahwa tren terkini mengenai Data Tertaut, teknologi semantik, dan penalaran skala besar adalah beberapa topik yang disorot oleh para ahli yang diwawancarai sehubungan dengan tantangan penelitian utama dan persyaratan teknologi utama untuk data besar. Bagian ini menyajikan tinjauan terkini mengenai analisis data besar dan literatur yang diterbitkan, menguraikan berbagai topik mulai dari bekerja secara efisien dengan data hingga pengelolaan data skala besar.

Skala Besar: Penalaran, Pembandingan, dan Pembelajaran Mesin

Ukuran dan heterogenitas web menghalangi dilakukannya penalaran penuh dan memerlukan solusi teknologi baru untuk memenuhi kemampuan inferensi yang diminta. Fitur yang diminta ini juga telah diperluas ke teknologi pembelajaran mesin dan teknologi ini diperlukan untuk mengekstrak informasi berguna dari data dalam jumlah besar. Secara

khusus, Franc¸ois Bancilhon menyebutkan dalam wawancaranya bagaimana pembelajaran mesin penting untuk deteksi topik dan klasifikasi dokumen di Data Publica. Kemudian, Ricardo Baeza-Yates dalam wawancaranya menyoroti perlunya standar dalam komputasi Big Data agar penyedia Big Data dapat membandingkan sistem mereka.

Penalaran Skala Besar

Janji penalaran seperti yang dipromosikan dalam konteks web semantik saat ini tidak sesuai dengan persyaratan data besar karena masalah skalabilitas. Penalaran ditentukan oleh prinsip-prinsip tertentu, seperti kesehatan dan kelengkapan, yang jauh dari dunia praktis dan karakteristik web, di mana data sering kali bertentangan, tidak lengkap, dan berukuran sangat besar. Selain itu, terdapat kesenjangan antara penalaran pada skala web dan penalaran yang lebih disesuaikan pada himpunan bagian logika tingkat pertama yang disederhanakan, karena fakta bahwa banyak aspek yang diasumsikan berbeda dari kenyataan (misalnya sekumpulan kecil aksioma dan fakta, kelengkapan dan kebenaran aturan inferensi).

Pendekatan modern (Fensel 2007) mengusulkan kombinasi metode penalaran dan pengambilan informasi berdasarkan teknik pencarian, untuk mengatasi masalah penalaran skala web. Penalaran yang tidak lengkap dan perkiraan ditonjolkan oleh Frank van Harmelen sebagai topik penting dalam wawancaranya. Kueri dan penalaran atas data terstruktur dapat didukung oleh model semantik yang secara otomatis dibangun dari pola kemunculan kata dari kumpulan teks besar (model semantik distribusional). Model semantik distribusi memberikan lapisan makna pelengkap untuk data terstruktur, yang dapat digunakan untuk mendukung perkiraan semantik untuk menanyakan dan menalar data heterogen.

Kombinasi penalaran berbasis logika dengan pengambilan informasi adalah salah satu aspek kunci dari pendekatan ini dan juga teknik pembelajaran mesin, yang memberikan trade-off antara aspek penalaran yang lengkap dan kepraktisan dari aspek-aspek tersebut di web. konteks. Ketika topik skalabilitas muncul, sistem penyimpanan juga memainkan peran penting, terutama teknik pengindeksan dan strategi pengambilan. Pertukaran antara penalaran online (mundur) dan penalaran offline (maju) disebutkan oleh Frank van Harmelen dalam wawancaranya. Peter Mika juga menguraikan pentingnya teknik pengindeksan yang efisien dalam wawancaranya. Dengan topik sistem skala besar, LarKC adalah proyek unggulan. LarKC4 adalah Proyek Integrasi Skala Besar FP7 UE dan tujuannya adalah untuk menangani sistem dan teknik penalaran skala besar yang menggunakan teknologi semantik.

Pembandingan untuk Repositori Berskala Besar

Pembandingan merupakan hal yang baru dalam bidang pemrosesan data semantik skala besar, dan faktanya saat ini baru diproduksi. Secara khusus, proyek *Linked Data Benchmark Council* (LDBC) bertujuan untuk membuat serangkaian tolok ukur untuk grafik skala besar dan pengelolaan data RDF *(Resource Description Framework)* serta membentuk otoritas independen untuk mengembangkan tolok ukur. Bagian dari rangkaian tolok ukur yang dibuat di LDBC adalah tolok ukur dan pengujian integrasi data dan fungsi penalaran yang didukung oleh sistem RDF. Tolok ukur ini difokuskan pada pengujian: (1) pencocokan instance dan ekstrak, transformasi, dan muat yang memainkan peran penting dalam integrasi data; dan (2) kemampuan penalaran mesin RDF yang ada. Kedua topik ini sangat penting dalam

praktiknya, dan keduanya sebagian besar diabaikan oleh tolok ukur pemrosesan data tertaut yang ada. Dalam membuat tolok ukur tersebut, LDBC menganalisis berbagai skenario yang tersedia untuk mengidentifikasi skenario yang paling dapat menampilkan integrasi data dan fungsi penalaran mesin RDF. Berdasarkan skenario ini, keterbatasan sistem RDF yang ada diidentifikasi untuk mengumpulkan serangkaian persyaratan untuk integrasi data RDF dan tolok ukur penalaran. Misalnya, sudah diketahui bahwa sistem yang ada tidak akan berfungsi dengan baik jika terdapat aturan penalaran yang tidak standar (misalnya penalaran tingkat lanjut yang mempertimbangkan negasi dan agregasi). Selain itu, para pemikir yang ada melakukan inferensi dengan mewujudkan penutupan kumpulan data (menggunakan Jaringan mundur atau maju). Namun, pendekatan ini mungkin tidak dapat diterapkan ketika aturan penalaran khusus aplikasi disediakan dan oleh karena itu kemungkinan besar peningkatan yang canggih akan berarti dukungan untuk strategi penalaran hibrid yang melibatkan rangkaian ke belakang dan ke depan, dan penulisan ulang kueri (yaitu menggabungkan rangkaian aturan dalam kueri).

Pembelajaran Mesin Skala Besar

Algoritme pembelajaran mesin menggunakan data untuk secara otomatis mempelajari cara melakukan tugas seperti prediksi, klasifikasi, dan deteksi anomali. Sebagian besar algoritme pembelajaran mesin telah dirancang untuk berjalan secara efisien pada satu prosesor atau inti. Perkembangan arsitektur multi-core dan komputasi grid telah menyebabkan meningkatnya kebutuhan akan pembelajaran mesin untuk memanfaatkan ketersediaan beberapa unit pemrosesan. Ada banyak antarmuka dan bahasa pemrograman yang didedikasikan untuk pemrograman paralel seperti Orca MPI atau OpenACC, yang berguna untuk pemrograman paralel tujuan umum. Namun, tidak selalu jelas bagaimana algoritma pembelajaran mesin yang ada dapat diimplementasikan secara paralel. Ada banyak penelitian tentang pembelajaran terdistribusi dan penambangan data (Bhaduri et al. 2011), yang mencakup algoritma pembelajaran mesin yang telah dirancang khusus untuk tujuan komputasi terdistribusi.

Daripada membuat algoritma versi paralel tertentu, pendekatan yang lebih umum melibatkan kerangka kerja untuk pemrograman pembelajaran mesin pada beberapa unit pemrosesan. Salah satu pendekatannya adalah dengan menggunakan abstraksi tingkat tinggi yang secara signifikan menyederhanakan desain dan implementasi kelas algoritma paralel yang terbatas. Secara khusus, abstraksi MapReduce telah berhasil diterapkan pada berbagai aplikasi pembelajaran mesin. Hal menunjukkan bahwa algoritme apa pun yang sesuai dengan model kueri statistik dapat ditulis dalam bentuk penjumlahan tertentu, yang dapat dengan mudah diimplementasikan dalam mode MapReduce dan mencapai percepatan mendekati linier dengan jumlah unit pemrosesan yang digunakan. Mereka menunjukkan bahwa hal ini berlaku untuk berbagai algoritma pembelajaran. Implementasi yang ditunjukkan dalam makalah ini menghasilkan versi pertama perpustakaan pembelajaran mesin MapReduce. Paradigma MapReduce membatasi pengguna untuk menggunakan asumsi pemodelan yang terlalu sederhana untuk memastikan tidak ada ketergantungan komputasi dalam pemrosesan data. Mereka mengusulkan abstraksi Graphlab yang mengisolasi pengguna dari kompleksitas

pemrograman paralel (yaitu data race, deadlock), sambil mempertahankan kemampuan untuk mengekspresikan ketergantungan komputasi yang kompleks menggunakan grafik data.

Bahasa pemrograman, toolkit, dan kerangka kerja yang dibahas memungkinkan banyak konfigurasi berbeda untuk melaksanakan pembelajaran mesin skala besar. Konfigurasi ideal yang digunakan bergantung pada aplikasi, karena aplikasi yang berbeda akan memiliki serangkaian persyaratan yang berbeda. Namun, salah satu kerangka kerja paling populer yang digunakan dalam beberapa tahun terakhir adalah Apache Hadoop, yang merupakan implementasi paradigma MapReduce bersumber terbuka dan gratis yang dibahas di atas. Andrazˇ Tori, salah satu narasumber kami, mengidentifikasi kesederhanaan Hadoop dan MapReduce sebagai pendorong utama kesuksesannya. Ia menjelaskan bahwa implementasi Hadoop dapat mengungguli dalam hal waktu komputasi, misalnya dengan implementasi yang menggunakan OpenMP, namun Hadoop menang dalam hal popularitas karena mudah digunakan. Upaya komputasi paralel yang dijelaskan di atas memungkinkan pemrosesan data dalam jumlah besar. Selain penerapan metode yang ada pada kumpulan data yang semakin besar, peningkatan daya komputasi juga mengarah pada pendekatan pembelajaran mesin berskala besar yang baru. Salah satu contohnya adalah karya terbaru dari Le et al. (2011) yang menggunakan kumpulan data sepuluh juta gambar untuk mengajarkan detektor wajah hanya dengan menggunakan data yang tidak berlabel. Penggunaan fitur yang dihasilkan dalam tugas pengenalan objek menghasilkan peningkatan kinerja sebesar 70% melebihi yang canggih. Memanfaatkan data dalam jumlah besar untuk mengatasi kebutuhan akan data pelatihan berlabel bisa menjadi tren penting. Dengan hanya menggunakan data yang tidak berlabel, salah satu hambatan terbesar dalam penerapan pembelajaran mesin secara luas dapat diatasi. Penggunaan metode pembelajaran tanpa pengawasan memiliki keterbatasan dan masih harus dilihat apakah teknik serupa juga dapat diterapkan di domain aplikasi lainnya.

Pengolahan Data Aliran

Penambangan data aliran disorot sebagai bidang penelitian yang menjanjikan oleh Ricardo Baeza-Yates dalam wawancaranya. Teknik ini berkaitan dengan kemampuan teknologi yang diperlukan untuk menangani aliran data dengan volume dan kecepatan tinggi, yang berasal dari jaringan sensor, atau aktivitas online lainnya yang melibatkan banyak pengguna.

Pencocokan Pola Aliran Data RDF

Termotivasi oleh banyaknya data terstruktur dan tidak terstruktur yang tersedia di web sebagai aliran berkelanjutan, teknik pemrosesan streaming menggunakan teknologi web baru-baru ini muncul. Untuk memproses aliran data di web, penting untuk mengatasi keterbukaan dan heterogenitas. Masalah inti dari sistem pemrosesan aliran data adalah memproses data dalam jangka waktu tertentu dan dapat membuat kueri pola. Fitur tambahan yang diinginkan mencakup dukungan data statis yang tidak akan berubah seiring waktu dan dapat digunakan untuk menyempurnakan data dinamis. Operator temporal dan jendela berbasis waktu juga biasanya ditemukan dalam sistem ini, digunakan untuk menggabungkan beberapa grafik RDF dengan ketergantungan waktu. Beberapa perkembangan besar di bidang ini adalah C-SPARQL (Barbieri et al. 2010) ETALIS (Anicic et al. 2011), dan SPARKWAVE (Komazec et al. 2012).

C-SPARQL adalah bahasa berdasarkan SPARQL (Protokol SPARQL dan Bahasa Kueri RDF) dan diperluas dengan definisi untuk aliran dan jendela waktu. Triple yang masuk pertama-tama diwujudkan berdasarkan RDFS dan kemudian dimasukkan ke dalam sistem evaluasi. C-SPARQL tidak memberikan evaluasi pola berkelanjutan yang sebenarnya, karena penggunaan snapshot RDF, yang dievaluasi secara berkala. Namun kekuatan C-SPARQL terletak pada situasi dengan sejumlah besar pengetahuan statis, yang perlu dikombinasikan dengan aliran data masuk yang dinamis.

ETALIS adalah sistem pemrosesan peristiwa di atas SPARQL. Karena komponen bahasa pola SPARQL diperluas dengan sintaksis pemrosesan peristiwa, bahasa pola tersebut disebut EP-SPARQL. Fitur yang didukung adalah operator temporal, evaluasi out-of-order, fungsi agregat, beberapa mode pengumpulan sampah, dan strategi konsumsi yang berbeda.

SPARKWAVE menyediakan pencocokan pola berkelanjutan melalui aliran data RDF yang disempurnakan dengan skema. Berbeda dengan C-SPARQL dan EP-SPARQL, SPARKWAVE bersifat tetap terkait skema yang digunakan dan tidak mendukung operator temporal atau fungsi agregat. Keuntungan memiliki skema yang tetap dan tidak ada alasan yang rumit adalah sistem dapat mengoptimalkan dan melakukan pra-perhitungan pada tahap inisialisasi struktur pola yang digunakan dalam memori, sehingga menghasilkan throughput yang tinggi saat memproses data RDF yang masuk.

Pemrosesan Peristiwa Kompleks

Salah satu wawasan dari wawancara tersebut adalah bahwa teknologi aliran data besar dapat diklasifikasikan menurut (1) mesin pemrosesan peristiwa yang kompleks, dan (2) infrastruktur pemrosesan aliran yang sangat skalabel. Mesin pemrosesan peristiwa yang kompleks fokus pada aspek bahasa dan eksekusi logika bisnis, sementara infrastruktur pemrosesan aliran menyediakan kerangka komunikasi untuk memproses pesan asinkron dalam skala besar.

Pemrosesan peristiwa kompleks (CEP) menggambarkan serangkaian teknologi yang mampu memproses peristiwa “dalam aliran”, yaitu berbeda dengan pemrosesan batch di mana data dimasukkan ke dalam database dan disurvei secara berkala untuk analisis lebih lanjut. Keuntungan sistem CEP adalah kemampuannya memproses kejadian dalam jumlah besar secara real-time. Nama pemrosesan peristiwa kompleks disebabkan oleh fakta bahwa peristiwa sederhana, misalnya. dari sensor atau data operasional lainnya, dapat dikorelasikan dan diproses menghasilkan kejadian yang lebih kompleks. Pemrosesan tersebut dapat terjadi dalam beberapa langkah, yang pada akhirnya menghasilkan peristiwa menarik yang memicu operator manusia atau intelijen bisnis.

Seperti yang ditunjukkan oleh para ahli, sistem berbasis peristiwa mencakup sejumlah besar fungsi pada berbagai tingkat teknologi (misalnya, bahasa, eksekusi, atau komunikasi)”. Mereka memberikan survei komprehensif yang membantu pemahaman dan klasifikasi sistem pemrosesan peristiwa yang kompleks.

Untuk analitik aliran data besar, ini adalah kemampuan utama yang dapat ditingkatkan skalanya oleh sistem pemrosesan peristiwa yang kompleks untuk memproses semua peristiwa yang masuk secara tepat waktu seperti yang dibutuhkan oleh domain aplikasi. Misalnya saja

data smart meter dari sebuah perusahaan utilitas besar dapat menghasilkan jutaan atau bahkan milyaran kejadian per detik yang dapat dianalisis untuk menjaga keandalan operasional jaringan listrik. Selain itu, mengatasi heterogenitas semantik di balik berbagai sumber data dalam lingkungan pembuatan peristiwa terdistribusi merupakan kemampuan mendasar untuk skenario data besar. Terdapat pendekatan pencocokan peristiwa semantik otomatis yang menargetkan skenario dengan tipe peristiwa yang heterogen. Contoh mesin pemrosesan kejadian yang kompleks mencakup SAP Sybase Event Stream Processor, IBM InfoSphere Stream, dan ruleCore, dan masih banyak lagi.

Penggunaan Data Tertaut dan Pendekatan Semantik pada Analisis Big Data

Menurut Tim Berners-Lee dan rekan-rekannya (Bizer dkk. 2009), "*Data Tertaut hanyalah tentang penggunaan Web untuk membuat tautan yang diketik antara data dari sumber berbeda"*. Data tertaut mengacu pada data yang dapat dibaca mesin, ditautkan ke kumpulan data lain dan dipublikasikan di web sesuai dengan serangkaian praktik terbaik yang dibangun berdasarkan teknologi web seperti *HTTP (Hypertext Transfer Protocol), RDF, dan URI (Uniform Resource Identifier)*. Teknologi semantik seperti SPARQL, OWL, dan RDF memungkinkan seseorang untuk mengelola dan menangani hal ini. Berdasarkan prinsip data tertaut, ruang data mengelompokkan semua sumber data yang relevan ke dalam repositori bersama yang terpadu (Heath dan Bizer 2011). Oleh karena itu, ruang data menawarkan solusi yang baik untuk menutupi heterogenitas web integrasi skala besar dan menangani tipe data yang luas dan spesifik.

Data terkait dan pendekatan semantik terhadap analisis Big Data telah disorot oleh sejumlah narasumber termasuk S€oren Auer, Franc¸ois Bancilhon, Richard Benjamins, Hjalmar Gislason, Frank van Harmelen, Jim Hendler, Peter Mika, dan Jeni Tennison. Teknologi-teknologi ini disorot karena mampu mengatasi tantangan-tantangan penting terkait Big Data, termasuk pengindeksan yang efisien, ekstraksi dan klasifikasi entitas, serta pencarian data yang ditemukan di web.

Peringkasan Entitas

Sejauh pengetahuan kami, peringkasan entitas pertama kali disebutkan dalam Cheng et al. (2008). Penulis menyajikan Falcons yang menyediakan pencarian berbasis kata kunci untuk entitas Web Semantik. Selain fitur-fitur seperti pencarian konsep, ontologi dan rekomendasi kelas, serta pencarian berbasis kata kunci, sistem ini juga menjelaskan pendekatan berbasis popularitas untuk pernyataan peringkat suatu entitas yang terlibat di dalamnya. Selanjutnya, penulis juga menjelaskan penggunaan teknik MMR (Carbonell dan Jade 1998) untuk mengurutkan ulang pernyataan-pernyataan untuk memperhitungkan keberagaman. Dalam publikasi selanjutnya (Cheng 2011), peringkasan entitas memerlukan "pemeringkatan elemen data berdasarkan seberapa besar elemen tersebut membantu mengidentifikasi entitas yang mendasarinya". Pernyataan ini menjelaskan definisi paling umum dari ringkasan entitas: pemeringkatan dan pemilihan pernyataan yang mengidentifikasi atau mendefinisikan suatu entitas.

Di Singhal (2012), penulis memperkenalkan Grafik Pengetahuan Google. Selain disambiguasi entitas dan penelusuran eksplorasi, grafik pengetahuan juga memberikan

ringkasan entitas, yaitu *"dapatkan ringkasan terbaik"*. Meskipun tidak dijelaskan secara rinci, Google menunjukkan bahwa mereka menggunakan permintaan pencarian pengguna untuk ringkasannya. Untuk ringkasan grafik pengetahuan, Google menggunakan kumpulan data unik dari jutaan kueri harian untuk memberikan ringkasan yang ringkas. Namun kumpulan data tersebut tidak tersedia untuk semua penyedia konten.

Sebagai alternatif, Thalhammer dkk. (2012b) menyarankan penggunaan data latar belakang pola konsumsi item untuk mendapatkan ringkasan entitas film. Idenya berasal dari bidang sistem pemberi rekomendasi di mana lingkungan item dapat diperoleh melalui perilaku konsumsi bersama pengguna (yaitu melalui analisis matriks item pengguna). Upaya pertama untuk membakukan evaluasi ringkasan entitas dilakukan oleh Thalhammer et al. (2012a). Penulis menyarankan permainan dengan tujuan (GWAP) untuk menghasilkan kumpulan data referensi untuk ringkasan entitas. Dalam deskripsinya, game ini dirancang sebagai kuis tentang entitas film dari Freebase. Dalam evaluasinya, penulis membandingkan ringkasan yang dihasilkan oleh Singhal (2012) dan ringkasan Thalhammer et al. (2012b).

Abstraksi Data Berdasarkan Ontologi dan Pola Alur Kerja Komunikasi

Masalah komunikasi di web, dan juga di luarnya, bukanlah hal yang sepele, mengingat pesatnya peningkatan jumlah saluran (platform berbagi konten, media dan jaringan sosial, beragam perangkat) dan audiens yang ingin dijangkau. Untuk mengatasi masalah ini, solusi teknologi sedang dikembangkan seperti yang dikemukakan oleh Fensel et al. (2012) berdasarkan semantik. Pengelolaan data melalui teknik semantik tentunya dapat memudahkan abstraksi komunikasi dan juga meningkatkan otomatisasi serta mengurangi upaya secara keseluruhan.

Terinspirasi oleh karya Mika (2005), pola alur kerja eKomunikasi (misalnya pola respons permintaan yang khas untuk komunikasi online), yang dapat digunakan dan disesuaikan dengan kebutuhan jaringan sosial, dapat didefinisikan (Stavrakantonakis 2013a, b). Selain itu, terdapat minat terhadap interaksi jaringan sosial (Fuentes-Fernandez dkk. 2012). Para penulis karya terakhirnya menciptakan *"properti sosial"* sebagai jaringan konsep teori aktivitas dengan makna tertentu. Properti sosial dianggap sebagai pola yang mewakili pengetahuan yang didasarkan pada ilmu-ilmu sosial tentang motivasi, perilaku, organisasi, interaksi. Hasil arahan penelitian ini dikombinasikan dengan pola alur kerja umum yang dijelaskan dalam Van Der Aalst dkk. (2003) sangat relevan dengan perwujudan pola komunikasi. Perancangan pola juga terkait dengan kolaborasi antar berbagai agen seperti yang dijelaskan dalam Dorn dkk. (2012) dalam lingkup alur kerja sosial. Selain sifat sosial, karya yang dijelaskan dalam Rowe et al. (2011) memperkenalkan penggunaan ontologi dalam pemodelan aktivitas pengguna sehubungan dengan konten dan sentimen. Dalam konteks pendekatan ini, perilaku pemodelan memungkinkan seseorang untuk mengidentifikasi pola masalah komunikasi dan memahami dinamika diskusi untuk menemukan cara untuk terlibat secara lebih efisien dengan publik di jaringan sosial. Beberapa peneliti telah mengusulkan realisasi alur kerja sadar konteks (Wieland et al. 2007) dan proses kolaborasi sosial (Liptchinsky et al. 2012), yang terkait dengan gagasan pemodelan aktor dan artefak terkait untuk memungkinkan adaptasi dan personalisasi dalam infrastruktur pola komunikasi.

5.4 TREN DAN TUNTUTAN MASA DEPAN UNTUK ANALISIS DATA BESAR

Persyaratan Masa Depan untuk Analisis Big Data

Teknologi Big Data saat ini seperti Apache Hadoop telah berkembang pesat selama bertahun-tahun menjadi platform yang banyak digunakan di berbagai industri. Beberapa orang yang kami wawancarai telah mengidentifikasi kebutuhan masa depan yang harus diatasi oleh teknologi Big Data generasi berikutnya:

1. **Menangani pertumbuhan Internet (Baeza-Yates)**—seiring dengan semakin banyaknya pengguna yang online, teknologi Big Data perlu menangani volume data yang lebih besar.
2. **Memproses tipe data yang kompleks (Baeza-Yates)**—data seperti data grafik dan kemungkinan tipe struktur data lain yang lebih rumit perlu diproses dengan mudah oleh teknologi Big Data.
3. **Pemrosesan real-time (Baeza-Yates)**—pemrosesan Big Data pada awalnya dilakukan dalam kumpulan data historis. Dalam beberapa tahun terakhir, sistem pemrosesan aliran seperti Apache Storm telah tersedia dan memungkinkan kemampuan aplikasi baru. Teknologi ini tergolong baru dan perlu dikembangkan lebih lanjut.
4. **Pemrosesan data secara bersamaan (Baeza-Yates)**—kemampuan memproses data dalam jumlah besar secara bersamaan sangat berguna untuk menangani pengguna dalam jumlah besar pada saat yang bersamaan.
5. **Orkestrasi layanan yang dinamis dalam konteks multi-server dan cloud (Tori)**—sebagian besar platform saat ini tidak cocok untuk cloud dan menjaga konsistensi data antar penyimpanan data yang berbeda merupakan sebuah tantangan.
6. **Pengindeksan yang efisien (Mika)**—pengindeksan merupakan hal mendasar dalam pencarian data secara online dan oleh karena itu penting dalam mengelola kumpulan besar dokumen dan metadata terkait.

Kesederhanaan

Kesederhanaan teknologi Big Data mengacu pada betapa mudahnya pengembang memperoleh teknologi dan menggunakannya di lingkungan spesifik mereka. Kesederhanaan penting karena mengarah pada adopsi teknologi yang lebih tinggi (Baeza-Yates). Beberapa orang yang kami wawancarai telah mengidentifikasi peran penting kesederhanaan dalam teknologi Big Data saat ini dan masa depan.

Keberhasilan Hadoop dan MapReduce terutama disebabkan oleh kesederhanaannya (Tori). Tersedia platform data besar lainnya yang dianggap lebih kuat, namun memiliki komunitas pengguna yang lebih kecil karena penerapannya lebih sulit untuk dikelola. Demikian pula, teknologi data tertaut, misalnya, RDF SPARQL, dilaporkan terlalu rumit dan mengandung kurva pembelajaran yang terlalu curam (Gislason). Teknologi seperti ini tampaknya dirancang secara berlebihan dan terlalu rumit hanya cocok untuk digunakan oleh para spesialis.

Secara keseluruhan, terdapat beberapa teknologi yang sangat matang untuk analisis data besar, namun teknologi ini perlu diindustrialisasikan dan dapat diakses oleh semua orang

(Benjamins). Orang-orang di luar komunitas inti Big Data harus menyadari kemungkinan-kemungkinan Big Data, untuk mendapatkan dukungan yang lebih luas. Big Data bergerak melampaui industri Internet dan memasuki industri non-teknis lainnya. Platform Big Data yang mudah digunakan akan membantu adopsi teknologi Big Data oleh industri non-teknis.

Data

Unsur utama yang jelas dalam solusi Big Data adalah data itu sendiri. Orang yang kami wawancarai mengidentifikasi beberapa masalah yang perlu ditangani. Perusahaan besar seperti Google dan Facebook sedang mengerjakan data besar dan mereka akan memfokuskan energi mereka pada bidang tertentu dan bukan pada bidang lain. Keterlibatan UE dapat mendukung ekosistem Big Data yang mendorong beragam pemain kecil, menengah, dan besar, dimana peraturannya efektif dan datanya terbuka.

Dalam melakukan hal ini, penting untuk menyadari bahwa terdapat jauh lebih banyak data di luar sana daripada yang disadari kebanyakan orang dan data ini dapat membantu kita mngambil keputusan yang lebih baik untuk mengidentifikasi ancaman dan melihat peluang. Banyak data yang dibutuhkan sudah ada, namun tidak mudah untuk menemukan dan menggunakan data tersebut. Mengatasi masalah ini akan membantu dunia usaha, pengambil kebijakan, dan pengguna akhir dalam pengambilan keputusan. Dengan menyediakan lebih banyak data dunia yang dapat diakses oleh banyak orang akan memberikan dampak yang besar secara keseluruhan. Item ini akan memiliki dampak yang signifikan dalam situasi darurat seperti gempa bumi dan bencana alam lainnya.

Namun, sulit untuk menyediakan data di perusahaan dan organisasi sebelum adanya Internet. Di perusahaan Internet, sejak awal terdapat fokus pada penggunaan data yang dikumpulkan untuk tujuan analitik. Perusahaan pra-Internet menghadapi masalah privasi, hukum serta teknis, dan pembatasan proses dalam menggunakan kembali data. Hal ini berlaku bahkan untuk data yang sudah tersedia dalam bentuk digital, seperti catatan detail panggilan untuk perusahaan telepon. Proses seputar penyimpanan dan penggunaan data tersebut tidak pernah diatur dengan tujuan menggunakan data untuk analisis.

Inisiatif data terbuka dapat memainkan peran penting dalam membantu perusahaan dan organisasi mendapatkan hasil maksimal dari data. Setelah kumpulan data melewati validasi yang diperlukan sehubungan dengan privasi dan pembatasan lainnya, kumpulan data tersebut dapat digunakan kembali untuk berbagai tujuan oleh berbagai perusahaan dan organisasi dan dapat berfungsi sebagai platform untuk bisnis baru (Hendler). Oleh karena itu, penting untuk berinvestasi dalam proses dan undang-undang yang mendukung inisiatif data terbuka. Mencapai kebijakan yang dapat diterima tampaknya merupakan sebuah tantangan. Seperti yang dicatat oleh salah satu narasumber kami, terdapat ketegangan yang melekat antara data terbuka dan privasi keduanya mungkin tidak dapat diperoleh (Tori). Namun kumpulan data tertutup juga harus ditangani.

Banyak informasi berharga, seperti data ponsel, yang saat ini ditutup dan dimiliki oleh industri telekomunikasi. UE harus mencari cara agar data tersebut tersedia bagi komunitas Big Data, sambil mempertimbangkan biaya yang terkait untuk membuat data tersebut terbuka. Selain itu, bagaimana industri telekomunikasi dapat mengambil manfaat dari membuka data

dengan tetap mempertimbangkan masalah privasi (Das). Web juga dapat berfungsi sebagai sumber data penting. Perusahaan seperti Data Publica mengandalkan snapshot web yang berukuran 60–70 terabyte, untuk mendukung layanan online. Tersedia versi snapshot web yang tersedia secara gratis, namun versi yang lebih terkini lebih disukai. Ini tidak harus gratis, tapi murah. Pemain web besar seperti Google dan Facebook memiliki akses ke data terkait pencarian dan jejaring sosial yang memiliki manfaat sosial yang penting. Misalnya, proses sosial yang dinamis seperti penyebaran penyakit atau tingkat lapangan kerja sering kali paling akurat dilacak melalui penelusuran Google. UE mungkin ingin memprioritaskan hal-hal serupa di Eropa yang serupa dengan cara Tiongkok mengkloning Google dan Twitter.

Ketika kumpulan data terbuka menjadi lebih umum, menemukan kumpulan data yang diperlukan menjadi semakin sulit. Sebuah prediksi memperkirakan bahwa pada tahun 2015 akan ada lebih dari 10 juta kumpulan data yang tersedia di web (Hendler). Pelajaran berharga dapat diambil dari bagaimana penemuan dokumen berkembang di web. Awalnya ada registri semua web dapat dicantumkan dalam satu halaman web; kemudian pengguna dan organisasi mempunyai daftarnya sendiri; lalu daftar daftar. Belakangan, Google mendominasi dengan menyediakan metrik tentang bagaimana dokumen tertaut ke dokumen lain. Jika dianalogikan dengan area data, saat ini berada di era registri. Dibutuhkan crawler untuk menemukan kumpulan data besar, metadata kumpulan data yang baik pada konten, tautan antar kumpulan data terkait, dan mekanisme pemeringkatan kumpulan data yang relevan (analog dengan peringkat halaman). Mekanisme penemuan yang hanya dapat bekerja dengan data berkualitas baik akan mendorong pemilik data untuk mempublikasikan datanya dengan cara yang lebih baik, serupa dengan cara optimasi mesin pencari (SEO) mendorong kualitas web saat ini (Tennison).

Bahasa

Sebagian besar teknologi Big Data berasal dari Amerika Serikat dan oleh karena itu dibuat dengan mempertimbangkan bahasa Inggris. Mayoritas perusahaan Internet melayani khalayak internasional dan banyak dari layanan mereka akhirnya diterjemahkan ke dalam bahasa lain. Sebagian besar layanan pada awalnya diluncurkan dalam bahasa Inggris dan hanya diterjemahkan setelah layanan tersebut mendapatkan popularitas. Selain itu, pengoptimalan teknologi terkait bahasa tertentu (misalnya pengoptimalan mesin pencari) mungkin berfungsi dengan baik untuk bahasa Inggris, namun tidak untuk bahasa lain. Bagaimanapun, bahasa harus diperhitungkan sejak awal, terutama di Eropa, dan harus memainkan peran penting dalam menciptakan arsitektur Big Data (Halevy).

Paradigma yang Muncul untuk Analisis Big Data

Komunitas

Munculnya Internet memungkinkan kita menjangkau khalayak dalam jumlah besar dengan cepat dan menumbuhkan komunitas seputar topik yang diminati. Big Data mulai memainkan peran yang semakin penting dalam perkembangan tersebut. Orang-orang yang kami wawancarai telah menyebutkan paradigma yang muncul ini dalam beberapa kesempatan.

1. **Bangkitnya Jurnalis Data**: Yang mampu menulis artikel menarik berdasarkan data yang diunggah publik ke infrastruktur seperti Google Fusion Tables. Jurnalis The Guardian, Simon Rogers, memenangkan penghargaan Jurnalis Internet Inggris Terbaik atas karyanya berdasarkan platform ini. Salah satu ciri penggunaan jurnalistik adalah bahwa blog data biasanya memiliki dampak diseminasi yang tinggi (Halevy).
2. **Keterlibatan Masyarakat Dalam Isu-Isu Politik Lokal**: Dua bulan setelah pembantaian sekolah di Connecticut warga setempat mulai melihat data terkait permohonan izin kepemilikan senjata di dua lokasi dan memaparkannya pada peta. Hal ini memicu diskusi besar-besaran mengenai isu-isu terkait.
3. **Keterlibatan Melalui Pengumpulan Dan Analisis Data Komunitas**: Perusahaan COSM (sebelumnya Pachube) telah mendorong sejumlah upaya yang dipimpin oleh komunitas. Gagasan utama di balik hal ini adalah bahwa cara pengumpulan data memperkenalkan pandangan tertentu tentang bagaimana data dapat diinterpretasikan dan digunakan. Melibatkan masyarakat mempunyai berbagai manfaat: jumlah titik pengumpulan data dapat ditingkatkan secara signifikan; masyarakat sering kali membuat alat yang disesuaikan dengan situasi tertentu dan untuk menangani masalah apa pun dalam pengumpulan data; dan keterlibatan warga meningkat secara signifikan. Salah satu contohnya adalah perusahaan tersebut melakukan pemantauan radiasi secara real-time di Jepang menyusul masalah reaktor di Fukushima. Kini terdapat ratusan feed terkait radiasi dari Jepang di Pachube, memantau kondisi secara real-time dan mendukung lebih dari setengah lusin aplikasi luar biasa berharga yang dibangun oleh orang-orang di seluruh dunia. Ini menggabungkan data "resmi", data "tidak resmi", dan juga pengukuran penghitung Geiger jaringan real-time yang disumbangkan oleh warga yang peduli (Haque).
4. **Keterlibatan Masyarakat Untuk Mendidik Dan Meningkatkan Keterlibatan Ilmiah**: Masyarakat bisa sangat berguna dalam mengumpulkan data. Partisipasi dalam proyek-proyek tersebut memungkinkan masyarakat untuk memperoleh pemahaman yang lebih baik tentang kegiatan ilmiah tertentu dan oleh karena itu membantu mendidik masyarakat tentang topik-topik ini. Peningkatan pemahaman tersebut akan lebih menstimulasi pengembangan dan apresiasi terhadap teknologi yang akan datang dan oleh karena itu menghasilkan siklus penguatan diri yang positif (Thompson).
5. **Crowdsourcing Untuk Meningkatkan Keakuratan Data**: Melalui crowdsourcing, keakuratan data yang dirilis Pemerintah Inggris mengenai lokasi halte bus meningkat secara dramatis (Hendler).

Upaya ini berperan baik dalam bagian persyaratan data di masa depan. Pendekatan berbasis komunitas dalam pembuatan kumpulan data akan mendorong kualitas data dan menghasilkan lebih banyak kumpulan data yang tersedia untuk umum.

Dampak Akademik

Ketersediaan kumpulan data yang besar akan berdampak pada akademisi (Tori) karena dua alasan. Pertama, kumpulan data publik dapat digunakan oleh peneliti dari berbagai disiplin ilmu seperti ilmu sosial dan ekonomi untuk mendukung kegiatan penelitian mereka.

Kedua, platform untuk berbagi kumpulan data akademis akan mendorong penggunaan kembali dan meningkatkan kualitas kumpulan data yang dipelajari. Berbagi kumpulan data juga memungkinkan orang lain menambahkan anotasi tambahan ke data, yang biasanya merupakan tugas yang mahal. Selain melihat teknologi Big Data mempengaruhi disiplin ilmu lainnya, disiplin ilmu lain juga dibawa ke dalam ilmu komputer. Perusahaan Internet besar seperti Yahoo mempekerjakan ilmuwan sosial, termasuk psikolog dan ekonom, untuk meningkatkan efektivitas alat analisis (Mika). Secara lebih umum, seiring dengan berlanjutnya analisis data di berbagai domain, kebutuhan akan pakar domain semakin meningkat.

5.5 STUDI KASUS SEKTOR UNTUK ANALISIS BIG DATA

Bagian ini menjelaskan beberapa studi kasus Big Data yang menguraikan pemangku kepentingan yang terlibat, jika memungkinkan, dan hubungan antara teknologi dan konteks sektor secara keseluruhan. Secara khusus, hal ini mencakup sektor-sektor berikut: sektor publik, sektor kesehatan, sektor ritel, logistik, dan terakhir sektor keuangan. Dalam banyak kasus, deskripsi tersebut didukung oleh wawancara yang telah dilakukan, dan menambah bukti lebih lanjut mengenai potensi besar dari Big Data.

Sektor Publik

Kota pintar menghasilkan data dari sensor, media sosial, laporan seluler warga, dan data kota seperti data pajak. Teknologi Big Data digunakan untuk memproses kumpulan data besar yang dihasilkan kota agar berdampak pada masyarakat dan bisnis (Baeza-Yates). Bagian ini membahas bagaimana teknologi Big Data memanfaatkan data kota pintar untuk menyediakan aplikasi dalam lalu lintas dan tanggap darurat.

Lalu Lintas

Sensor kota pintar yang dapat digunakan untuk aplikasi di lalu lintas antara lain deteksi loop induksi, kamera lalu lintas, dan kamera pengenalan plat nomor (LPR). Loop induksi dapat digunakan untuk menghitung volume lalu lintas pada suatu titik tertentu. Kamera lalu lintas dapat dikombinasikan dengan solusi analitik video untuk mengekstrak statistik secara otomatis seperti jumlah mobil yang lewat dan kecepatan rata-rata lalu lintas. Pengenalan plat nomor adalah teknologi berbasis kamera yang dapat melacak plat nomor di seluruh kota menggunakan beberapa kamera. Semua bentuk penginderaan ini membantu dalam memperkirakan statistik lalu lintas, meskipun tingkat keakuratan dan keandalannya berbeda-beda.

Penerapan teknologi tersebut di tingkat kota akan menghasilkan kumpulan data besar yang dapat digunakan untuk operasional sehari-hari, serta aplikasi seperti deteksi anomali dan dukungan dalam perencanaan operasional. Dalam hal analisis data besar, aplikasi yang paling menarik adalah deteksi anomali. Sistem dapat belajar dari data historis apa yang dianggap sebagai perilaku lalu lintas normal pada waktu tertentu dan hari dalam seminggu dan mendeteksi penyimpangan dari norma untuk menginformasikan operator di pusat komando dan kendali tentang kemungkinan insiden yang memerlukan perhatian (Thajchayapong dan Barria 2010). Pendekatan seperti itu menjadi lebih efektif ketika menggabungkan data dari beberapa lokasi menggunakan penggabungan data untuk mendapatkan perkiraan statistik

lalu lintas yang lebih akurat sehingga memungkinkan pendeteksian skenario yang lebih kompleks.

Tanggap Darurat

Kota-kota yang dilengkapi dengan sensor dapat memperoleh manfaat selama keadaan darurat dengan memperoleh informasi yang dapat ditindaklanjuti dan dapat membantu dalam pengambilan keputusan. Yang menarik adalah kemungkinan untuk menggunakan analisis media sosial selama tanggap darurat. Jaringan media sosial menyediakan aliran informasi yang konstan yang dapat digunakan sebagai jaringan penginderaan global berbiaya rendah untuk mengumpulkan informasi hampir real-time tentang keadaan darurat. Meskipun orang-orang memposting banyak informasi yang tidak berkaitan di jaringan media sosial, informasi apa pun mengenai keadaan darurat bisa sangat berharga bagi tim tanggap darurat. Data yang akurat dapat membantu memperoleh gambaran kesadaran situasional yang benar mengenai keadaan darurat, sehingga memungkinkan respons yang lebih efisien dan lebih cepat sehingga dapat mengurangi korban jiwa dan kerusakan secara keseluruhan (Van Kasteren dkk. 2014). Analisis media sosial digunakan untuk memproses postingan media sosial dalam jumlah besar, seperti tweet, untuk mengidentifikasi kelompok postingan yang berpusat pada topik yang sama (konten yang tumpang tindih), area yang sama (untuk postingan yang berisi tag GPS), dan pada waktu yang hampir bersamaan. Cluster postingan merupakan hasil dari tingginya aktivitas jejaring sosial di suatu daerah. Ini bisa menjadi indikasi suatu landmark (misalnya menara Eiffel), peristiwa yang direncanakan (misalnya pertandingan olahraga), atau peristiwa yang tidak direncanakan (misalnya kecelakaan). Situs-situs terkenal memiliki volume tweet yang tinggi sepanjang tahun dan oleh karena itu dapat dengan mudah disaring. Untuk kejadian lainnya, pengklasifikasi pembelajaran mesin digunakan untuk secara otomatis mengenali cluster mana yang diminati oleh operator tanggap darurat (Walther dan Kaisser 2013).

Penggunaan data media sosial untuk tujuan yang awalnya tidak dimaksudkan hanyalah satu contoh dampak signifikan yang dapat terjadi ketika data yang tepat disajikan kepada orang yang tepat pada waktu yang tepat. Beberapa orang yang kami wawancarai menjelaskan bahwa terdapat jauh lebih banyak data di luar sana daripada yang disadari kebanyakan orang dan data ini dapat membantu kami mengambil keputusan yang lebih baik dalam mengidentifikasi ancaman dan melihat peluang. Banyak data yang dibutuhkan sudah ada, namun tidak selalu mudah untuk menemukan dan menggunakan data ini (Gislason) (Halevy).

Kesehatan

Bagian sebelumnya membahas tentang data yang digunakan kembali dalam aplikasi yang sangat berbeda dari aplikasi asli yang menghasilkan data tersebut. Kasus serupa juga terjadi di sektor kesehatan. Misalnya, proses sosial yang dinamis seperti penyebaran penyakit dapat dilacak secara akurat melalui penelusuran Google (Bancilhon) dan catatan detail panggilan telepon dari Telefonica telah digunakan untuk mengukur dampak peringatan epidemi terhadap mobilitas manusia (Frias-Martinez et al. 2012).

Analisis data besar dapat digunakan untuk memecahkan masalah signifikan secara global. Oleh karena itu, UE disarankan untuk menghasilkan solusi yang menyelesaikan

permasalahan global daripada hanya berfokus pada permasalahan yang berdampak pada UE (Thompson). Contohnya adalah pembangunan sumur air bersih di Afrika. Keputusan mengenai lokasi sumur didasarkan pada spreadsheet yang mungkin berisi data yang belum diperbarui selama 2 tahun. Mengingat sumur baru dapat berhenti berfungsi setelah 6 bulan, hal ini menyebabkan kesulitan yang tidak perlu dan lebih banyak lagi (Halevy). Teknologi mungkin menawarkan solusi, baik dengan mengizinkan laporan masyarakat atau dengan menyimpulkan penggunaan sumur dari sumber data lain.

Dampaknya terhadap layanan kesehatan lokal diperkirakan akan sangat besar. Berbagai proyek teknologi ditujukan untuk mewujudkan layanan kesehatan di rumah, di mana setidaknya masyarakat dapat mencatat pengukuran terkait kesehatan di rumah mereka sendiri. Ketika digabungkan dengan proyek seperti solusi rumah pintar, dimungkinkan untuk membuat kumpulan data kaya yang terdiri dari data kesehatan dan semua jenis data perilaku yang dapat sangat membantu dalam menegakkan diagnosis, serta mendapatkan pemahaman yang lebih baik tentang penyakit. permulaan dan perkembangan.

Namun, terdapat kekhawatiran privasi yang sangat kuat di sektor layanan kesehatan yang kemungkinan besar akan menghambat banyak perkembangan ini sampai masalah tersebut terselesaikan. Profesor Marco Viceconti dari Universitas Sheffield menguraikan dalam wawancaranya bagaimana perkembangan terkini seperti k-anonymity dapat membantu melindungi privasi. Suatu dataset memiliki perlindungan k-anonymity jika informasi setiap individu dalam dataset tidak dapat dibedakan dari setidaknya k—1 individu yang informasinya juga muncul dalam dataset (Sweeney 2002). Profesor Viceconti membayangkan sistem masa depan yang secara otomatis dapat melindungi privasi dengan berfungsi sebagai membran antara pasien dan lembaga yang menggunakan data, di mana data dapat mengalir dua arah dan semua kebijakan privasi yang diperlukan serta proses anonimisasi dijalankan secara otomatis di antara keduanya. Sistem seperti ini akan menguntungkan pasien, dengan memberikan diagnosis yang lebih akurat, dan juga lembaga, dengan memungkinkan penelitian menggunakan data dunia nyata.

Ritel

O2 UK bersama dengan Telef´onica Digital baru-baru ini meluncurkan layanan bernama Telef´onica Dynamic Insights. Layanan ini mengambil semua data seluler Inggris, termasuk lokasi, waktu panggilan dan SMS, dan juga saat pelanggan berpindah dari satu tiang ke tiang lainnya. Data ini dipetakan dan digunakan kembali untuk industri ritel. Data pertama-tama dianonimkan, dikumpulkan, dan ditempatkan di cloud. Kemudian analitik dijalankan untuk menghitung di mana orang tinggal, di mana mereka bekerja, dan di mana mereka transit. Jika data ini kemudian digabungkan dengan data manajemen hubungan pelanggan (CRM) yang dianonimkan, data ini dapat menentukan jenis orang yang melewati toko tertentu pada titik waktu tertentu. Hal ini juga dapat menghitung jenis orang yang mengunjungi toko, di mana mereka tinggal, dan di mana lagi mereka berbelanja (disebut daerah tangkapan air).

Layanan ini mendukung pengelolaan real estat untuk pengecer dan sangat kontras dengan praktik saat ini. Apa yang dilakukan pengecer saat ini adalah mereka mempekerjakan siswa dengan clicker hanya untuk menghitung jumlah orang yang berjalan melewati toko,

sehingga menghasilkan data yang kurang detail. Dengan demikian, layanan ini memecahkan masalah yang ada dengan cara baru. Layanan ini dapat dijalankan secara mingguan atau harian dan memberikan peluang bisnis yang benar-benar baru. Selain ritel, layanan ini dapat dijalankan di sektor lain, misalnya, di sektor publik, layanan ini dapat menganalisis siapa yang berjalan melewati stasiun bawah tanah. Menggabungkan data seluler dengan data preferensi dapat membuka proposisi baru bagi industri yang sudah ada dan industri baru. Contoh ini adalah gambaran dari apa yang akan terjadi, yang secara total akan meningkatkan daya saing industri Eropa (Benjamins).

Logistik

Di Amerika Serikat, 45% buah-buahan dan sayur-sayuran sampai ke piring konsumen dan di Eropa 55% sampai ke piring konsumen. Hampir setengah dari apa yang diproduksi hilang. Ini adalah masalah data besar: mengumpulkan data pada keseluruhan Jaringan pasokan, menganalisis sistem yang berkaitan dengan makanan yang didistribusikan, dan mengidentifikasi kebocoran dan kemacetan dalam proses tersebut akan mempunyai dampak yang sangat besar. Jika diterapkan, maka akan ada penanganan harga yang lebih baik dan distribusi kekayaan yang lebih adil di antara semua pelaku dalam Jaringan pasokan pangan. Teknologi Big Data itu penting, begitu pula akses terhadap data dan sumber data yang tepat (Bancilhon).

Keuangan

Bank Dunia adalah organisasi yang bertujuan untuk mengakhiri kemiskinan ekstrem dan mendorong kesejahteraan bersama. Operasi mereka sangat bergantung pada informasi yang akurat dan mereka menggunakan analisis data besar untuk mendukung aktivitas mereka. Mereka berencana untuk menyelenggarakan kompetisi untuk mendorong kemampuan analitik guna mendapatkan ukuran alternatif mengenai kemiskinan dan untuk mendeteksi korupsi dan penipuan keuangan pada tahap awal.

Dalam kaitannya dengan kemiskinan, faktor pendorong yang penting adalah diperolehnya perkiraan kemiskinan yang lebih real-time, sehingga memungkinkan pengambilan keputusan jangka pendek yang lebih baik. Tiga contoh sumber informasi yang saat ini sedang dijajaki untuk memperoleh informasi yang dibutuhkan adalah: (1) Data Twitter dapat digunakan untuk mencari indikator kesejahteraan sosial dan ekonomi; (2) peta kemiskinan dapat digabungkan dengan sumber data alternatif seperti citra satelit untuk mengidentifikasi jalan beraspal dan mendukung keputusan dalam pembiayaan mikro; dan (3) data web dapat diambil untuk mendapatkan data harga dari supermarket yang membantu dalam estimasi kemiskinan.

Korupsi saat ini ditangani secara reaktif, artinya tindakan hanya diambil setelah korupsi dilaporkan ke Bank Dunia. Rata-rata hanya 30 % uang yang bisa diambil dalam kasus korupsi bila ditangani secara reaktif. Analisis Big Data akan memungkinkan pendekatan yang lebih proaktif, sehingga menghasilkan keuntungan yang lebih tinggi. Hal ini memerlukan penciptaan profil perusahaan dan mitra kerja yang lebih kaya. Pengumpulan data dari data profil mendalam ini bersama dengan sumber data lainnya akan memungkinkan untuk mengidentifikasi pola terkait risiko.

Secara keseluruhan, penting bagi Bank Dunia untuk dapat mengambil keputusan, memindahkan sumber daya, dan menyediakan pilihan investasi secepat mungkin melalui orang yang tepat pada waktu yang tepat. Melakukan hal ini berdasarkan kumpulan data lama yang terbatas tidak akan berkelanjutan dalam jangka menengah dan panjang. Informasi yang akurat dan real-time sangat penting dalam proses pengambilan keputusan. Misalnya, jika ada resesi yang akan terjadi, kita perlu mengambil tindakan sebelum hal itu terjadi. Jika terjadi bencana alam, pengambilan keputusan berdasarkan data yang tersedia langsung dari lapangan, dibandingkan berdasarkan data yang berumur 3 tahun, sangat diinginkan (Das).

Kesimpulan

Analisis Big Data adalah bagian mendasar dari Jaringan nilai Big Data. Kita dapat membuat karikatur proses ini menggunakan pepatah Inggris kuno yang mengatakan bahwa komponen ini menghasilkan "mengubah timah menjadi emas". Data dalam jumlah besar yang mungkin heterogen sehubungan dengan mekanisme pengkodean, format, struktur, semantik yang mendasarinya, asal, keandalan, dan kualitas diubah menjadi data yang dapat digunakan.

Oleh karena itu, analisis Big Data terdiri dari kumpulan teknik dan alat yang beberapa di antaranya merupakan mekanisme lama yang disusun kembali untuk menghadapi tantangan yang ditimbulkan oleh tiga V (misalnya penalaran skala besar) dan beberapa di antaranya baru (misalnya penalaran aliran).

Wawasan yang dikumpulkan mengenai analisis Big Data yang disajikan di sini didasarkan pada 19 wawancara dengan para pemain terkemuka di industri besar dan kecil serta para visioner dari Eropa dan Amerika Serikat. Pilihan yang diambil adalah dengan mewawancarai anggota staf senior yang memiliki peran kepemimpinan di perusahaan multinasional besar, ahli teknologi yang bekerja di bidang batubara yang menangani Big Data, pendiri dan CEO dari generasi baru UKM yang telah menghasilkan nilai dari Big Data, dan para pemimpin akademis di bidang Big Data. lapangan.

Dari analisis kami jelas bahwa memberikan analisis data yang sangat terukur dan mekanisme penalaran yang terkait dengan ekosistem alat yang dapat diakses dan digunakan akan menghasilkan manfaat yang signifikan bagi Eropa. Dampaknya akan bersifat ekonomi dan sosial. Model dan proses bisnis saat ini akan diubah secara radikal demi keuntungan ekonomi dan sosial. Studi kasus mengenai pengurangan jumlah makanan yang terbuang dalam siklus hidup produksi pangan global adalah contoh utama dari potensi Big Data.

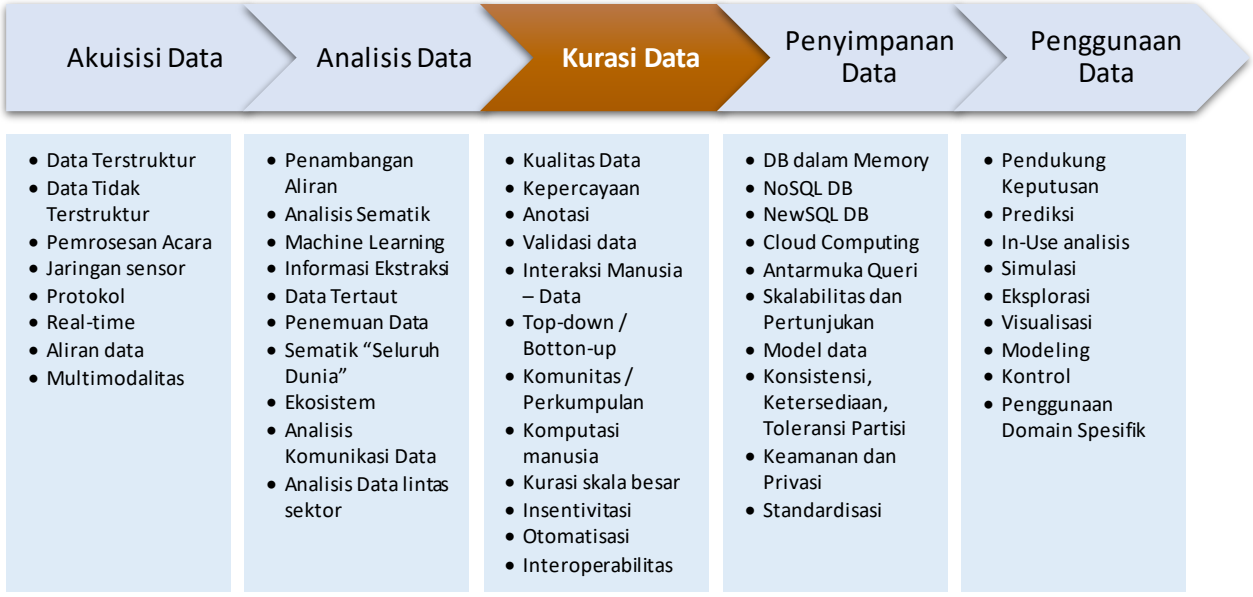
Ringkasnya, analisis Big Data merupakan bagian penting dari keseluruhan Jaringan nilai Big Data yang menjanjikan dampak ekonomi dan sosial yang signifikan di Uni Eropa dalam jangka pendek hingga menengah. Tanpa analisis Big Data, seluruh Jaringan tidak akan berfungsi. Seperti yang dinyatakan oleh salah satu orang yang kami wawancarai dalam diskusi baru-baru ini tentang hubungan antara analisis data dan analisis data:

BAB 6
KURASI BIG DATA

6.1 PENDAHULUAN

Salah satu prinsip utama analisis data adalah kualitas analisis bergantung pada kualitas informasi yang dianalisis. Gartner memperkirakan bahwa lebih dari 25% data penting di perusahaan-perusahaan terkemuka dunia memiliki kelemahan. Permasalahan kualitas data dapat berdampak signifikan terhadap operasional bisnis, terutama dalam proses pengambilan keputusan dalam organisasi.

Munculnya platform baru untuk pembuatan data yang terdesentralisasi seperti platform sensor dan seluler, meningkatnya ketersediaan data terbuka di web (Howe et al. 2008), menambah peningkatan jumlah sumber data di dalam organisasi (Brodie dan Liu 2010), menghadirkan volume data yang belum pernah ada sebelumnya untuk dikelola. Selain volume data, konsumen data di era Big Data juga perlu mengatasi keragaman data, sebagai konsekuensi dari generasi data yang terdesentralisasi, dimana data dibuat berdasarkan konteks dan kebutuhan yang berbeda. Mengonsumsi data pihak ketiga menimbulkan biaya intrinsik untuk menggunakan kembali, mengadaptasi, dan memastikan kualitas data untuk konteks barunya.



Gambar 6.1 Kurasi data dalam Jaringan nilai Big Data

Kurasi data memberikan dukungan metodologis dan teknologi manajemen data untuk mengatasi masalah kualitas data sehingga memaksimalkan kegunaan data. Menurut Cragin dkk. (2007), "Kurasi data adalah pengelolaan data secara aktif dan berkelanjutan melalui siklus hidup yang menarik dan berguna aktivitas kurasi memungkinkan penemuan dan pengambilan data, menjaga kualitas, menambah nilai, dan menyediakan penggunaan kembali seiring

waktu". Kurasi data muncul sebagai proses pengelolaan data utama di mana terdapat peningkatan jumlah sumber data dan platform untuk menghasilkan data.

Posisi kurasi Big Data dalam keseluruhan Jaringan nilai Big Data dapat dilihat pada Gambar 6.1. Proses kurasi data dapat dikategorikan ke dalam aktivitas berbeda seperti pembuatan konten, seleksi, klasifikasi, transformasi, validasi, dan pelestarian. Pemilihan dan penerapan proses kurasi data merupakan masalah multidimensi, bergantung pada interaksi antara dimensi insentif, ekonomi, standar, dan teknologi. Bab ini menganalisis dinamika data di mana kurasi data dimasukkan, menyelidiki kebutuhan masa depan dan tren yang muncul untuk kurasi data, dan secara singkat menjelaskan contoh studi kasus.

6.2 WAWASAN UNTUK KURASI BIG DATA

eScience dan eGovernment adalah inovatornya, sementara perusahaan biomedis dan media adalah pengguna awal. Tuntutan akan interoperabilitas data dan penggunaan kembali eScience serta tuntutan akan transparansi yang efektif melalui data terbuka dalam konteks eGovernment mendorong praktik dan teknologi kurasi data. Sektor-sektor ini berperan sebagai visioner dan inovator dalam siklus adopsi teknologi kurasi data. Dari perspektif industri, organisasi di bidang biomedis, seperti perusahaan farmasi, berperan sebagai pengguna awal, didorong oleh kebutuhan, untuk mengurangi waktu pemasaran dan menurunkan biaya jalur penemuan obat. Perusahaan media juga merupakan pengguna awal, didorong oleh kebutuhan untuk mengatur pengumpulan data besar yang tidak terstruktur, untuk mengurangi waktu dalam menciptakan produk baru, menggunakan kembali data yang sudah ada, dan untuk meningkatkan aksesibilitas dan visibilitas artefak informasi.

Dampak inti dari kurasi data adalah memungkinkan model berbasis data yang lebih lengkap dan berkualitas tinggi untuk organisasi pengetahuan. Model yang lebih lengkap mendukung jumlah jawaban yang lebih besar melalui analisis data. Praktik dan teknologi kurasi data akan semakin hadir dalam lingkungan manajemen data kontemporer, memfasilitasi organisasi dan individu untuk menggunakan kembali data pihak ketiga dalam konteks yang berbeda, sehingga mengurangi hambatan dalam menghasilkan konten dengan kualitas data yang tinggi. Kemampuan untuk mengatasi masalah kualitas data dan heterogenitas dalam skala besar secara efisien akan mendukung konsumen data dalam pembuatan model yang lebih canggih, yang sangat berdampak pada produktivitas organisasi berbasis pengetahuan.

Kurasi data bergantung pada pembuatan struktur insentif. Sebagai aktivitas yang baru muncul, masih terdapat ketidakjelasan dan pemahaman yang buruk mengenai peran kurasi data dalam siklus hidup Big Data. Di banyak proyek, biaya kurasi data tidak diperkirakan atau diremehkan. Individuasi dan pengakuan atas peran kurator data dan aktivitas kurasi data bergantung pada perkiraan realistis mengenai biaya yang terkait dengan produksi data berkualitas tinggi. Dewan pendanaan dapat mendukung proses ini dengan mewajibkan perkiraan eksplisit atas sumber daya kurasi data pada proyek-proyek yang didanai publik dengan hasil data dan dengan mewajibkan publikasi data berkualitas tinggi. Selain itu, peningkatan pelacakan dan pengenalan data dan infrastruktur sebagai kontribusi ilmiah kelas

satu juga merupakan pendorong mendasar bagi inovasi metodologi dan teknologi untuk kurasi data dan untuk memaksimalkan pengembalian investasi dan penggunaan kembali hasil-hasil ilmiah. Pengakuan serupa diperlukan dalam konteks perusahaan.

Model ekonomi yang berkembang dapat mendukung penciptaan infrastruktur kurasi data. Kemitraan pra-kompetitif dan kemitraan pemerintah-swasta merupakan model ekonomi baru yang dapat mendukung penciptaan infrastruktur kurasi data dan pembuatan data berkualitas tinggi. Selain itu, pembenaran atas investasi pada infrastruktur kurasi data dapat didukung oleh kuantifikasi yang lebih baik mengenai dampak ekonomi dari data berkualitas tinggi.

Kurasi dalam skala besar bergantung pada interaksi antara platform kurasi otomatis dan pendekatan kolaboratif yang memanfaatkan sejumlah besar kurator data. Meningkatkan skala kurasi data bergantung pada pengurangan biaya per tugas kurasi data dan meningkatkan kumpulan kurator data. Pendekatan kurasi data manusia-algoritmik hibrid dan kemampuan untuk menghitung ketidakpastian hasil pendekatan algoritmik merupakan hal mendasar untuk meningkatkan otomatisasi tugas kurasi yang kompleks. Pendekatan untuk mengotomatisasi tugas-tugas kurasi data seperti kurasi dengan demonstrasi dapat memberikan peningkatan yang signifikan dalam skala otomatisasi. Crowdsourcing juga memainkan peran penting dalam meningkatkan kurasi data, memungkinkan akses ke sejumlah besar kurator data potensial. Peningkatan platform crowdsourcing menuju platform yang lebih terspesialisasi, otomatis, andal, dan canggih serta peningkatan integrasi antara sistem organisasi dan platform crowdsourcing merupakan peluang yang dapat dieksploitasi di bidang ini.

Peningkatan interaksi manusia-data merupakan hal mendasar untuk kurasi data. Meningkatkan pendekatan di mana kurator dapat berinteraksi dengan data akan berdampak pada efisiensi kurasi dan mengurangi hambatan bagi pakar domain dan pengguna biasa dalam melakukan kurasi data. Contoh fungsi utama dalam interaksi manusia-data mencakup antarmuka bahasa alami, pencarian semantik, peringkasan dan visualisasi data, dan antarmuka transformasi data yang intuitif.

Mekanisme manajemen kepercayaan dan izin tingkat data sangat penting untuk mendukung infrastruktur manajemen data untuk kurasi data. Pengelolaan asal adalah kunci yang memungkinkan terjadinya kepercayaan dalam kurasi data, memberikan kurator konteks untuk memilih data yang mereka anggap dapat dipercaya dan memungkinkan mereka mengambil keputusan kurasi data. Kurasi data juga bergantung pada mekanisme untuk menetapkan izin dan hak digital di tingkat data. Standar data dan model konseptual sangat mengurangi upaya kurasi data. Representasi data berbasis standar mengurangi heterogenitas sintaksis dan semantik, sehingga meningkatkan interoperabilitas. Model data dan standar model konseptual (misalnya kosakata dan ontologi) tersedia di domain berbeda. Namun, adopsi mereka masih terus meningkat.

Terdapat kebutuhan untuk meningkatkan model teoritis dan metodologi untuk kegiatan kurasi data. Model dan metodologi teoretis untuk kurasi data harus berkonsentrasi pada mendukung kemudahan pengangkutan data yang dihasilkan dalam konteks yang

berbeda, memfasilitasi deteksi masalah kualitas data, dan meningkatkan otomatisasi alur kerja kurasi data. Diperlukan integrasi yang lebih baik antara pendekatan komputasi algoritmik dan manusia. Semakin matangnya teknik statistik berbasis data di bidang seperti *Natural Language Processing* (NLP) dan *Machine Learning* (ML) menggeser penggunaannya dari lingkungan akademis ke lingkungan industri. Banyak alat NLP dan ML memiliki tingkat ketidakpastian yang terkait dengan hasilnya dan bergantung pada pelatihan pada kumpulan data yang besar. Integrasi yang lebih baik antara pendekatan statistik dan platform komputasi manusia sangat penting untuk memungkinkan evolusi model statistik yang berkelanjutan melalui penyediaan data pelatihan tambahan dan juga untuk meminimalkan dampak kesalahan pada hasil.

6.3 SYARAT KURASI BIG DATA

Banyak skenario Big Data yang dikaitkan dengan penggunaan kembali dan pengintegrasian data dari sejumlah sumber data berbeda. Persepsi ini sering terjadi pada para ahli dan praktisi kurasi data dan tercermin dalam pernyataan seperti: *"banyak data besar adalah kumpulan data kecil", "sebagian besar data besar bukanlah suatu blok besar yang seragam", "masing-masing bagian datanya sangat kecil dan sangat berantakan, dan banyak hal yang kami lakukan di sana berkaitan dengan keragaman tersebut"* (Wawancara Kurasi Data: Paul Groth 2014).

Menggunakan kembali data yang dihasilkan berdasarkan kebutuhan yang berbeda memiliki konsekuensi tersendiri dalam mengatasi masalah kualitas data dan heterogenitas data. Data mungkin tidak lengkap atau mungkin perlu diubah agar dapat berguna. Kevin Ashley, direktur Pusat Kurasi Digital, merangkum pola pikir di balik penggunaan kembali data: *"ketika Anda hanya menggunakan apa yang ada, yang mungkin bukan apa yang Anda kumpulkan di dunia yang ideal, tetapi Anda mungkin dapat memperoleh pengetahuan yang berguna darinya*" (Kevin Ashley 2014). Dalam konteks ini, data bergeser dari sumber daya yang sejak awal disesuaikan untuk tujuan tertentu, menjadi bahan mentah yang perlu digunakan kembali dalam konteks berbeda untuk memenuhi kebutuhan tertentu.

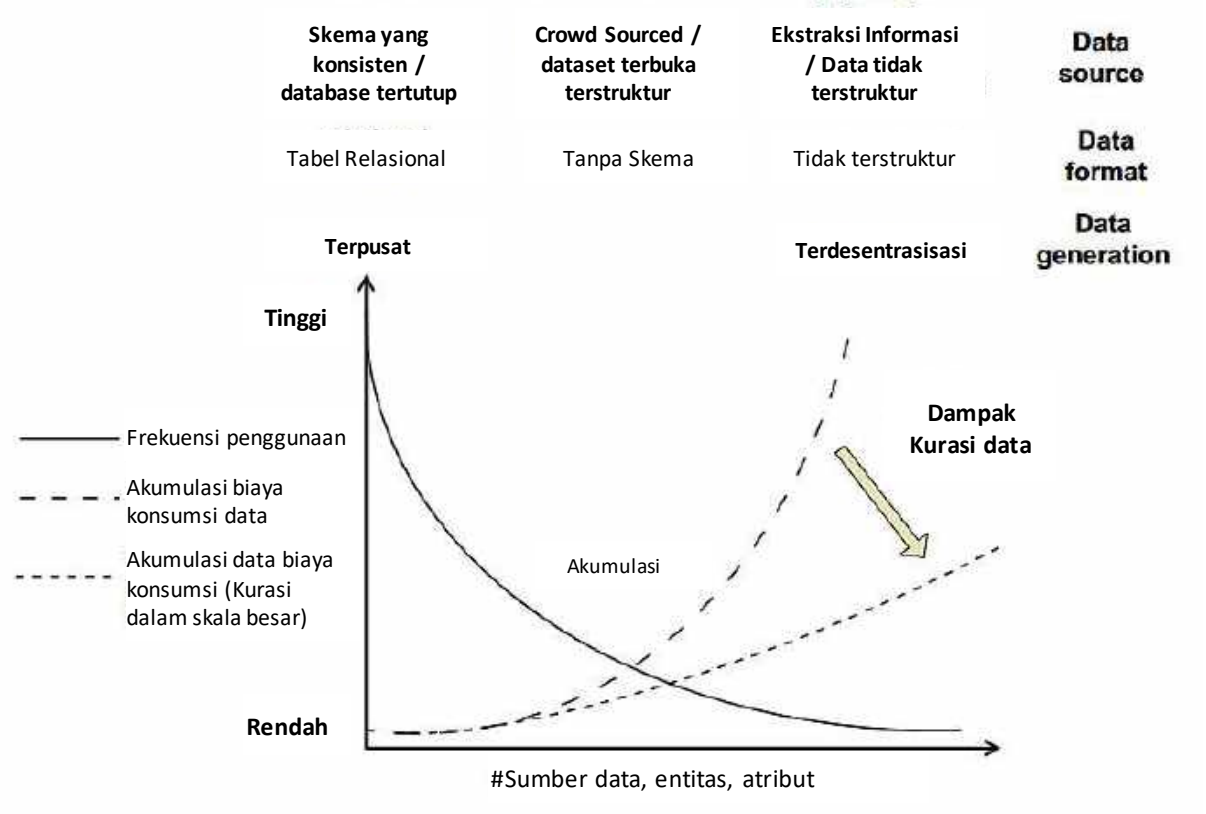
Dalam skenario ini kurasi data muncul sebagai aktivitas pengelolaan data utama. Kurasi data dapat dilihat dari perspektif pembuatan data (kurasi pada sumbernya), dimana data direpresentasikan sedemikian rupa sehingga memaksimalkan kualitasnya dalam konteks yang berbeda. Para ahli menekankan hal ini sebagai aspek penting dalam kurasi data: Dari aspek ilmu data, diperlukan metodologi untuk mendeskripsikan data agar benar-benar dapat digunakan kembali di luar konteks aslinya (Kevin Ashley 2014). Hal ini menunjukkan adanya kebutuhan untuk menyelidiki pendekatan yang memaksimalkan kualitas data dalam berbagai konteks dengan upaya kurasi minimal: *"kami akan melakukan kurasi data sedemikian rupa sehingga idealnya dapat digunakan untuk pertanyaan apa pun yang mungkin ditanyakan oleh seseorang. "* (Kevin Ashley 2014). Kurasi data juga dapat dilakukan di sisi konsumsi data di mana sumber daya data dipilih dan diubah agar sesuai dengan serangkaian persyaratan dari sisi konsumsi data.

Aktivitas kurasi data sangat bergantung pada tantangan skala, khususnya variasi data, yang muncul dalam konteks Big Data. James Cheney, peneliti di University of Edinburgh, mengamati *"Big Data sepertinya adalah tentang mengatasi tantangan skala, dalam hal seberapa cepat hal-hal yang Anda peroleh versus berapa banyak biaya yang harus dikeluarkan untuk mendapatkan nilai dari apa yang sudah Anda miliki"*. Mengatasi keragaman data bisa memakan banyak biaya, bahkan untuk jumlah data yang lebih kecil: *"Anda dapat menghadapi tantangan Big Data bukan hanya karena Anda memiliki data sebesar Petabyte, namun karena data sangat bervariasi sehingga menghabiskan banyak sumber daya untuk memahaminya".*

Sedangkan dalam konteks Big Data, variasi data ekspresi digunakan untuk mengungkapkan tren pengelolaan data dalam menghadapi data dari sumber yang berbeda, konsep kualitas data (Wang dan Strong 1996; Knight dan Burn 2005) dan heterogenitas data (Sheth 1999) telah ditetapkan dengan baik dalam literatur basis data dan memberikan dasar yang tepat untuk memahami tugas-tugas yang terlibat dalam kurasi data. Terlepas dari kenyataan bahwa heterogenitas data dan kualitas data merupakan kekhawatiran yang sudah ada sebelum era skala data besar (Wang dan Strong 1996; Knight dan Burn 2005), hal ini menjadi lebih umum dalam tugas pengelolaan data seiring dengan pertumbuhan jumlah sumber data. Pertumbuhan ini menimbulkan kebutuhan untuk mendefinisikan prinsip-prinsip dan pendekatan yang terukur untuk mengatasi masalah kualitas data. Hal ini juga membawa kurasi data dari aktivitas khusus, yang terbatas pada komunitas kecil ilmuwan dan analis dengan standar kualitas data yang tinggi, menjadi aktivitas pengelolaan data rutin, yang secara bertahap akan semakin hadir dalam lingkungan pengelolaan data rata-rata.

Pertumbuhan jumlah sumber data dan cakupan database menentukan keragaman data yang panjang (Curry dan Freitas 2014). Lingkungan manajemen data relasional tradisional berfokus pada data yang sering dipetakan ke proses bisnis dan cukup teratur untuk dimasukkan ke dalam model relasional. Variasi data ekor panjang (lihat Gambar 6.2) menunjukkan pergeseran ke arah perluasan cakupan data dalam lingkungan pengelolaan data menuju data yang lebih jarang digunakan, lebih terdesentralisasi, dan kurang terstruktur. Ekor panjang memungkinkan konsumen data memiliki model domain mereka yang lebih komprehensif yang dapat dicari, ditanyakan, dianalisis, dan dinavigasi.

Tantangan utama model kurasi data di era Big Data adalah menangani data yang panjang dan meningkatkan skalabilitas kurasi data, dengan mengurangi biaya kurasi data dan meningkatkan jumlah kurator data (Gambar 6.2), memungkinkan data tugas kurasi yang harus ditangani dalam batasan waktu yang terbatas. Meningkatkan kurasi data merupakan masalah multidisiplin yang memerlukan pengembangan model ekonomi, struktur sosial, model insentif, dan standar, yang dikoordinasikan dengan solusi teknologi. Hubungan antara dimensi ini dan skalabilitas kurasi data merupakan inti dari kebutuhan masa depan dan tren kurasi data di masa depan.



Gambar 6.2 Ekor panjang kurasi data dan skalabilitas aktivitas kurasi data

6.4 DAMPAK SOSIAL DAN EKONOMI DARI KURASI BIG DATA

Meningkatnya ketersediaan data memberikan peluang bagi masyarakat untuk menggunakannya sebagai informasi dalam proses pengambilan keputusan, sehingga konsumen data memiliki gambaran realitas yang lebih lengkap dan didukung data. Meskipun beberapa kasus penggunaan data besar didasarkan pada skema berskala besar namun kecil dan kumpulan data reguler, skenario pengambilan keputusan lainnya bergantung pada integrasi data yang kompleks, multi-domain, dan terdistribusi. Ekstraksi nilai dari informasi yang berasal dari sumber data yang berbeda bergantung pada kelayakan pengintegrasian dan analisis sumber data tersebut.

Pengambil keputusan dapat berkisar dari ahli biologi molekuler hingga pejabat pemerintah atau profesional pemasaran dan mereka memiliki kesamaan dalam kebutuhan untuk menemukan pola dan membuat model untuk mengatasi tugas atau tujuan bisnis tertentu. Model-model ini perlu didukung oleh bukti kuantitatif. Meskipun data tidak terstruktur (seperti sumber daya teks) dapat mendukung proses pengambilan keputusan, data terstruktur memberikan kemampuan analitis yang lebih besar kepada pengguna, dengan mendefinisikan representasi terstruktur yang terkait dengan data. Hal ini memungkinkan pengguna untuk membandingkan, menggabungkan, dan mengubah data. Dengan lebih banyak data yang tersedia, hambatan perolehan data berkurang. Namun, untuk mendapatkan manfaat darinya, data perlu diproses secara sistematis, diubah, dan digunakan kembali ke dalam konteks baru.

Area yang bergantung pada representasi model multi-domain dan kompleks memimpin siklus hidup teknologi kurasi data. Proyek eScience memimpin eksperimen dan inovasi dalam kurasi data dan didorong oleh kebutuhan untuk menciptakan infrastruktur untuk meningkatkan reproduktifitas dan kolaborasi multidisiplin skala besar dalam sains. Mereka berperan sebagai visioner dalam siklus hidup adopsi teknologi untuk teknologi kurasi data tingkat lanjut.

Pada fase pengguna awal siklus hidup ini, industri biomedis (khususnya, industri farmasi) merupakan pemain utama, didorong oleh kebutuhan untuk mengurangi biaya dan waktu pemasaran jalur penemuan obat (Wawancara Kurasi Data: Nick Lynch 2014). Bagi perusahaan farmasi, kurasi data sangat penting dalam pengelolaan data organisasi dan integrasi data pihak ketiga. Mengikuti serangkaian persyaratan yang berbeda, industri media juga diposisikan sebagai pengguna awal, menggunakan jalur kurasi data untuk mengklasifikasikan kumpulan besar sumber daya tidak terstruktur (teks dan video), meningkatkan pengalaman konsumsi data melalui aksesibilitas yang lebih baik, dan memaksimalkan penggunaan kembali sumber daya tersebut dalam konteks yang berbeda. Pengadopsi awal terbesar ketiga adalah pemerintah, yang menargetkan transparansi melalui proyek data terbuka (Shadbolt dkk. 2012).

Kurasi data memungkinkan ekstraksi nilai dari data, dan ini merupakan kemampuan yang diperlukan untuk area yang bergantung pada integrasi dan klasifikasi data yang kompleks dan/atau berkelanjutan. Peningkatan alat dan metode kurasi data secara langsung memberikan efisiensi yang lebih besar pada proses penemuan pengetahuan, memaksimalkan pengembalian investasi per item data melalui penggunaan kembali, dan meningkatkan transparansi organisasi.

6.5 KURASI BIG DATA YANG CANGGIH

Bagian ini berkonsentrasi pada penjelasan singkat tentang teknologi yang diadopsi secara luas dan pendekatan yang sudah mapan untuk kurasi data, sedangkan bagian berikutnya berfokus pada kebutuhan masa depan dan pendekatan yang muncul. Manajemen Data Master terdiri dari proses dan alat yang mendukung satu titik acuan untuk data suatu organisasi, sumber data resmi. Alat *Manajemen Data Master* (MDM) dapat digunakan untuk menghapus duplikat dan menstandardisasi sintaksis data, sebagai sumber data master yang otoritatif. MDM berfokus untuk memastikan bahwa suatu organisasi tidak menggunakan versi data master yang sama yang banyak dan tidak konsisten di berbagai bagian sistemnya. Proses dalam MDM meliputi identifikasi sumber, transformasi data, normalisasi, administrasi aturan, deteksi dan koreksi kesalahan, konsolidasi data, penyimpanan data, klasifikasi, layanan taksonomi, pemetaan skema, dan pengayaan semantik.

Manajemen data master sangat terkait dengan kualitas data. Menurut Morris dan Vesset (2005), tiga tujuan utama MDM adalah:

1. Menyinkronkan data master di beberapa contoh aplikasi perusahaan
2. Mengkoordinasikan pengelolaan data master pada saat migrasi aplikasi
3. Pelaporan manajemen kepatuhan dan kinerja di berbagai sistem analitik

Rowe (2012) memberikan analisis tentang bagaimana 163 organisasi menerapkan MDM dan dampak bisnisnya.

Kurasi di Sumber Kurasi belaka atau kurasi di sumber adalah pendekatan kurasi data di mana aktivitas kurasi ringan diintegrasikan ke dalam alur kerja normal mereka yang membuat dan mengelola data dan aset digital lainnya (Curry dkk. 2010). Aktivitas kurasi belaka dapat mencakup aktivitas kategorisasi ringan dan normalisasi. Contohnya adalah memeriksa atau "memberi peringkat" hasil proses kategorisasi yang dilakukan oleh algoritma kurasi. Aktivitas kurasi murni juga dapat digabungkan dengan aktivitas kurasi lainnya, sehingga memungkinkan akses lebih cepat ke data kurasi sekaligus memastikan kontrol kualitas yang hanya mungkin dilakukan oleh tim kurasi ahli.

Berikut ini adalah tujuan tingkat tinggi dari kurasi belaka yang dijelaskan oleh Hedges dan Blanke (2012):

- Hindari penyimpanan data dengan mengintegrasikan dengan alat alur kerja normal
- Menangkap informasi asal alur kerja
- Antarmuka yang mulus dengan infrastruktur kurasi data

Kurasi data crowdsourcing dapat menjadi tugas yang membutuhkan banyak sumber daya dan kompleks, yang dapat dengan mudah melampaui kapasitas satu individu. Sebagian besar upaya kurasi data yang tidak sepele bergantung pada pengaturan kurasi data kolektif, di ma na para peserta dapat berbagi biaya, risiko, dan tantangan teknis. Tergantung pada domain, skala data, dan jenis kegiatan kurasi, upaya kurasi data dapat memanfaatkan komunitas terkait melalui undangan atau kerumunan (Doan et al. 2011). Sistem ini dapat berkisar dari sistem dengan basis partisipasi yang besar dan terbuka seperti Wikipedia (berbasis kerumunan) hingga sistem atau kelompok pakar domain yang lebih terbatas, seperti Chemspider.

Gagasan mengenai *"kebijaksanaan orang banyak"* menganjurkan bahwa kelompok non-ahli yang berpotensi besar dapat memecahkan masalah-masalah kompleks yang biasanya dianggap hanya dapat diselesaikan oleh para ahli (Surowiecki 2005). Crowdsourcing telah muncul sebagai paradigma yang kuat untuk pekerjaan outsourcing dalam skala besar dengan bantuan orang-orang online (Doan et al. 2011). Crowdsourcing didorong oleh pesatnya perkembangan teknologi web yang memfasilitasi kontribusi dari jutaan pengguna online. Asumsi yang mendasarinya adalah bahwa tenaga kerja dalam skala besar dan murah dapat diperoleh melalui web.

Efektivitas crowdsourcing telah dibuktikan melalui situs web seperti Wikipedia, Amazon Mechanical Turk, dan Kaggle. Wikipedia mengikuti pendekatan crowdsourcing sukarela di mana masyarakat umum diminta untuk berkontribusi pada proyek pembuatan ensiklopedia demi kepentingan semua orang (Kittur dkk.2007). Amazon Mechanical Turk menyediakan pasar tenaga kerja untuk tugas crowdsourcing tanpa uang (Ipeirotis 2010). Kaggle memungkinkan organisasi untuk mempublikasikan masalah yang harus diselesaikan melalui kompetisi antar peserta untuk mendapatkan hadiah yang telah ditentukan. Meskipun berbeda dalam hal model insentif, semua situs web ini memungkinkan akses ke sejumlah besar pekerja, sehingga memungkinkan mereka digunakan sebagai platform rekrutmen untuk perhitungan manusia (Law dan von Ahn 2011). Platform layanan crowdsourcing tujuan umum

seperti CrowdFlower (CrowdFlower Whitepaper 2012) atau Amazon Mechanical Turk (Ipeirotis 2010) memungkinkan proyek merutekan tugas untuk kelompok berbayar. Pengguna layanan diabstraksi dari upaya mengumpulkan massa dan menawarkan tugasnya dengan harga tertentu di pasar pekerja massal. Platform layanan crowdsourcing memberikan model yang fleksibel dan dapat digunakan untuk menangani tugas kurasi data skala kecil ad hoc (seperti klasifikasi sederhana ribuan gambar untuk proyek penelitian), volume kurasi data puncak (misalnya pemetaan dan penerjemahan data dalam keadaan darurat). situasi respons), atau pada volume kurasi reguler (misalnya kurasi data berkelanjutan untuk suatu perusahaan).

Ruang kolaborasi seperti platform Wiki dan Sistem Manajemen Konten (CMS) memungkinkan pengguna untuk secara kolaboratif membuat dan mengkurasi data yang tidak terstruktur dan terstruktur. Meskipun CMS berfokus untuk memungkinkan kelompok yang lebih kecil dan lebih terbatas untuk secara kolaboratif mengedit dan mempublikasikan konten online (seperti platform Berita, blog, dan eCommerce), Wiki telah terbukti mampu menjangkau basis pengguna yang sangat besar. Pada tahun 2014, Wikipedia memiliki lebih dari 4.000.000 artikel dan memiliki komunitas dengan lebih dari 130.000 kontributor aktif terdaftar.

Wikipedia menggunakan wiki sebagai sistem utamanya untuk konstruksi konten. Wiki pertama kali diusulkan oleh Ward Cunningham pada tahun 1995 dan memungkinkan pengguna untuk mengedit konten dan berkolaborasi di web dengan lebih efisien. MediaWiki, platform wiki di balik Wikipedia, sudah banyak digunakan sebagai lingkungan kolaboratif di dalam organisasi. Kasus-kasus penting termasuk Intellipedia, penerapan platform MediaWiki yang mencakup 16 badan Intelijen AS, dan Wiki Proteins, sebuah lingkungan kolaboratif untuk penemuan dan anotasi pengetahuan.

Wikipedia mengandalkan cara yang sederhana namun sangat efektif untuk mengoordinasikan proses kurasinya, dan akun serta peran merupakan dasar dari sistem ini. Semua pengguna diperbolehkan mengedit konten Wikipedia. Administrator, bagaimanapun, memiliki izin tambahan dalam sistem (Curry et al. 2010. Sebagian besar platform Wiki dan CMS menargetkan konten data tidak terstruktur dan semi terstruktur, memungkinkan pengguna untuk mengklasifikasikan dan menghubungkan konten tidak terstruktur.

Platform Kurasi Data

a. **Penjinak Data**: Prototipe ini bertujuan untuk menggantikan proses ekstrak-transformasi-beban (ETL) yang berpusat pada pengembang dengan integrasi data otomatis. Sistem ini menggunakan serangkaian algoritme untuk secara otomatis memetakan skema dan menghilangkan duplikat entitas. Namun, pakar manusia dan orang banyak dimanfaatkan untuk memverifikasi pembaruan integrasi yang sangat sulit untuk algoritme.
b. **ZenCrowd**: Sistem ini mencoba mengatasi masalah menghubungkan entitas bernama dalam teks dengan basis pengetahuan. ZenCrowd menjembatani kesenjangan antara penautan otomatis dan manual dengan meningkatkan hasil penautan otomatis dengan manusia. Prototipe ini didemonstrasikan untuk menghubungkan entitas bernama dalam artikel berita dengan entitas di cloud data terbuka yang tertaut.

c. **CrowdDB**: Sistem database ini menjawab pertanyaan SQL yang tidak dapat dijawab oleh sistem manajemen database atau mesin pencari. Berbeda dengan operasi eksak dalam database, CrowdDB memungkinkan operasi fuzzy dengan bantuan manusia, misalnya, memeringkat item berdasarkan relevansi atau membandingkan kesetaraan gambar.
d. **Qurk**: Meskipun mirip dengan CrowdDB, sistem ini mencoba meningkatkan biaya dan latensi dari jenis dan gabungan yang didukung manusia. Dalam hal ini, Qurk menerapkan teknik seperti batching, filtering, dan output agreement.
e. **Bot Wikipedia**: Wikipedia menjalankan algoritme terjadwal untuk mengakses kualitas artikel teks, yang dikenal sebagai Bot. Bot ini juga menandai artikel yang memerlukan peninjauan lebih lanjut oleh para ahli. SuggestBot merekomendasikan artikel yang ditandai ke editor Wikipedia berdasarkan profil mereka.

6.6 TREN DAN SYARAT UNTUK KURASI BIG DATA

Bagian ini bertujuan untuk memberikan peta jalan kurasi data berdasarkan serangkaian persyaratan kurasi data di masa depan dan pendekatan kurasi data yang muncul untuk mengatasi persyaratan tersebut. Persyaratan masa depan dan pendekatan yang muncul dikumpulkan melalui analisis ekstensif terhadap pendekatan modern.

Persyaratan Masa Depan untuk Kurasi Big Data

Daftar persyaratan masa depan dikumpulkan dengan memilih dan mengkategorikan permintaan yang paling sering muncul dalam survei modern dan yang muncul dalam wawancara pakar domain sebagai arah mendasar untuk masa depan kurasi data. Setiap kebutuhan dikategorikan berdasarkan atribut berikut (Tabel 6.1):

- **Dimensi Persyaratan Inti**: Terdiri dari kategori utama yang diperlukan untuk memenuhi persyaratan. Dimensinya adalah teknis, sosial, insentif, metodologi, standardisasi, ekonomi, dan kebijakan.
- **Tingkat Dampak**: Terdiri dari dampak persyaratan bidang kurasi data. Berdasarkan konstruksinya, hanya persyaratan di atas ambang batas dampak tertentu yang dicantumkan. Nilai yang mungkin adalah sedang, sedang-tinggi, tinggi, sangat tinggi.
- **Daerah Yang Terkena Dampak**: Daftar daerah yang paling terkena dampak persyaratan ini. Nilai yang mungkin adalah ilmu pengetahuan, pemerintahan, sektor industri (keuangan, kesehatan, media dan hiburan, telekomunikasi, manufaktur), dan lingkungan hidup.
- **Prioritas**: Meliputi tingkat prioritas yang berhubungan dengan kebutuhan. Nilai yang mungkin adalah: jangka pendek (<3 tahun), jangka menengah (3–7 tahun), dan konsolidasi (>7 tahun).
- **Aktor Inti**: Meliputi aktor-aktor utama yang harus bertanggung jawab untuk memenuhi kebutuhan inti. Pelaku inti adalah pemerintah, industri, akademisi, organisasi non-pemerintah, dan komunitas pengguna.

Tabel. 6.1 Persyaratan kurasi data di masa depan

Kategori persyaratan	*Persyaratan*	*Dimensi kebutuhan inti*	*Tingkat dampak*	*Daerah yang terkena dampak*	*Prioritas*	*Aktor inti*
Penciptaan insentif	Penciptaan mekanisme insentif untuk pemeliharaan dan publikasi kumpulan data yang dikurasi	Ekonomi, sosial, kebijakan	Sangat tinggi	Ilmu pengetahuan, pemerintahan, lingkungan hidup, keuangan, kesehatan	Jangka pendek	Pemerintah
Model ekonomi	Definisi model ekonomi data	Kebijakan ekonomi	Sangat tinggi	Semua sektor	Jangka pendek	Pemerintah, industri
Mekanisme keterlibatan sosial	Pemahaman tentang mekanisme keterlibatan sosial	Sosial, teknis	Sedang	Sains, pemerintahan, lingkungan	Jangka panjang	Akademisi, LSM, industri
Kurasi dalam skala besar	Pengurangan biaya yang terkait dengan tugas kurasi data (skalabilitas)	Teknis, sosial, ekonomi	Sangat tinggi	Semua sektor	Jangka menengah	Akademisi, industri, komunitas pengguna
Interaksi manusia-data	Peningkatan aspek interaksi manusia-data. Memungkinkan pakar domain dan pengguna biasa untuk membuat kueri, menjelajahi, mengubah, dan mengkurasi data	Teknis	Sangat tinggi	Semua sektor	Jangka panjang	Akademisi, industri
Memercayai	Dimasukkannya mekanisme kepercayaan dalam kurasi data	Teknis	Tinggi	Semua sektor	Jangka pendek	Akademisi, industri
Standardisasi dan interoperabilitas	Integrasi dan interoperabilitas antar platform/standardisasi kurasi data	Teknis, sosial, kebijakan, metodologis	Sangat tinggi	Semua sektor	Jangka pendek	Komunitas pengguna, industri, akademisi
Model kurasi	Investigasi model teoretis dan spesifik domain untuk kurasi data	Teknis, metodologis	Sedang-tinggi	Semua sektor	Jangka panjang	Akademisi
Tidak terstruktur-integrasi terstruktur	Integrasi yang lebih baik antara data dan alat tidak terstruktur dan terstruktur	Teknis	Sedang	Sains, media, kesehatan, keuangan, pemerintahan	Jangka panjang	Akademisi, industri

Paradigma yang Muncul untuk Kurasi Big Data

Dalam analisis modern, pendekatan sosial, teknis, dan metodologis utama muncul untuk memenuhi kebutuhan masa depan. Pada bagian ini, pendekatan-pendekatan yang muncul ini dijelaskan serta cakupannya dalam kaitannya dengan kategori persyaratan. Pendekatan yang muncul didefinisikan sebagai pendekatan yang adopsinya terbatas. Pendekatan-pendekatan ini dirangkum dalam Tabel 6.2.

Tabel 6.2 Pendekatan yang muncul untuk mengatasi kebutuhan di masa depan

Kategori persyaratan	Pendekatan yang muncul	Adopsi/status	Contoh kasus penggunaan
Mekanisme penciptaan insentif dan keterlibatan sosial	Kebijakan data yang terbuka dan dapat dioperasikan	Tahap awal/Adopsi terbatas	Data.gov.uk
	Pengakuan yang lebih baik atas peran kurasi data	Kurangnya adopsi/ Meskipun terdapat kasus penggunaan yang patut dicontoh, peran kurator data masih belum diakui	Chemspider, Wikipedia, Protein Data Bank
	Atribusi dan pengakuan atas kontribusi data dan infrastruktur	Standar yang muncul/ Adopsi hilang	Altmetrics (Priem et al. 2010), ORCID
	Pemahaman yang lebih baik tentang mekanisme keterlibatan sosial	Tahap awal	GalaxyZoo (Forston et al. 2011), Foldit (Khatib et al. 2011)
Model ekonomi	Kemitraan pra-kompetitif	Kasus penggunaan penting	Pistoia Alliance (Barnes et al. 2009)
	Kemitraan publik-swasta	Kasus penggunaan penting	Geoconnections (Harper 2012)
	Kuantifikasi dampak ekonomi dari data	Kasus penggunaan penting	Technopolis Group (2011) ("Data centres: their use, value and impact")
Kurasi dalam skala besar	Komputasi manusia dan layanan Crowdsourcing	Adopsi/Layanan tingkat industri tersedia tetapi ada ruang untuk spesialisasi pasar	CrowdFlower, Amazon Mechanical Turk
	Model pengukuran ketidakpastian data berbasis bukti	Tahap penelitian	IBM Watson (Ferrucci et al. 2010)
	Pemrograman dengan demonstrasi, induksi alur kerja transformasi data	Tahap penelitian/bidang penelitian fundamental dikembangkan. Kurangnya penelitian terapan dalam konteks alur kerja dan kurasi data	Tuchinda et al. (2007), Tuchinda (2011)
	Kurasi di sumbernya	Kasus penggunaan yang ada baik dalam proyek akademik maupun industry	The New York Times

	Jalur kurasi data tujuan umum	Infrastruktur yang Tersedia	OpenRefine, Karma, Scientific Workflow management systems
	Validasi/anotasi algoritmik	Tahap awal	Wikipedia, Chemspider
Interaksi manusia-data	Fokus pada kemudahan interaktivitas	Alat penting tersedia	OpenRefine
	Antarmuka bahasa alami, kueri skema-agnostik	Tahap penelitian	IBM Watson (Ferrucci et al. 2010), Treo (Freitas and Curry 2014)
Memercayai	Pengambilan keputusan kurasi data	Standar sudah ada, instrumentasi aplikasi diperlukan	OpenPhacts
	Model dan alat manajemen izin yang terperinci	Infrastruktur kasar tersedia.	Qin and Atluri (2003), Ryutov et al. (2009), Kirrane et al. (2013), Rodrıguez-Doncel et al. (2013)
Standardisasi dan interoperabilitas	Model data standar	Standar tersedia	RDF(S), OWL
	Penggunaan kembali kosakata	Teknologi untuk mendukung penggunaan kembali kosakata diperlukan	Linked Open Data Web (Berners-Lee 2009)
	Integrasi dan komunikasi yang lebih baik antar alat	Rendah	N/A
	Representasi asal yang dapat dioperasikan	Standar yang ada/Adopsi standar masih belum ada	W3C PROV
Model kurasi	Definisi model informasi minimum untuk kurasi data	Adopsi yang rendah	MIRIAM (Laibe and Le Nove`re 2007)
	Publikasi nano	Konsep yang muncul	Mons and Velterop (2009), Groth et al. (2010)
	Investigasi prinsip teoritis dan model khusus domain untuk kurasi data	Konsep yang muncul	Pearl and Bareinboim (2011)
Tidak terstruktur-integrasi terstruktur	Saluran Pipa NLP	Alat tersedia, adopsi rendah	IBM Watson (Ferrucci et al. 2010)
	Pengenalan dan penyesuaianan entitas	Alat tersedia, adopsi rendah	DBpedia Spotlight (Mendes et al. 2011), IBM Watson (Ferrucci et al. 2010)

Insentif Sosial dan Mekanisme Keterlibatan

Kebijakan Data yang Terbuka dan Dapat Dioperasikan Permintaan akan data berkualitas tinggi adalah pendorong evolusi platform kurasi data. Upaya untuk menghasilkan dan memelihara data berkualitas tinggi perlu didukung oleh sistem insentif yang solid, yang pada saat ini belum sepenuhnya berjalan. Data terbuka berkualitas tinggi dapat menjadi salah satu pendorong dampak sosial dengan mendukung ilmu pengetahuan yang lebih efisien dan dapat direproduksi (eScience) (Norris 2007), serta pemerintahan yang lebih transparan dan efisien (eGovernment) (Shadbolt dkk. 2012). Sektor-sektor ini berperan sebagai inovator dan pengguna awal dalam siklus hidup adopsi teknologi kurasi data dan merupakan pendorong utama inovasi dalam alat dan metode kurasi data. Lembaga pendanaan dan pembuat kebijakan mempunyai peran mendasar dalam proses ini dan harus mengarahkan dan mendukung para ilmuwan dan pejabat pemerintah untuk menyediakan produk data mereka dengan cara yang dapat dioperasikan. Permintaan akan data berkualitas tinggi dan dapat dioperasikan dapat mendorong evolusi metode dan alat kurasi data.

Atribusi dan Pengakuan Kontribusi Data dan Infrastruktur Dari perspektif eScience, komite ilmiah dan editorial dari publikasi bergengsi memiliki kekuatan untuk mengubah lanskap metodologis komunikasi ilmiah, dengan menekankan reproduktifitas dalam proses peninjauan dan dengan mengharuskan publikasi didukung oleh sumber daya manusia yang tinggi. data berkualitas bila berlaku. Dari sudut pandang ilmuwan, publikasi yang didukung oleh data dapat memfasilitasi reproduktifitas dan menghindari pengerjaan ulang dan sebagai konsekuensinya meningkatkan efisiensi ilmiah dan dampak produk ilmiah. Selain itu, seiring dengan semakin lazimnya data sebagai produk ilmiah utama, data menjadi sumber daya yang dapat dimanfaatkan. Mekanisme seperti ORCID (Thomson Reuters Technical Report 2013) dan Altmetrics (Priem dkk. 2010) telah menyediakan elemen pendukung untuk mengidentifikasi, mengatribusikan, dan mengukur keluaran dampak seperti kumpulan data dan perangkat lunak. Pengakuan atas kontribusi data dan perangkat lunak dalam sistem evaluasi akademik merupakan elemen penting untuk mendorong data ilmiah berkualitas tinggi.

Pengakuan yang Lebih Baik atas Peran Kurasi Data Biaya penerbitan data berkualitas tinggi tidak dapat diabaikan dan harus menjadi bagian eksplisit dari perkiraan biaya proyek dengan hasil data. Selain itu, dampak metodologis dari kurasi data mengharuskan peran kurator data lebih dikenal di seluruh jalur ilmiah dan penerbitan. Beberapa organisasi dan proyek telah memiliki definisi yang jelas tentang berbagai peran kurator data. Contohnya adalah Wikipedia, New York Times (Curry et al. 2010), dan Chemspider (Pence dan Williams 2010). Pembaca dirujuk ke studi kasus untuk memahami aktivitas berbagai peran kurasi data.

Pemahaman yang Lebih Baik tentang Mekanisme Keterlibatan Sosial Meskipun sebagian dari struktur insentif mungkin dipicu oleh kebijakan publik, atau keuntungan finansial langsung, sebagian lainnya mungkin muncul dari manfaat langsung karena menjadi bagian dari proyek yang bermakna bagi komunitas pengguna. Proyek seperti Wikipedia, GalaxyZoo (Forston et al. 2011), atau FoldIt (Khatib et al. 2011) telah mengumpulkan sejumlah besar relawan kurator data yang mengeksplorasi serangkaian mekanisme insentif yang berbeda, yang dapat didasarkan pada visibilitas dan status sosial atau profesional. dampak

sosial, kebermaknaan, atau kesenangan. Pemahaman prinsip-prinsip ini dan pengembangan mekanisme di balik keterlibatan basis pengguna yang besar merupakan isu penting untuk memperkuat upaya kurasi data.

Model Ekonomi

Model ekonomi yang sedang berkembang dapat memberikan landasan finansial untuk mendukung pembuatan dan pemeliharaan data berkualitas tinggi serta infrastruktur kurasi data terkait. Kemitraan Pra-kompetitif untuk Kurasi Data Skema kolaborasi pra-kompetitif adalah salah satu model ekonomi di mana konsorsium organisasi, yang biasanya merupakan pesaing, berkolaborasi dalam bagian proses Penelitian & Pengembangan (R&D) yang tidak berdampak pada komersial mereka. keunggulan kompetitif. Hal ini memungkinkan mitra untuk berbagi biaya dan risiko yang terkait dengan bagian-bagian proses penelitian dan pengembangan. Salah satu contoh model ini adalah Pistoia Alliance (Barnes dkk. 2009), yang merupakan aliansi prakompetitif yang terdiri dari perusahaan-perusahaan ilmu hayati, vendor, penerbit, dan kelompok akademis yang bertujuan untuk menurunkan hambatan terhadap inovasi dengan meningkatkan interoperabilitas proses bisnis R&D. Aliansi Pistoia didirikan oleh perusahaan farmasi seperti AstraZeneca, GSK, Pfizer, dan Novartis, dan contoh sumber daya bersama mencakup data dan alat infrastruktur data.

Kemitraan Data Pemerintah-Swasta untuk Kurasi Model ekonomi lain yang muncul untuk kurasi data adalah kemitraan publik-swasta (KPS), yang mana perusahaan swasta dan sektor publik berkolaborasi menuju kemitraan yang saling menguntungkan. Dalam KPS, risiko, biaya, dan manfaat ditanggung bersama di antara para mitra, yang mempunyai kepentingan yang tidak saling bersaing dan saling melengkapi atas data tersebut. Data geospasial dan dampaknya yang besar terhadap sektor publik (lingkungan, administrasi) dan swasta (perusahaan sumber daya alam) merupakan salah satu kasus awal KPS. GeoConnections Canada adalah contoh inisiatif KPS yang diluncurkan pada tahun 1999, dengan tujuan mengembangkan *Infrastruktur Data Geospasial Kanada* (CGDI) dan mempublikasikan informasi geospasial di web (Harper 2012; Wawancara Kurasi Data: Joe Sewash 2014). GeoConnections telah dikembangkan dengan model kolaboratif yang melibatkan partisipasi lembaga federal, provinsi, dan teritorial, serta sektor swasta dan akademik.

Kuantifikasi Dampak Ekonomi Data Pengembangan pendekatan untuk mengukur dampak ekonomi, penciptaan nilai, dan biaya terkait di balik sumber daya data merupakan elemen mendasar untuk membenarkan investasi swasta dan publik dalam infrastruktur data. Salah satu contoh kuantifikasi nilai adalah studi JISC "Pusat data: penggunaan, nilai, dan dampaknya" (Technopolis Group 2011), yang memberikan perhitungan kuantitatif proses penciptaan nilai di delapan pusat data. Penciptaan ukuran keuangan kuantitatif dapat memberikan bukti yang diperlukan untuk mendukung investasi infrastruktur data baik pemerintah maupun swasta, menciptakan model bisnis berkelanjutan berdasarkan aset data, dan memperluas ekonomi data yang ada.

Kurasi dalam Skala Besar

Komputasi Manusia dan Layanan Crowdsourcing Platform crowdsourcing berkembang pesat namun masih terdapat peluang besar untuk diferensiasi dan pertumbuhan pasar.

CrowdFlower, misalnya, berkembang ke arah penyediaan API yang lebih baik, mendukung integrasi yang lebih baik dengan sistem eksternal.

Dalam platform crowdsourcing, orang-orang menunjukkan variabilitas dalam kualitas pekerjaan yang mereka hasilkan, serta jumlah waktu yang mereka gunakan untuk pekerjaan yang sama. Selain itu, keakuratan dan latensi prosesor manusia tidak seragam seiring berjalannya waktu. Oleh karena itu, diperlukan metode yang tepat untuk mengarahkan tugas kepada orang yang tepat pada waktu yang tepat (Hassan et al. 2012). Selain itu, menggabungkan pekerjaan yang dilakukan oleh orang-orang berbeda pada tugas yang sama juga dapat membantu meningkatkan kualitas pekerjaan (Law dan von Ahn 2009). Perekrutan manusia yang cocok untuk komputasi merupakan tantangan utama komputasi manusia.

Saat ini, sebagian besar platform ini terbatas pada tugas-tugas yang dapat didelegasikan kepada audiens umum berbayar. Peluang diferensiasi di masa depan mencakup: (1) dukungan untuk pakar domain yang sangat terspesialisasi, (2) lebih banyak fleksibilitas dalam pemilihan profil demografis, (3) penciptaan hubungan jangka panjang (lebih persisten) dengan tim pekerja, (4) penciptaan platform layanan crowdsourcing terbuka dengan tujuan umum untuk pekerjaan sukarela, dan (5) menggunakan data historis untuk memberikan produktivitas dan otomatisasi yang lebih besar bagi kurator data (Kittur dkk. 2007).

Menginstrumentasikan Aplikasi Populer untuk Kurasi Data Dalam kebanyakan kasus, kurasi data dilakukan dengan aplikasi perkantoran umum: spreadsheet biasa, editor teks, dan email (Wawancara Kurasi Data: James Cheney 2014). Alat-alat ini merupakan bagian intrinsik dari infrastruktur kurasi data yang ada dan pengguna sudah familiar dengannya. Namun, alat-alat ini tidak memiliki beberapa fungsi mendasar untuk kurasi data: (1) menangkap dan merepresentasikan tindakan pengguna; (2) mekanisme anotasi/penggunaan kembali kosakata; (3) kemampuan menangani data berskala besar; (4) kemampuan pencarian yang lebih baik; dan (5) integrasi dengan berbagai sumber data. Memperluas aplikasi dengan basis pengguna yang besar untuk kurasi data memberikan peluang untuk penetrasi fungsionalitas kurasi data yang rendah ke dalam infrastruktur kurasi data yang lebih ad hoc. Hal ini memungkinkan penggabungan proses kurasi data mendasar ke dalam aktivitas rutin yang ada tanpa gangguan besar pada proses kerja pengguna (Wawancara Kurasi Data: Carole Goble 2014).

Jalur Kurasi Data Bertujuan Umum Meskipun adaptasi dan instrumentasi alat reguler dapat memberikan solusi kurasi data generik yang berbiaya rendah, banyak proyek akan memerlukan penggunaan alat yang dirancang sejak awal untuk mendukung aktivitas kurasi data yang lebih canggih. Pengembangan kerangka kerja kurasi data tujuan umum yang mengintegrasikan fungsi kurasi data inti ke dalam platform kurasi data berskala besar merupakan elemen mendasar bagi organisasi yang melakukan kurasi data berskala besar. Platform seperti Open Refine4 dan Karma (Gil et al. 2011) memberikan contoh kerangka kurasi data yang sedang berkembang, dengan fokus pada transformasi dan integrasi data. Berbeda dengan kerangka kerja Extract Transform Load (ETL), platform kurasi data memberikan dukungan yang lebih baik untuk transformasi dan integrasi data ad hoc, dinamis, manual, lebih jarang (long tail), dan lebih sedikit skrip. Jalur pipa ETL dapat dilihat sebagai pemusatan

aktivitas berulang yang menjadi lebih formal ke dalam proses tertulis. Platform kurasi data untuk tujuan umum harus menargetkan pakar domain, mencoba menyediakan alat yang dapat digunakan oleh orang-orang di luar latar belakang ilmu komputer/teknologi informasi.

Validasi/Anotasi Algoritmik Arah utama lainnya untuk mengurangi biaya kurasi data terkait dengan otomatisasi aktivitas kurasi data. Algoritma menjadi lebih cerdas seiring dengan kemajuan pembelajaran mesin dan kecerdasan buatan. Kecerdasan mesin diharapkan mampu memvalidasi, memperbaiki, dan membuat anotasi data dalam hitungan detik, yang mungkin memerlukan waktu berjam-jam bagi manusia untuk melakukannya (Kong et al. 2011). Akibatnya, manusia akan terlibat sesuai kebutuhan, misalnya. untuk menentukan aturan kurasi, memvalidasi hard instance, atau menyediakan data untuk algoritma pelatihan (Hassan et al. 2012).

Bentuk otomatisasi paling sederhana terdiri dari aktivitas kurasi skrip yang berulang, menciptakan agen kurasi khusus. Pendekatan ini digunakan, misalnya, di Wikipedia (Wiki Bots) untuk membersihkan artikel dan mendeteksi vandalisme. Proses otomatisasi lainnya terdiri dari penyediaan pendekatan algoritmik untuk validasi atau anotasi data terhadap standar referensi (Wawancara Kurasi Data: Antony Williams 2014). Hal ini akan berkontribusi pada *"likesonomy"* di mana manusia dan algoritme dapat memberikan bukti lebih lanjut yang mendukung atau menentang data (Wawancara Kurasi Data: Antony Williams 2014). Pendekatan ini memberikan cara untuk mengotomatiskan bagian tugas kurasi yang lebih berulang dan dapat diterapkan saat ini di jalur kurasi mana pun (tidak ada hambatan teknologi yang besar). Namun, pembangunan basis algoritmik atau referensi ini memerlukan biaya yang tinggi (dalam hal konsumsi waktu dan keahlian), karena bergantung pada formalisasi eksplisit algoritma atau kriteria referensi (aturan).

Otomatisasi Kurasi Data Pendekatan otomatisasi yang lebih canggih yang dapat mengurangi kebutuhan formalisasi aktivitas kurasi secara eksplisit akan memainkan peran mendasar dalam mengurangi biaya kurasi data. Terdapat potensi signifikan penerapan pembelajaran mesin di bidang kurasi data. Dua bidang penelitian yang dapat memengaruhi otomatisasi kurasi data adalah:

- **Kurasi dengan Demonstrasi (CbD)/Induksi Alur Kerja Kurasi Data**: Pemrograman dengan contoh [atau pemrograman dengan demonstrasi (PbD)] (Cypher 1993; Flener dan Schmid 2008; Lieberman 2001) adalah serangkaian pendekatan pengembangan pengguna akhir di mana tindakan pengguna pada contoh nyata digeneralisasikan ke dalam sebuah program. PbD dapat digunakan untuk memungkinkan distribusi dan penguatan tugas pengembangan sistem dengan memungkinkan pengguna menjadi pemrogram. Meskipun merupakan bidang penelitian tradisional, dan dengan penelitian tentang integrasi data PbD (Tuchinda et al. 2007, 2011), metode PbD belum diterapkan secara luas ke dalam sistem kurasi data.
- **Model Pengukuran Ketidakpastian atas Data yang Berbasis Bukti**: Kuantifikasi dan estimasi model ketidakpastian yang umum dan spesifik domain dari basis bukti yang terdistribusi dan heterogen dapat memberikan dasar bagi keputusan tentang apa yang harus didelegasikan atau divalidasi oleh manusia dan apa yang harus didelegasikan atau

divalidasi oleh manusia. dapat didelegasikan ke pendekatan algoritmik. IBM Watson adalah contoh sistem yang menggunakan model statistik sebagai pusatnya untuk menentukan kemungkinan suatu jawaban benar (Ferrucci dkk. 2010). Model ketidakpastian juga dapat digunakan untuk mengarahkan tugas sesuai dengan tingkat keahlian, meminimalkan biaya, dan memaksimalkan kualitas kurasi data.

Interaksi Manusia-Data

Interaktivitas dan Kemudahan Tindakan Kurasi Pendekatan interaksi data yang memfasilitasi transformasi dan akses data merupakan hal mendasar untuk memperluas spektrum profil kurator data. Masih terdapat hambatan besar dalam berinteraksi dengan data terstruktur dan proses kueri, analisis, dan modifikasi data di dalam database dalam banyak kasus dimediasi oleh profesional TI atau aplikasi khusus domain. Mendukung pakar domain dan pengguna biasa dalam membuat kueri, menavigasi, menganalisis, dan mentransformasikan data terstruktur adalah fungsi mendasar dalam platform kurasi data.

Menurut Carole Goble "dari perspektif Big Data, tantangannya adalah menemukan potongan, pandangan, atau cara ke dalam kumpulan data yang memungkinkan Anda menemukan bagian yang perlu diedit, diubah" (Wawancara Kurasi Data: Carole Goble 2014). Oleh karena itu, peringkasan dan visualisasi data yang tepat penting tidak hanya dari sudut pandang penggunaan tetapi juga dari sudut pandang pemeliharaan (Hey dan Trefethen 2004). Khususnya, untuk metode pembersihan data kolaboratif, penting untuk memungkinkan penemuan anomali dalam data terstruktur dan tidak terstruktur. Selain itu, menjadikan aktivitas pengelolaan data lebih mobile dan interaktif diperlukan karena perangkat seluler melampaui desktop. Teknologi berikut memberikan arahan menuju interaksi yang lebih baik:

a. **Dokumen Berbasis Data (D3.js)**: D3.js adalah pustaka untuk menampilkan grafik interaktif dalam dokumen web. Perpustakaan ini mematuhi standar web terbuka seperti HTML5, SVG, dan CSS, untuk memungkinkan visualisasi yang kuat dengan lisensi sumber terbuka.
b. **Tableau**: Perangkat lunak ini memungkinkan pengguna untuk memvisualisasikan berbagai dimensi database relasional. Selain itu, ini memungkinkan visualisasi data tidak terstruktur melalui adaptor pihak ketiga. Tableau telah menerima banyak perhatian karena kemudahan penggunaannya dan akses gratis ke rencana publik.
c. **Open Refine**: Aplikasi open source ini memungkinkan pengguna untuk membersihkan dan mengubah data dari berbagai format seperti CSV, XML, RDF, JSON, dll. Open Refine sangat berguna untuk menemukan outlier dalam data dan memeriksa distribusi nilai dalam kolom melalui aspek. Memungkinkan rekonsiliasi data dengan sumber data eksternal seperti Freebase dan OpenCorporates.

Bahasa kueri terstruktur seperti SQL adalah pendekatan default untuk berinteraksi dengan database, bersama dengan antarmuka pengguna grafis yang dikembangkan sebagai fasad atas bahasa kueri terstruktur. Sintaks bahasa kueri dan kebutuhan untuk memahami skema database tidak sesuai bagi pakar domain untuk berinteraksi dan menjelajahi data. Membuat kueri terhadap basis data dan ruang data terstruktur yang semakin kompleks akan memerlukan pendekatan berbeda yang sesuai untuk tugas berbeda dan tingkat keahlian berbeda (Franklin dkk. 2005). Pendekatan baru untuk berinteraksi dengan data terstruktur

telah berkembang dari tahap penelitian awal dan dapat memberikan dasar bagi rangkaian alat baru yang dapat memfasilitasi interaksi antara pengguna dan data. Contohnya adalah pencarian kata kunci, antarmuka kueri visual, dan antarmuka kueri bahasa alami melalui database (Franklin et al. 2005; Freitas et al. 2012a, b; Kaufmann dan Bernstein 2007). Pendekatan fleksibel untuk kueri basis data bergantung pada kemampuan pendekatan untuk menafsirkan maksud kueri pengguna, mencocokkannya dengan elemen dalam basis data. Pendekatan ini pada akhirnya bergantung pada penciptaan model semantik yang mendukung pendekatan semantik (Freitas et al. 2011). Meskipun sudah melampaui tahap pembuktian konsep, fungsi dan pendekatan ini belum bermigrasi ke aplikasi tingkat komersial.

Kepercayaan

Pengelolaan Asal Seiring meningkatnya penggunaan kembali data, konsumen data pihak ketiga perlu memiliki mekanisme untuk memverifikasi kepercayaan dan kualitas data. Beberapa atribut kualitas data dapat dilihat dari data itu sendiri, sementara atribut lainnya bergantung pada pemahaman tentang konteks yang lebih luas di balik data tersebut, misalnya asal data, proses, artefak, dan aktor di balik pembuatan data. Menangkap dan mewakili konteks di mana data dihasilkan dan diubah serta membuatnya tersedia bagi konsumen data merupakan persyaratan utama kurasi data untuk kumpulan data yang ditargetkan untuk konsumen pihak ketiga. Standar asal seperti W3C PROV9 memberikan landasan bagi representasi data yang dapat dioperasikan. Namun, aplikasi kurasi data masih perlu diinstrumentasikan untuk mengetahui asal usulnya. Asalnya dapat digunakan untuk secara eksplisit menangkap dan mewakili keputusan kurasi yang dibuat (Wawancara Kurasi Data: Paul Groth 2014). Namun, penerapan penangkapan dan pengelolaan asal dalam aplikasi data masih relatif rendah. Selain itu, mengevaluasi kepercayaan dan kualitas data asal secara manual dapat menjadi proses yang memakan waktu. Representasi asal data perlu dilengkapi dengan pendekatan otomatis untuk memperoleh kepercayaan dan menilai kualitas data dari metadata asal, dalam konteks penerapan tertentu.

Model dan Alat Manajemen Izin yang Terperinci Mengizinkan sekelompok besar pengguna untuk berkolaborasi memerlukan penciptaan izin/hak yang terperinci terkait dengan peran kurasi. Sebagian besar sistem saat ini memiliki sistem izin yang sangat rumit, di mana pengurus sistem mengawasi kontributor umum. Meskipun mekanisme ini dapat sepenuhnya memenuhi kebutuhan beberapa proyek, terdapat permintaan yang jelas untuk sistem izin yang lebih terperinci, di mana izin dapat ditentukan pada tingkat item data (Qin dan Atluri 2003; Ryutov dkk. 2009) dan dapat ditentukan ditugaskan secara terdistribusi. Untuk mendukung pengendalian yang menyeluruh ini, penyelidikan dan pengembangan metode otomatis untuk inferensi dan penyebaran izin (Kirrane dkk. 2013), serta mekanisme penetapan izin terdistribusi yang mudah dilakukan, merupakan hal yang sangat penting. Secara analogi, metode serupa dapat diterapkan pada kontrol menyeluruh atas hak-hak digital.

Standardisasi dan Interoperabilitas

Model Data Standar dan Kosakata untuk Penggunaan Kembali Data Sebagian besar upaya kurasi data terdiri dari pengintegrasian dan penggunaan kembali data yang dibuat

dalam konteks berbeda. Dalam banyak kasus, integrasi ini dapat melibatkan ratusan sumber data. Standar model data seperti *Resource Description Framework* (RDF)10 memfasilitasi integrasi data pada tingkat model data. Penggunaan Pengidentifikasi Sumber Daya Universal (URI) dalam identifikasi entitas data berfungsi sebagai kunci asing terbuka berskala web, yang mendorong penggunaan kembali pengidentifikasi di kumpulan data yang berbeda, sehingga memfasilitasi proses integrasi data terdistribusi.

Penciptaan terminologi dan kosakata merupakan langkah metodologis penting dalam proyek kurasi data. Proyek seperti Indeks New York Times (NYT) (Curry et al. 2010) atau *Protein Data Bank* (PDB) (Bernstein et al. 1977) memprioritaskan penciptaan dan evolusi kosakata yang dapat berfungsi untuk mewakili dan memberi anotasi pada data domain. Dalam kasus PDB, kosakatanya mengungkapkan kebutuhan keterwakilan suatu komunitas. Penggunaan kosakata bersama adalah bagian dari visi web data yang terhubung (Berners-Lee 2009) dan merupakan salah satu alat metodologis yang dapat digunakan untuk memfasilitasi interoperabilitas semantik. Meskipun penciptaan kosakata lebih terkait dengan dimensi metodologis, pencarian semantik, pemetaan skema, atau pendekatan penyesuaianan ontologi (Shvaiko dan Euzenat 2005; Freitas et al. 2012a, b) sangat penting untuk mengurangi beban pemetaan kosakata manual pada sisi pengguna akhir, mengurangi beban penggunaan kembali terminologis (Freitas et al. 2012a, b).

Peningkatan Integrasi dan Komunikasi antar Alat Kurasi Data dibuat dan dikurasi dalam konteks berbeda dan menggunakan alat berbeda (yang dikhususkan untuk memenuhi kebutuhan kurasi data berbeda). Misalnya, pengguna dapat menganalisis kemungkinan ketidakkonsistenan data dengan alat visualisasi, melakukan pemetaan skema dengan alat lain, lalu memperbaiki data menggunakan platform crowdsourcing. Kemampuan untuk memindahkan data secara lancar antar alat yang berbeda dan menangkap keputusan kurasi pengguna serta transformasi data di berbagai platform merupakan hal mendasar untuk mendukung operasi kurasi data yang lebih canggih yang mungkin memerlukan alat yang sangat terspesialisasi agar hasil akhirnya dapat dipercaya (Wawancara Kurasi Data: Paul Groth 2014; Wawancara Kurasi Data: James Cheney 2014). Pembuatan model data dan kosakata yang terstandarisasi (seperti W3C PROV) dapat mengatasi sebagian permasalahan tersebut. Namun, aplikasi kurasi data perlu diadaptasi untuk menangkap dan mengelola sumber data dan untuk memberikan adopsi yang lebih baik dibandingkan standar yang ada.

Model Kurasi Data

Model Informasi Minimum untuk Kurasi Data Meskipun ada upaya baru-baru ini dalam mengenali dan memahami bidang kurasi data, proses di balik kurasi data masih perlu diformalkan dengan lebih baik. Penerapan metode seperti model informasi minimum (La Novere et al. 2005) dan perwujudannya dalam bentuk alat merupakan salah satu contoh perbaikan metodologi yang dapat memberikan standar kualitas minimum bagi kurator data. Dalam eScience, MIRIAM (informasi minimum yang diperlukan dalam anotasi model) (Laibe dan Le Nove`re 2007) adalah contoh upaya tingkat komunitas untuk menstandardisasi proses anotasi dan kurasi model kuantitatif sistem biologis.

Mengkurasi Publikasi Nano, Mengatasi Long Tail of Science Dengan meningkatnya jumlah komunikasi ilmiah, semakin sulit untuk menemukan, menghubungkan, dan menyusun pernyataan ilmiah. Publikasi nano adalah pernyataan ilmiah inti dengan konteks terkait, yang bertujuan menyediakan mekanisme sintetik untuk komunikasi ilmiah. Publikasi nano masih merupakan paradigma baru yang dapat memberikan jalan bagi penciptaan data semi-terstruktur yang terdistribusi baik dalam domain ilmiah maupun non-ilmiah.

Investigasi Prinsip Teoritis dan Model Spesifik Domain Model kurasi data harus berkembang dari praktik lapangan menjadi deskripsi yang lebih abstrak. Kemajuan algoritme kurasi data otomatis akan bergantung pada definisi model teoretis dan penyelidikan prinsip di balik kurasi data. Memahami mekanisme sebab akibat di balik alur kerja dan kondisi generalisasi di balik kemampuan pengangkutan data adalah contoh model teoretis yang dapat memengaruhi kurasi data, memandu pengguna menuju pembuatan dan representasi data yang dapat digunakan kembali dalam konteks yang lebih luas.

Integrasi Data Tidak Terstruktur dan Terstruktur

Pengenalan dan Tautan Entitas Sebagian besar informasi di web dan organisasi tersedia sebagai data tidak terstruktur (teks, video, dll.). Proses menjadikan informasi tersedia sebagai data tidak terstruktur memakan waktu: berbeda dengan data terstruktur, data tidak terstruktur tidak dapat dibandingkan, dikumpulkan, dan dioperasikan secara langsung. Pada saat yang sama, data tidak terstruktur menyimpan sebagian besar informasi dari variasi data ekor panjang (Gambar 6.2).

Mengekstraksi informasi terstruktur dari data tidak terstruktur adalah langkah mendasar untuk membuat data jangka panjang dapat dianalisis dan diinterpretasikan. Sebagian dari masalah ini dapat diatasi dengan pendekatan ekstraksi informasi (misalnya ekstraksi relasi, pengenalan entitas, dan ekstraksi ontologi). Alat-alat ini mengekstrak informasi dari teks dan dapat digunakan untuk secara otomatis membangun pengetahuan semi-terstruktur dari teks. Terdapat kerangka kerja ekstraksi informasi yang sudah matang untuk mengatasi permasalahan ekstraksi informasi tertentu, namun penerapannya masih terbatas pada pengguna awal.

Penggunaan Data Terbuka untuk Mengintegrasikan Data Terstruktur dan Tidak Terstruktur Pergeseran terbaru lainnya dalam bidang ini adalah ketersediaan sumber daya data terstruktur berskala besar, khususnya data terbuka, yang mendukung ekstraksi informasi. Misalnya, entitas dalam kumpulan data terbuka seperti DBpedia (Auer et al. 2007) dan Freebase (Bollacker et al. 2008) dapat digunakan untuk mengidentifikasi entitas bernama (orang, tempat, dan organisasi) dalam teks, yang dapat digunakan untuk mengkategorikan dan mengatur isi teks. Data terbuka dalam skenario ini berfungsi sebagai basis pengetahuan yang masuk akal bagi entitas dan dapat diperluas dengan entitas khusus domain di dalam lingkungan organisasi. Alat pengenalan dan penautan entitas bernama seperti DBpedia Spotlight (Mendes et al. 2011) dapat digunakan untuk menghubungkan data terstruktur dan tidak terstruktur.

Sebagai pelengkap, data tidak terstruktur dapat digunakan untuk memberikan deskripsi yang lebih komprehensif untuk data terstruktur, meningkatkan aksesibilitas konten

dan semantik. Model semantik distribusi, model semantik yang dibangun dari koleksi berskala besar (Freitas et al. 2012a, b), dapat diterapkan pada database terstruktur (Freitas dan Curry 2014) dan merupakan contoh pendekatan yang dapat digunakan untuk memperkaya semantik dari data.

Saluran Pemrosesan Bahasa Alami Komunitas Pemrosesan Bahasa Alami (NLP) memiliki pendekatan dan alat yang matang yang dapat langsung diterapkan pada proyek yang menangani data tidak terstruktur. Proyek sumber terbuka seperti Apache UIMA memfasilitasi integrasi fungsi NLP ke sistem lain. Selain itu, kasus penggunaan industri yang kuat seperti IBM Watson (Ferrucci et al. 2010), Thomson Reuters, The New York Times (Curry et al. 2010), dan Press Association (Wawancara Kurasi Data: Hellen Lippell) mengubah pola pikir persepsi teknik NLP dari bidang akademik hingga bidang industri.

6.7 STUDI KASUS SEKTOR UNTUK KURASI BIG DATA

Pada bagian ini, studi kasus dibahas yang mencakup berbagai proses kurasi data pada domain berbeda. Tujuan di balik studi kasus ini adalah untuk menangkap berbagai alur kerja yang telah diadopsi atau dirancang untuk menangani kurasi data dalam konteks Big Data.

Ilmu Kesehatan dan Kehidupan

► **Laba-laba Kimia**

ChemSpider adalah mesin pencari yang menyediakan akses gratis ke komunitas kimia yang berpusat pada struktur. Ini telah dirancang untuk mengumpulkan dan mengindeks struktur kimia dan informasi terkaitnya ke dalam satu repositori yang dapat dicari. ChemSpider berisi puluhan juta senyawa kimia dengan data terkait dan berfungsi sebagai penyedia data untuk situs web dan perangkat lunak. Tersedia sejak tahun 2007, ChemSpider telah mengumpulkan lebih dari 300 sumber data dari vendor bahan kimia, database pemerintah, laboratorium swasta, dan individu. Digunakan oleh ahli kimia untuk konversi dan prediksi pengidentifikasi, kumpulan data ChemSpider juga banyak dimanfaatkan oleh vendor bahan kimia dan perusahaan farmasi sebagai sumber daya pra-kompetitif untuk investigasi uji coba eksperimental dan klinis.

Kurasi data di ChemSpider terdiri dari anotasi manual dan koreksi data. Hal ini dapat mencakup perubahan pada struktur kimia suatu senyawa, penambahan atau penghapusan pengenal, mengaitkan hubungan antara senyawa kimia, sumber data terkait, dll. ChemSpider mendukung dua cara berbeda bagi kurator untuk membantu mengatur data di ChemSpider:

- Posting komentar pada catatan untuk menyoroti perlunya tindakan yang tepat oleh kurator utama.
- Sebagai anggota terdaftar yang mempunyai hak kurasi, kurasi data secara langsung atau hapus data yang salah.

ChemSpider mengadopsi model meritokratis untuk aktivitas kurasi mereka. Kurator biasa bertanggung jawab atas deposisi, yang diperiksa dan diverifikasi oleh kurator master. Kurator normal pada gilirannya dapat diundang untuk menjadi master setelah beberapa periode kontribusi yang memenuhi syarat. Platform ini memiliki proses kurasi berbasis manusia dan

komputer yang dipadukan. Kurasi robotik menggunakan algoritma untuk koreksi kesalahan dan validasi data pada waktu deposisi.

ChemSpider menggunakan campuran pendekatan komputasi untuk melakukan validasi data pada tingkat tertentu. Mereka telah membangun alat validasi data kimia mereka sendiri, yang disebut CVSP (platform validasi dan standardisasi kimia). CVSP membantu ahli kimia memeriksa bahan kimia untuk menentukan apakah bahan tersebut terwakili secara valid atau tidak, atau apakah ada masalah kualitas data sehingga mereka dapat menandai masalah kualitas tersebut dengan mudah dan efisien.

Dengan menggunakan model komunitas terbuka, ChemSpider mendistribusikan aktivitas kurasi ke seluruh komunitas menggunakan crowdsourcing untuk mengakomodasi tingkat pertumbuhan besar-besaran dan masalah kualitas. Mereka menggunakan pendekatan mirip wiki bagi orang-orang untuk berinteraksi dengan data, sehingga mereka dapat membuat anotasi, memvalidasi, mengkurasi, menandai, dan menghapusnya. ChemSpider sedang dalam proses menerapkan sistem pengenalan otomatis yang akan mengukur upaya kontribusi kurator melalui validasi data dan proses keterlibatan. Metrik kontribusi dapat dilihat secara publik dan dapat diakses melalui profil pusat untuk kurator data.

Bank Data Protein

Kolaborasi Penelitian untuk Bank Data Protein Bioinformatika Struktural (RCSB PDB) adalah kelompok yang didedikasikan untuk meningkatkan pemahaman fungsi sistem biologis melalui studi struktur 3D makromolekul biologis. PDB telah diunduh lebih dari 300 juta kumpulan data.

Sebagian besar proses kurasi di PDB terdiri dari penyediaan kosakata standar untuk menggambarkan hubungan antar entitas biologis, mulai dari jaringan organ hingga deskripsi struktur molekul. Penggunaan kosakata standar membantu tata nama yang digunakan untuk mendeskripsikan nama protein dan molekul kecil serta deskriptornya yang ada dalam entri struktur. Proses kurasi data meliputi identifikasi dan koreksi ketidakkonsistenan struktur protein 3D dan data eksperimen. Untuk menerapkan pendekatan tata kelola hierarki global pada alur kerja kurasi data, staf PDB meninjau dan memberi anotasi pada setiap entri yang dikirimkan sebelum kurasi robotik memeriksa masuk akal sebagai bagian dari penyimpanan, pemrosesan, dan distribusi data. Upaya kurasi data didistribusikan ke seluruh situs saudaranya.

Kurasi robot mengotomatiskan validasi dan verifikasi data. Kurator manusia berkontribusi pada definisi aturan untuk mendeteksi inkonsistensi. Proses kurasi juga dilakukan secara retrospektif, dimana kesalahan yang ditemukan pada data dikoreksi secara retrospektif pada arsip. Versi kumpulan data terkini dirilis setiap minggu untuk menjaga semua sumber tetap konsisten dengan standar saat ini dan untuk memastikan kualitas kurasi data yang baik.

Lipat

Foldit (Good dan Su 2011) adalah contoh populer komputasi manusia yang diterapkan pada masalah kompleks, misalnya menemukan pola pelipatan protein. Pengembang Foldit telah menggunakan gamifikasi untuk memungkinkan komputasi manusia. Melalui permainan

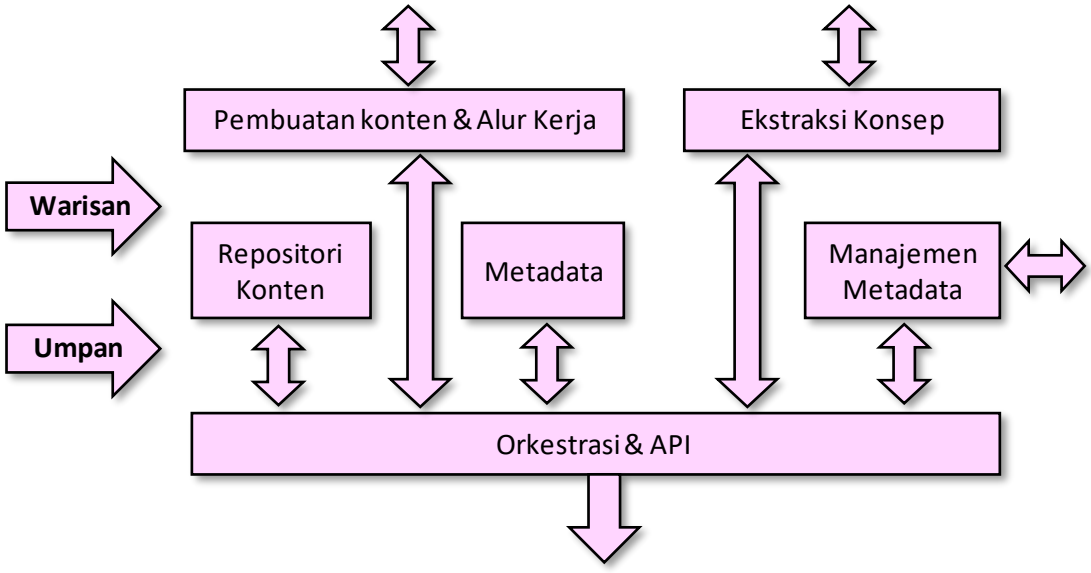
ini orang dapat memprediksi struktur protein yang mungkin membantu dalam menargetkan obat pada penyakit tertentu. Algoritme komputer saat ini tidak mampu menangani kemungkinan struktur protein dalam jumlah besar secara eksponensial. Untuk mengatasi masalah ini, Foldit menggunakan pelipatan protein kompetitif untuk menghasilkan protein terbaik.

Media dan Hiburan

- **Asosiasi Pers**

Press Association (PA) adalah kantor berita nasional untuk Inggris dan Irlandia dan penyedia konten multimedia terkemuka di web, seluler, penyiaran, dan media cetak. Selama 145 tahun terakhir, PA telah menyediakan feed (teks, data, foto, dan video) kepada media besar Inggris serta pelanggan korporat dan sektor publik.

Tujuan kurasi data di Press Association adalah untuk memilih informasi yang paling relevan bagi pelanggannya, mengklasifikasikan, memperkaya, dan mendistribusikannya dengan cara yang mudah dikonsumsi. Proses kurasi di Press Association mempekerjakan sejumlah besar kurator dalam proses klasifikasi konten, bekerja pada sejumlah besar sumber data. Kurator di dalam PA adalah seorang analis yang mengumpulkan, menggabungkan, mengklasifikasikan, menormalkan, dan menganalisis informasi mentah yang berasal dari berbagai sumber data. Karena sifat informasi yang dianalisis biasanya bervolume tinggi dan hampir real-time, kurasi data merupakan tantangan besar dalam perusahaan dan penggunaan alat otomatis memainkan peran penting dalam proses ini. Dalam proses kurasi, alat otomatis menyediakan triase dan klasifikasi tingkat pertama, yang selanjutnya disempurnakan dengan intervensi kurator manusia seperti yang ditunjukkan pada Gambar 6.3.



Gambar 6.3 Alur Kerja Pola Konten Dan Metadata Asosiasi Pers

Proses kurasi data dimulai dengan artikel yang dikirimkan ke platform yang menggunakan seperangkat aturan ekstraksi linguistik atas teks tidak terstruktur untuk secara otomatis mendapatkan tag untuk artikel tersebut, memperkayanya dengan data terstruktur yang dapat dibaca mesin. Kurator data kemudian memilih istilah yang lebih menggambarkan

konten dan menyisipkan tag baru jika perlu. Tag memperkaya teks asli dengan kategori umum dari konten yang dianalisis, sekaligus memberikan deskripsi entitas tertentu (tempat, orang, peristiwa, fakta) yang ada dalam teks. Manajer metadata kemudian meninjau klasifikasi dan konten dipublikasikan secara online.

The New York Times

The New York Times (NYT) adalah surat kabar metropolitan terbesar dan surat kabar terbesar ketiga di Amerika Serikat. Perusahaan ini memiliki sejarah panjang dalam kurasi artikelnya di repositori kurasi berusia 100 tahun (Indeks NYT).

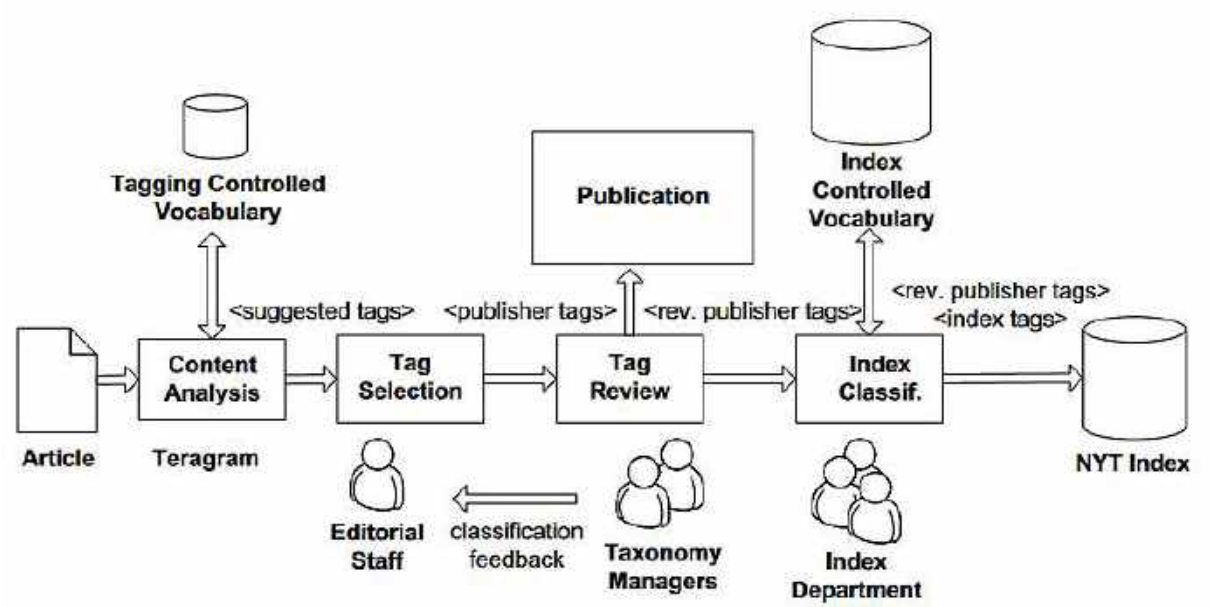
Jalur kurasi The New York Times (lihat Gambar 6.4) dimulai dengan sebuah artikel yang keluar dari ruang redaksi. Kurasi tingkat pertama terdiri dari proses klasifikasi konten yang dilakukan oleh staf redaksi yang terdiri dari beberapa ratus jurnalis. Dengan menggunakan aplikasi web, anggota staf editorial mengirimkan artikel baru melalui sistem ekstraksi informasi berbasis aturan (dalam hal ini, SAS Teragram). Teragram menggunakan seperangkat aturan ekstraksi linguistik, yang dibuat oleh pengelola taksonomi berdasarkan subset kosakata terkontrol yang digunakan oleh Departemen Indeks. Teragram menyarankan tag berdasarkan indeks kosakata yang berpotensi menggambarkan isi artikel. Anggota staf editorial kemudian memilih istilah yang lebih menggambarkan konten dan menyisipkan tag baru jika perlu.

Manajer taksonomi meninjau klasifikasi dan konten dipublikasikan secara online, memberikan umpan balik berkelanjutan ke dalam proses klasifikasi. Pada tahap selanjutnya, artikel menerima kurasi tingkat kedua oleh departemen indeks, yang menambahkan tag tambahan dan ringkasan artikel ke sumber daya yang disimpan.

Ritel

- **eBay**

eBay adalah salah satu pasar online paling populer yang melayani jutaan produk dan pelanggan. eBay telah menggunakan komputasi manusia untuk memecahkan dua masalah penting kualitas data: mengelola taksonomi produk dan menemukan pengidentifikasi dalam deskripsi produk. Pekerja crowdsourced membantu eBay dalam meningkatkan kecepatan dan kualitas algoritma klasifikasi produk dengan biaya lebih rendah.



Gambar 6.4 Alur Kerja Kurasi Klasifikasi Artikel New York Times

Unilever

Unilever adalah salah satu produsen barang konsumen terbesar di dunia, yang beroperasi secara global. Unilever memanfaatkan perhitungan manusia yang dilakukan secara crowdsourced dalam strategi pemasaran mereka untuk produk-produk baru. Komputasi manusia digunakan untuk mengumpulkan data yang memadai tentang umpan balik pelanggan dan menganalisis sentimen publik terhadap media sosial. Awalnya Unilever mengembangkan serangkaian algoritma pembelajaran mesin untuk melakukan analisis sentimen pelanggan di seluruh rangkaian produk mereka. Namun, algoritma analisis sentimen ini tidak mampu memperhitungkan perbedaan regional dan budaya di antara populasi sasaran. Oleh karena itu, Unilever secara efektif meningkatkan keakuratan algoritma analisis sentimen dengan crowdsourcing, dengan memverifikasi algoritma output dan mengumpulkan umpan balik dari platform crowdsourcing online, yaitu Crowdflower.

Kesimpulan

Dengan pertumbuhan jumlah sumber data dan pembuatan konten yang terdesentralisasi, memastikan kualitas data menjadi isu mendasar bagi lingkungan manajemen data di era Big Data. Evolusi metode dan alat kurasi data merupakan elemen penting untuk memastikan kualitas data pada skala Big Data.

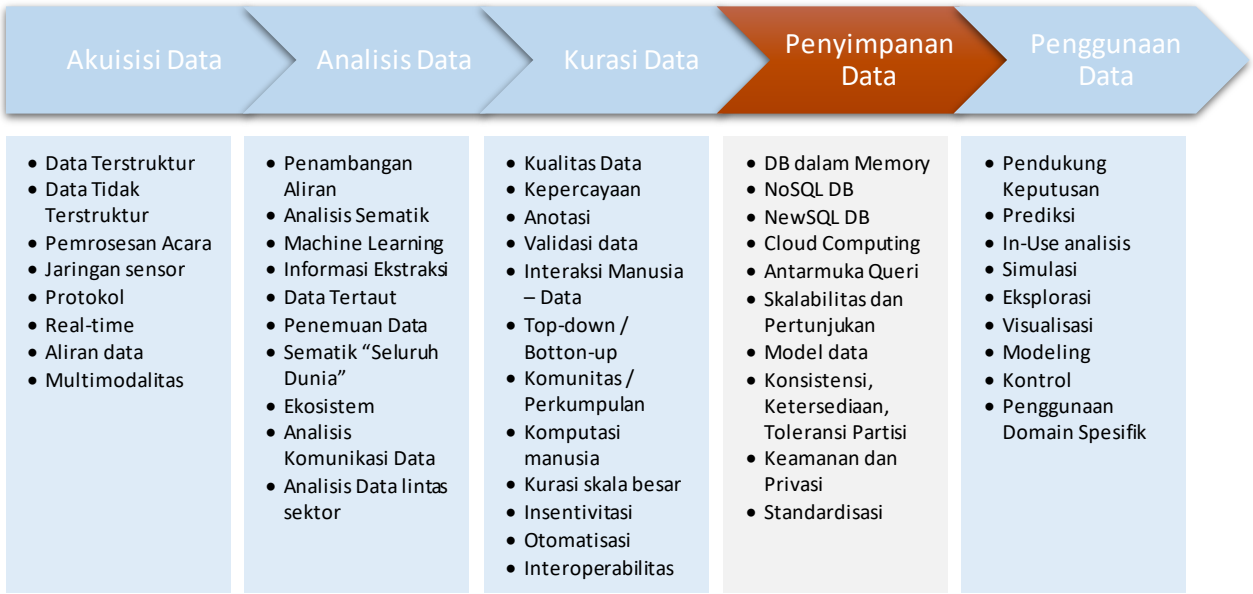
Berdasarkan bukti yang dikumpulkan melalui penyelidikan ekstensif yang mencakup analisis literatur komprehensif, survei, wawancara dengan pakar kurasi data, kuesioner, dan studi kasus, persyaratan masa depan dan tren kurasi data yang muncul telah diidentifikasi. Analisis ini dapat memberikan kepada kurator data, manajer teknis, dan peneliti pandangan terkini tentang tantangan, pendekatan, dan peluang kurasi data di era Big Data.

BAB 7
PENYIMPANAN BIG DATA

7.1 PENDAHULUAN

Bab ini memberikan gambaran umum tentang teknologi penyimpanan data besar yang berfungsi sebagai masukan terhadap pembuatan peta jalan lintas sektoral untuk pengembangan teknologi data besar di berbagai domain aplikasi berdampak tinggi. Daripada menguraikan masing-masing teknologi secara konkrit, bab ini memberikan gambaran umum tentang teknologi penyimpanan data sehingga pembaca dapat memperoleh pemahaman tingkat tinggi tentang kemampuan masing-masing teknologi dan bidang yang memerlukan penelitian lebih lanjut. Oleh karena itu, dampak sosial dan ekonomi dijelaskan, dan studi kasus terpilih yang menggambarkan penggunaan teknologi penyimpanan data besar juga disediakan. Hasil lengkap analisis penyimpanan data besar dapat ditemukan di Curry dkk. (2014).

Posisi penyimpanan Big Data dalam keseluruhan Jaringan nilai Big Data dapat dilihat pada Gambar 7.1. Penyimpanan data besar berkaitan dengan penyimpanan dan pengelolaan data dengan cara yang dapat diskalakan, memenuhi kebutuhan aplikasi yang memerlukan akses ke data. Sistem penyimpanan data besar yang ideal akan memungkinkan penyimpanan data dalam jumlah yang hampir tidak terbatas, mengatasi tingginya tingkat akses tulis dan baca acak, secara fleksibel dan efisien menangani berbagai model data yang berbeda, mendukung data terstruktur dan tidak terstruktur, dan untuk alasan privasi, hanya berfungsi pada data terenkripsi.



Gambar 7.1 Penyimpanan Data Dalam Jaringan Nilai Big Data

Tentu saja semua kebutuhan tersebut tidak dapat dipenuhi sepenuhnya. Namun dalam beberapa tahun terakhir, banyak sistem penyimpanan baru bermunculan yang setidaknya dapat mengatasi sebagian tantangan ini. Bab ini memberikan gambaran umum tentang teknologi penyimpanan data besar dan mengidentifikasi beberapa area yang memerlukan penelitian lebih lanjut. Teknologi penyimpanan data besar disebut sebagai teknologi penyimpanan yang secara khusus mengatasi tantangan volume, kecepatan, atau variasi dan tidak termasuk dalam kategori sistem basis data relasional. Ini tidak berarti bahwa sistem basis data relasional tidak mengatasi tantangan-tantangan ini, namun teknologi penyimpanan alternatif seperti penyimpanan kolom dan kombinasi cerdas dari sistem penyimpanan yang berbeda, misalnya. menggunakan Hadoop Distributed File System (HDFS), seringkali lebih efisien dan lebih murah.

Sistem penyimpanan data besar biasanya mengatasi tantangan volume dengan memanfaatkan arsitektur terdistribusi dan tidak berbagi apa pun. Hal ini memungkinkan mengatasi peningkatan kebutuhan penyimpanan dengan memperluas skala ke node baru yang menyediakan daya komputasi dan penyimpanan. Mesin baru dapat dengan mudah ditambahkan ke kluster penyimpanan dan sistem penyimpanan akan menangani pendistribusian data antar node secara transparan.

Solusi penyimpanan juga perlu mengatasi kecepatan dan variasi data. Kecepatan penting dalam kaitannya dengan latensi kueri, yaitu berapa lama waktu yang dibutuhkan untuk mendapatkan balasan atas suatu kueri? Hal ini sangat penting mengingat tingginya tingkat data yang masuk. Misalnya, akses tulis acak ke database dapat sangat memperlambat kinerja kueri jika diperlukan untuk memberikan jaminan transaksional. Sebaliknya, keragaman berkaitan dengan tingkat upaya yang diperlukan untuk mengintegrasikan dan bekerja dengan data yang berasal dari sejumlah besar sumber berbeda. Misalnya, database grafik adalah sistem penyimpanan yang cocok untuk mengatasi tantangan ini.

7.2 WAWASAN UTAMA UNTUK PENYIMPANAN BIG DATA

Sebagai hasil dari analisis teknologi penyimpanan data saat ini dan masa depan, sejumlah wawasan diperoleh terkait dengan teknologi penyimpanan data. Jelas terlihat bahwa penyimpanan data besar telah menjadi bisnis komoditas dan teknologi penyimpanan yang dapat diskalakan telah mencapai tingkat tingkat perusahaan yang dapat mengelola volume data yang hampir tidak terbatas. Buktinya adalah meluasnya penggunaan solusi berbasis Hadoop yang ditawarkan oleh vendor seperti Cloudera (2014a), Hortonworks (2014), dan MapR (2014) serta berbagai vendor database NoSQL, khususnya yang menggunakan in-memory dan kolumnar. teknologi penyimpanan. Dibandingkan dengan sistem manajemen basis data relasional tradisional yang mengandalkan penyimpanan berbasis baris dan strategi caching yang mahal, teknologi penyimpanan data besar yang baru ini menawarkan skalabilitas yang lebih baik dengan kompleksitas dan biaya operasional yang lebih rendah.

Meskipun terdapat kemajuan yang meningkatkan kinerja, skalabilitas, dan kegunaan teknologi penyimpanan, masih terdapat potensi signifikan yang belum dimanfaatkan dalam

teknologi penyimpanan data besar, baik untuk penggunaan maupun pengembangan lebih lanjut teknologi tersebut:

1. **Potensi untuk Mentransformasi Masyarakat dan Dunia Usaha di Seluruh Sektor:** Teknologi penyimpanan data besar merupakan faktor kunci dalam analisis tingkat lanjut yang memiliki potensi untuk mentransformasi masyarakat dan cara pengambilan keputusan bisnis yang penting. Hal ini sangat penting khususnya di sektor-sektor yang biasanya tidak berbasis TI seperti energi. Meskipun sektor-sektor ini menghadapi permasalahan non-teknis seperti kurangnya ahli Big Data yang terampil dan hambatan peraturan, teknologi penyimpanan data baru memiliki potensi untuk memungkinkan analisis baru yang menghasilkan nilai di dalam dan di berbagai sektor industri.
2. **Kurangnya Standar Merupakan Hambatan Utama:** Sejarah NoSQL didasarkan pada penyelesaian tantangan teknologi spesifik yang mengarah pada beragam teknologi penyimpanan yang berbeda. Banyaknya pilihan ditambah dengan kurangnya standar untuk menanyakan data membuat pertukaran penyimpanan data menjadi lebih sulit karena dapat mengikat kode spesifik aplikasi ke solusi penyimpanan tertentu.
3. **Tantangan Skalabilitas Terbuka dalam Penyimpanan Data Berbasis Grafik**: Memproses data berdasarkan struktur data grafik bermanfaat dalam semakin banyak aplikasi. Hal ini memungkinkan penangkapan semantik dan hubungan kompleks yang lebih baik dengan informasi lain yang berasal dari berbagai macam sumber data yang berbeda, dan berpotensi meningkatkan nilai keseluruhan yang dapat dihasilkan dengan menganalisis data. Meskipun database grafik semakin banyak digunakan untuk tujuan ini, masih sulit untuk mendistribusikan struktur data berbasis grafik secara efisien ke seluruh node komputasi.
4. **Privasi dan Keamanan Tertinggal**: Meskipun ada beberapa proyek dan solusi yang menangani privasi dan keamanan, perlindungan individu dan pengamanan data mereka tertinggal dibandingkan kemajuan teknologi sistem penyimpanan data. Penelitian yang cukup diperlukan untuk lebih memahami bagaimana data dapat disalahgunakan, bagaimana data tersebut perlu dilindungi dan diintegrasikan dalam solusi penyimpanan data besar.

Teknologi Penyimpanan Data

Selama dekade terakhir, kebutuhan untuk menghadapi ledakan data (Turner et al. 2014) dan peralihan perangkat keras dari pendekatan scale-up ke scale-out menyebabkan ledakan sistem penyimpanan data besar baru yang beralih dari basis data relasional tradisional. model. Pendekatan ini biasanya mengorbankan properti seperti konsistensi data untuk mempertahankan respons kueri yang cepat dengan meningkatnya jumlah data. Penyimpanan data besar digunakan dengan cara yang mirip dengan sistem manajemen basis data relasional tradisional, misalnya. untuk solusi pemrosesan transaksional online (OLTP) dan gudang data atas data terstruktur atau semi-terstruktur. Kekuatan khususnya adalah dalam menangani data tidak terstruktur dan semi terstruktur dalam skala besar.

Bagian ini menilai teknologi penyimpanan data terkini yang mampu menangani data dalam jumlah besar, dan mengidentifikasi tren terkait penyimpanan data. Berikut adalah berbagai jenis sistem penyimpanan:

- **Sistem File Terdistribusi**: Sistem file seperti Hadoop File System (HDFS) (Shvachko dkk. 2010) menawarkan kemampuan untuk menyimpan data tidak terstruktur dalam jumlah besar dengan cara yang andal pada perangkat keras komoditas. Meskipun ada sistem file dengan kinerja lebih baik, HDFS merupakan bagian integral dari kerangka Hadoop (White 2012) dan telah mencapai tingkat standar de-facto. Ini telah dirancang untuk file data besar dan sangat cocok untuk menyerap data dengan cepat dan pemrosesan massal.
- **Basis Data NoSQL**: Mungkin kelompok teknologi penyimpanan data besar yang paling penting adalah sistem manajemen basis data NoSQL. Basis data NoSQL menggunakan model data dari luar dunia relasional yang tidak selalu mematuhi sifat transaksional atomisitas, konsistensi, isolasi, dan daya tahan (ACID).
- **Basis Data NewSQL**: Bentuk modern dari basis data relasional yang bertujuan untuk skalabilitas yang sebanding dengan basis data NoSQL dengan tetap menjaga jaminan transaksional yang dibuat oleh sistem basis data tradisional.
- **Platform Kueri Big Data**: Teknologi yang menyediakan fasad kueri di depan penyimpanan data besar seperti sistem file terdistribusi atau database NoSQL. Perhatian utama adalah menyediakan antarmuka tingkat tinggi, misalnya. melalui SQL seperti bahasa kueri dan mencapai latensi kueri rendah.

Basis Data NoSQL

Basis data NoSQL dirancang untuk skalabilitas, seringkali dengan mengorbankan konsistensi. Dibandingkan dengan database relasional, database ini sering kali menggunakan antarmuka kueri non-standar tingkat rendah, sehingga lebih sulit untuk diintegrasikan ke dalam aplikasi yang sudah ada yang memerlukan antarmuka SQL. Kurangnya antarmuka standar membuat lebih sulit untuk berpindah vendor. Database NoSQL dapat dibedakan berdasarkan model data yang digunakannya.

- **Penyimpanan Nilai Kunci**: Penyimpanan nilai kunci memungkinkan penyimpanan data tanpa skema. Objek data bisa sepenuhnya tidak terstruktur atau terstruktur dan diakses dengan satu kunci. Karena tidak ada skema yang digunakan, objek data bahkan tidak perlu berbagi struktur yang sama.
- **Penyimpanan Kolom**: Menurut Wikipedia "DBMS berorientasi kolom adalah sistem manajemen basis data (DBMS) yang menyimpan tabel data sebagai bagian dari kolom data, bukan sebagai baris data, seperti kebanyakan DBMS relasional" (Wikipedia 2013). Basis data seperti ini biasanya berupa peta terurut multi-dimensi yang jarang, terdistribusi, dan persisten di mana data diindeks dengan tiga kunci baris, kunci kolom, dan stempel waktu. Nilai direpresentasikan sebagai tipe data string tidak terputus. Data diakses oleh kelompok kolom, yaitu sekumpulan kunci kolom terkait yang secara efektif memampatkan data yang jarang di kolom. Keluarga kolom dibuat sebelum data dapat disimpan dan jumlahnya diharapkan sedikit. Sebaliknya, jumlah kolom tidak

terbatas. Pada prinsipnya penyimpanan kolom kurang cocok bila semua kolom perlu diakses. Namun dalam praktiknya hal ini jarang terjadi, sehingga menghasilkan kinerja penyimpanan kolom yang unggul.

- **Basis Data Dokumen**: Berbeda dengan nilai dalam penyimpanan nilai kunci, dokumen terstruktur. Namun, tidak ada persyaratan untuk skema umum yang harus dipatuhi oleh semua dokumen seperti halnya catatan dalam database relasional. Jadi database dokumen disebut sebagai penyimpanan data semi-terstruktur. Mirip dengan penyimpanan nilai kunci, dokumen dapat dikueri menggunakan kunci unik. Namun, dimungkinkan untuk mengakses dokumen dengan menanyakan struktur internalnya, seperti meminta semua dokumen yang berisi bidang dengan nilai tertentu. Kemampuan antarmuka kueri biasanya bergantung pada format pengkodean yang digunakan oleh database. Pengkodean umum mencakup XML atau JSON.
- **Basis Data Grafik**: Basis data grafik, seperti Neo4J (2015), menyimpan data dalam struktur grafik sehingga cocok untuk menyimpan data yang sangat asosiatif seperti grafik jaringan sosial. Jenis database grafik tertentu adalah penyimpanan rangkap tiga seperti AllegroGraph (Franz 2015) dan Virtuoso (Erling 2009) yang dirancang khusus untuk menyimpan rangkap tiga RDF. Namun, teknologi triple store yang ada belum cocok untuk menyimpan kumpulan data yang sangat besar secara efisien.

Meskipun secara umum penyimpanan data NoSQL memiliki skala yang lebih baik daripada database relasional, skalabilitas menurun seiring dengan meningkatnya kompleksitas model data yang digunakan oleh penyimpanan data. Hal ini terutama berlaku untuk database grafik yang mendukung aplikasi yang intensif menulis dan membaca. Salah satu pendekatan untuk mengoptimalkan akses baca adalah dengan mempartisi grafik menjadi sub-grafik yang terhubung secara minimal satu sama lain dan mendistribusikan sub-grafik ini di antara node komputasi. Namun, ketika sisi-sisi baru ditambahkan ke suatu grafik, konektivitas antar sub-grafik dapat meningkat secara signifikan. Hal ini dapat menyebabkan latensi kueri yang lebih tinggi karena peningkatan lalu lintas jaringan dan komputasi non-lokal. Oleh karena itu, skema sharding yang efisien harus mempertimbangkan dengan cermat biaya overhead yang diperlukan untuk mendistribusikan ulang data grafik secara dinamis.

Basis Data NewSQL

Basis data NewSQL adalah bentuk modern dari basis data relasional yang bertujuan untuk skalabilitas yang sebanding dengan basis data NoSQL sambil mempertahankan jaminan transaksional yang dibuat oleh sistem basis data tradisional. Menurut Venkatesh dan Nirmala (2012) mereka memiliki ciri-ciri sebagai berikut:

- SQL adalah mekanisme utama untuk interaksi aplikasi
- Dukungan ACID untuk transaksi
- Mekanisme kontrol konkurensi non-locking
- Arsitektur yang memberikan kinerja per node yang jauh lebih tinggi
- Arsitektur yang diperluas dan tidak digunakan bersama, mampu berjalan pada sejumlah besar node tanpa mengalami hambatan

Harapannya adalah sistem NewSQL sekitar 50 kali lebih cepat dibandingkan OLTP RDBMS tradisional. Misalnya, VoltDB menskalakan secara linier dalam kasus kueri non-kompleks (partisi tunggal) dan menyediakan dukungan ACID. Ini menskalakan lusinan node di mana setiap node dibatasi oleh ukuran memori utama.

Platform Kueri Big Data

Platform kueri data besar menyediakan fasad kueri di atas penyimpanan data besar yang mendasarinya sehingga menyederhanakan kueri penyimpanan data yang mendasarinya. Mereka biasanya menawarkan antarmuka kueri mirip SQL untuk mengakses data, namun berbeda dalam pendekatan dan kinerjanya. Hive (Thusoo et al. 2009) menyediakan abstraksi di atas Hadoop Distributed File System (HDFS) yang memungkinkan file terstruktur untuk dikueri dengan bahasa kueri mirip SQL. Hive mengeksekusi kueri dengan menerjemahkan kueri dalam pekerjaan MapReduce. Akibatnya, kueri Hive memiliki latensi tinggi bahkan untuk kumpulan data kecil. Manfaat Hive mencakup antarmuka kueri mirip SQL dan fleksibilitas untuk mengembangkan skema dengan mudah. Hal ini dimungkinkan karena skema disimpan secara independen dari data dan data hanya divalidasi pada waktu kueri. Pendekatan ini disebut sebagai skema-on-baca dibandingkan dengan pendekatan skema-on-tulis pada database SQL. Oleh karena itu, mengubah skema merupakan operasi yang relatif murah. Toko kolumnar Hadoop HBase juga didukung oleh Hive.

Berbeda dengan Hive, Impala (Russel 2013) dirancang untuk mengeksekusi kueri dengan latensi rendah. Ia menggunakan kembali metadata dan antarmuka pengguna mirip SQL yang sama seperti Hive tetapi menggunakan mesin kueri terdistribusinya sendiri yang dapat mencapai latensi lebih rendah. Ini juga mendukung HDFS dan HBase sebagai penyimpanan data dasar.

Spark SQL (Shenker dkk. 2013) adalah fasad kueri latensi rendah lainnya yang mendukung antarmuka Hive. Proyek ini mengklaim bahwa "dapat mengeksekusi kueri Hive QL hingga 100 kali lebih cepat daripada Hive tanpa modifikasi apa pun pada data atau kueri yang ada" (Shenker dkk. 2013). Hal ini dicapai dengan mengeksekusi kueri menggunakan kerangka Spark (Zaharia et al. 2010) daripada kerangka MapReduce Hadoop.

Terakhir, Drill adalah implementasi open source dari Google Dremel (Melnik et al. 2002) yang mirip dengan Impala yang dirancang sebagai sistem kueri ad-hoc interaktif dan terukur untuk data bertumpuk. Drill menyediakan bahasa kueri mirip SQL, DrQL, yang kompatibel dengan Dremel, namun dirancang untuk mendukung bahasa kueri lain seperti Bahasa Kueri Mongo. Berbeda dengan Hive dan Impala, ini mendukung berbagai sumber data tanpa skema, seperti database HDFS, HBase, Cassandra, MongoDB, dan SQL.

Penyimpanan Awan

Seiring dengan semakin populernya komputasi awan, pengaruhnya terhadap data besar juga semakin meningkat. Sementara Amazon, Microsoft, dan Google membangun platform cloud mereka sendiri, perusahaan lain termasuk IBM, HP, Dell, Cisco, Rackspace, dll., membangun proposal mereka berdasarkan OpenStack, sebuah platform sumber terbuka untuk membangun sistem cloud (OpenStack 2014). Menurut IDC (Grady 2013), pada tahun

2020 40 % dunia digital "akan 'tersentuh' oleh komputasi awan", dan "mungkin sebanyak 15 % akan dikelola di awan".

Cloud secara umum, dan khususnya penyimpanan cloud, dapat digunakan oleh perusahaan dan pengguna akhir. Bagi pengguna akhir, menyimpan data mereka di cloud memungkinkan akses dari mana saja dan dari setiap perangkat dengan cara yang andal. Selain itu, pengguna akhir dapat menggunakan penyimpanan cloud sebagai solusi sederhana untuk pencadangan online data desktop mereka. Demikian pula bagi perusahaan, penyimpanan cloud menyediakan akses fleksibel dari berbagai lokasi dan skala kapasitas yang cepat dan mudah (Grady 2013) serta harga penyimpanan yang lebih murah dan dukungan yang lebih baik berdasarkan skala ekonomi (CloudDrive 2013) dengan efektivitas biaya yang sangat tinggi dalam suatu lingkungan di mana kebutuhan penyimpanan perusahaan berubah seiring berjalannya waktu.

Secara teknis solusi penyimpanan cloud dapat dibedakan antara penyimpanan objek dan penyimpanan blok. Penyimpanan objek "adalah istilah umum yang menggambarkan pendekatan untuk menangani dan memanipulasi unit penyimpanan terpisah yang disebut objek" (Margaret Rouse 2014a). Sebaliknya, data penyimpanan blok disimpan dalam volume yang juga disebut sebagai blok. Menurut Margaret Rouse (2014b), "setiap blok bertindak sebagai hard drive individual" dan memungkinkan akses acak ke bit-bit data sehingga berfungsi baik dengan aplikasi seperti database.

Selain penyimpanan objek dan blok, platform utama menyediakan dukungan untuk penyimpanan berbasis database relasional dan non-relasional serta penyimpanan dalam memori dan penyimpanan antrian. Dalam penyimpanan cloud, terdapat perbedaan signifikan yang perlu dipertimbangkan dalam fase perencanaan aplikasi:

- Karena *penyimpanan cloud* adalah sebuah layanan, aplikasi yang menggunakan penyimpanan ini memiliki kontrol yang lebih kecil dan mungkin mengalami penurunan kinerja akibat jaringan. Perbedaan kinerja ini perlu diperhitungkan pada tahap desain dan implementasi.
- *Keamanan* adalah salah satu perhatian utama terkait cloud publik. Hasilnya, CTO Amazon memperkirakan bahwa dalam lima tahun semua data di cloud akan dienkripsi secara default (Vogels 2013).
- Cloud kaya fitur seperti AWS mendukung kalibrasi latensi, redundansi, dan tingkat throughput untuk akses data, sehingga memungkinkan pengguna menemukan trade-off yang tepat antara biaya dan kualitas.

Masalah penting lainnya ketika mempertimbangkan penyimpanan cloud adalah model konsistensi yang didukung (dan skalabilitas, ketersediaan, toleransi partisi, dan latensi terkait). Meskipun Simple Storage Service (S3) Amazon mendukung konsistensi akhir, penyimpanan blob Microsoft Azure mendukung konsistensi yang kuat dan pada saat yang sama ketersediaan tinggi serta toleransi partisi. Microsoft menggunakan dua lapisan: (1) lapisan aliran "yang memberikan ketersediaan tinggi dalam menghadapi partisi jaringan dan kegagalan lainnya", dan (2) lapisan partisi yang "memberikan jaminan konsistensi yang kuat".

Privasi dan Keamanan

Privasi dan keamanan merupakan tantangan umum dalam Big Data. Kelompok Kerja Big Data CSA menerbitkan daftar 10 Tantangan Keamanan dan Privasi Big Data Teratas (Mora dkk. 2012). Berikut adalah lima tantangan yang sangat penting bagi penyimpanan data besar.

Praktik Terbaik Keamanan untuk Penyimpanan Data Non-relasional

Ancaman keamanan terhadap database NoSQL mirip dengan RDBMS tradisional dan oleh karena itu praktik terbaik yang sama harus diterapkan (Winder 2012). Namun, banyak langkah keamanan yang diterapkan secara default dalam RDBMS tradisional tidak ada dalam database NoSQL (Okman et al. 2011). Langkah-langkah tersebut akan mencakup enkripsi data sensitif, proses sandboxing, validasi input, dan otentikasi pengguna yang kuat. Beberapa pemasok NoSQL merekomendasikan penggunaan database di lingkungan tepercaya tanpa tindakan keamanan atau otentikasi tambahan. Namun, pendekatan ini tidak masuk akal ketika memindahkan penyimpanan data besar ke cloud.

Keamanan database NoSQL semakin mendapat perhatian dari para peneliti keamanan dan peretas, dan keamanan akan semakin meningkat seiring dengan semakin matangnya pasar. Misalnya, terdapat inisiatif untuk menyediakan kemampuan kontrol akses untuk database NoSQL berdasarkan modul otentikasi Kerberos (Winder 2012).

Penyimpanan Data dan Log Transaksi yang Aman

Tantangan keamanan khusus untuk penyimpanan data muncul karena distribusi data. Dengan auto-tiering, operator menyerahkan kendali penyimpanan data kepada algoritma untuk mengurangi biaya. Keberadaan data, pergerakan tingkatan, dan perubahan harus diperhitungkan dalam log transaksi.

Strategi auto-tiering harus dirancang secara hati-hati untuk mencegah data sensitif dipindahkan ke tingkat yang kurang aman sehingga lebih murah; mekanisme pemantauan dan pencatatan harus ada untuk mendapatkan gambaran yang jelas tentang penyimpanan data dan pergerakan data dalam solusi auto-tiering (Mora dkk. 2012).

Skema enkripsi ulang proxy (Blaze et al. 2006) dapat diterapkan pada penyimpanan multi-tier dan berbagi data untuk memastikan kerahasiaan dan keaslian (Shucheng et al. 2010). Namun, kinerja harus ditingkatkan untuk aplikasi data besar. Log transaksi untuk sistem operasi multi-tingkat masih hilang.

Kontrol Akses yang Diberlakukan Secara Kriptografis dan Komunikasi yang Aman

Saat ini, data sering kali disimpan tidak terenkripsi, dan kontrol akses hanya bergantung pada penegakan seperti gerbang. Namun, data hanya boleh diakses oleh entitas yang berwenang dengan jaminan kriptografi—demikian pula dalam penyimpanan dan transmisi. Untuk tujuan ini, diperlukan mekanisme kriptografi baru yang menyediakan fungsionalitas yang diperlukan dengan cara yang efisien dan terukur.

Meskipun penyedia penyimpanan cloud mulai menawarkan enkripsi, materi kunci kriptografi harus dibuat dan disimpan di klien dan tidak pernah diserahkan ke penyedia cloud. Beberapa produk menambahkan fungsi ini ke lapisan aplikasi penyimpanan data besar, misalnya zNcrypt, Protegrity Big Data Protection untuk Hadoop, dan Distribusi Intel untuk Apache Hadoop (sekarang bagian dari Cloudera). Enkripsi berbasis atribut adalah teknologi

yang menjanjikan untuk mengintegrasikan kriptografi dengan kontrol akses untuk penyimpanan data besar.

Tantangan Keamanan dan Privasi untuk Kontrol Akses Granular

Keberagaman data merupakan tantangan besar karena persyaratan keamanan yang sama beragamnya, misalnya pembatasan hukum, kebijakan privasi, dan kebijakan perusahaan lainnya. Mekanisme kontrol akses yang menyeluruh diperlukan untuk memastikan kepatuhan terhadap persyaratan ini. Komponen Big Data utama menggunakan Kerberos (Miller et al. 1987) bersamaan dengan autentikasi berbasis token, dan Daftar Kontrol Akses (ACL) berdasarkan pengguna dan pekerjaan. Namun, mekanisme yang lebih terperinci, misalnya Kontrol Akses Berbasis Atribut (ABAC) dan Bahasa Markup Kontrol Akses eXtensible (XACLM), diperlukan untuk memodelkan keragaman asal data dan analitik yang sangat beragam. penggunaan.

Asal Data

Integritas dan riwayat objek data dalam Jaringan nilai sangatlah penting. Asal tradisional sebagian besar mengatur kepemilikan dan penggunaan. Namun dengan Big Data, kompleksitas metadata asal akan meningkat (Glavic 2014).

Upaya awal telah dilakukan untuk mengintegrasikan sumber ke dalam ekosistem Big Data (Ikeda dkk. 2011; Sherif dkk. 2013); namun, sumber yang aman memerlukan jaminan integritas dan kerahasiaan data asal dalam semua bentuk penyimpanan data besar dan masih merupakan tantangan terbuka. Selain itu, analisis grafik asal yang sangat besar memerlukan komputasi yang intensif dan memerlukan algoritma yang cepat.

Tantangan Privasi dalam Penyimpanan Big Data

Para peneliti telah menunjukkan (Acquisti dan Gross 2009) bahwa analisis Big Data dari informasi yang tersedia untuk umum dapat dimanfaatkan untuk menebak nomor jaminan sosial seseorang. Beberapa produk secara selektif mengenkripsi bidang data untuk menciptakan anonimitas yang dapat dibalik, bergantung pada hak akses.

Menganonimkan dan menghapus identifikasi data mungkin tidak cukup karena banyaknya data yang memungkinkan dilakukannya identifikasi ulang. Sebuah diskusi meja bundar (Bollier dan Firestone 2010) menganjurkan transparansi dalam penanganan data dan algoritma serta kesepakatan baru mengenai data besar untuk memberdayakan pengguna akhir sebagai pemilik data. Kedua opsi tersebut tidak hanya melibatkan transparansi organisasi, namun juga perangkat teknis seperti Security & Privacy by Design dan hasil proyek EEXCESS EU FP7.

7.3 PENYIMPANAN BIG DATA DI MASA DEPAN

Bagian ini memberikan gambaran umum tentang kebutuhan masa depan dan tren yang muncul.

Persyaratan Masa Depan untuk Penyimpanan Big Data

Tiga bidang utama telah diidentifikasi yang diharapkan dapat mengatur teknologi penyimpanan data besar di masa depan. Hal ini mencakup standarisasi antarmuka kueri,

peningkatan dukungan terhadap keamanan data, perlindungan privasi pengguna, dan dukungan model data semantik.

Antarmuka Kueri Standar

Dalam jangka menengah hingga jangka panjang, database NoSQL akan mendapatkan keuntungan besar dari antarmuka kueri terstandarisasi, mirip dengan SQL untuk sistem relasional. Saat ini tidak ada standar untuk masing-masing jenis penyimpanan NoSQL selain API standar de-facto untuk database grafik (Blueprints 2014) dan bahasa manipulasi data SPARQL (Aranda et al. 2013) yang didukung oleh vendor triplestore. Basis data NoSQL lainnya biasanya menyediakan bahasa deklaratif atau API mereka sendiri, dan standarisasi untuk bahasa deklaratif ini tidak ada.

Meskipun untuk beberapa kategori database (kunci/nilai, dokumen, dll.) standarisasi bahasa deklaratif masih belum ada, terdapat upaya untuk mendiskusikan kebutuhan standardisasi. Misalnya, Kelompok Studi ISO/IEC JTC mengenai data besar baru-baru ini merekomendasikan agar komite standar ISO/IEC yang ada harus menyelidiki lebih lanjut "definisi antarmuka standar untuk mendukung penyimpanan data non-relasional" (Lee dkk. 2014).

Definisi antarmuka standar akan memungkinkan pembuatan lapisan virtualisasi data yang akan memberikan abstraksi sistem penyimpanan data heterogen seperti yang biasa digunakan dalam kasus penggunaan data besar. Beberapa persyaratan lapisan virtualisasi data telah dibahas secara online di artikel blog Infoworld (Kobielus 2013).

Keamanan dan Privasi

Wawancara dilakukan dengan konsultan dan pengguna akhir penyimpanan data besar yang memiliki tanggung jawab atas keamanan dan privasi, untuk mendapatkan pandangan dan wawasan pribadi. Berdasarkan wawancara ini dan kesenjangan yang diidentifikasi, beberapa persyaratan masa depan untuk keamanan dan privasi dalam penyimpanan data besar diidentifikasi.

Data Commons dan Norma Sosial Data yang disimpan dalam jumlah besar akan dibagikan serta karya turunannya untuk memaksimalkan manfaat Big Data. Saat ini, pengguna tidak menyadari bagaimana Big Data memproses datanya (transparansi), dan tidak jelas bagaimana pengguna Big Data dapat berbagi dan memperoleh data secara efisien. Selain itu, batasan hukum terkait privasi dan hak cipta dalam Big Data saat ini belum sepenuhnya jelas di UE. Misalnya, data besar memungkinkan analisis baru berdasarkan data gabungan dari berbagai sumber. Bagaimana pendekatan ini mempengaruhi informasi pribadi? Bagaimana peraturan dan regulasi untuk remix dan karya turunannya dapat diterapkan pada Big Data? Ketidakpastian seperti ini dapat menyebabkan kerugian bagi UE dibandingkan Amerika.

Privasi Data Penyimpanan data besar harus mematuhi peraturan privasi UE seperti Directive 95/46/EC ketika informasi pribadi disimpan. Saat ini, penerapan arahan ini yang heterogen membuat penyimpanan informasi pribadi dalam data besar menjadi sulit. Peraturan Perlindungan Data Umum (GDRP) yang pertama kali diusulkan pada tahun 2012—merupakan upaya berkelanjutan untuk menyelaraskan perlindungan data di antara negara-negara anggota UE. GDRP diperkirakan akan mempengaruhi kebutuhan penyimpanan data

besar di masa depan. Pada tahun 2014, GDRP harus menjalani negosiasi yang menyulitkan perkiraan aturan akhir dan awal penegakan hukum. Misalnya, versi draf tahun 2013 memungkinkan subjek data (orang) meminta pengontrol data untuk menghapus data pribadi, yang sering kali tidak dipertimbangkan secara memadai oleh solusi penyimpanan data besar.

Penelusuran dan Asal Data Penelusuran dan asal data menjadi semakin penting dalam penyimpanan data besar karena dua alasan: (1) pengguna ingin memahami dari mana data berasal, apakah data tersebut benar dan dapat dipercaya, serta apa yang terjadi pada hasilnya dan (2) penyimpanan data besar akan tunduk pada aturan kepatuhan karena data besar memasuki proses bisnis dan Jaringan nilai yang penting. Oleh karena itu, penyimpanan data besar harus mempertahankan metadata asal, menyediakan asal di sepanjang Jaringan pemrosesan data, dan menawarkan cara yang mudah digunakan untuk memahami dan melacak penggunaan data.

Sandboxing dan Virtualisasi Sandboxing dan virtualisasi analitik data besar menjadi lebih penting selain kontrol akses. Menurut skala ekonomi, analisis Big Data mendapat manfaat dari pembagian sumber daya. Namun, pelanggaran keamanan pada komponen analitik bersama menyebabkan kunci akses kriptografi dan akses penyimpanan penuh disusupi. Oleh karena itu, pekerjaan dalam analisis data besar harus dimasukkan ke dalam kotak pasir (sandbox) untuk mencegah peningkatan pelanggaran keamanan dan oleh karena itu akses tidak sah ke penyimpanan.

Model Data Semantik

Banyaknya sumber data yang heterogen meningkatkan biaya pengembangan, karena aplikasi memerlukan pengetahuan tentang format data individual dari masing-masing sumber. Tren yang muncul adalah web semantik dan khususnya web sensor semantik yang mencoba mengatasi tantangan ini. Banyak proyek penelitian yang berkaitan dengan semua tingkat pemodelan dan komputasi semantik. Sebagaimana dirinci dalam buku ini, kebutuhan akan anotasi semantik misalnya telah diidentifikasi untuk sektor kesehatan. Oleh karena itu, persyaratan penyimpanan data adalah untuk mendukung penyimpanan dan pengelolaan model data semantik dalam skala besar. Khususnya trade-off antara ekspresivitas dan penyimpanan serta kueri yang efisien perlu dieksplorasi lebih lanjut.

Paradigma yang Muncul untuk Penyimpanan Big Data

Ada beberapa paradigma baru yang muncul untuk penyimpanan kumpulan data yang besar dan kompleks. Paradigma baru ini mencakup, antara lain, peningkatan penggunaan database NoSQL, konvergensi dengan kerangka analitik, dan pengelolaan data di pusat data pusat.

Peningkatan Penggunaan Database NoSQL

Basis data NoSQL, terutama basis data grafik dan penyimpanan kolom, semakin banyak digunakan sebagai pengganti atau pelengkap sistem relasional yang sudah ada. Misalnya, kebutuhan untuk menggunakan model data semantik dan menghubungkan data dengan banyak sumber data dan informasi yang berbeda sangat mendorong kebutuhan untuk dapat menyimpan dan menganalisis data dalam jumlah besar menggunakan model berbasis grafik. Namun, hal ini memerlukan mengatasi keterbatasan sistem berbasis grafik saat ini seperti

dijelaskan di atas. Misalnya, Jim Webber menyatakan "Teknologi grafik akan menjadi sangat penting" (Webber 2013). Dalam wawancara lainnya, Ricardo Baeza-Yates, VP Riset Eropa dan Amerika Latin di Yahoo!, juga menyatakan pentingnya menangani data grafik berskala besar (Baeza-Yates 2013). Proyek penelitian Microsoft Trinity mencapai terobosan signifikan dalam bidang ini. Trinity adalah penyimpanan data dalam memori dan platform pemrosesan terdistribusi. Dengan mengembangkan kemampuan traversal grafik yang sangat cepat, peneliti Microsoft memperkenalkan pendekatan baru untuk mengatasi kueri grafik. Proyek lainnya termasuk grafik pengetahuan Google dan pencarian grafik Facebook yang menunjukkan meningkatnya relevansi dan semakin matangnya teknologi grafik.

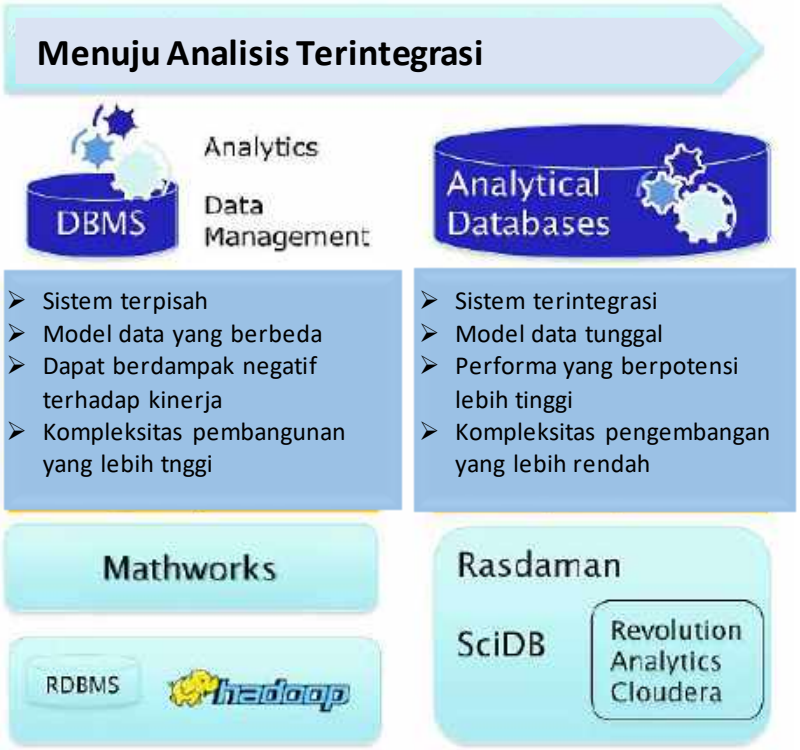
Desain Dalam Memori dan Berorientasi Kolom

Banyak database NoSQL modern berkinerja tinggi didasarkan pada desain kolom. Keuntungan utamanya adalah dalam sebagian besar aplikasi praktis hanya diperlukan beberapa kolom untuk mengakses data. Akibatnya menyimpan data dalam kolom memungkinkan akses lebih cepat. Selain itu, database berorientasi kolom sering kali tidak mendukung operasi gabungan yang mahal dari dunia relasional. Sebaliknya, pendekatan umum adalah dengan menggunakan tabel kolom lebar tunggal yang menyimpan data berdasarkan skema yang sepenuhnya didenormalisasi.

Menurut Michael Stonebraker *"Semua vendor SQL akan pindah ke penyimpanan kolom, karena mereka jauh lebih cepat daripada penyimpanan baris"* (Stonebraker 2012a). Basis data dalam memori berkinerja tinggi seperti SAP HANA biasanya menggabungkan teknik dalam memori dengan desain berbasis kolom. Berbeda dengan sistem relasional yang menyimpan data dalam memori, database dalam memori dapat menggunakan teknik seperti anti-caching (DeBrabant et al. 2013). Harizopoulos dkk. telah menunjukkan bahwa sebagian besar waktu untuk mengeksekusi query dihabiskan untuk tugas-tugas administratif seperti manajemen buffer dan penguncian (Harizopoulos et al. 2008).

Konvergensi dengan Kerangka Analytics

Selama pelaksanaan proyek, banyak skenario telah diidentifikasi yang memerlukan analisis yang lebih baik terhadap data yang tersedia untuk meningkatkan operasi di berbagai sektor. Secara teknis, hal ini berarti meningkatnya kebutuhan akan analisis kompleks yang melampaui agregasi dan statistik sederhana. Stonebraker menunjukkan bahwa kebutuhan akan analisis yang kompleks akan sangat berdampak pada solusi penyimpanan data yang ada (Stonebraker 2012b). Karena analisis spesifik kasus penggunaan adalah salah satu komponen paling penting yang menciptakan nilai bisnis aktual, maka menjadi semakin penting untuk meningkatkan analisis yang memenuhi persyaratan kinerja, namun juga mengurangi kompleksitas dan biaya pengembangan secara keseluruhan. Gambar 7.2 menunjukkan beberapa perbedaan antara penggunaan sistem terpisah untuk pengelolaan data dan analisis dibandingkan dengan database analitik terintegrasi.



Gambar 7.2 Pergeseran paradigma dari sistem penyimpanan data murni ke database analitik terintegrasi

Pusat Data

Pusat data pusat yang mengintegrasikan semua data dalam suatu perusahaan adalah sebuah paradigma yang mempertimbangkan pengelolaan semua data perusahaan secara keseluruhan, bukan dalam database berbeda dan terisolasi yang dikelola oleh unit organisasi berbeda. Manfaat dari data hub terpusat adalah data dapat dianalisis secara keseluruhan, menghubungkan berbagai dataset yang dimiliki perusahaan sehingga menghasilkan wawasan yang lebih mendalam.

Implementasi teknis umumnya didasarkan pada sistem berbasis Hadoop yang mungkin menggunakan HDFS atau HBase (Apache 2014) untuk menyimpan kumpulan data master terintegrasi. Di satu sisi, kumpulan data master ini dapat digunakan sebagai kebenaran dasar dan cadangan untuk sistem manajemen data yang ada, namun juga memberikan dasar untuk analisis tingkat lanjut yang menggabungkan kumpulan data yang sebelumnya terisolasi.

Perusahaan seperti Cloudera menggunakan paradigma ini untuk memasarkan distribusi Hadoop mereka (Cloudera 2014b). Banyak kasus penggunaan hub data perusahaan yang sudah ada. Studi kasus di sektor keuangan dijelaskan pada bagian berikutnya.

7.4 STUDI KASUS UNTUK PENYIMPANAN BIG DATA

Pada bagian ini dijelaskan tiga kasus penggunaan terpilih yang menggambarkan potensi dan kebutuhan teknologi penyimpanan di masa depan. Kasus penggunaan kesehatan menggambarkan bagaimana analisis berbasis media sosial diaktifkan oleh teknologi penyimpanan NoSQL. Kasus penggunaan kedua dari sektor keuangan menggambarkan

munculnya paradigma pusat data terpusat. Kasus penggunaan terakhir dari sektor energi menggambarkan manfaat pengelolaan data Internet of Things (IoT) yang terperinci untuk analisis tingkat lanjut. Ikhtisar karakteristik utama dari use case dapat ditemukan pada Tabel 7.1. Studi kasus lebih lanjut disajikan dalam Curry dkk. (2014).

Tabel 7.1 Karakteristik utama dari studi kasus penyimpanan data besar yang dipilih

STUDI KASUS	SEKTOR	VOLUME	TEKNOLOGI PENYIMPANAN	PERSYARATAN UTAMA
Treato: Kecerdasan pengobatan berbasis media sosial	Kesehatan	>150 TBC	HBase	Efisiensi biaya, batasan skalabilitas DB relasional
Pusat data terpusat	Keuangan	Antara beberapa petabyte dan lebih dari 150 PB	Hadoop/HDFS	Membangun model yang lebih akurat, skala data, kesesuaian untuk data tidak terstruktur
Jaringan pintar	Energi	Puluhan TBC per hari	Hadoop	Volume data, tantangan operasional

Sektor Kesehatan: Kecerdasan Pengobatan Berbasis Media Sosial

Treato adalah perusahaan Israel yang mengkhususkan diri dalam menambang konten buatan pengguna dari blog dan forum untuk memberikan layanan intelijen merek kepada perusahaan farmasi. Saat Treato menganalisis web sosial, Treato termasuk dalam kategori "klasik" dalam menganalisis data tidak terstruktur dalam jumlah besar, sebuah area aplikasi yang sering kali meminta solusi penyimpanan data besar. Layanan Treato sebagai kasus penggunaan menunjukkan nilai penggunaan teknologi penyimpanan data besar. Informasi tersebut didasarkan pada studi kasus yang diterbitkan oleh Cloudera (2012), perusahaan penyedia distribusi Hadoop yang telah digunakan Treato.

Saat membuat prototipenya, Treato menemukan "bahwa efek samping dapat diidentifikasi melalui media sosial jauh sebelum perusahaan farmasi atau Badan Pengawas Obat dan Makanan (FDA) mengeluarkan peringatan mengenai hal tersebut. Misalnya, ketika melihat diskusi tentang Singulair, obat asma, Treato menemukan bahwa hampir separuh UGC membahas gangguan mental; efek sampingnya akan dapat diidentifikasi empat tahun sebelum peringatan resmi dikeluarkan." (Cloudera 2012). Treato awalnya menghadapi dua tantangan besar: Pertama, mereka perlu mengembangkan kemampuan analitis untuk menganalisis bahasa sehari-hari pasien dan memetakannya ke dalam terminologi medis yang sesuai untuk memberikan wawasan kepada pelanggannya. Kedua, penting untuk menganalisis sumber data dalam jumlah besar secepat mungkin memberikan informasi akurat secara real time.

Tantangan pertama, mengembangkan analitik, pada awalnya telah diatasi dengan sistem non-Hadoop yang berbasis pada database relasional. Dengan sistem tersebut, Treato menghadapi keterbatasan yaitu hanya dapat menangani "pengumpulan data dari lusinan situs web dan hanya dapat memproses beberapa juta postingan per hari" (Cloudera 2012). Oleh

karena itu, Treato mencari platform analitik hemat biaya yang dapat memenuhi persyaratan utama berikut:

1. Penyimpanan yang andal dan terukur
2. Infrastruktur pemrosesan yang andal dan terukur
3. Kemampuan mesin pencari untuk mengambil postingan dengan ketersediaan tinggi
4. Penyimpanan real-time yang dapat diskalakan untuk mengambil statistik dengan ketersediaan tinggi

Hasilnya Treato memutuskan sistem berbasis Hadoop yang menggunakan HBase untuk menyimpan daftar URL yang akan diambil. Postingan yang tersedia di URL ini dianalisis dengan menggunakan pemrosesan bahasa alami bersama dengan ontologi kepemilikannya. Selain itu "setiap postingan diindeks, statistik dihitung, dan tabel HBase diperbarui" (Cloudera 2012).

Menurut laporan studi kasus, solusi berbasis Hadoop menyimpan lebih dari 150 TB data termasuk 1,1 miliar postingan online dari ribuan situs web termasuk lebih dari 11.000 obat dan lebih dari 13.000 kondisi. Treato mampu memproses 150–200 juta postingan pengguna per hari.

Bagi Treato, dampak dari infrastruktur penyimpanan dan pemrosesan berbasis Hadoop adalah mereka memperoleh sistem yang dapat diskalakan, andal, dan hemat biaya yang bahkan dapat menciptakan wawasan yang tidak mungkin diperoleh tanpa infrastruktur ini. Studi kasus mengklaim bahwa dengan Hadoop, Treato meningkatkan waktu eksekusi setidaknya enam kali lipat. Hal ini memungkinkan Treato menanggapi permintaan pelanggan tentang pengobatan baru dalam satu hari.

Sektor Keuangan: Pusat Data Terpusat

Sebagaimana dipetakan dalam uraian peta jalan sektoral, sektor keuangan menghadapi tantangan sehubungan dengan peningkatan volume data dan beragamnya sumber data baru seperti media sosial. Di sini kasus penggunaan untuk sektor keuangan dijelaskan berdasarkan ringkasan solusi Cloudera (Cloudera 2013).

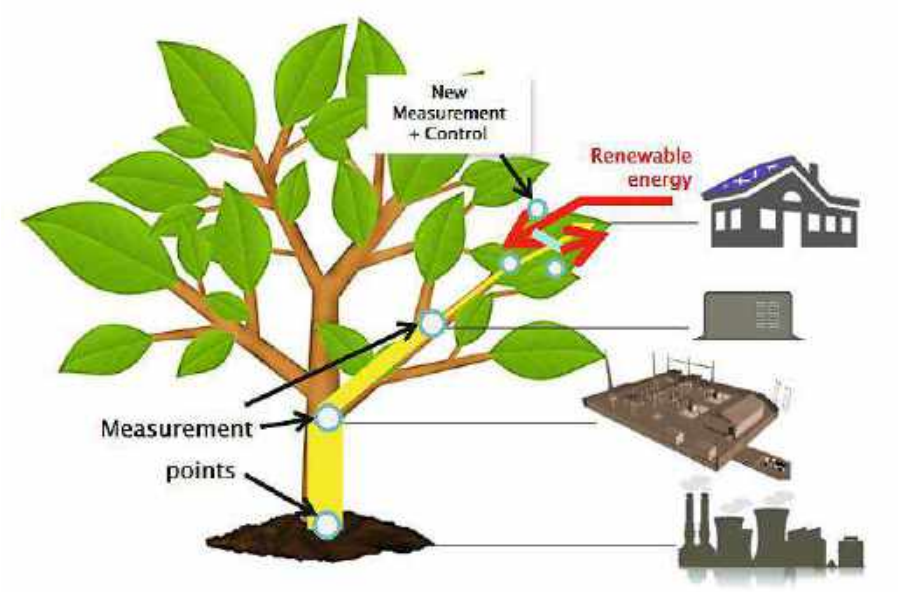
Produk keuangan semakin terdigitalisasi termasuk perbankan dan perdagangan online. Karena akses online dan seluler menyederhanakan akses terhadap produk keuangan, terdapat peningkatan tingkat aktivitas yang menghasilkan lebih banyak data. Potensi Big Data dalam skenario ini adalah menggunakan seluruh data yang tersedia untuk membangun model akurat yang dapat membantu sektor keuangan mengelola risiko keuangan dengan lebih baik. Menurut ringkasan solusi, perusahaan memiliki akses ke beberapa petabyte data. Menurut Larry Feinsmith, direktur pelaksana JPMorgan Chase, perusahaannya menyimpan lebih dari 150 petabyte secara online dan menggunakan Hadoop untuk mendeteksi penipuan (Cloudera 2013).

Kedua, sumber data baru menambah volume dan variasi data yang tersedia. Secara khusus, data tidak terstruktur dari weblog, media sosial, blog, dan feed berita lainnya dapat membantu dalam manajemen hubungan pelanggan, manajemen risiko, dan bahkan mungkin perdagangan algoritmik. Menggabungkan semua data dalam pusat data terpusat memungkinkan analisis lebih rinci yang dapat memberikan keunggulan kompetitif. Namun sistem tradisional tidak dapat mengimbangi skala, biaya, dan integrasi rumit dari proses

ekstraksi, transformasi, pemuatan (ETL) tradisional yang menggunakan skema data tetap, dan juga tidak mampu menangani data tidak terstruktur. Namun sistem penyimpanan data besar berskala sangat baik dan dapat memproses data terstruktur dan tidak terstruktur.

Energi: Pengukuran Tingkat Perangkat

Di sektor energi, pengelolaan smart grid dan smart meter merupakan bidang yang menjanjikan manfaat ekonomi dan lingkungan yang tinggi. Seperti digambarkan pada Gambar 7.3, penggunaan energi terbarukan seperti sistem fotovoltaik yang diterapkan pada rumah-rumah dapat menyebabkan ketidakstabilan jaringan listrik. Saat ini operator jaringan listrik hanya mempunyai sedikit pengetahuan tentang last mile kepada konsumen energi. Oleh karena itu, mereka tidak mampu bereaksi secara tepat terhadap ketidakstabilan yang terjadi di bagian paling ujung jaringan jaringan listrik. Dengan menganalisis sampel data meter pintar pada interval kedua, perkiraan kebutuhan energi jangka pendek dan pengelolaan permintaan perangkat seperti pemanas dan mobil listrik menjadi mungkin dilakukan, sehingga menstabilkan jaringan listrik. Jika diterapkan pada jutaan rumah tangga, volume data dapat mencapai skala petabyte, sehingga mendapatkan manfaat besar dari teknologi penyimpanan baru. Tabel 7.2 menunjukkan volume data hanya untuk data mentah yang dikumpulkan selama satu hari.



Gambar 7.3 Pengenalan energi terbarukan di lokasi konsumen mengubah topologi jaringan energi dan memerlukan titik pengukuran baru di bagian luar jaringan

Tabel 7.2 Perhitungan jumlah data yang diambil sampelnya dengan smart meter

Tingkat pengambilan sampel	1Hz
Rekam ukuran	50 Byte
Data mentah per hari dan rumah tangga	4,1 MB
Data mentah per hari untuk 10 pelanggan Mio	~39 TB

Proyek *Peer Energy Cloud* (PEC) adalah proyek yang didanai publik yang telah menunjukkan bagaimana data meteran pintar dapat dianalisis dan digunakan untuk perdagangan energi di lingkungan lokal, sehingga meningkatkan stabilitas jaringan listrik secara keseluruhan. Selain itu, penelitian ini telah berhasil menunjukkan bahwa dengan mengumpulkan data yang lebih terperinci, misalnya memantau konsumsi energi masing-masing perangkat di rumah tangga, keakuratan prediksi konsumsi energi rumah tangga dapat ditingkatkan secara signifikan (Ziekow dkk. 2013). Seiring dengan meningkatnya volume data, penanganan data dengan database relasional lama menjadi semakin sulit (Strohbach dkk. 2011).

Kesimpulan

Bab ini berisi ikhtisar teknologi penyimpanan data besar saat ini serta paradigma yang muncul dan kebutuhan masa depan. Ikhtisar ini secara khusus mencakup teknologi dan pendekatan yang berkaitan dengan privasi dan keamanan. Daripada berfokus pada deskripsi rinci masing-masing teknologi, gambaran umum diberikan, dan aspek teknis yang berdampak pada penciptaan nilai dari sejumlah besar data disorot. Dampak sosial dan ekonomi dari teknologi penyimpanan data besar telah dijelaskan, dan tiga studi kasus terpilih di tiga sektor berbeda juga dijelaskan secara rinci, yang menggambarkan perlunya teknologi terukur yang mudah digunakan.

Dapat disimpulkan bahwa sudah ada banyak sekali tawaran teknologi penyimpanan data besar. Teknologi-teknologi tersebut telah mencapai tingkat kematangan yang cukup tinggi sehingga para pengadopsi awal di berbagai sektor sudah menggunakan atau berencana menggunakannya. Penyimpanan data besar sering kali memiliki keuntungan berupa skalabilitas yang lebih baik dengan harga yang lebih rendah dan kompleksitas operasional. Keadaan terkini mencerminkan bahwa pengelolaan data yang efisien pada hampir semua ukuran data bukanlah sebuah tantangan. Oleh karena itu, hal ini mempunyai potensi besar untuk mengubah bisnis dan masyarakat di banyak bidang.

Dapat juga disimpulkan bahwa terdapat kebutuhan yang kuat untuk meningkatkan kematangan teknologi penyimpanan sehingga dapat memenuhi kebutuhan masa depan dan mengarah pada adopsi yang lebih luas, khususnya di perusahaan berbasis non-IT. Peningkatan teknis yang diperlukan mencakup skalabilitas database grafik yang akan memungkinkan penanganan hubungan kompleks yang lebih baik, serta meminimalkan latensi kueri pada kumpulan data besar, misalnya dengan menggunakan database dalam memori. Hambatan besar lainnya adalah kurangnya antarmuka standar ke sistem database NoSQL. Kurangnya standardisasi mengurangi fleksibilitas dan memperlambat adopsi. Terakhir, diperlukan peningkatan besar dalam hal keamanan dan privasi. Teknologi penyimpanan yang aman perlu dikembangkan lebih lanjut untuk melindungi privasi pengguna. Rincian lebih lanjut tentang teknologi penyimpanan data besar dapat ditemukan di Curry dkk. (2014). Laporan ini, bersama dengan analisis sektor publik dan 10 sektor industri (Zillner et al. 2014), telah digunakan sebagai dasar untuk mengembangkan peta jalan lintas sektoral yang dijelaskan dalam buku ini.

BAB 8
PENGGUNAAN DATA BESAR

8.1 PENDAHULUAN

Salah satu tugas bisnis inti dari penggunaan data tingkat lanjut adalah mendukung keputusan bisnis. Penggunaan data adalah bidang luas yang dibahas dalam bab ini dengan melihat penggunaan data dari berbagai perspektif, termasuk tumpukan teknologi yang mendasarinya, tren di berbagai sektor, dampaknya terhadap model bisnis, dan persyaratan interaksi manusia-komputer.

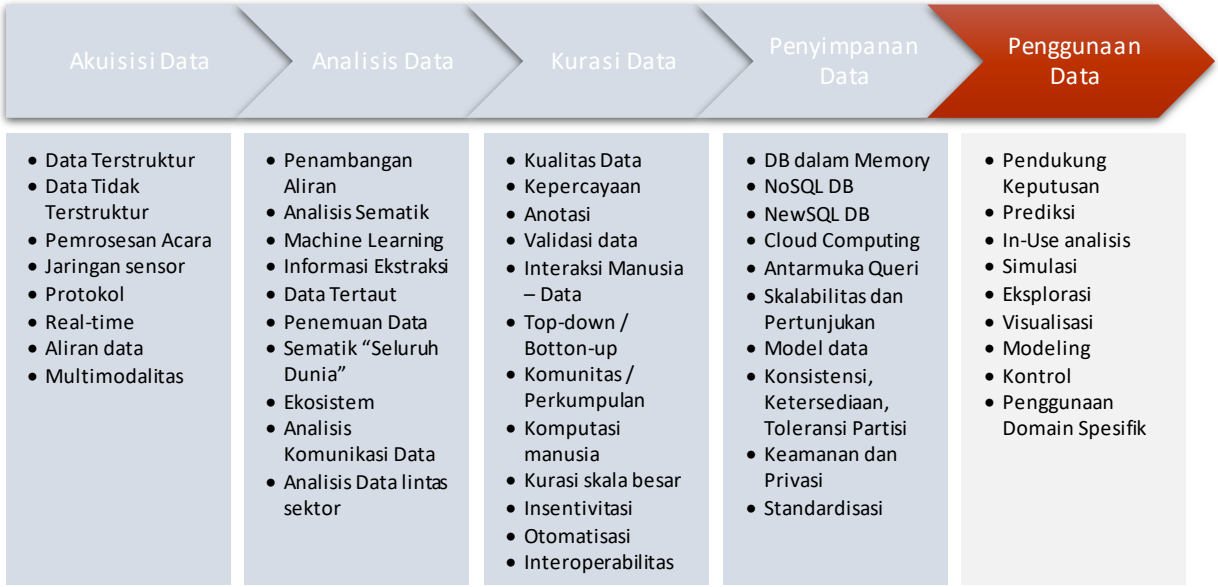
Siklus hidup informasi secara lengkap dibahas dalam buku ini, dengan bab-bab sebelumnya mencakup akuisisi, penyimpanan, analisis, dan kurasi data. Posisi penggunaan big data dalam keseluruhan Jaringan nilai big data dapat dilihat pada Gambar 8.1. Penggunaan data mencakup tujuan bisnis yang memerlukan akses ke data tersebut, analisisnya, dan alat yang diperlukan untuk mengintegrasikan analisis dalam pengambilan keputusan bisnis.

Proses pengambilan keputusan meliputi pelaporan, eksplorasi data (browsing dan pencarian), dan pencarian eksplorasi (menemukan korelasi, perbandingan, skenario bagaimana-jika, dll). Nilai bisnis dari logistik informasi tersebut ada dua: (1) kendali atas Jaringan nilai dan (2) transparansi Jaringan nilai. Yang pertama umumnya tidak bergantung pada data besar; namun hal terakhir ini memberikan peluang dan persyaratan bagi pasar dan layanan data. Big data mempengaruhi validitas pengambilan keputusan berdasarkan data di masa depan. Faktor-faktor yang mempengaruhinya adalah (1) rentang waktu pengambilan keputusan/rekomendasi, dari jangka pendek hingga jangka panjang dan (2) berbagai basis data (dalam arti non-teknis) dari data masa lalu, data historis hingga data terkini dan terkini. Aplikasi baru berbasis data akan sangat mempengaruhi perkembangan pasar baru. Potensi penghambat perkembangan tersebut adalah kebutuhan akan jaringan mitra baru (kombinasi dari kemampuan yang saat ini terpisah), proses bisnis, dan pasar.

Area khusus penggunaan big data adalah sektor manufaktur, transportasi, dan logistik. Sektor-sektor ini sedang mengalami perubahan transformasional sebagai bagian dari tren industri yang disebut “Industri 4.0”, yang berasal dari digitalisasi dan keterkaitan produk, fasilitas produksi, dan infrastruktur transportasi sebagai bagian dari perkembangan “Internet of Things”. Penggunaan data mempunyai dampak besar pada sektor-sektor ini, misalnya. Penerapan analisis prediktif dalam pemeliharaan mengarah pada model bisnis baru karena produsen mesin berada pada posisi terbaik untuk menyediakan pemeliharaan berbasis data besar. Munculnya sistem cyber-fisik (CPS) untuk produksi, transportasi, logistik, dan sektor lainnya membawa tantangan baru untuk simulasi dan perencanaan, untuk pemantauan, pengendalian, dan interaksi (oleh para ahli dan non-ahli) dengan mesin atau penggunaan data aplikasi.

Dalam skala yang lebih besar, diperlukan layanan baru dan infrastruktur layanan baru. Di bawah judul “data pintar” dan layanan data pintar, persyaratan untuk pasar data dan juga layanan dirumuskan. Selain infrastruktur teknologi untuk interaksi dan kolaborasi layanan dari berbagai sumber, terdapat permasalahan hukum dan peraturan yang perlu ditangani.

Infrastruktur layanan yang sesuai juga merupakan peluang bagi UKM untuk mengambil bagian dalam skenario penggunaan data besar dengan menawarkan layanan tertentu, misalnya melalui pasar layanan penggunaan data.



Gambar 8.1 Penggunaan data dalam Jaringan nilai big data

Akses terhadap penggunaan data diberikan melalui alat khusus dan pada gilirannya melalui bahasa kueri dan skrip yang biasanya bergantung pada penyimpanan data yang mendasarinya, mesin eksekusi, API, dan model pemrogramannya. Serangga, tumpukan teknologi yang berbeda dan beberapa trade-off yang terlibat dibahas. Aspek-aspek umum pendukung keputusan, diikuti dengan diskusi mengenai akses khusus terhadap hasil analisis melalui visualisasi dan antarmuka eksploratif baru. Interaksi manusia-komputer akan memainkan peran yang semakin besar dalam mendukung pengambilan keputusan karena banyak kasus tidak dapat mengandalkan model korelasi yang sudah ada sebelumnya. Dalam kasus seperti ini, antarmuka pengguna (misalnya dalam visualisasi data untuk analisis visual) harus mendukung eksplorasi data dan potensi koneksinya. Tren yang muncul dan kebutuhan masa depan disajikan di sub bab 8.6 dengan penekanan khusus pada Industri 4.0 dan meningkatnya kebutuhan akan data pintar dan layanan pintar.

8.2 WAWASAN PENTING UNTUK PENGGUNAAN BIG DATA

Wawasan utama untuk penggunaan big data yang diidentifikasi adalah sebagai berikut: Analisis Prediktif Contoh utama penerapan analisis prediktif adalah dalam pemeliharaan prediktif berdasarkan data sensor dan konteks untuk memprediksi penyimpangan dari interval pemeliharaan standar. Jika data menunjukkan sistem yang stabil, interval pemeliharaan dapat diperpanjang sehingga menurunkan biaya pemeliharaan. Jika data menunjukkan adanya masalah sebelum pemeliharaan terjadwal tercapai, penghematan dapat lebih tinggi jika kerusakan, biaya perbaikan, dan waktu henti dapat dihindari. Sumber informasi lebih dari sekadar data sensor dan cenderung mencakup data lingkungan dan konteks, termasuk

informasi penggunaan (misalnya beban tinggi) mesin. Karena analisis prediktif bergantung pada sensor baru dan infrastruktur pemrosesan data, produsen besar mengubah model bisnis mereka dan berinvestasi pada infrastruktur baru (menyadari dampak skala dalam perjalanan) dan menyewakan mesin kepada pelanggan mereka.

Industri 4.0 Tren yang berkembang di bidang manufaktur adalah penggunaan sistem cyber-fisik. Hal ini membawa evolusi pada proses manufaktur lama, di satu sisi menyediakan sejumlah besar sensor dan data lainnya dan di sisi lain menghadirkan kebutuhan untuk menghubungkan semua data yang tersedia melalui jaringan komunikasi dan skenario penggunaan yang memberikan potensi manfaat. Industri 4.0 mewakili masuknya TI ke dalam industri manufaktur dan membawa serta sejumlah tantangan dalam dukungan TI. Hal ini mencakup layanan untuk beragam tugas seperti perencanaan dan simulasi, pemantauan dan pengendalian, penggunaan mesin secara interaktif, logistik dan perencanaan sumber daya perusahaan (ERP), analisis prediktif, dan pada akhirnya analisis preskriptif di mana proses pengambilan keputusan dapat dikontrol secara otomatis melalui analisis data.

Integrasi Data dan Layanan Cerdas Saat mengembangkan lebih lanjut skenario Industri 4.0 di atas, layanan yang menyelesaikan tugas-tugas yang ada menjadi fokus. Untuk mengaktifkan penerapan layanan pintar untuk menangani masalah penggunaan data besar, ada masalah teknis dan organisasi. Masalah perlindungan data dan privasi, masalah peraturan, dan tantangan hukum baru (misalnya terkait masalah kepemilikan data turunan) harus diatasi.

Pada tingkat teknis, terdapat beberapa dimensi di mana interaksi layanan harus dimungkinkan: pada tingkat perangkat keras mulai dari mesin individual, hingga fasilitas, hingga jaringan; pada tingkat konseptual dari perangkat cerdas hingga sistem dan keputusan cerdas; pada tingkat infrastruktur mulai dari IaaS hingga PaaS dan SaaS hingga layanan baru untuk penggunaan data besar dan bahkan hingga proses bisnis dan pengetahuan sebagai layanan.

Eksplorasi Interaktif Saat bekerja dengan volume data yang besar dan variasi yang besar, model yang mendasari hubungan fungsional seringkali hilang. Ini berarti analis data mempunyai kebutuhan yang lebih besar untuk mengeksplorasi kumpulan data dan analisis. Hal ini diatasi melalui analisis visual dan cara visualisasi data yang baru dan dinamis, namun diperlukan antarmuka pengguna baru dengan kemampuan baru untuk eksplorasi data. Lingkungan penggunaan data yang terintegrasi memberikan dukungan, misalnya melalui mekanisme riwayat dan kemampuan untuk membandingkan analisis yang berbeda, pengaturan parameter yang berbeda, dan model yang bersaing.

8.3 DAMPAK SOSIAL DAN EKONOMI DARI PENGGUNAAN BIG DATA

Salah satu dampak paling penting dari skenario penggunaan data besar adalah ditemukannya hubungan dan ketergantungan baru dalam data yang pada permukaannya mengarah pada peluang ekonomi dan efisiensi yang lebih besar. Pada tingkat yang lebih

dalam, penggunaan big data dapat memberikan pemahaman yang lebih baik tentang ketergantungan ini, menjadikan sistem lebih transparan dan mendukung proses pengambilan keputusan ekonomi dan sosial (Manyika dkk. 2011). Dimanapun data tersedia untuk umum, pengambilan keputusan sosial didukung; jika data yang relevan tersedia pada tingkat individu, pengambilan keputusan pribadi akan didukung. Potensi transparansi melalui penggunaan big data memiliki sejumlah persyaratan:

(1) peraturan dan perjanjian mengenai akses, kepemilikan, perlindungan, dan privasi data,
(2) tuntutan terhadap kualitas data (misalnya kelengkapan, keakuratan, dan ketepatan waktu) data), dan
(3) akses terhadap data mentah serta akses terhadap alat atau layanan yang sesuai untuk penggunaan data besar.

Transparansi dengan demikian memiliki dimensi ekonomi, sosial, dan pribadi. Apabila persyaratan yang tercantum di atas dapat dipenuhi, keputusan menjadi transparan dan dapat dibuat dengan cara yang lebih obyektif dan dapat direproduksi, dimana proses pengambilan keputusan terbuka untuk melibatkan lebih banyak pihak.

Penggerak ekonomi penggunaan big data saat ini adalah perusahaan-perusahaan besar yang memiliki akses terhadap infrastruktur yang lengkap. Hal ini mencakup sektor-sektor seperti periklanan di perusahaan-perusahaan Internet dan data sensor dari infrastruktur besar (misalnya jaringan pintar atau kota pintar) atau untuk mesin yang kompleks (misalnya mesin pesawat terbang). Dalam contoh terakhir, terdapat kecenderungan ke arah integrasi penggunaan data yang lebih erat di perusahaan-perusahaan besar karena kemampuan big data tetap berada di tangan produsen (dan bukan pelanggan), misalnya ketika mesin hanya disewa dan infrastruktur data besar dimiliki dan dikelola oleh produsen.

Ada peningkatan kebutuhan akan standar dan pasar data yang dapat diakses serta layanan untuk mengelola, menganalisis, dan mengeksploitasi penggunaan data lebih lanjut. Jika persyaratan tersebut terpenuhi, peluang tercipta bagi UKM untuk berpartisipasi dalam kasus penggunaan data besar yang lebih kompleks.

8.4 PENGGUNAAN BIG DATA YANG CANGGIH

Bagian ini memberikan ikhtisar tentang kecanggihan penggunaan data besar saat ini, membahas secara singkat aspek-aspek utama tumpukan teknologi yang digunakan dan subbidang pendukung keputusan, analisis prediktif, simulasi, eksplorasi, visualisasi, dan aspek teknis lainnya.

Tumpukan Teknologi Penggunaan Big Data

Aplikasi big data bergantung pada Jaringan nilai data lengkap yang tercakup dalam proyek BIG, mulai dari akuisisi data, termasuk kurasi, penyimpanan, analisis, dan digabungkan untuk penggunaan data. Di sisi teknologi, aplikasi penggunaan data besar bergantung pada keseluruhan teknologi yang mencakup penyimpanan data dan aksesnya hingga mesin eksekusi pemrosesan yang digunakan oleh antarmuka dan bahasa kueri.

Perlu ditekankan bahwa tumpukan teknologi big data yang lengkap dapat dilihat lebih luas, yaitu mencakup infrastruktur perangkat keras, seperti sistem penyimpanan, server,

infrastruktur jaringan pusat data, organisasi data terkait dan perangkat lunak manajemen, serta keseluruhannya. berbagai layanan mulai dari konsultasi dan outsourcing hingga dukungan dan pelatihan di sisi bisnis serta sisi teknologi.

Akses pengguna sebenarnya terhadap penggunaan data diberikan melalui alat khusus dan pada gilirannya melalui bahasa kueri dan skrip yang biasanya bergantung pada penyimpanan data yang mendasarinya, mesin eksekusi, API, dan model pemrogramannya. Beberapa contohnya termasuk SQL untuk sistem manajemen basis data relasional klasik (RDBMS), Dremel dan Sawzall untuk sistem file Google (GFS), dan MapReduce, Hive, Pig, dan Jaql untuk pendekatan berbasis Hadoop, Scope untuk Dryad dan CosmosFS dari Microsoft, dan banyak lainnya persembahan, mis. Meteor/Sopremo Stratosfer dan AQL/Algebricks ASTERIX.

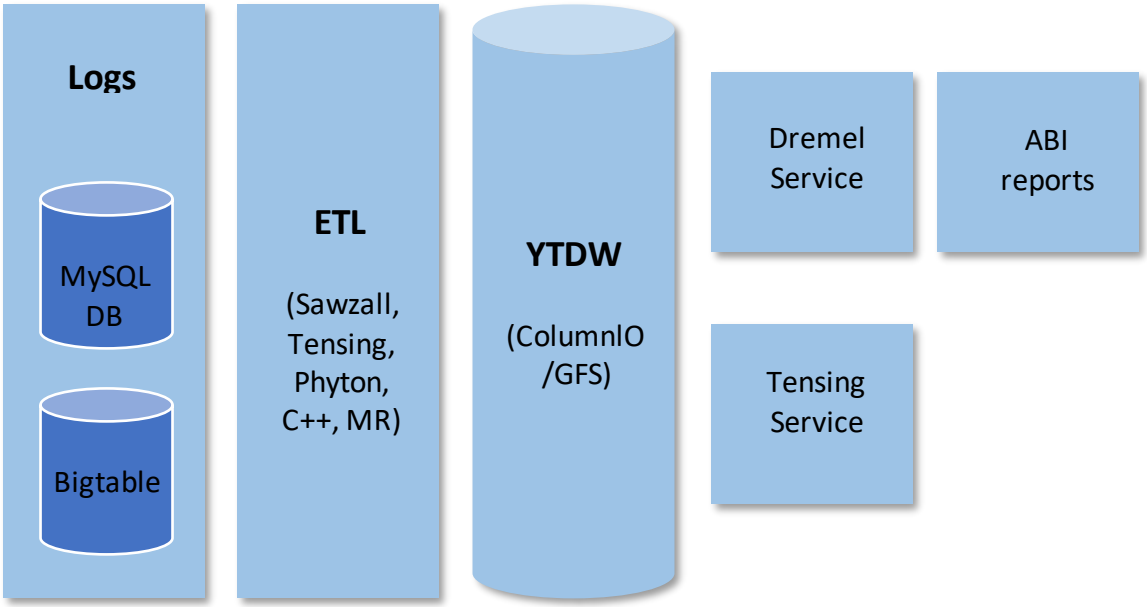
Alat analisis yang relevan untuk penggunaan data mencakup SystemT (IBM, untuk penambangan data dan ekstraksi informasi) dan Matlab (U. Auckland dan Mathworks, perwakilan untuk analisis matematis dan statistik), alat untuk intelijen dan analitik bisnis (SAS Analytics (SAS), Vertica (HP), SPSS (IBM)), alat untuk pencarian dan pengindeksan (Lucene dan Solr (Apache)), dan alat khusus untuk visualisasi (Tableau, Tableau Software). Masing-masing alat ini memiliki area penerapan spesifik dan mencakup berbagai aspek data besar.

Alat untuk penggunaan big data mendukung aktivitas bisnis yang dapat dikelompokkan menjadi tiga kategori: pencarian, pembelajaran, dan investigasi. Batasannya terkadang tidak jelas dan pembelajaran serta penyelidikan dapat dikelompokkan sebagai contoh pencarian eksplorasi. Pendukung pengambilan keputusan memerlukan akses terhadap data dalam berbagai cara, dan karena big data lebih sering memungkinkan pendeteksian korelasi yang sebelumnya tidak diketahui, akses data harus lebih sering dilakukan dari antarmuka yang memungkinkan pencarian eksplorasi dan bukan sekadar akses ke laporan yang telah ditentukan sebelumnya.

Pertukaran dalam Teknologi Penggunaan Big Data

Analisis studi kasus mendalam terhadap aplikasi big data yang lengkap dilakukan untuk menentukan keputusan yang terlibat dalam mempertimbangkan kelebihan dan kekurangan berbagai komponen tumpukan teknologi big data yang tersedia. Gambar 8.2 menunjukkan infrastruktur yang digunakan untuk Gudang Data YouTube (YTDW) Google sebagaimana dirinci dalam Chattopadhyay (2011). Beberapa pelajaran inti yang dipelajari oleh tim YouTube mencakup pertukaran fungsionalitas yang dapat diterima saat memberikan prioritas pada kueri berlatensi rendah. Hal ini membenarkan keputusan untuk tetap menggunakan ([alat Dremel (untuk menanyakan kumpulan data besar) yang memiliki kelemahan yang dapat diterima dalam kekuatan ekspresif (jika dibandingkan dengan alat berbasis SQL), namun memberikan hasil latensi rendah dan skala sesuai dengan apa yang dianggap Google sebagai “sedang” Namun perlu diperhatikan bahwa Google menggunakan “triliunan baris dalam hitungan detik”, dan berjalan pada “ribuan CPU dan petabyte data”, memproses “kuadriliun data per bulan”. Meskipun Google menganggap ini sebagai skala menengah, hal ini mungkin cukup untuk banyak aplikasi yang jelas-jelas bergerak di bidang big data. Tabel 8.1 menunjukkan perbandingan berbagai komponen teknologi penggunaan data yang digunakan

dalam YTDW, dimana latensi mengacu pada waktu yang dibutuhkan sistem untuk menjawab permintaan; skalabilitas hingga kemudahan penggunaan kumpulan data yang semakin besar; SQL mengacu pada kemampuan (yang sering kali lebih disukai) untuk menggunakan kueri SQL (atau serupa); dan kekuatan mengacu pada kekuatan ekspresif kueri penelusuran.



Gambar 8.2 Infrastruktur Gudang Data YouTube (YTDW). Berasal dari Chattopadhyay (2011)

Tabel 8.1 Perbandingan teknologi penggunaan data yang digunakan di YTDW. Sumber: Chattopadhyay (2011)

	Sawzall	menegangkan	Tubuh
Latensi	Tinggi	Sedang	Rendah
Skalabilitas	Tinggi	Tinggi	Sedang
SQL	Tidak ada	Tinggi	Sedang
Kekuatan	Tinggi	Sedang	Rendah

Pendukung Keputusan

Sistem pendukung keputusan saat ini—sejauh mengandalkan laporan statis—menggunakan teknik ini tetapi tidak memungkinkan penggunaan dinamis yang memadai untuk mendapatkan potensi penuh dari penelusuran eksplorasi. Namun, dengan kompleksitas yang semakin meningkat, kelompok-kelompok ini mencakup tujuan bisnis berikut:

- **Pencarian**: Pada tingkat kompleksitas terendah, data hanya diambil untuk berbagai tujuan. Ini termasuk pengambilan fakta dan pencarian item yang diketahui, misalnya. untuk tujuan verifikasi. Fungsi tambahan mencakup navigasi melalui kumpulan data dan transaksi.
- **Pembelajaran**: Pada tingkat berikutnya, fungsi-fungsi ini dapat mendukung perolehan pengetahuan dan interpretasi data, sehingga memungkinkan pemahaman. Fungsi

pendukungnya meliputi perbandingan, agregasi, dan integrasi data. Komponen tambahan mungkin mendukung fungsi sosial untuk pertukaran data. Contoh pembelajaran mencakup pencarian sederhana untuk item tertentu (perolehan pengetahuan), misalnya. seorang selebriti dan penggunaannya dalam periklanan (ritel). Sebuah aplikasi pencarian big data diharapkan dapat menemukan seluruh data terkait dan menyajikan tampilan yang terintegrasi.

- **Investigasi**: Pada sistem pendukung keputusan tingkat tertinggi, data dapat dianalisis, dikumpulkan, dan disintesis. Ini termasuk dukungan alat untuk pengecualian, negasi, dan evaluasi. Pada tingkat analisis ini, penemuan sebenarnya didukung dan alatnya memengaruhi perencanaan dan perkiraan. Investigasi (penemuan) tingkat yang lebih tinggi akan berupaya menemukan korelasi penting, misalnya pengaruh musim dan/atau cuaca terhadap penjualan produk tertentu pada acara tertentu. Contoh lebih lanjut, khususnya penggunaan data besar untuk keputusan bisnis strategis tingkat tinggi, diberikan di sub bab 8.6 tentang persyaratan masa depan.

Pada tingkat yang lebih tinggi lagi, fungsi-fungsi ini mungkin (sebagian) diotomatisasi untuk memberikan analisis prediktif dan bahkan normatif. Yang terakhir mengacu pada keputusan yang diperoleh dan dilaksanakan secara otomatis berdasarkan hasil analisis otomatis (atau manual). Namun, fungsi-fungsi tersebut berada di luar cakupan sistem pendukung keputusan pada umumnya dan lebih cenderung disertakan dalam lingkungan pemrosesan peristiwa kompleks (CEP) di mana latensi rendah dari keputusan otomatis dianggap lebih tinggi daripada keselamatan tambahan manusia di dalam proses. loop yang disediakan oleh sistem pendukung keputusan.

Analisis Prediktif

Contoh utama analisis prediktif adalah pemeliharaan prediktif berdasarkan penggunaan data besar. Interval pemeliharaan biasanya ditentukan sebagai keseimbangan antara pemeliharaan yang mahal dan berfrekuensi tinggi serta bahaya kegagalan sebelum pemeliharaan yang sama besarnya. Tergantung pada skenario penerapannya, masalah keselamatan sering kali memerlukan pemeliharaan yang sering, misalnya dalam industri dirgantara. Namun, pada kasus lain, kerugian akibat kegagalan mesin tidak terlalu besar dan menentukan interval perawatan hanya merupakan tindakan ekonomis.

Asumsi yang mendasari analisis prediktif adalah bahwa dengan informasi sensor yang memadai dari mesin tertentu dan database sensor serta data kegagalan yang cukup besar dari mesin ini atau jenis mesin umum, waktu spesifik hingga kegagalan mesin dapat diprediksi dengan lebih akurat. Pendekatan ini menjanjikan penurunan biaya karena:

- Interval pemeliharaan yang lebih lama karena gangguan produksi (atau pekerjaan) yang “tidak perlu” dapat dihindari bila waktu pemeliharaan rutin telah tercapai. Model prediktif memungkinkan perpanjangan interval pemeliharaan, berdasarkan data sensor saat ini.
- Jumlah kegagalan yang lebih rendah karena jumlah kegagalan yang terjadi lebih awal dari pemeliharaan terjadwal dapat dikurangi berdasarkan data sensor dan pemeliharaan prediktif yang memerlukan pekerjaan pemeliharaan lebih awal.

- Menurunkan biaya kegagalan karena potensi kegagalan dapat diprediksi melalui pemeliharaan prediktif dengan waktu peringatan awal tertentu, memungkinkan penjadwalan pekerjaan pemeliharaan/pertukaran, sehingga menurunkan waktu pemadaman.

Model Bisnis Baru

Penerapan analitik prediktif memerlukan ketersediaan data sensor untuk mesin tertentu (di mana "mesin" digunakan sebagai istilah yang cukup umum) serta kumpulan data komprehensif dari data sensor yang dikombinasikan dengan data kegagalan.

Melengkapi mesin yang ada dengan sensor tambahan, menambahkan jalur komunikasi dari sensor ke layanan pemeliharaan prediktif, dll., bisa menjadi sebuah proposisi yang mahal. Berdasarkan keengganan pelanggan mereka dalam berinvestasi, sejumlah perusahaan (terutama produsen mesin) telah mengembangkan model bisnis baru untuk mengatasi permasalahan ini.

Contoh utama adalah turbin angin GE dan mesin pesawat Rolls Royce. Mesin Rolls Royce semakin banyak ditawarkan untuk disewa, dengan kontrak layanan penuh termasuk pemeliharaan, yang memungkinkan pabrikan memperoleh manfaat dari penerapan pemeliharaan prediktif. Dengan mengkorelasikan konteks operasional dengan data sensor mesin, kegagalan dapat diprediksi sejak dini, mengurangi (biaya) penggantian, memungkinkan dilakukannya pemeliharaan terencana, bukan hanya pemeliharaan terjadwal. Solusi GE OnPoint menawarkan paket layanan serupa yang dijual bersama dengan mesin GE.

Eksplorasi

Menjelajahi kumpulan data besar dan hasil analisis terkait dapat didistribusikan ke berbagai sumber dan format (misalnya portal baru, blog perjalanan, jejaring sosial, layanan web, dll.). Untuk menjawab pertanyaan kompleks—mis. "Astronot mana saja yang pernah ke bulan?", "Di mana restoran Italia selanjutnya dengan rating tinggi?", "Tempat wisata mana yang harus saya kunjungi dan urutannya bagaimana?"—pengguna harus mengajukan banyak permintaan ke berbagai sumber dan media yang berbeda. Terakhir, hasilnya harus digabungkan secara manual.

Dukungan terhadap pendekatan coba-coba manusia dapat menambah nilai dengan menyediakan metode cerdas untuk ekstraksi dan agregasi informasi otomatis untuk menjawab pertanyaan-pertanyaan kompleks. Metode seperti ini dapat mengubah proses analisis data menjadi eksploratif dan iteratif. Pada tahap pertama, data yang relevan diidentifikasi dan kemudian konteks tahap pembelajaran kedua ditambahkan untuk data tersebut. Fase eksplorasi ketiga memungkinkan berbagai operasi untuk mengambil keputusan dari data atau mengubah dan memperkaya data.

Analisis Iteratif

Pemrosesan aliran data berulang yang efisien dan paralel membawa sejumlah tantangan teknis. Proses analisis data berulang biasanya menghitung hasil analisis dalam serangkaian langkah. Dalam setiap langkah, hasil atau keadaan antara yang baru dihitung dan diperbarui. Mengingat tingginya volume dalam aplikasi data besar, komputasi dijalankan

secara paralel, mendistribusikan, menyimpan, dan mengelola keadaan secara efisien di beberapa mesin. Banyak algorita memerlukan jumlah iterasi yang tinggi untuk menghitung hasil akhir, sehingga memerlukan iterasi latensi rendah untuk meminimalkan waktu respons secara keseluruhan. Namun, dalam beberapa aplikasi, upaya komputasi berkurang secara signifikan antara iterasi pertama dan terakhir. Sistem berbasis batch seperti Map/Reduce (Dean dan Ghemawat 2008) dan Spark (Apache 2014) mengulangi semua komputasi di setiap iterasi bahkan ketika hasil (parsial) tidak berubah.

Sistem aliran data yang benar-benar berulang seperti Stratosphere (Stratosphere 2014) atau sistem grafik khusus seperti GraphLab (Low et al. 2012) dan Google Pregel (Malewicz et al. 2010) memanfaatkan properti tersebut dan mengurangi biaya komputasi di setiap iterasi.

Visualisasi

Memvisualisasikan hasil analisis termasuk penyajian tren dan prediksi lainnya dengan alat visualisasi yang memadai merupakan aspek penting dalam penggunaan big data. Pemilihan parameter, subset, dan fitur yang relevan merupakan elemen penting dalam penambangan data dan pembelajaran mesin dengan banyak siklus yang diperlukan untuk menguji berbagai pengaturan. Karena pengaturan dievaluasi berdasarkan hasil analisis yang disajikan, visualisasi berkualitas tinggi memungkinkan evaluasi kualitas hasil secara cepat dan tepat, misalnya dalam memvalidasi kualitas prediktif suatu model dengan membandingkan hasilnya dengan kumpulan data pengujian. Tanpa visualisasi yang mendukung, hal ini dapat menjadi proses yang mahal dan lambat, sehingga menjadikan visualisasi sebagai faktor penting dalam analisis data.

Untuk menggunakan hasil analisis data pada langkah selanjutnya dari skenario penggunaan data, misalnya, memungkinkan ilmuwan data dan pengambil keputusan bisnis menarik kesimpulan dari analisis, presentasi visual yang dipilih dengan baik dapat menjadi sangat penting untuk menghasilkan rangkaian hasil yang besar. dapat dikelola dan efektif. Bergantung pada kompleksitas visualisasi, hal ini dapat memakan biaya komputasi yang mahal dan menghambat penggunaan visualisasi secara interaktif.

Namun, pencarian eksploratif dalam hasil analisis sangat penting dalam banyak kasus penggunaan data besar. Dalam beberapa kasus, hasil analisis big data hanya akan diterapkan pada satu contoh saja, misalnya mesin pesawat terbang. Namun, dalam banyak kasus, kumpulan data analisis sama rumitnya dengan data yang mendasarinya, mencapai batas teknik visualisasi statistik klasik dan memerlukan eksplorasi dan analisis interaktif (Spence 2006; Ward et al. 2010). Dalam karya penting Shneiderman tentang visualisasi (Shneiderman 1996), ia mengidentifikasi tujuh jenis tugas: ikhtisar, zoom, filter, detail sesuai permintaan, menghubungkan, sejarah, dan mengekstrak.

Namun area visualisasi lainnya berlaku untuk model data yang digunakan dalam banyak algoritma pembelajaran mesin dan berbeda dari aplikasi penambangan dan pelaporan data tradisional. Jika model data tersebut digunakan untuk klasifikasi, pengelompokan, rekomendasi, dan prediksi, kualitasnya diuji dengan kumpulan data yang dipahami dengan baik. Visualisasi mendukung validasi dan konfigurasi model serta parameternya.

Terakhir, besarnya ukuran kumpulan data merupakan tantangan berkelanjutan bagi alat visualisasi yang didorong oleh kemajuan teknologi dalam GPU, layar, dan lambatnya adopsi lingkungan visualisasi imersif seperti gua, VR, dan AR. Aspek-aspek ini tercakup dalam bidang visualisasi ilmiah dan informasi.

Analisis Visual

Definisi analisis visual, diambil dari Keim et al. (2010) mengingat istilah ini pertama kali disebutkan pada tahun 2004. Baru-baru ini, istilah ini digunakan dalam konteks yang lebih luas, menggambarkan bidang multidisiplin baru yang menggabungkan berbagai bidang penelitian termasuk visualisasi, interaksi manusia-komputer, analisis data, manajemen data, pemrosesan data geo-spasial dan temporal, dukungan keputusan spasial dan statistik.

"Vs" dari big data mempengaruhi analisis visual dalam beberapa cara. Volume data yang besar menciptakan kebutuhan untuk memvisualisasikan data berdimensi tinggi dan analisisnya serta untuk menampilkan berbagai tipe data seperti grafik tertaut. Dalam banyak kasus, visualisasi interaktif dan lingkungan analisis diperlukan yang mencakup visualisasi yang terhubung secara dinamis. Kecepatan data dan sifat dinamis dari big data memerlukan visualisasi dinamis yang diperbarui lebih sering dibandingkan alat pelaporan statis sebelumnya. Variasi data menghadirkan tantangan baru untuk kokpit dan dasbor.

Aspek dan tren baru yang utama adalah:

- Interaktivitas, kueri visual, eksplorasi (visual), interaksi multimodal (layar sentuh, perangkat input, AR/VR)
- Animasi
- Adaptasi pengguna (personalisasi)
- Semi-otomatisasi dan peringatan, CEP (pemrosesan peristiwa kompleks), dan BRE (mesin aturan bisnis)
- Tipe data yang sangat beragam, termasuk grafik, animasi, diagram mikro (Tufte), pengukur (seperti kokpit)
- Kumpulan data spasial dan aplikasi data besar yang menangani sistem informasi geografis (GIS)
- Visualisasi hampir real-time. Sektor industri keuangan (perdagangan), manufaktur (dasbor), minyak/gas—CEP, BAM (pemantauan aktivitas bisnis)
- Perincian data sangat bervariasi
- Semantik

Kasus penggunaan untuk analisis visual mencakup beberapa sektor, misalnya sektor swasta. pemasaran, manufaktur, layanan kesehatan, media, energi, transportasi, tetapi juga segmen pasar tambahan seperti rekayasa perangkat lunak. Contoh khusus analisis visual yang dipelopori oleh komunitas intelijen AS adalah visualisasi untuk keamanan siber. Karena sifat dari segmen pasar ini, rinciannya mungkin sulit diperoleh; namun ada publikasi yang tersedia, misalnya konferensi VizSec.

8.5 TREN DAN SYARAT PENGGUNAAN BIG DATA

Bagian ini memberikan gambaran umum mengenai kebutuhan masa depan dan tren yang muncul sebagai hasil penelitian gugus tugas.

Persyaratan Masa Depan untuk Penggunaan Big Data

Seiring dengan semakin pentingnya penggunaan data besar, terdapat permasalahan pada asumsi mendasar yang menjadi lebih penting. Masalah utamanya adalah validasi data yang mendasarinya. Kutipan berikut ini berasal dari Ronald Coase, pemenang Hadiah Nobel di bidang ekonomi pada tahun 1991, yang menjadikannya sebagai lelucon yang mengacu pada inkuisisi: "Jika Anda menyiksa data cukup lama, [mereka] akan mengakui apa pun".

Pada catatan yang lebih serius, ada beberapa kesalahpahaman umum dalam penggunaan data besar:

1. Mengabaikan pemodelan dan lebih mengandalkan korelasi daripada pemahaman sebab-akibat.
2. Asumsi bahwa dengan ketersediaan data yang cukup—atau bahkan seluruhnya (lihat poin berikutnya), maka tidak diperlukan model apa pun (Anderson 2008).
3. Bias sampel. Tersirat dalam big data adalah harapan bahwa semua data (pada akhirnya) akan dijadikan sampel. Hal ini jarang sekali benar; perolehan data bergantung pada pengaruh teknis, ekonomi, dan sosial yang menciptakan bias sampel.
4. Melebih-lebihkan keakuratan analisis: hasil positif palsu mudah diabaikan.

Untuk mengatasi masalah ini, persyaratan masa depan berikut ini akan menjadi penting:

- Sertakan lebih banyak pemodelan, gunakan simulasi, dan koreksi (lihat poin berikutnya) untuk bias sampel.
- Memahami sumber data dan bias sampel yang timbul dalam konteks perolehan data. Buat model kumpulan data total yang nyata untuk mengoreksi bias sampel.
- Transparansi data dan analisis: Jika data dan analisis yang diterapkan diketahui, kita dapat menilai seberapa besar peluang (secara statistik) bahwa korelasi tersebut tidak hanya "signifikan secara statistik" tetapi juga jumlah korelasi yang diuji dan mungkin tidak ada. cukup besar untuk membuat temuan korelasi hampir tidak bisa dihindari.

Dengan latar belakang peringatan umum ini, bidang-bidang utama yang diharapkan akan menentukan masa depan penggunaan big data telah diidentifikasi:

- Kualitas data dalam penggunaan big data
- Kinerja alat
- Keputusan bisnis yang strategis
- Sumber daya manusia, posisi spesifik big data

Poin terakhir dicontohkan oleh laporan pasar kerja Inggris dalam data besar (e-skills 2013) dimana permintaan meningkat dengan kuat. Secara khusus, meningkatnya jumlah administrator yang dicari menunjukkan bahwa big data sedang berkembang dari status eksperimental menjadi unit bisnis inti.

Persyaratan Khusus

Beberapa tren umum sudah dapat diidentifikasi dan dapat dikelompokkan ke dalam persyaratan berikut:

- Penggunaan data besar untuk tujuan pemasaran

- Mendeteksi kejadian abnormal pada data yang masuk secara real time
- Penggunaan data besar untuk meningkatkan efisiensi (dan efektivitas) dalam operasi inti
    - Mewujudkan penghematan selama operasi melalui ketersediaan data real-time, data yang lebih terperinci, dan pemrosesan otomatis
    - Basis data yang lebih baik untuk perencanaan rincian operasional dan proses bisnis baru
    - Transparansi untuk keperluan internal dan eksternal (pelanggan).
- Kustomisasi, adaptasi situasi, kesadaran konteks, dan personalisasi
- Integrasi dengan kumpulan data tambahan
    - Data terbuka
    - Data diperoleh melalui berbagi dan pasar data
- Masalah kualitas data ketika data tidak dikurasi atau diberikan di bawah tekanan, misalnya, untuk memperoleh akun di jaringan sosial yang tujuan penggunaannya bersifat anonim
- Masalah privasi dan kerahasiaan, kontrol akses data
- Antarmuka
    - Analisis ad hoc yang interaktif dan fleksibel untuk memberikan reaksi yang adaptif terhadap situasi dan sadar konteks, misalnya. rekomendasi
    - Antarmuka yang sesuai untuk menyediakan akses ke penggunaan data besar di lingkungan non-kantor, misalnya di kantor. situasi seluler, lantai pabrik, dll.
    - Alat untuk visualisasi, pembuatan kueri, dll.
- Kesenjangan antara pengetahuan teknis yang diperlukan untuk melaksanakan analisis data (staf teknis) dan penggunaannya dalam pengambilan keputusan bisnis (oleh staf non-teknis)
- Perlunya alat yang memungkinkan penerapan dini. Karena perkembangan dalam industri dianggap semakin cepat, langkah awal dari penerapan awal juga dianggap semakin penting dan semakin meningkatkan keunggulan kompetitif.

Industri 4.0

Untuk penerapan big data di berbagai bidang seperti manufaktur, energi, transportasi, dan bahkan kesehatan, di mana pun mesin cerdas terlibat dalam proses bisnis, terdapat kebutuhan untuk menyelaraskan teknologi perangkat keras (yaitu mesin dan sensor) dengan teknologi perangkat lunak (yaitu mesin dan sensor). Representasi data, komunikasi, penyimpanan, analisis, dan pengendalian mesin). Perkembangan masa depan dalam sistem tertanam yang berkembang menjadi "sistem cyber-fisik" perlu menyinkronkan pengembangan bersama antara perangkat keras (komputasi, penginderaan, dan jaringan) dan perangkat lunak (format data, sistem operasi, serta sistem analisis dan kontrol).

Pemasok industri mulai mengatasi masalah ini. Perangkat lunak GE mengidentifikasi "Betapapun majunya teknologi industri, kepentingan jangka pendek dan jangka panjang ini tidak dapat diwujudkan hanya dengan menggunakan teknologi saat ini. Perangkat lunak dan

perangkat keras pada mesin industri saat ini sangat saling bergantung dan saling berkaitan erat, sehingga sulit untuk meningkatkan perangkat lunak tanpa meningkatkan perangkat keras, dan sebaliknya" (Chauhan 2013).

Di satu sisi hal ini menambah ketergantungan baru pada penggunaan big data, yaitu ketergantungan pada sistem perangkat keras serta pengembangan dan pembatasannya. Di sisi lain, hal ini membuka peluang baru untuk mengatasi sistem yang lebih terintegrasi dengan aplikasi penggunaan data besar sebagai inti pendukung keputusan bisnis.

Aliran Data Iteratif

Ada dua bidang persyaratan utama untuk implementasi penggunaan data besar yang efisien dan kuat yang berkaitan dengan arsitektur dan teknologi yang mendasari dalam pemrosesan kumpulan data besar dan aliran data besar yang terdistribusi dan berlatensi rendah.

- **Pipelining Dan Materialisasi**: Kecepatan data yang tinggi menimbulkan tantangan khusus untuk pemrosesan aliran data. Arsitektur yang mendasarinya didasarkan pada pendekatan jalur pipa di mana data yang diproses dapat diteruskan ke langkah pemrosesan berikutnya dengan penundaan yang sangat rendah untuk menghindari kemacetan jalur pipa. Jika algoritma tersebut tidak ada, data dikumpulkan dan disimpan sebelum diproses. Pendekatan seperti ini disebut "materialisasi". Latensi rendah untuk kueri biasanya hanya dapat diwujudkan dalam pendekatan pipeline.
- **Toleransi Kesalahan**: Toleransi kesalahan dan minimalisasi kesalahan merupakan tantangan penting bagi sistem perpipaan. Kegagalan pada node komputasi sering terjadi dan dapat menyebabkan sebagian hasil analisis hilang. Sistem paralel harus dirancang dengan cara yang kuat untuk mengatasi kesalahan tersebut tanpa kegagalan. Pendekatan yang umum adalah titik pemeriksaan berkelanjutan di mana hasil antara disimpan, memungkinkan rekonstruksi keadaan sebelumnya jika terjadi kesalahan. Menyimpan data di pos pemeriksaan mudah diterapkan, namun menimbulkan biaya eksekusi yang tinggi karena kebutuhan sinkronisasi dan biaya penyimpanan saat menyimpan ke penyimpanan persisten. Algoritma alternatif baru menggunakan pendekatan optimis yang dapat menciptakan kembali keadaan valid yang memungkinkan kelanjutan komputasi. Pendekatan seperti ini menambah biaya hanya jika terjadi kesalahan namun hanya dapat diterapkan pada kasus tertentu.

Visualisasi

Ada sejumlah tren masa depan yang perlu diatasi dalam bidang visualisasi dan analisis visual dalam jangka menengah hingga masa depan, misalnya (Keim dkk. 2010):

- Persepsi visual dan aspek kognitif
- "Desain" (seni visual)
- Kualitas data, data yang hilang, asal data
- Kolaborasi multipihak, misalnya dalam skenario darurat
- Analisis visual pasar massal dan pengguna akhir

Selain itu, Markl dkk. (2013) menyusun daftar panjang pertanyaan penelitian yang berikut ini sangat penting bagi penggunaan dan visualisasi data:

- ⌘ Bagaimana visualisasi dapat mendukung proses pembuatan model data untuk prediksi dan klasifikasi?
- ⌘ Teknologi visualisasi manakah yang dapat mendukung analis dalam analisis eksploratif?
- ⌘ Bagaimana audio dan video (animasi) dikumpulkan dan dihasilkan secara otomatis untuk analisis visual?
- ⌘ Bagaimana meta-informasi seperti semantik, kualitas data, dan asal dapat dimasukkan ke dalam proses visualisasi?

Paradigma yang Muncul dalam Penggunaan Big Data

Sejumlah paradigma yang muncul untuk penggunaan data besar telah diidentifikasi dan terbagi dalam dua kategori. Kategori pertama mencakup semua aspek integrasi penggunaan data besar ke dalam proses bisnis yang lebih besar dan evolusi menuju tren baru yang disebut "data pintar". Tren kedua lebih bersifat lokal dan menyangkut alat antarmuka untuk bekerja dengan data besar. Alat eksplorasi baru akan memungkinkan ilmuwan dan analis data secara umum mengakses lebih banyak data dengan lebih cepat dan mendukung pengambilan keputusan dengan menemukan tren dan korelasi dalam kumpulan data yang dapat didasarkan pada model proses bisnis yang mendasarinya.

Ada sejumlah tren teknologi yang muncul (misalnya database dalam memori) yang memungkinkan analisis yang cukup cepat untuk memungkinkan analisis data eksploratif dan dukungan keputusan. Pada saat yang sama, layanan baru sedang berkembang, menyediakan analisis data, integrasi, dan transformasi data besar menjadi pengetahuan organisasi.

Seperti di semua pasar digital baru, perkembangan ini sebagian didorong oleh start-up yang mengisi ceruk teknologi baru; namun, dominasi pemain besar sangatlah penting karena mereka memiliki akses yang lebih mudah terhadap data besar. Transfer teknologi ke UKM lebih cepat dibandingkan revolusi digital sebelumnya; namun, kasus bisnis yang sesuai untuk UKM tidak mudah untuk dirancang secara terpisah dan biasanya melibatkan integrasi ke dalam jaringan atau pasar yang lebih besar.

Data Cerdas

Konsep smart data diartikan sebagai penerapan big data yang efektif dan berhasil mendatangkan manfaat yang terukur dan mempunyai makna yang jelas (semantik), kualitas dan keamanan data yang terukur (termasuk standar privasi data).

Skenario data cerdas merupakan perluasan alami dari penggunaan data besar dalam konteks apa pun yang layak secara ekonomi. Hal ini dapat berupa model bisnis baru yang dimungkinkan oleh penerapan analisis data yang inovatif, atau peningkatan efisiensi/profitabilitas model bisnis yang sudah ada. Yang terakhir ini mudah untuk dimulai karena data tersedia dan, karena data tersebut tertanam dalam proses bisnis yang ada, sudah memiliki makna (semantik) dan struktur bisnis yang ditetapkan. Oleh karena itu, nilai tambah dari jaminan kualitas data dan metadata yang adalah yang dapat menjadikan penggunaan big data menjadi sebuah smart data.

Di luar tantangan teknis, munculnya data pintar juga membawa tantangan tambahan:

1. Menyelesaikan permasalahan regulasi mengenai kepemilikan data dan privasi data (Bitkom 2012).
2. Membuat data lebih mudah diakses dengan melakukan penataan melalui penambahan metadata, memungkinkan integrasi silo data terpisah (Bertolucci 2013).
3. Meningkatkan manfaat dari data terbuka dan sumber data terkait yang sudah tersedia. Potensi pasar mereka saat ini belum sepenuhnya terealisasi (Groves et al. 2013).

Potensi utama penggunaan data, menurut Lo (2012), terdapat pada optimalisasi proses bisnis, peningkatan manajemen risiko, dan pengembangan produk yang berorientasi pasar. Tujuan peningkatan penggunaan big data sebagai data pintar adalah untuk memecahkan tantangan sosial dan ekonomi di banyak sektor, termasuk energi, manufaktur, kesehatan, dan media.

Bagi UKM, fokusnya adalah pada integrasi ke dalam Jaringan nilai yang lebih besar yang memungkinkan banyak perusahaan berkolaborasi untuk memberi UKM akses terhadap dampak skala yang mendasari potensi penggunaan data besar. Mengembangkan kolaborasi semacam itu dimungkinkan oleh data cerdas ketika makna data bersifat eksplisit, sehingga memungkinkan kombinasi data informasi perencanaan, pengendalian, produksi, dan keadaan di luar batas masing-masing perusahaan mitra.

Data pintar menciptakan persyaratan di empat bidang: semantik, kualitas data, keamanan dan privasi data, dan metadata. Semantik Memahami dan menyediakan makna kumpulan data memungkinkan langkah-langkah penting dalam pemrosesan data cerdas:

- Interoperabilitas
- Pemrosesan yang cerdas
- Integrasi data
- Analisis data adaptif

Metadata Sebagai sarana untuk menyandikan dan menyimpan makna (semantik) data. Metadata juga dapat digunakan untuk menyimpan informasi lebih lanjut tentang kualitas data, asal, hak penggunaan, dll. Saat ini ada banyak usulan namun belum ada standar yang ditetapkan untuk metadata. Kualitas Data Kualitas dan asal data adalah salah satu persyaratan yang dipahami dengan baik untuk data besar (terkait dengan salah satu "V", yaitu "kebenaran").

Keamanan dan Privasi Data Masalah-masalah yang terpisah namun terkait ini sangat dipengaruhi oleh standar peraturan yang ada. Pelanggaran undang-undang privasi data dapat dengan mudah timbul dari pemrosesan data pribadi, misalnya. profil pergerakan, data kesehatan, dll. Meskipun data tersebut bisa sangat bermanfaat, pelanggaran terhadap undang-undang privasi data akan dikenakan hukuman yang berat. Selain menghilangkan peraturan tersebut, metode anonimisasi (ICO 2012) dan nama samaran (Gowing dan Nickson 2010) dapat dikembangkan dan digunakan untuk mengatasi masalah ini.

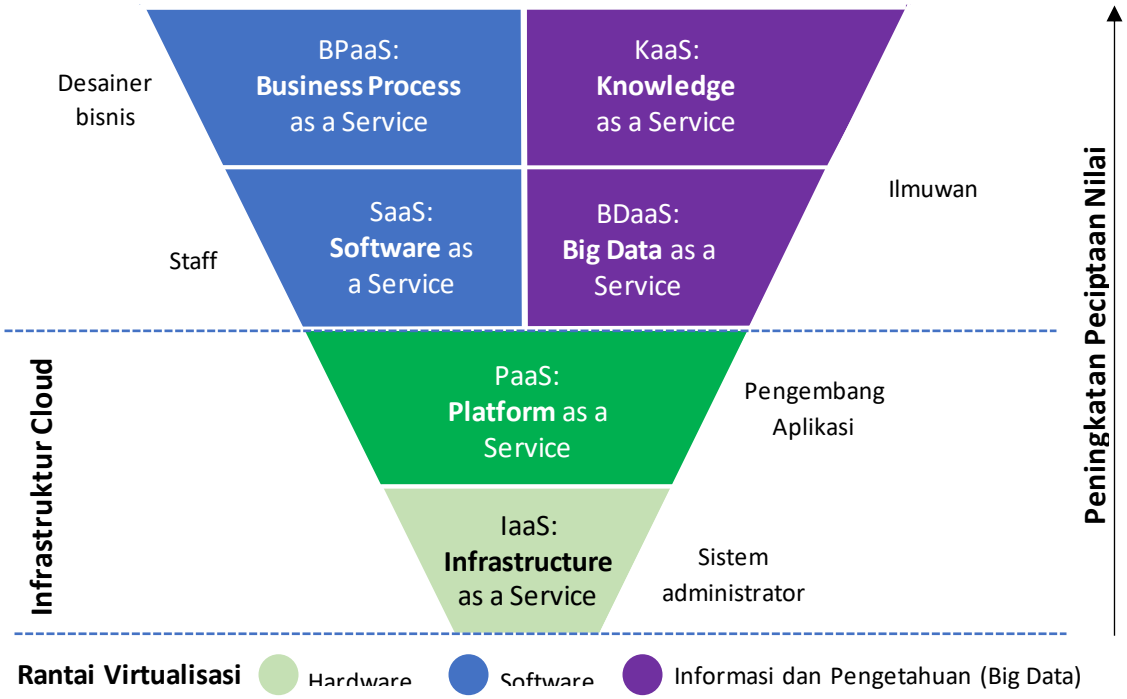
Penggunaan Big Data dalam Lingkungan Terintegrasi dan Berbasis Layanan

Integrasi layanan digital yang berkelanjutan (Internet of Services), produk digital pintar (Internet of Things), dan lingkungan produksi (Internet of Things, Industri 4.0) mencakup penggunaan data besar di sebagian besar langkah integrasi. Sebuah studi terbaru yang dilakukan General Electric meneliti berbagai dimensi integrasi dalam industri penerbangan

(Evans dan Annunziata 2012). Produk pintar seperti turbin diintegrasikan ke dalam mesin yang lebih besar, dan contoh pertama adalah pesawat terbang. Pesawat pada gilirannya merupakan bagian dari keseluruhan armada yang beroperasi di jaringan kompleks bandara, hanggar pemeliharaan, dll. Pada setiap langkah, integrasi proses bisnis saat ini diperluas dengan integrasi data besar. Manfaat pengoptimalan dapat diperoleh di setiap tingkat (aset, fasilitas, armada, dan seluruh jaringan) dan dengan mengintegrasikan pengetahuan dari data di semua langkah.

Integrasi Layanan

Infrastruktur dimana penggunaan big data akan diterapkan akan beradaptasi dengan kecenderungan integrasi ini. Perangkat keras dan perangkat lunak akan ditawarkan sebagai layanan, semuanya terintegrasi untuk mendukung penggunaan data besar. Lihat Gambar 8.3 untuk gambaran konkrit dari rangkaian layanan yang akan menyediakan lingkungan untuk "Di luar standar dan protokol teknis, platform baru yang memungkinkan perusahaan untuk membangun aplikasi spesifik berdasarkan kerangka/arsitektur bersama [diperlukan]", seperti yang diperkirakan oleh studi GE atau "Ada juga kebutuhan akan inovasi berkelanjutan dalam teknologi dan teknik yang akan membantu individu dan organisasi untuk mengintegrasikan, menganalisis, memvisualisasikan, dan memanfaatkan aliran data besar yang terus meningkat", seperti yang digambarkan oleh studi McKinsey (Manyika dkk.2011).



Gambar 8.3 Big data dalam konteks infrastruktur layanan yang diperluas. W. Wahlster (2013, Komunikasi Pribadi)

Gambar 8.3 menunjukkan big data sebagai bagian dari infrastruktur layanan tervirtualisasi. Pada tingkat terbawah, infrastruktur perangkat keras yang ada saat ini akan divirtualisasikan dengan teknologi komputasi awan; infrastruktur perangkat keras serta platform akan disediakan sebagai layanan. Selain infrastruktur berbasis cloud ini, perangkat lunak sebagai layanan (SaaS) dan proses bisnis sebagai layanan (BPaaS) juga dapat dibangun.

Secara paralel, big data akan ditawarkan sebagai layanan dan tertanam sebagai prasyarat untuk layanan pengetahuan, misalnya. integrasi teknologi semantik untuk analisis data tidak terstruktur dan agregat. Perhatikan bahwa data besar sebagai layanan dapat dilihat sebagai perluasan lapisan antara PaaS dan SaaS.

Jaringan virtualisasi dari perangkat keras ke perangkat lunak hingga informasi dan pengetahuan juga mengidentifikasi keterampilan yang dibutuhkan untuk memelihara infrastruktur. Pekerja pengetahuan atau data scientist diperlukan untuk menjalankan big data dan layanan pengetahuan.

Eksplorasi Kompleks

Alat eksplorasi data besar mendukung kumpulan data yang kompleks dan analisisnya melalui banyak pendekatan baru, misalnya. Bagian selanjutnya tentang visualisasi. Metode eksplorasi data dan hasil analisis saat ini memiliki kelemahan utama yaitu pengguna hanya dapat mengikuti eksplorasinya secara selektif dalam satu arah. Jika mereka memasuki jalan buntu atau keadaan yang tidak memuaskan, mereka harus mundur ke keadaan sebelumnya, seperti dalam algoritma pencarian mendalam pertama atau mendaki bukit. Antarmuka pengguna yang muncul untuk eksplorasi paralel (CITE) lebih serbaguna dan dapat dibandingkan dengan pencarian terbaik pertama atau pencarian sinar: pengguna dapat mengikuti dan membandingkan beberapa rangkaian eksplorasi pada saat yang bersamaan.

Contoh awal dari pendekatan ini telah dikembangkan dengan nama “antarmuka subjungtif” (Lunzer dan Hornbæk 2008) dan diterapkan pada kumpulan data geografis (Javed et al. 2012) dan sebagai “penelusuran segi paralel” (Buschbeck et al. 2013). Pendekatan terakhir mengasumsikan data terstruktur tetapi berlaku untuk semua jenis kumpulan data, termasuk hasil analisis dan CEP (pemrosesan peristiwa kompleks).

Alat eksplorasi yang kompleks ini mengatasi bahaya yang melekat dalam analisis data besar yang muncul ketika kumpulan data besar secara otomatis mencari korelasi: semakin banyak korelasi yang tampaknya signifikan secara statistik akan ditemukan dan perlu diuji untuk mengetahui penyebab mendasarnya dalam suatu model atau oleh analisis manusia yang ahli. Eksplorasi yang kompleks dapat mendukung proses pengecekan dengan memungkinkan eksplorasi paralel terhadap variasi pola dan konsekuensi yang diharapkan dari asumsi sebab akibat.

8.6 STUDI KASUS SEKTOR UNTUK PENGGUNAAN BIG DATA

Pada bagian ini disajikan ikhtisar studi kasus yang menunjukkan nilai aktual dan potensial dari penggunaan big data. Rincian lebih lanjut dapat ditemukan di Zillner dkk. (2013, 2014). Kasus penggunaan yang dipilih di sini memberikan contoh aspek tertentu yang tercakup dalam laporan tersebut.

Pelayanan Kesehatan: Pendukung Keputusan Klinis

Deskripsi Aplikasi pendukung keputusan klinis (CDS) bertujuan untuk meningkatkan efisiensi dan kualitas operasi perawatan dengan membantu dokter dan profesional kesehatan dalam proses pengambilan keputusan mereka. Aplikasi CDS memungkinkan akses informasi yang bergantung pada konteks dengan memberikan informasi pra-diagnosis, atau dengan

memvalidasi dan mengoreksi data. Oleh karena itu, sistem CDS mendukung dokter dalam pengambilan keputusan, yang sekali lagi membantu mengurangi kesalahan pengobatan serta membantu meningkatkan efisiensi. Dengan mengandalkan teknologi big data, aplikasi pendukung keputusan klinis di masa depan akan menjadi jauh lebih cerdas. Contoh kasus penggunaan adalah pra-diagnosis gambar medis, dengan rekomendasi pengobatan yang mencerminkan pedoman medis yang ada. Prasyarat intinya adalah integrasi data yang komprehensif dan kualitas data tingkat tinggi yang diperlukan agar dokter dapat benar-benar mengandalkan dukungan keputusan otomatis.

Pemantauan dan Pengawasan terhadap Penyelenggara Perjudian Online

Deskripsi Skenario masa depan ini mewakili kebutuhan yang jelas. Tujuan utamanya adalah deteksi penipuan yang sulit dilakukan karena jumlah data yang diterima secara real time, harian dan bulanan, tidak dapat diproses dengan alat database standar. Data real-time diterima dari operator perjudian setiap lima menit. Saat ini, pengawas harus menentukan kasus yang akan menerapkan analisis offline terhadap data yang dipilih.

Prasyarat intinya adalah kebutuhan untuk mengeksplorasi data secara interaktif, membandingkan berbagai model dan pengaturan parameter berdasarkan teknologi, misalnya. pemrosesan peristiwa kompleks yang memungkinkan analisis real-time dari kumpulan data tersebut. Kasus penggunaan ini berkaitan dengan masalah analisis dan eksplorasi visual, serta analisis prediktif.

Telco, Media, dan Hiburan: Peningkatan Bandwidth Dinamis

Deskripsi Pengenalan penawaran Telco baru (misalnya aplikasi game baru) dapat menyebabkan masalah dengan alokasi bandwidth. Skenario seperti ini sangat penting bagi penyedia telekomunikasi, karena lebih banyak keuntungan yang didapat dari layanan data dibandingkan dengan layanan suara. Untuk mengetahui dengan tepat penyebab masalah bandwidth, transkrip percakapan pusat panggilan dapat diperoleh untuk mengidentifikasi pelanggan dan game yang terlibat dengan informasi waktu, menerapkan langkah-langkah infrastruktur untuk secara dinamis mengubah bandwidth yang disediakan sesuai dengan penggunaan. Prasyarat inti terkait dengan analisis prediktif. Jika masalah dapat dideteksi saat masalah tersebut sedang terjadi, maka puncak dapat dihindari sama sekali. Jika pendukung keputusan dapat diotomatisasi, skenario ini dapat diperluas hingga bersifat preskriptif analisis.

Manufaktur: Analisis Prediktif

Deskripsi Jika data sensor, data kontekstual dan lingkungan tersedia, kemungkinan kegagalan mesin dapat diprediksi. Prediksi didasarkan pada nilai sensor abnormal yang sesuai dengan model kegagalan fungsional. Lebih jauh lagi, informasi konteks seperti kesimpulan mengenai penggunaan berat atau ringan tergantung pada tugas yang dilaksanakan (diambil, misalnya dari sistem ERP) dan informasi yang berkontribusi seperti kondisi cuaca, dll., dapat diperhitungkan.

Prasyarat inti, selain persyaratan klasik seperti integrasi data dari berbagai sumber data yang sebagian tidak terstruktur, adalah model prediksi yang transparan dan kumpulan

data yang cukup besar untuk mengaktifkan algoritme pembelajaran mesin yang mendasarinya.

Kesimpulan

Bab ini memberikan informasi terkini serta persyaratan masa depan dan tren penggunaan data besar yang sedang berkembang. Kegunaan utama aplikasi big data adalah untuk mendukung keputusan, dalam analisis prediktif (misalnya untuk pemeliharaan prediktif), dan dalam simulasi dan pemodelan. Tren baru bermunculan dalam visualisasi (analitik visual) dan sarana baru dalam eksplorasi dan perbandingan analisis alternatif dan analisis pesaing.

Area khusus penggunaan big data adalah sektor manufaktur, transportasi, dan logistik dengan tren baru "Industri 4.0". Munculnya sistem cyber-fisik untuk produksi, transportasi, logistik, dan sektor lainnya membawa tantangan baru bagi simulasi dan perencanaan, pemantauan, pengendalian, dan interaksi (oleh para ahli dan non-ahli) dengan mesin atau aplikasi penggunaan data besar. Dalam skala yang lebih besar, diperlukan layanan baru dan infrastruktur layanan baru. Di bawah judul "data pintar" dan layanan data pintar, persyaratan untuk pasar data dan juga layanan dirumuskan. Selain infrastruktur teknologi untuk interaksi dan kolaborasi layanan dari berbagai sumber, terdapat permasalahan hukum dan peraturan yang perlu ditangani. Infrastruktur layanan yang sesuai juga merupakan peluang bagi UKM untuk mengambil bagian dalam skenario penggunaan data besar dengan menawarkan layanan tertentu, misalnya melalui pasar layanan data.

BAB 9
INOVASI BERBASIS BIG DATA DI SEKTOR INDUSTRI

9.1 PENDAHULUAN

Apa pun bentuknya, data mempunyai potensi untuk menyampaikan cerita, mengidentifikasi penghematan dan efisiensi biaya, koneksi dan peluang baru, serta memungkinkan peningkatan pemahaman tentang masa lalu untuk membentuk masa depan yang lebih baik (US Chamber of Commerce Foundation 2014). Big data berarti sejumlah besar informasi termasuk data yang dihasilkan pengguna dari platform media sosial (yaitu data Internet); data mesin, seluler, dan GPS serta Internet of Things (data industri dan sensor); data bisnis termasuk pelanggan, inventaris, dan data transaksional (data perusahaan); kumpulan data yang dihasilkan atau dikumpulkan oleh lembaga pemerintah, serta universitas dan organisasi nirlaba (data publik) (US Chamber of Commerce Foundation 2014). Bagi banyak perusahaan dan pemerintah di berbagai belahan dunia, teknik untuk memproses dan menganalisis data dalam jumlah besar (big data) merupakan sumber daya penting untuk mendorong penciptaan nilai, mengembangkan produk, proses, dan pasar baru, serta memungkinkan terciptanya pengetahuan baru (OECD 2014). Pada tahun 2013 saja, perekonomian yang berbasis data menambah nilai baru sebesar Rp.67 miliar pada perekonomian Australia, setara dengan 4,4 % dari produk domestik bruto atau seluruh sektor ritelnya (Stone dan Wang 2014).

Sebagai sumber pertumbuhan dan pembangunan ekonomi, big data merupakan sumber daya infrastruktur yang dapat digunakan dalam beberapa cara untuk menghasilkan produk dan layanan yang berbeda. Hal ini juga memungkinkan terciptanya pengetahuan yang penting untuk mengendalikan fenomena alam, sistem sosial, atau proses organisasi dan mendukung pengambilan keputusan yang kompleks (OECD 2014). Dalam hal ini, komunitas pembangunan internasional dan Perserikatan Bangsa-Bangsa sedang mencari dukungan politik di tingkat tertinggi dalam memanfaatkan inovasi berbasis data untuk mendukung pembangunan berkelanjutan, khususnya di bawah Tujuan Pembangunan Berkelanjutan (SDGs) global yang baru (Kelompok Penasihat Ahli Independen tentang Pembangunan Berkelanjutan). Revolusi Data 2014). Kota-kota seperti Helsinki, Manchester, Amsterdam, Barcelona, dan Chicago juga memanfaatkan data yang besar dan terbuka dari jaringan sensor terbuka, proses sektor publik, dan data sosial yang dikumpulkan dari berbagai sumber untuk meningkatkan mobilitas, mendorong penciptaan layanan publik digital, dan secara umum memungkinkan kecerdasan kota yang lebih baik untuk mendukung perencanaan dan pembangunan kota yang lebih efektif.

Pada saat yang sama, terdapat peningkatan pemahaman mengenai tantangan yang terkait dengan eksploitasi big data di masyarakat. Tantangan-tantangan ini berkisar dari kurangnya kapasitas yang diperlukan (misalnya literasi data) hingga dilema etika dalam menangani big data dan bagaimana memberikan insentif bagi partisipasi pemangku

kepentingan penting lainnya dalam mengadopsi dan memanfaatkan inovasi berbasis big data untuk mengatasi tantangan sosial.

Bab ini menjelaskan hal-hal yang terlibat dalam inovasi berbasis data besar, memberikan contoh inovasi berbasis data besar di berbagai sektor, dan menyatukan faktor-faktor pendukung dan tantangan yang terkait dengan pengembangan ekosistem inovasi data besar. Bab ini ditutup dengan menawarkan rekomendasi (kebijakan) praktis tentang cara mengembangkan program dan inisiatif inovasi big data yang layak.

9.2 INOVASI BERBASIS BIG DATA

Inovasi adalah suatu proses berulang yang bertujuan untuk menciptakan produk, proses, pengetahuan, atau layanan baru dengan menggunakan pengetahuan baru atau bahkan yang sudah ada (Kusiak 2009). Inovasi berbasis data memerlukan eksploitasi segala jenis data dalam proses inovasi untuk menciptakan nilai (Stone dan Wang 2014). Tren inovasi berbasis data besar yang muncul mengarah pada pengembangan barang dan jasa berbasis data dan dapat memungkinkan perencanaan berbasis data, pemasaran berbasis data, dan operasi berbasis data di seluruh sektor dan domain industri. Dari perspektif ekonomi, data sebagai barang milik bersama yang tidak saling bersaing seperti minyak bumi berfungsi sebagai sumber daya infrastruktur (dari perspektif fungsional) yang dapat dieksploitasi secara bersamaan oleh banyak pengguna atau pelaku untuk tujuan yang saling bersaing atau saling melengkapi. Permintaan akan data dalam hal ini menurut OECD (2014) terutama didorong oleh kegiatan produktif hilir yang memerlukan data sebagai masukan dan, pada kenyataannya, merupakan modal yang tidak sepele. Selain itu, penulis yang sama menegaskan bahwa sumber daya data dapat digunakan sebagai masukan untuk berbagai macam barang, termasuk barang swasta, publik, dan sosial. Dengan kata lain, big data berpotensi memberikan keuntungan yang signifikan terhadap skala dan cakupan.

Inovasi berbasis data besar secara implisit dikaitkan dengan model Jaringan nilai atau lebih tepatnya "Jaringan nilai virtual" yang menentukan bagaimana data yang diminati akan dikumpulkan, diorganisasikan, dipilih, diubah menjadi produk atau layanan, dan didistribusikan (Rayport dan Sviokla 1995; Piccoli 2012). Jaringan nilai big data seperti yang dibahas di Bab. 3 merupakan inti dari penyampaian inovasi berbasis data dengan menggunakan teknologi big data. Pada tingkat organisasi, setidaknya ada dua kategori inisiatif strategis yang dapat dihasilkan dari inovasi berbasis big data dan Jaringan nilai big data yang mendasarinya. Kategori inisiatif pertama bertujuan untuk menyediakan informasi mengenai aspek proses dan layanan organisasi untuk memungkinkan perbaikan. Secara umum, dengan menginstrumentasikan operasi organisasi, sejumlah besar data (yaitu data besar) dihasilkan yang menginformasikan atau mendorong perubahan yang diperlukan (Piccoli 2012). Rangkaian inisiatif kedua bersifat eksternal dan melibatkan eksploitasi data pelanggan seperti pencarian dan log pengguna, catatan transaksi, dan konten lain yang dihasilkan pelanggan untuk mendorong pemasaran jangka panjang, rekomendasi yang ditargetkan dan dipersonalisasi, peningkatan penjualan, dan kepuasan pelanggan. Contoh populer dari hal ini adalah algoritma pemfilteran kolaboratif Netflix untuk memprediksi rating film pengguna

(Chen dan Storey 2012). Contoh lainnya adalah penggunaan perilaku penelusuran pengguna oleh Google untuk menargetkan iklan (US Chamber of Commerce Foundation 2014).

Di Amerika Serikat, ratusan perusahaan memanfaatkan data terbuka dan besar (seperti data cuaca dan GPS) sebagai sumber daya utama untuk menghasilkan nilai di berbagai sektor termasuk keuangan dan investasi, pendidikan, lingkungan dan cuaca, perumahan dan real estat, serta pangan dan pertanian (US Chamber of Commerce Foundation 2014). Bagian berikutnya menguraikan sejumlah transformasi berbasis data di berbagai sektor termasuk telekomunikasi, layanan kesehatan, sektor publik, keuangan dan asuransi, media dan hiburan, energi, dan transportasi.

9.3 TRANSFORMASI DI SEKTOR BISNIS

Proyek BIG mengkaji bagaimana teknologi big data dapat memungkinkan inovasi dan transformasi bisnis di berbagai sektor dengan mengumpulkan kebutuhan big data dari sektor industri vertikal, termasuk kesehatan, sektor publik, keuangan, asuransi, telekomunikasi, media, hiburan, manufaktur, ritel, energi, dan mengangkut. Ada sejumlah tantangan yang perlu diatasi sebelum inovasi berbasis data besar diadopsi secara umum. Big data hanya dapat berhasil mendorong inovasi jika sebuah bisnis menerapkan strategi data yang terdefinisi dengan baik sebelum mulai mengumpulkan dan memproses informasi. Tentu saja, investasi di bidang teknologi memerlukan strategi untuk menggunakannya sesuai dengan ekspektasi komersial; jika tidak, lebih baik tetap mempertahankan sistem dan prosedur yang ada. Organisasi-organisasi di berbagai sektor kini mulai meluangkan waktu untuk memahami ke mana arah strategi ini.

Hasil lengkap analisis ini tersedia di Zillner et al. (2014). Bagian III buku ini memberikan ringkasan singkat mengenai temuan-temuan utama dari sejumlah sektor terpilih. Bagian selanjutnya dari bab ini memberikan ringkasan eksekutif mengenai temuan-temuan dari masing-masing sektor beserta pembahasan dan analisisnya.

Pelayanan Kesehatan

Investigasi sektor kesehatan di bab 10 mengungkapkan beberapa perkembangan, seperti meningkatnya biaya layanan kesehatan, meningkatnya kebutuhan akan cakupan layanan kesehatan, dan pergeseran tren penggantian biaya penyedia layanan kesehatan, yang telah memicu permintaan akan teknologi big data. Di sektor ini, ketersediaan dan akses data kesehatan terus meningkat, teknologi big data yang diperlukan (seperti teknologi integrasi dan analisis data tingkat lanjut) sudah tersedia, dan penerapan praktik terbaik telah menunjukkan potensi big data. teknologi. Namun, revolusi big data di bidang layanan kesehatan masih berada pada tahap awal dan masih banyak potensi penciptaan nilai dan pengembangan bisnis yang belum dimanfaatkan dan dieksplorasi. Hambatan saat ini menuju inovasi berbasis data besar adalah sistem insentif yang sudah mapan dalam sistem layanan kesehatan yang menghambat kolaborasi dan, dengan demikian, pertukaran dan pertukaran data. Tren pemberian layanan kesehatan berbasis nilai akan mendorong kolaborasi para pemangku kepentingan untuk meningkatkan nilai pengobatan pasien, dan dengan demikian akan secara signifikan meningkatkan kebutuhan akan aplikasi data besar.

Sektor Publik

Investigasi sektor publik di Gambar 11.1 menunjukkan bahwa sektor ini menghadapi beberapa tantangan penting kurangnya produktivitas dibandingkan dengan sektor lain, keterbatasan anggaran, dan masalah struktural lainnya akibat populasi menua yang akan menyebabkan peningkatan permintaan terhadap layanan medis dan sosial, serta kurangnya layanan medis dan sosial yang diperkirakan akan berdampak buruk pada sektor ini misalnya tenaga kerja muda di masa depan.

Sektor publik semakin sadar akan potensi nilai yang dapat diperoleh dari inovasi berbasis data besar melalui peningkatan efektivitas dan efisiensi serta dengan alat analisis baru. Pemerintah menghasilkan dan mengumpulkan data dalam jumlah besar melalui aktivitas sehari-hari, seperti mengelola pembayaran pensiun dan tunjangan, pengumpulan pajak, dan lain-lain. Persyaratan utama, yang sebagian besar bersifat non-teknis, dari sektor publik adalah:

- **Interoperabilitas**: Hambatan dalam mengeksploitasi aset data karena fragmentasi kepemilikan data dan silo data yang diakibatkannya.
- **Dukungan Legislatif Dan Kemauan Politik**: Proses pembuatan undang-undang baru seringkali terlalu lambat untuk mengimbangi perkembangan teknologi dan peluang bisnis yang berkembang pesat.
- **Masalah Privasi Dan Keamanan**: Pengumpulan data melintasi batas administratif tanpa berdasarkan permintaan merupakan tantangan nyata.
- **Keahlian Big Data**: Selain tenaga teknis, pengetahuan mengenai potensi big data juga kurang dimiliki oleh orang-orang yang berorientasi bisnis.

Keuangan dan Asuransi

Seperti yang akan dibahas dalam Bab. 12 sektor keuangan dan asuransi adalah contoh paling jelas dari industri berbasis data. Big data mewakili peluang unik bagi sebagian besar organisasi perbankan dan jasa keuangan untuk memanfaatkan data pelanggan mereka guna mentransformasikan bisnis mereka, mewujudkan peluang pendapatan baru, mengelola risiko, dan mengatasi loyalitas pelanggan. Namun, serupa dengan teknologi baru lainnya, big data pasti menciptakan tantangan baru dan gangguan data bagi industri yang sudah dihadapkan pada persyaratan tata kelola, keamanan, dan peraturan, serta tuntutan dari basis pelanggan yang semakin sadar akan privasi.

Saat ini, tidak semua perusahaan pembiayaan siap menerima big data, infrastruktur informasi lama, dan faktor organisasi yang menjadi hambatan terbesar dalam penerapan big data di sektor ini. Penerapan solusi big data harus selaras dengan tujuan bisnis agar adopsi teknologi berhasil guna mengembalikan nilai bisnis maksimal.

Energi dan Transportasi

Bab 13 mengkaji sektor-sektor energi dan transportasi yang dari sudut pandang infrastruktur, serta dari sudut pandang efisiensi sumber daya dan kualitas hidup, sangat penting bagi Eropa. Infrastruktur fisik yang berkualitas tinggi dan daya saing global para pemangku kepentingan perlu dipertahankan sehubungan dengan transformasi digital dan inovasi berbasis data besar.

Analisis terhadap sumber data yang tersedia di bidang energi serta kasus penggunaannya dalam berbagai kategori untuk nilai big data: efisiensi operasional, pengalaman pelanggan, dan model bisnis baru memperjelas bahwa pemanfaatan teknologi big data yang sudah ada saja sudah cukup. Adaptasi spesifik domain dan perangkat diperlukan untuk digunakan dalam sistem cyber-fisik minyak, gas, listrik, dan transportasi. Inovasi mengenai privasi dan kerahasiaan, menjaga pengelolaan dan analisis data merupakan perhatian utama semua pemangku kepentingan energi dan transportasi yang menangani data pelanggan, baik itu bisnis-ke-konsumen atau bisnis-ke-bisnis. Tanpa memenuhi kebutuhan akan privasi dan kerahasiaan, akan selalu ada ketidakpastian seputar peraturan dan penerimaan pelanggan terhadap penawaran baru berbasis data.

Meningkatnya kecerdasan yang tertanam dalam infrastruktur akan memungkinkan analisis data "di lapangan" untuk menghasilkan "data cerdas". Hal ini nampaknya perlu, karena analisis yang terlibat akan memerlukan algoritma yang jauh lebih rumit dibandingkan sektor lain seperti ritel. Selain itu, taruhannya juga sangat tinggi karena peluang optimalisasi berada pada infrastruktur penting.

Media dan Hiburan

Industri media dan hiburan sering kali menjadi yang terdepan dalam mengadopsi teknologi baru. Bab 14 merinci permasalahan bisnis utama yang mendorong perusahaan media untuk melihat inovasi berbasis data besar sebagai cara untuk mengurangi biaya operasional dalam lanskap yang semakin kompetitif, dan pada saat yang sama, kebutuhan untuk meningkatkan pendapatan dari penyampaian konten. Menerbitkan surat kabar atau menyiarkan program televisi saja tidak lagi cukup operator masa kini harus meningkatkan nilai aset mereka di setiap tahap siklus hidup data.

Pemutar media juga lebih terhubung dengan pelanggan dan pesaing mereka dibandingkan sebelumnya berkat dampak disintermediasi, konten dapat dibuat, dibagikan, dikurasi, dan diterbitkan ulang oleh siapa saja. Ini berarti bahwa kemampuan teknologi big data untuk menyerap dan memproses berbagai sumber data, dan jika diperlukan bahkan secara real-time, merupakan aset berharga yang siap diinvestasikan oleh perusahaan.

Seperti halnya industri telekomunikasi, aspek hukum dan peraturan dalam beroperasi di Eropa tidak dapat diabaikan. Sebagai salah satu contoh, penting bahwa meskipun secara teknis dimungkinkan untuk mengumpulkan sejumlah besar detail tentang pelanggan dari penggunaan layanan mereka, interaksi pusat panggilan, pembaruan media sosial, dan sebagainya, bukan berarti hal tersebut etis tanpa transparan tentang bagaimana data akan digunakan. Eropa memiliki peraturan perlindungan data yang jauh lebih ketat dibandingkan Amerika Serikat, yang berarti privasi individu dan daya saing global perlu diseimbangkan.

Telekomunikasi

Sektor telekomunikasi nampaknya yakin akan potensi teknologi big data. Kombinasi manfaat dalam pemasaran dan manajemen penawaran, hubungan pelanggan, penerapan layanan, dan operasional dapat diringkas sebagai pencapaian keunggulan operasional bagi para pemain telekomunikasi.

Ada sejumlah platform komersial khusus telekomunikasi big data yang tersedia di pasar yang menyediakan dasbor, laporan untuk membantu proses pengambilan keputusan, dan dapat diintegrasikan dengan sistem pendukung bisnis (BSS). Aktuasi otomatis pada jaringan sebagai hasil analisis belum datang. Selain platform-platform ini, Data as a Service (DaaS) adalah tren yang diikuti oleh beberapa operator, yang terdiri dari memberikan wawasan analitis kepada perusahaan dan organisasi sektor publik yang memungkinkan pihak ketiga menjadi lebih efektif.

Faktor lain yang sangat penting dalam sektor ini terkait dengan kebijakan. Kerangka Connected Continent, yang bertujuan untuk memberikan manfaat bagi pelanggan dan mendorong terciptanya infrastruktur yang dibutuhkan Eropa untuk menjadi komunitas yang terhubung, pada pandangan pertama, kemungkinan besar akan menghasilkan peraturan yang lebih ketat bagi para pemain telekomunikasi. Kerangka kerja yang jelas dan stabil sangat penting untuk mendorong investasi di bidang teknologi, termasuk solusi big data.

Ritel

Sektor ritel akan bergantung pada pengumpulan data di dalam toko, data produk, dan data pelanggan. Agar sukses di masa depan, pengecer harus memiliki kemampuan untuk mengekstrak informasi yang benar dari kumpulan data besar yang diperoleh di lingkungan ritel yang terinstrumentasi secara real-time. Kecerdasan bisnis yang ada untuk analisis ritel harus ditata ulang untuk memahami perilaku pelanggan dan untuk mampu membangun alat rekomendasi yang lebih peka konteks, berorientasi pada konsumen dan tugas untuk pemasaran dialog pengecer-konsumen.

Manufaktur

Persyaratan inti di sektor manufaktur adalah penyesuaian produk dan produksi ukuran lot satu integrasi produksi dalam Jaringan nilai produk yang lebih besar, dan pengembangan produk cerdas.

Industri manufaktur sedang mengalami perubahan radikal dengan diperkenalkannya teknologi IT dalam skala besar. Perkembangan di bawah “Industri 4.0” mencakup peningkatan jumlah sensor dan konektivitas di seluruh aspek proses produksi. Oleh karena itu, akuisisi data berkaitan dengan membuat data yang sudah tersedia dapat dikelola, yaitu standardisasi dan integrasi data merupakan persyaratan terbesar. Analisis data sudah diterapkan dalam aplikasi intra-mural dan akan diperlukan untuk aplikasi yang lebih terintegrasi yang mencakup Jaringan logistik lengkap di seluruh pabrik dalam Jaringan produksi dan bahkan hingga penggunaan produk (pintar) pasca-penjualan. Perencanaan produksi perlu didukung oleh simulasi berbasis data dari lingkungan yang lengkap tersebut.

Mesin yang kompleks dan cerdas, misalnya mesin pesawat terbang, bisa mendapatkan keuntungan dari pemeliharaan prediktif berbasis data besar di mana sensor dan informasi konteks digunakan dengan algoritma pembelajaran mesin untuk menghindari pemeliharaan yang tidak perlu dan untuk menjadwalkan perbaikan pelindung ketika diperkirakan terjadi kegagalan. Mengingat biaya infrastruktur tambahan, produsen menggunakan model bisnis baru dimana mesin disewakan dan tidak dijual; dan pada gilirannya, data dan layanan sensor

dimiliki dan dilaksanakan oleh pabrikan dan bukan oleh pengguna mesin. Hal ini menimbulkan tantangan dalam peraturan dan kontrak mengenai kepemilikan data.

Sektor manufaktur Eropa dapat menjadi pemimpin pasar yang menggunakan big data dalam konteks Industri 4.0, sekaligus menjadi pasar terdepan, di mana big data manufaktur terintegrasi dalam Jaringan nilai produk yang lebih besar dan produk-produk cerdas dapat dimanfaatkan.

9.4 PEMBAHASAN DAN ANALISIS

Analisis terhadap temuan-temuan utama di berbagai sektor menunjukkan bahwa penting untuk membedakan perspektif teknis dari perspektif bisnis. Dari perspektif teknologi, penerapan big data mewakili sebuah langkah evolusi. Teknologi data besar, seperti jaringan terdesentralisasi dan komputasi terdistribusi untuk penyimpanan data terukur dan analisis data terukur, teknologi semantik dan ontologi, pembelajaran mesin, pemrosesan bahasa alami, dan teknik penambangan data lainnya telah menjadi fokus proyek penelitian selama bertahun-tahun. Kini teknik-teknik ini digabungkan dan diperluas untuk mengatasi tantangan teknis yang dihadapi dalam paradigma big data.

Ketika dianalisis dari perspektif bisnis, menjadi jelas bahwa penerapan big data memiliki dampak revolusioner bahkan terkadang disruptif terhadap praktik bisnis seperti biasa yang ada di industri. Jika dipikirkan matang-matang: muncul pemain baru yang lebih cocok untuk menawarkan layanan berdasarkan data massal. Proses bisnis yang mendasarinya berubah secara mendasar. Misalnya dalam bidang layanan kesehatan, teknologi big data dapat digunakan untuk menghasilkan wawasan baru tentang efektivitas pengobatan dan pengetahuan ini dapat digunakan untuk meningkatkan kualitas layanan. Namun, untuk mendapatkan manfaat dari penerapan big data ini, industri memerlukan model penggantian biaya baru yang menghargai kualitas dibandingkan kuantitas perawatan. Perubahan serupa juga diperlukan dalam industri energi: data penggunaan energi dari pengguna akhir akan memberikan manfaat bagi banyak pemangku kepentingan seperti pengecer energi, operator jaringan distribusi, dan pemain baru seperti penyedia dan agregator respons permintaan, serta penyedia layanan efisiensi energi. Namun siapa yang akan berinvestasi pada teknologi yang dapat memanfaatkan data energi? Dibutuhkan jaringan nilai bisnis yang partisipatif dan bukan Jaringan nilai yang statis.

Di semua industri, 3 Vs data besar, volume, kecepatan, dan variasi, memiliki relevansi. Selain itu, sektor-sektor industri yang telah meninjau diri mereka sendiri sehubungan dengan era big data menambahkan V lebih lanjut untuk mencerminkan aspek-aspek spesifik sektoral dan untuk menyesuaikan paradigma big data dengan kebutuhan khusus mereka. Banyak dari ekstensi tersebut, seperti privasi data, kualitas data, kerahasiaan data, dll., mengatasi tantangan tata kelola data, sementara ekstensi lainnya, seperti nilai, mengatasi fakta bahwa potensi nilai bisnis dari aplikasi big data masih belum dieksplorasi dan mungkin tidak dipahami dengan baik dalam sektor ini.

Di semua sektor industri, menjadi jelas bahwa bukan ketersediaan teknologi, namun kurangnya kasus bisnis dan model bisnis yang menghambat implementasi big data. Biasanya,

kasus bisnis perlu didefinisikan dengan jelas dan meyakinkan sebelum investasi dilakukan pada aplikasi baru. Namun, dalam konteks penerapan big data, pengembangan kasus bisnis yang konkrit merupakan tugas yang sangat menantang. Hal ini disebabkan oleh dua alasan. Pertama, karena dampak penerapan big data bergantung pada agregasi tidak hanya satu tapi juga berbagai macam sumber data heterogen di luar batas-batas organisasi, maka diperlukan kerja sama yang efektif dari berbagai pemangku kepentingan dengan potensi kepentingan yang berbeda atau awalnya ortogonal. Oleh karena itu, kepentingan dan kendala masing-masing pemangku kepentingan yang seringkali merupakan target yang berubah-ubah perlu tercermin dalam kasus bisnis. Kedua, pendekatan yang ada untuk mengembangkan model bisnis dan kasus bisnis biasanya berfokus pada organisasi tunggal dan tidak memberikan panduan untuk jaringan nilai dinamis dari banyak pemangku kepentingan dalam pasar tunggal digital.

Kesimpulan dan Rekomendasi

Inovasi berbasis data berpotensi memberikan dampak pada semua sektor perekonomian. Namun untuk mewujudkan hal ini, calon pembuat kebijakan perlu mengembangkan kebijakan yang koheren dalam penggunaan data. Hal ini dapat dicapai dengan: (1) mendukung pendidikan yang berfokus pada keterampilan ilmu data, (2) menghilangkan hambatan untuk menciptakan pasar tunggal digital, (3) menstimulasi lingkungan investasi yang diperlukan untuk teknologi big data, (4) membuat data publik dapat diakses melalui data terbuka dan menghilangkan silo data, (5) menyediakan infrastruktur teknis yang kompetitif, dan (6) mendorong peraturan perundang-undangan yang seimbang, dan pada saat yang sama, kebijakan harus mengatasi permasalahan seperti privasi dan keamanan, kepemilikan dan transfer, serta infrastruktur. dan data kewarganegaraan (Hemerly 2013). Sehubungan dengan hal ini, ada seruan untuk membentuk magna carta data guna menjawab pertanyaan tentang bagaimana teknologi big data dapat memfasilitasi diskriminasi dan marginalisasi; bagaimana memastikan bahwa kontrak antara individu dan perusahaan big data atau pemerintah bersifat adil; dan di mana menempatkan tanggung jawab atas keamanan data (Insight Center for Data Analytics 2015). Menurut pendapat kami, kemajuan lebih lanjut dan berkelanjutan dalam inovasi berbasis data besar bergantung pada tindakan pemerintah yang bekerja sama dengan pemangku kepentingan utama lainnya dalam mengembangkan lingkungan kebijakan dan peraturan yang tepat berdasarkan bukti empiris dari penelitian sistematis seputar beberapa pertanyaan yang diajukan di atas.

BAB 10
BIG DATA DI BIDANG KESEHATAN

10.1 PENDAHULUAN

Beberapa perkembangan di bidang layanan kesehatan, seperti meningkatnya biaya layanan kesehatan, meningkatnya kebutuhan akan cakupan layanan kesehatan, dan perubahan tren penggantian biaya penyedia layanan kesehatan, memicu permintaan akan teknologi big data untuk meningkatkan efisiensi dan kualitas pemberian layanan secara keseluruhan. Misalnya, Studi McKinsey Company (2011) menunjukkan dampak finansial yang tinggi dari penerapan big data di bidang layanan kesehatan, yang bernilai sekitar Rp.300 miliar per tahun hanya di AS. Angka mengesankan serupa juga diberikan oleh IBM: dalam Laporan Eksekutif IBM Global Business Services (Korster dan Seider 2010), penulis menggambarkan sistem layanan kesehatan sebagai sistem yang sangat tidak efisien, yaitu sekitar Rp. 2,5 triliun terbuang sia-sia setiap tahunnya dan efisiensi dapat ditingkatkan sebesar 35%. Dibandingkan dengan industri lain, hal ini merupakan peluang terbesar untuk peningkatan efisiensi. Selain itu, pemain-pemain besar berinvestasi di pasar obat-obatan yang sedang berkembang untuk populasi lanjut usia, misalnya Google mendirikan perusahaan baru Calico untuk mengatasi masalah kesehatan terkait usia. Kesimpulannya, penerapan big data dalam layanan kesehatan memiliki potensi dan peluang masa depan yang tinggi.

Namun, sejauh pengetahuan kami, saat ini hanya ada sedikit skenario penerapan berbasis data besar yang dapat ditemukan. Meskipun aplikasi analisis layanan kesehatan yang tidak canggih seperti analisis untuk peningkatan akuntansi, pengendalian kualitas, atau penelitian klinis tersedia secara luas, aplikasi ini tidak memanfaatkan potensi teknologi data besar. Hal ini terutama disebabkan oleh fakta bahwa data kesehatan tidak dapat diakses dengan mudah. Diperlukan investasi dan upaya yang besar untuk memungkinkan pengelolaan data kesehatan yang efisien dan akses data kesehatan yang lancar sebagai landasan bagi aplikasi data besar. Akibatnya, kasus-kasus bisnis yang meyakinkan sulit diidentifikasi karena beban investasi awal sangat mengurangi ekspektasi keuntungan. Dengan kata lain, salah satu tantangan terbesar dalam bidang layanan kesehatan untuk merealisasikan penerapan big data adalah kenyataan bahwa diperlukan investasi, standar, dan kerangka kerja yang tinggi serta teknologi pendukung baru agar data kesehatan tersedia untuk keperluan selanjutnya. aplikasi analisis data besar. Oleh karena itu, pengelolaan dan integrasi data kesehatan yang efisien merupakan persyaratan utama untuk aplikasi big data di bidang layanan kesehatan yang perlu ditangani.

Investigasi (Zillner dkk. 2014a, b) dalam bab ini menemukan bahwa dampak terbesar dari penerapan big data di bidang layanan kesehatan diperkirakan akan terjadi ketika kita tidak hanya dapat mengandalkan satu sumber data saja, namun berbagai sumber data sehingga aspek-aspek yang berbeda dari data tersebut dapat diandalkan. berbagai domain dapat dihubungkan. Oleh karena itu, ketersediaan dan integrasi semua sumber data kesehatan

terkait, seperti data klinis, klaim, data biaya dan administrasi, data farmasi dan penelitian, data pemantauan pasien, serta data kesehatan di web, merupakan hal yang sangat relevan.

Data kesehatan merupakan salah satu bentuk "Big data" bukan hanya karena volumenya yang besar namun juga karena kompleksitas, keragaman, dan ketepatan waktunya. Meskipun data terstruktur dalam jumlah besar sudah tersedia saat ini, volume data tidak terstruktur, seperti data biometrik, laporan teks, dan gambar medis, akan melampaui seluruh kebutuhan volume data. Hal ini erat kaitannya dengan tantangan penanganan data kesehatan yang sangat beragam, yaitu tidak hanya data yang sangat heterogen, seperti gambar, laporan terstruktur, catatan tidak terstruktur, dan lain-lain, memerlukan bentuk pemrosesan (pra-) baru tetapi juga semantik. berbagai domainnya, seperti keuangan, administrasi, penelitian, kesehatan pasien atau masyarakat, perlu dicerminkan. Nilai aplikasi big data bergantung pada identifikasi kasus bisnis yang meyakinkan. Karena dampak dan keberhasilan kasus-kasus bisnis layanan kesehatan bergantung pada kerja sama berbagai pemangku kepentingan yang sering kali mempunyai kepentingan berbeda, hal ini menjadi sulit untuk diidentifikasi.

10.2 ANALISIS KEBUTUHAN INDUSTRI BIDANG KESEHATAN

Wawancara dan investigasi pada bagian ini menunjukkan bahwa tuntutan tingkat tinggi akan peningkatan efisiensi dan kualitas layanan kesehatan saat ini sering kali dipandang bertentangan. Mayoritas layanan kesehatan berkualitas tinggi bergantung pada analisis data dan konten dalam jumlah besar. Hal ini secara otomatis menyebabkan peningkatan biaya perawatan mengingat sarana untuk analisis data otomatis, seperti teknologi big data, masih belum ada. Namun, dengan analisis data besar, pasien dapat disegmentasikan ke dalam beberapa kelompok dan selanjutnya menentukan perbedaan antar kelompok pasien. Daripada menanyakan pertanyaan "Apakah pengobatan ini efektif?", kita bisa menjawab pertanyaan "Untuk pasien manakah pengobatan ini efektif?" Peralihan dari layanan kesehatan berbasis rata-rata ke layanan kesehatan individual mempunyai potensi untuk meningkatkan kualitas layanan secara keseluruhan dengan cara yang efisien. Oleh karena itu, informasi apa pun yang dapat membantu meningkatkan kualitas dan efisiensi layanan kesehatan pada saat yang sama dianggap paling relevan dan berguna. Wawasan berdampak tinggi hanya dapat diwujudkan jika analisis data dilakukan pada kumpulan data heterogen yang mencakup data dari domain klinis, administratif, keuangan, dan publik. Hal ini mengharuskan berbagai pemangku kepentingan yang memiliki data tersebut bersedia membagikan aset datanya. Namun, ada persaingan yang kuat antara pemangku kepentingan yang terlibat dalam industri kesehatan. Ini adalah persaingan untuk mendapatkan sumber daya dan sumber dayanya terbatas. Masing-masing pemangku kepentingan berfokus pada kepentingan keuangannya masing-masing, yang sering kali menghasilkan keputusan pengobatan yang tidak optimal. Akibatnya, pasienlah yang paling menderita saat ini. Kepentingan dan peran berbagai kelompok pemangku kepentingan dapat diringkas sebagai berikut:

- Pasien mempunyai minat terhadap layanan kesehatan yang terjangkau, berkualitas tinggi, dan luas. Saat ini, data mengenai kondisi kesehatan pasien yang tersedia sangat

terbatas dan peluang pasien untuk terlibat aktif dalam proses tersebut sangat terbatas.

- Operator rumah sakit berusaha mengoptimalkan pendapatan mereka dari perawatan medis, yaitu mereka mempunyai kepentingan yang kuat dalam peningkatan efisiensi layanan, seperti rutinitas akuntansi otomatis, peningkatan proses, atau peningkatan pemanfaatan sumber daya.
- Dokter dan dokter tertarik pada proses rutin yang lebih otomatis dan tidak memerlukan banyak tenaga kerja, seperti tugas pengkodean, agar memiliki lebih banyak waktu untuk dan bersama pasien. Selain itu, mereka tertarik untuk mengakses data kesehatan yang dikumpulkan, dianalisis, dan disajikan secara ringkas yang memungkinkan pengambilan keputusan dan keputusan pengobatan berkualitas tinggi.
- Pembayar, seperti perusahaan asuransi kesehatan pemerintah atau swasta. Saat ini, sebagian besar sistem penggantian biaya yang ada mengelola pembayaran berbasis fee-for-service atau Diagnose-related Group (DRG) dengan menggunakan negosiasi TI sederhana dan proses pertukaran data antara pembayar dan penyedia layanan kesehatan dan tidak bergantung pada analisis data. Ketika pembayar memutuskan layanan kesehatan mana (yaitu pengobatan apa, diagnosis apa, atau tes pencegahan mana) yang akan ditanggung atau tidak, posisi dan pengaruh mereka mengenai penerapan pengobatan dan praktik inovatif cukup kuat. Namun, saat ini hanya tersedia data yang terbatas dan terfragmentasi mengenai efektivitas dan nilai layanan kesehatan; alasan cakupan pengobatan sering kali masih belum jelas dan terkadang terkesan sewenang-wenang.
- Farmasi, ilmu hayati, bioteknologi, dan penelitian klinis: Di sini penemuan pengetahuan baru menjadi perhatian dan fokus utama. Saat ini, berbagai domain yang disebutkan sebagian besar tidak terhubung dan menyelesaikan analisis datanya pada sumber data tunggal. Dengan mengintegrasikan sumber data yang heterogen dan terdistribusi, dampak solusi analitik data diperkirakan akan meningkat secara signifikan di masa depan.
- Penyedia produk medis tertarik untuk mengakses dan menganalisis data klinis untuk mempelajari kinerja produk mereka dibandingkan dengan produk pesaing guna meningkatkan pendapatan dan/atau meningkatkan posisi pasar mereka sendiri.

Untuk mengubah sistem layanan kesehatan saat ini menjadi sistem yang preventif, proaktif, dan berbasis nilai, diperlukan pertukaran dan pembagian data kesehatan yang lancar. Hal ini sekali lagi memerlukan kerja sama yang efektif antar pemangku kepentingan. Namun, saat ini kondisi layanan kesehatan sebagian besar ditentukan oleh insentif yang menghambat kerja sama. Untuk mendorong penerapan dan adaptasi aplikasi big data yang komprehensif di sektor layanan kesehatan, insentif dan peraturan mendasar yang menjelaskan kondisi dan kendala yang mendasari interaksi dan kerja sama berbagai pemangku kepentingan perlu diubah.

10.3 POTENSI PENERAPAN BIG DATA UNTUK KESEHATAN

Analisis sektor kesehatan (Zillner et al. 2014b) menunjukkan bahwa terdapat beberapa skenario penerapan big data yang bertujuan untuk menyelaraskan kebutuhan akan peningkatan kualitas, yang secara umum berarti peningkatan biaya layanan, dengan kebutuhan peningkatan efisiensi layanan. Hal yang umum pada semua aplikasi big data yang teridentifikasi adalah kenyataan bahwa semuanya memerlukan sarana untuk mendeskripsikan dan menyelaraskan berbagai sumber data yang heterogen secara semantik, sarana untuk memastikan kualitas data yang tinggi, sarana yang mengatasi privasi dan keamanan data, serta sarana untuk analisis data secara terintegrasi. kumpulan data.

Misalnya, aplikasi Analisis Kesehatan Masyarakat menunjukkan potensi peluang serta persyaratan teknis terkait dengan teknologi data besar. Aplikasi kesehatan masyarakat bergantung pada pengelolaan data kesehatan yang komprehensif dan longitudinal dari penyakit kronis (misalnya diabetes, gagal jantung kongestif) atau penyakit parah (misalnya kanker) dari populasi pasien tertentu untuk mengumpulkan dan menganalisis data pengobatan dan hasil. Pengetahuan yang diperoleh sangat berharga karena membantu mengurangi komplikasi, memperlambat perkembangan penyakit, serta meningkatkan hasil pengobatan. Misalnya, sejak tahun 1970 Swedia terus berinvestasi dalam inisiatif analitik kesehatan masyarakat yang menghasilkan 90 registrasi yang saat ini mencakup 90 % dari seluruh data pasien Swedia dengan karakteristik yang dipilih (beberapa bahkan mencakup data longitudinal) (Soderland dkk. 2012). Sebuah studi terkait (PricewaterhouseCoopers (2009)) menunjukkan bahwa Swedia memiliki hasil layanan kesehatan terbaik di Eropa berdasarkan rata-rata biaya layanan kesehatan (9 % dari produk domestik bruto (PDB)). Untuk mencapai hal ini, data kesehatan (yang disimpan secara terstruktur (misalnya laporan laboratorium) maupun data tidak terstruktur (misalnya laporan medis, gambar medis)) perlu diperkaya secara semantik (Pengayaan Data Semantik) untuk membuat semantik implisit. data kesehatan yang dapat dipahami oleh seluruh organisasi dan pemangku kepentingan yang terlibat. Selain itu, diperlukan infrastruktur umum dengan standar umum yang memungkinkan berbagi data tanpa hambatan (Berbagi Data) serta integrasi fisik berbagai sumber data ke dalam satu platform (Integrasi Data). Agar dapat mematuhi persyaratan keamanan data dan privasi yang tinggi yang diperlukan untuk melindungi sifat sensitif data kesehatan longitudinal, kerangka hukum umum serta sarana teknis untuk anonimisasi data (Keamanan dan Privasi Data) harus ada. Selain itu, untuk memastikan komparabilitas kumpulan data kesehatan, diperlukan proses yang memastikan kualitas data yang tinggi melalui dokumentasi standar serta analisis sistematis data kesehatan dan hasil dari populasi pasien tertentu (Kualitas Data).

Dalam hal penanganan data, skenario aplikasi lain yang teridentifikasi menghasilkan persyaratan teknis yang sangat mirip. Misalnya, aplikasi Penelitian Efektivitas Komparatif bertujuan untuk membandingkan efektivitas intervensi klinis dan finansial untuk meningkatkan efisiensi dan kualitas layanan perawatan klinis. Untuk mencapai hal ini, kumpulan data besar yang mencakup data klinis (informasi tentang karakteristik pasien), data keuangan (data biaya), dan data administratif (pengobatan dan layanan yang dicapai) dianalisis secara kritis untuk mengidentifikasi metode yang paling efektif dan juga secara

klinis. sebagai perawatan paling hemat biaya yang bekerja paling baik untuk pasien tertentu. Aplikasi Clinical Operation Intelligence bertujuan untuk mengidentifikasi pemborosan dalam proses klinis untuk mengoptimalkannya. Dengan menganalisis prosedur medis, peluang kinerja, seperti peningkatan proses klinis, penyesuaian, dan adaptasi pedoman klinis, dapat diwujudkan. Contoh lainnya adalah aplikasi Clinical Decision Support (CDS) yang berupaya meningkatkan efisiensi dan kualitas operasi perawatan dengan membantu dokter dan profesional layanan kesehatan dalam proses pengambilan keputusan dengan memungkinkan akses informasi yang bergantung pada konteks, dengan menyediakan informasi pra-diagnostik atau dengan memvalidasi dan mengoreksi data yang diberikan. Kategori skenario selanjutnya adalah penerapan yang menangani Penggunaan Data Kesehatan Sekunder yang mengandalkan agregasi, analisis, dan presentasi ringkas data klinis, keuangan, administratif, serta data kesehatan terkait lainnya untuk menemukan pengetahuan baru yang berharga, misalnya. Misalnya, untuk mengidentifikasi tren, memprediksi hasil, atau untuk mempengaruhi perawatan pasien, pengembangan obat, dan pilihan terapi. Terakhir, Aplikasi Keterlibatan Pasien berfokus pada pembentukan platform/portal pasien yang mendorong keterlibatan aktif pasien dalam proses layanan kesehatan. Aplikasi kesehatan apa pun yang berjalan di platform pasien mengandalkan integrasi data kesehatan episodik dari pengaturan klinis serta data non-episodik yang diambil oleh perangkat untuk memantau parameter terkait kesehatan, seperti aktivitas, pola makan, tidur, atau berat badan.

10.4 PENDORONG DAN KENDALA BIG DATA DI BIDANG KESEHATAN

Keberhasilan realisasi big data di bidang kesehatan mempunyai beberapa pendorong dan kendala.

Pengemudi

Penggerak berikut telah diidentifikasi untuk big data di sektor kesehatan:

- **Meningkatnya Volume Data Kesehatan Elektronik**: Dengan meningkatnya penerapan teknologi catatan kesehatan elektronik (EHR) (yang sudah terjadi di AS), dan kemajuan teknologi di bidang pengurutan generasi berikutnya dan segmentasi citra medis, semakin banyak pula data kesehatan elektronik yang tersedia. data kesehatan akan tersedia.
- **Perlunya Peningkatan Efisiensi Operasional**: Untuk mengatasi peningkatan jumlah pasien (populasi lanjut usia) dan mengurangi biaya layanan kesehatan yang sangat tinggi, diperlukan transparansi mengenai efisiensi operasional.
- **Pemberian Layanan Kesehatan Berbasis Nilai**: Layanan kesehatan berbasis nilai bergantung pada keselarasan pengobatan dan kesuksesan finansial. Untuk mendapatkan wawasan tentang korelasi antara efektivitas dan biaya perawatan, diperlukan solusi analisis data pada kumpulan data layanan kesehatan yang terintegrasi, heterogen, kompleks, dan besar.
- **Perundang-Undangan AS**: Reformasi Layanan Kesehatan AS, juga dikenal sebagai Obamacare, mendorong penerapan teknologi EHR serta analisis data kesehatan. Hal

ini mempunyai dampak yang signifikan terhadap pasar internasional untuk aplikasi data kesehatan yang besar.

- **Peningkatan Keterlibatan Pasien**: Aplikasi seperti "PatientsLikeMe" menunjukkan kesediaan pasien untuk terlibat aktif dalam proses layanan kesehatan.
- **Insentif Baru**: Sistem insentif yang ada saat ini menerapkan pengobatan yang "berkualitas tinggi" dan "jumlah yang banyak". Meskipun jelas bahwa tidak ada seorang pun yang mau membayar untuk perawatan yang tidak efektif, hal ini masih terjadi di banyak sistem medis. Untuk menghindari penggantian biaya yang berkualitas rendah, insentif sistem medis perlu diselaraskan dengan hasil yang diperoleh. Beberapa inisiatif, seperti Accountable Care Organizations (ACO) (Centers for Medicare and Medicaid Services 2010), atau Diagnose-related Groups (DRG) (Ma Ching-To Albert 1994), telah diterapkan untuk memberi penghargaan pada kualitas, bukan penghargaan pada kualitas. kuantitas perawatan.

Kendala

Kendala big data di bidang kesehatan dapat diringkas sebagai berikut:

- **Digitalisasi data kesehatan**: Hanya sebagian kecil data terkait kesehatan yang tersedia dalam format digital.
- **Kurangnya data kesehatan yang terstandarisasi**: Pembagian data yang lancar mengharuskan data kesehatan di seluruh rumah sakit dan pasien dikumpulkan dengan cara yang terstandarisasi dan terpadu.
- **Silo data**: Data layanan kesehatan sering kali disimpan dalam silo data terdistribusi, sehingga membuat analisis data menjadi rumit dan tidak stabil.
- **Isolasi organisasi**: Karena hilangnya insentif, kerja sama antar organisasi, dan kadang-kadang bahkan antar departemen dalam satu organisasi, jarang terjadi dan merupakan hal yang luar biasa.
- **Keamanan dan privasi data**: Kerangka hukum yang menjelaskan masalah dan strategi akses, keamanan, dan privasi data tidak ada, sehingga menghambat pembagian dan pertukaran data.
- **Investasi tinggi**: Mayoritas aplikasi big data di sektor layanan kesehatan bergantung pada ketersediaan data layanan kesehatan berskala besar, berkualitas tinggi, dan longitudinal. Pengumpulan dan pemeliharaan sumber data yang komprehensif tersebut tidak hanya memerlukan investasi yang besar, namun juga waktu (tahun) hingga kumpulan data tersebut cukup komprehensif untuk menghasilkan hasil analisis yang baik.
- **Kasus bisnis yang tidak ada dan model bisnis yang tidak jelas**: Teknologi inovatif apa pun yang tidak selaras dengan kasus bisnis yang konkrit, termasuk tanggung jawab terkait, kemungkinan besar akan gagal. Hal ini juga berlaku untuk solusi data besar. Oleh karena itu, keberhasilan penerapan solusi big data memerlukan transparansi tentang: (a) siapa yang membayar solusi tersebut, (b) siapa yang memperoleh manfaat dari solusi tersebut, dan (c) siapa yang mendorong solusi tersebut. Misalnya, penerapan solusi analitik data menggunakan data klinis memerlukan investasi dan

sumber daya yang tinggi untuk mengumpulkan dan menyimpan data pasien, misalnya melalui solusi catatan kesehatan elektronik (EHR). Meskipun sudah jelas bagaimana pemangku kepentingan yang terlibat dapat memperoleh manfaat dari kumpulan data yang dikumpulkan, masih belum jelas apakah pemangku kepentingan bersedia membiayai, atau mendorong, penerapan tersebut.

10.5 SUMBER DATA KESEHATAN YANG TERSEDIA

Sistem layanan kesehatan memiliki beberapa kumpulan data kesehatan utama yang disimpan oleh berbagai pemangku kepentingan/pihak:

1. Data klinis, yang dimiliki oleh penyedia layanan (seperti rumah sakit, pusat perawatan, dokter, dll.) dan mencakup informasi apa pun yang disimpan dalam sistem informasi rumah sakit klasik atau EHR, seperti rekam medis, gambar medis, hasil laboratorium, data genetik , dll.
2. Data klaim, biaya, dan administrasi, yang dimiliki oleh penyedia layanan dan pembayar dan mencakup kumpulan data apa pun yang relevan untuk masalah penggantian biaya, seperti pemanfaatan layanan, perkiraan biaya, klaim, dll.
3. Data penelitian, yang dimiliki oleh perusahaan farmasi, laboratorium penelitian/akademisi, dan pemerintah yang mencakup uji klinis, studi klinis, data populasi dan penyakit, dll.
4. Data pemantauan pasien, yang dimiliki oleh pasien atau produsen perangkat pemantauan dan mencakup segala informasi terkait perilaku dan preferensi pasien.
5. Data kesehatan di web: situs web seperti “PatientsLikeMe” menjadi semakin populer. Dengan secara sukarela berbagi data tentang penyakit langka atau pengalaman luar biasa mengenai penyakit umum, komunitas dan penggunanya menghasilkan sejumlah besar data kesehatan dengan konten yang berharga.

Peningkatan kualitas layanan dapat diatasi jika berbagai dimensi data kesehatan digabungkan dalam analisis data kesehatan otomatis. Dimensi data meliputi (a) data klinis yang menggambarkan status kesehatan dan riwayat pasien, (b) data administrasi dan proses klinis, (c) pengetahuan tentang penyakit serta data populasi terkait (yang dianalisis), dan (d) pengetahuan tentang perubahan. Jika analisis data dibatasi hanya pada satu dimensi data, misalnya data administratif dan keuangan, maka akan dimungkinkan untuk memperbaiki proses pengelolaan dan penggantian biaya yang telah ditetapkan; namun tidak mungkin untuk mengidentifikasi standar baru untuk perawatan individual. Oleh karena itu, dampak klinis terbesar dari pendekatan big data pada bidang layanan kesehatan dapat dicapai jika data dari empat dimensi dikumpulkan, dibandingkan, dan dihubungkan.

Karena setiap kumpulan data dimiliki oleh pemangku kepentingan/pihak yang berbeda, data di bidang kesehatan menjadi sangat terfragmentasi. Namun, integrasi berbagai kumpulan data yang heterogen merupakan prasyarat penting bagi penerapan data kesehatan yang besar dan memerlukan keterlibatan dan interaksi yang efektif dari berbagai pemangku kepentingan. Oleh karena itu, diperlukan insentif sistem yang memadai, yang mendukung kelancaran pembagian dan pertukaran data kesehatan.

10.6 PERSYARATAN SEKTOR KESEHATAN

Forum Sektoral Layanan Kesehatan mampu mengidentifikasi dan menyebutkan beberapa persyaratan yang perlu dipenuhi oleh penerapan big data di bidang layanan kesehatan. Berikut ini akan dibedakan persyaratan non teknis dan teknis.

Persyaratan Non-teknis

Persyaratan terkait bisnis disebut persyaratan non-teknis dan mencakup prasyarat dan kebutuhan penting untuk penerapan data kesehatan besar, seperti kebutuhan akan investasi tinggi, insentif sistem berbasis nilai, atau kasus bisnis multi-pemangku kepentingan. Perlunya Investasi Tinggi Karena sifat big data kesehatan yang berskala besar, pengembangan dan pemeliharaan aplikasi big data di bidang layanan kesehatan serta kumpulan datanya sendiri memerlukan investasi yang besar. Aplikasi data kesehatan besar terutama bergantung pada data layanan kesehatan berskala besar, berkualitas tinggi, dan seringkali bersifat longitudinal, yang memerlukan pengumpulan data selama beberapa tahun untuk menghasilkan kumpulan data komprehensif yang dapat dianalisis untuk menghasilkan hasil yang akurat dan berwawasan luas. Investasi sebesar itu jarang dapat dibiayai oleh satu pihak saja, namun perlu melibatkan banyak pemangku kepentingan, yang akan mengarah langsung pada persyaratan non-teknis berikutnya.

Kasus Bisnis Multi-Pemangku Kepentingan Karena tingginya kebutuhan investasi yang dijelaskan di atas, seringkali penting bagi beberapa pemangku kepentingan untuk bekerja sama untuk menutupi biaya investasi. Di sini kepentingan para pemangku kepentingan seringkali berbeda. Persoalan penting lainnya adalah bahwa penerima manfaat utama dari sebuah solusi seringkali bukanlah mereka yang mampu atau bersedia membiayai solusi yang menyeluruh (misalnya pasien). Meskipun demikian, meskipun sudah terlihat jelas bagaimana para pemangku kepentingan yang terlibat dapat memperoleh manfaat dari solusi big data tertentu dengan kumpulan kumpulan data berkualitas tinggi, sering kali masih belum jelas apakah para pemangku kepentingan tersebut mampu atau bersedia untuk mendorong atau membayar solusi tersebut.

Perlunya Insentif Sistem Berbasis Nilai Untuk meningkatkan efektivitas perawatan medis, perlu untuk menghindari penggantian biaya yang berkualitas rendah. Artinya, situasi saat ini yang memerlukan pengobatan dalam jumlah besar dan bukan pengobatan berkualitas tinggi. Karena tidak ada seorang pun yang mau membayar untuk pengobatan yang tidak efektif, insentif sistem kesehatan harus selaras dengan hasil (misalnya sistem pembiayaan dan penggantian biaya berbasis kinerja) dan, sebagai tambahan, kerja sama antar pemangku kepentingan perlu dihargai.

Persyaratan Teknis

Persyaratan teknis adalah persyaratan yang berkaitan dengan teknologi tertentu. Hal tersebut mencakup pengayaan data semantik, integrasi dan pembagian data, privasi dan keamanan data, serta kualitas data. Prasyarat utama untuk aplikasi dan analisis big data adalah ketersediaan data dalam bentuk digital yang sesuai. Banyak teknologi tepat guna yang tersedia untuk memenuhi dan mendukung kebutuhan ini (misalnya pengenalan suara). Oleh karena itu, tidak ada penekanan pada digitalisasi data. Kurangnya data digital yang tepat

dalam layanan kesehatan sebagian besar disebabkan oleh terbatasnya penerapan pendekatan digitalisasi data dalam rutinitas sehari-hari dan alur kerja yang biasa dilakukan para dokter.

Pengayaan Data Semantik Seperti yang diperkirakan oleh lembaga riset pasar IDC, sekitar 90 % data kesehatan akan tersedia secara tidak terstruktur di tahun-tahun mendatang (Lu¨nendonk GmbH 2013). Untuk memfasilitasi dan menjamin kelancaran pemrosesan data tersebut, diperlukan pengayaan data semantik. Artinya, data kesehatan, seperti laporan medis, gambar, video, atau komunikasi, perlu diperkaya dengan apa yang disebut label semantik. Tantangan utama pengayaan data semantik adalah kemajuan teknologi perlu dicapai dengan analisis beberapa jenis data yang berbeda.

Integrasi dan Pembagian Data Untuk menghindari silo data atau kuburan data, big data harus diintegrasikan secara efisien dari berbagai sumber data yang berbeda dan dibagikan dengan lancar. Saat ini, adopsi teknologi untuk pertukaran data masih tertinggal di Eropa (Accenture 2012). Di Inggris, kurang dari 46% penyedia layanan kesehatan melakukan pertukaran informasi layanan kesehatan, dan di Jerman dan Perancis angka ini bahkan lebih rendah lagi (sekitar 25%). Persyaratan ini sejalan dengan kebutuhan akan data terstruktur atau diperkaya secara semantik agar data mudah diakses. Prasyarat utama untuk penelitian medis adalah kemungkinan untuk mengintegrasikan data dari berbagai sumber berbeda untuk memperoleh gambaran longitudinal mengenai riwayat pasien.

Keamanan dan Privasi Data Ketika berbicara tentang pemrosesan, pengintegrasian, atau pembagian data medis, penekanan yang kuat harus diberikan pada keamanan dan privasi data. Data medis dikategorikan sebagai data pribadi yang sangat sensitif dan oleh karena itu perlindungan dari akses tidak sah, manipulasi, atau kerusakan harus dijamin. Oleh karena itu, sifat big data mungkin mengabaikan pendekatan perlindungan privasi yang sudah ada (misalnya saat menggabungkan big data dari sumber data yang berbeda). Aplikasi data kesehatan yang besar perlu lebih fokus pada privasi dan keamanan data. Misalnya, anonimisasi dikenal sebagai pendekatan populer untuk menghilangkan identifikasi data pribadi terkait kesehatan. Dengan menggabungkan data besar dari berbagai sumber data yang berbeda, data yang dianonimkan dapat diidentifikasi ulang secara tidak sengaja. Oleh karena itu, metode peningkatan privasi yang ada perlu dievaluasi untuk mengetahui apakah metode tersebut dapat memenuhi semua persyaratan privasi bahkan ketika berhadapan dengan data besar. Jika privasi data tidak dapat dijamin dengan metode tertentu, metode ini perlu diadaptasi untuk memenuhi kebutuhan privasi atau metode dan pendekatan baru perlu dikembangkan. Terlepas dari tantangan teknis, kerangka hukum internasional bersama dengan pedoman perlu ditetapkan untuk memberikan landasan bersama bagi pertukaran internasional dan integrasi data besar yang berhubungan dengan kesehatan.

Kualitas Data Kumpulan data yang tersedia berkualitas tinggi merupakan prasyarat utama untuk aplikasi big data di bidang layanan kesehatan. Manfaat suatu aplikasi berkorelasi kuat dengan kualitas datanya. Dalam bidang layanan kesehatan, kualitas data yang tersedia seringkali tidak jelas. Frekuensi nilai yang hilang atau salah merupakan indikator kualitas data. Biasanya kualitas data meningkat ketika data ditangkap dan diproses menggunakan alat teknologi informasi (TI) berkualitas tinggi. Alat tersebut dapat diintegrasikan ke dalam

rutinitas kerja sehari-hari dan melakukan pemeriksaan kualitas data tertentu (misalnya pemeriksaan masuk akal) selama proses pengambilan atau pemasukan data. Untuk menghasilkan hasil yang berharga atau dukungan keputusan ketika menganalisis data kesehatan, aplikasi big data harus memenuhi standar kualitas yang tinggi.

10.7 PETA JALAN TEKNOLOGI BIG DATA DI BIDANG KESEHATAN

Peta jalan berikut menguraikan dan menjelaskan teknologi dan pertanyaan penelitian yang mendasarinya, yang memenuhi persyaratan yang ditentukan di bagian sebelumnya.

Pengayaan Data Semantik

Untuk memperkaya data medis secara semantik, diperlukan suatu kerangka kerja. Oleh karena itu, diperlukan teknik pengayaan semantik yang lebih dari sekadar ekstraksi informasi relevan dari teks tidak terstruktur atau gambar medis. Label semantik, yang mengungkapkan dan mendefinisikan makna informasi, membuat konten asli dapat diakses secara semantik serta dapat diproses dan dibaca mesin secara otomatis. Misalnya, prosedur medis dan entitas diagnosis dalam teks tidak terstruktur seperti laporan medis dikenali dan bagian penjelasannya dihubungkan. Oleh karena itu diperlukan teknik analisis teks yang canggih (Bretschneider et al. 2013). Selain itu, diperlukan kerangka pengayaan yang terstandarisasi, yang mendukung integrasi teknis. Untuk memfasilitasi dan meningkatkan pengayaan semantik data medis, diperlukan kemajuan dalam teknologi berikut:

- Ekstraksi informasi dari teks kedokteran menghadirkan tantangan baru terhadap teknik ekstraksi informasi klasik, karena negasi, temporalitas, dan fitur kontekstual lebih lanjut perlu diperhitungkan. Beberapa penelitian (Fan dan Friedman 2011; Savova et al. 2010) menunjukkan kemajuan menuju kebutuhan khusus dalam penguraian teks medis. Karena penelitian yang sedang berlangsung terutama berfokus pada teks klinis dalam bahasa Inggris, diperlukan adaptasi ke bahasa Eropa lainnya.
- Algoritme pemahaman gambar yang secara formal menangkap informasi gambar yang terdeteksi secara otomatis, seperti struktur anatomi, struktur abnormal, dan anotasi gambar semantik, diperlukan. Oleh karena itu, diperlukan penelitian tambahan yang menargetkan dan mempertimbangkan kompleksitas tubuh manusia serta teknologi pencitraan medis yang berbeda.
- Kerangka kerja anotasi medis terstandar yang mencakup pemrosesan teks medis terstandarisasi dan mendukung integrasi teknis teknologi anotasi. Meskipun terdapat beberapa kerangka kerja yang tersedia (misalnya UIMA), adaptasi diperlukan untuk memenuhi tantangan dan persyaratan spesifik dalam bidang layanan kesehatan.

Berbagi dan Integrasi Data

Integrasi data yang efisien dan pembagian yang lancar bergantung pada skema dan terminologi pengkodean standar serta model data. Sistem pengkodean yang terstandarisasi saat ini digunakan untuk pengkodean informasi tingkat tinggi (misalnya penyakit, nilai laboratorium, obat-obatan) atau tidak digunakan secara internasional. Banyak informasi yang tidak tersedia dalam format kode sama sekali. Untuk penggunaan model data standar, Model Informasi Referensi HL (RIM) dianggap menjadi model data standar untuk implementasi EHR.

Namun demikian, sebagian besar penyedia teknologi masih mengandalkan model data mereka sendiri dalam hal integrasi data. Untuk memajukan integrasi dan pembagian data, skema pengkodean serta model data perlu ditingkatkan dan distandarisasi.

1. Model data semantik memungkinkan representasi data yang tidak ambigu. Model yang ada (misalnya RIM HL) memiliki beberapa masalah yang membuatnya sulit untuk diterapkan. Kegiatan penelitian lebih lanjut, seperti Model for Clinical Information (MCI) (Oberkampf et al. 2013) yang mengintegrasikan model pasien berdasarkan ontologi, sedang berlangsung.
2. Model pengetahuan semantik seperti ontologi dan terminologi domain biomedis digunakan dalam kombinasi dengan model data semantik dan membantu memfasilitasi interoperabilitas semantik. Ada beberapa model berbeda (misalnya SNOMED CT) yang tersedia, namun diperlukan penelitian lebih lanjut untuk meningkatkan standar ini, serta mengembangkan standar baru.
3. Informasi konteks diperlukan untuk memberikan informasi tentang asal, penggunaan, atau kepemilikan data. Oleh karena itu diperlukan standar untuk mendeskripsikan informasi konteks.

Privasi dan Keamanan Data

Untuk memenuhi tingginya permintaan akan privasi dan keamanan data kesehatan yang besar, berbagai aspek perlu dipertimbangkan. Selain undang-undang perlindungan data nasional, kerangka hukum bersama untuk Uni Eropa juga diperlukan untuk memfasilitasi pendekatan atau kerja sama internasional. Ketika berbicara tentang privasi dan keamanan data kesehatan yang besar, seringkali diperlukan identifikasi ulang pasien (misalnya untuk menilai status kesehatan pasien secara longitudinal). Agregasi data dari berbagai sumber data menghadirkan dua tantangan besar bagi privasi dan keamanan big data. Pertama, pengumpulan data dari sumber data yang heterogen sulit dilakukan dan data untuk pasien harus diselaraskan dengan benar. Sifat data besar juga dapat mengabaikan metode peningkatan privasi tertentu ketika menggabungkan data dari berbagai sumber data yang berbeda. Oleh karena itu diperlukan kemajuan untuk teknologi berikut:

- Algoritma hash sering digunakan sebagai metode enkripsi. Fungsi satu arahnya juga dapat digunakan untuk menghasilkan pengenal semu dan karenanya memfasilitasi nama samaran yang aman. Namun algoritma hash harus kuat dan tahan benturan.
- Pertukaran data yang aman antar lembaga dan negara sangat penting untuk mewujudkan beberapa visi menarik dalam bidang layanan kesehatan (misalnya EHR internasional). Oleh karena itu, profil Integrating the Healthcare Enterprise (IHE)6 banyak digunakan (misalnya berbagi dokumen lintas perusahaan IHE) meskipun profil tersebut masih menjadi fokus kegiatan penelitian.
- Algoritma de-identifikasi, seperti anonimisasi atau pseudonimisasi, perlu ditingkatkan untuk menjamin privasi data bahkan ketika menggabungkan data besar dari sumber data yang berbeda. K-anonymity (El Emam dan Dankar 2008) adalah pendekatan menjanjikan yang bertujuan untuk memastikan anonimitas bahkan dalam konteks big data.

Kualitas Data

Kualitas data yang baik merupakan faktor kunci dalam aplikasi data kesehatan yang besar. Hal ini bergantung pada empat aspek berbeda: (1) kualitas data dari sumber data asli, (2) cakupan dan tingkat detail data yang dikumpulkan, (3) semantik umum seperti yang dijelaskan sebelumnya, dan (4) penanganan media. Untuk meningkatkan kualitas data keempat aspek tersebut, diperlukan kemajuan teknologi sebagai berikut:

- Peningkatan pengelolaan asal usul diperlukan untuk memungkinkan kurasi data kesehatan yang andal. Oleh karena itu, mekanisme manajemen kepercayaan dan izin di tingkat data perlu diterapkan.
- Teknologi interaksi manusia-data [misalnya antarmuka bahasa alami, formulasi kueri skema-agnostik (Freitas dan Curry 2014)] meningkatkan kualitas data karena memfasilitasi interaksi kemudahan penggunaan yang terintegrasi sempurna dalam alur kerja tertentu.
- Pendekatan ekstraksi informasi yang andal diperlukan untuk memfasilitasi pemrosesan data medis yang tidak terstruktur (misalnya laporan medis, gambar medis). Oleh karena itu, pendekatan yang ada (misalnya pemrosesan bahasa alami) harus ditingkatkan guna mengatasi karakteristik spesifik informasi dan data kesehatan.

Pengembangan peta jalan biasanya dilakukan untuk satu perusahaan. Terdapat kebutuhan untuk mengembangkan peta jalan untuk pasar Eropa yang bergantung pada (a) sejauh mana persyaratan non-teknis akan dipenuhi dan (b) sejauh mana organisasi-organisasi Eropa bersedia berinvestasi dalam pengembangan dan penggunaan big data. implementasi kasus. Oleh karena itu, tidak mungkin untuk memberikan garis waktu yang pasti mengenai pencapaian teknologi, namun perkiraan garis waktunya dapat dilihat pada Tabel 10.1.

Kesimpulan dan Rekomendasi Bidang Kesehatan

Teknologi big data dan analisis data kesehatan menyediakan sarana untuk mengatasi tantangan efisiensi dan kualitas di bidang kesehatan. Misalnya, dengan mengumpulkan dan menganalisis data kesehatan dari berbagai sumber, seperti data klinis, keuangan, dan administratif, hasil pengobatan terkait pemanfaatan sumber daya dapat dipantau. Analisis ini pada gilirannya membantu meningkatkan efisiensi perawatan. Selain itu, identifikasi pasien berisiko tinggi dengan model prediktif mengarah pada perawatan pasien proaktif yang memungkinkan pemberian layanan berkualitas tinggi.

Tabel 10.1 Jangka waktu hasil utama yang diharapkan pada sektor kesehatan

PERSYARATAN TEKNIS	TAHUN 1	TAHUN 2	TAHUN 3	TAHUN 4	TAHUN 5
Pengayaan data		Format dan antarmuka standar untuk modul anotasi	Algoritma ekstraksi informasi berbasis pengetahuan	Algoritma untuk deteksi anomali pada gambar. Teknologi pengayaan data tersedia untuk sejumlah besar jenis teks berbeda dan berbagai bahasa	Definisi dan implementasi kerangka anotasi medis

Integrasi data	Representasi konteks untuk repositori data	Model dan terminologi pengetahuan semantik yang selaras	Model data semantik umum untuk data pasien terstruktur	Model data semantik umum untuk data pasien tidak terstruktur	Representasi konteks untuk semua data pasien
Keamanan dan privasi data		Profil IHE untuk pertukaran data yang aman		Peningkatan privasi melalui algoritma hash	Pendekatan anonimisasi, nama samaran, dan k-anonimitas untuk data besar
Kualitas data	Metode untuk manajemen kepercayaan dan izin		UI bahasa alami dan kueri skema agnostik	Alur kerja terintegrasi untuk manajemen kepercayaan dan izin	Integrasi data tidak terstruktur yang sadar konteks

Analisis yang komprehensif terhadap kebutuhan dan persyaratan domain menunjukkan bahwa dampak terbesar dari penerapan big data di domain layanan kesehatan dapat dicapai ketika dimungkinkan untuk tidak hanya memperoleh data dari satu sumber, namun dari berbagai sumber data sehingga berbagai aspek dapat digabungkan untuk memperoleh manfaat. wawasan baru. Oleh karena itu, ketersediaan dan integrasi semua sumber data kesehatan terkait, seperti data klinis, klaim, data biaya dan administrasi, data farmasi dan penelitian dan pengembangan, perilaku pasien, dan data sentimen serta data kesehatan di web, sangatlah penting. relevansi.

Namun, akses terhadap data kesehatan saat ini hanya dapat dilakukan dengan cara yang sangat terbatas. Untuk memungkinkan akses yang lancar terhadap data layanan kesehatan, beberapa persyaratan teknis perlu dipenuhi, termasuk (1) konten data kesehatan yang tidak terstruktur (seperti gambar atau laporan) disempurnakan dengan anotasi semantik; (2) silo data diatasi melalui teknologi efisien untuk berbagi dan pertukaran data semantik; (3) sarana teknis yang didukung oleh kerangka hukum memastikan pembagian dan pertukaran data kesehatan diatur; dan (4) tersedia teknik untuk menilai dan meningkatkan kualitas data.

Ketersediaan teknologi tidak akan cukup untuk mendorong adopsi big data di bidang layanan kesehatan. Batu sandungan utama adalah kurangnya kasus bisnis dan model bisnis. Ketika big data menumbuhkan dimensi proposisi nilai baru dalam pemberian layanan kesehatan, dengan wawasan tentang efektivitas pengobatan untuk meningkatkan kualitas layanan secara signifikan, diperlukan model penggantian biaya baru yang menghargai kualitas, bukan kuantitas perawatan.

BAB 11
BIG DATA DI SEKTOR PUBLIK

11.1 PENDAHULUAN

Sektor publik menjadi semakin sadar akan potensi nilai yang dapat diperoleh dari big data. Pemerintah menghasilkan dan mengumpulkan data dalam jumlah besar melalui aktivitas sehari-hari, seperti mengelola pembayaran pensiun dan tunjangan, pengumpulan pajak, sistem kesehatan nasional, pencatatan data lalu lintas, dan penerbitan dokumen resmi. Bab ini memperhitungkan tren sosio-ekonomi dan teknologi saat ini, termasuk peningkatan produktivitas di lingkungan dengan keterbatasan anggaran yang signifikan, meningkatnya permintaan akan layanan medis dan sosial, serta standardisasi dan interoperabilitas sebagai persyaratan penting bagi teknologi dan aplikasi sektor publik. Beberapa contoh manfaat potensial adalah sebagai berikut:

1. **Pemerintahan Yang Terbuka Dan Pembagian Data**: Arus informasi yang bebas dari organisasi ke masyarakat mendorong kepercayaan dan transparansi yang lebih besar antara masyarakat dan pemerintah, sejalan dengan inisiatif data terbuka.
2. **Analisis Sentimen Warga**: Informasi dari media sosial tradisional dan baru (situs web, blog, twitter, dll.) dapat membantu pembuat kebijakan untuk memprioritaskan layanan dan memperhatikan kepentingan dan pendapat warga.
3. **Segmentasi Dan Personalisasi Warga Negara Sambil Menjaga Privasi**: Menyesuaikan layanan pemerintah untuk individu dapat meningkatkan efektivitas, efisiensi, dan kepuasan warga negara.
4. **Analisis Ekonomi**: Korelasi berbagai sumber data akan membantu ekonom pemerintah menghasilkan perkiraan keuangan yang lebih akurat.
5. **Agen Pajak**: Algoritme otomatis untuk menganalisis kumpulan data besar dan integrasi data terstruktur dan tidak terstruktur dari media sosial dan sumber lain akan membantu mereka memvalidasi informasi atau menandai potensi penipuan.
6. **Penerapan Kota Pintar Dan Internet Of Things (IoT)**: Sektor publik semakin ditandai dengan penerapan yang mengandalkan pengukuran sensor terhadap fenomena fisik seperti volume lalu lintas, pencemaran lingkungan, tingkat penggunaan wadah limbah, lokasi kendaraan kota, atau deteksi perilaku tidak normal. Analisis terpadu terhadap sumber data IoT bervolume tinggi dan berkecepatan tinggi ini berpotensi meningkatkan pengelolaan perkotaan secara signifikan dan memberikan dampak positif terhadap keselamatan dan kualitas hidup warganya.
7. **Keamanan Dunia Maya**: Mengumpulkan, mengatur, dan menganalisis sejumlah besar data dari jaringan komputer pemerintah dengan data sensitif atau layanan penting, untuk memberikan kemampuan yang lebih besar kepada para pembela dunia maya dalam mendeteksi dan melawan serangan jahat.

Big Data untuk Sektor Publik

Hingga saat ini, belum ada penerapan big data secara luas di sektor publik. Dibandingkan dengan sektor lain, sektor publik belum secara tradisional menggunakan

teknologi data mining secara intensif. Namun, terdapat peningkatan minat di sektor publik terhadap potensi big data untuk memperbaiki kondisi keuangan saat ini.

Beberapa contoh peningkatan kesadaran global adalah Satuan Tugas Gabungan Industri/Pemerintah untuk mendorong pengembangan big data di Irlandia, yang diumumkan oleh Menteri Pekerjaan, Perusahaan, dan Inovasi Irlandia pada bulan Juni 2013 (Pemerintah Irlandia 2013), atau pengumuman dibuat oleh pemerintahan Obama (Gedung Putih 2012), mengenai "Inisiatif Penelitian dan Pengembangan Big Data" di mana enam departemen dan lembaga Federal mengumumkan komitmen baru senilai lebih dari Rp.200 juta untuk meningkatkan alat dan teknik yang diperlukan untuk mengakses, mengatur, dan mengumpulkan penemuan dari sejumlah besar data digital.

Dampak Pasar dari Big Data

Tidak ada dampak pasar atau persaingan langsung, karena sektor publik bukanlah sektor produktif, meskipun pengeluarannya mewakili 49,3 % PDB pada tahun 2012 di Uni Eropa28. Sebagian besar pendapatan sektor ini dikumpulkan melalui pajak dan kontribusi sosial. Oleh karena itu, dampak teknologi big data adalah dalam hal efisiensi: semakin efisien sektor publik, maka semakin baik pula kesejahteraan masyarakatnya, karena semakin sedikit sumber daya (pajak) yang dibutuhkan untuk menyediakan tingkat layanan yang sama. Oleh karena itu, semakin efektif sektor publik, semakin besar dampak positifnya terhadap perekonomian, transisi ke sektor-sektor produktif lainnya, dan semakin besar dampak positifnya terhadap masyarakat. Selain itu, kualitas layanan yang diberikan, misalnya pendidikan, kesehatan, layanan sosial, kebijakan aktif, dan keamanan, juga dapat ditingkatkan dengan memanfaatkan teknologi big data.

11.2 ANALISIS KEBUTUHAN INDUSTRI DI SEKTOR PUBLIK

Manfaat big data di sektor publik dapat dikelompokkan menjadi tiga bidang besar, berdasarkan klasifikasi jenis manfaatnya:

> *Analisis Big Data Area ini mencakup aplikasi yang hanya dapat dilakukan melalui algoritme otomatis untuk analisis tingkat lanjut guna menganalisis kumpulan data besar guna pemecahan masalah yang dapat mengungkap wawasan berbasis data. Kemampuan tersebut dapat digunakan untuk mendeteksi dan mengenali pola atau menghasilkan perkiraan.*

Penerapan di bidang ini mencakup deteksi penipuan (McKinsey Global Institute 2011); pengawasan terhadap kegiatan yang diatur oleh sektor swasta; analisis sentimen konten Internet untuk menentukan prioritas layanan publik (Oracle 2012); deteksi ancaman dari sumber data eksternal dan internal untuk pencegahan kejahatan, intelijen, dan keamanan (Oracle 2012); dan prediksi untuk keperluan perencanaan pelayanan publik (Yiu 2012).

Peningkatan Efektivitas Meliputi penerapan big data untuk memberikan transparansi internal yang lebih baik. Warga negara dan dunia usaha dapat mengambil keputusan yang lebih baik dan lebih efektif, dan bahkan menciptakan produk dan layanan baru berkat informasi yang diberikan. Beberapa contoh penerapan dalam bidang ini mencakup

ketersediaan data di seluruh silo organisasi (McKinsey Global Institute 2011); berbagi informasi melalui organisasi sektor publik [misalnya menghindari masalah akibat kurangnya database identitas tunggal (misalnya di Inggris) (Yiu 2012)]; pemerintahan terbuka dan data terbuka memfasilitasi arus bebas informasi dari organisasi publik ke masyarakat dan dunia usaha, menggunakan kembali data untuk memberikan layanan baru dan inovatif kepada masyarakat (McKinsey Global Institute 2011; Ojo et al. 2015).

Peningkatan Efisiensi Area ini mencakup aplikasi yang memberikan layanan yang lebih baik dan perbaikan berkelanjutan berdasarkan personalisasi layanan dan pembelajaran dari kinerja layanan tersebut. Beberapa contoh penerapan di bidang ini adalah personalisasi layanan publik untuk menyesuaikan dengan kebutuhan masyarakat dan peningkatan layanan publik melalui analisis internal berdasarkan analisis indikator kinerja.

11.3 POTENSI PENERAPAN BIG DATA UNTUK SEKTOR PUBLIK

Empat aplikasi potensial untuk sektor publik dijelaskan dan dikembangkan di Zillner et al. (2013, 2014) untuk mendemonstrasikan penggunaan teknologi big data di sektor publik (Tabel 11.1).

Tabel 11.1 Ringkasan skenario penerapan untuk sektor publik

Nama	Pemantauan Dan Pengawasan Terhadap Aktivitas Yang Diatur Bagi Operator Perjudian Online
Latar Belakang	Banyaknya data yang tersedia menyulitkan pengaturan dan pengawasan kegiatan secara efektif
Ringkasan	Untuk memantau operator perjudian online untuk mengendalikan aktivitas yang diatur dan mendeteksi penipuan. Pengguna aplikasi ini merupakan badan publik yang membidangi kegiatan pengawasan. Prosedur ini merupakan kewajiban regulasi dari administrasi publik; operator perjudian online harus memberikan informasi kepada masyarakat regulator melalui saluran komunikasi tertentu. Data real-time diterima dari operator perjudian setiap 5 menit.
Tujuan bisnis	Memastikan kepatuhan terhadap peraturan, pencegahan dan deteksi penipuan, dan investigasi kriminal.
Nama	Efisiensi operasional di agen tenaga kerja
Latar Belakang	Ekstrak nilai dari sejumlah besar data yang tidak terpakai
Ringkasan	Mengaktifkan rangkaian layanan baru yang dipersonalisasi, meningkatkan layanan pelanggan, dan memangkas biaya operasional di agen Tenaga Kerja Federal Jerman. Seluruh pekerja yang menganggur menerima standar layanan yang sama meskipun memiliki profil yang berbeda. Data historis pelanggan mereka dianalisis, termasuk profil, intervensi, dan waktu yang dibutuhkan untuk mencari pekerjaan. Berdasarkan analisis ini, segmentasi pelanggan dikembangkan.
Tujuan bisnis	Mengurangi biaya dan meningkatkan kualitas layanan: kini mereka dapat mendapatkan pekerjaan baru dalam waktu yang lebih singkat.
Nama	Keamanan Publik di Kota Cerdas

Latar Belakang	Data dalam jumlah besar yang tersedia dari sensor, media sosial, dan panggilan darurat dapat digabungkan untuk memberikan keamanan publik yang efektif.
Ringkasan	Kota pintar yang dilengkapi dengan sensor dan infrastruktur komunikasi membantu sektor publik menjaga kota dan warganya tetap aman. Memiliki informasi yang akurat dan terkini memungkinkan respons yang lebih baik dan lebih cepat selama keadaan darurat serta mengurangi kerusakan dan korban jiwa. Sumber umum untuk memperoleh informasi tersebut dapat berasal dari panggilan tanggap darurat, kamera pengintai, dan pasukan bergerak (seperti mobil patroli polisi) yang tiba di suatu lokasi. Dalam beberapa tahun terakhir, media sosial telah menunjukkan potensi menarik untuk mengumpulkan informasi yang membantu memperoleh gambaran kesadaran situasional yang akurat (van Kasteren dkk. 2014). Semua informasi yang dikumpulkan dikumpulkan di pusat komando dan kendali di mana operator dapat memutuskan bagaimana mengarahkan pasukan bergerak yang tersedia.
Tujuan bisnis	Respon cepat terhadap keadaan darurat, pencegahan kerusakan, dan lebih sedikit korban jiwa.
Nama	Pemolisian prediktif menggunakan data terbuka
Latar Belakang	Penggunaan kembali data terbuka publik untuk memberikan kebijakan prediktif
Ringkasan	Pemerintah di seluruh dunia telah memulai inisiatif data terbuka untuk membuat data sektor publik tersedia bagi publik demi transparansi dan memungkinkan pihak ketiga menawarkan layanan berdasarkan data tersebut. Salah satu layanan tersebut dapat digambarkan sebagai kebijakan prediktif dimana data kejahatan historis digunakan untuk secara otomatis menemukan tren dan pola. Pola yang teridentifikasi membantu memperoleh wawasan mengenai masalah terkait kejahatan yang dihadapi sebuah kota dan memungkinkan penempatan pasukan polisi yang lebih efektif dan efisien (Wang dkk. 2013; PredPol 2013).
Tujuan bisnis	Penurunan kejahatan secara signifikan, penggunaan pasukan bergerak secara efisien.

11.4 PENDORONG DAN KENDALA BIG DATA DI SEKTOR PUBLIK

Faktor pendorong dan kendala utama teknologi big data di sektor publik adalah:

Pengemudi

Penggerak berikut ini diidentifikasi untuk big data di sektor publik:

- Pemerintah dapat bertindak sebagai katalis dalam pengembangan ekosistem data melalui pembukaan kumpulan data mereka sendiri, dan secara aktif mengelola penyebaran dan penggunaannya (World Economic Forum 2012).
- Inisiatif data terbuka adalah titik awal untuk meningkatkan pasar data yang dapat memanfaatkan informasi terbuka (konten) dan teknologi data besar. Oleh karena itu, kebijakan aktif di bidang data terbuka dapat menguntungkan sektor swasta, dan sebagai imbalannya memfasilitasi pertumbuhan industri ini di Eropa. Pada akhirnya hal ini akan menguntungkan anggaran publik dengan peningkatan pendapatan pajak dari industri data Eropa yang sedang berkembang.

Kendala

Kendala big data di sektor publik dapat diringkas sebagai berikut:

- Kurangnya kemauan politik untuk membuat sektor publik memanfaatkan teknologi ini. Hal ini memerlukan perubahan pola pikir para pejabat senior di sektor publik.
- Kurangnya orang-orang terampil yang berorientasi pada bisnis yang menyadari di mana dan bagaimana big data dapat membantu memecahkan tantangan-tantangan sektor publik, dan siapa yang dapat membantu mempersiapkan kerangka peraturan untuk keberhasilan pengembangan solusi big data.
- Peraturan Perlindungan Data Umum yang baru dan arahan PSI menunjukkan ketidakpastian mengenai dampak penerapan inisiatif big data dan data terbuka di sektor publik. Secara khusus, data terbuka diharapkan menjadi katalisator dari sektor publik hingga sektor swasta untuk membangun industri data yang kuat.
- Mendapatkan momentum adopsi. Saat ini, terdapat lebih banyak pemasaran seputar big data di sektor publik dibandingkan pengalaman nyata untuk mempelajari aplikasi mana yang lebih menguntungkan, dan bagaimana penerapannya. Hal ini memerlukan pengembangan serangkaian standar solusi big data untuk sektor ini.
- Banyak badan dalam administrasi publik (terutama yang sangat terdesentralisasi), begitu banyak energi yang hilang dan akan tetap demikian sampai strategi bersama terwujud untuk penggunaan kembali platform lintas teknologi.

11.5 SUMBER DAYA DATA SEKTOR PUBLIK YANG TERSEDIA

Dalam Petunjuk 2003/98/EC (Parlemen Eropa dan Dewan Uni Eropa 2003), mengenai penggunaan kembali informasi sektor publik, informasi sektor publik (PSI) didefinisikan sebagai berikut: “Ini mencakup setiap representasi tindakan, fakta atau informasi – dan setiap kompilasi dari tindakan, fakta atau informasi tersebut – apapun medianya (tertulis di atas kertas, atau disimpan dalam bentuk elektronik atau sebagai rekaman suara, visual atau audio visual), yang dimiliki oleh badan publik. Dokumen yang dipegang oleh badan sektor publik adalah dokumen yang badan sektor publiknya mempunyai hak untuk mengizinkan penggunaannya kembali.” Menurut Correia (2004), mengenai ketersediaan informasi yang dihasilkan oleh badan-badan publik tersebut, dan jika tidak ada pedoman khusus, badan penghasil informasi bebas memutuskan bagaimana menyediakannya: langsung kepada pengguna akhir, mendirikan badan publik/ kemitraan swasta, atau melakukan outsourcing eksploitasi komersial atas informasi tersebut kepada operator swasta. Directive 2003/98/EC mengklarifikasi bahwa aktivitas yang berada di luar tugas publik: “biasanya mencakup penyediaan dokumen yang dibuat dan dikenakan biaya secara eksklusif atas dasar komersial dan bersaing dengan pihak lain di pasar”.

Mengenai sifat PSI yang tersedia, ada beberapa pendekatan. Makalah Hijau tentang PSI (Komisi Eropa 1998) mengusulkan beberapa klasifikasi seperti:

- PSI membedakan antara administratif dan non-administratif
- Keistimewaan PSI mengenai relevansinya bagi publik

Selain itu, data ini dapat dibedakan berdasarkan potensi nilai pasarnya, dan dalam beberapa kasus berdasarkan konten data pribadi:

Pembedaan PSI berdasarkan anonimitasnya

Jumlah data terpenting yang dihasilkan oleh sektor publik adalah tekstual atau numerik, dibandingkan sektor lain seperti layanan kesehatan yang menghasilkan gambar elektronik dalam jumlah besar. Sebagai hasil dari inisiatif e-Government selama 15 tahun terakhir, sebagian besar data dibuat dalam bentuk digital, 90 % menurut McKinsey (McKinsey Global Institute 2011).

Menurut survei yang dilakukan untuk perumusan Kemitraan Nilai Big Data Eropa kepada perwakilan sektor publik (Zillner et al. 2014), aset data utama adalah keseluruhan sistem sektor publik, pencatatan, basis data, dan sistem informasi, yang mana yang paling signifikan adalah:

1. Warga negara, dunia usaha, dan properti (misalnya pencatatan dasar, transaksi)
2. Data fiskal
3. Keamanan data
4. Pengelolaan dokumen terutama seiring dengan berkembangnya transaksi elektronik
5. Pengadaan dan pengeluaran pemerintah
6. Badan publik dan pegawai
7. Data geografis terutama berkaitan dengan kadaster
8. Konten terkait budaya, pendidikan, dan pariwisata
9. Dokumen legislatif
10. Data statistik (data sosial-ekonomi yang dapat digunakan oleh sektor swasta)
11. Data geospasial

11.6 PERSYARATAN SEKTOR PUBLIK

Persyaratan sektor publik dipecah menjadi persyaratan non-teknis dan teknis.

Persyaratan Non-teknis

Masalah Privasi dan Keamanan Pengumpulan data melintasi batas administratif tanpa berdasarkan permintaan merupakan tantangan nyata, karena informasi ini dapat mengungkapkan informasi pribadi dan keamanan yang sangat sensitif ketika digabungkan dengan berbagai sumber data lainnya, tidak hanya membahayakan privasi individu tetapi juga keamanan sipil. Hak akses ke kumpulan data yang diperlukan untuk suatu operasi harus dibenarkan dan diperoleh. Ketika operasi baru dilakukan pada data yang ada, pemberitahuan atau lisensi harus diperoleh dari Badan Privasi Data. Anonimitas harus dijaga dalam kasus ini, sehingga diperlukan disosiasi data. Masalah privasi individu dan keamanan publik harus diatasi sebelum pemerintah dapat diyakinkan untuk membagikan data secara lebih terbuka, tidak hanya secara publik namun juga berbagi secara terbatas dengan pemerintah lain atau entitas internasional. Dimensi lainnya adalah regulasi penggunaan cloud computing sedemikian rupa sehingga sektor publik dapat mempercayai penyedia cloud. Selain itu, kurangnya penyedia komputasi awan big data Eropa di pasar Eropa juga menjadi hambatan dalam penerapannya.

Keterampilan Big Data Masih kurangnya ilmuwan dan ahli teknologi data yang terampil yang dapat menangkap dan memproses sumber data baru ini. Ketika teknologi big data semakin diadopsi dalam bisnis, profesional big data yang terampil akan semakin sulit ditemukan. Badan-badan publik dapat melakukan banyak hal dengan keterampilan yang mereka miliki, namun mereka perlu memastikan bahwa keterampilan tersebut meningkat (1105 Government Information Group n.d.). Selain orang-orang yang berorientasi teknis, pengetahuan orang-orang yang berorientasi bisnis dan sadar akan manfaat big data untuk membantu mereka memecahkan tantangan-tantangan sektor publik masih kurang. Persyaratan Lainnya Persyaratan non teknis lainnya antara lain:

1. Kesediaan untuk memasok dan mengadopsi teknologi big data, dan juga mengetahui cara menggunakannya.
2. Perlunya pendekatan (kebijakan) nasional atau Eropa yang sama—seperti kebijakan Eropa untuk interoperabilitas dan data terbuka.
3. Kurangnya kepemimpinan di bidang ini.
4. Ketidaksesuaian umum antara intelijen bisnis pada umumnya dan data besar pada khususnya di sektor publik.

Persyaratan Teknis

Di bawah ini adalah penjelasan rinci masing-masing dari delapan persyaratan teknis yang disaring dari empat aplikasi big data yang dipilih untuk Forum Sektor Publik. Penemuan Pola Mengidentifikasi pola dan persamaan untuk mendeteksi perilaku kriminal atau ilegal tertentu dalam skenario penerapan pemantauan dan pengawasan operator perjudian online (dan juga untuk skenario pemantauan serupa di sektor publik). Persyaratan ini juga berlaku dalam skenario peningkatan efisiensi operasional di agen tenaga kerja, dan dalam skenario kebijakan prediktif.

Berbagi Data/Integrasi Data Diperlukan untuk mengatasi kurangnya standarisasi skema data dan fragmentasi kepemilikan data. Integrasi sumber data yang beragam dan beragam ke dalam platform data besar. Wawasan Waktu Nyata Memungkinkan analisis data baru/waktu nyata untuk pengambilan keputusan instan, untuk memperoleh wawasan waktu nyata dari data.

Keamanan dan Privasi Data Prosedur hukum dan sarana teknis yang memungkinkan pembagian data yang aman dan menjaga privasi. Solusi terhadap kebutuhan ini dapat membuka peluang meluasnya penggunaan big data di sektor publik. Kemajuan dalam perlindungan dan privasi data merupakan kunci bagi sektor publik, karena hal ini memungkinkan analisis sejumlah besar data yang dimiliki oleh sektor publik tanpa mengungkapkan informasi sensitif. Masalah privasi dan keamanan ini menghalangi penggunaan infrastruktur cloud (pemrosesan, penyimpanan) oleh banyak lembaga publik yang menangani data sensitif. Transmisi Data Real Time Karena kemampuan penempatan sensor semakin meningkat dalam skenario aplikasi kota pintar, terdapat permintaan yang tinggi untuk transmisi data real-time. Diperlukan kemampuan pemrosesan dan pembersihan terdistribusi untuk sensor gambar agar tidak merusak saluran komunikasi dan hanya

memberikan informasi yang diperlukan untuk analisis real-time, yang akan memperkuat sistem kesadaran situasional bagi para pengambil keputusan.

Natural Language Analytics Ekstrak informasi dari sumber online yang tidak terstruktur (misalnya media sosial) untuk memungkinkan penambangan sentimen. Pengenalan data dari input bahasa alami seperti teks, audio, dan video.

Analisis Prediktif Sebagaimana dijelaskan dalam skenario penerapan kepolisian prediktif, yang tujuannya adalah untuk mendistribusikan pasukan keamanan dan sumber daya sesuai dengan prediksi insiden, memberikan prediksi berdasarkan pembelajaran dari situasi sebelumnya untuk memperkirakan alokasi sumber daya yang optimal untuk layanan publik. Pemodelan dan Simulasi Alat khusus domain untuk pemodelan dan simulasi peristiwa berdasarkan data peristiwa masa lalu untuk mengantisipasi hasil keputusan yang diambil untuk mempengaruhi kondisi saat ini secara real-time, misalnya dalam skenario keselamatan publik.

11.7 PETA JALAN TEKNOLOGI UNTUK BIG DATA DI SEKTOR PUBLIK

Untuk setiap kebutuhan di sektor ini, bagian ini menyajikan teknologi yang dapat diterapkan dan pertanyaan penelitian yang akan dikembangkan (Gambar 11.1). Semua referensi yang disajikan di sini berasal dari Curry dkk. (2014).

Penemuan Pola

1. **Teknologi Analisis Data**: Teknologi pola semantik termasuk pencocokan pola aliran. Pertanyaan Penelitian: Pencocokan pola kompleks yang dapat diskalakan. Mencapai triliunan melalui kumpulan data akan memakan waktu 5 tahun.
2. **Teknologi Kurasi Data**: Validasi keluaran analisis pola dengan manusia melalui kurasi. Pertanyaan Penelitian: Kurasi dalam skala besar bergantung pada interaksi antara platform kurasi otomatis dan pendekatan kolaboratif yang memanfaatkan kumpulan besar kurator data. Hasil penerapan komersial dapat dicapai dalam 6–10 tahun.
3. **Teknologi Penyimpanan Data**: Database Analitik, Hadoop, Spark, Mahout. Pertanyaan Penelitian: Bahasa Kueri Array Standar. Saat ini terdapat kekurangan bahasa kueri standar tetapi upaya seperti ArrayQL sedang dalam proses. Saat ini belum ada adopsi yang luas dan DB yang ada (SciDB, Rasdaman) digunakan dalam komunitas ilmiah. Hal ini mungkin berubah dalam 3–5 tahun dari sekarang.

Berbagi Data/Integrasi Data

1. **Teknologi Akuisisi Data**: Untuk memfasilitasi integrasi dan analisis. – Pertanyaan Penelitian: Pemilihan fragmen data, pengambilan sampel, dan skalabilitas. Solusi akan dihasilkan oleh komputer kuantum (diperkirakan akan tersedia dalam 5–10 tahun, namun 15–20 tahun tampaknya lebih realistis.)
2. **Teknologi Analisis Data**: Data tertaut menyediakan rangkaian teknologi terbaik untuk berbagi data di Web. Data dan ontologi yang terhubung menyediakan mekanisme untuk mengintegrasikan data (memetakan ke ontologi yang sama; memetakan antar ontologi/skema/contoh).

a. Pertanyaan Penelitian: Skalabilitas, berhubungan dengan kecepatan data yang tinggi dan variasi yang tinggi. Menangani triliunan node akan memakan waktu 3–5 tahun.
b. Pertanyaan Penelitian: Membuat sistem semantik mudah digunakan oleh para ahli non-semantik (logika). Diperlukan waktu setidaknya 5 tahun untuk mendapatkan dukungan peralatan yang komprehensif.

Gambar 11.1 Persyaratan pemetaan untuk pertanyaan penelitian di sektor publik

2) **Teknologi Kurasi/Penyimpanan Data**: Metadata dan kerangka asal data.

Pertanyaan Penelitian: Apa standar format penelusuran data umum? Asalnya pada jenis penyimpanan tertentu, mis. database grafik, masih mahal secara komputasi. Integrasi kesadaran asal ke dalam alat yang ada dapat dicapai

dalam jangka pendek (2–3 tahun) setelah alat ini mencapai permintaan pasar yang penting.

Wawasan Waktu Nyata

1) **Teknologi Analisis Data**: Teknologi data dan pembelajaran mesin yang terhubung dapat mendukung analisis otomatis, yang diperlukan untuk memperoleh wawasan waktu nyata.
   Pertanyaan Penelitian: Performa tinggi saat menghadapi 3 V (volume, Variety, dan Velocity). Analisis mendalam secara real-time membutuhkan waktu lebih dari 5 tahun lagi.
2) **Teknologi Penyimpanan Data**: Google Data Flow, Amazon Kinesis, Spark, Drill, Impala, database dalam memori.
   Pertanyaan Penelitian: Bagaimana kueri ad hoc dan streaming pada kumpulan data besar dapat dieksekusi dengan latensi minimal? Ini adalah bidang penelitian yang aktif dan mungkin akan mencapai kematangan lebih lanjut dalam waktu beberapa tahun.

Keamanan dan Privasi Data

1) **Teknologi Penyimpanan Data**: Penyimpanan dan DB terenkripsi; enkripsi ulang proxy antar domain; perlindungan privasi otomatis (misalnya privasi diferensial).
   Pertanyaan Penelitian: Kemajuan dalam "privasi berdasarkan desain" untuk menghubungkan kebutuhan analitik dengan kontrol perlindungan dalam pemrosesan dan penyimpanan. Kerangka hukum, misalnya Peraturan Perlindungan Data Umum (GDPR), harus diselaraskan di antara negara-negara anggota UE. Selain undang-undang, data dan kepentingan bersama juga diperlukan (Curry dkk. 2014). Ini akan memerlukan setidaknya 3 tahun penelitian lebih lanjut.

Transmisi Data Waktu Nyata

1) **Teknologi Akuisisi Data**: Kafka, Flume, Storm, dll., Curry dkk. (2014). – Pertanyaan Penelitian: Pemrosesan dan pembersihan terdistribusi. Pendekatan saat ini harus dapat memberi tahu pengguna jenis sumber daya yang mereka perlukan untuk melakukan tugas yang ditentukan oleh pengguna (misalnya proses 10 GB/s). Pendekatan pertama untuk mencapai tujuan ini sedang muncul dan akan tersedia di pasar dalam 5 tahun ke depan.
2) **Teknologi Penyimpanan Data**: Praktik terbaik saat ini: menulis solusi penyimpanan yang dioptimalkan (misalnya HDFS), penyimpanan berbentuk kolom.
   Pertanyaan Penelitian: Bagaimana meningkatkan kinerja baca/tulis acak teknologi basis data. Arsitektur Lambda yang dijelaskan oleh Marz dan Warren mencerminkan standar praktik terbaik saat ini untuk menyimpan data berkecepatan tinggi. Secara efektif hal ini mengatasi kekurangan kinerja acak/baca tulis yang tidak mencukupi dari teknologi DB yang ada. Peningkatan kinerja akan berkelanjutan dan bertahap serta menyederhanakan pengembangan tumpukan teknologi big data secara keseluruhan. Teknologi dapat mencapai tingkat kematangan yang mengarah pada penyederhanaan cetak biru arsitektur dalam 3–4 tahun.

Analisis Bahasa Alami

1) **Teknologi Analisis Data**: Ekstraksi informasi, pengenalan entitas bernama, pembelajaran mesin, data tertaut. Penautan entitas dan resolusi referensi bersama.
   Pertanyaan Penelitian: Meningkatkan skalabilitas dan ketahanan. Solusi yang kuat dan terukur setidaknya membutuhkan waktu 3–5 tahun lagi.
2) **Teknologi Kurasi Data**: Validasi keluaran Natural Language Analytics (NLA) dengan manusia melalui kurasi.
   Pertanyaan Penelitian: Kurasi dalam skala besar bergantung pada interaksi antara platform kurasi otomatis dan pendekatan kolaboratif yang memanfaatkan sejumlah besar kurator data. Secara teknis, integrasi ini dapat dicapai dalam jangka pendek (2–3 tahun).

Analisis Prediktif

1) **Teknologi Penyimpanan Data**: Database analitis.
   Pertanyaan Penelitian: Bagaimana database dapat mendukung analitik prediktif secara efisien? Dari sudut pandang penyimpanan, database analitis mengatasi masalah kinerja yang lebih baik karena DB sendiri mampu mengeksekusi kode analitis. Saat ini terdapat kekurangan bahasa kueri standar tetapi upaya seperti ArrayQL sedang dalam proses. Hal ini mungkin berubah dalam 3–5 tahun dari sekarang.

Pemodelan dan Simulasi

1) Teknologi Penyimpanan Data: Praktik terbaik; pemrosesan batch dan in-stream (arsitektur Lambda), database temporal.
   Pertanyaan Penelitian: Bagaimana data deret waktu dikelola secara umum untuk analisis yang efektif? Basis data spatiotemporal adalah bidang penelitian aktif dan hasilnya mungkin melebihi skala waktu 5 tahun.
2) Teknologi Penggunaan Data: Standar dalam pemodelan (semantik); penerapan simulasi dalam perencanaan (misalnya perencanaan pabrik).
   Pertanyaan Penelitian: Membuat model eksplisit dan/atau transparan. Ini adalah pertanyaan penelitian dengan jangka waktu yang panjang (di luar tahun 2020).

Kesimpulan dan Rekomendasi untuk Sektor Publik

Temuan setelah menganalisis persyaratan dan teknologi yang tersedia saat ini menunjukkan bahwa ada sejumlah pertanyaan penelitian terbuka yang harus dijawab untuk mengembangkan teknologi sehingga solusi yang kompetitif dan efektif dapat dibangun. Perkembangan utama diperlukan di bidang skalabilitas analisis data, penemuan pola, dan aplikasi real-time. Yang juga diperlukan adalah peningkatan sumber data untuk pembagian dan integrasi data dari sektor publik. Menyediakan mekanisme keamanan dan privasi yang terintegrasi dalam aplikasi big data juga sangat penting, karena sektor publik mengumpulkan data sensitif dalam jumlah besar. Di banyak negara, undang-undang membatasi penggunaan data hanya untuk tujuan awal data tersebut diperoleh. Bagaimanapun, menghormati privasi warga negara adalah kewajiban wajib di Uni Eropa.

Bidang lain, yang khususnya menarik dalam penerapan keselamatan di sektor publik, adalah analisis bahasa alami, yang dapat berguna sebagai metode untuk mengumpulkan

umpan balik tidak terstruktur dari masyarakat, misalnya. dari media sosial dan jaringan. Pengembangan analisis prediktif yang efektif, serta alat pemodelan dan simulasi untuk analisis data historis, merupakan tantangan utama yang harus diatasi oleh penelitian di masa depan.

BAB 12
BIG DATA DI SEKTOR KEUANGAN DAN ASURANSI

12.1 PENDAHULUAN

Sektor keuangan dan asuransi pada dasarnya telah menjadi industri yang sangat bergantung pada data selama bertahun-tahun, dengan lembaga keuangan yang mengelola data pelanggan dalam jumlah besar dan menggunakan analisis data di berbagai bidang seperti perdagangan pasar modal. Bisnis asuransi didasarkan pada analisis data untuk memahami dan mengevaluasi risiko secara efektif. Aktuaris dan profesional penjamin emisi bergantung pada analisis data untuk dapat menjalankan peran inti mereka; sehingga dapat dikatakan bahwa data ini merupakan kekuatan dominan di sektor ini.

Namun terdapat peningkatan prevalensi data yang termasuk dalam domain data besar, yaitu aset informasi bervolume tinggi, berkecepatan tinggi, dan beragam yang dihasilkan dari munculnya data pelanggan, pasar, dan peraturan baru yang muncul dari berbagai sumber. Yang menambah kompleksitas adalah keberadaan data terstruktur dan tidak terstruktur secara berdampingan. Data tidak terstruktur di industri jasa keuangan dan asuransi dapat diidentifikasi sebagai area dimana terdapat banyak sekali nilai bisnis yang belum dieksploitasi. Misalnya, ada banyak nilai komersial yang dapat diperoleh dari sejumlah besar dokumentasi klaim asuransi yang sebagian besar berbentuk teks dan berisi deskripsi yang dimasukkan oleh operator pusat panggilan, catatan yang terkait dengan klaim dan kasus individual. Dengan bantuan teknologi big data, nilai tidak hanya dapat diekstraksi secara lebih efisien dari sumber data tersebut, namun analisis bentuk data tidak terstruktur ini digabungkan dengan beragam kumpulan data untuk mengekstraksi nilai komersial yang lebih cepat dan tepat sasaran. Karakteristik penting dari big data dalam industri ini adalah nilai — bagaimana sebuah bisnis tidak hanya mengumpulkan dan mengelola big data, namun bagaimana data yang memiliki nilai dapat diidentifikasi dan bagaimana organisasi dapat merekayasa ke depan (bukan mengevaluasi secara retrospektif) nilai komersial dari data.

Dampak Pasar dari Big Data

Pasar teknologi big data di bidang keuangan dan asuransi adalah salah satu yang paling menjanjikan. Menurut perkiraan TechNavio (Technavio 2013), pasar data besar global di sektor jasa keuangan akan tumbuh pada CAGR sebesar 56,7% selama periode 2012–2016. Salah satu faktor utama yang berkontribusi terhadap pertumbuhan pasar ini adalah kebutuhan untuk memenuhi peraturan keuangan, namun kurangnya sumber daya terampil untuk mengelola data besar dapat menimbulkan tantangan.

Vendor utama yang mendominasi bidang ini termasuk Hewlett-Packard, IBM, Microsoft, dan Oracle yang merupakan pemain global yang mapan dengan profil generalis. Namun, daya tarik pasar akan menjadi faktor penarik pendatang baru di tahun-tahun mendatang. Karena data merupakan aset yang paling penting, teknologi ini sangat menguntungkan dan membedakan organisasi jasa keuangan, seperti yang dinyatakan dalam laporan IBM Institute for Business Value "Analytics: The real-world use of big data in financial

services" (IBM 2013). Dengan memanfaatkan aset ini, bank dan perusahaan pasar keuangan dapat memperoleh pemahaman komprehensif tentang pasar, pelanggan, saluran, produk, peraturan, pesaing, pemasok, dan karyawan sehingga mereka dapat bersaing dengan lebih baik. Oleh karena itu, hal ini merupakan tren positif di pasar dan diharapkan dapat mendorong pertumbuhan pasar big data global di sektor jasa keuangan.

Dalam hal strategi data, organisasi jasa keuangan mengambil pendekatan berbasis bisnis terhadap big data: persyaratan bisnis diidentifikasi terlebih dahulu dan kemudian sumber daya dan kapasitas internal yang ada diselaraskan untuk mendukung peluang bisnis, sebelum berinvestasi pada sumber data dan infrastruktur. Namun, tidak semua lembaga keuangan mempunyai kecepatan yang sama. Menurut laporan IBM, meskipun 26% organisasi fokus pada pemahaman prinsip-prinsip utama (dibandingkan dengan 24% organisasi global), mayoritas organisasi sudah menentukan peta jalan terkait big data (47%) atau sudah melakukan uji coba dan implementasi big data (27%).

Ketertinggalan mereka dibandingkan rekan-rekan lintas industri adalah dalam penggunaan tipe data yang lebih bervariasi dalam implementasi big data mereka. Lebih dari 21 % dari perusahaan-perusahaan ini menganalisis data audio (sering kali diproduksi dalam jumlah besar di call center bank ritel), sementara lebih dari 27 % melaporkan menganalisis data sosial (dibandingkan dengan 38 % dan 43 %, masing-masing, dari data lintas negara). rekan industri). Kurangnya fokus pada data tidak terstruktur disebabkan oleh perjuangan yang sedang berlangsung untuk mengintegrasikan data terstruktur yang sangat besar dalam organisasi.

12.2 ANALISIS KEBUTUHAN INDUSTRI SEKTOR KEUANGAN DAN ASURANSI

Munculnya big data dalam layanan keuangan dapat membawa banyak keuntungan bagi lembaga keuangan. Manfaat yang memiliki dampak komersial terbesar disoroti sebagai berikut:

Peningkatan Tingkat Wawasan, Keterlibatan, dan Pengalaman Pelanggan Dengan digitalisasi produk dan layanan keuangan serta meningkatnya tren interaksi pelanggan dengan merek atau organisasi di ruang digital, terdapat peluang bagi organisasi jasa keuangan untuk meningkatkan tingkat keterlibatan pelanggan dan secara proaktif meningkatkan pengalaman pelanggan. Banyak yang berargumentasi bahwa ini adalah area yang paling penting bagi lembaga keuangan untuk mulai memanfaatkan teknologi big data agar tetap menjadi yang terdepan, atau bahkan sekadar bersaing dalam persaingan. Untuk membantu mencapai hal ini, teknologi big data dan teknik analisis dapat membantu memperoleh wawasan dari sumber-sumber baru yang tidak terstruktur seperti media sosial.

Peningkatan Kemampuan Deteksi dan Pencegahan Penipuan Lembaga jasa keuangan selalu rentan terhadap penipuan. Ada individu dan organisasi kriminal yang berupaya menipu lembaga keuangan dan kecanggihan serta kompleksitas skema ini terus berkembang seiring berjalannya waktu. Di masa lalu, bank hanya menganalisis sampel kecil transaksi untuk mendeteksi penipuan. Hal ini dapat menyebabkan beberapa aktivitas penipuan lolos dan "hal-hal positif palsu" lainnya akan disorot. Pemanfaatan data besar berarti organisasi-organisasi

ini kini dapat menggunakan kumpulan data yang lebih besar untuk mengidentifikasi tren yang mengindikasikan penipuan guna membantu meminimalkan paparan terhadap risiko tersebut.

Analisis Perdagangan Pasar yang Ditingkatkan Perdagangan pasar keuangan mulai menjadi ruang digital beberapa tahun yang lalu, didorong oleh meningkatnya permintaan untuk eksekusi perdagangan yang lebih cepat. Strategi perdagangan yang menggunakan algoritme canggih untuk memperdagangkan pasar keuangan dengan cepat merupakan salah satu penyumbang utama data besar.

Data pasar dapat dianggap sebagai data besar. Volumenya tinggi, dihasilkan dari berbagai sumber, dan dihasilkan dengan kecepatan yang fenomenal. Namun, data besar ini tidak serta merta menghasilkan informasi yang dapat ditindaklanjuti. Manfaat nyata dari big data terletak pada ekstraksi informasi yang dapat ditindaklanjuti secara efektif dan mengintegrasikan informasi tersebut dengan sumber lain. Data pasar dari berbagai pasar dan geografi serta berbagai kelas aset dapat diintegrasikan dengan sumber terstruktur dan tidak terstruktur lainnya untuk menciptakan kumpulan data hibrid yang diperkaya (kombinasi data terstruktur dan tidak terstruktur). Ini memberikan pandangan yang komprehensif dan terintegrasi tentang keadaan pasar dan dapat digunakan untuk berbagai aktivitas seperti pembuatan sinyal, eksekusi perdagangan, pelaporan laba dan rugi (P&L), dan pengukuran risiko, semuanya secara real-time sehingga memungkinkan perdagangan lebih efektif.

12.3 POTENSI PENERAPAN BIG DATA DI BIDANG KEUANGAN DAN ASURANSI

Tiga aplikasi potensial untuk sektor keuangan dan asuransi dijelaskan dan dikembangkan di Zillner et al. (2013, 2014) sebagai perwakilan penerapan teknologi big data di sektor tersebut (Tabel 12.1).

Tabel 12.1 Ringkasan skenario penerapan big data untuk sektor keuangan dan asuransi

Nama	Deteksi manipulasi pasar.
Latar Belakang	Deteksi rumor palsu yang mencoba memanipulasi pasar.
Ringkasan	Pasar keuangan sering kali dipengaruhi oleh rumor. Terkadang rumor palsu sengaja dibuat untuk mengalihkan perhatian dan menyesatkan pelaku pasar lainnya. Perilaku ini berbeda berdasarkan hasil manipulasi yang diharapkan. Contoh penyalahgunaan pasar adalah market sounding (penyebaran ilegal informasi tidak benar tentang perusahaan yang sahamnya diperdagangkan di bursa) dan pump and dump (laporan positif palsu dipublikasikan tentang perusahaan yang sahamnya dapat diperdagangkan dengan tujuan mendorong pelaku pasar lainnya untuk melakukan hal yang sama. Membeli saham di perusahaan terkait; peningkatan permintaan akan menyebabkan harga saham naik ke tingkat yang artifisial).
Tujuan bisnis	Mengidentifikasi hoax dan menilai konsistensi informasi baru dengan sumber lain yang dapat dipercaya.
Nama	Manajemen risiko reputasi.

Latar Belakang	Penilaian paparan risiko reputasi terkait dengan layanan konsultasi yang ditawarkan bank kepada nasabahnya.
Ringkasan	Persepsi negatif dapat berdampak buruk terhadap kemampuan bank dalam mempertahankan keberadaannya, membangun hubungan bisnis baru, atau melanjutkan akses terhadap sumber pendanaan. Peningkatan kemungkinan gagal bayar (risiko kredit penerbit), ketidakstabilan harga, dan kesulitan untuk menukar produk keuangan tertentu di pasar yang dibatasi semuanya berkontribusi pada peningkatan risiko reputasi dan operasional yang terkait dengan layanan perantara dan konsultasi. Bank dan lembaga keuangan biasanya menawarkan produk keuangan pihak ketiga. Hal ini menyiratkan bahwa kurangnya kinerja produk pihak ketiga dapat berdampak nyata pada hubungan antara bank dan nasabahnya.
Tujuan bisnis	Memantau reputasi pihak ketiga dan dampak gangguan reputasi terhadap hubungan langsung antara bank dan nasabah.
Nama	Pialang ritel.
Latar Belakang	Temukan tren topik, deteksi peristiwa, atau dukung optimalisasi portofolio/alokasi aset.
Ringkasan	Tren umum di seluruh industri pialang ritel dan data pasar adalah menghadirkan fungsi yang menawarkan informasi yang dapat ditindaklanjuti. Fokusnya tidak lagi pada angka-angka berdasarkan data historis kuantitatif, misalnya. angka-angka kunci atau data kinerja. Sebaliknya, investor mencari sinyal yang memiliki elemen prediktif namun mudah dipahami. Dalam hal ini, ekstraksi sentimen dan topik dari sumber tekstual merupakan tambahan yang sempurna untuk data dan fungsi konvensional yang sudah ditawarkan oleh perusahaan pialang ritel.
Tujuan bisnis	Mengumpulkan dan meninjau berbagai sumber informasi keuangan (di pasar, perusahaan, atau lembaga keuangan) berulang kali dengan mengotomatisasi tugas ini.

12.4 PENDORONG DAN KENDALA BIG DATA DI SEKTOR KEUANGAN DAN ASURANSI

Keberhasilan realisasi big data di bidang keuangan dan asuransi memiliki beberapa pendorong dan kendala.

Pengemudi

Penggerak big data di sektor keuangan dan asuransi telah diidentifikasi sebagai berikut:

- Pertumbuhan Data: Volume transaksi keuangan meningkat, menyebabkan pertumbuhan data di perusahaan jasa keuangan. Di pasar modal, kehadiran perdagangan elektronik menyebabkan peningkatan jumlah perdagangan. Pertumbuhan data tidak terbatas pada bisnis pasar modal. Studi Pembayaran Global Capgemini/RBS pada tahun 2012 (Capgemini 2012) memperkirakan bahwa jumlah transaksi pembayaran elektronik secara global berjumlah sekitar 260 miliar dan tumbuh antara 15 dan 22 % di negara-negara berkembang.

- Peningkatan pengawasan dari regulator: Regulator industri sekarang memerlukan pandangan yang lebih transparan dan akurat mengenai bisnis keuangan dan asuransi, yang berarti mereka tidak lagi menginginkan laporan; mereka membutuhkan data mentah. Oleh karena itu, lembaga keuangan perlu memastikan bahwa mereka mampu menganalisis data mentah mereka dengan tingkat rincian yang sama dengan regulator.
- Kemajuan teknologi berarti peningkatan aktivitas: Berkat digitalisasi produk dan layanan keuangan, kemudahan dan keterjangkauan dalam melakukan transaksi keuangan secara online telah menyebabkan peningkatan aktivitas dan ekspansi ke pasar-pasar baru. Individu dapat melakukan lebih banyak perdagangan, lebih sering, di lebih banyak jenis akun, karena mereka dapat melakukannya hanya dengan mengklik tombol dalam kenyamanan rumah mereka sendiri.
- Perubahan model bisnis: Didorong oleh faktor-faktor yang disebutkan di atas, lembaga keuangan mendapati diri mereka berada di pasar yang secara fundamental berbeda dari pasar beberapa tahun yang lalu. Penerapan analisis big data diperlukan untuk membantu membangun model bisnis bagi lembaga keuangan yang bertujuan untuk mempertahankan pangsa pasar dari meningkatnya persaingan yang datang dari sektor lain.
- Wawasan pelanggan: Saat ini hubungan antara bank dan konsumen telah terbalik: konsumen kini memiliki hubungan sementara dengan banyak bank. Bank tidak lagi memiliki gambaran lengkap tentang preferensi, pola pembelian, dan perilaku nasabahnya. Oleh karena itu, teknologi big data memainkan peran penting dalam mewujudkan sentrisitas pelanggan dalam paradigma baru ini.

Kendala

Kendala big data di sektor keuangan dan asuransi dapat diringkas sebagai berikut:

- Budaya dan infrastruktur lama: Banyak bank masih bergantung pada infrastruktur TI lama yang kaku, dengan data yang terisolasi dan banyak sekali sistem lama. Oleh karena itu, big data merupakan sebuah tambahan, bukan inisiatif mandiri yang benar-benar baru. Budaya merupakan hambatan yang lebih besar terhadap penerapan big data. Banyak organisasi keuangan gagal menerapkan program big data karena mereka tidak mampu memahami bagaimana analisis data dapat meningkatkan bisnis inti mereka.
- Kurangnya keterampilan: Beberapa organisasi telah menyadari data dan peluang yang ada; namun mereka kekurangan sumber daya manusia dengan tingkat keterampilan yang tepat untuk menjembatani kesenjangan antara data dan peluang potensial. Keterampilan yang “hilang” adalah keterampilan seorang data scientist.
- “Kemampuan untuk Ditindaklanjuti” Data: Tantangan utama berikutnya dapat dilihat dalam membuat big data dapat ditindaklanjuti. Teknologi big data dan teknik analisis memungkinkan lembaga jasa keuangan mendapatkan wawasan mendalam tentang perilaku dan pola nasabah, namun tantangannya masih terletak pada kemampuan organisasi untuk mengambil tindakan spesifik berdasarkan data ini.

- Privasi dan keamanan data: Data pelanggan terus menjadi perhatian. Regulasi masih belum diketahui secara luas: apa yang diperbolehkan dan apa yang tidak diperbolehkan secara hukum dalam kepemilikan dan penggunaan data pelanggan masih belum jelas, dan hal ini merupakan faktor penghambat adopsi yang cepat dan berskala besar.

12.5 SUMBER DAYA DATA KEUANGAN DAN ASURANSI YANG TERSEDIA

Sistem jasa keuangan memiliki beberapa kumpulan data utama yang dimiliki oleh pemangku kepentingan/pihak yang berbeda. Data diklasifikasikan menjadi tiga kategori utama:

1. **Data Terstruktur**

   Ini mengacu pada informasi dengan tingkat pengorganisasian yang tinggi, sehingga penyertaannya dalam database relasional dapat dilakukan dengan lancar dan mudah dicari dengan algoritma mesin pencari yang sederhana dan lugas, atau operasi pencarian lainnya. Contoh sumber data terstruktur keuangan adalah:

   a. Sistem perdagangan (data transaksi)
   b. Sistem akun (data kepemilikan dan pergerakan akun)
   c. Memasarkan data dari penyedia eksternal
   d. Data referensi sekuritas
   e. Informasi harga
   f. Indikator teknis

2. **Data Tidak Terstruktur**

   Meskipun industri keuangan sebelumnya berfokus pada data pasar berkecepatan tinggi, kini industri keuangan beralih ke data tidak terstruktur untuk mengubah dinamika perdagangan. Contoh data keuangan tidak terstruktur adalah:

   a. Umpan stok harian
   b. Pengumuman perusahaan (berita ad-hoc)
   c. Media berita online
   d. Artikel/blog
   e. Umpan balik/pengalaman pelanggan

3. **Data Semi Terstruktur**

   Suatu bentuk data terstruktur yang tidak sesuai dengan struktur formal model data yang terkait dengan database relasional atau bentuk tabel data lainnya, namun meskipun demikian mengandung tag atau penanda untuk memisahkan elemen semantik dan menegakkan hierarki catatan dan bidang di dalamnya data. Contoh data semi terstruktur dinyatakan dalam bahasa meta (kebanyakan berbasis XML) seperti:

   - Bahasa Markup Produk Keuangan (FpML)
   - Pertukaran Informasi Keuangan (FIX)
   - Pertukaran Keuangan Interaktif (IFX)
   - Bahasa Definisi Data Pasar (MDDL)

- Pertukaran Data Elektronik Keuangan (FEDI)
- Buka Pertukaran Finansial (OFX)
- Bahasa Pelaporan Bisnis yang Dapat Diperluas (XBRL)
- Standar SWIFT

Saat ini jumlah informasi tidak terstruktur di perusahaan adalah sekitar 80–85%. Industri keuangan dan asuransi memiliki gudang data terstruktur yang sangat besar dibandingkan dengan industri lain, dan sebagian besar informasi ini berasal dari dalam organisasi.

12.6 PERSYARATAN SEKTOR KEUANGAN DAN ASURANSI

Persyaratan Non-teknis

Perlindungan Data dan Privasi Khususnya di UE, ada banyak masalah perlindungan data dan privasi yang perlu dipertimbangkan ketika melakukan analisis big data. Persyaratan peraturan menyatakan bahwa data pribadi harus diproses untuk tujuan yang ditentukan dan sah serta pemrosesan tersebut harus memadai, relevan, dan tidak berlebihan. Dampak dari prinsip-prinsip ini terhadap organisasi jasa keuangan sangatlah signifikan, dimana individu dapat meminta organisasi jasa keuangan untuk menghapus atau tidak memproses data pribadi mereka dalam keadaan tertentu.

Persyaratan ini dapat menyebabkan peningkatan biaya bagi organisasi jasa keuangan, karena mereka menangani permintaan individu. Penghapusan data ini juga dapat menyebabkan kumpulan data menjadi tidak tepat, karena kelompok masyarakat tertentu akan lebih aktif dan sadar akan hak-hak mereka dibandingkan kelompok masyarakat lainnya.

Kerahasiaan dan Persyaratan Peraturan Segala informasi yang terkait dengan pihak ketiga yang tunduk pada analisis big data kemungkinan besar merupakan informasi rahasia. Oleh karena itu, organisasi jasa keuangan perlu memastikan bahwa mereka mematuhi kewajiban mereka dan bahwa setiap penggunaan data tersebut tidak menimbulkan pelanggaran terhadap kerahasiaan atau kewajiban peraturan mereka.

Masalah Tanggung Jawab Hanya karena big data berisi sejumlah besar informasi, bukan berarti big data tersebut mencerminkan sampel populasi yang mewakili. Oleh karena itu, ada risiko salah menafsirkan informasi yang dihasilkan dan tanggung jawab mungkin timbul jika mengandalkan informasi tersebut. Ini adalah faktor yang harus dipertimbangkan oleh organisasi jasa keuangan ketika mempertimbangkan penggunaan big data dalam model analitis dan memastikan bahwa ketergantungan pada output disertai dengan penafian yang relevan.

Persyaratan Teknis

Ekstraksi Data dan Klasifikasi Sentimen Meskipun definisi sentimen tidak jelas, secara umum sentimen terhadap suatu objek adalah pandangan, sikap, emosi, atau penilaian positif atau negatif terhadap atau dari penulis atau aktor dokumen.

Sentimen sering kali diungkapkan dengan cara yang spesifik untuk domain tertentu, dan penggunaan kosakata yang tidak spesifik untuk domain dapat menyebabkan kesalahan klasifikasi. Tujuannya adalah untuk mengekstrak fakta dan sentimen mengenai kasus penggunaan keuangan: instrumen keuangan, situasi, kondisi, indikator, dan penilaian para ahli

mengenai instrumen tersebut, serta sentimen investor, dll. Klasifikasi sentimen dapat dilakukan pada beberapa tingkatan: kata, frasa, kalimat, paragraf, dokumen, dan bahkan beberapa dokumen, lalu digabungkan.

Ekstraksi data perlu mengatasi gangguan, misinformasi, ironi, bias, atau ketidakpastian. Selain itu, dengan sentimen, penting tidak hanya untuk menentukan sentimen suatu informasi, namun bagaimana kata-kata mempengaruhi orientasi semantik dan bagaimana sentimen berubah.

Kualitas Data Semakin tepat waktu, akurat, dan relevan data (serta analisis yang baik), semakin baik penilaian terhadap kondisi keuangan saat ini. Hal ini memerlukan proses yang lebih baik dalam mengidentifikasi dan memelihara sumber data yang diinginkan, memverifikasi, membersihkan, mengubah, mengintegrasikan, dan menghapus duplikasi data. Karena banyaknya data yang tersedia, diperlukan otomatisasi dan skalabilitas proses. Metode deteksi bahasa juga perlu disempurnakan untuk meningkatkan presisi dan keandalan.

Akuisisi Data Bagi bank dan penyedia jasa keuangan, volume data yang dihasilkan, dikonsumsi, disimpan, dan diakses akan meningkat secara eksponensial dari tahun ke tahun. Penerapannya bergantung pada perolehan dan akses sejumlah besar informasi historis yang heterogen dan informasi langsung yang tidak terstruktur, semi terstruktur, dan terstruktur. Sejumlah besar data berasal dari data terstruktur internal, meskipun ada kecenderungan yang meningkat terhadap data eksternal yang tidak terstruktur (dari berita, blog, artikel, jejaring sosial, dan situs web). Meskipun terdapat beragam sumber data yang dapat diakses, sumber data aktual yang diperlukan bergantung pada desain aplikasi tertentu.

Integrasi/Berbagi Data Ini menjelaskan tugas untuk mengatasi heterogenitas sumber data yang berbeda dalam hal perangkat keras, perangkat lunak, sintaksis, dan/atau semantik dengan menyediakan alat akses yang memungkinkan interoperabilitas. Data biasanya tersebar di antara sumber-sumber heterogen yang berbeda dengan representasi konseptual yang berbeda (struktur dan semantik data yang berbeda) tetapi data tersebut dikemas menjadi satu sumber data yang homogen bagi pengguna akhir.

Motivasi integrasi mungkin didasarkan pada pertimbangan strategis atau operasional. Mengenai pertimbangan dan analisis strategis, mungkin tidak diperlukan untuk terus-menerus mengintegrasikan data namun untuk mengintegrasikan cuplikan data pada titik waktu tertentu. Untuk analisis operasional, integrasi informasi terkini mungkin diperlukan secara real-time.

Biasanya integrasi data bukanlah konversi yang dilakukan sekali saja namun merupakan tugas yang berkelanjutan, oleh karena itu menimbulkan kendala tambahan bahwa solusi yang dipilih harus kuat dalam hal kemampuan beradaptasi, ekstensibilitas, dan skalabilitas. Pendekatan yang memanfaatkan standar seperti eXtensible Business Reporting Language (XBRL) dan Linked Data menunjukkan hasil yang menjanjikan (O'Ria´in dkk. 2012).

Aliran informasi berkelanjutan yang cepat ini telah menantang kemampuan penyimpanan, komputasi, dan komunikasi dalam sistem komputasi, karena hal tersebut memerlukan kebutuhan sumber daya yang tinggi pada sistem pemrosesan aliran data. Sistem Pendukung Keputusan (DSS) DSS berbasis model menekankan akses dan manipulasi model

statistik, finansial, optimalisasi, dan/atau simulasi. Model menggunakan data dan parameter untuk membantu pengambil keputusan dalam menganalisis suatu situasi, misalnya, menilai dan mengevaluasi alternatif keputusan dan memeriksa dampak perubahan. Hal ini memerlukan pengintegrasian informasi dari basis pengetahuan ke dalam model deteksi peristiwa keuangan, model visualisasi, model keputusan, dan pelaksanaan model-model ini secara terukur.

Untuk beberapa skenario aplikasi, respons sistem harus mendukung wawasan real-time atau mendekati real-time. Kecepatan respons bergantung pada kebutuhan pengguna akhir. Dalam DSS, visualisasi adalah alat yang sangat berguna untuk memberikan gambaran umum dan wawasan terhadap sejumlah besar data untuk mendukung proses pengambilan keputusan.

Privasi dan Keamanan Data Prioritas utama sektor keuangan saat ini mencakup kepatuhan terhadap peraturan yang berkelanjutan [misalnya. Sarbanes-Oxley (SOX) Act, Pemerintah AS (2002); Arahan perlindungan data UE, Parlemen (1995); arahan keamanan siber, Parlemen (2013)] dan mitigasi risiko, adaptasi berkelanjutan terhadap harapan konsumen terhadap layanan di mana saja/kapan saja, mengurangi biaya operasional, dan meningkatkan efisiensi melalui penggunaan layanan berbasis cloud.

Perbankan dan lembaga keuangan perlu mengamankan penyimpanan, transit, dan penggunaan data perusahaan dan pribadi di seluruh aplikasi bisnis, termasuk perbankan online dan komunikasi elektronik atas informasi dan dokumen sensitif. Sifat industri yang semakin global dan interkonektivitas yang tinggi mengharuskan kita untuk menangani peraturan privasi dan keamanan data internasional secara komprehensif, dari depan hingga belakang, dan di sepanjang Jaringan pasokan, termasuk pihak ketiga. Data tidak selalu disimpan secara internal tetapi pada pihak ketiga. Penggunaan layanan “cloud” komersial sebagai lokasi penyimpanan data menimbulkan potensi masalah privasi dan keamanan karena persyaratan layanan untuk produk ini sering kali kurang dipahami.

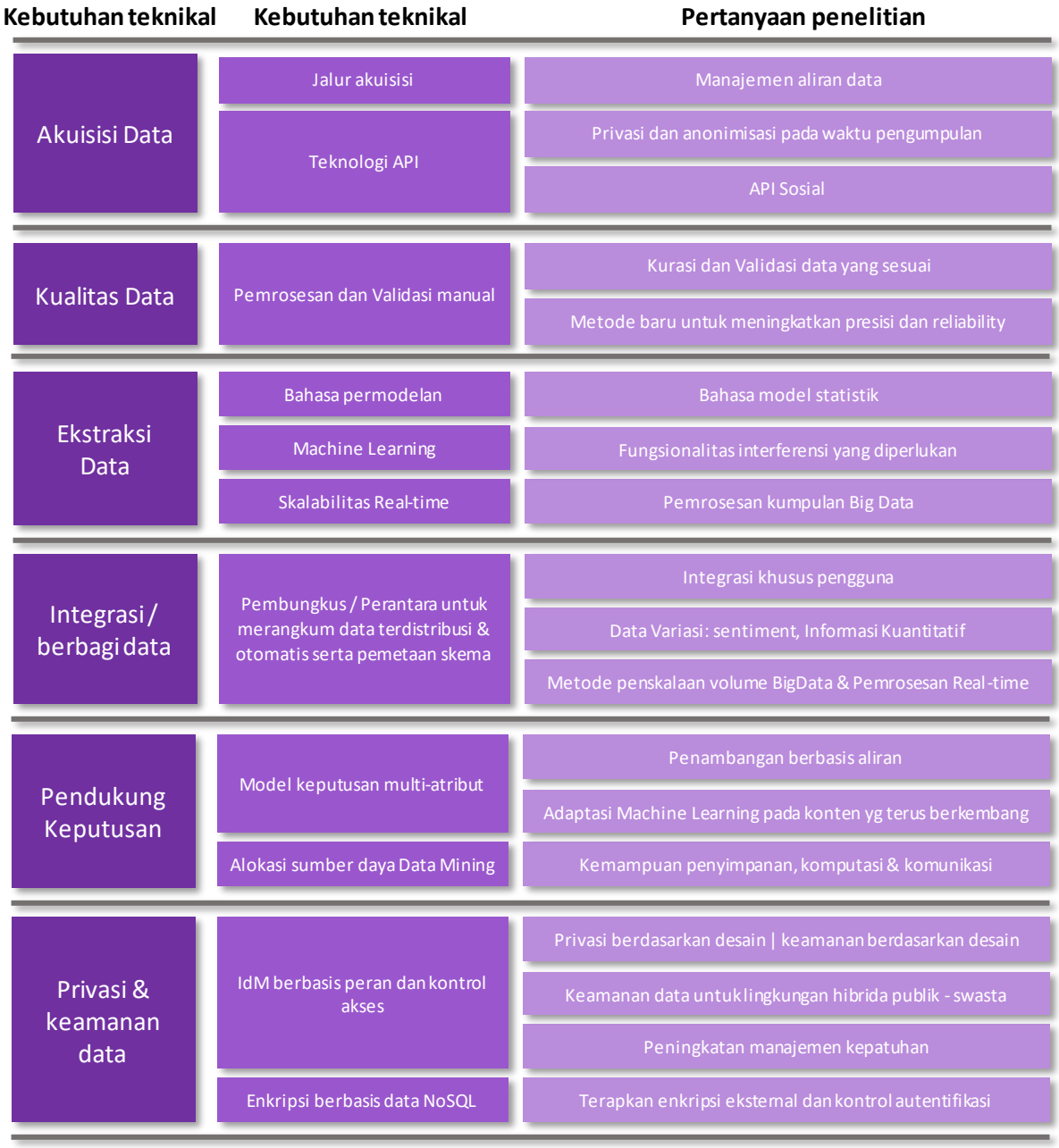
12.7 PETA JALAN TEKNOLOGI BIG DATA DI SEKTOR KEUANGAN DAN ASURANSI

Untuk setiap kebutuhan di sektor ini, bagian ini menyajikan teknologi yang dapat diterapkan dan pertanyaan penelitian yang akan dikembangkan (Gambar 12.1; Tabel 12.2). Semua referensi yang disajikan di sini berasal dari Curry dkk. (2014).

Tabel 12.2 Jangka waktu hasil utama yang diharapkan dari peta jalan big data untuk sektor keuangan dan asuransi

PERSYARATAN TEKNIS	TAHUN 1	TAHUN 2	TAHUN 3	TAHUN 4	TAHUN 5
Akuisisi data			API Sosial		Manajemen aliran data. Privasi dan anonimisasi pada waktu pengumpulan
Kualitas data			Kurasi dan validasi data yang skalabel		Metode baru untuk meningkatkan presisi dan keandalan
Ekstraksi data		Model bahasa statistik	Teknik pembelajaran mesin baru untuk memenuhi		Pemrosesan kumpulan data besar

			fungsionalitas inferensi baru yang dibutuhkan		
Integrasi/berbagi data					Integrasi khusus pengguna Keanekaragaman data: sentimen, informasi kuantitatif Metode penskalaan untuk volume data besar dan pemrosesan hampir real-time
Pendukung keputusan		Penambangan data berbasis aliran			Adaptasi pembelajaran mesin terhadap konten yang terus berkembang Peningkatan kemampuan penyimpanan, komputasi, dan komunikasi
Privasi dan keamanan data		Terapkan enkripsi eksternal dan kontrol autentikasi	Privasi berdasarkan desain | Keamanan berdasarkan desain		Keamanan data untuk lingkungan hibrida publik-swasta Peningkatan manajemen kepatuhan



Gambar 12.1 Pemetaan pertanyaan penelitian di sektor keuangan dan asuransi

Akuisisi Data

1) **Teknologi saluran akuisisi.**
   Pertanyaan Penelitian: Manajemen aliran data. Analisis data saat ini dalam domain data yang disimpan perlu beralih ke pengelolaan data dalam aliran data itu sendiri.
2) **Teknologi API yang dipatenkan**.
   Pertanyaan Penelitian: Privasi dan anonimisasi pada waktu pengumpulan. Proses pengumpulan data memerlukan anonimisasi data intrinsik dan/atau pemisahan data pribadi dari data yang berasal dari proses bisnis atau lainnya.

Pertanyaan Penelitian: API Sosial. Melangkah lebih maju dari API kepemilikan (atau bahkan terbuka) yang sudah ada, API sosial ke dalam kumpulan data layanan keuangan perlu diselidiki.

Kualitas Data

Teknologi pemrosesan dan validasi manual.

Pertanyaan Penelitian: Kurasi dan validasi data yang skalabel.

Pertanyaan Penelitian: Metode baru untuk meningkatkan presisi dan keandalan.

Ekstraksi Data

1) **Teknologi pemodelan bahasa.**

   Pertanyaan Penelitian: Mendapatkan kata kunci dan frase kunci dengan menggunakan model bahasa statistik.

2) **Teknologi Pembelajaran Mesin**.

   Pertanyaan Penelitian: Besarnya kumpulan data dalam layanan keuangan mengharuskan teknik pembelajaran mesin baru untuk memenuhi fungsi inferensi baru yang diperlukan.

3) **Skalabilitas dalam teknologi real-time**: Informasi real-time merupakan hal yang menarik dalam beberapa skenario penerapan jasa keuangan.

   Pertanyaan Penelitian: Tantangan dalam memproses kumpulan data yang besar merupakan persyaratan bagi penelitian dalam skalabilitas pemrosesan data secara real-time seiring dengan bertambahnya jumlah dan ukuran kumpulan data.

Integrasi/Berbagi Data

Pembungkus/mediator untuk merangkum data terdistribusi dan data otomatis serta teknologi pemetaan skema: Sumber data dalam industri jasa keuangan dapat didistribusikan ke seluruh organisasi, atau melintasi ruang dan waktu.

Pertanyaan Penelitian: Integrasi khusus pengguna. Integrasi data untuk kepentingan pengguna tertentu (yaitu, proses bisnis, atau organisasi pengguna akhir sasaran).

Pertanyaan Penelitian: Variasi data: sentimen, informasi kuantitatif.

Pertanyaan Penelitian: Metode penskalaan untuk volume data besar dan pemrosesan hampir real-time. Tantangan penelitian ini terkait dengan “skalabilitas dalam waktu nyata” yang dijelaskan sebelumnya, dalam “ekstraksi data”.

Pendukung Keputusan

1) **Teknologi model keputusan multi-atribut:** Ketersediaan informasi dari berbagai sumber akan menyediakan berbagai jenis atribut yang tersedia untuk disertakan dalam model keputusan.

   Pertanyaan Penelitian: Penambangan data berbasis aliran.

   Pertanyaan Penelitian: Adaptasi pembelajaran mesin terhadap konten yang terus berkembang.

2) **Alokasi sumber daya dalam teknologi aliran data penambangan:** Komputasi elastis saat ini memungkinkan alokasi sumber daya dinamis sesuai kebutuhan. Perbaikan mungkin diperlukan dalam alokasi sumber daya untuk mendukung pengambilan keputusan hampir secara real-time.

Pertanyaan Penelitian: Peningkatan kemampuan penyimpanan, komputasi, dan komunikasi.

Privasi dan Keamanan Data

1) **Manajemen identitas berbasis peran dan teknologi kontrol akses**: kontrol akses dalam konteks kumpulan data yang besar akan menimbulkan masalah ketika data sensitif (yang terkait dengan proses bisnis) mulai dieksploitasi dalam kumpulan data yang besar dan diintegrasikan dengan data lain, serta diakses oleh pihak ketiga.
   Pertanyaan Penelitian: Privasi berdasarkan desain
2) **Keamanan berdasarkan desain**. Kemajuan dalam "privasi berdasarkan desain" untuk menghubungkan kebutuhan analitik dengan kontrol perlindungan dalam pemrosesan dan penyimpanan.
   Pertanyaan Penelitian: Keamanan Data untuk lingkungan hibrida publik-swasta.
   Namun, munculnya penyimpanan cloud dan layanan komputasi mengorbankan keamanan data dan privasi pengguna.
   Pertanyaan Penelitian: Peningkatan manajemen Kepatuhan (perlindungan data, lainnya). Penelitian telah dimulai, namun perlu terus menyediakan metodologi dan infrastruktur yang memfasilitasi pemantauan, penegakan, dan audit indikator-indikator terukur mengenai keamanan proses bisnis.
3) **Teknologi enkripsi basis data:** Konsep keamanan basis data NoSQL umumnya bergantung pada mekanisme penegakan hukum eksternal.
   Pertanyaan Penelitian: Tinjau arsitektur keamanan dan kebijakan sistem secara keseluruhan dan terapkan enkripsi eksternal dan kontrol otentikasi untuk melindungi database NoSQL. Keamanan data untuk lingkungan hibrida publik-swasta.

Kesimpulan dan Rekomendasi untuk Sektor Keuangan dan Asuransi

Analisis sektor keuangan dan asuransi untuk peta jalan ini didasarkan pada empat skenario penerapan utama berdasarkan pemanfaatan data milik bank dan perusahaan asuransi untuk menciptakan nilai bisnis baru. Temuan analisis ini menunjukkan bahwa masih terdapat tantangan penelitian untuk mengembangkan teknologi secara maksimal guna memberikan solusi yang kompetitif dan efektif. Tantangan-tantangan ini muncul di semua tingkat Jaringan nilai big data dan melibatkan serangkaian teknologi berbeda, sehingga memerlukan prioritas investasi dalam penelitian dan pengembangan. Secara umum tampaknya terdapat kesepakatan umum mengenai aspek real-time, teknik kualitas data yang lebih baik, skalabilitas pengelolaan dan pemrosesan data, metode klasifikasi sentimen yang lebih baik, dan kepatuhan terhadap persyaratan keamanan di sepanjang Jaringan pasokan. Namun, perlu disebutkan pentingnya skenario penerapan dan kebutuhan nyata pengguna akhir untuk menentukan prioritas ini. Pada saat yang sama, selain aspek teknologi, terdapat faktor organisasi, budaya, dan hukum yang akan memainkan peran penting dalam cara pasar jasa keuangan memanfaatkan data besar untuk operasional dan pengembangan bisnisnya.

BAB 13
BIG DATA DI SEKTOR ENERGI DAN TRANSPORTASI

13.1 PENDAHULUAN

Sektor energi dan transportasi saat ini sedang mengalami dua transformasi utama: digitalisasi dan liberalisasi. Kedua transformasi ini mengedepankan karakteristik khas skenario big data: sensor, komunikasi, komputasi, dan kemampuan kontrol melalui peningkatan digitalisasi dan otomatisasi infrastruktur untuk efisiensi operasional yang menghasilkan data bervolume tinggi dan berkecepatan tinggi. Di pasar yang sudah liberal, potensi big data dapat diwujudkan dalam skenario konsumerisasi dan ketika beragam data lintas batas organisasi dimanfaatkan.

Di kedua sektor tersebut, terdapat konotasi bahwa istilah "big data" saja tidak cukup: meningkatnya sumber daya komputasi yang tertanam dalam infrastruktur juga dapat dimanfaatkan untuk menganalisis data guna menghasilkan "smart data". Taruhannya besar, karena peluang optimalisasi multimoda berada pada infrastruktur penting seperti sistem tenaga listrik dan perjalanan udara, yang dapat membahayakan nyawa manusia, bukan hanya sumber pendapatan.

Untuk mengidentifikasi kebutuhan dan persyaratan industri terhadap teknologi big data, analisis dilakukan terhadap sumber data yang tersedia di bidang energi dan transportasi serta kasus penggunaannya dalam berbagai kategori untuk nilai big data: efisiensi operasional, pengalaman pelanggan, dan model bisnis baru. Sektor energi dan transportasi hampir sama dalam hal karakteristik utama mengenai kebutuhan dan persyaratan big data serta tren masa depan. Kawasan khusus adalah kawasan perkotaan dimana seluruh kompleksitas dan optimalisasi potensi sektor energi dan transportasi difokuskan pada kawasan regional yang terkonsentrasi.

Kebutuhan utama sektor ini adalah representasi virtual dari sistem fisik yang mendasarinya melalui sensor, perangkat pintar, atau yang disebut perangkat elektronik cerdas serta pemrosesan dan analisis data dari perangkat tersebut. Penerapan teknologi big data yang sudah ada seperti yang digunakan oleh para big data native saja tidak akan cukup. Teknologi big data khusus domain diperlukan dalam sistem cyber-fisik untuk energi dan transportasi. Privasi dan kerahasiaan yang menjaga pengelolaan dan analisis data merupakan perhatian utama semua pemangku kepentingan energi dan transportasi yang menangani data pelanggan. Tanpa memenuhi kebutuhan akan privasi dan kerahasiaan, akan selalu ada ketidakpastian peraturan dan hambatan dalam penerimaan pelanggan terhadap penawaran baru berbasis data.

13.2 BIG DATA DI SEKTOR ENERGI DAN TRANSPORTASI

Bagian berikut ini mengkaji dimensi big data di bidang energi dan transportasi untuk mengidentifikasi kebutuhan bisnis dan pengguna akhir sehubungan dengan teknologi big data dan penggunaannya.

Data Industri Ketenagalistrikan berasal dari gardu induk digitalisasi, gardu trafo, dan gardu distribusi lokal dalam infrastruktur jaringan listrik yang kepemilikannya tidak dibundel. Informasi dapat datang dalam bentuk laporan servis dan pemeliharaan dari kru lapangan tentang perbaikan rutin dan tidak terduga, data sensor kesehatan dari aset yang dipantau sendiri, data penggunaan akhir dan pasokan daya dari meter pintar, dan data real-time beresolusi tinggi. dari unit pengukuran fasor yang disinkronkan GPS atau perangkat proteksi dan relai cerdas. Contoh kasus penggunaan berasal dari E´ lectricite´ de France (EDF) (Picard 2013), di mana mereka saat ini melakukan pembacaan meter standar sebulan sekali. Dengan smart meter, utilitas harus memproses data dengan interval 15 menit. Jumlah ini merupakan peningkatan sekitar 3000 kali lipat dalam pemrosesan data harian untuk sebuah perusahaan utilitas, dan ini hanyalah gelombang pertama dari banjir data. Data berasal dari kurva beban individual, data cuaca, informasi kontrak; data jaringan 1 diukur setiap 10 menit untuk target 35 juta pelanggan.

Perkiraan volume data tahunan adalah 1.800 miliar catatan atau 120 TB data mentah. Gelombang kedua akan mencakup data granular dari peralatan pintar, kendaraan listrik, dan titik pengukuran lainnya di seluruh jaringan listrik. Hal ini akan meningkatkan jumlah data yang dihasilkan secara eksponensial.

Data Industri Minyak dan Gas Bumi berasal dari stasiun penyimpanan dan distribusi digital, namun sumur, kilang, dan stasiun pengisian juga menjadi sumber data dalam infrastruktur cerdas perusahaan minyak dan gas terintegrasi. Sensor lubang bawah dari lokasi produksi mengirimkan data secara real-time termasuk pengukur tekanan, suhu, dan getaran, pengukur aliran, akustik dan elektromagnetik, padatan sirkulasi. Data lain berasal dari sumber seperti vendor, kru layanan pelacakan, pengukuran lalu lintas truk, peralatan dan rekahan hidrolik, penggunaan air; Data Supervisory Control and Data Acquisition (SCADA) dari kejadian katup dan pompa, parameter pengoperasian aset, alarm di luar kondisi; data cadangan tidak terstruktur, data geospasial, catatan insiden keselamatan, dan aliran video pengawasan. Contoh kasus penggunaan berasal dari Shell (Mearian 2012) di mana "serat optik yang dipasang ke sensor lubang bawah menghasilkan sejumlah besar data yang disimpan di bagian pribadi yang terisolasi di Amazon Web Services. Mereka telah mengumpulkan 46 petabyte data dan pengujian pertama yang mereka lakukan di satu sumur minyak menghasilkan 1 petabyte informasi. Mengetahui bahwa mereka ingin menyebarkan sensor tersebut ke sekitar 10.000 sumur minyak, kita berbicara tentang 10 Exabytes data, atau 10 hari dari semua data yang dibuat di Internet. Karena kumpulan data yang sangat besar ini, Shell mulai melakukan uji coba dengan Hadoop di Amazon Virtual Private Cloud". Contoh lain dalam industri ini termasuk (Nicholson 2012): "Bukti konsep Chevron menggunakan Hadoop untuk pemrosesan data seismik; Proyek Cloudera Seismic Hadoop yang menggabungkan Seismic Unix dengan Apache Hadoop; Server Data Seismik PointCross dan Server Data Pengeboran menggunakan Hadoop dan NoSQL".

Transportasi Di bidang transportasi jumlah sumber data meningkat pesat. Pelabuhan udara dan laut, stasiun kereta api dan bus, pusat logistik, dan gudang semakin banyak menggunakan sensor: Electronic on board recorder (EOBR) di truk yang mengirimkan data

tentang waktu bongkar/muat, waktu perjalanan, jam kerja pengemudi, catatan pengemudi truk, tag palet atau trailer yang mengirimkan data tentang waktu transit dan waktu tunggu, informasi tentang pemogokan di pelabuhan, jadwal angkutan umum, sistem tarif dan kartu pintar, survei pengendara, pembaruan GPS dari armada kendaraan, volume data tradisional yang lebih banyak dari sumber yang sudah mapan seperti program frequent flyer , dll. Contoh kasus penggunaan berasal dari Kota Dublin (Tabbitt 2014) di mana "departemen jalan dan lalu lintas kini dapat menggabungkan aliran data besar dari berbagai sumber—jadwal bus, detektor lalu lintas loop induktif, sirkuit tertutup kamera televisi, dan pembaruan GPS yang dikirimkan oleh masing-masing dari 1000 bus kota setiap 20 detik—untuk membuat peta digital kota yang dilapis dengan posisi bus Dublin secara real-time menggunakan komputasi aliran dan data geospasial. Beberapa intervensi telah menghasilkan pengurangan waktu perjalanan sebesar 10–15%".

13.3 ANALISIS KEBUTUHAN INDUSTRI SEKTOR ENERGI DAN TRANSPORTASI

Kebutuhan bisnis dapat diperoleh dari dimensi big data sebelumnya dan contoh-contoh dari sektor energi dan transportasi, Kemudahan penggunaan teknologi big data pada umumnya akan memastikan adopsi skala luas. Teknologi big data menerapkan paradigma baru dan sebagian besar menawarkan akses terprogram. Pengguna memerlukan keterampilan pengembangan perangkat lunak dan pemahaman mendalam tentang paradigma komputasi terdistribusi serta pengetahuan tentang penerapan algoritma analisis data dalam lingkungan terdistribusi tersebut. Hal ini di luar keahlian sebagian besar pengguna bisnis. Semantik korelasi dan anomali yang dapat ditemukan dan divisualisasikan melalui analisis big data perlu dibuat dapat diakses. Saat ini hanya pakar domain dan data yang dapat menafsirkan outlier data; pengguna bisnis sering kali hanya menebak-nebak saat melihat hasil analisis data.

Kebenaran data perlu dijamin sebelum digunakan dalam aplikasi energi dan transportasi. Karena peningkatan data yang akan digunakan untuk aplikasi ini akan jauh lebih besar, aturan sederhana atau pemeriksaan masuk akal secara manual tidak lagi berlaku. Data pintar sering digunakan oleh pemangku kepentingan industri untuk menekankan bahwa pengguna bisnis industri memerlukan data yang disempurnakan tidak harus seluruh data mentah (data besar) tetapi tanpa kehilangan informasi dengan hanya berkonsentrasi pada data kecil yang relevan saat ini. Dalam sistem cyber-fisik dibandingkan dengan bisnis online, terdapat teknologi informasi dan komunikasi (TIK) yang tertanam di seluruh sistem, bukan hanya di backend TI perusahaan. Operator infrastruktur memiliki peluang untuk melakukan pra-pemrosesan data di lapangan, mengumpulkan data, dan mendistribusikan kecerdasan untuk analisis data di seluruh infrastruktur TIK guna memanfaatkan sumber daya komputasi dan komunikasi sebaik-baiknya guna menangani volume dan kecepatan data sensor massa.

Dukungan keputusan dan otomatisasi menjadi kebutuhan inti seiring dengan perubahan kecepatan dan struktur bisnis. Operator jaringan listrik di Eropa saat ini perlu melakukan intervensi hampir setiap hari untuk mencegah potensi pemadaman listrik skala besar, misalnya pada pemadaman listrik skala besar. karena integrasi pasar energi terbarukan atau liberalisasi. Sistem manajemen lalu lintas menjadi semakin rumit seiring dengan

meningkatnya jumlah elemen yang terdigitalisasi dan dapat dikontrol. Pengguna bisnis membutuhkan lebih banyak informasi daripada "ada sesuatu yang salah". Visualisasi bisa sangat berguna, namun pertanyaan tentang apa yang perlu dilakukan masih harus dijawab baik secara real-time atau sebelum terjadinya suatu peristiwa, yaitu dengan cara yang prediktif.

Analitik tingkat lanjut yang terukur akan mendorong teknologi canggih. Misalnya, analisis data pengukuran cerdas (Picard 2013) mencakup segmentasi berdasarkan kurva beban, perkiraan wilayah lokal, penilaian kerugian non-teknis, pengenalan pola dalam kurva beban, pemodelan prediktif, dan analisis waktu nyata dengan cara yang cepat dan andal. untuk mengendalikan sistem yang rumit dan kompleks seperti jaringan listrik (Heyde dkk. 2010). Di sektor transportasi AS, nilai bisnis dari analitik real-time yang dapat diskalakan telah diperoleh dengan menggunakan sistem data besar untuk aplikasi otomasi skala penuh, misalnya. penjadwalan ulang otomatis yang membantu kereta beradaptasi secara dinamis terhadap peristiwa dan tepat waktu di area yang luas. Analisis data besar menawarkan banyak perbaikan bagi pengguna akhir. Efisiensi operasional pada akhirnya berarti efisiensi dan ketepatan waktu energi dan sumber daya, yang akan meningkatkan kualitas hidup—terutama di lingkungan mobilitas perkotaan.

Pengalaman pelanggan dan model bisnis baru yang terkait dengan skenario big data sepenuhnya didasarkan pada pelayanan yang lebih baik kepada pengguna akhir energi dan mobilitas. Namun, kedua skenario memerlukan data yang dipersonalisasi dalam resolusi lebih tinggi. Ada manfaat yang signifikan dalam menggabungkan berbagai data, yang pada sisi negatifnya dapat membuat nama samaran atau bahkan anonimisasi menjadi tidak efektif dalam melindungi identitas dan pola perilaku individu, atau pola bisnis dan strategi perusahaan. Model bisnis baru yang didasarkan pada monetisasi data yang dikumpulkan, dengan peraturan yang saat ini tidak jelas, membuat pengguna akhir tidak mendapat informasi sama sekali, dan tidak terlindungi dari penggunaan sekunder data mereka untuk tujuan yang mungkin tidak mereka setujui, misalnya untuk keperluan bisnis. klasifikasi asuransi, peringkat kredit, dll.

Transparansi terbalik berada di urutan teratas daftar keinginan pengguna akhir yang melek data. Analisis data perlu memberdayakan pengguna akhir untuk memahami penggunaan jalur data mereka. Akses dan penggunaan data pengguna akhir harus dapat dikonfigurasi secara efisien dan dinamis oleh pengguna akhir. Pengguna akhir memerlukan akses praktis terhadap informasi tentang data apa yang digunakan oleh siapa, dan untuk tujuan apa dengan cara yang mudah digunakan dan dikelola. Peraturan dan regulasi diperlukan untuk memberikan transparansi kepada pengguna akhir.

Akses, pertukaran, dan berbagi data untuk bisnis dan pengguna akhir. Di pasar listrik atau mobilitas antarmoda yang kompleks saat ini, hampir tidak ada skenario di mana semua data yang diperlukan untuk menjawab pertanyaan bisnis, atau teknik, berasal dari database satu departemen. Meskipun demikian, sebagian besar infrastruktur meteran canggih yang saat ini terpasang sudah mengunci data penggunaan energi yang diperoleh ke sistem penagihan perusahaan utilitas. Penguncian ini mempersulit penggunaan data energi untuk

analisis berharga lainnya. Penyimpanan data ini berakar dari masa ketika sebagian besar bisnis infrastruktur di Eropa merupakan perusahaan yang terintegrasi secara vertikal. Selain itu, jumlah data yang dipertukarkan jauh lebih sedikit, sehingga antarmuka, protokol, dan proses pertukaran data masih belum sempurna.

13.4 POTENSI PENERAPAN BIG DATA UNTUK SEKTOR ENERGI DAN TRANSPORTASI

Dalam upaya mengumpulkan berbagai persyaratan sektoral menuju ekonomi big data Eropa dan peta jalan teknologinya, penerapan big data di bidang energi dan transportasi telah dianalisis. Temuan yang sejalan dengan penelitian Gartner tentang kemajuan analitik (Kart 2013) adalah bahwa aplikasi big data dapat dikategorikan sebagai "efisiensi operasional", "pengalaman pelanggan", dan "model bisnis baru".

Efisiensi operasional adalah pendorong utama (Kart 2013) di balik investasi digitalisasi dan otomatisasi. Kebutuhan akan efisiensi operasional bermacam-macam, seperti meningkatkan margin pendapatan, kewajiban terhadap peraturan, atau mengatasi hilangnya pekerja terampil yang pensiun. Setelah percontohan teknologi big data disiapkan untuk menganalisis sejumlah besar data demi tujuan efisiensi operasional, perusahaan menyadari bahwa mereka sedang membangun peta digital untuk bisnis, produk, dan infrastruktur mereka dan bahwa peta tersebut dikombinasikan dengan berbagai sumber data. juga dapat memberikan wawasan tambahan di bidang bisnis lainnya seperti kondisi aset, pola penggunaan akhir, dll. Bagian selanjutnya dari bagian ini merinci skenario big data dan tantangan utama yang mencegah penerapan skenario ini di Eropa.

Efisiensi Operasional

Efisiensi operasional mencakup semua kasus penggunaan yang melibatkan peningkatan pemeliharaan dan pengoperasian secara real-time, atau secara prediktif, berdasarkan data yang berasal dari infrastruktur, stasiun, aset, dan konsumen. Vendor teknologi yang mengembangkan sensorisasi infrastruktur menjadi faktor pendukung utama. Permintaan pasar terhadap teknologi yang ditingkatkan semakin meningkat, karena hal ini membantu bisnis di sektor energi untuk mengelola risiko dengan lebih baik. Kompleksitas pasar listrik yang saling terhubung di seluruh Eropa, dengan integrasi energi terbarukan dan liberalisasi perdagangan listrik, memerlukan visibilitas yang lebih besar terhadap sistem yang mendasarinya dan aliran energi secara real-time. Sebagai aturan praktis, segala sesuatu yang memiliki kata sifat "pintar" termasuk dalam kategori ini: jaringan pintar, meteran cerdas, kota pintar, dan ladang pintar (minyak, gas). Beberapa contoh kasus penggunaan big data dalam efisiensi operasional adalah sebagai berikut:

- Analisis gangguan pada sistem tenaga listrik secara prediktif dan real-time serta tindakan penanggulangan yang hemat biaya.
- Sistem perencanaan, pemantauan, dan pengendalian kapasitas operasional untuk pasokan dan jaringan energi, penetapan harga yang dinamis.
- Mengoptimalkan jaringan multimoda di bidang energi dan transportasi khususnya di wilayah perkotaan, seperti logistik kota atau eCar-sharing sehingga konsumsi energi

dan feed-in ke pusat-pusat transportasi dapat dioptimalkan secara silang dengan logistik.

Semua skenario dalam kategori ini memiliki tantangan besar dalam menghubungkan silo data: baik antar departemen dalam perusahaan yang terintegrasi secara vertikal, atau antar organisasi di sepanjang Jaringan nilai ketenagalistrikan. Kasus penggunaan big data dalam skenario efisiensi operasional memerlukan integrasi data, komunikasi, dan analitik yang lancar di berbagai sumber data, yang dimiliki oleh berbagai pemangku kepentingan.

Pengalaman Pelanggan

Memahami peluang big data mengenai kebutuhan dan keinginan pelanggan merupakan hal yang sangat menarik bagi perusahaan di pasar konsumen yang sudah diliberalisasi seperti listrik, dimana hambatan masuk bagi pemain baru serta margin semakin berkurang. Loyalitas pelanggan dan peningkatan layanan berkelanjutan memungkinkan pemain energi tumbuh di pasar-pasar ini.

Beberapa contoh penggunaan big data untuk meningkatkan pengalaman pelanggan adalah sebagai berikut:

- Peningkatan layanan berkelanjutan dan inovasi produk, misalnya. penawaran tarif individual berdasarkan segmentasi pelanggan terperinci menggunakan meteran pintar atau data konsumsi tingkat perangkat.
- Manajemen siklus hidup aset yang prediktif, yaitu data dari mesin dan perangkat yang digabungkan dengan perencanaan sumber daya perusahaan dan data teknis untuk menawarkan layanan seperti logistik suku cadang cerdas sesuai permintaan.
- Manajemen sisi permintaan industri, yang memungkinkan produksi hemat energi dan meningkatkan daya saing bisnis manufaktur.

Tantangan utamanya adalah menangani kerahasiaan dan privasi pelanggan domestik dan bisnis sekaligus mengetahui dan mengantisipasi kebutuhan mereka. Pencetus data, pemilik data, dan pengguna data merupakan pemangku kepentingan berbeda yang perlu berkolaborasi dan berbagi data untuk mewujudkan skenario penerapan big data ini.

Model Bisnis Baru

Model bisnis baru berkisar pada monetisasi sumber data yang tersedia dan layanan data yang ada dengan cara baru. Ada beberapa kasus di mana sumber data atau analisis dari satu sektor mewakili wawasan para pemangku kepentingan di sektor lain. Analisis data energi dan mobilitas yang baru dimulai menunjukkan bahwa ada cara baru untuk menghasilkan nilai bisnis jika pengguna akhir memiliki sumber daya tersebut. Kemudian bisnisnya sepenuhnya berorientasi pada pelanggan dan layanan; sedangkan infrastruktur energi dan transportasi dengan pemangku kepentingan yang ada dimanfaatkan sebagai bagian dari layanan tersebut. Ini disebut model bisnis perantara.

Profil segmen konsumen energi, seperti profil prosumer untuk pasokan daya dari fotovoltaik, atau gabungan unit panas dan daya; atau profil sisi permintaan yang dikelola secara aktif, dll., dari penyedia layanan pengukuran juga dapat ditawarkan kepada pengecer energi kecil, operator jaringan, atau utilitas yang dapat memperoleh manfaat dari perbaikan

profil standar penggunaan energi namun belum memiliki akses terhadap resolusi tinggi data energi pelanggan mereka sendiri.

Tantangan utamanya adalah memberikan peraturan yang jelas seputar penggunaan sekunder data energi dan mobilitas. Pengguna akhir yang terhubung merupakan prasyarat minimal untuk model bisnis baru yang berfokus pada konsumen ini. Segmen pasar baru didiversifikasi melalui perusahaan rintisan energi data besar seperti Next: Kraftwerke, yang "menggabungkan data dari berbagai sumber seperti data operasional dari pembangkit listrik virtual kami, data cuaca dan jaringan terkini, serta data pasar langsung. Hal ini memberikan Next Kraftwerke keunggulan dibandingkan pedagang listrik konvensional" (Kraftwerke 2014).

Dalam transportasi, 95% waktu mobil diparkir (Barter, 2013) dan menurut penelitian terbaru, satu kendaraan berbagi mobil menggantikan 32 pembelian kendaraan baru (AlixPartners 2014). Bisnis yang sebelumnya hanya berkisar pada produk kini beralih ke layanan berbasis data. Berbeda dengan sektor energi, langkah berani ini menunjukkan kesiapan para pelaku industri transportasi untuk memanfaatkan potensi nilai big data dari bisnis berbasis data.

13.5 PENDORONG DAN KENDALA BIG DATA DI BIDANG ENERGI DAN TRANSPORTASI

Pengemudi

Faktor pendorong utama di sektor energi dan transportasi adalah sebagai berikut:

- Peningkatan efisiensi infrastruktur energi dan transportasi serta operasional terkait.
- Sumber energi terbarukan telah mengubah seluruh kebijakan energi nasional, misalnya bahasa Jerman "Energiewende". Integrasi energi terbarukan memerlukan optimalisasi di berbagai bidang (misalnya jaringan listrik, pasar, dan penggunaan akhir atau penyimpanan) dan meningkatkan ketergantungan elektrifikasi pada prakiraan cuaca dan cuaca.
- Digitalisasi dan otomatisasi dapat secara signifikan meningkatkan efisiensi pengoperasian jaringan aliran seperti jaringan listrik, gas, air, atau transportasi. Jaringan infrastruktur ini akan semakin tersensor, sehingga menambah volume, kecepatan, dan variasi data industri secara signifikan.
- Komunikasi dan konektivitas diperlukan untuk mengumpulkan data guna optimalisasi dan otomatisasi kontrol. Perlu ada konektivitas dua arah dan multiarah antar perangkat lapangan, misalnya. perangkat elektronik cerdas di gardu induk jaringan listrik atau lampu lalu lintas.
- Data terbuka: Publikasi data operasional pada platform transparansi oleh operator jaringan listrik, oleh pasar pertukaran energi, dan oleh operator sistem transmisi gas merupakan kewajiban peraturan yang mendorong proyek-proyek akar rumput. Open Weather Map dan Open Street Map adalah contoh penyediaan data gratis yang dibuat pengguna dan sangat penting bagi kedua sektor.
- "Pergeseran keterampilan": Sebagai akibat dari pensiunnya pekerja terampil, seperti pengemudi truk atau operator jaringan listrik, timbullah kekurangan pengetahuan yang perlu segera diisi. Hal ini secara langsung berarti kenaikan harga bagi pelanggan,

karena gaji yang lebih tinggi perlu dibayar untuk menarik pekerja terampil yang tersisa di pasar. Dalam jangka menengah dan panjang, peningkatan efisiensi dan lebih banyak otomatisasi akan menjadi tren yang berlaku: seperti kendaraan tanpa pengemudi truk dalam transportasi atau sistem perlindungan dan pengendalian pemantauan area luas di bidang energi.

Kendala

Kendala pada sektor energi dan transportasi adalah sebagai berikut:

- Keahlian: Hanya sedikit orang yang dapat menerapkan pengetahuan manajemen big data dan analisis serta pengetahuan domain dalam sektor-sektor tersebut.
- Interpretasi: Model implisit atau diam-diam ada di kepala para pekerja terampil (pensiun). Ekstraksi model domain yang dapat diskalakan menjadi kuncinya, misalnya. dalam sistem manajemen lalu lintas, basis peraturan berkembang selama bertahun-tahun hingga menjadi semakin kompleks dan tidak dapat dikelola.
- Digitalisasi belum mencapai titik kritis: Digitalisasi dan otomasi infrastruktur memerlukan investasi awal, yang tidak dipertimbangkan dengan baik, atau bahkan sama sekali, oleh peraturan insentif yang mengikat para operator infrastruktur. Data real-time dengan resolusi lebih tinggi masih belum tersedia secara luas.
- Ketidakpastian mengenai hak-hak digital dan undang-undang perlindungan data: Ketidakjelasan pandangan mengenai kepemilikan data menghambat big data bagi pengguna akhir di segmen sektor energi dan transportasi (misalnya infrastruktur meteran pintar).
- Uni Eropa yang “terpecah secara digital”: Eropa memiliki yurisdiksi yang terfragmentasi dalam hal hak-hak digital.
- “Bisnis seperti biasa” mengalahkan “bisnis berbasis data”: Dalam bisnis yang sudah mapan, sangat sulit untuk mengubah Jaringan nilai bisnis yang sudah berjalan. Petahana perlu menghadapi banyak perubahan: perubahan dalam siklus inovasi panjang yang ada, perubahan terhadap sistem tertutup dan silo, dan perubahan pola pikir agar TIK menjadi faktor pendukung (enabler) atau bahkan kompetensi inti (core) dalam inovasi perusahaan mereka.
- Kurangnya penerimaan pengguna akhir: Di sektor energi sering kali dikatakan bahwa masyarakat tidak tertarik dengan data penggunaan energi. Namun, ketika ada argumen yang menyebutkan hilangnya penerimaan pengguna akhir terhadap suatu teknologi, hal ini lebih merupakan pernyataan bahwa layanan berguna yang menggunakan teknologi ini belum diterapkan.
- Hilangnya kepercayaan: Kepercayaan adalah masalah yang dapat dan harus diatasi dengan perlindungan data teknologi dan kerangka peraturan (yaitu undang-undang perlindungan privasi yang sesuai).

13.6 SUMBER DAYA DATA ENERGI DAN TRANSPORTASI YANG TERSEDIA

Ketika potensi big data dieksplorasi dalam kedua sektor tersebut, semakin jelas bahwa daftar sumber data yang tersedia akan bertambah dan masih belum lengkap. Pengamatan

utamanya adalah bahwa keragaman sumber data yang digunakan untuk menemukan jawaban atas pertanyaan bisnis atau teknik merupakan pembeda dari bisnis seperti biasa.

1. Data infrastruktur mencakup jalur transmisi dan distribusi tenaga listrik, serta jaringan pipa untuk minyak, gas, atau air. Di bidang transportasi, infrastruktur terdiri dari jalan raya, kereta api, udara dan laut. Pertanyaan pendorongnya adalah kapasitas. Apakah jalanan macet? Apakah saluran listrik kelebihan beban?
2. Stasiun dianggap sebagai bagian dari infrastruktur. Dalam urusan bisnis dan teknik, sektor ini memainkan peranan khusus karena mencakup aset utama infrastruktur di wilayah yang padat, dan mempunyai nilai ekonomi yang tinggi. Pertanyaan pendorong utama adalah status saat ini dan tingkat pemanfaatan, yaitu kapasitas efektif infrastruktur. Apakah saluran transmisi terbuka atau tertutup? Apakah ditutup karena ada kesalahan pada saluran? Apakah kereta bawah tanah tertunda? Apakah karena kendala teknis?
3. Data yang diberi stempel waktu dan geo-tag diperlukan dan semakin banyak tersedia, terutama data yang tersinkronisasi GPS di kedua sektor, namun juga data GSM untuk menelusuri mobilitas dan mengekstraksi pola mobilitas.
4. Data cuaca, selain data geolokasi, merupakan sumber data yang paling banyak digunakan di kedua sektor. Sebagian besar konsumsi energi disebabkan oleh pemanasan dan pendinginan, yang merupakan pola konsumsi yang sangat bergantung pada cuaca. Dengan sumber daya energi terbarukan, pasokan listrik ke jaringan listrik menjadi bergantung pada cuaca.
5. Data dan pola penggunaan, indikator, dan nilai turunan dari penggunaan akhir dari masing-masing sumber daya dan infrastruktur, baik energi maupun transportasi, dapat diperoleh dengan berbagai cara, misalnya melalui transportasi. dalam infrastruktur pintar, melalui pengukuran di stasiun-stasiun di tepi jaringan, atau perangkat pintar.
6. Pola perilaku mempengaruhi penggunaan energi dan pola mobilitas serta dapat diprediksi. Aspek etika dan sosial menjadi perhatian dan batu sandungan utama. Dampak positif seperti pengalaman konsumen yang lebih baik, efisiensi energi, transparansi yang lebih baik, dan penetapan harga yang adil harus dibandingkan dengan dampak negatifnya.
7. Sumber data dalam lanskap TI horizontal, termasuk data yang berasal dari sumber seperti alat CRM, perangkat lunak akuntansi, dan data historis yang berasal dari sistem bisnis biasa. Potensi nilai dari penggabungan silang data historis dengan sumber data baru yang berasal dari peningkatan digitalisasi dan otomasi dalam sistem energi dan transportasi sangatlah tinggi.
8. Pada akhirnya, banyak sekali data pihak ketiga eksternal atau sumber data terbuka yang penting untuk skenario big data di sektor energi dan transportasi, termasuk data makro-ekonomi, data lingkungan (jasa meteorologi, model/simulasi cuaca global), data pasar (informasi perdagangan, spot and forward, berita bisnis), aktivitas manusia (web, telepon, dll.), informasi penyimpanan energi, data geografis, prediksi

berdasarkan Facebook dan Twitter, dan komunitas informasi seperti Open Energy Information.

13.7 PERSYARATAN SEKTOR ENERGI DAN TRANSPORTASI

Kebutuhan pengguna bisnis dan pengguna akhir yang dianalisis, serta berbagai jenis kebutuhan berbagi data secara langsung diterjemahkan ke dalam persyaratan teknis dan non-teknis.

Persyaratan Non-teknis

Beberapa persyaratan non-teknis di sektor-sektor tersebut diidentifikasi:

a. Investasi dalam komunikasi dan keterhubungan: Komunikasi broadband, atau ICT secara umum, perlu tersedia secara luas di seluruh Eropa dan bersamaan dengan infrastruktur energi dan transportasi untuk akses data real-time. Keterhubungan perlu diperluas ke pengguna akhir agar mereka dapat terus terhubung.
b. Uni Eropa yang bersatu secara digital: Biaya roaming telah menghalangi pengguna akhir di Eropa untuk menggunakan aplikasi intensif data melintasi batas negara. Penyedia layanan berbasis data di Eropa—terutama perusahaan rintisan yang mencari skalabilitas model bisnis mereka—sebagian besar berfokus pada pasar AS, dan bukan 27 negara anggota UE lainnya karena perbedaan peraturan terkait data. Para pemangku kepentingan di Eropa memerlukan peraturan dan regulasi yang dapat diandalkan dan minimal konsisten mengenai hak dan regulasi digital. Undang-undang hak digital seperti yang diserukan oleh penemu Web, Tim Berners-Lee, secara global merupakan langkah yang tepat dan harus didukung oleh Eropa.
c. Diperlukan tempat berkembang biak yang lebih baik bagi start-up dan budaya start-up, terutama bagi perubahan paradigma tekno-ekonomi seperti big data dan penyebaran digitalisasi, di mana bisnis baru sangat menyimpang dari bisnis seperti biasanya. Startup di bidang energi dan mobilitas memerlukan lebih dari sekedar investasi finansial namun juga kebebasan untuk melakukan eksplorasi dan eksperimen dengan data. Tanpa kebebasan ini, inovasi tidak akan mempunyai peluang, kecuali tentu saja teknik analisis pelestarian privasi yang disebutkan di atas tidak dapat dilakukan.
d. Data terbuka dalam hal ini merupakan peluang besar; Namun, standardisasi diperlukan. Jalur migrasi yang praktis diperlukan untuk menyederhanakan penerapan standar-standar mutakhir. Model data dan standar representasi akan memungkinkan pertumbuhan ekosistem data dengan penambangan data kolaboratif, granularitas data yang dapat dibagikan, dan teknik penyerta yang mencegah de-anonimisasi.
e. Orang yang ahli di bidang data: Pemrograman, statistik, dan alat terkait harus menjadi bagian dari pendidikan teknik. Analis data tradisional perlu memahami paradigma komputasi terdistribusi, misalnya. cara merancang algoritma yang berjalan pada sistem paralel yang sangat besar, cara memindahkan algoritma ke data, atau cara merekayasa jenis algoritma yang benar-benar baru.

Persyaratan Teknis

Beberapa persyaratan teknis diidentifikasi di sektor-sektor:

- Abstraksi: infrastruktur big data yang sebenarnya diperlukan untuk memungkinkan (a) kemudahan penggunaan, dan (b) ekstensibilitas dan fleksibilitas. Kasus penggunaan yang dianalisis memiliki persyaratan yang beragam sehingga tidak ada satu pun platform atau solusi big data yang akan memberdayakan bisnis utilitas di masa depan.
- Data adaptif dan model sistem diperlukan agar pengetahuan baru yang diambil dari analisis domain dan sistem dapat diterapkan kembali ke dalam kerangka analisis data tanpa mengganggu bisnis sehari-hari. Lapisan abstraksi harus mengakomodasi model adaptif plug-in.
- Interpretabilitas data harus terjamin tanpa keterlibatan pakar domain secara terus-menerus. Hasilnya harus dapat dilacak dan dijelaskan. Pengetahuan pakar dan domain harus dipadukan ke dalam manajemen data dan analitik.
- Analisis data diperlukan sebagai bagian dari setiap langkah mulai dari akuisisi data hingga penggunaan data. Dalam akuisisi data, analisis lapangan yang tertanam dapat meningkatkan kebenaran data dan dapat mendukung pengaturan privasi dan kerahasiaan yang berbeda pada sumber data yang sama untuk pengguna data yang berbeda, misalnya. penyedia jasa.
- Analisis real-time diperlukan untuk mendukung pengambilan keputusan, yang harus dibuat dalam rentang waktu yang semakin singkat. Dalam pengaturan smart grid, kontrol dinamis hampir real-time memerlukan wawasan pada sumber datanya.
- Data lake diperlukan dalam hal teknologi penyimpanan siap pakai (off-the-shelf) berbiaya rendah yang dikombinasikan dengan kemampuan untuk menyebarkan model data sesuai permintaan (“schema-on-read”) secara efisien, dibandingkan dengan solusi data warehouse yang biasa berupa ekstrak-transformasi -beban (ETL).
- Pasar data, data terbuka, logistik data, protokol standar yang mampu menangani keragaman, volume, dan kecepatan data, serta platform data diperlukan untuk berbagi data dan pertukaran data lintas batas organisasi.

13.8 PETA JALAN TEKNOLOGI SEKTOR ENERGI DAN TRANSPORTASI

Jaringan nilai big data untuk sistem bisnis energi dan transportasi yang berpusat pada infrastruktur dan sumber daya terdiri dari tiga fase utama: akuisisi data, pengelolaan data, dan penggunaan data. Analisis data, seperti yang ditunjukkan oleh kebutuhan pengguna bisnis, secara implisit diperlukan dalam semua langkah dan bukan merupakan fase terpisah.

Peta jalan teknologi untuk memenuhi persyaratan utama sepanjang Jaringan nilai data untuk sektor energi dan transportasi berfokus pada teknologi yang belum tersedia dan memerlukan penelitian dan pengembangan lebih lanjut untuk memenuhi persyaratan aplikasi energi dan transportasi yang lebih ketat (Gambar 13.1).

Gambar 13.1 Persyaratan pemetaan untuk pertanyaan penelitian di sektor energi dan transportasi

Akses dan Berbagi Data

Energi dan transportasi adalah bisnis infrastruktur yang berpusat pada sumber daya. Akses ke data penggunaan menciptakan peluang untuk menganalisis penggunaan produk atau layanan untuk meningkatkannya, atau mendapatkan efisiensi dalam penjualan dan operasi. Data penggunaan perlu digabungkan dengan data lain yang tersedia untuk menghasilkan model prediktif yang andal. Saat ini terdapat trade-off antara meningkatkan kemampuan interpretasi data dan menjaga privasi dan kerahasiaan. Contoh data penggunaan mobilitas berikut yang dikombinasikan dengan berbagai data lainnya menunjukkan tantangan privasi. de Montjoye dkk. (2013) menunjukkan bahwa "titik spatio-temporal (perkiraan tempat dan waktu) cukup untuk mengidentifikasi secara unik 95% dari 1,5 juta orang dalam database mobilitas. Studi lebih lanjut menyatakan bahwa kendala ini tetap ada bahkan ketika resolusi kumpulan data rendah". Penelitian ini menunjukkan bahwa kumpulan data mobilitas yang dikombinasikan dengan metadata dapat menghindari anonimitas.

Pada saat yang sama, pilihan perlindungan privasi yang tidak memadai dapat menghambat sumber data besar, seperti yang ditunjukkan oleh pengalaman penerapan pengukuran cerdas di bisnis energi. Di UE hanya 10 % rumah yang memiliki meteran pintar (Nunez 2012). Meskipun ada mandat bahwa teknologi ini dapat menjangkau 80% rumah pada tahun 2020, penerapannya di Eropa masih stagnan. Sebuah survei yang dilakukan pada tahun 2012 (Departemen Energi dan Perubahan Iklim 2012) menemukan bahwa "dengan meningkatnya frekuensi membaca, yaitu dari bulanan menjadi harian, menjadi setengah jam, dan seterusnya, data konsumsi energi mulai terasa lebih sensitif seiring dengan semakin

detailnya tingkat tampak mengganggu... Selain itu, tidak jelas bagi beberapa [peserta] mengapa ada orang yang menginginkan tingkat detail yang lebih tinggi, meninggalkan celah yang harus diisi dengan spekulasi yang mengakibatkan beberapa [peserta] menjadi lebih tidak nyaman". Kemajuan diperlukan untuk teknologi berikut untuk akses dan berbagi data:

- Data tertaut adalah praktik ringan untuk mengekspos dan menghubungkan potongan data, informasi, atau pengetahuan menggunakan standar web dasar. Hal ini menjanjikan untuk membuka kepemilikan data yang tersembunyi dan sudah menjadi faktor yang memungkinkan terjadinya data terbuka dan berbagi data. Namun, dengan semakin banyaknya sumber data yang terhubung, berbagai jenis data baru yang akan berasal dari infrastruktur cerdas, dan pengguna akhir yang selalu terhubung dalam hal energi dan mobilitas, skalabilitas dan efisiensi biaya menjadi sebuah permasalahan. Salah satu pertanyaan penelitian terbuka adalah bagaimana (semi-) otomatis mengekstraksi hubungan data untuk meningkatkan skalabilitas saat ini.
- Penyimpanan data terenkripsi dapat memungkinkan keamanan tingkat data yang terintegrasi. Ketika penyimpanan cloud menjadi hal yang lumrah bagi pengguna akhir domestik dan komersial, perlindungan data yang lebih baik dan ramah pengguna menjadi faktor pembeda (Tanner 2014). Untuk menjaga privasi dan kerahasiaan, penggunaan penyimpanan data terenkripsi akan menjadi pendukung dasar berbagi data dan analisis bersama. Namun, analisis pada data terenkripsi masih menjadi pertanyaan penelitian. Penelitian yang paling banyak dilakukan disebut enkripsi homomorfik sepenuhnya. Enkripsi homomorfik secara teoritis memungkinkan operasi dilakukan pada teks sandi. Hasilnya adalah ciphertext yang bila didekripsi cocok dengan hasil operasi pada plaintext. Saat ini hanya operasi dasar yang layak dilakukan.
- Asal data adalah seni menelusuri data melalui seluruh transformasi, analisis, dan interpretasi. Provenance memastikan bahwa data yang digunakan untuk menciptakan wawasan yang dapat ditindaklanjuti dapat diandalkan. Metadata yang dihasilkan untuk mewujudkan asal muasal berbagai macam kumpulan data dari berbagai sumber juga meningkatkan kemampuan interpretasi data, yang pada gilirannya dapat meningkatkan ekstraksi informasi otomatis. Namun, penskalaan asal data di seluruh dimensi big data masih merupakan pertanyaan penelitian yang terbuka.
- Privasi diferensial (Dwork dan Roth 2014) adalah definisi privasi yang ketat secara matematis (dan kerugiannya) dengan algoritma yang menyertainya. Hukum dasar pemulihan informasi (Dwork dan Roth 2014) menyatakan bahwa terlalu banyak pertanyaan dengan terlalu sedikit kesalahan akan mengungkap informasi sebenarnya. Tujuan dari pengembangan algoritme yang lebih baik adalah untuk mendorong peristiwa ini sejauh mungkin. Gagasan ini sangat mirip dengan kesadaran arus utama saat ini bahwa tidak ada keamanan yang tidak dapat dipecahkan, namun hambatan yang ada perlu diperbaiki dan ditingkatkan jika dipatahkan. Penelitian mutakhir tentang privasi diferensial mempertimbangkan database terdistribusi dan komputasi pada aliran data, memungkinkan skalabilitas linier dan pemrosesan real-time untuk analitik yang menjaga privasi. Oleh karena itu, teknik ini dapat membantu analitik yang

menjaga privasi pada data besar, sehingga memungkinkan data besar diterima oleh pengguna dalam hal mobilitas dan energi.

Analisis Real-Time dan Multidimensi

Analisis real-time dan multi-dimensi memungkinkan analisis streaming, energi spatiotemporal, dan data transportasi secara real-time dan multi-arah. Contoh dari sistem cyber-fisik yang dinamis dan kompleks seperti jaringan listrik menunjukkan adanya mandat bisnis yang jelas. Pengeluaran global untuk analisis data utilitas listrik diperkirakan mencapai Rp.20 miliar dalam 9 tahun ke depan, dengan pengeluaran tahunan sebesar Rp.3,8 miliar secara global pada tahun 2020 (GTM Research 2012). Namun efektivitas biaya dari teknologi yang dibutuhkan perlu dibuktikan. Pemantauan secara real-time tidak akan membenarkan biaya yang dikeluarkan jika tindakan tidak dapat dilakukan secara real-time. Teknologi pengukuran fasor, yang memungkinkan tampilan resolusi tinggi terhadap status jaringan listrik saat ini secara real-time, merupakan teknologi yang ditemukan 30 tahun yang lalu. Kemungkinan penerapannya telah diteliti selama lebih dari satu dekade. Awalnya, hal ini tidak diperlukan secara bisnis, karena sistem tenaga listrik pada masa itu dirancang dengan baik dan terstruktur dengan baik, bersifat hierarkis, statis, dan dapat diprediksi. Dengan meningkatnya dinamika melalui liberalisasi pasar dan integrasi teknologi pembangkit listrik dari sumber terbarukan seperti angin dan matahari, pandangan real-time terhadap jaringan listrik menjadi sangat diperlukan.

Kemajuan diperlukan untuk teknologi berikut:

- Komputasi aliran terdistribusi kini semakin populer. Ada dua jenis penelitian dan pengembangan komputasi aliran yang berbeda: (1) komputasi aliran seperti dalam pemrosesan peristiwa kompleks (CEP), yang memiliki fokus utama pada analisis data dengan variasi tinggi dan kecepatan tinggi, dan (2) terdistribusi komputasi aliran, dengan fokus pada pemrosesan data bervolume tinggi dan berkecepatan tinggi. Melengkapi dimensi ketiga, volume dan variasi yang hilang pada kedua strain adalah arah penelitian saat ini. Ada pendapat bahwa komputasi aliran terdistribusi, yang telah memiliki skalabilitas linier dan kemampuan pemrosesan waktu nyata, akan mengatasi tantangan data dengan variasi tinggi dengan teknik semantik (Hasan dan Curry 2014) dan data Tertaut. Pertanyaan terbuka selanjutnya adalah bagaimana memudahkan pengembangan dan penerapan algoritma yang menggunakan komputasi aliran terdistribusi serta solusi komputasi dan penyimpanan lainnya, seperti gudang data lama dan RDBMS. Karena efektivitas biaya adalah faktor utama yang mendukung nilai big data, elastisitas tingkat lanjut dengan komputasi dan penyimpanan sesuai permintaan sesuai kebutuhan algoritma juga harus diatasi.
- Pembelajaran mesin adalah kemampuan mendasar yang diperlukan ketika berhadapan dengan data besar dan sistem dinamis, ketika manusia tidak mungkin meninjau semua data, atau ketika manusia tidak memiliki pengalaman atau kemampuan untuk mendefinisikan pola. Sistem menjadi semakin dinamis dengan efek jaringan yang kompleks. Dalam sistem ini, manusia tidak mampu mengekstraksi

petunjuk yang dapat diandalkan secara real-time—namun hanya mampu melihat ke belakang selama analisis data post-mortem (yang dapat memakan waktu lama jika dilakukan oleh ilmuwan data manusia). Pembelajaran mendalam, bidang penelitian yang mendapatkan momentum, berkonsentrasi pada model data non-linier yang lebih kompleks dan berbagai transformasi data. Beberapa representasi data lebih baik untuk menjawab pertanyaan spesifik dibandingkan yang lain, artinya beberapa representasi data yang sama dalam dimensi berbeda mungkin diperlukan untuk memenuhi keseluruhan aplikasi. Pertanyaan terbukanya adalah: bagaimana merepresentasikan data energi dan mobilitas tertentu, mungkin dalam berbagai dimensi— dan bagaimana merancang algoritma yang mempelajari jawaban atas pertanyaan spesifik di bidang energi dan mobilitas dengan lebih baik daripada yang bisa dilakukan oleh operator manusia—dan melakukannya dengan cara yang dapat diverifikasi. . Pertanyaan utama dalam pembelajaran mesin adalah penyimpanan dan komputasi yang hemat biaya untuk sejumlah besar data dengan sampel tinggi, desain struktur data baru yang efisien, dan algoritme seperti pemodelan tensor dan jaringan saraf konvolusional.

Analisis Preskriptif

Analisis preskriptif memungkinkan otomatisasi pengambilan keputusan secara real-time dalam sistem energi dan mobilitas. Semakin kompleks dan dinamis suatu sistem, semakin cepat pula masukan dari data yang diperlukan untuk meningkatkan pengambilan keputusan. Dengan meningkatnya penggunaan ICT pada infrastruktur cerdas energi dan transportasi, otomatisasi pengambilan keputusan menjadi mungkin dilakukan. Namun, dengan meningkatnya digitalisasi, kondisi pengoperasian normal, ketika semua perangkat lapangan yang didigitalkan memberikan informasi yang dapat ditindaklanjuti tentang cara beroperasi dengan lebih efisien, akan membebani operator manusia. Satu-satunya kesimpulan logis adalah memiliki algoritma keputusan otomatis yang dapat diandalkan, atau mengabaikan wawasan per detik yang tidak dapat ditangani oleh operator manusia secara wajar sehingga mengakibatkan berkurangnya efisiensi operasional.

Kemajuan diperlukan untuk teknologi berikut:

- Analisis preskriptif: Teknologi yang memungkinkan analisis real-time merupakan dasar analisis preskriptif dalam sistem cyber-fisik dengan infrastruktur yang berpusat pada sumber daya seperti energi dan transportasi. Dengan analisis preskriptif, model prediktif sederhana ditingkatkan dengan tindakan yang mungkin dilakukan dan hasilnya, serta evaluasi terhadap hasil tersebut. Dengan cara ini, analisis preskriptif tidak hanya menjelaskan apa yang mungkin terjadi, namun juga menyarankan serangkaian tindakan optimal. Simulasi dan pengoptimalan adalah alat analisis yang mendukung analisis preskriptif.
- Model sistem dan rekayasa yang dapat dibaca mesin: Saat ini banyak model sistem yang tidak dapat dibaca mesin. Model rekayasa di sisi lain bersifat semi-terstruktur karena alat digital semakin banyak digunakan untuk merekayasa suatu sistem. Penelitian dan inovasi dalam bidang pekerjaan ini akan memastikan bahwa algoritma pembelajaran mesin dapat memanfaatkan pengetahuan sistem yang saat ini hanya

terbatas pada manusia. Data yang terhubung akan memfasilitasi penggandengan semantik pengetahuan pada waktu desain dan implementasi, dengan pengetahuan yang ditemukan dari data pada waktu operasi, sehingga menghasilkan model data dan algoritma pembelajaran mesin yang lebih baik (Curry dkk. 2013).

- Komputasi tepi (edge computing): Infrastruktur cerdas di sektor energi dan mobilitas memiliki kemampuan TIK yang tertanam di dalamnya, yang berarti terdapat penyimpanan dan daya komputasi di seluruh infrastruktur siber-fisik sistem kelistrikan dan transportasi, tidak hanya di ruang kontrol dan pusat data di perusahaan -tingkat. Analisis tertanam, dan analisis data terdistribusi, yang memfasilitasi analisis dalam jaringan dan di lapangan (terkadang disebut sebagai edge-computing) bersama dengan analisis yang dilakukan di tingkat perusahaan, akan menjadi pemicu inovasi di bidang energi dan transportasi.

Abstraksi

Abstraksi dari teknologi big data yang mendasarinya diperlukan untuk memungkinkan kemudahan penggunaan bagi ilmuwan data, dan pengguna bisnis. Banyak teknik yang diperlukan untuk analisis preskriptif real-time, seperti pemodelan prediktif, pengoptimalan, dan simulasi, memerlukan data dan komputasi yang intensif. Dikombinasikan dengan data besar, hal ini memerlukan penyimpanan terdistribusi dan komputasi paralel atau terdistribusi. Pada saat yang sama, banyak algoritma pembelajaran mesin dan penambangan data tidak mudah untuk diparalelkan. Sebuah survei baru-baru ini menemukan bahwa "walaupun 49% responden data scientist tidak dapat lagi memasukkan data mereka ke dalam database relasional, hanya 48% yang telah menggunakan Hadoop atau Spark dan 76% dari mereka mengatakan bahwa mereka tidak dapat bekerja secara efektif karena untuk masalah platform".

Ini merupakan indikator bahwa komputasi data besar terlalu rumit untuk digunakan tanpa pengetahuan ilmu komputer yang canggih. Salah satu arah kemajuan adalah mengembangkan abstraksi dan prosedur tingkat tinggi yang menyembunyikan kompleksitas komputasi terdistribusi dan pembelajaran mesin dari ilmuwan data. Arah lainnya tentu saja adalah ilmuwan data yang lebih terampil, yang melek komputasi terdistribusi, atau pakar komputasi terdistribusi yang menjadi lebih melek ilmu data dan statistik. Kemajuan diperlukan untuk teknologi berikut:

- Abstraksi adalah alat umum dalam ilmu komputer. Setiap teknologi pada awalnya rumit. Abstraksi mengelola kompleksitas sehingga pengguna (misalnya, pemrogram, ilmuwan data, atau pengguna bisnis) dapat bekerja mendekati tingkat pemecahan masalah manusia, tanpa mengabaikan detail realisasi praktis. Dalam evolusi teknologi data besar, beberapa abstraksi telah menyederhanakan penggunaan sistem file terdistribusi dengan mengekstraksi bahasa kueri mirip SQL untuk menjadikannya serupa dengan basis data, atau dengan mengadaptasi gaya pemrosesan ke kerangka pemrosesan analitis online yang sudah dikenal.
- Data tertaut adalah salah satu pendukung canggih untuk mewujudkan tingkat abstraksi pada sumber data berskala besar. Keterkaitan semantik data tanpa

pengetahuan sebelumnya dan terus menghubungkan dengan pengetahuan yang ditemukan akan memungkinkan pemodelan dan pengambilan pengetahuan yang terukur dalam pengaturan data besar. Pertanyaan terbuka selanjutnya adalah bagaimana mengelola berbagai sumber data dengan cara yang terukur. Penelitian di masa depan harus membangun pemahaman menyeluruh tentang arsitektur agnostik tipe data.

Kesimpulan dan Rekomendasi untuk Sektor Energi dan Transportasi

Sektor energi dan transportasi, dari perspektif infrastruktur serta dari sudut pandang efisiensi sumber daya, daya saing global, dan kualitas hidup, sangat penting bagi Eropa. Analisis terhadap sumber data yang tersedia di bidang energi serta kasus penggunaannya dalam berbagai kategori nilai big data, efisiensi operasional, pengalaman pelanggan, dan model bisnis baru membantu dalam mengidentifikasi kebutuhan dan persyaratan industri terhadap teknologi big data. Dalam penyelidikan persyaratan ini, menjadi jelas bahwa pemanfaatan teknologi big data yang ada seperti yang digunakan oleh bisnis data online saja tidak akan cukup. Adaptasi spesifik domain dan perangkat untuk digunakan dalam energi cyber-fisik dan sistem transportasi diperlukan. Inovasi mengenai privasi dan kerahasiaan yang menjaga pengelolaan dan analisis data merupakan perhatian utama para pemangku kepentingan di sektor energi dan transportasi. Tanpa memenuhi kebutuhan akan privasi dan kerahasiaan, akan selalu ada ketidakpastian peraturan, dan ketidakpastian mengenai penerimaan pengguna terhadap penawaran baru berbasis data. Di antara para pemangku kepentingan di sektor energi dan transportasi, ada perasaan bahwa “data besar” tidak akan cukup. Meningkatnya kecerdasan yang tertanam dalam infrastruktur akan mampu menganalisis data untuk menghadirkan “data cerdas”. Hal ini nampaknya perlu, karena analisis yang terlibat akan memerlukan algoritma yang jauh lebih rumit dibandingkan sektor lain. Selain itu, pertaruhan dalam skenario big data energi dan transportasi sangat tinggi, karena peluang optimalisasi akan berdampak pada infrastruktur penting. Ada beberapa contoh di sektor energi dan transportasi, dimana teknologi untuk akuisisi data, yaitu perangkat pintar, telah ada selama bertahun-tahun, atau para pemangku kepentingan telah mengukur dan menangkap sejumlah besar data. Namun kebutuhan dunia usaha tidak jelas, sehingga sulit untuk membenarkan investasi. Dengan kemajuan terkini, data dapat dikomunikasikan, disimpan, dan diproses dengan biaya yang efektif. Oleh karena itu, beberapa pemangku kepentingan menghadapi risiko jika tidak mengakui adanya dorongan teknologi. Di sisi lain, peraturan yang tidak jelas tentang penggunaan data apa yang diperbolehkan menghalangi mereka untuk bereksperimen.

Banyak teknologi big data tercanggih yang menunggu adaptasi dan penggunaan di sektor tradisional ini. Peta jalan teknologi mengidentifikasi dan menguraikan persyaratan dan teknologi berprioritas tinggi yang akan membawa sektor energi dan transportasi melampaui teknologi terkini, sehingga mereka dapat berkonsentrasi dalam menghasilkan nilai dengan mengadaptasi dan menerapkan teknologi tersebut dalam domain dan nilai penerapan spesifiknya. -menambahkan kasus penggunaan.

BAB 14
BIG DATA DI SEKTOR MEDIA DAN HIBURAN

14.1 PENDAHULUAN

Industri media dan hiburan sering kali menjadi yang terdepan dalam mengadopsi teknologi baru. Permasalahan bisnis utama yang mendorong perusahaan media untuk mempertimbangkan kemampuan big data adalah kebutuhan untuk mengurangi biaya operasional dalam lanskap yang semakin kompetitif dan, pada saat yang sama, kebutuhan untuk menghasilkan pendapatan dari penyampaian konten dan data melalui beragam platform dan produk.

Tidak cukup lagi hanya menerbitkan surat kabar harian atau menyiarkan program televisi. Operator kontemporer harus mendorong nilai dari aset mereka di setiap tahap siklus hidup data. Operator media paling gesit saat ini bahkan mungkin tidak membuat konten asli sendiri. Dua layanan streaming video internasional terbesar, Netflix dan Amazon, sebagian besar merupakan agregator konten lain, meskipun mereka juga menawarkan konten yang awalnya dipesan untuk menarik pelanggan baru dan lama.

Para pelaku industri media kini semakin terhubung dengan pelanggan dan pesaing mereka dibandingkan sebelumnya. Berkat dampak disintermediasi, konten dapat dibuat, dibagikan, dikurasi, dan diterbitkan ulang oleh siapa saja yang memiliki perangkat yang mendukung Internet. Pendapatan global dari perangkat tersebut, termasuk ponsel pintar, tablet, PC desktop, TV, konsol game, e-reader, gadget yang dapat dipakai, dan bahkan drone diperkirakan mencapai sekitar Rp.750 miliar pada tahun 2014 (Deloitte 2014). Artinya, kemampuan teknologi big data untuk menyerap, menyimpan, dan memproses berbagai sumber data yang berbeda, dan secara real-time, merupakan aset berharga bagi perusahaan yang siap berinvestasi di dalamnya.

Sektor Media dalam banyak hal merupakan pengguna awal teknologi big data, namun masih banyak evolusi yang harus dilakukan agar potensi penuhnya dapat terwujud. Integrasi yang lebih baik antara solusi-solusi di sepanjang Jaringan nilai data akan sangat penting untuk meyakinkan para pengambil keputusan agar berinvestasi dalam inovasi, terutama pada saat perekonomian sedang tidak menentu. Selain itu, pasar solusi didominasi oleh perusahaan-perusahaan AS, dan semakin banyak perusahaan-perusahaan Asia. Oleh karena itu, terdapat keharusan ekonomi bagi Eropa untuk mengembangkan dan menggunakan teknologi big data secara lebih luas. Konten dan platform media dan hiburan memiliki jangkauan global yang membuat iri banyak perusahaan di sektor lain, bahkan ritel dan manufaktur.

Studi kasus mengenai kesuksesan proyek big data di media cenderung berasal dari sisi kiri Jaringan nilai data (yaitu akuisisi dan analisis data). Namun, terdapat kebutuhan untuk mengidentifikasi contoh dan kesenjangan dalam kurasi dan penggunaan big data, karena hal ini merupakan bidang keunggulan kompetitif yang signifikan bagi organisasi media. Big data berkontribusi terhadap keuntungan dengan memungkinkan organisasi melakukan transformasi digital. Menurut PWC (2014), hal ini membentuk kepercayaan konsumen,

menciptakan kepercayaan diri untuk berinovasi dengan cepat dan gesit, serta memberdayakan inovasi.

Berbeda dengan beberapa sektor lainnya, sebagian besar data yang dapat ditindaklanjuti di sektor media sudah dalam bentuk digital (dan produk analog seperti surat kabar telah diciptakan melalui teknologi digital selama beberapa tahun terakhir). Namun, hal ini tidak berarti bahwa organisasi memperoleh manfaat finansial atau efisiensi biaya semaksimal mungkin baik dari data yang ada maupun sumber data baru. Ada semakin banyak bukti bahwa masih banyak pekerjaan yang harus dilakukan di tingkat penelitian dan kebijakan untuk mendukung ekosistem beragam bisnis yang sedang berkembang yang terlibat dalam analisis, peningkatan, dan penyampaian konten dan data.

14.2 ANALISIS KEBUTUHAN INDUSTRI DI SEKTOR MEDIA DAN HIBURAN

Sektor media selalu menghasilkan data, baik dari penelitian, penjualan, database pelanggan, file log, dan sebagainya. Demikian pula, sebagian besar penerbit dan lembaga penyiaran selalu menghadapi kebutuhan untuk bersaing sejak awal kemunculan surat kabar di abad kedelapan belas. Bahkan pemerintah atau badan media yang didanai publik harus terus membuktikan relevansinya dengan khalayak, agar tetap relevan di dunia yang penuh pilihan dan untuk mendapatkan pendanaan di masa depan. Namun pola pikir, solusi teknis, dan strategi big data menawarkan kemampuan untuk mengelola dan menyebarkan data dengan kecepatan dan skala yang belum pernah ada sebelumnya.

Ada tiga bidang utama dimana big data berpotensi mengganggu status quo dan merangsang pertumbuhan ekonomi di sektor media dan hiburan:

1. **Produk dan Layanan**: Bisnis media berbasis data besar memiliki kemampuan untuk mempublikasikan konten dengan cara yang lebih canggih. Keahlian manusia, misalnya, kurasi, editorial, dan psikologi dapat dilengkapi dengan wawasan kuantitatif yang diperoleh dari analisis kumpulan data yang besar dan heterogen. Namun hal ini didasarkan pada kemudahan penggunaan alat analisis data besar bagi ilmuwan data dan pengguna bisnis.
2. **Pelanggan dan Pemasok**: Perusahaan media yang ambisius akan menggunakan data besar untuk mengetahui lebih banyak tentang pelanggan mereka—preferensi, profil, sikap mereka—dan mereka akan menggunakan informasi tersebut untuk membangun hubungan yang lebih terlibat. Dengan alat media sosial dan pengambilan data yang kini tersedia secara luas bagi siapa saja, individu juga menjadi pemasok konten kembali ke perusahaan media. Banyak organisasi kini memasukkan analisis media sosial ke dalam proses jurnalisme ortodoks mereka, sehingga konsumen memiliki hubungan yang lebih kaya dan interaktif dengan berita. Tanpa aplikasi big data, akan ada pendekatan yang sia-sia dan acak dalam menemukan konten yang paling menarik.
3. **Infrastruktur dan Proses**: Meskipun perusahaan rintisan dan UKM dapat beroperasi secara efisien dengan infrastruktur open source dan cloud, bagi pemain yang lebih besar dan lebih tua, memperbarui infrastruktur TI lama merupakan sebuah tantangan. Produk dan standar lama masih perlu didukung dalam transisi ke cara berpikir dan

bekerja big data. Proses dan budaya organisasi mungkin juga perlu mengimbangi ekspektasi dari apa yang ditawarkan oleh big data. Kegagalan dalam mentransformasikan budaya dan keterampilan staf dapat berdampak pada perusahaan yang saat ini meraih keuntungan namun tidak dapat beradaptasi dengan model bisnis berbasis data.

14.3 POTENSI PENERAPAN BIG DATA UNTUK SEKTOR MEDIA DAN HIBURAN

Enam skenario penerapan untuk sektor media dijelaskan dan dikembangkan lebih lanjut dalam Zillner dkk. (2013, 2014a). Semua skenario ini mewakili model bisnis nyata bagi organisasi; namun, tanpa dukungan teknologi big data, perusahaan tidak akan mampu mengembangkan proyek percontohan atau proyek berskala kecil yang ada menjadi peluang pendapatan di masa depan (Tabel 14.1).

Tabel 14.1 Ringkasan enam skenario penerapan big data untuk sektor media

Nama	Jurnalisme data
Latar belakang	Data dalam jumlah besar tersedia untuk organisasi media.
Ringkasan	Kumpulan data tunggal atau ganda memerlukan analisis untuk memperoleh wawasan, menemukan cerita menarik, dan menghasilkan materi. Hal ini kemudian dapat ditingkatkan dan pada akhirnya dimonetisasi dengan menjual kepada pelanggan.
Tujuan bisnis	– Meningkatkan kualitas jurnalisme dan karenanya meningkatkan merek – Analisis data secara lebih menyeluruh dengan biaya lebih sedikit – Memungkinkan analisis data dilakukan oleh lebih banyak pengguna
Nama	Penerbitan semantik dinamis
Latar belakang	Pemrosesan konten yang skalabel untuk penargetan yang efisien
Ringkasan	Menggunakan teknologi semantik untuk memproduksi dan menargetkan konten dengan lebih efisien
Tujuan bisnis	– Kelola konten dan sumber daya staf yang langka dengan lebih efisien – Menambah nilai pada data untuk membedakan layanan dari pesaing
Nama	Analisis media sosial
Latar belakang	Memproses kumpulan data konten buatan pengguna dalam jumlah besar.
Ringkasan	Analisis batch dan real-time terhadap jutaan tweet, gambar, pembaruan status untuk mengidentifikasi tren dan konten yang dapat dikemas dalam layanan bernilai tambah.
Tujuan bisnis	– Menciptakan layanan bernilai tambah untuk klien – Melakukan pemrosesan data skala besar dengan cara yang hemat biaya
Nama	Penjualan silang produk terkait
Latar belakang	Mengembangkan mesin rekomendasi menggunakan berbagai sumber data.
Ringkasan	Aplikasi yang mengeksploitasi pemfilteran kolaboratif, pemfilteran berbasis konten, dan gabungan kedua pendekatan.
Tujuan bisnis	– Menghasilkan lebih banyak pendapatan dari pelanggan
Nama	Pengembangan produk
Latar belakang	Menggunakan analisis prediktif untuk menugaskan layanan baru

Ringkasan	Penambangan data untuk mendukung pengembangan produk baru dan lebih baik untuk pasar
Tujuan bisnis	– Menawarkan produk dan layanan baru yang inovatif – Memungkinkan pembangunan dengan cara yang lebih kuantitatif dibandingkan yang mungkin dilakukan saat ini
Nama	Wawasan audiens
Latar belakang	Menggunakan data dari berbagai sumber untuk membangun pandangan 360◦ yang komprehensif tentang pelanggan
Ringkasan	Perpanjangan skenario "Pengembangan Produk"—penambangan data eksternal organisasi untuk informasi tentang kebiasaan dan preferensi pelanggan
Tujuan bisnis	– Mengurangi biaya retensi dan akuisisi pelanggan – Gunakan wawasan untuk membantu penerapan produk dan layanan baru – Memaksimalkan pendapatan dari pelanggan

14.4 PENDORONG DAN KENDALA BIG DATA DI SEKTOR MEDIA DAN HIBURAN

Seperti semua bisnis, perusahaan media bertujuan untuk memaksimalkan pendapatan, meminimalkan biaya, dan meningkatkan pengambilan keputusan dan proses bisnis.

Pengemudi

Khusus untuk sektor media dan hiburan, terdapat faktor-faktor pendorong berikut ini: Bertujuan untuk memahami pelanggan pada tingkat yang sangat mendetail, sering kali dengan menganalisis berbagai jenis interaksi (misalnya penggunaan produk, interaksi layanan pelanggan, media sosial, dll.).

Beroperasi di sub-sektor yang ramai seperti pemasaran digital atau penerbitan buku, di mana hanya sedikit pemain yang mendominasi, dan preferensi konsumen serta mode dapat berubah dengan sangat cepat.

Diversifikasi penawaran layanan sedapat mungkin. Sebagian besar perusahaan media besar di Eropa beroperasi di banyak bidang, misalnya penerbit surat kabar, situs web, dan aplikasi komersial; atau lembaga penyiaran juga dapat menjual akses broadband.

Berkomunikasi untuk membangun pengaruh dalam masyarakat, misalnya. secara politis. Hal ini kurang nyata dibandingkan sekadar menjual produk, namun dipandang sama pentingnya oleh pemilik media atau pemerintah.

Kendala

Kendala big data di sektor media dan hiburan dapat diringkas sebagai berikut:

- Meningkatnya kesadaran dan kekhawatiran konsumen tentang bagaimana data pribadi digunakan. Ada ketidakpastian peraturan bagi perusahaan-perusahaan di Eropa yang menangani data pribadi, yang berpotensi menempatkan mereka pada posisi yang dirugikan dibandingkan dengan, misalnya, perusahaan-perusahaan Amerika yang beroperasi dalam lanskap hukum yang jauh lebih longgar.
- Kurangnya akses terhadap pendanaan bagi perusahaan rintisan (start-up) media dan UKM. Meskipun relatif mudah untuk memulai perusahaan baru yang memproduksi

aplikasi, game, atau jejaring sosial, akan lebih sulit untuk mengembangkannya tanpa investor yang berkomitmen.

- Pasar tenaga kerja di seluruh Eropa tidak menyediakan cukup tenaga profesional data yang mampu memanipulasi aplikasi big data, misalnya. untuk jurnalisme data dan manajemen produk.
- Ketakutan akan pembajakan dan pengabaian konsumen terhadap hak cipta dapat menghambat orang-orang dan perusahaan kreatif untuk mengambil risiko dalam meluncurkan produk dan layanan media dan budaya baru.
- Pemain besar AS mendominasi industri konten dan data. Perusahaan seperti Apple, Amazon, dan Google memiliki dominasi besar di banyak sub-sektor termasuk musik, periklanan, penerbitan, dan media konsumen elektronik.
- Perbedaan penetrasi penyediaan broadband berkecepatan tinggi antar negara anggota, di perkotaan, dan di pedesaan. Hal ini merupakan disinsentif bagi perusahaan yang ingin mengirimkan konten yang memerlukan bandwidth tinggi, misalnya. streaming film, karena mengurangi basis pelanggan potensial.

14.5 SUMBER DATA MEDIA DAN HIBURAN YANG TERSEDIA

Tabel 14.2 dimaksudkan untuk memberikan gambaran tentang sumber data yang rutin ditangani oleh sebagian besar perusahaan media. Tabel pertama mencantumkan beberapa kategori data yang dihasilkan oleh perusahaan itu sendiri, sedangkan tabel kedua menunjukkan sumber pihak ketiga yang sedang atau dapat diproses oleh pihak-pihak yang berkecimpung di sektor media, bergantung pada lini bisnis mereka.

Setiap jenis sumber data disesuaikan dengan karakteristik utama big data. Biasanya, industri teknologi membicarakan "tiga V data besar", yaitu volume, variasi, dan kecepatan. Kobielus (2013) juga membahas karakteristik keempat—kejujuran. Hal ini penting bagi sektor media karena produk dan layanan konsumen dapat dengan cepat gagal jika kontennya tidak memiliki kredibilitas, atau kualitasnya buruk, atau sumbernya tidak pasti. Menurut IBM (2014), 27% responden survei di AS tidak yakin seberapa besar data mereka tidak akurat—menunjukkan bahwa skala permasalahannya terlalu diremehkan.

Tabel 14.2 Sumber daya data media dipetakan ke karakteristik "V" dari big data

Data yang dihasilkan secara internal	Karakteristik kunci "V".
Detail profil konsumen termasuk interaksi layanan pelanggan.	Volume—Data dalam jumlah besar untuk disimpan dan berpotensi ditambang. Variasi berlaku ketika mempertimbangkan berbagai cara pelanggan berinteraksi dengan penyedia layanan media—dan karenanya peluang bagi bisnis untuk "bergabung
Pencatatan jaringan (misalnya untuk perusahaan web atau hiburan yang mengoperasikan jaringannya sendiri).	Kecepatan—Masalah jaringan harus diidentifikasi secara real-time untuk menyelesaikan masalah dan mempertahankan kepercayaan konsumen.
Organisasi memiliki layanan data kepada pengguna akhir.	Karakteristiknya akan bergantung pada tujuan bisnis data tersebut, misalnya, kantor berita akan

	memprioritaskan kecepatan pengiriman ke pelanggan, lembaga penyiaran akan fokus pada streaming konten dalam berbagai format ke berbagai jenis perangkat.
Preferensi konsumen disimpulkan dari sumber termasuk data aliran klik, perilaku penggunaan produk, riwayat pembelian, dll.	Volume—Data dalam jumlah besar dapat dikumpulkan. Kecepatan akan menjadi relevan jika layanan harus responsif terhadap tindakan pengguna, misalnya jaringan game online yang menjual fitur tambahan kepada pemain.
Data pihak ketiga	Karakteristik kunci "V".
Umpan data komersial, misalnya data olahraga, kantor berita kantor berita.	Kecepatan—Menjadi yang pertama menggunakan data seperti olahraga atau acara berita akan membangun keunggulan kompetitif.
Informasi jaringan (tempat jaringan eksternal digunakan, misalnya aplikasi perpesanan yang mendukung jaringan seluler).	Kecepatan—Masalah jaringan harus diidentifikasi secara real-time untuk memastikan kelangsungan layanan.
Kumpulan data terbuka sektor publik.	Kebenaran—Data terbuka mungkin mempunyai masalah kualitas, asal usul, dan kelengkapan.
Data terstruktur dan/atau tertaut gratis, misalnya Wikidata/DBpedia	Kebenaran—data yang dikumpulkan dari sumber daya manusia (crowdsourced) mungkin mempunyai masalah kualitas, sumber, dan kelengkapan.
Data media sosial, misalnya pembaruan, video, gambar, tautan, dan sinyal seperti "suka".	Volume, variasi, kecepatan, dan kebenaran—Perusahaan media harus memprioritaskan pemrosesan berdasarkan kasus penggunaan yang diharapkan. Sebagai salah satu contoh, jurnalisme data memerlukan sejumlah besar data yang harus dipersiapkan untuk dianalisis dan diinterpretasikan. Di sisi lain, bisnis pemasaran media mungkin lebih mementingkan keragaman data sosial di banyak saluran.

14.6 PERSYARATAN SEKTOR MEDIA DAN HIBURAN

Forum Sektoral Media dan Hiburan mampu mengidentifikasi dan menyebutkan beberapa persyaratan yang perlu dipenuhi oleh penerapan big data di domain tersebut. Persyaratannya dibedakan antara persyaratan non teknis dan teknis.

Persyaratan Non-teknis

Penting untuk dicatat bahwa meluasnya penggunaan big data dalam industri media tidak hanya bergantung pada keberhasilan penerapan teknologi dan solusi tertentu. Dalam Zillner dkk. (2014b), sebuah survei dilakukan terhadap para manajer menengah dan senior di Eropa dari sektor media (dan juga sektor telekomunikasi, di mana para pemain besar semakin banyak yang pindah ke bidang-bidang yang dulunya dianggap murni bidang penyiaran, penerbit, dll.). Responden diminta untuk mengurutkan beberapa prioritas big data berdasarkan seberapa penting hal tersebut bagi organisasi mereka.

Sangat mengejutkan bahwa semua peserta survei mengidentifikasi perlunya kerangka kerja Eropa untuk standar bersama, lanskap peraturan yang jelas, dan ekosistem kolaboratif — yang menyiratkan bahwa dunia usaha kurang percaya diri terhadap kemampuan mereka untuk mengatasi hype dan benar-benar mencapai tujuan. mengatasi big data di perusahaan mereka. Bidang lain yang dinilai sangat penting oleh sebagian besar responden adalah membuat solusi bermanfaat dan menarik bagi pengguna bisnis (bukan hanya data scientist).

Persyaratan Teknis

Tabel 14.3 mencantumkan 37 persyaratan yang disaring dari kerja Forum Sektor Media. Setiap persyaratan disesuaikan dengan tujuan bisnis (meskipun tentu saja dalam praktiknya beberapa persyaratan dapat memenuhi lebih dari satu tujuan). Lima kolom di sisi kanan tabel menempatkan setiap persyaratan pada tempat yang sesuai di sepanjang Jaringan nilai big data. Media, sebagai sektor ekonomi yang paling banyak berhubungan dengan pelanggan dan menghasilkan pendapatan, memiliki banyak kebutuhan penting dalam kurasi dan penggunaan data.

Tabel 14.3 Persyaratan teknis big data pada sektor media

Persyaratan data besar	Tujuan bisnis	Akuisisi	Analisis	Kurasi	Penyimpanan	Penggunaan
Kurasi sumber data yang heterogen dengan cara yang tidak bergantung pada konten dan asal	Memperbaiki proses bisnis	X		X		
Secara terprogram menginterogasi data untuk mengetahui tren	Memperbaiki proses bisnis		X			
Mulai memproses tipe data baru dengan cepat saat diperlukan	Memperbaiki proses bisnis	X	X	X	X	
Analisis data tidak terstruktur berkenaan dengan sentimen, topik, dan aspek teks tak berwujud lainnya	Memperbaiki proses bisnis		X	X		
Mengubah dan menambah data terbuka dari sektor publik dalam hal format, semantik, dan kualitas	Memperbaiki proses bisnis	X	X	X	X	
Alat yang dapat diskalakan untuk aplikasi pencarian dan penemuan	Memperbaiki proses bisnis				X	X
Visualisasikan data untuk analitik dan metrik (terutama untuk pengguna teknis bisnis)	Memperbaiki proses bisnis		X			X
Secara otomatis membuat dan menerapkan metadata ke kumpulan data	Memperbaiki proses bisnis	X	X	X		
Memproses data dengan cepat dan akurat hampir secara real-time	Meningkatkan pengambilan keputusan	X	X		X	
Terapkan model dan ontologi ke data untuk mengekstrak hubungan	Meningkatkan pengambilan keputusan	X	X		X	

Ubah aliran dari sensor menjadi tampilan yang dapat ditindaklanjuti	Meningkatkan pengambilan keputusan	X			X	
Alat analitik yang memungkinkan kueri dan manipulasi canggih oleh non-programmer atau ahli statistik	Meningkatkan pengambilan keputusan		X		X	
Mesin inferensi untuk menganalisis data grafik semantik	Meningkatkan pengambilan keputusan		X	X		
Dapatkan nilai dari kumpulan data kepemilikan	Meningkatkan pendapatan	X	X	X	X	X
Dapatkan nilai dari kumpulan data terbuka publik	Meningkatkan pendapatan		X	X		X
Memberikan data dan konten yang disesuaikan kepada pelanggan	Meningkatkan pendapatan			X		X
Editorialisasi aliran data hasil kurasi yang berpusat pada manusia	Meningkatkan pendapatan		X	X		X
Algoritma untuk mengolah data untuk menghasilkan rekomendasi yang lebih menarik dibandingkan "rekomendasi yang lebih sama"	Meningkatkan pendapatan					X
Alat manajemen algoritma untuk pengguna non-teknis	Meningkatkan pendapatan			X		X
Memperkaya konten multimedia seperti gambar dan video dengan metadata semantik	Meningkatkan pendapatan		X	X		X
Padukan konten buatan pengguna dengan media yang diproduksi secara komersial untuk menciptakan produk digital baru	Meningkatkan pendapatan	X		X		X
Menghasilkan wawasan dari data untuk mengaktifkan model bisnis baru (misalnya penjualan silang berdasarkan kebiasaan menonton)	Meningkatkan pendapatan		X			X
Tingkatkan konversi dari aktivitas pemasaran offline (misalnya surat langsung) dengan menganalisis data online	Meningkatkan pendapatan		X			X
Solusi analitik prediktif yang dapat mengidentifikasi tren, segmen, dan pola tanpa harus dimodelkan secara eksplisit	Meningkatkan pendapatan		X			
Memberikan hasil pencarian yang lebih relevan kepada konsumen	Meningkatkan pendapatan					X
aplikasi menggunakan analisis semantik						
Solusi database yang dapat diatur lebih cepat dibandingkan dengan aplikasi tradisional	Mengurangi biaya				X	
Kemampuan untuk menggunakan kurasi data crowdsourcing untuk melengkapi keahlian materi pelajaran internal	Mengurangi biaya			X		
Kelola data berskala besar dalam database grafik	Mengurangi biaya				X	
Terjemahkan data tidak terstruktur (misalnya teks atau suara) ke satu atau beberapa bahasa	Mengurangi biaya	X	X		X	

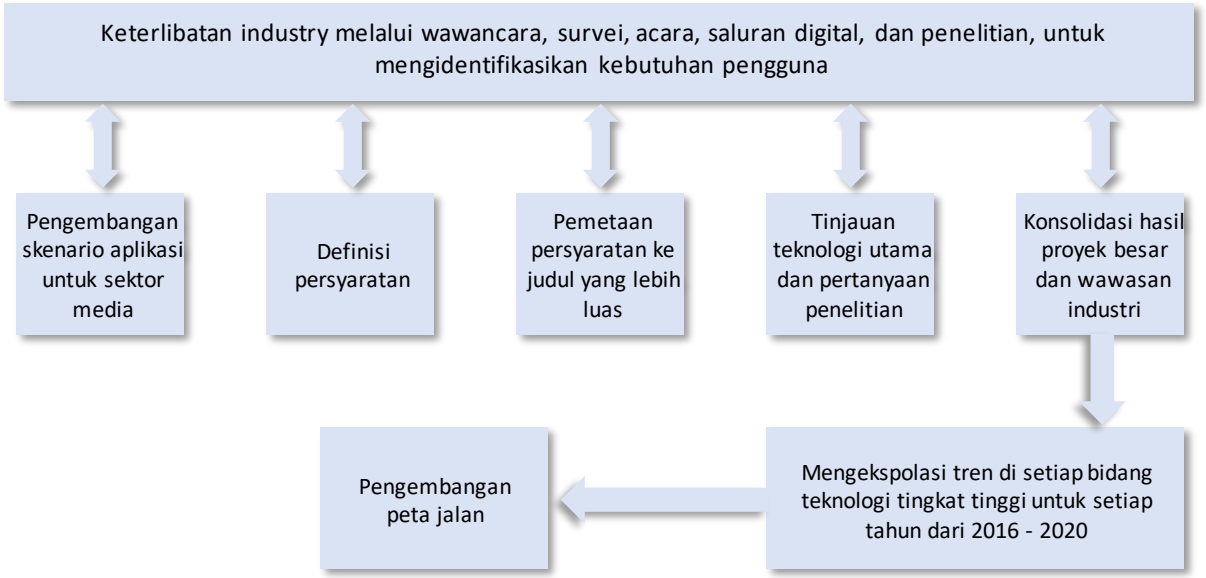
Alat pengikisan dan perayapan data bervolume tinggi	Mengurangi biaya	X			X	
Identifikasi pola dalam data untuk mendorong wawasan tentang perilaku konsumen	Pahami pelanggan		X			
Pertimbangkan banyak faktor (misalnya lokasi, perangkat, profil pengguna, konteks penggunaan) untuk menargetkan pengiriman konten dengan lebih baik	Pahami pelanggan	X				
Hubungkan data dari semua interaksi pelanggan untuk membentuk tampilan 360◦	Pahami pelanggan	X	X		X	
Menyerap data dari kelas perangkat baru (misalnya perangkat yang dapat dikenakan)	Pahami pelanggan	X				
Telusuri perilaku konsumen secara lebih rinci	Pahami pelanggan		X			X
Bina hubungan yang lebih terlibat dengan audiens dan pelanggan melalui analisis data sosial yang tidak terstruktur	Pahami pelanggan					X
Arah kebijakan yang jelas mengenai penggunaan data pribadi di UE	Pahami pelanggan				X	X

14.7 PETA JALAN TEKNOLOGI BIG DATA DI SEKTOR MEDIA DAN HIBURAN

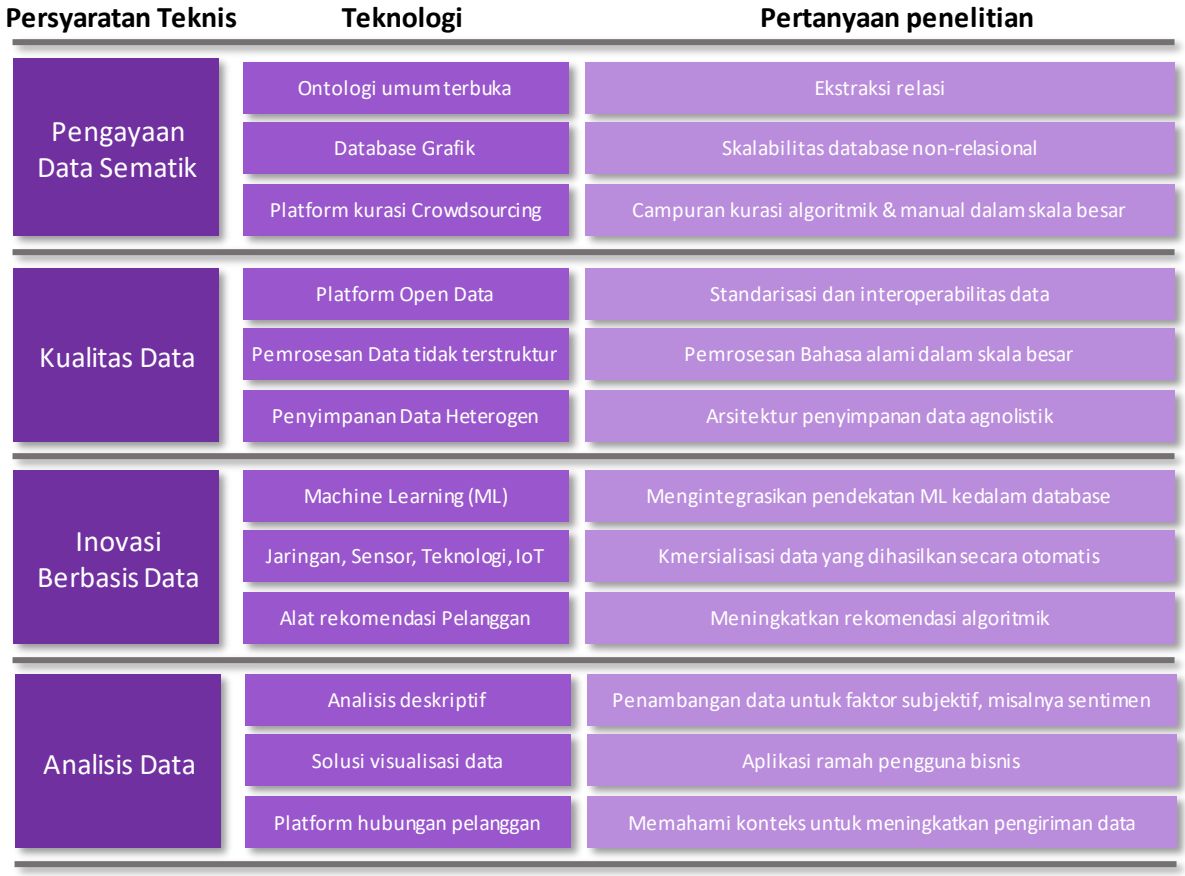
Dari semua sektor yang dibahas dalam buku ini, media bisa dibilang merupakan sektor yang paling sering dan tiba-tiba berubah. Paradigma baru dapat muncul dengan sangat cepat dan menjadi penting secara komersial dalam waktu singkat (misalnya Twitter baru didirikan pada tahun 2006 dan kini memiliki kapitalisasi pasar sebesar miliaran dolar). Tahun 2015 dan seterusnya akan menyaksikan banyak pemain media dan konsumen bereksperimen dengan drone (lebih tepatnya, "kendaraan udara tak berawak", atau UAV) untuk melihat apakah rekaman yang diambil dapat dimonetisasi baik secara langsung sebagai konten atau secara tidak langsung untuk menarik iklan.

Gambar 14.2 dan Tabel 14.4 mengkonsolidasikan hasil penelitian yang diselesaikan oleh Zillner dkk. (2013, 2014a), beserta latar belakang penelitian tambahan. Gambar 14.1 memetakan metodologi yang digunakan untuk menyusun peta jalan sektor, yang menunjukkan bagaimana keterlibatan berulang dengan industri mendukung definisi kebutuhan dan teknologi seputar big data untuk sektor media di setiap tahap.

Setiap peta jalan harus menyadari risiko bahwa peta jalan tersebut akan ketinggalan zaman bahkan sebelum dipublikasikan. Meskipun demikian, judul-judul utama yang ditampilkan dalam angka-angka di bagian ini diperkirakan tetap relevan dengan sektor ini karena alasan-alasan berikut.



Gambar 14.1 Metodologi untuk menyusun peta jalan sektor media



Gambar 14.2 Persyaratan pemetaan untuk pertanyaan penelitian di sektor media

Pengayaan Data Semantik

Semantik adalah bidang yang telah lama berdiri dan kini berkembang pesat yang akhirnya memenuhi janji akademisnya. Aplikasi media utama seperti “asisten pribadi cerdas”,

misalnya. Siri dan Cortana, didukung oleh "kecerdasan buatan" dan teknologi analisis semantik. Pengembangan lebih lanjut diperlukan untuk membantu organisasi komersial di Eropa memanfaatkan potensi ontologi, database grafik, dan platform kurasi.

Kualitas Data

Perkembangan teknologi utama di bidang ini mencakup data terbuka dan standar data secara umum untuk membantu interoperabilitas. Yang juga penting adalah kemampuan untuk memproses aliran data yang tidak terstruktur (terutama bahasa alami). Terakhir, terdapat kebutuhan akan sistem back-end yang dapat menyerap berbagai jenis data dengan sesedikit mungkin hambatan, dengan meminimalkan kebutuhan untuk menentukan skema data terlebih dahulu.

Inovasi Berbasis Data

Tiga teknologi utama yang mendasari dorongan inovasi berkualitas tinggi adalah pembelajaran mesin pada skala perusahaan; Internet of Things (IoT), yang secara eksponensial akan meningkatkan volume dan keragaman aliran data yang tersedia bagi siapa pun yang terlibat dalam media atau penyampaian cerita berbasis data; dan terakhir, alat untuk menafsirkan interaksi pelanggan dengan produk dan layanan dengan lebih baik.

Analisis Data

Perusahaan media dan hiburan perlu menganalisis data tidak hanya pada tingkat pelanggan dan produk, namun juga pada tingkat jaringan dan infrastruktur (misalnya pemasok video streaming, bisnis internet, lembaga penyiaran televisi, dan sebagainya). Teknologi utama di tahun-tahun mendatang adalah analitik deskriptif, solusi manajemen hubungan pelanggan yang lebih canggih, dan yang terakhir adalah solusi visualisasi data yang dapat diakses oleh berbagai pengguna di perusahaan. Hanya dengan "memanusiakan" alat-alat inilah maka big data akan mampu memberikan manfaat yang semakin dibutuhkan oleh bisnis berbasis data (Tabel 14.4).

Tabel 14.4 Peta jalan teknologi big data untuk sektor media

Persyaratan teknis	Tahun 1	Tahun 2	Tahun 3	Tahun 4	Tahun 5
Pengayaan data semantik	Perusahaan media besar dengan sumber daya membuat dan menerbitkan ontologi terbuka	Ontologi umum untuk kasus penggunaan tertentu di industri media dan hiburan	Alat manajemen dan manipulasi ontologi tersedia untuk berbagai penggunaan komersial	Teknologi ekstraksi relasi tersedia dalam skala dan keterjangkauan	Inferensi semantik untuk mendukung analisis prediktif data, misalnya perilaku pengguna, pelacakan berita
Kualitas data	Terbatasnya data terbuka yang tersedia bagi perusahaan yang ingin	Data terbuka diterbitkan dalam format yang dapat dibaca mesin	Alat pemrosesan bahasa alami yang dapat diskalakan untuk data dalam	Standarisasi protokol akuisisi data	Arsitektur data-agnostik memungkinkan beragam aliran data dianalisis

	menghasilkan model bisnis baru	oleh lebih banyak badan sektor publik dan swasta	jumlah besar (termasuk ucapan)		secara bersamaan dan secara real-time
Inovasi berbasis data	Platform kurasi untuk meningkatkan nilai tambah produk data	Alat rekomendasi yang skalabel untuk pengguna non-teknis	Kerangka pembelajaran mesin- tertanam ke dalam alat pengambilan keputusan	Agregasi aliran waktu nyata yang dihasilkan oleh jaringan, sensor, perangkat yang dikenakan di tubuh	Platform pengembangan produk untuk iterasi cepat layanan berbasis data
Analisis data	Segmentasi pelanggan yang lebih rinci berdasarkan faktor subjektif	Alat visualisasi data intuitif untuk aplikasi interaktif	Konvergensi intelijen bisnis dan aplikasi analisis produk	Analisis prediktif peristiwa atau tren yang dapat ditindaklanjuti di seluruh aliran data yang besar dan tersebar	Pendekatan analitik yang dapat digabungkan memungkinkan wawasan mendalam tentang pola berdasarkan konteks

14.8 KESIMPULAN DAN REKOMENDASI UNTUK SEKTOR MEDIA DAN HIBURAN

Eropa memiliki banyak hal yang bisa ditawarkan dalam budaya dan konten ke pasar global. Penerbit dan perusahaan TV Eropa terkenal secara global, namun belum ada pesaing berbasis di Uni Eropa yang mampu menandingi raksasa multinasional seperti Google, Amazon, Apple, atau Facebook. Perbedaan antara perekonomian Eropa dan AS, seperti kemudahan akses terhadap modal ventura, tampaknya menghalangi hal ini terjadi. Oleh karena itu, cara terbaik bagi Eropa untuk maju adalah dengan membangun kekuatan kreativitas dan pergerakan bebas manusia dan jasa, guna menyatukan komunitas pelaku industri, peneliti, dan pemerintah untuk menangani prioritas berikut:

- Memahami aliran data, baik teks, gambar, video, sensor, dan sebagainya. Produk dan layanan yang canggih dapat dikembangkan dengan mengekstraksi nilai dari sumber yang heterogen.
- Memanfaatkan perubahan langkah big data dalam kemampuan menyerap dan memproses data mentah, sehingga meminimalkan risiko dalam menghadirkan penawaran baru berbasis data ke pasar.
- Mengkurasi informasi berkualitas dari aliran data yang sangat besar, menggunakan pendekatan yang dapat diskalakan secara algoritmik dan memadukannya dengan pengetahuan manusia melalui platform kurasi.
- Mempercepat adopsi bisnis terhadap data besar. Kesadaran konsumen semakin meningkat dan peningkatan teknis terus mengurangi biaya penyimpanan dan alat analisis. Oleh karena itu, sangatlah penting bagi dunia usaha untuk memiliki keyakinan bahwa mereka memahami apa yang mereka inginkan dari big data dan bahwa aspek non-teknis seperti sumber daya manusia dan peraturan sudah ada.

BAB 15
ANALISIS PERSYARATAN LINTAS SEKTORAL UNTUK RISET BIG DATA

15.1 PENDAHULUAN

Bab ini mengidentifikasi persyaratan lintas sektoral untuk penelitian big data yang diperlukan untuk menentukan peta jalan penelitian yang diprioritaskan berdasarkan dampak yang diharapkan. Tujuan dari peta jalan ini adalah untuk memaksimalkan dan mempertahankan dampak teknologi dan aplikasi big data di berbagai sektor industri dengan mengidentifikasi dan mendorong peluang di Eropa. Target audiens dari peta jalan ini adalah berbagai pemangku kepentingan yang terlibat dalam ekosistem big data termasuk industri pengguna aplikasi big data, penyedia teknis solusi big data, regulator, pembuat kebijakan, peneliti, dan pengguna akhir.

Langkah pertama menuju peta jalan ini adalah menetapkan daftar persyaratan dan tujuan bisnis lintas sektoral dari masing-masing sektor industri yang tercakup dalam sebagian buku ini dan dalam Zillner dkk. (2014). Hasil konsolidasi tersebut terdiri dari serangkaian persyaratan lintas sektor yang diprioritaskan yang digunakan untuk menentukan peta jalan teknologi, bisnis, kebijakan, dan masyarakat serta rekomendasi tindakan. Bab ini menyajikan versi ringkas dari persyaratan konsolidasi lintas sektoral. Bab ini membahas masing-masing persyaratan tingkat tinggi dan sub-tingkat serta tantangan terkait yang perlu diatasi. Terakhir, bab ini diakhiri dengan penentuan prioritas persyaratan lintas sektoral. Sedapat mungkin, peta jalan telah dikuantifikasi untuk memungkinkan penetapan prioritas dan rencana aksi yang beralasan (misalnya kebijakan).

15.2 PERSYARATAN KONSOLIDASI LINTAS SEKTORAL

Untuk membangun pemahaman umum tentang persyaratan serta deskripsi teknologi di seluruh domain, label persyaratan spesifik sektor diselaraskan. Setiap sektor menyediakan persyaratannya dengan kebutuhan pengguna yang terkait, dan persyaratan serupa dan terkait digabungkan, diselaraskan, atau direstrukturisasi untuk menciptakan satu set yang homogen. Meskipun sebagian besar persyaratan ada di setiap sektor, tingkat kepentingan persyaratan di setiap sektor berbeda-beda. Untuk analisis lintas sektor, setiap persyaratan yang diidentifikasi oleh setidaknya dua sektor sebagai persyaratan signifikan untuk sektor tersebut dimasukkan ke dalam definisi peta jalan lintas sektor. Dengan demikian, daftar awal dari 13 persyaratan tingkat tinggi dan 28 persyaratan sub-tingkat dikurangi menjadi 5 persyaratan tingkat tinggi dan 12 persyaratan sub-tingkat (lihat Tabel 15.1). Dalam bab ini, pembahasan mengenai setiap persyaratan lintas sektoral telah diringkas dan pembaruan kecil telah diterapkan. Rincian lengkap tersedia di Becker dkk. (2014).

Teknik Manajemen Data

Rekayasa manajemen data persyaratan tingkat tinggi bertujuan untuk mencapai strategi yang efisien untuk mengelola sumber data dan teknologi yang heterogen. Rekayasa manajemen data memiliki empat sub-persyaratan:

- Pengayaan data
- Integrasi data
- Berbagi data
- Transmisi data waktu nyata

Pengayaan Data

Pengayaan data sub-persyaratan bertujuan agar data tidak terstruktur dapat dipahami di seluruh domain, aplikasi, dan rantai nilai.

Tabel 15.1 Konsolidasi persyaratan lintas sektoral (dan sektor-sektor yang menuntut)

Persyaratan teknologi	Jumlah sektor yang menuntut	Kesehatan	Umum	Keuangan dan asuransi	Energi dan transportasi	Telekomunikasi dan media	Ritel	Manufaktur
Teknik manajemen data	3		X			X	X	
Pengayaan data	2	X				X		
Integrasi data	5	X	X	X		X	X	
Berbagi data	4	X	X	X	X			
Real-time Transmisi data	3		X				X	X
Kualitas Data	3	X		X			X	
Peningkatan Data	2						X	
Keamanan & Privasi data	7	X	X	X	X	X	X	X
Visualisasi data dan pengalaman pengguna	2						X	X
Analisis data mendalam	3		X	X			X	
Simulasi permodelan	3		X				X	X
Analisis bahasa alami			X				X	X
Penemuan Pola	3		X	X		X		
Analisis prediktif	2		X			X		
Analisis prespektif	3				X	X	X	
Wawasan waktu nyata	5		X		X	X	X	X
Analisis penggunan	2			X		X		

Di sektor kesehatan, pengayaan data mempunyai relevansi yang tinggi, karena 90% data kesehatan hanya tersedia dalam format tidak terstruktur tanpa label semantik yang memberi informasi kepada aplikasi tentang konten datanya. Secara khusus, pendekatan untuk anotasi semantik pada gambar medis dan teks medis diperlukan.

Di sektor telekomunikasi dan media, pengayaan data mencakup ontologi (misalnya eTOM SID), transformasi data, penambahan metadata, format, dll, dengan mempertimbangkan bahwa sumber datanya heterogen (termasuk informasi media sosial, audio, data pelanggan, dan data lalu lintas, misalnya). Data yang berasal dari sumber berbeda

dan format berbeda, dihasilkan oleh sistem heterogen, harus diproses bersama. Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Ekstraksi informasi dari teks
- Algoritma pemahaman gambar
- Kerangka anotasi standar

Berbagi dan Integrasi Data

Pembagian dan integrasi data sub-persyaratan bertujuan untuk membangun dasar bagi integrasi berbagai sumber data yang beragam dan lancar ke dalam platform data besar. Kurangnya skema data yang terstandarisasi, model data semantik, serta fragmentasi kepemilikan data merupakan aspek penting yang perlu diatasi. Saat ini, kurang dari 30% data kesehatan dibagikan antar penyedia layanan kesehatan (Accenture 2012). Untuk memungkinkan pertukaran data yang lancar di bidang kesehatan dan domain lainnya, diperlukan sistem pengkodean dan terminologi standar serta model data.

Di sektor telekomunikasi, data telah dikumpulkan selama bertahun-tahun dan diklasifikasikan menurut standar bisnis berdasarkan eTOM (2014), namun model referensi data belum mempertimbangkan penyertaan data media sosial. Diperlukan suatu sistem informasi terpadu yang mencakup data baik dari operator telekomunikasi maupun pelanggan. Setelah model informasi ini tersedia, model tersebut harus dimasukkan ke dalam model referensi eTOM SID dan diperhitungkan dalam solusi khusus telekomunikasi big data agar semua data (sosial dan non-sosial) dapat diintegrasikan.

Di sektor ritel, ontologi produk standar diperlukan untuk memungkinkan berbagi data antara produsen produk dan pengecer. Layanan untuk mengoptimalkan keputusan operasional di ritel hanya dapat dilakukan dengan data produk yang dianotasi secara semantik. Di sektor publik, berbagi dan integrasi data penting untuk mengatasi kurangnya standarisasi skema data dan fragmentasi kepemilikan data, untuk mencapai integrasi sumber data yang beragam dan beragam ke dalam platform data besar. Hal ini diperlukan ketika analisis data harus dilakukan dari data yang dimiliki oleh domain dan pemilik berbeda (misalnya lembaga berbeda di sektor publik) atau mengintegrasikan data eksternal yang heterogen (dari data terbuka, jaringan sosial,sensor, dll).

Di sektor keuangan, beberapa faktor telah menempatkan organisasi pada situasi di mana sejumlah besar kumpulan data tidak memiliki interkoneksi dan integrasi. Organisasi keuangan menyadari nilai potensial dari keterkaitan kumpulan data tersebut untuk mengekstrak informasi yang berguna untuk mengoptimalkan operasi, meningkatkan layanan kepada pelanggan, atau bahkan menciptakan model bisnis baru. Teknologi yang ada dapat mencakup sebagian besar kebutuhan industri jasa keuangan, namun teknologi tersebut masih belum diterapkan secara luas.

Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Data semantik dan model pengetahuan
- Informasi konteks
- Pencocokan entitas
- Penyimpanan rangkap tiga yang dapat diskalakan, penyimpanan kunci/nilai

- Memfasilitasi integrasi inti pada akuisisi data
- Praktik terbaik untuk berbagi data berkecepatan tinggi dan beragam
- Kegunaan sistem semantik
- Kerangka kerja metadata dan asal data
- Mekanisme pemetaan data/skema otomatis yang dapat diskalakan

Transmisi Data Waktu Nyata

Transmisi data real-time sub-persyaratan bertujuan untuk memperoleh informasi (sensor dan peristiwa) secara real-time. Di sektor publik, hal ini berkaitan erat dengan peningkatan kemampuan penerapan sensor dan skenario Internet of Things, seperti di bidang keselamatan publik dan kota pintar. Sensor gambar telah mengikuti Hukum Moore, yang menggandakan kepadatan megapiksel per dolar setiap 2 tahun (PWC 2014). Kemampuan pemrosesan dan pembersihan yang terdistribusi diperlukan untuk sensor gambar guna menghindari kelebihan beban pada saluran transmisi (Jobling 2013) dan menyediakan analisis real-time yang diperlukan untuk memberi masukan pada sistem kesadaran situasional bagi para pengambil keputusan.

Di sektor manufaktur, data sensor harus diperoleh dengan kecepatan sampel yang tinggi dan perlu dikirim mendekati waktu nyata agar dapat digunakan secara efektif. Keputusan dapat diambil pada titik perencanaan, komando, dan kendali pusat, atau dapat dibuat pada tingkat lokal secara terdistribusi. Transmisi data harus cukup mendekati waktu nyata, sehingga jauh lebih baik dibandingkan interval yang panjang (setiap jam atau lebih) saat pengambilan sampel data inventaris. Lingkungan kerja yang tidak bersahabat di bidang manufaktur dapat menghambat transmisi data.

Untuk sektor ritel, data dari sensor di dalam toko harus diperoleh secara real-time. Ini mencakup data visual dari kamera dan lokasi pelanggan dari sensor posisi.

Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Pemrosesan dan pembersihan data terdistribusi
- Solusi penyimpanan baca/tulis yang dioptimalkan untuk data berkecepatan tinggi
- Pemrosesan aliran data hampir secara real-time

Kualitas Data

Persyaratan tingkat tinggi, kualitas data, menggambarkan kebutuhan untuk menangkap dan menyimpan data berkualitas tinggi sehingga aplikasi analitik dapat menggunakan data tersebut sebagai masukan yang andal untuk menghasilkan wawasan yang berharga. Kualitas data memiliki satu sub-persyaratan:

- Peningkatan data

  Penerapan big data di sektor kesehatan harus memenuhi standar kualitas data yang tinggi agar dapat memperoleh wawasan yang dapat diandalkan untuk pengambilan keputusan terkait kesehatan. Misalnya, fitur dan daftar parameter yang digunakan untuk menggambarkan status kesehatan pasien perlu distandarisasi agar memungkinkan perbandingan kumpulan data pasien (populasi) yang andal.

Di sektor telekomunikasi dan media, meskipun data telah dikumpulkan selama bertahun-tahun, masih terdapat permasalahan kualitas data yang membuat informasi tidak dapat dieksploitasi tanpa pengolahan terlebih dahulu. Di sektor keuangan, kualitas data bukanlah masalah utama dalam kumpulan data yang dihasilkan secara internal, namun informasi yang dikumpulkan dari sumber eksternal mungkin tidak sepenuhnya dapat diandalkan.

Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Pengelolaan asal usulnya
- Interaksi data manusia
- Integrasi data tidak terstruktur

- **Perbaikan Data**

  Penyempurnaan data sub-persyaratan bertujuan untuk menghilangkan noise/data redundan, memeriksa keterpercayaan, dan menambahkan data yang hilang. Di sektor telekomunikasi dan media, hal ini berkaitan dengan kemampuan untuk meningkatkan penawaran komersial penyedia layanan berdasarkan informasi yang tersedia dalam sistem tradisional, serta teknik-teknik canggih seperti analisis prediktif, ucapan, atau preskriptif.

  Di sektor ritel, baik data sensor maupun data yang diambil dari sumber web (yaitu data produk dan data pelanggan) rentan terhadap kesalahan dan perlu diperiksa kelayakannya. Oleh karena itu diperlukan prosedur perbaikan data yang membantu menghilangkan data dan noise yang salah/berlebihan.

  - Validasi manusia melalui kurasi
  - Penghapusan kebisingan dalam jumlah besar secara otomatis dalam skala besar
  - Validasi semantik yang dapat diskalakan

- **Keamanan dan Privasi Data**

  Persyaratan keamanan dan privasi data tingkat tinggi menggambarkan kebutuhan untuk melindungi data bisnis dan pribadi yang sangat sensitif dari akses tidak sah. Oleh karena itu, hal ini membahas ketersediaan prosedur hukum dan sarana teknis yang memungkinkan pembagian data secara aman. Dalam aplikasi layanan kesehatan, penekanan kuat harus diberikan pada privasi dan keamanan data karena beberapa pendekatan perlindungan privasi yang umum dapat diabaikan oleh sifat data besar. Misalnya, dalam hal data terkait kesehatan, anonimisasi adalah pendekatan yang sudah mapan untuk menghilangkan identifikasi data pribadi. Meskipun demikian, data yang dianonimkan dapat diidentifikasi ulang (El Emam dkk. 2014) ketika menggabungkan data besar dari sumber data yang berbeda.

  Aplikasi data besar di ritel memerlukan penyimpanan informasi pribadi pelanggan agar pengecer dapat memberikan layanan yang disesuaikan. Itu sangat penting agar data ini disimpan dengan aman untuk memastikan perlindungan privasi pelanggan. Di sektor manufaktur, terdapat konflik kepentingan dalam penyimpanan data produk untuk kemudahan pengambilan dan perlindungan data dari pengambilan yang tidak sah. Data yang dikumpulkan selama produksi dan penggunaan mungkin

berisi informasi kepemilikan mengenai proses bisnis internal. Kekayaan intelektual perlu dilindungi sejauh hal tersebut dikodekan dalam data produk dan produksi. Peraturan mengenai kepemilikan data perlu ditetapkan, misalnya, akses apa yang mungkin dimiliki oleh produsen mesin produksi terhadap data penggunaannya.

Perlindungan privasi bagi pekerja yang berinteraksi di lingkungan Industri 4.0 perlu ditetapkan. Enkripsi data dan kontrol akses ke dalam memori objek perlu diintegrasikan. Peraturan Eropa dan seluruh dunia perlu diselaraskan. Perlunya peraturan privasi data dan perlindungan privasi yang transparan. Di sektor telekomunikasi dan media, salah satu kekhawatiran utama adalah bahwa kebijakan big data berlaku untuk data pribadi, yaitu data yang berkaitan dengan orang yang teridentifikasi atau dapat diidentifikasi. Namun, tidak jelas apakah prinsip privasi inti peraturan tersebut berlaku untuk pengetahuan yang baru ditemukan atau informasi yang berasal dari data pribadi, terutama ketika data tersebut telah dianonimkan atau digeneralisasikan dengan diubah menjadi profil grup. Privasi adalah masalah utama yang dapat membahayakan kepercayaan pengguna akhir, yang penting agar data besar dapat dieksploitasi oleh penyedia layanan. Survei Wawasan Konsumen Ovum (2013) mengungkapkan bahwa 68% pengguna Internet di 11 negara di seluruh dunia akan memilih fitur "Jangan Lacak" jika fitur tersebut tersedia dengan mudah. Hal ini jelas menyoroti sejumlah antipati pengguna akhir terhadap pelacakan online. Privasi dan kepercayaan merupakan penghalang penting karena data harus kaya agar bisnis dapat menggunakannya.

Menemukan solusi untuk memastikan keamanan dan privasi data dapat membuka potensi besar big data di sektor publik. Kemajuan dalam perlindungan dan privasi data merupakan kunci bagi sektor publik, karena hal ini memungkinkan analisis sejumlah besar data yang dimiliki oleh sektor publik tanpa mengungkapkan informasi sensitif. Dalam banyak kasus, peraturan sektor publik membatasi penggunaan data untuk berbagai tujuan pengumpulan data. Masalah privasi dan keamanan juga menghalangi penggunaan infrastruktur cloud (misalnya pemrosesan, penyimpanan) oleh banyak lembaga publik yang menangani data sensitif. Pendekatan baru terhadap keamanan dalam infrastruktur cloud dapat menghilangkan hambatan ini.

Persyaratan keamanan dan privasi data muncul di sektor keuangan dalam rangka membangun model bisnis baru berdasarkan data yang dikumpulkan oleh lembaga jasa keuangan dari pelanggannya (individu). Layanan inovatif dapat diciptakan dengan teknologi yang menyelaraskan penggunaan data dan persyaratan privasi. Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Algoritma hash
- Pertukaran data yang aman
- Algoritma de-identifikasi dan anonimisasi
- Teknologi penyimpanan data ke penyimpanan terenkripsi dan DB; enkripsi ulang proxy antar domain; perlindungan privasi otomatis

- Kemajuan dalam "privasi berdasarkan desain" untuk menghubungkan kebutuhan analitik dengan kontrol perlindungan dalam pemrosesan dan penyimpanan
- Asal data untuk memungkinkan transparansi penggunaan dan metadata untuk informasi privasi

Visualisasi Data dan Pengalaman Pengguna

Visualisasi data persyaratan tingkat tinggi dan pengalaman pengguna menggambarkan kebutuhan untuk menyesuaikan visualisasi dengan pengguna. Hal ini dimungkinkan dengan mengurangi kompleksitas data, keterkaitan data, dan hasil analisis data. Di ritel, sangat penting untuk menyesuaikan visualisasi informasi dengan pelanggan tertentu. Contohnya adalah iklan yang disesuaikan, yang sesuai dengan profil pelanggan.

Dalam bidang manufaktur, pengambilan keputusan dan bimbingan oleh manusia perlu didukung di semua tingkatan: mulai dari lantai produksi hingga manajemen tingkat tinggi. Alat visualisasi data yang sesuai harus tersedia dan terintegrasi untuk mendukung penelusuran, pengendalian, dan pengambilan keputusan dalam proses perencanaan dan pelaksanaan. Hal ini berlaku terutama untuk data besar secara umum, namun juga meluas dan mencakup visualisasi khusus dari aspek spatiotemporal dari proses manufaktur untuk analisis spasial dan temporal.

Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Menerapkan teknik pemodelan pengguna pada analisis visual
- Visualisasi performa tinggi
- Visualisasi skala besar berdasarkan kerangka semantik adaptif
- Antarmuka multimoda di lingkungan kerja yang tidak bersahabat
- Pemrosesan bahasa alami untuk konteks yang sangat bervariasi
- Visualisasi interaktif dan pertanyaan visual

Analisis Data Mendalam

Analisis data mendalam dengan persyaratan tingkat tinggi adalah penerapan teknik pemrosesan data canggih untuk menghasilkan informasi dari beberapa, biasanya kumpulan data besar yang terdiri dari data tidak terstruktur dan semi terstruktur. Analisis data mendalam memiliki tujuh sub-persyaratan:

- Pemodelan dan simulasi mencakup alat khusus domain untuk pemodelan dan simulasi peristiwa berdasarkan perubahan dari peristiwa masa lalu.
- Analisis bahasa alami bertujuan untuk mengekstraksi informasi dari sumber tidak terstruktur (misalnya media sosial) untuk memungkinkan analisis lebih lanjut (misalnya penambangan sentimen).
- Penemuan pola bertujuan untuk mengidentifikasi pola dan persamaan.
- Wawasan waktu nyata memungkinkan analisis data waktu nyata untuk pengambilan keputusan secara instan.
- Analisis penggunaan memberikan analisis penggunaan produk, layanan, sumber daya, proses, dll.

- Analisis prediktif memanfaatkan berbagai teknik statistik, pemodelan, penambangan data, dan pembelajaran mesin untuk mempelajari data terkini dan historis guna membuat prediksi tentang masa depan.
- Analisis preskriptif fokus pada menemukan tindakan terbaik untuk situasi tertentu.

Analisis preskriptif termasuk dalam portofolio kemampuan analitik yang mencakup analisis deskriptif dan prediktif. Meskipun analisis deskriptif bertujuan untuk memberikan wawasan tentang apa yang telah terjadi, dan analisis prediktif membantu memodelkan dan memperkirakan apa yang mungkin terjadi, analisis preskriptif berupaya menentukan solusi atau hasil terbaik di antara berbagai pilihan, berdasarkan parameter yang diketahui.

Di sektor publik, analisis data mendalam dapat membantu dalam beberapa skenario ketika informasi harus diambil dari data. Dalam skenario pemantauan dan pengawasan operator perjudian online, tantangannya adalah mendeteksi perilaku kriminal atau ilegal tertentu menggunakan penemuan pola untuk memberikan wawasan waktu nyata. Pemahaman serupa juga diperlukan dalam pengawasan pasar yang diatur oleh sektor publik (energi, telekomunikasi, pasar saham, dll.).

Skenario penerapan lainnya juga memerlukan analisis data yang mendalam, seperti dalam kasus keselamatan publik di kota pintar, di mana wawasan waktu nyata dapat memungkinkan analisis data baru atau waktu nyata untuk pengambilan keputusan secara instan. Dalam skenario ini, sistem kesadaran situasional dapat dibangun menggunakan data real-time yang disediakan oleh jaringan sensor dan data hampir real-time yang diambil dari jaringan sosial melalui analisis bahasa alami. Kesadaran situasi kota pintar juga dapat menerapkan alat pemodelan dan simulasi untuk mengelola peristiwa (misalnya mengelola kerumunan besar orang di acara publik) untuk mengantisipasi hasil dari keputusan yang diambil untuk mempengaruhi kondisi saat ini secara real-time.

Skenario penerapan lainnya seperti kebijakan prediktif mungkin memerlukan penggunaan analisis prediktif untuk memberikan wawasan berdasarkan pembelajaran dari situasi sebelumnya. Hal ini akan memungkinkan alokasi sumber daya keamanan yang optimal, sesuai dengan prediksi insiden, yang mungkin didasarkan pada pola waktu atau terkait dengan peristiwa tertentu dalam bentuk apa pun (acara olahraga, kondisi cuaca, atau variabel lainnya). Untuk sektor telekomunikasi dan media, analisis data mendalam diperlukan untuk meningkatkan pengalaman pelanggan, baik dengan menyesuaikan penawaran, meningkatkan layanan pelanggan, atau dengan secara proaktif mengadaptasi sumber daya (misalnya jaringan) untuk memenuhi harapan pelanggan dalam hal penyampaian layanan. Hal ini dapat dicapai dengan memperoleh tampilan pelanggan 360◦, yang memungkinkan pemahaman yang lebih baik tentang pelanggan dan memprediksi kebutuhan atau permintaan mereka. Pelanggan tingkat lanjut dan fleksibel segmentasi, mengetahui suka dan tidak suka pelanggan, menganalisis secara mendalam kebiasaan pengguna, interaksi pelanggan, dll., membantu penyedia layanan komunikasi dan konten menemukan pola dan sentimen dari data, memungkinkan penjualan silang berdasarkan berbagai faktor. Karena Kualitas Pengalaman (QoE) dan kepuasan pelanggan dapat berbeda dengan sangat cepat (seperti halnya suasana

hati), idealnya analitik harus menyediakan sarana untuk menghitung dan mengotomatiskan tindakan terbaik berikutnya secara real-time.

Pemrosesan analitik data besar secara historis dan online akan diadopsi karena wawasan yang diperoleh akan membuat perencanaan dan pengoperasian menjadi lebih tepat. Di sisi lain, analisis real-time masih menghadapi beberapa tantangan teknologi, yang mungkin menjadi alasan kurangnya penerapan analisis real-time di bidang energi dan transportasi. Langkah-langkah manual dalam proses analisis data pada umumnya, seperti perselisihan data, misalnya, tidak menyesuaikan kecepatan dan volume data yang akan dianalisis dalam skenario efisiensi operasional dalam optimalisasi energi dan transportasi.

Di sektor ritel, keputusan operasional dapat dioptimalkan dengan menganalisis data tidak terstruktur dari web. Ini bisa berupa informasi tentang peristiwa regional yang akan datang, data cuaca, atau bahkan potensi bencana alam yang dapat diambil dari jejaring sosial menggunakan analisis bahasa alami. Data, seperti data visual dari kamera, yang diperoleh dari sensor di dalam toko perlu dianalisis untuk mengekstrak pola tertentu, seperti pola pergerakan pelanggan. Segmentasi pelanggan dimungkinkan dengan menganalisis interaksi pelanggan-produk dan pelanggan-staf. Informasi ini juga dapat digunakan untuk menjalankan analisis preskriptif. Hal ini diperlukan untuk memungkinkan inventaris cerdas, penjadwalan staf cerdas, dan denah lantai/pengoptimalan lokasi produk.

Untuk memenuhi kebutuhan ini, tantangan-tantangan berikut perlu diatasi:

- Integrasi data, menghubungkan, dan semantik
- Analisis sentimen
- Pembelajaran mesin
- Mengintegrasikan semantik ke dalam lingkungan pemodelan dan simulasi skala besar
- Meningkatkan skalabilitas dan ketahanan ekstraksi informasi, pengenalan entitas bernama, pembelajaran mesin, data tertaut, penautan entitas, dan resolusi referensi bersama
- Validasi keluaran analisis pola dan keluaran analisis bahasa alami dengan manusia melalui kurasi
- Integrasi analisis bahasa alami ke dalam skenario penggunaan data
- Teknologi pola semantik termasuk pencocokan pola aliran dan pencocokan pola kompleks yang dapat diskalakan
- Basis data analitis untuk mendukung analisis prediktif secara efisien
- Menggabungkan penalaran skala besar dengan pendekatan statistik
- Pemeliharaan prediktif: memprediksi kegagalan, menentukan interval pemeliharaan Mendukung analisis kegagalan
- Memperluas analisis prediktif ke analisis preskriptif
- Pemrosesan peristiwa kompleks menerapkan aturan bisnis (atau kerangka kerja lainnya) secara terus-menerus pada interval aliran data real-time yang ditentukan (pendek) dengan latensi rendah

- Teknologi dalam memori, teknik visualisasi dan interaksi baru, reaksi sistem otomatis untuk memungkinkan kueri ad hoc pada kumpulan data besar dieksekusi dengan latensi minimal
- Pemrosesan analitis real-time dan in-stream

15.3 PRIORITAS PERSYARATAN LINTAS SEKTORAL

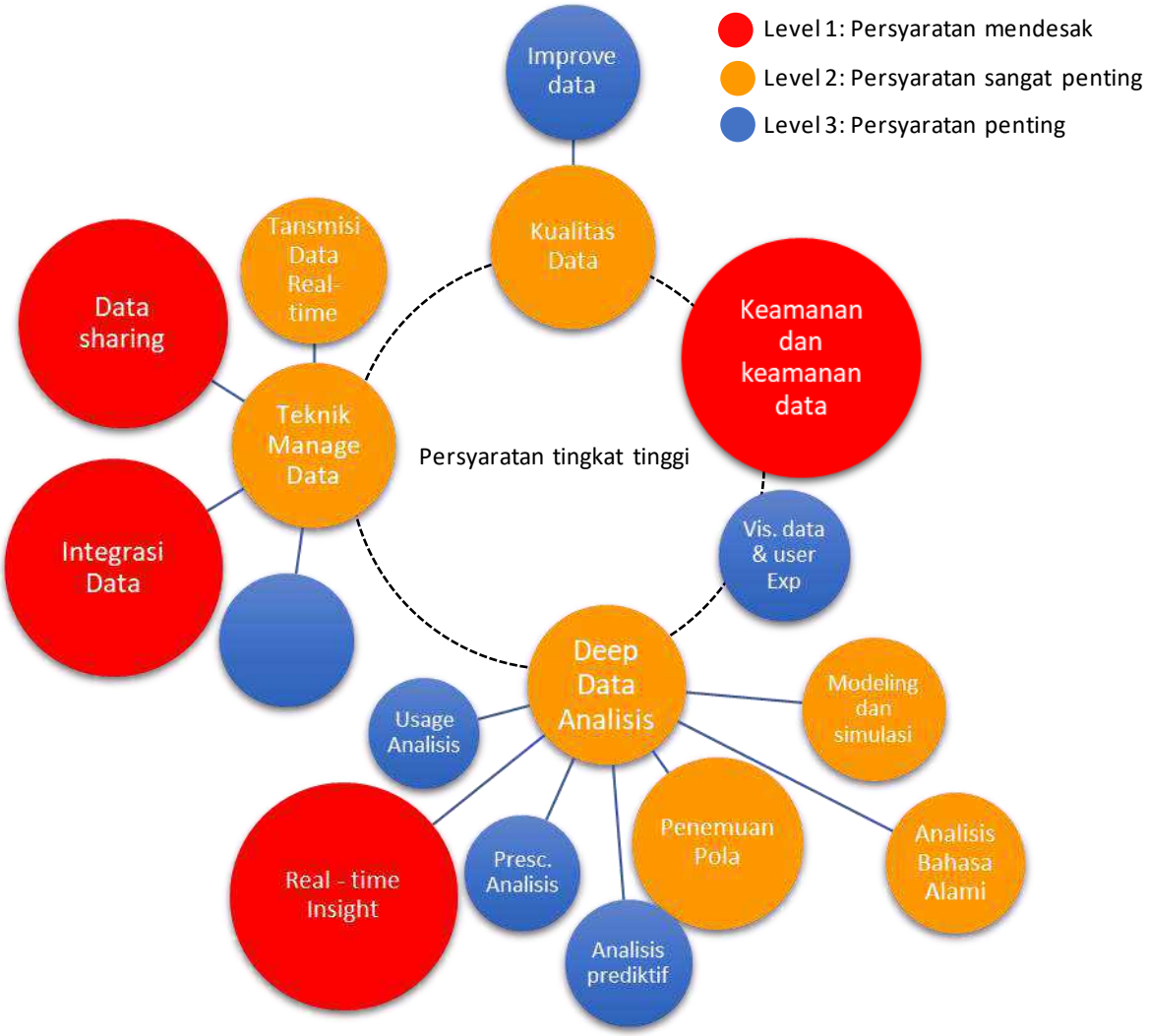
Peta jalan yang dapat ditindaklanjuti harus memiliki kriteria pemilihan yang jelas mengenai prioritas seluruh tindakan. Berbeda dengan peta jalan teknologi untuk konteks satu perusahaan, peta jalan teknologi Eropa perlu mencakup perkembangan di berbagai sektor. Proses penentuan peta jalan mencakup analisis pasar big data dan masukan yang diterima dari para pemangku kepentingan. Melalui analisis ini, dapat diketahui karakteristik apa yang menunjukkan potensi persyaratan teknis big data yang lebih tinggi atau lebih rendah.

Sebagai dasar pemeringkatan, digunakan pendekatan berbasis tabel yang mengevaluasi setiap kandidat berdasarkan sejumlah parameter yang berlaku. Dalam setiap kasus, parameter dikumpulkan dengan tujuan agar sektor ini independen. Parameter kuantitatif digunakan jika memungkinkan dan tersedia.

Melalui konsultasi dengan pemangku kepentingan, parameter berikut digunakan untuk menentukan peringkat berbagai persyaratan teknis. Parameter peringkat meliputi:

- Jumlah sektor yang terkena dampak
- Ukuran sektor yang terkena dampak dalam persentase PDB
- Perkiraan tingkat pertumbuhan sektor-sektor tersebut
- Kemungkinan perkiraan tingkat pertumbuhan sektor ini akibat teknologi big data
- Perkiraan potensi ekspor sektor tersebut
- Perkiraan manfaat lintas sektoral
- Hasil jangka pendek yang tergantung rendah

Dengan menggunakan wawasan ini, sebuah prioritas yang terdiri dari beberapa parameter dibuat, yang memberikan gambaran relatif mengenai persyaratan teknologi mana yang mungkin siap untuk mencapai keuntungan lebih besar dan mana yang akan menghadapi hambatan paling kecil. Peringkat persyaratan teknis lintas sektor disajikan pada Tabel 15.2 dan diilustrasikan pada Gambar 15.1, dimana warna menunjukkan tingkat kepentingan yang diperkirakan, dan ukuran gelembung perkiraan sektor industri yang terkena dampak. Penting untuk dicatat bahwa indeks-indeks ini tidak memberikan gambaran lengkap, namun memberikan gambaran yang masuk akal mengenai potensi ketersediaan dan cakupannya di seluruh sektor. Ada batasan tertentu pada pendekatan ini. Tidak semua angka dan masukan yang relevan tersedia karena kecepatan pengembangan dan adopsi teknologi bergantung pada beberapa faktor. Pemeringkatan ini bergantung pada prakiraan dan perkiraan dari pihak ketiga dan tim proyek. Akibatnya, tidak selalu mungkin untuk menentukan angka pasti mengenai jangka waktu dan dampak spesifik. Penyelidikan lebih lanjut terhadap pertanyaan-pertanyaan ini akan diperlukan untuk penelitian di masa depan. Rincian lengkap proses pemeringkatan tersedia di (Becker, T., Jentzsch, A., & Palmetshofer, W. 2014).



Gambar 15.1 Persyaratan lintas sektoral diprioritaskan

Tabel 15.2 Prioritas persyaratan teknis lintas sektoral

Prioritas	Persyaratan teknologi	Skor
Tingkat 1: Mendesak		
	Keamanan dan privasi data	78
	Rekayasa manajemen data—integrasi data	69.25
	Analisis data mendalam—wawasan waktu nyata	61.5
	Rekayasa manajemen data—berbagi data	48.5
Tingkat 2: Sangat penting		
	Kualitas data	40.5
	Rekayasa manajemen data—transmisi data waktu nyata	37
	Analisis data mendalam—simulasi pemodelan	37
	Analisis data mendalam—analisis bahasa alami	37
	Analisis data mendalam—penemuan pola	34.25
	Analisis data mendalam	31.75
	Rekayasa manajemen data	31.5
Tingkat 3: Penting		

	Rekayasa manajemen data—pengayaan data	29.5
	Visualisasi data dan pengalaman pengguna	29.5
	Analisis data mendalam—analisis preskriptif	29.5
	Analisis data mendalam—analisis penggunaan	26.75
	Kualitas data—peningkatan data	24
	Analisis data mendalam—analisis prediktif	20.75

Ringkasan

Tujuan dari peta jalan lintas sektor ini adalah untuk memaksimalkan dan mempertahankan dampak teknologi dan aplikasi big data di berbagai sektor industri dengan mengidentifikasi dan mendorong peluang di Eropa. Meskipun sebagian besar persyaratan yang diidentifikasi ada dalam beberapa bentuk di setiap sektor, tingkat pentingnya persyaratan antar sektor tertentu berbeda-beda. Untuk persyaratan lintas sektor, setiap persyaratan yang diidentifikasi oleh setidaknya dua sektor sebagai persyaratan signifikan bagi sektor tersebut dimasukkan ke dalam definisi peta jalan lintas sektor. Hal ini menghasilkan identifikasi 5 persyaratan tingkat tinggi dan 12 persyaratan sub-tingkat dengan tantangan terkait yang perlu diatasi.

Setiap kebutuhan lintas sektoral diprioritaskan berdasarkan dampak yang diharapkan. Hasil konsolidasi ini terdiri dari serangkaian persyaratan lintas sektor yang diprioritaskan yang digunakan untuk menentukan peta jalan lintas sektoral beserta rekomendasi tindakan terkait.

BAB 16
PETA JALAN UNTUK TEKNOLOGI, BISNIS, KEBIJAKAN, DAN MASYARAKAT

16.1 PENDAHULUAN

Tujuan utama dari proyek BIG adalah untuk menentukan peta jalan big data yang mempertimbangkan aspek teknis, bisnis, kebijakan, dan masyarakat. Bab ini menguraikan peta jalan dan rencana aksi lintas sektoral yang terpadu.

Tujuan kedua dari proyek BIG adalah untuk membentuk inisiatif yang dipimpin industri seputar manajemen informasi cerdas dan data besar untuk berkontribusi terhadap daya saing UE dan menempatkannya di Horizon 2020. Tujuan ini dicapai melalui kerja sama dengan Platform Teknologi Eropa NESSI dengan peluncuran dari Asosiasi Nilai Data Besar (BDVA). Pada akhirnya, implementasi peta jalan memerlukan mekanisme untuk mengubah peta jalan menjadi agenda nyata yang didukung oleh sumber daya yang diperlukan (investasi ekonomi dari pemangku kepentingan pemerintah dan swasta). Hal ini dibuktikan dengan penandatanganan cPPP Nilai Data Besar (BDVcPPP) antara BDVA dan Komisi Eropa. CPPP ditandatangani oleh Wakil Presiden Neelie Kroes, yang saat itu menjabat sebagai Komisaris Agenda Digital UE, dan Jan Sundelin, Presiden Big Data Value Association (BDVA), pada 13 Oktober 2014 di Brussels. CPPP BDV memberikan kerangka kerja yang menjamin kepemimpinan industri, investasi, dan komitmen pihak swasta dan publik untuk membangun perekonomian berbasis data di seluruh Eropa. Tujuan strategis BDVcPPP adalah untuk menguasai pembangkitan nilai dari data besar dan menciptakan keunggulan kompetitif yang signifikan bagi industri Eropa yang akan meningkatkan pertumbuhan ekonomi dan lapangan kerja. BDVA telah menghasilkan Agenda Riset & Inovasi Strategis (SRIA) tentang Nilai Big Data yang pada awalnya didukung oleh makalah teknis dan peta jalan BIG dan diperluas dengan masukan dari konsultasi publik yang mencakup ratusan pemangku kepentingan tambahan yang mewakili kedua belah pihak. dan sisi permintaan.

Bab ini menjelaskan peta jalan teknologi, bisnis, kebijakan, dan masyarakat yang ditentukan oleh proyek BIG. Bagian ini kemudian memperkenalkan Big Data Value Association dan kontrak Kemitraan Pemerintah Swasta Big Data Value dan menjelaskan peran yang dimainkan oleh proyek BIG dalam pendiriannya. BDVA dan cPPP BDV akan memberikan kerangka kerja yang diperlukan bagi kepemimpinan industri, investasi, dan komitmen pihak swasta dan publik untuk membangun perekonomian berbasis data di seluruh Eropa.

16.2 MENGAKTIFKAN EKOSISTEM BIG DATA

Big data kini menjadi praktik yang ada di mana-mana, baik di sektor publik maupun swasta. Ini bukan solusi yang berdiri sendiri dan bergantung pada banyak lapisan seperti infrastruktur, Internet of Things, broadband, jaringan, sumber terbuka, dan masih banyak lagi.

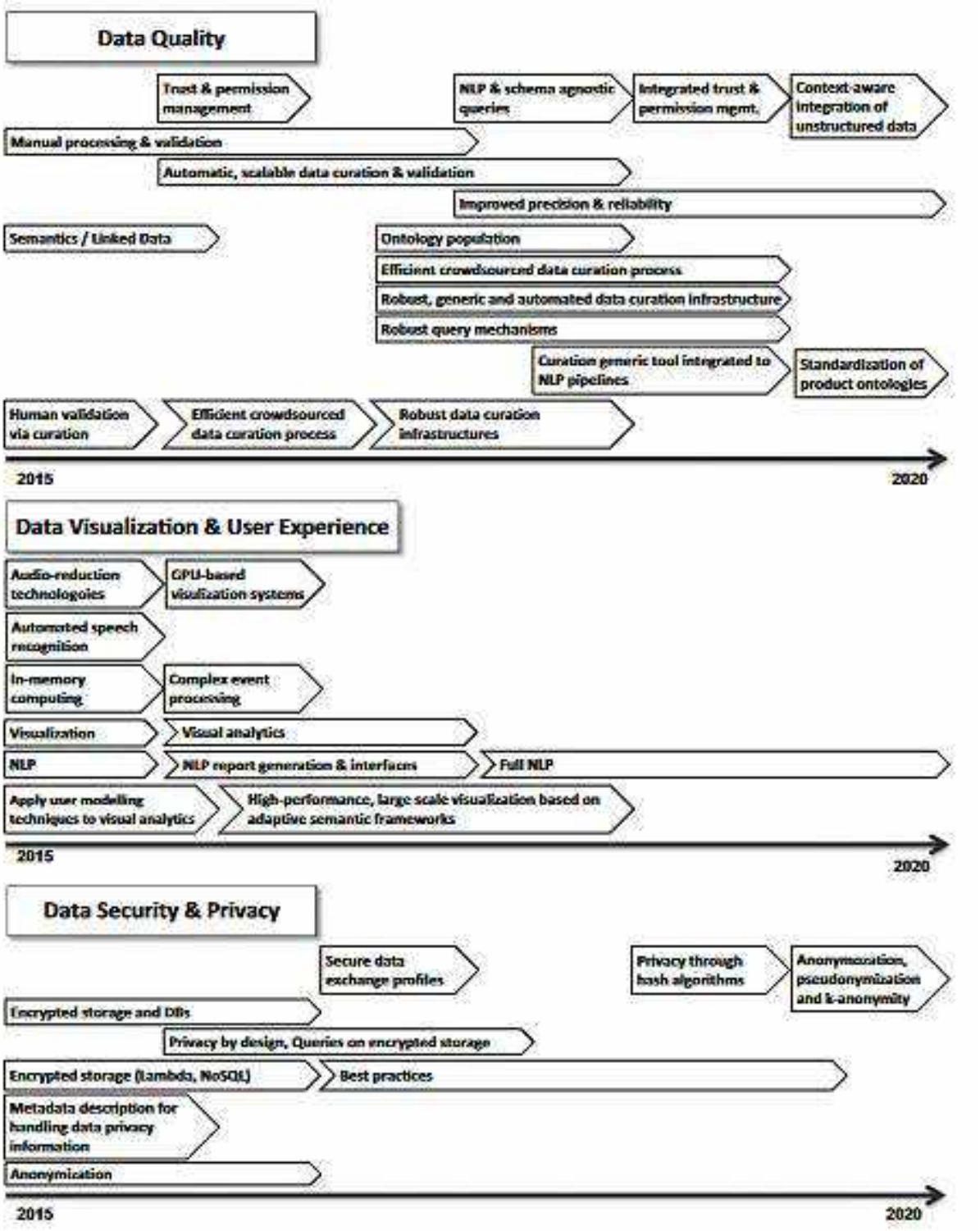
Selain itu, isu-isu penting lainnya adalah isu-isu non-teknis termasuk kebijakan, keterampilan, peraturan, dan model bisnis.

Big data harus dimasukkan dalam agenda bisnis Eropa. Oleh karena itu, para pembuat kebijakan perlu bertindak tepat waktu untuk menciptakan lingkungan yang mendukung organisasi-organisasi yang ingin mengambil manfaat dari perkembangan yang tak terelakkan ini dan peluang-peluang yang ada. Kegagalan mengembangkan ekosistem big data yang komprehensif dalam beberapa tahun ke depan membawa risiko hilangnya keunggulan kompetitif dibandingkan kawasan global lainnya.

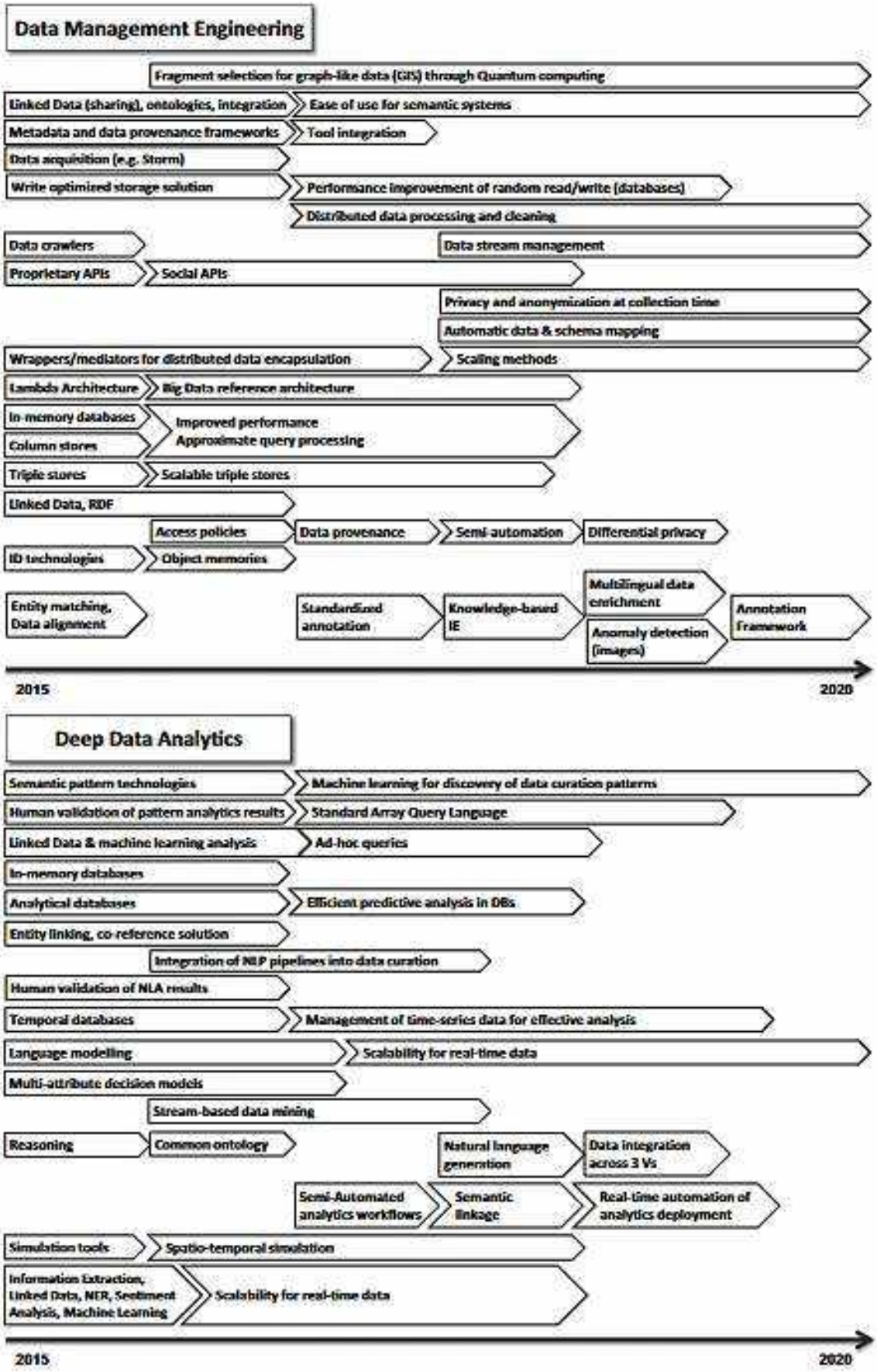
Peta jalan yang dijelaskan dalam bab ini menguraikan isu-isu paling mendesak dan menantang untuk big data di Eropa. Hal ini didasarkan pada penelitian selama lebih dari 2 tahun dan masukan dari berbagai pemangku kepentingan terkait kebijakan, bisnis, masyarakat, dan teknologi. Peta jalan ini akan mendorong terciptanya ekosistem big data. Hal ini akan memungkinkan perusahaan, dunia usaha (baik besar maupun kecil), pengusaha, start-up, dan masyarakat untuk memperoleh manfaat dari big data di Eropa. Bab ini menyajikan ringkasan peta jalan; deskripsi lengkap tersedia di Becker dkk. (2014).

16.3 PETA JALAN TEKNOLOGI UNTUK BIG DATA

menentukan teknologi mana yang dibutuhkan pada titik waktu tertentu diperlukan pendekatan sistematis untuk memprediksi perkembangan teknologi. Peta jalan teknologi spesifik sektor yang dikembangkan menetapkan kerangka kerja tersebut dengan menyelaraskan kebutuhan pengguna dan persyaratan terkait dengan kemajuan teknologi dan pertanyaan penelitian terkait. Berbeda dengan peta jalan teknologi yang dikembangkan dalam konteks satu perusahaan, pendekatan yang diambil di sini mencakup pengembangan peta jalan teknologi untuk pasar Eropa. Sebagai konsekuensinya, tidak mungkin untuk menentukan garis waktu yang tepat mengenai pencapaian teknologi, karena kecepatan pengembangan teknologi dan penerapannya bergantung pada (a) pada sejauh mana persyaratan non-teknis yang teridentifikasi akan dipenuhi dan (b) mengenai sejauh mana organisasi-organisasi Eropa bersedia berinvestasi dan memanfaatkan big data. Gambar 16.1 menggambarkan peta jalan teknologi konsolidasi untuk big data. Untuk peta jalan teknologi spesifik sektor, lihat Bagian II buku ini dan Zillner dkk. (2014). Untuk penjelasan lebih rinci tentang peta jalan teknologi konsolidasi, lihat Becker dkk. (2014).



Gambar 16.1 Peta jalan teknologi untuk big data



Gambar 16.1 (lanjutan)

16.4 PETA JALAN BISNIS UNTUK BIG DATA

Peran dunia usaha sangat penting dalam penerapan big data di Eropa. Dunia usaha perlu memahami potensi teknologi big data dan memiliki kemampuan untuk menerapkan strategi dan teknologi yang tepat untuk keuntungan komersial. Peta jalan bisnis big data disajikan pada Tabel 16.1.

Tabel 16.1 Peta Jalan Bisnis Untuk Big Data

Bisnis	2015	2019 atau sebelumnya
1. Sikap perubahan dan jiwa wirausaha	Perubahan pada manajemen tingkat atas dimulai dan aktivitas kewirausahaan didorong.	Manajemen tingkat atas di bisnis-bisnis Eropa memiliki pola pikir yang berbasis data besar.
2. Model bisnis	Menjelajahi model bisnis yang didorong oleh data besar.	Berhasil mengeksploitasi model bisnis big data baru.
3. Privasi berdasarkan desain	Mulai terapkan privasi berdasarkan desain.	Privasi berdasarkan desain secara default.
4. Pendidikan tenaga kerja	Program pendidikan tenaga kerja baru mengenai big data.	Peningkatan signifikan dalam jumlah karyawan yang paham big data di semua departemen.
5. Standardisasi	Identifikasi standardisasi penting yang diperlukan.	Langkah-langkah besar dalam standardisasi telah tercapai.
6. Meningkatkan penelitian dan pengembangan	Meningkatkan pengeluaran penelitian dan pengembangan data besar.	Peningkatan minimal 25% dalam belanja penelitian dan pengembangan big data

Sikap terhadap Perubahan dan Semangat Kewirausahaan Mayoritas perusahaan di Eropa dan para pemimpinnya perlu mengatasi permasalahan inti dalam penggunaan data untuk menggerakkan organisasi mereka. Hal ini mengharuskan inovasi berbasis data menjadi prioritas di tingkat atas organisasi, tidak hanya di departemen TI. Jiwa kewirausahaan diperlukan dalam tim kepemimpinan untuk menghadapi perubahan yang cepat dan ketidakpastian dalam dunia bisnis big data. Perubahan, bahkan dengan kemungkinan konsekuensi kegagalan, harus diterima.

Model Bisnis Di tahun-tahun mendatang, lingkungan bisnis akan mengalami perubahan besar akibat transformasi big data. Model bisnis yang ada bisa berubah dan model baru akan muncul. Dunia usaha masih belum mengetahui analisis data apa yang relevan dan bermanfaat bagi bisnis mereka, dan laba atas investasi sering kali tidak jelas. Namun, mereka menyadari perlunya menganalisis data yang mereka kumpulkan untuk mendapatkan keunggulan kompetitif dan menciptakan peluang bisnis baru. Adaptasi terhadap perubahan ini sangat penting bagi keberhasilan banyak organisasi. Privasi berdasarkan Desain Privasi berdasarkan desain dapat memperoleh lebih banyak kepercayaan dari pelanggan dan pengguna. Eropa perlu mengambil peran utama dalam menggabungkan privasi dengan operasi bisnis di semua sektornya.

Pendidikan Tenaga Kerja Ada perang untuk memperebutkan talenta big data. Dunia usaha harus fokus pada pelatihan dan pendidikan seluruh staf mereka, tidak hanya dari departemen TI, dengan keterampilan terkait big data yang diperlukan. Standardisasi Bisnis perlu bekerja sama dengan pemangku kepentingan dan organisasi lain untuk menciptakan standar teknologi dan data yang diperlukan untuk memungkinkan ekosistem data besar. Kurangnya standar, karena tidak adanya interoperabilitas, misalnya database NoSQL dan database SQL, merupakan hambatan utama dalam penerapan big data secara lebih cepat.

Meningkatkan Penelitian dan Pengembangan Perusahaan harus fokus untuk tidak kehilangan keunggulan dan berinvestasi dalam penelitian dan pengembangan data besar untuk mendapatkan keunggulan kompetitif bagi organisasi mereka. Dukungan yang tepat harus diberikan baik pada sektor publik maupun swasta untuk mendorong penelitian dan inovasi yang diperlukan untuk nilai big data.

16.5 PETA JALAN KEBIJAKAN UNTUK BIG DATA

Kebijakan dan agenda Eropa sangat penting untuk memastikan bahwa big data dapat mencapai potensi maksimalnya di Eropa. Peta jalan kebijakan untuk big data tersedia pada Tabel 16.2. Pendidikan dan Keterampilan Pengakuan dan promosi literasi digital sebagai keterampilan penting abad kedua puluh satu adalah salah satu bidang paling penting bagi keberhasilan jangka panjang big data di Eropa. Saat ini terdapat kekurangan tenaga profesional TI dan big data yang sangat besar, dan Eropa diperkirakan akan menghadapi kekurangan hingga 900.000 tenaga profesional ICT pada tahun 2020. Kekurangan keterampilan ini membahayakan potensi pertumbuhan dan daya saing digital. Menurut sejumlah penelitian, permintaan akan pekerja big data tertentu (misalnya ilmuwan data, insinyur data, arsitek, analis) akan semakin meningkat hingga 240 % dalam 5 tahun ke depan yang dapat mengakibatkan tambahan 100.000 pekerja terkait data. lapangan kerja pada tahun 2020. Masalah ini tidak hanya berdampak pada domain big data, namun juga seluruh lanskap digital dan harus ditangani secara umum, luas, dan mendesak. Literasi data dan kode harus diintegrasikan ke dalam kurikulum standar sejak usia dini. Keterampilan data besar yang spesifik seperti teknik data, ilmu data, teknik statistik, dan disiplin ilmu terkait harus diajarkan di institusi pendidikan tinggi. Akses yang lebih mudah terhadap izin kerja bagi warga non-Eropa juga harus dipertimbangkan untuk membantu memacu perekonomian big data Eropa.

Pasar Tunggal Digital Eropa Terlepas dari kenyataan bahwa ekonomi digital telah ada sejak lama, pasar tunggal UE masih berfungsi paling baik di bidang-bidang yang lebih tradisional seperti perdagangan barang fisik. Sejauh ini negara tersebut gagal beradaptasi terhadap berbagai tantangan ekonomi digital.

Pasar tunggal digital yang mapan dapat memimpin dunia dalam teknologi digital. Para pengambil kebijakan perlu mendorong harmonisasi. Hal ini berarti menggabungkan 28 sistem peraturan yang berbeda, menghilangkan hambatan, mengatasi fragmentasi, dan meningkatkan standar teknis dan interoperabilitas. Mencapai tujuan ini pada tahun 2019 merupakan hal yang cukup ambisius, namun ini merupakan langkah penting menuju bidang data bersama Eropa di masa depan.

Tabel 16.2 Peta Jalan Kebijakan Untuk Big Data

Kebijakan	2015	2019 atau sebelumnya
1 Pendidikan dan keterampilan	Kekurangan dalam pendidikan big data telah diatasi.	Benua terbaik untuk pendidikan big data.
2 Pasar tunggal digital	Fokus pada penciptaan pasar data tunggal Eropa.	Pasar data tunggal Eropa untuk 500 juta pengguna terbentuk.
3 Pendanaan untuk teknologi big data	Pertahankan tingkat pendanaan saat ini (850 juta).	Dua kali lipat ukuran dunia modal ventura di Eropa pada tahun 2015.
4 Buka data dan silo data	Pembahasan mengenai data pemerintahan terbuka secara default.	Eropa memimpin dalam data terbuka. Silo data yang diminimalkan.
5 Privasi dan hukum	Memulai perdebatan publik, Perlindungan Data UE ditandatangani.	Keseimbangan yang tepat untuk masyarakat dan bisnis tercapai.
6 Membina infrastruktur teknis	Terus membina lingkungan TI.	Infrastruktur Eropa bersaing atau melampaui Amerika/Asia.

Pendanaan untuk Teknologi Big Data Ciptakan lingkungan start-up yang lebih ramah dengan peningkatan akses terhadap pendanaan. Kurangnya pendanaan yang memadai untuk penelitian dan inovasi. Dukungan dan pendanaan publik harus ditingkatkan. Namun, mengingat keterbatasan anggaran saat ini di Eropa, pendekatan alternatif juga perlu dipertimbangkan (seperti memberikan insentif hukum untuk investasi pada big data, European Investment Bank, dll.).

Eropa juga kurang memiliki atmosfer kewirausahaan (yaitu modal ventura yang dibelanjakan per kapita dibandingkan dengan AS atau Israel). Membina lingkungan pembiayaan swasta yang lebih baik bagi perusahaan rintisan dan UKM sangatlah penting. Privasi dan Hukum Memberikan aturan yang jelas, dapat dimengerti, dan masuk akal mengenai privasi data. Terkait hak privasi dan big data, terdapat tantangan ganda yang dihadapi, yaitu tidak adanya Pasar Tunggal Digital Eropa, dan tidak adanya hak pengguna yang terpadu. Hal ini perlu segera diatasi, karena kepercayaan dan adopsi teknologi big data bergantung pada kepercayaan pengguna. Berdasarkan indikasi terbaru, Perlindungan Data UE diperkirakan akan ditandatangani pada tahun 2015, namun diskusi yang lebih luas masih diperlukan. Hal lain yang perlu diperhatikan adalah hak cipta dan apakah ada hak kepemilikan data.

Tidak peduli seberapa cepat kemajuan teknologi, masyarakat tetap mempunyai kewenangan untuk memastikan bahwa inovasi didorong dan nilai-nilai dilindungi melalui hukum, kebijakan, dan praktik yang didorong di sektor publik dan swasta. Untuk itu, pembuat kebijakan harus menetapkan aturan yang jelas mengenai privasi data sehingga organisasi mengetahui data pribadi apa yang dapat mereka simpan dan untuk berapa lama, serta data apa yang secara eksplisit dilindungi oleh peraturan privasi. Pembuat kebijakan perlu

memajukan undang-undang konsumen dan privasi untuk memastikan konsumen memiliki standar yang jelas, dapat dipahami, dan masuk akal mengenai bagaimana informasi pribadi mereka digunakan di era big data.

Data Terbuka dan Silo Data Data terbuka dapat menciptakan perubahan budaya dalam organisasi menuju berbagi data dan kerja sama. Mulai dari mengurangi biaya pengelolaan data hingga menciptakan peluang bisnis baru, banyak organisasi memperoleh manfaat dari keterbukaan dan berbagi data perusahaan terpilih. Pemerintahan di Eropa perlu memulai diskusi mengenai keterbukaan. Memanfaatkan data sebagai sumber daya publik untuk meningkatkan penyampaian layanan publik. Semakin cepat pemerintah Eropa membuka datanya, semakin tinggi pula keuntungannya. Big Open Data harus menjadi tujuan jika memungkinkan.

Menumbuhkan Infrastruktur Teknis Big data bukanlah solusi yang berdiri sendiri dan bergantung pada banyak lapisan seperti infrastruktur, Internet of Things, akses broadband untuk pengguna, jaringan, open source, dan banyak lagi. Pemupukan silang dari lapisan-lapisan ini sangat penting untuk keberhasilan data besar. Dorongan teknologi diperlukan untuk memperkuat penyedia teknologi Eropa untuk menyediakan infrastruktur big data yang kompetitif atau terdepan jika dibandingkan dengan kawasan lain.

16.6 PETA JALAN MASYARAKAT UNTUK BIG DATA

Selain peta jalan bisnis dan kebijakan yang disajikan, peta jalan untuk masyarakat di Eropa juga telah ditetapkan. Tanpa dukungan masyarakat Eropa, penggunaan teknologi big data akan tertunda dan peluang yang ada akan hilang. Kampanye untuk meningkatkan kesadaran akan manfaat big data akan berguna untuk memotivasi warga dan masyarakat Eropa. Kampanye ini dapat mencakup promosi tokoh-tokoh teladan (terutama perempuan, dan orang-orang dengan latar belakang beragam) dan dampak positif jangka panjang dari sektor TI dan inovasi. Peta jalan masyarakat untuk big data disajikan pada Tabel 16.3.

Tabel 16.3 Peta Jalan Masyarakat Untuk Big Data

Masyarakat	2015	2019 atau sebelumnya
1. Pendidikan dan keterampilan	Apakah Anda sudah membuat kode?	Jumlah orang yang ahli dalam bidang coding dan big data di Eropa empat kali lipat dibandingkan tahun 2014.
2. Jaringan kolaboratif	Apakah kamu terhubung?	Benua terdepan dalam hal komunitas data besar yang demokratis.
3. Data terbuka	Apakah Anda sudah terlibat dalam data terbuka?	Eropa adalah masyarakat data terbuka terkemuka.
4. Kewirausahaan	Apakah Anda seorang inovator data?	Peningkatan signifikan dalam kewirausahaan big data.

5. Keterlibatan sipil	Apakah Anda memberikan suara atau tetap berhubungan dengan Anggota Parlemen Eropa (MEP) Anda?	Eropa adalah masyarakat yang paling banyak menggunakan data besar secara digital dan politik.
6. Privasi dan kepercayaan	Apa pendapat Anda tentang privasi? Apakah Anda mempercayai data besar?	Benua Eropa terdepan dalam privasi. Peningkatan signifikan dalam kepercayaan terhadap data besar.

Pendidikan dan Keterampilan Pengetahuan matematika dan statistik, dipadukan dengan keterampilan coding dan data merupakan dasar dari literasi big data. Meningkatkan literasi big data penting bagi masyarakat berbasis data. Penting bagi anggota masyarakat untuk mengembangkan kefasihan dalam memahami cara pengumpulan dan pembagian data, bagaimana algoritma digunakan, dan untuk tujuan apa. Penting untuk memastikan warga negara dari segala usia memiliki kemampuan dan alat yang diperlukan untuk melindungi diri mereka secara memadai dari penggunaan dan penyalahgunaan data. Inisiatif seperti "Code Week for Europe" adalah contoh yang baik untuk acara serupa di domain big data.

Jaringan Kolaboratif Semua segmen masyarakat, mulai dari hacker hingga start-up, dari UKM hingga bisnis besar, dari investor hingga politisi di Brussels, harus bekerja sama untuk memajukan agenda big data di Eropa. Eropa berpeluang menjadi benua yang merangkul big data melalui proses demokrasi dari bawah ke atas. Data Terbuka Data terbuka adalah cara yang baik untuk melibatkan masyarakat dan menggambarkan manfaat positif data besar bagi perubahan organisasi, efisiensi, dan transparansi (tentu saja hanya dengan data terbuka pemerintah yang non-pribadi). Tujuannya adalah menjadi data terbuka yang besar untuk Eropa.

Kewirausahaan Perkembangan TI dan big data saat ini memberikan dampak yang luar biasa terhadap dunia bisnis dan masyarakat secara keseluruhan. Kesempatan untuk mengubah keadaan menjadi lebih baik bagi masyarakat perlu diambil. Akses yang terjangkau terhadap alat, data, teknologi, dan layanan diperlukan untuk menumbuhkan ekosistem dukungan bagi wirausahawan komersial dan sosial untuk memanfaatkan potensi data besar untuk menciptakan produk dan layanan baru, mendirikan perusahaan rintisan, dan mendorong penciptaan lapangan kerja baru.

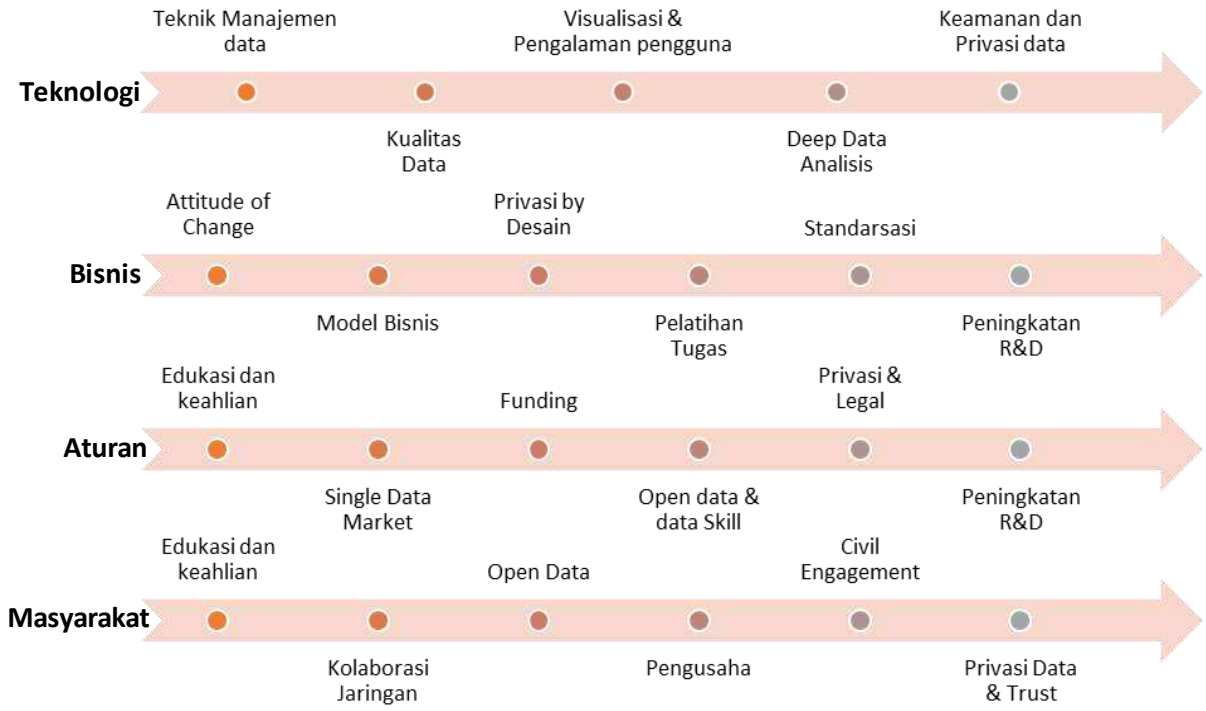
Keterlibatan Masyarakat Setiap orang di Eropa dapat mengubah cara Eropa menangani dampak big data dengan mempengaruhi politik dan kebijakan di Brussels. Warga negara perlu memahami bahwa "Eropa adalah Anda" dan bahwa partisipasi mereka dalam kehidupan politik komunitas Eropa di era transisi digital ini sangat diperlukan. Masyarakat sipil harus memainkan peran penting, yang bergantung pada setiap warga negara untuk menjadi warga negara yang terlibat.

Privasi dan Kepercayaan Hal penting dalam keberhasilan big data di Eropa adalah perlunya diskusi terbuka mengenai pro dan kontra big data dan privasi untuk membangun kepercayaan warga. Perbedaan sudut pandang yang ada di negara-negara anggota Eropa dan

warganya perlu diatasi. Kepercayaan harus dibangun di pasar data tunggal digital Eropa di mana kebebasan konsumen dan sipil dilindungi. Warga negara harus bersuara; jika tidak, tuntutan mereka tidak akan didengar dalam diskusi yang sedang berlangsung mengenai privasi.

16.7 PETA JALAN BIG DATA EROPA

Langkah terakhir adalah membuat peta jalan terpadu yang mempertimbangkan aspek teknis, bisnis, kebijakan, dan masyarakat. Peta jalan big data Eropa yang dihasilkan merupakan konsensus yang mencerminkan peta jalan dengan prioritas dan tindakan yang ditentukan yang diperlukan untuk big data di Eropa. Peta jalan (seperti yang diilustrasikan pada Gambar 16.2) adalah hasil analisis ekstensif dan keterlibatan pemangku kepentingan dalam ekosistem big data selama lebih dari 2 tahun. Penting untuk dicatat bahwa meskipun tindakan divisualisasikan secara berurutan, pada kenyataannya banyak tindakan yang dapat dan harus ditangani pada waktu yang sama secara paralel, sebagaimana dirinci dalam peta jalan tertentu.



Gambar 16.2 Peta Jalan Big Data Eropa

16.8 MENUJU PEREKONOMIAN BERBASIS DATA UNTUK EROPA

Dalam pidatonya sebagai Komisaris Digital Eropa, Neelie Kroes menyerukan tindakan dari para pemangku kepentingan Eropa untuk memobilisasi seluruh masyarakat, industri, akademisi, dan penelitian guna mewujudkan ekonomi big data Eropa. VP Kroes mengidentifikasi perlunya membangun dan mendukung kerangka kerja untuk memastikan terdapat cukup pekerja data berketerampilan tinggi (analis, pemrogram, insinyur, ilmuwan, jurnalis, politisi, dll.) untuk dapat menghasilkan teknologi, produk, dan teknologi masa depan.

layanan yang diperlukan untuk rantai nilai big data dan untuk memastikan komunitas pemangku kepentingan yang berkelanjutan di masa depan.

Tujuan utama dari proyek BIG adalah untuk menciptakan dan meningkatkan koneksi baru dalam ekosistem big data di seluruh Eropa saat ini, dengan mendorong terciptanya kemitraan baru yang lintas sektor dan domain. Eropa perlu membangun pemain-pemain yang kuat untuk menjadikan seluruh ekosistem nilai big data, dan akibatnya perekonomian Eropa, kuat, bersemangat, dan berharga. BIG menyadari perlunya menciptakan wadah yang memungkinkan terjadinya interkoneksi dan interaksi ide-ide serta kemampuan big data yang akan mendukung keberlanjutan, akses, dan pengembangan platform komunitas big data dalam jangka panjang. Keterhubungan antar pemangku kepentingan akan membentuk dasar bagi ekosistem berbasis data besar sebagai sumber peluang dan inovasi bisnis baru. Pemupukan silang antar pemangku kepentingan merupakan elemen kunci untuk memajukan perekonomian big data yang berkelanjutan.

16.9 ASOSIASI NILAI BIG DATA

Forum Publik Swasta Big Data, demikian sebutan awalnya, dimaksudkan untuk menciptakan jalan menuju implementasi peta jalan. Jalan tersebut memerlukan dua elemen utama: (1) mekanisme untuk mengubah peta jalan menjadi agenda nyata yang didukung oleh sumber daya yang diperlukan (investasi ekonomi dari pemangku kepentingan publik dan swasta) dan (2) komunitas yang berkomitmen untuk melakukan investasi dan berkolaborasi. penilaian terhadap implementasi agenda.

Konsorsium BIG yakin bahwa untuk mencapai hasil ini diperlukan kesadaran dan komitmen yang luas di luar proyek. BIG mengambil langkah-langkah yang diperlukan untuk menghubungi pemain-pemain besar dan bekerja sama dengan Platform Teknologi Eropa NESSI untuk bersama-sama berupaya mewujudkan upaya ini. Kolaborasi ini dimulai pada musim panas tahun 2013 dan memungkinkan mitra BIG untuk membangun koneksi tingkat tinggi yang diperlukan baik di tingkat industri maupun politik. Tujuan ini dicapai melalui kerja sama dengan NESSI dengan peluncuran Big Data Value Association (BDVA) dan kontrak Big Data Value Public Private Partnership (BDV cPPP) dalam Horizon 2020.

BDVA adalah organisasi nirlaba yang sepenuhnya dibiayai sendiri berdasarkan hukum Belgia dengan 24 anggota pendiri dari industri besar dan kecil serta penelitian, termasuk banyak mitra proyek BIG. BDVA adalah komunitas perwakilan pemangku kepentingan yang dipimpin oleh industri yang siap berkomitmen pada cPPP nilai data besar dengan kemauan untuk menginvestasikan uang dan waktu.

Tujuan BDVA adalah untuk meningkatkan penelitian, pengembangan, dan inovasi nilai big data Eropa. Hal ini bertujuan untuk:

a) Memperkuat daya saing dan memastikan kepemimpinan industri dalam penyedia dan pengguna akhir sistem dan layanan berbasis teknologi nilai besar
b) Mempromosikan penggunaan teknologi dan layanan nilai big data secara luas dan paling efektif untuk penggunaan profesional dan pribadi
c) Menetapkan keunggulan ilmiah sebagai landasan penciptaan nilai dari big data

BDVA akan melaksanakan beberapa kegiatan untuk mencapai tujuannya, antara lain:

1 Mengembangkan tujuan strategis untuk penelitian dan inovasi nilai big data Eropa, dan mendukung implementasinya
2 Meningkatkan daya saing industri Eropa melalui teknologi, aplikasi, layanan, dan solusi nilai big data yang inovatif
3 Memperkuat aktivitas jaringan komunitas nilai data besar Eropa
4 Mempromosikan penawaran dan organisasi nilai big data Eropa
5 Menjangkau pengguna baru dan lama
6 Berkontribusi pada pengembangan kebijakan, pendidikan, dan percabangan teknologi di bidang etika, hukum, dan kemasyarakatan

16.10 NILAI BIG DATA PEMERINTAH SWASTA

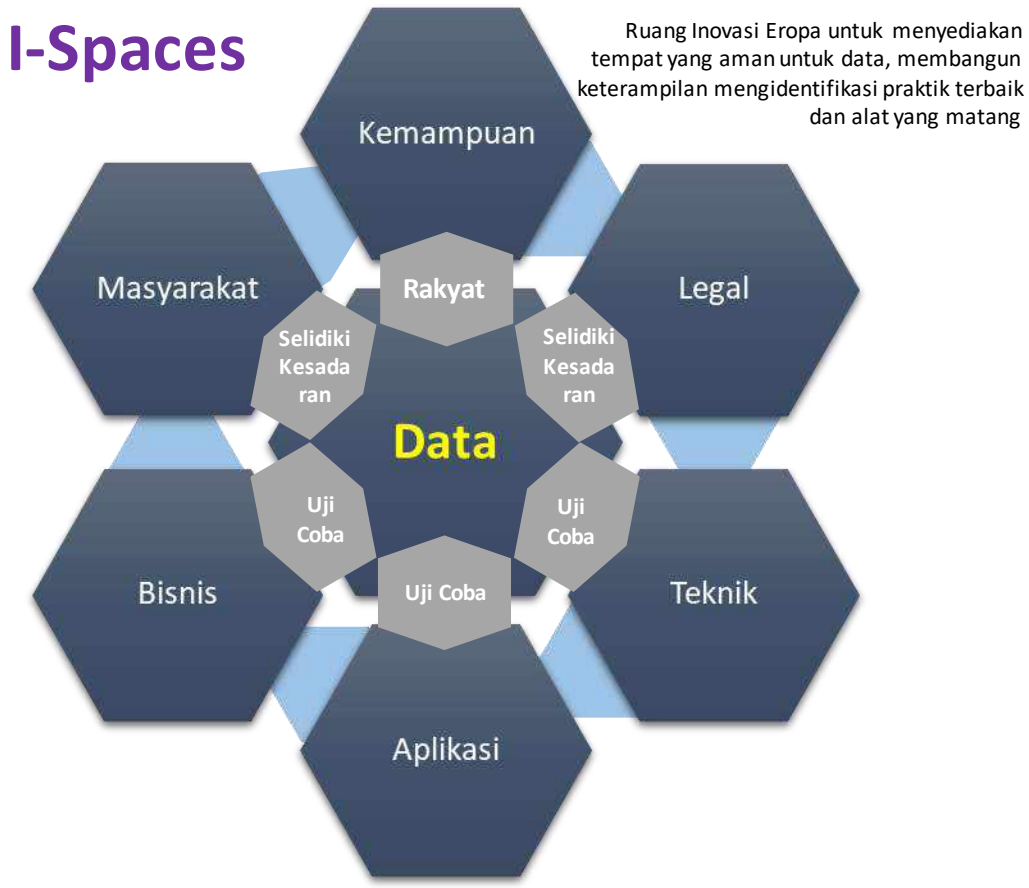
BDVA mengembangkan Agenda Riset & Inovasi Strategis (SRIA) mengenai Nilai Big Data (BDVA 2015) yang pada awalnya didukung oleh makalah teknis dan peta jalan BIG dan diperluas dengan masukan dari konsultasi publik yang mencakup ratusan pemangku kepentingan tambahan yang mewakili kedua belah pihak. dan sisi permintaan. BDVA kemudian mengembangkan proposal cPPP (KPS kontraktual) sebagai langkah formal untuk menyiapkan KPS pada nilai big data. Proposal cPPP dibangun berdasarkan SRIA dengan menambahkan elemen konten tambahan seperti instrumen potensial yang dapat digunakan untuk implementasi agenda.

Peran penting dalam lanskap big data Eropa akan dipenuhi oleh kontrak Kemitraan Pemerintah dan Swasta Nilai Big Data (BDV cPPP). Pada tanggal 13 Oktober 2014 penandatanganan cPPP BDV dilakukan di Brussels, oleh Wakil Presiden Komisi Eropa saat itu Neelie Kroes dan Presiden BDVA Jan Sundelin, TIE Kinetix. BDVA adalah mitra kontrak Komisi Eropa yang dipimpin industri untuk penerapan cPPP BDV. Peran utama BDVA adalah memperbarui SRIA Nilai Big Data secara berkala, menentukan dan memantau metrik cPPP BDV, dan berpartisipasi dengan Komisi Eropa dalam dewan kemitraan cPPP BDV.

Penandatanganan cPPP BDV adalah langkah pertama menuju pembangunan komunitas data yang berkembang di UE. CPPP BDV didorong oleh keyakinan bahwa penelitian dan inovasi yang berfokus pada kombinasi kebutuhan bisnis dan penggunaan adalah strategi jangka panjang terbaik untuk memberikan nilai dari big data serta menciptakan lapangan kerja dan kesejahteraan. Tujuan strategis cPPP BDV sebagaimana tertuang dalam BDV SRIA (BDVA 2015) adalah:

- **Data**: Untuk mengakses, menyusun, dan menggunakan data dengan cara yang sederhana dan jelas yang memungkinkan transformasi data menjadi informasi.
- **Keterampilan**: Untuk berkontribusi pada kondisi pengembangan keterampilan di industri dan akademisi.
- **Hukum dan Kebijakan**: Untuk berkontribusi pada proses kebijakan untuk menemukan lingkungan peraturan Eropa yang menguntungkan, dan mengatasi permasalahan privasi dan inklusi warga.

- **Teknologi**: Untuk menumbuhkan kepemimpinan teknologi BDV Eropa untuk penciptaan lapangan kerja dan kesejahteraan dengan menciptakan basis teknologi dan aplikasi di seluruh Eropa dan membangun kompetensi. Selain itu, aktifkan penelitian dan inovasi, termasuk dukungan interoperabilitas dan standardisasi, untuk landasan masa depan penciptaan BDV di Eropa.
- **Aplikasi**: Untuk memperkuat kepemimpinan dan kemampuan industri Eropa agar berhasil bersaing di tingkat global dalam pasar solusi nilai data dengan memajukan aplikasi yang diubah menjadi peluang baru bagi bisnis.
- **Bisnis:** Untuk memfasilitasi percepatan ekosistem bisnis dan model bisnis yang sesuai dengan fokus khusus pada UKM, yang ditegakkan dengan tolok ukur penggunaan, efisiensi, dan manfaat di seluruh Eropa.
- **Sosial**: Memberikan solusi sukses terhadap tantangan-tantangan sosial utama yang dihadapi Eropa seperti kesehatan, energi, transportasi, dan lingkungan hidup. Dan untuk meningkatkan kesadaran tentang manfaat BDV bagi dunia usaha dan sektor publik, sekaligus melibatkan masyarakat sebagai prosumer untuk mempercepat penerimaan dan penyerapan.



Gambar 16.3 Tantangan yang saling berhubungan dari cPPP BDV dalam i-Spaces

Mengingat luasnya cakupan tujuan yang berfokus pada berbagai aspek nilai big data, diperlukan strategi implementasi yang komprehensif. BDVA SRIA (BDVA 2015) merinci

pendekatan implementasi interdisipliner yang mengintegrasikan keahlian dari berbagai bidang yang diperlukan untuk mencapai tujuan strategis dan spesifik cPPP BDV. Strategi ini berisi sejumlah jenis mekanisme yang berbeda, termasuk lingkungan lintas organisasi dan lintas sektoral yang dikenal sebagai i-Spaces, seperti diilustrasikan pada Gambar 16.3, yang akan memungkinkan tantangan ditangani secara interdisipliner sekaligus berfungsi sebagai penghubung bagi kegiatan penelitian dan inovasi, proyek mercusuar yang akan meningkatkan kesadaran akan peluang yang ditawarkan oleh data besar dan nilai aplikasi berbasis data untuk berbagai sektor, proyek teknis yang akan menangani aspek-aspek prioritas teknis yang ditargetkan, dan proyek untuk mendorong dan mendukung kerja sama dan koordinasi yang efisien di seluruh kegiatan cPPP BDV.

Kesimpulan

Tujuan utama proyek BIG adalah untuk menentukan peta jalan big data Eropa yang mempertimbangkan aspek teknis, bisnis, kebijakan, dan masyarakat. Bab ini merinci hasil peta jalan lintas sektoral dan rencana aksi terkait.

Tujuan kedua dari proyek BIG adalah untuk membentuk inisiatif yang dipimpin oleh industri seputar manajemen informasi cerdas dan data besar untuk berkontribusi terhadap daya saing UE dan menempatkannya di Horizon 2020. Forum Swasta Publik Data Besar (Big Data Public Private Forum), sebagaimana awalnya disebut, dimaksudkan untuk untuk menciptakan jalan menuju implementasi peta jalan. Jalur ini memerlukan dua elemen utama: (1) mekanisme untuk mengubah peta jalan menjadi agenda nyata yang didukung oleh sumber daya yang diperlukan (investasi ekonomi dari pemangku kepentingan publik dan swasta) dan (2) komunitas yang berkomitmen untuk melakukan investasi dan berkolaborasi dalam implementasinya. dari agenda-agenda tersebut. Tujuan ini dicapai melalui kerja sama dengan platform teknologi NESSI dengan peluncuran Big Data Value Association (BDVA) dan kontrak Big Data Value Public Private Partnership (BDV cPPP) dalam kerangka

Horizon 2020. BDVA dan BDV cPPP memberikan kerangka kerja yang diperlukan yang menjamin kepemimpinan industri, investasi, dan komitmen pihak swasta dan publik untuk membangun ekonomi berbasis data di seluruh Eropa. Tujuan strategis dari cPPP BDV adalah untuk menguasai pembangkitan nilai dari data besar dan menciptakan keunggulan kompetitif yang signifikan bagi industri Eropa yang akan meningkatkan pertumbuhan ekonomi dan lapangan kerja.

DAFTAR PUSTAKA

A. (2013). Worldwide Big Data Technology and Services 2013–2017 Forecast (IDC #244979). IDC Mark Anal 34.

Accenture. (2012). Connected health: The drive to integrated healthcare delivery. Online: www. acccenture.com/connectedhealthstudy

Accenture. (2012). Connected health: The drive to integrated healthcare delivery. Online: www. acccenture.com/connectedhealthstudy

ACFE Association of Certified Fraud Examiners. (2014). Report to the nations on occupational fraud and abuse, Global fraud Study 2014. Available online at: http://www.acfe.com/rttn/docs/ 2014-report-to-nations.pdf

Acquisti, A., & Gross, R. (2009). Predicting social security numbers from public data. Proceedings of the National Academy of Sciences, 106(27), 10975–10980.

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. Harvard Business Review, 84, 98–107.

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. Harvard Business Review, 84, 98–107.

AlixPartners. (2014). AlixPartners car sharing outlook study. Retrieved from: http://www. alixpartners.com/en/MediaCenter/PressReleases/tabid/821/articleType/ArticleView/ articleId/ 950/AlixPartners-Study-Indicates-Greater-Negative-Effect-of-Car-Sharing-on-Vehicle-Pur chases.aspx and Trends in Theoretical Computer Science, 9(3–4), 211–407. doi:10.1561/0400000042.

Anderson, C. The end of theory. Wired, 16.07, 2008. Available at http://archive.wired.com/ science/discoveries/magazine/16-07/pb_theory

Anicic, D., Fodor, P., Rudolph, S., Stu¨hmer, R., Stojanovic, N., & Studer, R. (2011). ETALIS: Rule-based reasoning in event processing. In: Reasoning in event-based distributed systems (pp. 99–124). Studies in Computational Intelligence, vol. 347, Springer. Apache Spark. http://spark.apache.org/, (last retrieved April 2014).

Apache. (2014). Apache HBase Project Website. http://hbase.apache.org. Accessed Nov 21, 2014. Aranda, C. B., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., et al. (2013). SPARQL 1.1 overview. In W3C Recommendation.

article]. Available: http://www.computerworld.com/article/2502623/cloud-computing/shell-oil-targets-hybrid-cloud-as-fix-for-energy-saving--agile-it.html

Ashley, K. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference (pp. 722–735).

Bank of America et al. AMQP v1.0. (2011). Available online at http://www.amqp.org/sites/amqp. org/files/amqp.pdf

Barbieri, D. F., Braga, D., Ceri, S., Della Valle, E., & Grossniklaus, M. (2010). C-SPARQL: A continuous query language for RDF data streams. International Journal of Semantic Comput- ing, 4(1), 3–125.

Barnes, M. R., Harland, L., Foord, S. M., Hall, M. D., Dix, I., Thomas, S., et al. (2009). Lowering industry firewalls: Pre-competitive informatics initiatives in drug discovery. Nature Reviews Drug Discovery, 8(9), 701–708.

Barter, P. (2013, February 22). 'Cars are parked 95% of the time'. Let's check! [Online article]. Available: http://www.reinventingparking.org/2013/02/cars-are-parked-95-of-time-lets-check. html

BDVA. (2015). N. de Lama, J. Marguerite, K. D. Platte, J. Urban, S. Zillner, E. Curry (eds) European Big Data Value strategic research and innovation agenda. Big Data Value Association.

Becker, T., Jentzsch, A., & Palmetshofer, W. (2014). D2.5 Cross-sectorial roadmap consolidation. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Bennett, D., & Harvey, A. (2009). Publishing Open Government Data. W3C, Technical Report, 2009. Available online at: http://www.w3.org/TR/gov-data/

Berners-Lee, T. (2009). Linked data design issues. http://www.w3.org/DesignIssues/LinkedData. html

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., & Rodgers, J. R. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. Journal of Molecular Biology, 112(3), 535–542.

Bertolucci, J. IBM's Predictions: 6 Big Data Trends In 2014, December 2013. Available at http://www.informationweek.com/big-data/big-data-analytics/ibms-predictions-6-big-data-trends- in-2014-/d/d-id/1113118

Bhaduri, K., Das, K., Liu, K., Kargupta, H., & Ryan, J. (2011). Distributed data mining bibliography. http://www.cs.uinbc.edu/-hillol/DDM-BIB. BIG deliverable D2.4.1 – First draft of sector's roadmaps.

Bitkom. (Ed.). (2012). Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Available at http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online%281%29.pdf

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

Blaze, M., Bleumer, G., & Strauss, M. (2006). Divertible protocols and atomic proxy cryptography. In Proceedings of Eurocrypt (pp. 127–144).

Blueprints. (2014). Blueprints Project Homepage. https://github.com/tinkerpop/blueprints/wiki. Accessed Feb 4, 2015.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 1247–1250). New York, NY.

Bollier, D., & Firestone, C. M. (2010). The promise and peril of big data. Washington, DC: Aspen Institute, Communications and Society Program.

Bradic, A. (2011) S4: Distributed stream computing platform, Slides@Software Scalability Belgrad. Available online at: http://de.slideshare.net/alekbr/s4-stream-computing-platform Carzaniga, A., Rosenblum, D. S., & Wolf, A. L. (2000). Achieving scalability and expressiveness

Bretschneider, C., Zillner, S., & Hammon, M. (2013). Grammar-based lexicon enhancement for aligning German radiology text and images. In Proceedings of the Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria.

Brodie, M. L., & Liu, J. T. (2010). The power and limits of relational technology in the age of information ecosystems. On the Move Federated Conferences.

Buneman, P., Cheney, J., Tan, W., & Vansummeren, S. (2008). Curated databases. In Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Data- base Systems.

Buschbeck, S., Jameson, A., Spirescu, A., Schneeberger, T., Troncy, R., & Khrouf, H., et al. (2013). Parallel faceted browsing. In Extended Abstracts of CHI 2013, the Conference on Human Factors in Computing Systems (Interactivity Track).

C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of American Medical Informat- ics Association, 17(5), 507–513.

C. W., Schubmehl, D., Stolarski, K., Turner, M. J., Wardley, M., Webster, M., & Zaidi, Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolvsvold, A., McKelvie, S. et al. (2011). Windows azure storage: a highly available cloud storage service with strong consistency. In: Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (143–157). New York, NY: ACM.

Capgemini. (2012). World payments report 2012. Available at http://www.capgemini.com/resource-file-access/resource/pdf/The_8th_Annual_World_Payments_Report_2012.pdf. Accessed 2014.

Carbonell, J., & Jade, G. (1998). The use of MMR, diversity-based reranking for reodering documents and producing summaries. SIGIR (pp. 335–336). Melbourne, Australia: ACM. Cheng, G. T. (2011). RELIN: Relatedness and informativeness-based centrality for entity sum

Centers for Medicare and Medicaid Services. (2010). Medicare accountable care organizations – Shared savings program – New Section 1899 of Title XVIII, Preliminary questions and answers. Online retrieved January 10, 2010.

Chattopadhyay, B. (Google), Youtube Data Warehouse, Latest technologies behind Youtube, including Dremel and Tenzing, XLDB 2011, Stanford.

Chauhan, N. (2013). Modernizing machine-to-machine interactions: A platform for Igniting the Next Industrial Revolution, GE Software. Available at http://www.gesoftware.com/sites/ default/files/GE-Software-Modernizing-Machine-to-Machine-Interactions.pdf

Chen, H., Chiang, R.H.L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. In MIS Quarterly, 36(4), 1165–1188.

Cheney, J. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Cheng, G., Ge, W., & Qu, Y. (2008). Falcons: Searching and browsing entities on the semantic web. In Proceedings of the 17th International Conference on World Wide Web (pp. 1101–1102). Beijing, China: ACM.

Choo, C. W. (1996). The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. International Journal of Information Man- agement, 16, 329–340. doi:10.1016/0268-4012(96)00020-5.

Chu, C., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G., Ng, A. Y., et al. (2007). Map-reduce for machine learning on multicore. Advances in Neural Information Processing Systems, 19, 281.

CloudDrive. (2013). Advantages of cloud data storage. http://www.clouddrive.com.au/download/ www.clouddrive.com.au-WhitePaper.pdf. Accessed Nov 20, 2013.

Cloudera. (2012). Treato Customer Case Study. https://www.cloudera.com/content/dam/cloudera/ Resources/PDF/casestudy/Cloudera_Customer_Treato_Case_Study.pdf

Cloudera. (2013). Identifying fraud, managing risk and improving compliance in financial services.

Cloudera. (2014a). Cloudera Company Web page. www.cloudera.com. Accessed May 6, 2015. Cloudera. (2014b). Rethink data. http://www.cloudera.com/content/cloudera/en/new/. Accessed Feb 4, 2014. Communications of the ACM, 51(1), 107–113.

Correia, Z. P. (2004). Toward a stakeholder model for the co-production of the public-sector information system. Information Research, 10(3), paper 228. Retrieved February 27, 2013, from InformationR.net: http://InformationR.net/ir/10-3/paper228.html

Cragin, M., Heidorn, P., Palmer, C. L., & Smith, L. C. (2007). An educational program on data curation, ALA science & technology section conference. cross-references in Systems Biology. BMC Systems Biology, 1, 58.

CrowdFlower. (2012). Crowdsourcing: Utilizing the cloud-based workforce (Whitepaper). Curry, E., & Freitas, A. (2014). Coping with the long tail of data variety. Athens: European Data Forum.

Cugola, G., & Margara, A. (2012). Processing flows of information. ACM Computing Surveys, 44

Curry, E., Freitas, A., & O'Ria´in, S. (2010). The role of community-driven data curation for enterprise. In D. Wood (Ed.), Linking enterprise data (pp. 25–47). Boston, MA: Springer US.

Curry, E., Ngonga, A., Domingue, J., Freitas, A., Strohbach, M., Becker, T. et al. (2014). D2.2.2. Final version of the technical white paper. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Curry, E., O'Donnell, J., Corry, E., et al. (2013). Linking building data in the cloud: Integrating cross-domain building data using linked data. Advanced Engineering Informatics, 27, 206–219.

Cypher, A. (1993). Watch what i do: Programming by demonstration. Cambridge, MA: MIT Press.

D2.3.2. Final version of the sectorial requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

DeBrabant, J., Pavlo, A., Tu, S., Stonebraker, M., & Zdonik, S. (2013). Anti-caching: a new approach to database management system architecture. In Proceedings of the VLDB Endow- ment (pp 1942–1953).

Deloitte. (2014). Technology, media and telecommunications predictions 2014. Retrieved from http://www.deloitte.co.uk/tmtpredictions/assets/downloads/Deloitte-TMT-Predictions-2014.pdf

Department of Energy and Climate Change. (2012, December). Smart metering data access and privacy – Public attitudes research. [Whitepaper]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/43045/7227-sm-data-access-privacy-public-att.pdf Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations.

DHL Solutions & Innovation Trend Research. (2013). Big Data in Logistics. A DHL perspective on how to move beyond the hype. Available online at: http://www.delivering-tomorrow.com/ wp-content/uploads/2014/02/CSI_Studie_BIG_DATA_FINAL-ONLINE.pdf.

Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Official Journal L 345, 31/12/2003 P. 0090 - 0096. Brussels: The European Parliament and the Council of the European Union.

Doan, A., Ramakrishnan, R., & Halevy, A. (2011). Crowdsourcing systems on the world-wide web. Communications of the ACM, 54(4), 86–96.

Dorn, C., Taylor, R., & Dustdar, S. (2012). Flexible social workflows: Collaborations as human architecture. IEEE Internet Computing, 16(2), 72–77.

Eiben, C. B., et al. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nature Biotechnology, 30, 190–192.

El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. Journal of the American Medical Information Association, 15(5), 627–37.

El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., et al. (2014). De-identification methods for open health data: The case of the Heritage Health Prize claims dataset. Journal of Medical Internet Research, 14(1), e33.

enterprises. In D. Wood (Ed.), Linking enterprise data (pp. 25–47). Boston, MA: Springer US. DG Connect. (2013). A European strategy on the data value chain.

Erling, O. (2009). Virtuoso, a Hybrid RDBMS/Graph Column Store. In: Virtuoso Open-Source Wiki.http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSArticleVirtuosoAHybridRDBMSGraphColumnStore. Accessed Feb 6, 2015.

eTOM. (2014). TM forum. Retrieved from Business Process Framework: http://www.tmforum.org/BestPracticesStandards/BusinessProcessFramework/1647/Home.html

European Commission. (1998). COM(1998)585. Public sector information: A key resource for Europe. Green paper on public sector information in the information society. European Commission.

European Commission. (2010). Communication from the Commission: Europe 2020 – A European strategy for smart, sustainable and inclusive growth. COM 2020.

European Commission. (2013). Digital Agenda for Europe, Session Reports, ICT for Industrial Leadership: Innovating by exploiting big and open data and digital content.

European Commission. (2014). Towards a thriving data-driven economy, Communication from the commission to the European Parliament, the council, the European economic and social Committee and the committee of the regions, Brussels.

European Commission. (2014). Towards a thriving data-driven economy, Communication from the commission to the European Parliament, the council, the European economic and social Committee and the committee of the regions, Brussels.

Evans, P. C., Annunziata, M. (2012). Industrial internet: Pushing the boundaries of minds and machines, GE, November 26, 2012.

Fan, J. W., & Friedman, C. (2011). Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. Journal of Biomedical Informatics, 44 (5), 805–14.

Fensel, A., Fensel, D., Leiter, B., Thalhammer, A. (2012). Effective and efficient online communication: The channel model. In Proceedings of International Conference on Data Technolo- gies and Applications (DATA'12) (pp. 209–215), SciTePress, Rome, Italy, 25–27 July, 2012. Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., et al. (2008). Towards LarKC: A platform for web-scale reasoning. Los Alamitos, CA: IEEE

Flener, P., & Schmid, U. (2008). An introduction to inductive programming. Artificial Intelligence Review, 29, 45–62.

Flick, U. (2004). Triangulation in qualitative research. In U. Flick, E. V. Kardorff, & I. Steinke (Eds.), A companion to qualitative research (p. 432). London: Sage.

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., & Lintott, C., et al. (2011). Galaxy Zoo: Morphological classification and citizen science, machine learning and mining for astronomy. Chapman & Hall.

Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces: A new abstraction for information management. ACM SIGMOD Record, 34(4), 27–33.

Franz. (2015). Allegrograph product web page. http://franz.com/agraph/allegrograph/. Accessed Feb 6, 2015.

Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. In Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI), Haifa.

Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. In Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI), Haifa.

Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. 18th International Conference on Intelligent User Interfaces (pp. 279–288). Haifa, Israel: ACM.

Freitas, A., Carvalho, D., Pereira da Silva, J. C., O'Riain, S., & Curry, E. (2012a). A semantic best-effort approach for extracting structured discourse graphs from Wikipedia. In Proceedings of the 1st Workshop on the Web of Linked Entities (WoLE 2012) at the 11th International Semantic Web Conference (ISWC).

Freitas, A., Curry, E., Oliveira, J. G., & O'Riain, S. (2012b). Querying heterogeneous datasets on the linked data web: Challenges, approaches and trends. IEEE Internet Computing, 16(1), 24–33.

Freitas, A., Oliveira, J. G., O'Riain, S., Curry, E., & Pereira da Silva, J. C. (2011). Querying Linked data using semantic relatedness: A vocabulary independent approach. In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB).

Frias-Martinez, V., Rubio, A., & Frias-Martinez, E. (2012). Measuring the impact of epidemic alerts on human mobility (Vol. 12). Pervasive Urban Applications – PURBA.

Fuentes-Fernandez, R., Gomez-Sanz, J. J., & Pavon, J. (2012). User-oriented analysis of interactions in online social networks. IEEE Intelligent Systems, 27, 18–25.

Gabriel, G. (2012) Storm: The Hadoop of Realtime Stream Processing. PyConUs. Available online at http://pyvideo.org/video/675/storm-the-hadoop-of-realtime-stream-processing

Gil, Y., Szekely, P., Villamizar, S., Harmon, T. C., Ratnakar, V., Gupta, S., et al. (2011). Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows. In Proceedings of the 10th International Semantic Web Conference (ISWC).

Gislason, H. (2013). Datamarket.com. BIG Project Interviews Series. Halevy, A. (2013). Google. BIG Project Interviews Series.

Goble, C. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Good, B. M., & Su, A. I. (2011). Games with a scientific purpose. Genome Biology, 12(12), 135. Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. Infermation

Gossain, S., & Kandiah, G. (1998). Reinventing value: The new business ecosystem. Strategy and Leadership, 26, 28–33.

Government of Ireland. Department of Jobs, Enterprise and Innovation. (2013, June 24). Joint Industry/Government Task Force to drive development of Big Data in Ireland – Minister Bruton. Retrieved February 17, 2013, from Working for Jobs, Enterprise and Innovation: http://www.djei.ie/press/2013/20130624.htm

Gowing, W., & Nickson, J. (2010) Pseudonymisation Technical White Paper, NHS connecting for health, March 2010.

Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006). Attribute-based encryption for fine-grained access control of encrypted data. In Proceedings of the 13th ACM Conference on Computer and Communications Security (pp. 89–98).

Grady, J. (2013). Is enterprise cloud storage a good fit for your business? http://www.1cloudroad. com/is-enterprise-cloud-storage-a-good-fit-for-your-business

Groth, P. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The 'big data' revolution in healthcare, January 2013. Available at http://www.mckinsey.com/insights/health_systems/~/media/7764A72F70184C8EA88D805092D72D58.ashx

GTM Research. (December 2012). The soft grid 2013-2020: Big data and utility analytics for smart grid. [Online]. Available: www.sas.com/news/analysts/Soft_Grid_2013_2020_Big_Data_Utility_Analytics_Smart_Grid.pdf

Han, X., Sun, L., & Zhao, J. (2011). Collective entity linking in web text: A graph-based method. In Proceedings of the 34th International ACM SIGIR Conference on Research and Develop- ment in Information Retrieval.

Haque, U. (2013). Cosm. BIG Project Interviews Series. Hasan, S., & Curry, E. (2014). Approximate semantic matching of events for the internet of things.

Harizopoulos, S., Abadi, D. J., Madden, S., & Stonebraker, M. (2008). OLTP through the looking glass, and what we found there. In Proceedings of the 2008 ACM SIGMOD International Conference on Management data (SIGMOD '08).

Harper, D. (2012). GeoConnections and the Canadian Geospatial Data Infrastructure (CGDI): An SDI Success Story, Global Geospatial Conference.

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013). A discussion of privacy challenges in user profiling with big data techniques: The EEXCESS use case. IEEE Interna- tional Congress on Big Data (pp. 25–30).

Hasan, S., & Curry, E. (2014a). Approximate semantic matching of events for the internet of things. ACM Transactions on Internet Technology 14(1):1–23. doi:10.1145/2633684.

Hasan, S., & Curry, E. (2014b). Thematic event processing. In Proceedings of the 15th Interna- tional Middleware Conference on – Middleware '14, ACM Press, New York, NY, Hasan, S., & Curry, E. (2014b). Thematic event processing. In Proceedings of the 15th inter- national middleware conference on - middleware'14, ACM Press, New York, NY, pp 109–120. doi:10.1145/2663165.2663335.

Heath, T., & Bizer, C. (2011) Linked data: Evolving the web into a global data space (1st edn). In Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1): 1–136. Morgan & Claypool.

Hey, T., Tansley, S., & Tolle, K. M. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Heyde, C. O., Krebs, R., Ruhle, O., & Styczynski, Z. A. (2010). Dynamic voltage stability assessment using parallel computing. In Proceeding of: Power and energy society general meeting, 2010 IEEE.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., & Yon Rhee, S. (2008). Big data: The future of biocuration. Nature, 455(7209), 47–50.

IBM. (2013). Architecture of the IBM Big Data Platform. Available online at http://public.dhe. ibm.com/software/data/sw-library/big-data/ibm-bigdata-platform-19-04-2012.pdf

IBM. (2014). Infographic – The four V's of big data. Retrieved from http://www.ibmbigdatahub. com/enlarge-infographic/1642

ICO. (2012). Anonymisation: Managing data protection risk code of practice. Wilmslow: Infor- mation Commissioner's Office.

Ikeda, R., Park, H., & Widom, J. (2011). Provenance for generalized map and reduce workflows. in an internet-scale event notification service. In Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing, pp 219–27.

Independent Expert Advisory Group on Data Revolution. (2014). A world that counts – mobilizing the data revolution for sustainable development. New York.

Insight Centre for Data Analytics. (2015). Towards a Magna Carta for Data Insight Centre for Data Analytics White Paper.

Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads, The ACM Magazine for Students, 17(2), 16–21.

Jacobs, A. (2009). The pathologies of big data. Communications of the ACM, 52, 36–44. doi:10.1145/1536616.1536632.

Javed, W., Ghani, S., & Elmqvist, N. (2012). PolyZoom: Multiscale and multifocus explora-tion in 2D visual spaces. In Human factors in computing systems: CHI 2012 conference proceedings. New York: ACM.

Jobling, C. (2013, July 31). Capturing, processing, and transmitting video: Opportunities and challenges. Retrieved from Military embedded systems: http://mil-embedded.com/articles/ capturing-processing-transmitting-video-opportunities-challenges/

Kamara, S., & Lauter, K. (2010). Cryptographic cloud storage. In Financial Cryptography and Data Security (pp. 136–149).

Kart, L. (April 2013). Advancing analytics. Online Presentation, p. 6. Available: http://meetings2. informs.org/analytics2013/Advancing%20Analytics_LKart_INFORMS%20Exec%20Forum_ April%202013_final.pdf

Kaufmann, E., & Bernstein, A. (2007). How useful are natural language interfaces to the semantic web for casual end-users? In Proceedings of the 6th International The Semantic Web Confer- ence (pp. 281–294).

Keim, D., Kohlhammer, J., & Ellis, G. (eds.) (2010) Mastering the information age: solving problems with visual analytics. Eurographics Association.

Khatib, F., DiMaio, F., Foldit Contenders Group, Foldit Void Crushers Group, Cooper, S., Kazmierczyk, M. et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature Structural and Molecular Biology, 18, 1175–1177.

Kirrane, S., Abdelrahman, A., Mileo, S., & Decker, S. (2013). Secure manipulation of linked data.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. World Wide Web, 1(2), 19.

Knight, S. A., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. Informing Science, 8, 159–172.

Kobielus, J. (2013) Big Data needs data virtualization. In: InfoWorld Webpage. http://www.infoworld.com/article/2611579/big-data/big-data-needs-data-virtualization.html. Accessed Nov 18, 2014.

Kobielus, J. (2013). Measuring the business value of big data. Retrieved from http://www.ibmbigdatahub.com/blog/measuring-business-value-big-data

Komazec, S., Cerri, D., & Fensel, D. (2012). Sparkwave: Continuous schema-enhanced pattern matching over RDF data streams. In Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems (DEBS '12) (pp. 58–68). New York, NY: ACM. doi:10.1145/ 2335484.2335491.

Kong, N., Hanrahan, B., Weksteen, T., Convertino, G., & Chi, E. H. (2011). VisualWikiCurator: Human and machine intelligence for organizing wiki content. In Proceedings of the 16th International Conference on Intelligent User Interfaces (pp. 367–370).

Koomey, J. G. (2008). Worldwide electricity used in data centers. Environmental Research Letters, 3, 034008. doi:10.1088/1748-9326/3/3/034008.

Korster, P., & Seider, C. (2010). The world's 4 trillion dollar challenge. Executive Report of IBM Global Business Services, online.

Kraftwerke. (2014). http://www.next-kraftwerke.com/ Mearian, L. (2012, April 4). Shell oil targets hybrid cloud as fix for energy-saving, agile IT [Online

Krishnamurthy, K. (2013). Leveraging big data to revolutionize fraud detection, information week bank systems & technology. Available online at: http://www.banktech.com/leveraging-big- data-to-revolutionize-fraud-detection/a/d-id/1296473?

Kroes, N. (2013). Big data for Europe – ICT 2013 Event – Session on Innovating by exploiting big and open data and digital content, Vilnius.

Kusiak, A. (2009). Innovation: A data-driven approach. International Journal of Production Economics, 122(1), 440–448. doi:10.1016/j.ijpe.2009.06.025.

La Novere, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology, 23(12), 1509–1515.

Laibe, C., & Le Nove`re, N. (2007). MIRIAM resources: Tools to generate and resolve robust

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group.

Law, E., & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (vol. 4, pp. 1197–1206).

Law, E., & von Ahn, L. (2011). Human computation. Synthesis Lectures on Artificial Intelligence and Machine Learning, 5, 1–121.

Le, Q. V, Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., et al. (2011). Building high-level features using large-scale unsupervised learning. International Conference on Machine Learning.

Lee, C.-C., Chung, P.-S., & Hwang, M.-S. (2013). A survey on attribute-based encryption schemes of access. International Journal of Network Security, 15, 231–240.

Lee, K., Manning, S., Melton, J., Boyd, D., Grady, N., & Levin, O. (2014). Final SGBD report to JTC1, ISO/IEC JTC1 SGBD document no. N0095.

Li, M., Yu, S., Zheng, Y., Ren, K., Lou, W., et al. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. IEEE Transactions on Parallel and Distributed Systems, 24, 131–143.

Lieberman, H. (2001). Your wish is my command: Programming By example. San Francisco, CA: Morgan Kaufmann.

Lippell, H. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Liptchinsky, V., Khazankin, R., Truong, H., & Dustdar, S. (2012). A novel approach to mod eling context-aware and social collaboration processes. In: Advanced Information Systems Engi- neering, (pp. 565–580), Springer.

Lo, S. (2012). Big data facts and figures, November 2012. Available at http://blogs.sap.com/ innovation/big-data/big-data-facts-figures-02218

Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., & Hellerstein, J. M. (2012). Distributed Graph-Lab: A framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment, 5(8), 716–727.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J. M. (2010). GraphLab: A new framework for parallel machine learning. The 26th Conference on Uncer- tainty in Artificial Intelligence (UAI 2010), Catalina Island, California, July 8–11.

Lu¨nendonk GmbH. (2013). Trendpapier 2013: Big Data bei Krankenversicherungen. Bewa¨ltigung der Datenmengen in einem vera¨nderten Gesundheitswesen. Online.

Luckham, D. (2002). The power of events: An introduction to complex event processing in distributed enterprise systems. Boston, MA: Addison-Wesley Longman Publishing Co.

Lunzer, A., & Hornbæk, K. (2008). Subjunctive interfaces: Extending applications to support parallel setup, viewing and control of alternative scenarios. ACM Transactions on Computer- Human Interaction, 14(4), 17.

Lynch, N. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Ma Ching-To Albert (1994). Health care payment systems: Cost and quality incentives. Journal of Economics and Management Strategy, 3(1), Spring

Madsen, K. (2012) Storm: Comparison-introduction-concepts, slides, March. Available online at: http://de.slideshare.net/KasperMadsen/storm-12024820

Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., et al. (2010). Pregel: a system for large-scale graph processing. In Proceedings of the 2010 ACM SIGMOD Inter- national Conference on Management of data, ACM (pp. 135–146).

Manyika, J. et al. (2011). Big data: The next frontier for innovation, competition, and productivity.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, p. 156.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, p. 156.

MapR. (2014). MapR Company Website. https://www.mapr.com/. Accessed Feb 6, 2014. Margaret Rouse. (2014a). Object Storage. http://searchstorage.techtarget.com/definition/object-

Marz, N., & Warren, J. (2014). A new paradigm for Big Data. In N. Marz & J. Warren (Eds.), Big Data: Principles and best practices of scalable real-time data systems. Shelter Island, NY: Manning Publications.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The management revolution. Harvard Business Review, 90(10), 60–66. Available online at http://automotivedigest.com/wp-content/uploads/ 2013/01/BigDataR1210Cf2.pdf.

McKinsey & Company. Markl, V., Hoeren, T., & Krcmar, H. (2013). Innovationspotenzialanalyse fu¨r die neuen Technologien fu¨r das Verwalten und Analysieren von großen Datenmengen, November 2013.

McKinsey Company. (2011). Big data: The next frontier for innovation, competition, and productivity, online.

McKinsey Global Institute. (2011, June). Big Data: The next frontier for innovation, competition, and productivity. McKinsey & Company.

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In Proceedings of the 18th International Conference Data Engineering. IEEE Computer Society (pp. 117–128).

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. (pp. 522–536). The Semantic Web - ISWC.

Miller, S. P., Neuman, B. C., Schiller, J. I., & Saltzer, J. H. (1987, December 21). Section E.2.1: Kerberos authentication and authorization system. MIT Project Athena, Cambridge, MA.

Mons, B., & Velterop, J. (2009). Nano-Publication in the e-science era, International Semantic Web Conference.

Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., et al. (2008). Calling on a million minds for community annotation in WikiProteins. Genome Biology, 9(5), R89.

Moore, J. F. (1996). The death of competition: Leadership and strategy in the age of business ecosystems. New York: HarperCollins.

Moore, J. F. (2006). Business ecosystems and the view from the firm. Antitrust Bulletin, 51, 31–75. NESSI. (2012). Big data: A new world of opportunities. NESSI White Paper.

Mora, A. C., Chen, Y., Fuchs, A., Lane, A., Lu, R., Manadhata, P. et al. (2012). Top Ten Big Data Security and Privacy Challenges. Cloud Security Alliance (CSA).

Morris, H. D., & Vesset, D. (2005). Managing Master Data for Business Performance Management: The Issues and Hyperion's Solution, Technical Report.

Neumeyer, L. (2011). Apache S4: A distributed stream computing platform, Slides Stanford Infolab, Nov. Available online at: http://de.slideshare.net/leoneu/20111104-s4-overview

Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2011). S4: Distributed stream computing platform, KDCloud. Available online at: http://www.4lunas.org/pub/2010-s4.pdf

Nicholson, R.(2012). Big data in the oil and gas industry. IDC energy insights. Presentation. Retrieved from https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/RICK%20-%20IDC_ Calgary_Big_Data_Oil_and-Gas/$file/RICK%20-%20IDC_Calgary_ Big_Data_Oil_and-Gas.pdf Nunez, C. (2012, December 12). Who's watching? Privacy concerns persist as smart meters roll out [Online article]. Available: http://news.nationalgeographic.com/news/energy/2012/12/

Norris, R. P. (2007). How to make the dream come true: The astronomers' data manifesto. Data Science Journal, 6, S116–S124.

Novacek, V., Handschuh, S., & Decker S., (2011). Getting the meaning right: A complementary distributional layer for the web semantics. International Semantic Web Conference (1):504–519.

Oberkampf, H., Zillner, S., Bauer, B., & Hammon, M. (2013). An OGMS-based Model for Clinical Information (MCI). In Proceedings of International Conference on Biomedical Ontology, Montreal, Canada.

Ojo, A., Curry, E., & Sanaz-Ahmadi, F. (2015). A tale of open data innovations in five smart cities. In 48th Annual Hawaii international conference on system sciences (HICSS-48) (pp. 2326–2335). IEEE. doi:10.1109/HICSS.2015.280.

Ojo, A., Curry, E., & Sanaz-Ahmadi, F. (2015). A tale of open data innovations in five smart cities.

Okman, L., Gal-Oz, N., Gonen, Y., et al. (2011). Security issues in nosql databases. In 2011 I.E. 10th International Conference on Trust, Security and Privacy in Computing and Communica- tions (TrustCom) (pp. 541–547).

Oppenheim, A. N. (1992). Questionnaire design, interviewing and attitude measurement. London: Continuum.

Ovum. (2013). Retrieved from http://ovum.com/press_releases/ovum-predicts-turbulence-for-the- internet-economy-as-more-than-two-thirds-of-consumers-say-no-to-internet-tracking/

Palmer, C. L., et al. (2013). Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data.

Paradigm 4. (2014, July 1). Leaving data on the table: New survey shows variety, not volume, is the bigger challenge of analyzing big data. Survey. Available: http://www.paradigm4.com/wp- content/uploads/2014/06/P4PR07012014.pdf

Parliament, E. (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at http://eur-lex.europa.eu/legal-content/en/ ALL/?uri¼CELEX:31995L0046. Accessed 2014.

Parliament, E. (2013). Proposal for a Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union, from European Commission. Available at

http://ec.europa.eu/information_
society/newsroom/cf/dae/document.cfm?doc_id¼1666. Accessed 2014.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In Proceedings of the 25th National Conference on Artificial Intelligence (AAAI).

PEC. (2014). Peer Energy Cloud Project Website. http://www.peerenergycloud.de/. Accessed Feb 6, 2015.

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. Library and Archives Journal, 1, 1–3.

Piccoli, G. (2012). Information systems for managers – text and cases (2nd ed.). Wiley. Rayport, J. F., & Sviokla, J. J. (1995). Exploiting the virtual value chain. Harvard Business

PredPol. (n.d.). Retrieved September 08, 2013, from PredPol Web site: http://www.predpol.com/ The European Parliament and the Council of The European Union. (2003, November 17).

PricewaterhouseCoopers. (2009). Transforming healthcare through secondary use of health data. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute,

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. http:// altmetrics.org/manifesto/

PWC. (2014). Global entertainment and media outlook 2014-2018 – key industry themes. Retrieved from http://www.pwc.com/gx/en/global-entertainment-media-outlook/insights-and- analysis.jhtml

PWC. (2014). Image sensor: Steady growth for new capabilities. Retrieved from PWC: http://www.pwc.com/gx/en/technology/mobile-innovation/image-sensor-steady-growth-new- capabilities.jhtml

Qin, L., & Atluri, V. (2003). Concept-level access control for the Semantic Web. In Proceedings of the ACM Workshop on XML Security – XMLSEC '03. ACM Press.

Rodrıguez-Doncel, V., Gomez-Perez, A., & Mihindukulasooriya, N. (2013). Rights declaration in Linked Data. In Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, 2013.

Rowe, M., Angeletou, S., & Alani, H. (2011). Predicting discussions on the social semantic web (pp. 405–420). The Semantic Web: Research and Applications.

Rowe, N. (2012). The state of master data management, building the foundation for a better enterprise. Aberdeen Group.

Ryutov, T., Kichkaylo, T., & Neches, R. (2009). Access control policies for semantic networks. In 2009 I.E. International Symposium on Policies for Distributed Systems and Networks (pp. 150–157).

Schutz, A., & Buitelaar, P. (2005). RelExt: A tool for relation extraction from text in ontology extension. In Proceedings of the 4th International Semantic Web Conference.

Sensmeier, L. (2013). How Big Data is revolutionizing Fraud Dedection in Financial Services. Hortonworks Blog. Available online at: http://hortonworks.com/blog/how-big-data-is-revolu tionizing-fraud-detection-in-financial-services/Vivisimo. (2012). Big Data White Paper. Services and Use, 30, 1–2. 51–56.

Sewash, J. (2014). Data curation interview. Expert interview series for the EU-project BIG (318062; ICT-2011.4.4). http://big-project.eu/text-interviews

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., et al. (2012). Linked open government data: Lessons from Data.gov.uk. IEEE Intelligent Systems, 27(3), Spring Issue, 16–24.

Shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (pp. 1–8). New York: ACM.

Shenker, S., Stoica, I., Zaharia, M., & Xin, R. (2013). Shark: SQL and rich analytics at scale. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 13–24).

Sherif, A., Sohan, R., & Hopper, A. (2013). HadoopProv: Towards provenance as a first class citizen in MapReduce. In Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance.

Sheth, A. (1999). Changing focus on interoperability in information systems: From System, Syntax, Structure to Semantics. Interoperating Geographic Information Systems The Springer International Series in Engineering and Computer Science (vol. 495, pp. 5–29).

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of Visual Languages.

Shucheng, Y., Wang, C., Ren, K., & Lou, W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. INFOCOM (pp. 1–9).

Shvachko, K. H. K., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In IEEE 26th Symposium on Mass Storage Systems and Technologies (pp. 1–10).

Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. Journal on Data Semantics, IV, 146–171.

Singhal, A. (2012). Introducing the knowledge graph. Retrieved from googleblog: http:// googleblog.blogspot.com/2012/05/introducing-knowledge-graphthings

Soderland, N., Kent, J., Lawyer, P., & Larsson, S. (2012). Progress towards value-based health care. Lessons from 12 countries. The Boston Consulting Group, online.

Spence, R. (2006). Information visualization – design for interaction (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Stavrakantonakis, I. (2013a). Personal data and user modelling in tourism. In: Information and Communication Technologies in Tourism 2013 (pp 507–518).

Stavrakantonakis, I. (2013b). Semantically assisted Workflow Patterns for the Social Web. In Proceedings of the 10th Extended Semantic Web Conference ESWC 2013 PhD Symposium track. (pp. 692–696).

Stonebraker, M. (2012). What does 'big data' mean. Communications of the ACM, BLOG@ ACM. Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. Ecology, 16, 284–307.

Stonebraker, M. (2012a) What Does "Big Data" Mean? In: BLOG@ACM. http://cacm.acm.org/ blogs/blog-cacm/155468-what-does-big-data-mean/fulltext. Accessed Feb 5, 2015.

Stonebraker, M. (2012b). What Does "Big Data" Mean? (Part 2). In: BLOG@ACM. http://cacm. acm.org/blogs/blog-cacm/156102-what-does-big-data-mean-part-2/fulltext. Accessed Apr 25, 2013.

Strohbach, M., Ziekow, H., Gazis, V., & Akiva, N. (2011). Towards a big data analytics framework for IoT and smart city applications. In F. Xhafa & P. Papajorgji (Eds.), Modeling and processing for next generation big data technologies with applications (pp. 257–282). Berlin: Springer.

Surowiecki, J. (2005). The wisdom of crowds. New York: Random House LLC. Technopolis Group. (2011). Data centres: Their use, value and impact (JISC Report).

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570.

Tabbitt, S. (2014, 17 February). Big data analytics keeps Dublin moving [Online article]. Available: http://www.telegraph.co.uk/sponsored/sport/rugby-trytracker/10630406/ibm-big-data- analytics-dublin.html

Tanner, A. (2014, July 11). The wonder (and woes) of encrypted cloud storage [Online article]. Available: http://www.forbes.com/sites/adamtanner/2014/07/11/the-wonder-and-woes-of- encrypted-cloud-storage/

Technavio. (2013). Global big data market in the financial services sector 2012-2016. Retrieved from http://www.technavio.com. Accessed 2014.

Thajchayapong, S., & Barria, J. A. (2010). Anomaly detection using microscopic traffic variables on freeway segments. Transportation Research Board of the National Academies, 10-2393.

Thalhammer, A., Knuth, M., & Sack, H. (2012a). Evaluating entity summarization using a game-based ground truth. International Semantic Web Conference (2) (pp. 350–361). Boston: Springer.

Thalhammer, A., Toma, I., Roa-Valverde, A. J., Fensel, D. (2012b). Leveraging usage data for linked data movie entity summarization. In Proceedings of the 2nd International Workshop on Usage Analysis and the Web of Data. Lyon, France: USEWOD co-located with WWW.

The White House. (2012, March 29). Big Data is a Big Deal. Retrieved January 18, 2013, from The White House: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

Thomson Reuters Technical Report, ORCID: The importance of proper identification and attribu- tion across the scientific literature ecosystem. (2013).

Thusoo, A., Sarma, J., Jain, N., Shao, Z., Chakka, P., Anthony, S., et al. (2009). Hive – a warehousing solution over a map-reduce framework. Statistics and Operations Research Transactions, 2, 1626–1629.

Tori, A. (2013). BIG Project Interviews Series. Turney, P. D., & Pantel, P. (2010). From frequency to meaning: vector space models of semantics.

Tuchinda, R., Knoblock, C. A., & Szekely, P. (2011). Building Mashups by demonstration. ACM Transactions on the Web (TWEB), 5(3), Art. 16.

Tuchinda, R., Szekely, P., & Knoblock, C. A. (2007). Building data integration queries by demonstration. In Proceedings of the International Conference on Intelligent User Interface. Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers.

Turner, V., Gantz, J. F., Reinsel, D., & Minton, S. (2014). The digital universe of opportunities: rich data and the increasing value of the internet of things. Rep. from IDC EMC.

Turner, V., Gantz, J. F., Reinsel, D., & Minton, S. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. Framingham, MA: International Data Corporation (IDC). http://idcdocserv.com/1678. Accessed Aug 19, 2015.

U.S. Government. (2002). Sarbanes-Oxley Act of 2002. Available at 107th Congress Public Law

Van Kasteren, T., Ulrich, B., Srinivasan, V., & Niessen, M. (2014). Analyzing tweets to aid situational awareness. In 36th European Conference on Information Retrieval.

van Kasteren, T., Ulrich, B., Srinivasan, V., & Niessen, M. (2014). Analyzing tweets to aid situational awareness. 36th European Conference on Information Retrieval.

Venkatesh, P., & Nirmala, S. (2012). NewSQL – The new way to handle big data. In: Blog. http:// www.opensourceforu.com/2012/01/newsql-handle-big-data/. Accessed Nov 18, 2014.

Voisard, A., & Ziekow, H. (2011). ARCHITECT: A layered framework for classifying technol- ogies of event-based systems. Information Systems, 36(6), 937–957. doi:10.1016/j.is.2011.03. 006.

VoltDB. (2014). VoltDB Company Website. http://www.voltdb.com. Accessed Nov 21, 2014. Webber, J. (2013). BIG Interview by John Dominque.

Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the twitter stream. Advances in Information Retrieval, 356–367.

Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Detecting patterns of crime with series finder. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Ward, M., Grinstein, G. G., & Keim, D. (2010). Interactive data visualization: Foundations, techniques, and applications. Natick, MA: Taylor & Francis.

White, T. (2012). Hadoop: The Definitive Guide. O'Reilly.

Wieland, M., Kopp, O., Nicklas, D., & Leymann, F (2007). Towards context-aware workflows. In: CAiSE. pp. 11–15.

Wikipedia. (2013). Column-oriented DBMS. http://en.wikipedia.org/wiki/Column-oriented_DBMS. Accessed Apr 25, 2013.

Winder, D. (2012). Securing NoSQL applications: Best practises for big data security. Computer Weekly.

World Economic Forum. (2012). Big Data, Big Impact: New possibilities for international development. Geneva: The World Economic Forum.

Wu, C., & Guo, Y. (2013). Enhanced user data privacy with pay-by-data model. In Proceeding of 2013 I.E. International Conference on Big Data (pp. 53–57).

Yin, R. K. (2009). Case study research: Design and methods. In L. Bickman & D. J. Rog (Eds.), Essential guide to qualitative methods in organization research. London: Sage. doi:10.1097/ FCH.0b013e31822dda9e.

Yiu, C. (2012). The Big Data Opportunity. Making government faster, smarter and more personal.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark : Cluster computing with working sets. In HotCloud'10 Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (p. 10).

Ziekow, H., Goebel, C., Struker, J., & Jacobsen, H.-A. (2013). The potential of smart home sensors in forecasting household electricity demand. In Smart Grid Communications (SmartGridComm), 2013 I.E. International Conferences (pp. 229–234).

Zillner, S., Bretschneider, C., Oberkampf, H., Neurerer, S., Munne´, R., Lippell, H. et al. (2014). D2.4.2 Final version of sectors roadmap. Public deliverable of the EU-project BIG (318062; ICT-2011.4.4).

Zillner, S., Bretschneider, C., Oberkampf, H., Neurerer, S., Munne´, R., Lippell, H., et al. (2014).

Zillner, S., Lasierra, N., Faix, W., & Neururer, S. (2014a). User needs and requirements analysis for big data healthcare applications. In Proceeding of the 25th European medical informatics conference (MIE 2014), Istanbul, Turkey, September 2014.

Zillner, S., Neurerer, S., Munne´, R., Lippell, H., Vilela, L., Prieto, E. et al. (2013). D2.4.1 first draft

Zillner, S., Neurerer, S., Munne´, R., Lippell, H., Vilela, L., Prieto, E. et al. (2014a). D2.3.2. Final version of the sectorial requisites. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Zillner, S., Neurerer, S., Munne´, R., Lippell, H., Vilela, L., Prieto, E. et al. (2014b). D2.4.2 Final version of sectors roadmap. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Zillner, S., Neururer, S., Munne, R., Prieto, E., Strohbach, M., van Kasteren, T., et al. (2014). D2.3.2. Final Version of the Sectorial Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

TEORI EKONOMI BERBASIS BIG DATA

Dr. Agus Wibowo, M.Kom, M.Si, MM

BIO DATA PENULIS

Penulis memiliki berbagai disiplin ilmu yang diperoleh dari Universitas Diponegoro (UNDIP) Semarang. dan dari Universitas Kristen Satya Wacana (UKSW) Salatiga. Disiplin ilmu itu antara lain teknik elektro, komputer, manajemen dan ilmu sosiologi. Penulis memiliki pengalaman kerja pada industri elektronik dan sertifikasi keahlian dalam bidang Jaringan Internet, Telekomunikasi, Artificial Intelligence, Internet Of Things (IoT), Augmented Reality (AR), Technopreneurship, Internet Marketing dan bidang pengolahan dan analisa data (komputer statistik).

Penulis adalah pendiri dari Universitas Sains dan Teknologi Komputer (Universitas STEKOM) dan juga seorang dosen yang memiliki Jabatan Fungsional Akademik Lektor Kepala (Associate Professor) yang telah menghasilkan puluhan Buku Ajar ber ISBN, HAKI dari beberapa karya cipta dan Hak Paten pada produk IPTEK. Penulis juga terlibat dalam berbagai organisasi profesi dan industri yang terkait dengan dunia usaha dan industri, khususnya dalam pengembangan sumber daya manusia yang unggul untuk memenuhi kebutuhan dunia kerja secara nyata.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
JL. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8120-76-5 (PDF)
9 786238 120765

Dr. Agus Wibowo, M.Kom, M.Si, MM

TEORI EKONOMI BERBASIS BIG DATA

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
JL. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id