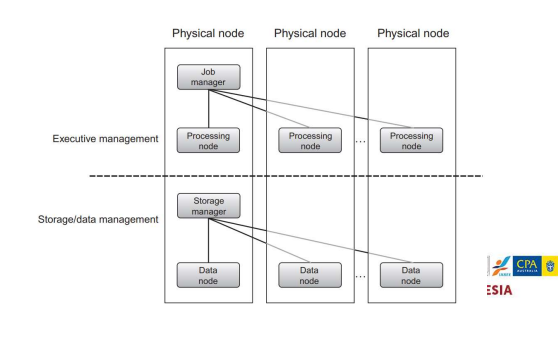
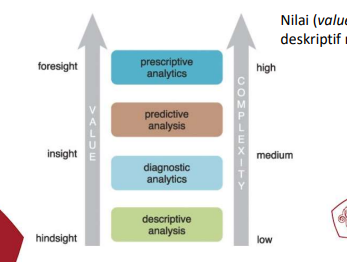
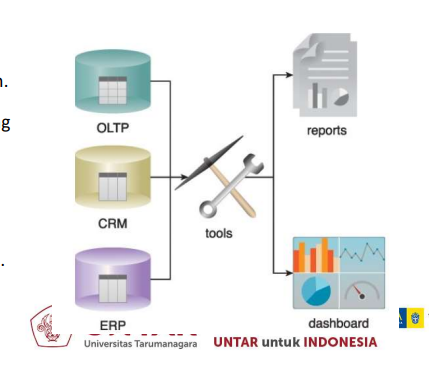
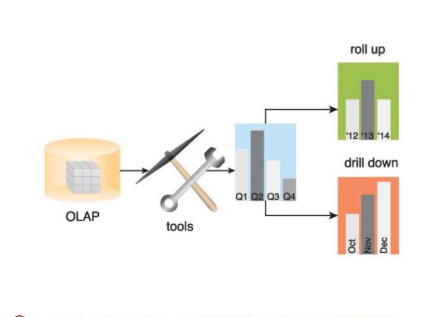
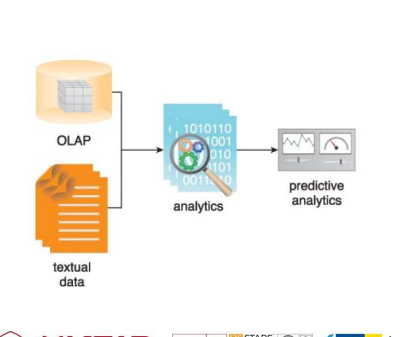
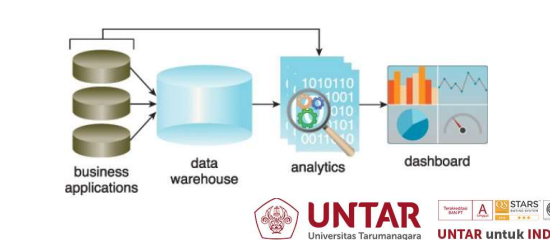
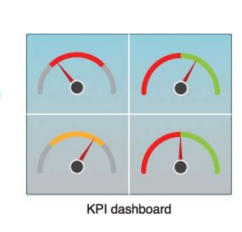
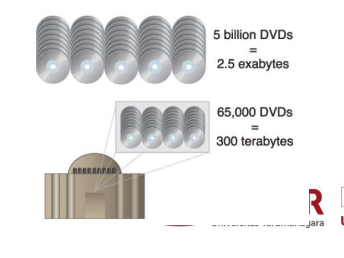
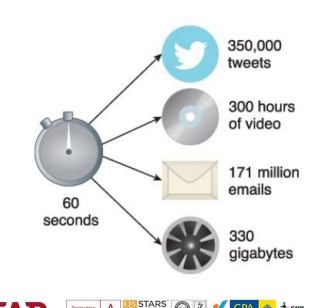
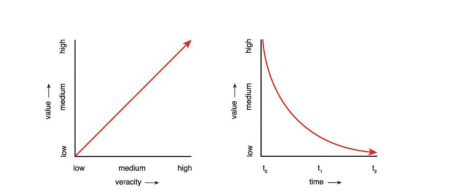
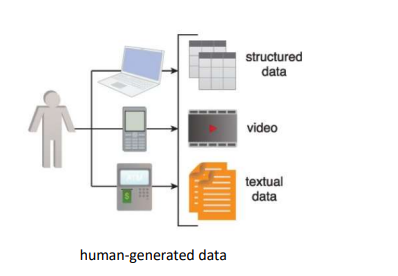
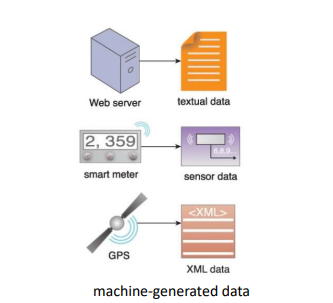
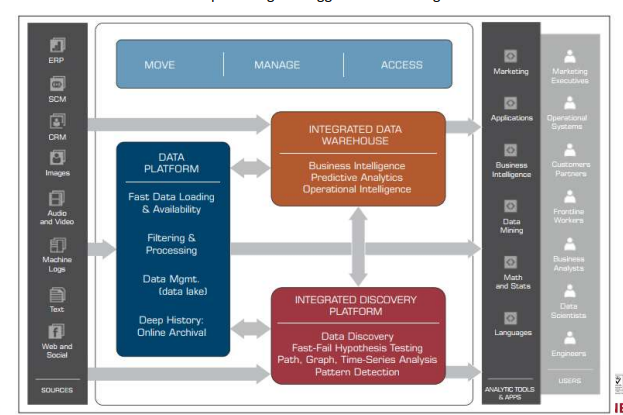
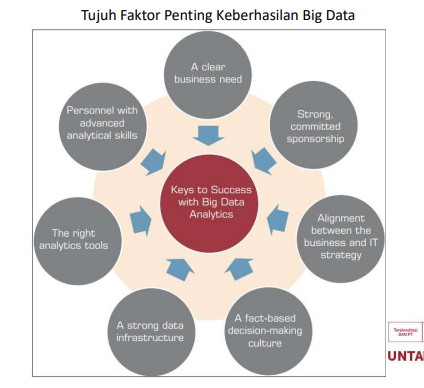
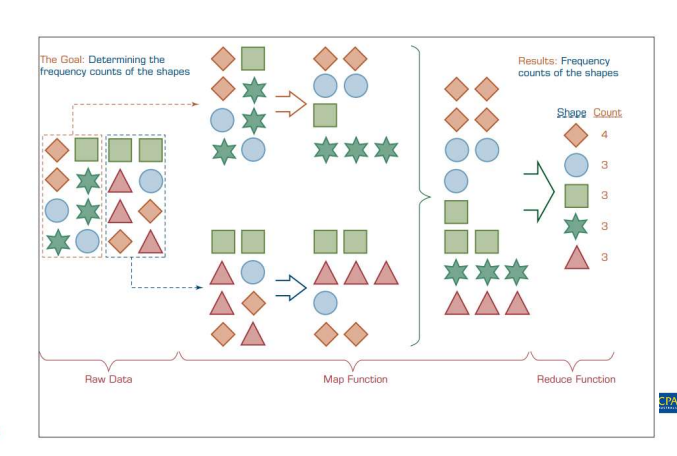
Bab 01 – Pengenalan Big Data

* Big Data adalah bidang yang didedikasikan untuk analisis, pemrosesan, dan penyimpanan kumpulan data besar yang sering kali berasal dari sumber yang berbeda.
* “Big data adalah aset informasi bervolume tinggi (high-volume), berkecepatan tinggi (high-velocity), dan beragam (high-variety) yang menuntut bentuk pemrosesan informasi yang hemat biaya dan inovatif untuk meningkatkan wawasan dan pengambilan keputusan.” – Gartner
* “Sementara perusahaan berjuang untuk mengkonsolidasikan sistem dan meruntuhkan database yang berlebihan untuk memungkinkan konsistensi operasional, analitis, dan kolaboratif yang lebih besar, perubahan kondisi ekonomi telah membuat pekerjaan ini lebih sulit. E-commerce, khususnya, telah meledakkan tantangan manajemen data dalam tiga dimensi: volume, kecepatan, dan variasi. Pada tahun 2001/02, organisasi TI banyak menyusun berbagai pendekatan yang mereka miliki untuk menangani masing-masing.” – Doug Lane
* Apa yang membuat Big Data dibutuhkan?
* 
* Sumber Daya Komputasi Utama untuk Big Data
  + Kemampuan pemrosesan: CPU, prosesor, atau node
  + Memory
  + Storage
  + Jaringan
  + 
* Konsep dan Terminologi: Datasets
  + Kumpulan atau kelompok data terkait umumnya disebut sebagai kumpulan data.
  + Setiap kelompok atau anggota dataset (datum) berbagi set atribut atau properti yang sama dengan yang lain dikumpulan data yang sama.
  + Beberapa contoh kumpulan data adalah:
    - ⎫tweet disimpan dalam file
    - ⎫kumpulan file gambar dalam direktori
    - ⎫ekstrak baris dari tabel database yang disimpan dalam file berformat CSV
    - ⎫pengamatan cuaca historis yang disimpan sebagai file XML
  + Datasets dapat ditemukan dalam berbagai format.
  + Data analysis adalah proses pemeriksaan data untuk menemukan fakta, hubungan, pola,wawasan dan/atau tren.
  + Tujuan keseluruhan dari analisis data adalah untuk mendukung pengambilan keputusan yang lebih baik.
  + Contoh analisis data sederhana adalah analisis data penjualan es krim untuk mengetahui hubungan jumlah es krim yang terjual dengan suhu harian.
  + Hasil analisis semacam itu akan mendukung keputusan yang terkait dengan berapa banyak es krim di toko harus disediakan sehubungan dengan informasi ramalan cuaca. • Melakukan analisis data membantu membangun pola dan hubungan di antara data yang dianalisis
  + Data analytics adalah istilah yang lebih luas yang mencakup analisis data.
  + Data analytics adalah disiplin yang mencakup pengelolaan siklus hidup data yang lengkap, yang meliputi pengumpulan, pembersihan, pengorganisasian, penyimpanan, analisis, dan pengaturan data.
  + Istilah ini mencakup pengembangan metode analisis, teknik ilmiah, dan otomatisasi peralatan.
  + Di lingkungan Big Data, data analytics telah mengembangkan metode yang memungkinkan data analisis terjadi melalui penggunaan teknologi terdistribusi yang sangat skalabel dan kerangka kerja yang mampu menganalisis data dalam jumlah besar dari berbagai sumber Konsep dan Terminologi: Data Analytics
  + Siklus hidup analitik Big Data umumnya melibatkan identifikasi, pengadaan, persiapan, dan menganalisis sejumlah besar data mentah dan tidak terstruktur untuk mengekstrak informasi bermakna yang dapat berfungsi sebagai masukan untuk mengidentifikasi pola, memperkaya data perusahaan yang ada dan melakukan pencarian skala besar
  + Berbagai jenis organisasi menggunakan alat dan teknik analisis data dengan cara yang berbeda.
  + Contoh:
    - ⎫Dalam lingkungan berorientasi bisnis, hasil analisis data dapat menurunkan biaya operasional dan memfasilitasi pengambilan keputusan strategis
    - ⎫Dalam domain ilmiah, analitik data dapat membantu mengidentifikasi penyebab suatu fenomena untuk meningkatkan akurasi prediksi.
    - ⎫Dalam lingkungan berbasis layanan seperti organisasi sektor publik, analisis data dapat membantu memperkuat fokus dalam memberikan layanan berkualitas tinggi dengan menurunkan biaya.
  + Data analytics memungkinkan pengambilan keputusan berdasarkan data dengan dukungan ilmiah sehingga keputusan dapat didasarkan pada data faktual dan bukan hanya pada pengalaman atau intuisi masa lalu.
  + Ada empat kategori umum analitik yang dibedakan berdasarkan hasil yang diperoleh:
    - ⎫descriptive analytics
    - ⎫diagnostic analytics
    - ⎫predictive analytics
    - ⎫prescriptive analytics
  + Jenis analitik yang berbeda memanfaatkan teknik dan algoritma analisis yang berbeda.
  + Ini menyiratkan bahwa mungkin ada berbagai data, penyimpanan, dan persyaratan pemrosesan untuk memfasilitasi penyampaian berbagai jenis hasil analitik.
  + Gambar di bawah ini menggambarkan kenyataan bahwa generasi hasil analitik bernilai tinggi meningkatkan kompleksitas dan biaya analitik lingkungan. Nilai (value) dan kompleksitas (complexity) meningkat dari analitik deskriptif menjadi preskriptif
  + 
* Descriptive Analytics
  + Analisis deskriptif dilakukan untuk menjawab pertanyaan tentang peristiwa yang telah terjadi.
  + Bentuk analitik ini mengontekstualisasikan data untuk menghasilkan informasi.
  + Contoh pertanyaan dapat mencakup:
    - ⎫Berapa volume penjualan selama 12 bulan terakhir?
    - ⎫Berapa jumlah “telepon darurat” yang diterima yang dikategorikan berdasarkan tingkat keparahan dan lokasi geografis?
    - ⎫Berapa komisi bulanan yang diperoleh setiap agen penjualan?
  + Diperkirakan bahwa 80% dari hasil analitik yang dihasilkan bersifat deskriptif.
  + Dari segi nilai, analitik deskriptif memberikan nilai yang paling rendah dan membutuhkan keahlian yang relatif mendasar.
  + Sistem operasional, gambar kiri, ditanyakan melalui analitik deskriptif alat untuk menghasilkan laporan atau dasbor, gambar kanan.
  + 
  + Laporan umumnya bersifat statis dan menampilkan data historis yang disajikan dalam bentuk grid data atau grafik.
  + Kueri dieksekusi pada data operasional toko dari dalam suatu perusahaan, misalnya sistem Customer Relationship Management (CRM) atau sistem Enterprise Resource Planning (ERP).
  + Online transaction processing (OLTP).
* Analisis Diagnostik (Diagnostic Analytics)
  + Analisis diagnostik bertujuan untuk menentukan penyebab dari suatu fenomena yang terjadi di masa lalu menggunakan pertanyaan yang berfokus pada alasan di balik peristiwa tersebut.
  + Tujuannya adalah untuk menentukan informasi apa yang terkait dengan fenomena untuk memungkinkan menjawab pertanyaan yang berusaha untuk menentukan mengapa sesuatu telah terjadi.
  + Pertanyaan tersebut meliputi:
    - ⎫Mengapa penjualan Q2 kurang dari penjualan Q1?
    - ⎫Mengapa ada lebih banyak “telepon darurat” yang berasal dari wilayah Timur dari pada dari wilayah barat?
    - ⎫Mengapa ada peningkatan tingkat penerimaan kembali pasien selama tiga bulan terakhir?
  + Analitik diagnostik memberikan nilai lebih dari pada analitik deskriptif tetapi membutuhkan lebih banyak keahlian tingkat lanjut.
  + Analitik diagnostik biasanya memerlukan pengumpulan data dari beberapa sumber dan menyimpannya dalam struktur yang cocok untuk melakukan drill-down dan rollup analisis Analisis Diagnostik (Diagnostic Analytics)
  + Analisis drill-down berarti mendapatkan data rinci tentang ringkasan informasi yang menekankan pada hal tertentu.
  + Analisis roll-up untuk mendapatkan ringkasan.
  + 
  + Hasil analitik diagnostik dilihat melalui interaktif alat visualisasi yang memungkinkan pengguna untuk mengidentifikasi tren dan pola.
  + Kueri yang dieksekusi lebih kompleks dibandingkan dengan analitik deskriptif dan dilakukan pada data multidimensi yang disimpan dalam sistem pemrosesan analitik.
* Predictive Analytics
  + Analisis prediktif dilakukan dalam upaya untuk menentukan hasil dari suatu peristiwa yang mungkin terjadi di masa depan.
  + Dengan analitik prediktif, informasi ditingkatkan dengan makna untuk menghasilkan pengetahuan yang menyampaikan bagaimana informasi itu terkait.
  + Kekuatan dan besarnya asosiasi membentuk dasar model yang digunakan untuk menghasilkan masa depan prediksi berdasarkan peristiwa masa lalu.
  + Penting untuk dipahami bahwa model yang digunakan untuk analitik prediktif memiliki ketergantungan implisit pada kondisi di mana masa lalu peristiwa terjadi.
  + Jika kondisi yang mendasari ini berubah, maka model yang membuat prediksi perlu diperbarui.
  + Pertanyaan biasanya dirumuskan dengan menggunakan what-if, seperti berikut:
    - ⎫Berapa peluang seorang nasabah akan gagal membayar pinjaman jika mereka melewatkan pembayaran bulanan?
    - ⎫Berapa tingkat kelangsungan hidup pasien jika Obat B diberikan sebagai pengganti Obat A?
    - ⎫Jika seorang pelanggan telah membeli Produk A dan B, berapa peluang mereka akan membeli Produk C?
  + Analitik prediktif mencoba memprediksi hasil peristiwa, dan prediksi dibuat berdasarkan pada pola, tren, dan pengecualian yang ditemukan dalam data historis dan terkini.
  + 
  + Hal ini dapat menyebabkan identifikasi resiko dan peluang.
  + Analisis semacam ini melibatkan penggunaan kumpulan data besar yang terdiri dari data internal dan eksternal dan berbagai teknik analisis data.
  + Hal ini memberikan nilai yang lebih besar dan membutuhkan lebih banyak keahlian lanjutan dari pada analitik deskriptif dan diagnostik.
  + Alat analitik prediktif dapat menyediakan antarmuka front-end yang ramah pengguna.
* Prescriptive Analytics
  + Analitik preskriptif dibangun berdasarkan hasil analitik prediktif dengan menentukan tindakan yang harus diambil.
  + Fokusnya tidak hanya pada pilihan mana yang terbaik untuk diikuti, tapi kenapa.
  + Dengan kata lain, analitik preskriptif memberikan hasil yang dapat dinalar karena mereka menanamkan unsur pemahaman situasional.
  + Dengan demikian, analitik semacam ini dapat digunakan untuk mendapatkan keuntungan atau mengurangi risiko.
  + Analitik preskriptif memberikan nilai lebih dari pada jenis analitik lainnya dan karenanya memerlukan keahlian paling canggih, serta perangkat lunak dan peralatan khusus.
  + Pendekatan bergeser dari penjelasan ke nasihat dan dapat mencakup simulasi berbagai skenario.
  + Contoh pertanyaan dapat mencakup:
    - Di antara tiga obat, mana yang memberikan hasil terbaik?
    - Kapan waktu terbaik untuk memperdagangkan saham tertentu? Prescriptive Analytics
  + Analisis semacam ini menggabungkan data internal dengan data eksternal.
  + Data internal termasuk data penjualan saat ini dan data historis, informasi pelanggan, data produk, dan aturan bisnis.
  + Data eksternal dapat mencakup data media sosial, prakiraan cuaca, dan data demografi yang dihasilkan pemerintah.
  + Analitik preskriptif melibatkan penggunaan aturan bisnis dan sejumlah besar data internal dan eksternal untuk mensimulasikan hasil dan meresepkan tindakan yang terbaik.
  + Analitik preskriptif melibatkan penggunaan aturan bisnis dan internal dan/atau data eksternal untuk melakukan analisis mendalam
* Business Intelligence (BI)
* 
  + BI memungkinkan organisasi untuk mendapatkan wawasan tentang kinerja perusahaan dengan menganalisis data yang dihasilkan oleh proses bisnis dan sistem informasinya.
  + Hasil analisis dapat digunakan oleh manajemen untuk mengarahkan bisnis dalam upaya memperbaiki masalah yang terdeteksi atau meningkatkan kinerja organisasi.
  + BI menerapkan analitik ke sejumlah besar data di seluruh perusahaan, yang biasanya telah dikonsolidasikan ke dalam gudang data perusahaan untuk menjalankan kueri analitik. Business Intelligence (BI)
  + Keluaran BI dapat ditampilkan ke dasbor yang memungkinkan manajer mengakses dan menganalisis hasilnya dan berpotensi menyaring kueri analitik untuk mengeksplorasi data lebih lanjut.
* Key Performance Indicators (KPI)
* 
  + KPI adalah metrik yang dapat digunakan untuk mengukur kesuksesan dalam konteks bisnis tertentu.
  + KPI terkait dengan tujuan dan sasaran strategis perusahaan secara keseluruhan.
  + KPI sering digunakan untuk mengidentifikasi masalah kinerja bisnis dan menunjukkan kepatuhan terhadap peraturan.
  + Oleh karena itu, KPI bertindak sebagai titik referensi yang dapat diukur untuk mengukur aspek tertentu dari suatu kinerja keseluruhan bisnis.
  + Dasbor menggabungkan tampilan beberapa KPI dan membandingkan pengukuran aktual dengan nilai ambang batas yang menentukan rentang nilai yang dapat diterima dari KPI.
* Karakteristik Big Data
  + 
  + Karakteristik Big Data: Volume
  + 
    - Volume data yang tinggi memaksakan permintaan penyimpanan dan pemrosesan data yang berbeda, serta proses persiapan, kurasi, dan pengelolaan data tambahan.
    - Sumber data yang menghasilkan volume data yang tinggi dapat mencakup:
      * ⎫transaksi online, seperti point-of-sale dan perbankan
      * ⎫eksperimen ilmiah dan penelitian
      * ⎫sensor, seperti sensor GPS, RFID, pengukur pintar, dan telematika
      * ⎫media sosial, seperti Facebook dan Twitter
  + Karakteristik Big Data: Velocity
  + 
    - Di lingkungan Big Data, data dapat tiba dengan kecepatan tinggi, dan kumpulan data yang sangat besar dapat terakumulasi dalam waktu yang sangat singkat.
    - Dari sudut pandang perusahaan, kecepatan data diterjemahkan ke dalam jumlah waktu yang dibutuhkan untuk data diproses sekali itu memasuki perimeter perusahaan.
    - Mengatasi arus masuk data yang cepat perusahaan harus untuk merancang solusi pemrosesan data yang sangat elastis dan tersedia dan kemampuan penyimpanan data yang sesuai.
    - Contoh kumpulan data Big Data berkecepatan tinggi yang dihasilkan setiap menit termasuk tweet, video, email, dan GB yang dihasilkan dari mesin jet.
  + Karakteristik Big Data: Variety
  + 
    - Keragaman data mengacu pada berbagai format dan jenis data yang perlu didukung oleh solusi Big Data.
    - Keragaman data membawa tantangan bagi perusahaan dalam hal data integrasi, transformasi, pemrosesan, dan penyimpanan.
    - Contoh kumpulan data Big Data yang sangat beragam termasuk data terstruktur, tekstual, gambar, video, audio, XML, JSON, data sensor dan metadata.
  + Karakteristik Big Data: Veracity
    - Veracity mengacu pada kualitas data.
    - Data yang masuk ke lingkungan Big Data perlu dinilai kualitasnya, yang dapat menyebabkan kegiatan pemrosesan data mengatasi masalah data yang tidak valid dan menghilangkan noise.
    - Sehubungan dengan kebenaran, data dapat menjadi bagian dari sinyal atau noise dari kumpulan data.
    - Noise adalah data yang tidak dapat diubah menjadi informasi dan karenanya tidak memiliki nilai, sedangkan signals memiliki nilai dan mengarah pada informasi yang bermakna.
    - Data dengan rasio signal-to-noise tinggi memiliki lebih banyak kebenaran daripada data dengan rasio yang lebih rendah.
    - Data yang didapat secara terkontrol, misalnya melalui pendaftaran pelanggan online, biasanya berisi lebih sedikit noise daripada data yang diperoleh melalui sumber yang tidak terkontrol, seperti posting blog.
    - Jadi rasio signal-to-noise data tergantung pada sumber data dan jenisnya.
  + Karakteristik Big Data: Value
  + 
    - Value didefinisikan sebagai kegunaan data untuk suatu perusahaan.
    - Sifat nilainya adalah secara intuitif terkait dengan karakteristik kejujuran karena semakin tinggi fidelitas data, semakin banyak nilai yang dimiliki untuk bisnis.
    - Value tergantung pada berapa lama pemrosesan data berlangsung karena hasil analitik memiliki masa simpan; misalnya, penawaran saham tertunda 20 menit memiliki sedikit atau tidak ada nilai untuk melakukan perdagangan dibandingkan dengan penawaran yang berumur 20 milidetik.
    - Seperti yang ditunjukkan, nilai dan waktu berbanding terbalik.
    - Semakin lama waktu yang dibutuhkan untuk data berubah menjadi informasi yang berarti, semakin sedikit nilainya untuk bisnis.
    - Data yang memiliki veracity yang tinggi dan dapat dianalisis dengan cepat memiliki nilai lebih untuk bisnis.
    - Terlepas dari veracity dan waktu, nilai juga dipengaruhi oleh siklus hidup berikut yang terkait kekhawatiran:
      * ⎫Seberapa baik data telah disimpan?
      * ⎫Apakah atribut berharga dari data dihapus selama pembersihan data?
      * ⎫Apakah jenis pertanyaan yang tepat diajukan selama analisis data?
      * ⎫Apakah hasil analisis dikomunikasikan secara akurat kepada pihak pembuat keputusan yang tepat
* Berbagai Jenis Data
  + Data yang diproses oleh solusi Big Data dapat dibuat oleh manusia atau dibuat oleh mesin, meskipun pada akhirnya merupakan tanggung jawab mesin untuk menghasilkan hasil analitik.
  + Data yang dihasilkan manusia adalah hasil interaksi manusia dengan sistem, seperti layanan online dan perangkat digital.
  + Data yang dihasilkan mesin dihasilkan oleh program perangkat lunak dan perangkat keras direspon terhadap peristiwa dunia nyata.
  + Misalnya, file log menangkap keputusan otorisasi dibuat oleh layanan keamanan, dan sistem point-of-sale menghasilkan transaksi terhadap persediaan untuk mencerminkan barang yang dibeli oleh pelanggan.
  + Dari perspektif perangkat keras, dan contoh data yang dihasilkan mesin adalah informasi yang disampaikan dari banyak sensor di ponsel yang mungkin melaporkan informasi, termasuk posisi dan menara seluler kekuatan sinyal.
  + human-generated data machine-generated data. Contoh data yang dihasilkan manusia termasuk media sosial, posting blog, email, berbagi foto, dan perpesanan.
  + Contoh data yang dihasilkan mesin termasuk log web, data sensor,data telemetri, data smart meter dan data penggunaan alat.
  + Data yang dibuat oleh manusia dan yang dibuat oleh mesin dapat berasal dari berbagai sumber dan direpresentasikan dalam berbagai format atau jenis.
  + Jenis data utama adalah:
    - ⎫structured data
    - ⎫unstructured data
    - ⎫semi-structured data
    - Structured Data
      * Data terstruktur (structured data) sesuai dengan model atau skema data dan sering disimpan dalam bentuk tabel.
      * Data terstruktur (structured data) digunakan untuk menangkap hubungan antara entitas yang berbeda dan paling sering disimpan dalam database relasional.
      * Data terstruktur (structured data) sering dibuat oleh perusahaan aplikasi dan sistem informasi seperti ERP dan sistem CRM.
      * Karena alat melimpah dan database yang secara asli mendukung data terstruktur, jarang memerlukan pertimbangan dalam hal pengolahan atau penyimpanan.
      * Contoh dari jenis data ini meliputi: transaksi perbankan, faktur, dan catatan pelanggan.
    - Unstructured Data
      * Data yang tidak sesuai dengan model data atau skema data dikenal sebagai data tidak terstruktur (unstructured data).
      * Diperkirakan bahwa data tidak terstruktur membentuk 80% dari data yang diberikan perusahaan.
      * Data tidak terstruktur memiliki tingkat pertumbuhan yang lebih cepat daripada data terstruktur.
      * Bentuk data ini adalah tekstual atau biner (binary files) dan sering disampaikan melalui file yang mandiri dan non-relasional.
      * Sebuah file teks mungkin berisi konten berbagai tweet atau posting blog.
      * File biner sering menjadi media file yang berisi data gambar, audio atau video.
      * Secara teknis, file teks dan biner memiliki: struktur ditentukan oleh format file itu sendiri, tetapi aspek ini diabaikan, dan gagasan tentang menjadi tidak terstruktur dalam kaitannya dengan format data yang terkandung dalam file itu sendiri.
      * Logika khusus biasanya diperlukan untuk memproses dan menyimpan data yang tidak terstruktur.
      * Misalnya, untuk memutar file video, penting tersedia codec (decoder-coder).
      * Data tidak terstruktur tidak dapat langsung diproses atau di-query menggunakan SQL.
      * Jika diperlukan untuk disimpan dalam database relasional, data disimpan dalam tabel sebagai Binary Large Object (BLOB).
      * Not-only SQL (NoSQL) adalah database non-relasional yang dapat digunakan untuk menyimpan data tidak terstruktur bersama data terstruktur.
    - Semi-structured Data
    - Data semi-terstruktur (semi-structured data) memiliki tingkat struktur dan konsistensi yang ditentukan, tetapi tidak relasional.
    - Sebaliknya, data semi-terstruktur bersifat hierarkis atau berbasis grafik.
    - Data semacam ini biasanya disimpan dalam file yang berisi teks.
    - XML dan file JSON adalah bentuk umum dari data semi terstruktur.
    - Karena sifat tekstual ini data dan kesesuaiannya dengan beberapa tingkat struktur, lebih mudah diproses daripada data tidak terstruktur. Semi-structured Data
    - Contoh sumber umum dari data semi-terstruktur meliputi: electronic data interchange (EDI) files, spreadsheets, RSS feeds and sensor data.
    - Data semi-terstruktur sering memiliki persyaratan pra-pemrosesan dan penyimpanan khusus, terutama jika format dasarnya tidak berbasis teks.
    - Contoh pra-pemrosesan data semi-terstruktur adalah validasi dari file XML untuk memastikan bahwa file sesuai dengan definisi skemanya.
  + Metadata
    - Metadata memberikan informasi tentang karakteristik dan struktur kumpulan data.
    - Jenis data ini sebagian besar dihasilkan oleh mesin dan dapat ditambahkan ke data.
    - Pelacakan metadata sangat penting untuk pemrosesan, penyimpanan, dan analisis Big Data karena menyediakan informasi tentang silsilah data dan asalnya selama pemrosesan.
    - Contoh metadata meliputi:
      * ⎫tag XML memberikan penulis dan tanggal pembuatan dokumen.
      * ⎫atribut yang menyediakan ukuran file dan resolusi foto digital.
  + Latar Belakang Studi Kasus
    - Ensure to Insure (ETI) adalah perusahaan asuransi terkemuka yang menyediakan berbagai paket asuransi di sektor kesehatan, bangunan, kelautan dan penerbangan kepada 25 juta pelanggannya yang tersebar di seluruh dunia.
    - Lingkungan TI ETI terdiri dari kombinasi platform client-server dan mainframe yang mendukung pelaksanaan sejumlah sistem, termasuk kutipan kebijakan, administrasi kebijakan, manajemen klaim, penilaian risiko, manajemen dokumen, penagihan, enterprise resource planning (ERP) dan customer relationship management (CRM).
  + Hambatan Bisnis •
    - Selama beberapa dekade terakhir, profitabilitas perusahaan telah menurun.
    - Temuan komite mengungkapkan bahwa alasan utama di balik memburuknya posisi keuangan perusahaan adalah meningkatnya jumlah klaim penipuan dan pembayaran terkait yang dilakukan terhadap mereka.
    - Contoh Studi Kasus
    - ETI memilih Big Data untuk implementasi tujuan strategisnya, tetapi ETI tidak memiliki keterampilan Big Data internal dan perlu memilih antara menyewa konsultan Big Data atau mengirim tim TI-nya ke pelatihan Big Data.
    - Perusahaan memutuskan tim TI senior yang dikirim ke pelatihan untuk mengantisipasi solusi jangka panjang yang hemat biaya di mana anggota tim yang terlatih akan menjadi sumber daya Big Data internal permanen yang dapat dikonsultasikan kapan saja dan juga dapat melatih tim junior anggota untuk lebih meningkatkan keahlian Big Data internal.
    - Setelah menerima pelatihan Big Data, anggota tim yang terlatih menekankan perlunya istilah kosakata umum sehingga seluruh tim berada di halaman yang sama ketika berbicara tentang Big Data.
    - Mereka mengadopsi example-driven approach.
    - Saat membahas kumpulan data, beberapa kumpulan data terkait yang ditunjukkan oleh anggota tim mencakup klaim, kebijakan, kutipan, data profil pelanggan, dan data sensus.
    - Meskipun konsep data analysis dan data analytics cepat dipahami, beberapa anggota tim yang tidak memiliki banyak eksposur bisnis mengalami kesulitan memahami BI dan penetapan KPI yang sesuai.
    - Salah satu anggota tim TI yang terlatih menjelaskan BI dengan menggunakan proses pembuatan laporan bulanan untuk mengevaluasi kinerja bulan sebelumnya sebagai contoh.
    - Proses ini melibatkan mengimpor data dari sistem operasional ke EDW (Enterprise Data Warehouse) dan menghasilkan KPI seperti: polis yang dijual dan klaim yang diajukan, diproses, diterima, dan ditolak yang ditampilkan di dasbor dan kartu skor yang berbeda.
    - Dalam hal analitik, ETI menggunakan analitik deskriptif dan diagnostik.
    - Analisis deskriptif meliputi query sistem administrasi polis untuk menentukan jumlah polis yang terjual setiap hari, query sistem manajemen klaim untuk mengetahui berapa banyak klaim yang diajukan setiap hari dan query sistem penagihan untuk mengetahui berapa banyak pelanggan yang terlambat membayar premi.
    - Diagnostic analytics dilakukan sebagai bagian dari berbagai aktivitas BI, seperti melakukan query untuk menjawab pertanyaan seperti mengapa target penjualan bulan lalu tidak tercapai.
    - Ini termasuk melakukan operasi penelusuran untuk mengelompokkan penjualan menurut jenis dan lokasi sehingga dapat ditentukan lokasi mana yang berkinerja buruk untuk jenis kebijakan tertentu.
  + Contoh Studi Kasus: Mengidentifikasi Karakteristik Data
    - Volume
      * Sejumlah besar data transaksi dihasilkan sebagai hasil dari pemrosesan klaim, penjualan polis baru, dan perubahan polis yang ada.
      * Volume data tidak terstruktur yang besar, baik di dalam maupun di luar perusahaan, mungkin terbukti membantu dalam mencapai tujuan ETI.
      * Data ini meliputi catatan kesehatan, dokumen yang diserahkan oleh pelanggan pada saat mengajukan aplikasi asuransi, jadwal properti, data media sosial dan data cuaca.
    - Velocity
      * Data tergolong low velocity, seperti data pengajuan klaim dan data polis baru yang diterbitkan.
      * Data log server web dan kutipan asuransi adalah data berkecepatan tinggi.
      * Data media sosial dan data cuaca tiba dengan cepat.
      * Manajemen bencana dan deteksi klaim penipuan, data perlu diproses dengan cukup cepat untuk meminimalkan kerugian.
    - Variety
      * Kumpulan data yang mencakup catatan kesehatan, data kebijakan, data klaim, data kutipan, data media sosial, catatan agen pusat panggilan, catatan penyesuaian klaim, foto insiden, laporan cuaca, sensus data, log server web, dan email
    - Veracity
      * Sampel data yang diambil dari sistem operasional dan EDW menunjukkan tanda-tanda kebenaran yang tinggi (high veracity).
      * Validasi data dilakukan pada beberapa tahap termasuk validasi pada saat entri data, validasi di berbagai titik saat aplikasi memproses data, seperti validasi input tingkat fungsi, dan validasi yang dilakukan oleh database saat data disimpan.
      * Data yang diperoleh dari sosial media dan data cuaca memerlukan proses validasi dan pembersihan data.
    - Value
      * Perlu menentukan nilai maksimum dari kumpulan data yang tersedia dengan memastikan kumpulan data disimpan dalam bentuk aslinya dan sesuai dengan jenis analitik yang tepat.
    - Data terstruktur (structured data): data polis, data klaim, data profil pelanggan dan data penawaran.
    - Data tidak terstruktur (unstructured data): data media sosial, dokumen aplikasi asuransi, panggilan catatan agen pusat, catatan penyesuai klaim, dan foto-foto insiden.
    - Data semi terstruktur (semi-structured data): catatan kesehatan, data profil pelanggan, laporan cuaca,data sensus, log server web, dan email

Bab 2 – Konsep dan Alat Big Data

* Definisi Big Data
  + Karena semakin banyak data tersedia dalam berbagai bentuk dan mode, pemrosesan data yang tepat waktu dengan cara tradisional menjadi tidak praktis.
  + Fenomena ini disebut Big Data.
  + Big Data menerima liputan pers yang substansial dan menarik minat baik dari pengguna bisnis maupun profesional IT.
  + Big Data menjadi kata kunci pemasaran yang berlebihan dan terlalu sering digunakan.
  + Big Data memiliki arti yang berbeda bagi orang-orang dengan latar belakang dan minat yang berbeda.
  + Big Data dan analitik memicu cara baru untuk mengubah proses, organisasi, seluruh industri, dan bahkan masyarakat secara bersamaan.
  + Untuk sebagian besar bisnis, istilah "Big" relatif tergantung pada ukuran organisasi.
  + Intinya lebih pada menemukan nilai baru di dalam dan di luar sumber data konvensional.
  + Dari mana Big Data berasal?
  + Jawaban sederhananya adalah “di mana-mana”.
  + Big Data dapat berasal dari berbagai sumber:
    - ⎫ log Web
    - ⎫ radio-frequency identification(RFID)
    - ⎫ global positioning systems (GPS)
    - ⎫ jaringan sensor
    - ⎫ jaringan sosial
    - ⎫ dokumen teks berbasis Internet
    - ⎫ indeks pencarian Internet
    - ⎫ catatan panggilan detail
    - ⎫ astronomi
    - ⎫ ilmu atmosfer
    - ⎫ biologi
    - ⎫ genomik
    - ⎫ fisika nuklir
    - ⎫ eksperimen biokimia
    - ⎫ catatan medis
    - ⎫ penelitian ilmiah
    - ⎫ pengawasan militer
    - ⎫ arsip fotografi
    - ⎫ arsip video
    - ⎫ praktik e-commerce skala besar
  + Big Data bukanlah hal baru.
  + Yang baru adalah definisi dan struktur Big Data terus berubah. •
  + Perusahaan telah menyimpan dan menganalisis data dalam jumlah besar sejak munculnya data warehouse di awal 1990-an. •
  + Dulu terabyte identik dengan Big Data, sekarang menjadi exabyte, dan laju pertumbuhan volume data terus meningkat ketika organisasi berusaha untuk menyimpan dan menganalisis tingkat detail transaksi yang lebih besar, serta data yang dihasilkan oleh Web dan mesin, untuk mendapatkan pemahaman yang lebih baik tentang perilaku pelanggan dan penggerak bisnis. •
  + Big Data is not just “big”.
* "V" yang Mendefinisikan Big Data
* Big Data biasanya didefinisikan oleh tiga "V":
  + ⎫volume
  + ⎫variety
  + ⎫velocity
* VOLUME
  + Volume jelas merupakan sifat paling umum dari Big Data. •
  + Banyak faktor yang berkontribusi terhadap peningkatan eksponensial dalam volume data, seperti data berbasis transaksi yang disimpan selama bertahun-tahun, data teks yang terus mengalir dari media sosial, peningkatan jumlah data sensor yang dikumpulkan, data RFID dan GPS yang dihasilkan secara otomatis, dan sebagainya. •
  + Seperti disebutkan sebelumnya, “big” adalah istilah yang relative, berubah dari waktu ke waktu dan dirasakan secara berbeda oleh organisasi yang berbeda. •
  + Massa data tertinggi yang dulu disebut petabyte (PB) telah meninggalkan tempatnya menjadi zettabytes (ZB), yaitu satu triliun gigabyte (GB) atau satu miliar terabyte (TB).
* VARIETY •
  + Data saat ini hadir dalam semua jenis format—mulai dari database tradisional hingga penyimpanan data hierarkis yang dibuat oleh pengguna akhir dan sistem OLAP (Online analytical processing) hingga dokumen teks, email, XML, data yang dikumpulkan meter dan ditangkap sensor, hingga video, audio, dan data transaksi saham. •
  + Diperkirakan, 80%-85% dari semua data organisasi berada dalam semacam format tidak terstruktur atau semi terstruktur (format yang tidak cocok untuk skema basis data tradisional). •
  + Tetapi tidak dapat disangkal nilainya, maka harus dimasukkan dalam analisis untuk mendukung pengambilan keputusan.
* VELOCITY •
  + Menurut Gartner, velocity (kecepatan) berarti seberapa cepat data diproduksi dan seberapa cepat data harus diproses (yaitu, ditangkap, disimpan, dan dianalisis) untuk memenuhi kebutuhan atau permintaan. •
  + Tag RFID, sensor otomatis, perangkat GPS, dan pengukur pintar mendorong peningkatan kebutuhan untuk menangani aliran data yang hampir realtime. •
  + Velocity mungkin merupakan karakteristik Big Data yang paling diabaikan. •
  + Bereaksi cukup cepat untuk menghadapi velocity (kecepatan) merupakan tantangan bagi sebagian besar organisasi. •
  + Untuk lingkungan yang sensitif terhadap waktu, jam biaya data mulai berdetak saat data dibuat. •
  + Seiring berjalannya waktu, proposisi nilai data menurun dan akhirnya menjadi tidak berharga.
* VERACITY •
  + Veracity adalah istilah yang diciptakan oleh IBM yang digunakan sebagai "V" keempat untuk menggambarkan Big Data. •
  + Ini mengacu pada kesesuaian dengan fakta: akurasi, kualitas, kebenaran, atau kepercayaan data. •
  + Alat dan teknik sering digunakan untuk menangani kebenaran Big Data dengan mengubah data menjadi wawasan yang berkualitas dan dapat dipercaya.
* VARIABILITY •
  + Selain kecepatan dan variasi data yang meningkat, aliran data bisa sangat tidak konsisten dengan puncak periodik. •
  + Apakah sesuatu yang besar sedang tren di media sosial? •
  + Pemuatan data puncak harian, musiman, dan dipicu oleh peristiwa bisa sangat bervariasi dan dengan demikian menantang untuk dikelola—terutama dengan keterlibatan media sosial.
* VALUE PROPOSITION •
  + Kegembiraan seputar Big Data adalah proposisi nilainya.
  + Gagasan yang terbentuk sebelumnya tentang "Big" data adalah bahwa data tersebut berisi atau memiliki potensi yang lebih besar untuk memuat lebih banyak pola dan anomali yang menarik dari pada “small“ data. •
  + Jadi, dengan menganalisis data yang besar dan kaya fitur, organisasi dapat memperoleh nilai bisnis yang lebih besar yang mungkin tidak mereka miliki sebaliknya. •
  + Meskipun pengguna dapat mendeteksi pola dalam kumpulan data kecil menggunakan metode statistik dan pembelajaran mesin sederhana atau alat kueri dan pelaporan ad hoc, Big Data berarti "big“ analitik. •
  + Analisis besar berarti wawasan yang lebih besar dan keputusan yang lebih baik, sesuatu yang dibutuhkan setiap organisasi.
  + Arsitektur Konseptual Tingkat Tinggi untuk Solusi Big Data
    - 
* Fundamentals of Big Data Analytics, Tujuh Faktor Penting Keberhasilan Big Data
  + 
  + Kebutuhan bisnis yang jelas (keselarasan dengan visi dan strategi) •
    - Investasi bisnis harus dilakukan untuk kebaikan bisnis, bukan demi kemajuan teknologi belaka. •
    - Oleh karena itu, pendorong utama untuk analitik Big Data haruslah kebutuhan bisnis, di tingkat manapun - strategis, taktis, dan operasi.
  + Sponsor yang kuat dan berkomitmen •
    - Ini adalah fakta yang terkenal bahwa jika Anda tidak memiliki sponsor eksekutif yang kuat dan berkomitmen, sulit (bukan tidak mungkin) untuk berhasil. •
    - Jika ruang lingkupnya adalah satu atau beberapa aplikasi analitis, sponsorship dapat berada di tingkat departemen. •
    - Jika targetnya adalah transformasi organisasi di seluruh perusahaan, yang sering terjadi pada inisiatif Big Data, sponsor harus berada di tingkat tertinggi dan di seluruh organisasi.
  + Keselarasan antara bisnis dan strategi TI •
    - Sangat penting untuk memastikan bahwa pekerjaan analitik selalu mendukung strategi bisnis, dan bukan sebaliknya. •
    - Analytics harus memainkan peran yang memungkinkan dalam menjalankan strategi bisnis dengan sukses.
  + Budaya pengambilan keputusan berbasis fakta •
    - Dalam budaya pengambilan keputusan berbasis fakta, angka-angka intuisi, firasat, atau pengandaian mendorong pengambilan keputusan. •
    - Ada juga budaya eksperimen untuk melihat apa yang berhasil dan apa yang tidak. •
    - Untuk menciptakan budaya pengambilan keputusan berbasis fakta, manajemen senior perlu:
      * ⎫Mengetahui bahwa beberapa orang tidak dapat atau tidak mau menyesuaikan
      * ⎫Menjadi pendukung vocal
      * ⎫Menekankan bahwa metode yang sudah ketinggalan zaman harus dihentikan
      * ⎫Meminta untuk melihat analitik apa yang digunakan untuk mengambil keputusan
      * ⎫Menghubungkan insentif dan kompensasi dengan perilaku yang diinginkan
  + Infrastruktur data yang kuat •
    - Data warehouse telah menyediakan infrastruktur data untuk analitik. • Infrastruktur ini berubah dan ditingkatkan di era Big Data dengan teknologi baru.
    - Sukses butuhkan menggabungkan yang lama dengan yang baru untuk infrastruktur holistik yang bekerja secara sinergis.
  + Komputasi kinerja tinggi (high-performance computing) •
    - In-memory analytics: •
      * Memecahkan masalah kompleks hampir secara real-time dengan wawasan yang sangat akurat dengan memungkinkan komputasi analitik dan Big Data diproses dalam memori dan didistribusikan ke seluruh kumpulan node khusus. •
    - In-database analytics: •
      * Mempercepat waktu dan memungkinkan tata kelola data yang lebih baik dengan melakukan integrasi data dan fungsi analitik di dalam database sehingga Anda tidak perlu memindahkan atau mengonversi data berulang kali. Komputasi kinerja tinggi (high-performance computing) •
    - Grid computing: •
      * Mempromosikan efisiensi, biaya lebih rendah, dan kinerja yang lebih baik dengan memproses pekerjaan sumber daya IT yang dikelola secara terpusat. •
    - Appliances: •
      * Menyatukan perangkat keras dan perangkat lunak dalam satu unit fisik yang tidak hanya cepat tetapi juga dapat diskalakan sesuai kebutuhan.
  + Tantangan Menerapkan Analitik Big Data •
    - Data volume: •
      * Kemampuan untuk menangkap, menyimpan, dan memproses sejumlah besar data dengan kecepatan yang dapat diterima sehingga informasi terbaru tersedia bagi pengambil keputusan saat mereka membutuhkannya. •
    - Data integration: •
      * Kemampuan untuk menggabungkan data yang tidak serupa dalam struktur atau sumber dan untuk melakukannya dengan cepat dan dengan biaya yang masuk akal.
    - Processing capabilities: •
      * Kemampuan untuk memproses data dengan cepat, seperti yang ditangkap. •
      * Cara tradisional untuk mengumpulkan dan memproses data mungkin tidak berhasil. •
      * Dalam banyak situasi, data perlu dianalisis segera setelah diambil untuk memanfaatkan nilanya. •
    - Data governance: •
      * Kemampuan untuk mengikuti masalah keamanan, privasi, kepemilikan, dan kualitas Big Data. •
      * Seiring dengan perubahan volume, keragaman (format dan sumber), dan kecepatan data, demikian pula kemampuan praktik tata kelola.
    - Skills availability: •
      * Big Data memanfaatkan alat-alat baru dan melihat dengan cara yang berbeda. •
      * Ilmuwan data dengan keterampilan untuk melakukan pekerjaan tersebut. •
    - Solution cost : •
      * Karena Big Data telah membuka dunia dengan kemungkinan peningkatan bisnis, banyak eksperimen dan penemuan dilakukan untuk menentukan pola yang penting dan wawasan yang berubah menjadi nilai. •
      * Untuk memastikan pengembalian investasi yang positif pada proyek Big Data, sangat penting untuk mengurangi biaya dalam mencari solusi.
  + Masalah Bisnis yang Ditangani oleh Big Data Analytics •
    - Efisiensi proses dan pengurangan biaya •
    - Brand management •
    - Maksimalisasi pendapatan, cross-selling, dan up-selling •
    - Pengalaman pelanggan yang ditingkatkan •
    - Identifikasi perekrutan pelanggan •
    - Peningkatan layanan pelanggan •
    - Mengidentifikasi produk baru dan peluang pasar •
    - Manajemen risiko •
    - Kepatuhan terhadap peraturan •
    - Kemampuan keamanan yang ditingkatkan
  + Big Data Technologies •
    - MapReduce •
    - Hadoop •
    - NoSQL MapReduce •
    - MapReduce adalah teknik yang dipopulerkan oleh Google yang mendistribusikan pemrosesan file data multi-struktur yang sangat besar di sekelompok mesin besar. •
    - Performa tinggi dicapai dengan memecah pemrosesan menjadi unit kerja kecil yang dapat dijalankan secara paralel di ratusan, berpotensi ribuan, node dalam cluster.
    - MapReduce
      * “MapReduce adalah model pemrograman dan implementasi terkait untuk memproses dan menghasilkan kumpulan data besar. Program yang ditulis dalam gaya fungsional ini secara otomatis diparalelkan dan dieksekusi pada sekelompok besar mesin komoditas. Hal ini memungkinkan pemrogram tanpa pengalaman dengan sistem paralel dan terdistribusi untuk dengan mudah memanfaatkan sumber daya dari sistem terdistribusi besar”. (Dekan & Ghemawat, 2004)
      * Poin kunci yang perlu diperhatikan dari kutipan ini adalah bahwa MapReduce adalah model pemrograman, bukan bahasa pemrograman, yaitu, dirancang untuk digunakan oleh programmer, bukan pengguna bisnis.
      * 
      * Tujuannya adalah untuk menghitung jumlah kotak setiap warna. •
      * Programmer bertanggung jawab untuk mengkodekan map dan mereduksi program; sisa pemrosesan ditangani oleh sistem perangkat lunak yang mengimplementasikan model pemrograman MapReduce. •
      * Sistem MapReduce pertama-tama membaca file input dan membaginya menjadi beberapa bagian. •
      * Dalam contoh ini, ada dua pemisahan, tetapi dalam skenario kehidupan nyata, jumlah pemisahan biasanya akan jauh lebih tinggi. •
      * Pemisahan ini kemudian diproses oleh beberapa program map yang berjalan secara paralel pada node cluster. •
      * Peran masing-masing program map dalam hal ini adalah mengelompokkan data secara terpisah berdasarkan warna. •
      * Sistem MapReduce kemudian mengambil output dari setiap program map dan menggabungkan (mengacak/mengurutkan) hasilnya untuk input ke program reduksi, yang menghitung jumlah kuadrat dari setiap warna.
    - Mengapa Menggunakan MapReduce? •
      * MapReduce membantu organisasi dalam memproses dan menganalisis data multistruktur dalam jumlah besar. •
      * Contoh aplikasi termasuk pengindeksan dan pencarian, analisis grafik, analisis teks, pembelajaran mesin, transformasi data, dan sebagainya. •
      * Jenis aplikasi ini seringkali sulit diimplementasikan menggunakan SQL standar yang digunakan oleh DBMS relasional. •
      * Sifat prosedural MapReduce membuatnya mudah dipahami oleh programmer yang terampil. •
      * Hal ini juga memiliki keuntungan bahwa pengembang tidak perlu khawatir dengan penerapan komputasi paralel—ditangani secara transparan oleh sistem.
    - Hadoop •
      * Hadoop adalah open source framework untuk memproses, menyimpan, dan menganalisis sejumlah besar data terdistribusi dan tidak terstruktur. •
      * Awalnya dibuat oleh Doug Cutting di Yahoo! •
      * Hadoop dirancang untuk menangani petabyte dan exabyte data yang didistribusikan melalui beberapa node secara paralel. •
      * Cluster Hadoop berjalan pada perangkat keras komoditas yang murah sehingga proyek dapat ditingkatkan tanpa merusak bank. •
      * Hadoop sekarang menjadi proyek Apache Software Foundation, di mana ratusan kontributor terus meningkatkan teknologi inti.
      * Komponen Teknis Hadoop •
        + Sebuah " stack " Hadoop terdiri dari sejumlah komponen, yang meliputi: •

Hadoop Distributed File System (HDFS) •

Name Node •

Secondary Node •

Job Tracker •

Slave Nodes

Hadoop Distributed File System (HDFS): •

Lapisan penyimpanan default di setiap cluster Hadoop tertentu. •

Name Node: •

Node dalam cluster Hadoop yang menyediakan informasi klien tentang di mana dalam cluster, data tertentu disimpan; dan jika ada node yang gagal. •

Secondary Node: •

Cadangan ke Name Node, secara berkala mereplikasi dan menyimpan data dari Name Node jika gagal.

Job Tracker •

Node dalam cluster Hadoop yang memulai dan mengoordinasikan pekerjaan MapReduce atau pemrosesan data. •

Slave Nodes: •

Grunts dari setiap cluster Hadoop, node slave menyimpan data dan mengambil arah untuk memprosesnya dari Job Tracker.

* + - * Bagaimana Hadoop Bekerja? •
        + Klien mengakses data tidak terstruktur dan semi terstruktur dari sumber termasuk file log, umpan media sosial, dan penyimpanan data internal. •
        + Hadoop memecah data menjadi “parts", yang kemudian dimuat ke dalam sistem file yang terdiri dari beberapa node yang berjalan pada perangkat keras komoditas. •
        + Penyimpanan file default di Hadoop adalah Hadoop Distributed File System, atau HDFS. •
        + Sistem file seperti HDFS mahir dalam menyimpan volume besar data tidak terstruktur dan semi terstruktur karena tidak memerlukan data untuk diatur ke dalam baris dan kolom relasional. •
        + Setiap “parts" direplikasi beberapa kali dan dimuat ke dalam sistem file sehingga jika sebuah node gagal, node lain memiliki salinan data yang terdapat pada node yang gagal
        + Name Node bertindak sebagai fasilitator, mengkomunikasikan kembali ke informasi klien seperti node mana yang tersedia, di mana data tertentu berada di cluster, dan node mana yang gagal. •
        + Setelah data dimuat ke dalam cluster, data siap untuk dianalisis melalui framework MapReduce. •
        + Klien mengirimkan “Map"—biasanya kueri yang ditulis dalam Java—ke salah satu node dalam cluster yang dikenal sebagai Job Tracker. •
        + Job Tracker mengacu pada Name Node untuk menentukan data mana yang perlu diakses untuk menyelesaikan pekerjaan dan di mana dalam cluster data tersebut berada. •
        + Setelah ditentukan, Job Tracker mengirimkan kueri ke node yang relevan. •
        + Daripada membawa semua data kembali ke lokasi pusat untuk diproses, pemrosesan terjadi pada setiap node secara bersamaan, atau secara paralel. •
        + Ini adalah karakteristik penting dari Hadoop.
        + Ketika setiap node telah selesai memproses tugas yang diberikan, node menyimpan hasilnya. •
        + Klien memulai tugas " Reduce " melalui Job Tracker di mana hasil dari map phase yang disimpan secara lokal pada masing-masing node dikumpulkan untuk menentukan "jawaban" untuk kueri asli, dan kemudian dimuat ke node lain di cluster. •
        + Klien mengakses hasil ini, yang kemudian dapat dimuat ke salah satu dari sejumlah lingkungan analitik untuk analisis. •
        + Pekerjaan MapReduce sekarang telah selesai.
        + Setelah fase MapReduce selesai, data yang diproses siap untuk dianalisis lebih lanjut oleh data scientists dan lainnya dengan keterampilan analisis data tingkat lanjut. •
        + Data scientists dapat memanipulasi dan menganalisis data menggunakan salah satu dari sejumlah alat untuk sejumlah kegunaan, termasuk mencari wawasan dan pola tersembunyi, atau digunakan sebagai dasar untuk membangun aplikasi analitik yang dihadapi pengguna. •
        + Data juga dapat dimodelkan dan ditransfer dari cluster Hadoop ke database relasional yang ada, data warehouses, dan sistem IT tradisional lainnya untuk analisis lebih lanjut dan/atau untuk mendukung pemrosesan transaksional.
      * Subprojek yang paling sering dirujuk untuk Hadoop
        + ⎫ HIVE
        + ⎫ AMBARI
        + ⎫ PIG
        + ⎫ AVRO
        + ⎫ HBASE
        + ⎫ MAHOUT
        + ⎫ FLUME
        + ⎫ SQOOP
        + ⎫ OOZIE
        + ⎫ HCATALOG
    - NoSQL •
      * Gaya database baru yang disebut NoSQL (Not Only SQL) telah muncul, seperti Hadoop, memproses volume besar data multi-struktur. •
      * Basis data NoSQL ditujukan untuk menyajikan data diskrit yang disimpan di antara volume besar data multi-struktur ke pengguna akhir dan aplikasi Big Data otomatis. •
      * Kemampuan ini sangat kurang dari teknologi database relasional, yang tidak dapat mempertahankan tingkat kinerja aplikasi yang dibutuhkan pada skala Big Data. •
      * Kelemahan dari sebagian besar database NoSQL saat ini adalah memperdagangkan kepatuhan ACID (atomicity, consistency, isolation, durability) untuk kinerja dan skalabilitas. •
      * Basis data NoSQL yang saat ini tersedia antara lain HBase, Cassandra, MongoDB, Accumulo, Riak, CouchDB, dan DynamoDB.
      * Kegunaan Hadoop •
        + Hadoop adalah gudang dan penyulingan untuk data mentah. •
        + Hadoop adalah arsip yang kuat, ekonomis, dan aktif.
        + Hadoop sebagai gudang (repository) dan penyulingan (refinery)

Menuangkan data ke dalam HDFS memberikan fleksibilitas yang sangat dibutuhkan arsitek. •

Arsitek tidak hanya dapat menangkap 100-an terabyte dalam sehari, tetapi mereka juga dapat menyesuaikan konfigurasi Hadoop ke atas atau ke bawah untuk memenuhi lonjakan dan jeda dalam penyerapan data. •

Ini dicapai dengan biaya serendah mungkin per gigabyte karena ekonomi sumber terbuka dan memanfaatkan perangkat keras komoditas. •

Karena data disimpan di penyimpanan lokal, akses data Hadoop sering kali jauh lebih cepat, dan tidak menyumbat jaringan dengan perpindahan data terabyte. •

Setelah data mentah ditangkap, Hadoop digunakan untuk memperbaikinya.

* + - * + Hadoop sebagai arsip aktif •

Meskipun mungkin diperlukan waktu bertahun-tahun lagi untuk menghentikan arsip pita magnetik, saat ini beberapa bagian dari beban kerja pita telah dialihkan ke cluster Hadoop. •

Ada 2 alsan: •

Pertama, meskipun mungkin tampak murah untuk menyimpan data pada tape, biaya sebenarnya datang dengan kesulitan pengambilan. •

Kedua, telah ditunjukkan bahwa ada nilai dalam menyimpan data historis secara online dan dapat diakses.

* + - * + Kegunaan Data Warehousing •

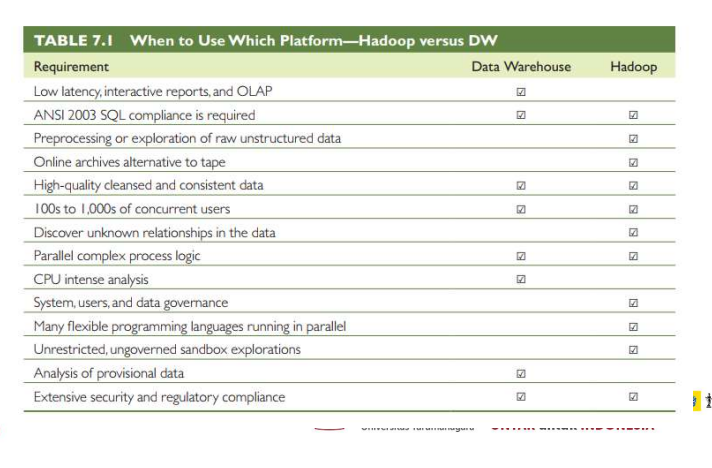
Setelah hampir 30 tahun melakukan investasi, penyempurnaan, dan pertumbuhan, daftar fitur yang tersedia di gudang data cukup mengejutkan. •

Dibangun di atas teknologi database relasional menggunakan skema dan mengintegrasikan alat BI, perbedaan utama dalam arsitektur ini adalah: •

Kinerja data warehouse •

Data terintegrasi yang memberikan nilai bisnis •

Alat BI interaktif untuk pengguna akhir

* + - * 
        + Vendor dan Platform Big Data •

Dua pemimpin pasar dalam hal distribusi adalah Cloudera (cloudera.com) dan Hortonworks (hortonworks.com). •

Cloudera dimulai oleh para pakar Big Data termasuk pencipta Hadoop Doug Cutting dan mantan ilmuwan data Facebook Jeff Hammerbacher. •

Hortonworks dikeluarkan oleh Yahoo! •

Selain distribusi, kedua perusahaan menawarkan pelatihan/layanan tingkat perusahaan berbayar dan perangkat lunak manajemen Hadoop berpemilik. •

MapR (mapr.com), start-up Valley lainnya, menawarkan distribusi Hadoop sendiri yang melengkapi HDFS dengan sistem file jaringan (NFS) miliknya untuk meningkatkan kinerja. •

Open-source tools seperti bahasa pemrograman R memiliki banyak fungsi yang diimplementasikan untuk memanfaatkan paralelisasi eksekusi melalui sebuah cluster. •

Misalnya, Treasata menawarkan aplikasi Big-Data-sebagai-layanan untuk beberapa industri.

IBM InfoSphere BigInsights •

InfoSphere BigInsights IBM adalah platform berbasis open source. •

Proyek Apache Hadoop untuk menganalisis data terstruktur tradisional yang ditemukan dalam database lama bersama dengan data semi dan tidak terstruktur seperti teks, video, audio, gambar, media sosial, log Web, dan clickstreams . •

Teradata Aster •

Teradata Aster adalah platform Big Data untuk penyimpanan terdistribusi dan pemrosesan kumpulan data multistruktur besar. •

Teradata Aster telah digunakan untuk pengoptimalan pemasaran, deteksi penipuan, analisis olahraga, analisis jejaring sosial, analisis data mesin, analisis energi, analisis perawatan kesehatan, dan banyak aplikasi lainnya.

* + - * + Big Data dan Stream Analytics •

Analisis aliran (juga disebut analitik data-dalam-gerak dan analitik data real-time) adalah istilah yang umum digunakan untuk proses analitik mengekstraksi informasi yang dapat ditindak lanjuti dari data yang terus mengalir/streaming. •

Aliran didefinisikan sebagai urutan elemen data yang berkesinambungan (Zikopoulos et al., 2013). •

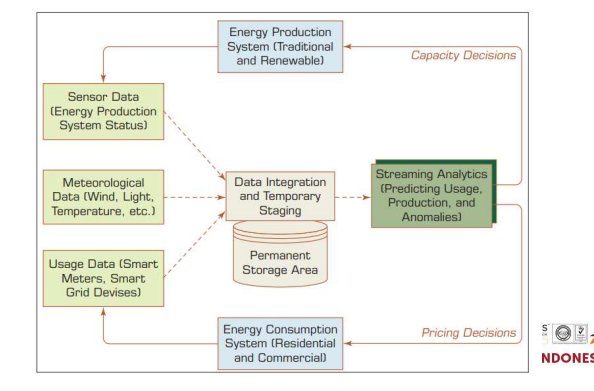
Elemen data dalam stream sering disebut tupel. •

Dalam pengertian basis data relasional, tuple mirip dengan deretan data (catatan, objek, instance). •

Namun, dalam konteks data semi terstruktur atau tidak terstruktur, tuple adalah abstraksi yang mewakili paket data, yang dapat dicirikan sebagai sekumpulan atribut untuk objek tertentu.

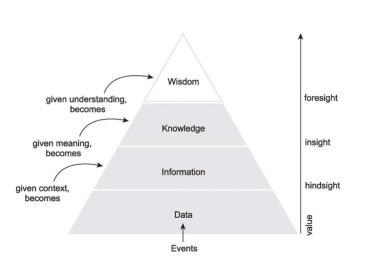
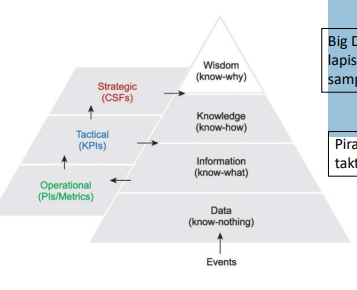
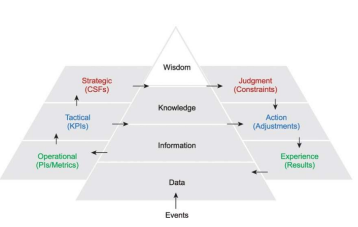
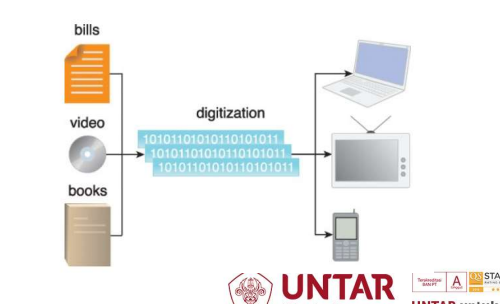
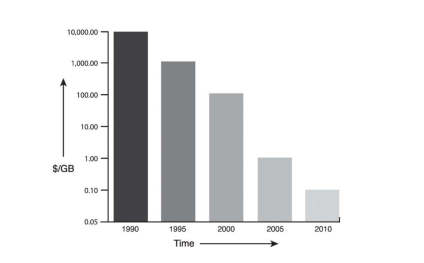
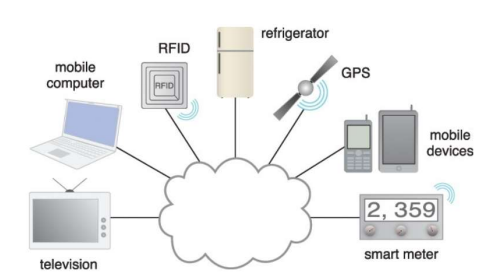
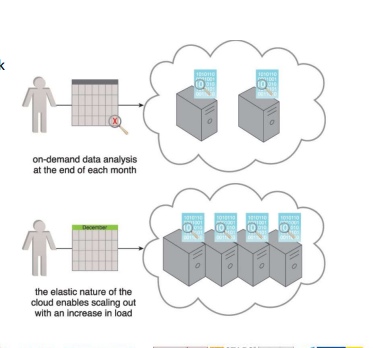
Penggambaran kasus penggunaan umum untuk analitik streaming di industri energi (smart grid application).

Tujuannya adalah untuk secara akurat memprediksi permintaan dan produksi listrik secara real-time dengan menggunakan data streaming yang berasal dari smart meter, sensor sistem produksi, dan model meteorologi.

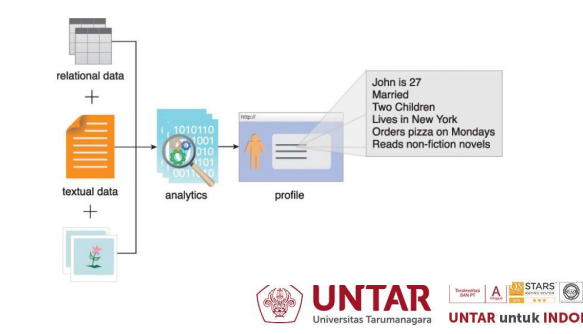
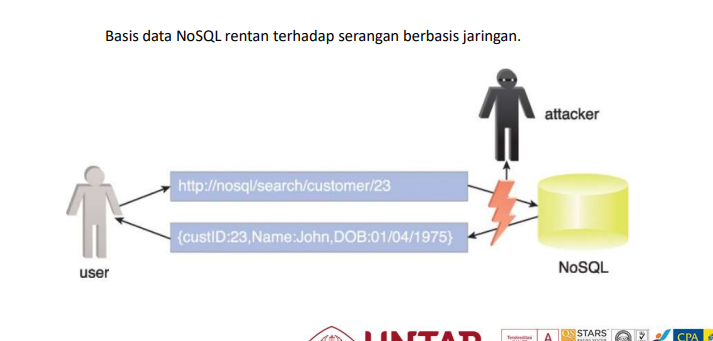
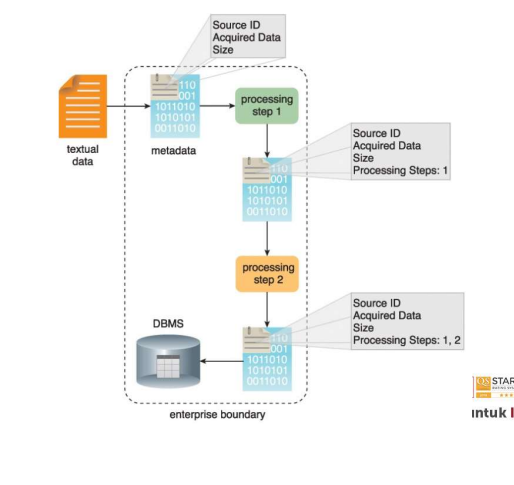
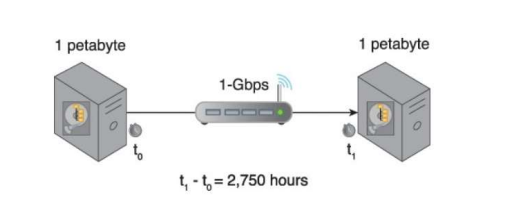
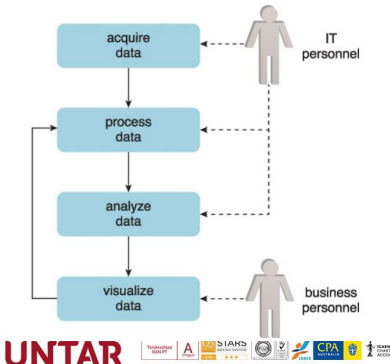
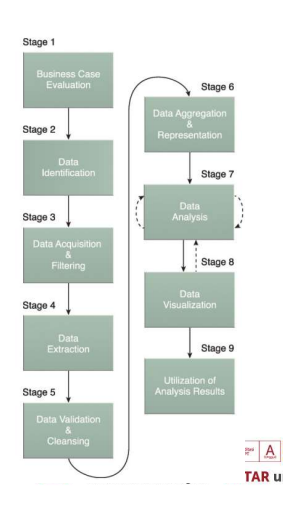
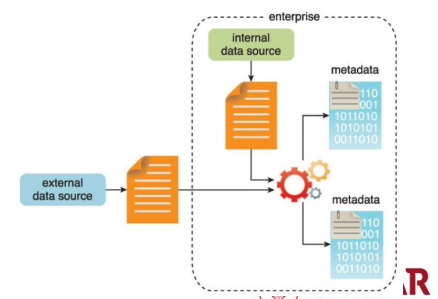
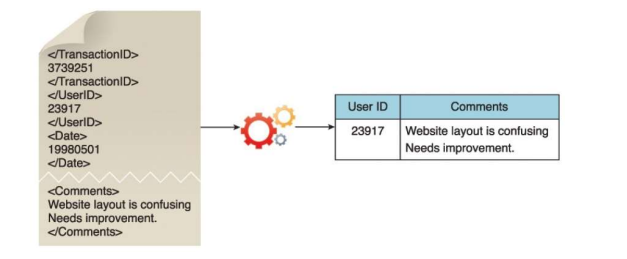
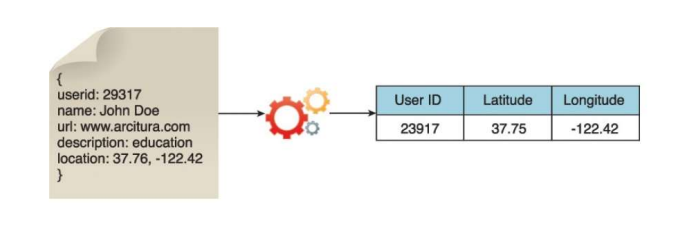
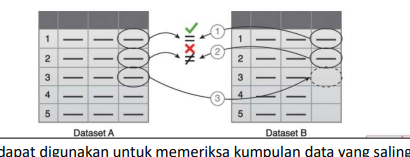
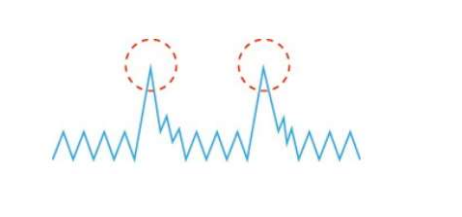
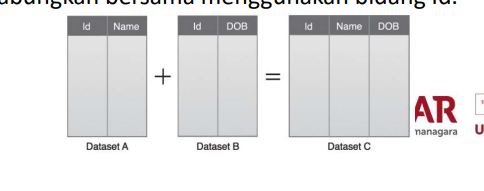
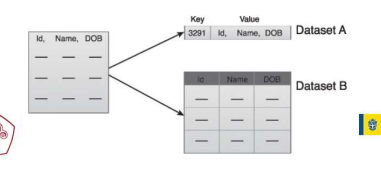
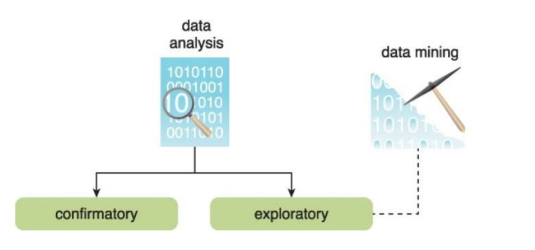
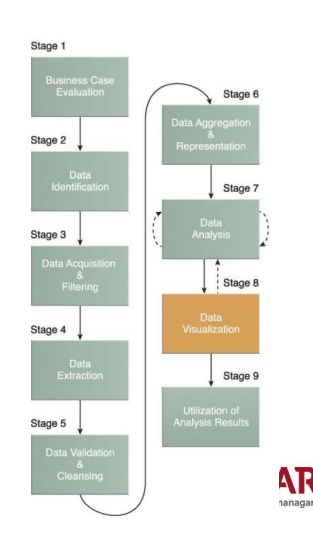


* + - * Pemrosesan peristiwa kritis (critical event processing) •
        + Pemrosesan peristiwa kritis adalah metode menangkap, melacak, dan menganalisis aliran data untuk mendeteksi peristiwa (di luar kejadian normal) dari jenis tertentu yang layak untuk dilakukan. •
        + Peristiwa penting ini mungkin terjadi di berbagai lapisan organisasi seperti prospek penjualan, pesanan, atau panggilan layanan pelanggan. •
        + Lebih luas lagi mungkin berupa berita, pesan teks, posting media sosial, feed pasar saham, laporan lalu lintas, kondisi cuaca, atau jenis anomali lain yang mungkin berdampak signifikan pada kesejahteraan organisasi.
      * Penambangan Aliran Data (Data Stream Mining) •
        + Penambangan aliran data, sebagai teknologi yang memungkinkan untuk analitik aliran, adalah proses mengekstraksi pola baru dan struktur pengetahuan dari catatan data yang cepat dan berkelanjutan. •
        + Dalam banyak aplikasi penambangan aliran data, tujuannya adalah untuk memprediksi kelas atau nilai instance baru dalam aliran data yang diberikan pengetahuan tentang keanggotaan kelas atau nilai instance sebelumnya dalam aliran data. •
        + Teknik pembelajaran mesin khusus (kebanyakan turunan dari teknik pembelajaran mesin tradisional) dapat digunakan untuk mempelajari tugas prediksi ini dari contoh berlabel secara otomatis.
      * Aplikasi Analisis Aliran •
        + perdagangan elektronik (e-Commerce) •
        + Telekomunikasi •
        + Penegakan Hukum dan Keamanan Siber •
        + Industri Tenaga Listrik •
        + Layanan Keuangan •
        + Ilmu Kesehatan

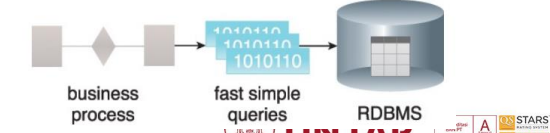
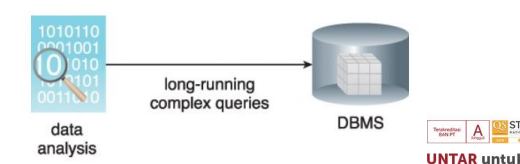
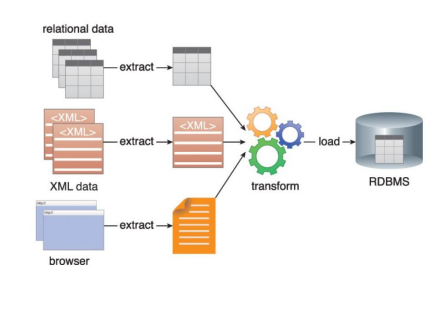
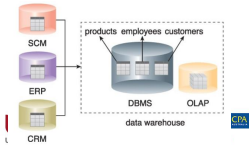
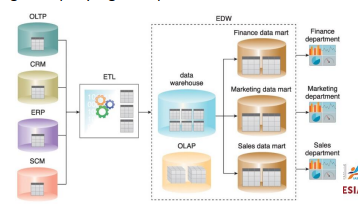
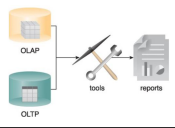
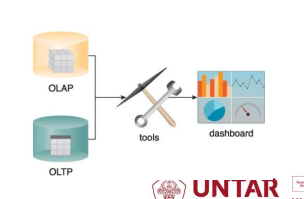
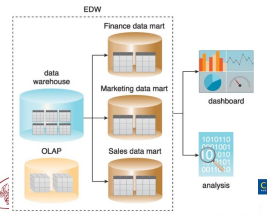
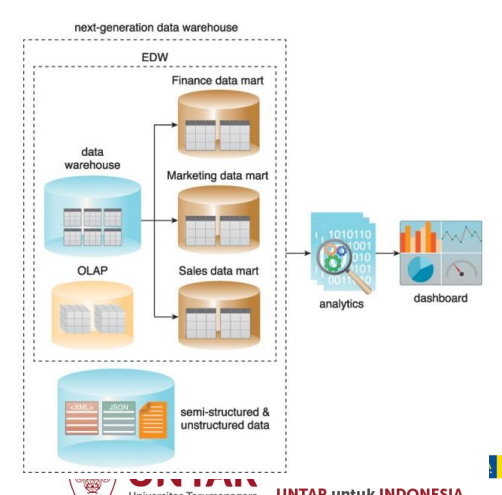
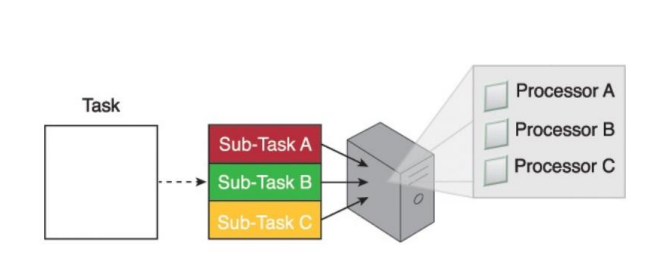
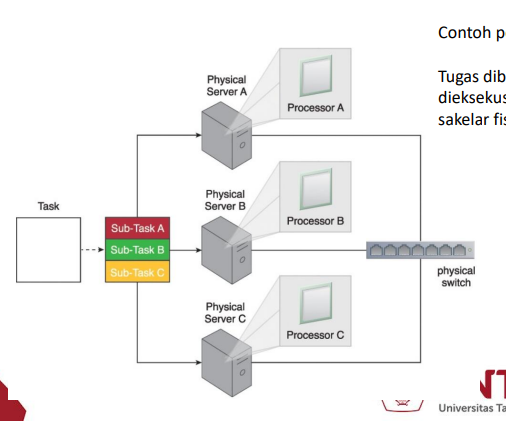
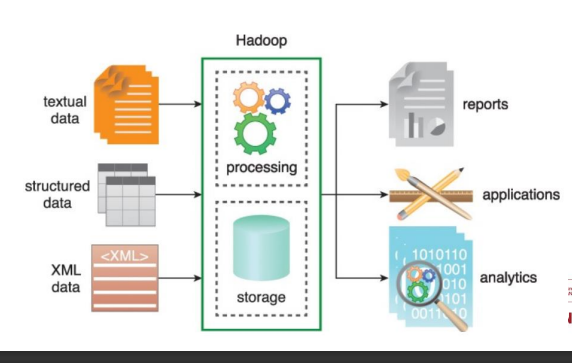
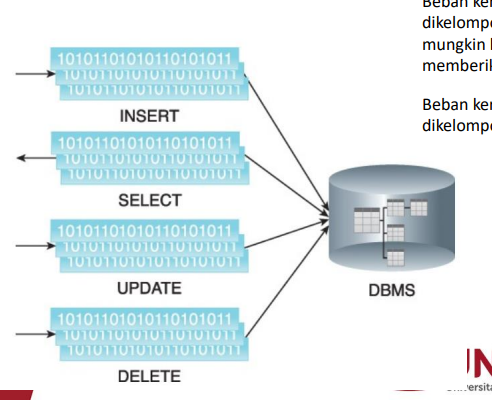
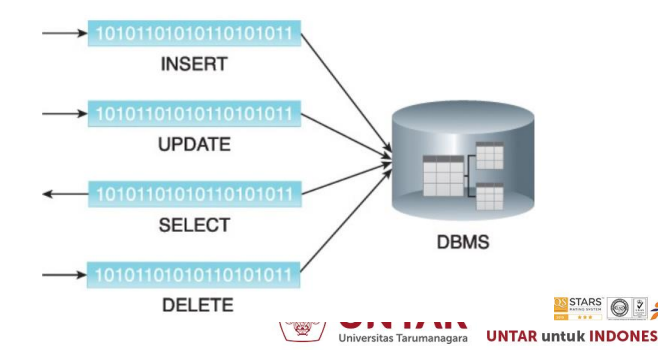
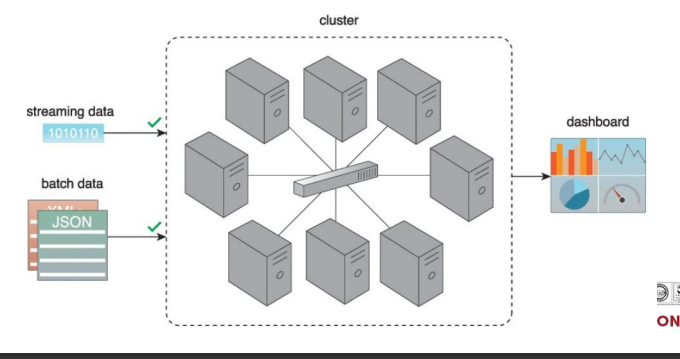
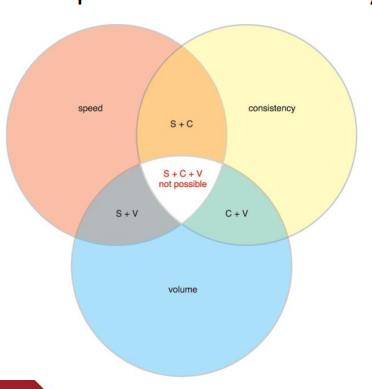
Bab 3 - Motivasi Bisnis dan Pendorong Big Data

* Dinamika Pasar •
  + Ada perubahan mendasar dalam cara pandang bisnis dan pasar. •
  + Dalam 15 tahun terakhir, dua koreksi besar pasar saham telah terjadi— yang pertama adalah ledakan dot-com bubble pada tahun 2000, dan yang kedua adalah resesi global yang dimulai pada tahun 2008. •
  + Bisnis bekerja untuk meningkatkan efisiensi dan efektivitas untuk menstabilkan profit dengan mengurangi biaya. •
  + Ketika ekonomi global pulih dari resesi, perusahaan mulai fokus ke luar, mencari pelanggan baru dan menjaga pelanggan yang sudah ada agar tidak membelot ke pesaing pasar. •
  + Inovasi membawa harapan bagi perusahaan dalam menemukan cara baru untuk mencapai keunggulan kompetitif di pasar, akibatnya meningkatkan pendapatan.
  + Perekonomian global dapat mengalami periode ketidakpastian karena berbagai faktor. •
  + Dunia bisnis mengubah perspektif tentang identitas dan kemandirian karena menyadari bahwa mereka juga terkait dalam jaringan produk dan layanan yang rumit. •
  + Untuk alasan ini, perusahaan perlu memperluas kegiatan Business Intelligence. •
  + Perusahaan perlu membuka diri terhadap sumber data eksternal sebagai sarana untuk merasakan pasar dan posisi mereka di dalamnya. •
  + Data eksternal dan data internal memungkinkan perusahaan untuk meningkatkan rantai nilai analitik. •
  + Dengan alat yang tepat untuk mendukung kemampuan simulasi yang canggih, perusahaan dapat mengembangkan hasil analitik yang menguntungkan. •
  + Inilah kekuatan Big Data— memperkaya perspektif perusahaan.
  + 
  + Di bagian atas segitiga, kebijaksanaan ditampilkan sebagai garis besar untuk menunjukkan bahwa itu ada tetapi biasanya tidak dihasilkan melalui sistem ICT (Information and Communication Technologies).
  + Knowledge worker memberikan wawasan dan pengalaman untuk membingkai pengetahuan yang ada sehingga dapat diintegrasikan untuk membentuk kearifan.
  + Dalam lingkungan bisnis, teknologi digunakan untuk mendukung manajemen pengetahuan, dan personel bertanggung jawab untuk menerapkan kompetensi dan kebijakan.
  + Piramida DIKW menunjukkan bagaimana data dapat diperkaya dengan konteks untuk menciptakan informasi, informasi dapat diberikan makna untuk menciptakan pengetahuan dan pengetahuan dapat diintegrasikan untuk membentuk kearifan.
* Arsitektur Bisnis •
  + Dalam upaya bersaing dengan TI, arsitektur bisnis telah muncul sebagai disiplin pelengkap. •
  + Tujuannya adalah arsitektur enterprise akan menghadirkan pandangan yang seimbang antara arsitektur bisnis dan teknologi. •
  + Arsitektur bisnis menyediakan sarana cetak biru (blueprinting) atau secara konkret mengekspresikan desain bisnis. •
  + Arsitektur bisnis membantu organisasi menyelaraskan visi strategis dengan pelaksanaan, sebagai sumber daya teknis atau modal manusia. •
  + Arsitektur bisnis mencakup keterkaitan dari konsep abstrak seperti misi bisnis, visi, strategi, dan tujuan ke yang lebih konkret seperti layanan bisnis, struktur organisasi, indikator kinerja utama, dan layanan aplikasi. •
  + Keterkaitan ini penting karena memberikan panduan tentang bagaimana menyelaraskan bisnis dan teknologi informasinya.
  + Bisnis beroperasi sebagai sistem berlapis. •
  + Lapisan teratas adalah lapisan strategis yang ditempati oleh eksekutif dan kelompok penasehat. •
  + Lapisan tengah adalah lapisan taktis atau manajerial yang berusaha mengarahkan organisasi agar sejalan dengan strategi. •
  + Lapisan bawah adalah lapisan operasi tempat bisnis menjalankan proses intinya dan memberikan nilai kepada pelanggannya.
  + Pemantauan aktivitas bisnis di lapisan operasi menghasilkan Performance Indicators (PI) dan metrik, baik untuk layanan maupun proses. •
  + Mereka membuat Key Performance Indicators (KPI) yang digunakan pada lapisan taktis. •
  + KPI ini dapat disejajarkan dengan Critical Success Factors (CSFs) pada lapisan strategis untuk membantu mengukur kemajuan yang dicapai menuju pencapaian tujuan dan sasaran strategis.
  + 
  + Piramida DIKW menggambarkan keselarasan dengan strategi, taktis dan tingkat operasional korporat.
  + Big Data memiliki ikatan dengan arsitektur bisnis di setiap lapisan organisasi, seperti yang digambarkan pada piramida di samping.
  + Big Data meningkatkan nilai karena menyediakan konteks tambahan melalui integrasi perspektif eksternal untuk membantu mengubah data menjadi informasi dan memberikan makna untuk menghasilkan pengetahuan dari informasi. •
  + Misalnya, di tingkat operasional, metrik yang dihasilkan hanya melaporkan apa yang terjadi dalam bisnis. •
  + Big Data mengonversi data melalui konsep dan konteks bisnis untuk menghasilkan informasi. •
  + Di tingkat manajerial, informasi dapat diperiksa melalui lensa kinerja perusahaan untuk menjawab pertanyaan tentang bagaimana kinerja bisnis (memberi makna pada informasi). •
  + Informasi ini dapat diperkaya lebih lanjut untuk menjawab pertanyaan tentang mengapa bisnis berkinerja pada tingkat itu. •
  + Lapisan strategis dapat memberikan wawasan lebih lanjut untuk membantu menjawab pertanyaan tentang strategi mana yang perlu diubah atau diadopsi untuk memperbaiki atau meningkatkan kinerja.
  + Pada sistem berlapis, lapisan tidak semuanya berubah pada kecepatan yang sama. •
  + Dalam kasus perusahaan bisnis, lapisan strategis adalah lapisan yang bergerak paling lambat, dan lapisan operasional adalah lapisan yang bergerak paling cepat. •
  + Lapisan yang bergerak lebih lambat memberikan stabilitas dan arah ke lapisan yang bergerak lebih cepat. •
  + Dalam hierarki organisasi tradisional, lapisan manajemen bertanggung jawab untuk mengarahkan lapisan operasional sejalan dengan strategi yang dibuat oleh tim eksekutif.
  + Karena variasi dalam hal perubahan kecepatan, memungkinkan untuk ketiga lapisan tersebut masing-masing bertanggung jawab atas eksekusi strategi, eksekusi bisnis, dan eksekusi proses. •
  + Masing-masing lapisan bergantung pada metrik dan ukuran yang berbeda, yang disajikan melalui fungsi visualisasi dan pelaporan yang berbeda.
  + 
  + Siklus yang baik untuk menyelaraskan organisasi di seluruh lapisan melalui loop(perulangan) umpan balik.
  + Varian diagram yang dihasilkan oleh Joe Gollner dalam “The Anatomy of Knowledge,” menunjukkan bagaimana sebuah organisasi dapat menghubungkan dan menyelaraskan lapisan organisasinya dengan menciptakan siklus yang baik melalui loop umpan balik.
  + Di sisi kanan gambar, lapisan strategis mendorong respons melalui penerapan penilaian dengan membuat keputusan mengenai strategi, kebijakan, dan sasaran perusahaan yang dikomunikasikan sebagai kendala pada lapisan taktis.
  + Lapisan taktis memanfaatkan pengetahuan ini untuk menghasilkan prioritas dan tindakan yang sesuai dengan arah perusahaan.
  + Tindakan ini menyesuaikan pelaksanaan bisnis di lapisan operasional.
  + Hal ini harus menghasilkan perubahan terukur dalam pengalaman pemangku kepentingan internal dan pelanggan eksternal saat memberikan dan menggunakan layanan bisnis
  + Hasil ini harus muncul dan terlihat dalam data dalam bentuk PI yang diubah yang kemudian digabungkan ke dalam KPI.
  + KPI adalah metrik yang dapat dikaitkan dengan faktor penentu keberhasilan yang memberi tahu tim eksekutif apakah strategi mereka berhasil atau tidak.
  + Seiring waktu, penilaian dan tindakan lapisan strategis dan manajemen ke dalam lingkaran akan berfungsi untuk menyempurnakan penyampaian layanan bisnis.
* Manajemen Proses Bisnis •
  + Bisnis memberikan nilai kepada pelanggan dan pemangku kepentingan lainnya melalui pelaksanaan proses bisnis. •
  + Proses bisnis adalah deskripsi tentang bagaimana pekerjaan dilakukan dalam sebuah organisasi. •
  + Hubungan antara kegiatan mungkin temporal; misalnya, aktivitas A dijalankan sebelum aktivitas B. •
  + Hubungan dapat menggambarkan apakah pelaksanaan kegiatan bersyarat, berdasarkan output atau kondisi yang dihasilkan oleh kegiatan lain.
  + Business Process Management Systems (BPMS) menyediakan platform perangkat lunak berbasis model yang menjadi Business Application Development Environment (BADE). •
  + Aplikasi bisnis perlu: menengahi antara manusia dan sumber daya lain yang dimiliki oleh teknologi, dijalankan sesuai dengan kebijakan perusahaan, dan memastikan distribusi pekerjaan yang adil kepada karyawan. •
  + Pada BADE, model proses bisnis digabungkan dengan model dari peran struktur organisasi, entitas bisnis, aturan bisnis, dan antarmuka pengguna. •
  + Lingkungan pengembangan mengintegrasikan model-model ini bersama- sama untuk membuat aplikasi bisnis yang mengelola screenflow dan workflow serta menyediakan manajemen beban kerja. •
  + Proses individu atau semua proses dapat diinterogasi dan divisualisasikan melalui Business Activity Monitoring (BAM).
* Teknologi Informasi dan Komunikasi •
  + Data Analytics and Data Science •
  + Digitization •
  + Affordable Technology and Commodity Hardware •
  + Social Media •
  + Hyper-Connected Communities and Devices •
  + Cloud Computing •
  + Internet of Everything (IoE)
* Data Analytics dan Data Science •
  + Perusahaan mengumpulkan, mendapatkan, menyimpan, mengatur, dan memproses data dalam jumlah yang semakin banyak. •
  + Hal ini terjadi dalam upaya untuk menemukan wawasan baru yang dapat mendorong operasi yang lebih efisien dan efektif, mengarahkan bisnis secara proaktif dan merumuskan inisiatif strategis dengan lebih baik. •
  + Perusahaan mencari cara baru untuk mendapatkan keunggulan kompetitif. •
  + Dengan demikian kebutuhan akan teknik dan teknologi yang dapat mengekstrak informasi dan wawasan yang bermakna telah meningkat.
  + Pendekatan komputasi, teknik statistik, dan data warehouse bergabung, masing-masing membawa teknik dan alat khusus yang memungkinkan kinerja analisis Big Data. •
  + Kematangan bidang praktik menginspirasi dan memungkinkan banyak fungsi inti yang diharapkan dari solusi, lingkungan, dan platform Big Data kontemporer.
* Digitization •
  + Bagi banyak bisnis, media digital telah menggantikan media fisik sebagai komunikasi de facto dan mekanisme pengiriman. •
  + Penggunaan artefak digital menghemat waktu dan biaya karena distribusi didukung oleh infrastruktur internet yang sudah ada. •
  + Saat konsumen terhubung ke bisnis melalui interaksi dengan pengganti digital mengarah pada peluang untuk mengumpulkan data "sekunder“. •
  + Misalnya, meminta pelanggan untuk memberikan umpan balik, menyelesaikan survei, atau sekadar memberikan pengait untuk menampilkan iklan yang relevan dan melacak rasio klik-tayangnya. •
  + Mengumpulkan data sekunder penting bagi bisnis karena menambang data dapat memungkinkan pemasaran yang disesuaikan, rekomendasi otomatis, dan pengembangan fitur produk yang dioptimalkan. Digitization
  + 
  + Contoh digitalisasi termasuk perbankan online, televisi on-demand dan video streaming.
* Teknologi Terjangkau dan Komoditas Perangkat Keras •
  + Teknologi yang mampu menyimpan dan memproses sejumlah besar data yang beragam menjadi semakin terjangkau. •
  + Big Data sering memanfaatkan perangkat lunak open-source yang dijalankan pada perangkat keras komoditas, yang selanjutnya mengurangi biaya. •
  + Penggunaan perangkat keras membuat Big Data dapat diakses oleh bisnis tanpa investasi modal yang besar.
  + 
  + Contoh penurunan harga terkait dengan harga penyimpanan data selama 20 tahun terakhir
* Media sosial •
  + Munculnya media sosial telah memberdayakan pelanggan untuk memberikan umpan balik hampir secara realtime melalui media terbuka dan publik. •
  + Pergeseran ini telah memaksa bisnis untuk mempertimbangkan umpan balik pelanggan tentang layanan dan penawaran produk dalam perencanaan strategis. •
  + Bisnis menyimpan semakin banyak data tentang interaksi pelanggan dalam sistem customer relationship management (CRM) dan dari mengumpulkan ulasan pelanggan, keluhan, dan pujian dari situs media sosial. •
  + Informasi ini memberi masukan bagi algoritma analisis Big Data yang memunculkan suara pelanggan dalam upaya untuk memberikan tingkat layanan yang lebih baik, meningkatkan penjualan, memungkinkan pemasaran yang ditargetkan, dan bahkan menciptakan produk dan layanan baru.
  + Bisnis telah menyadari bahwa aktivitas branding tidak lagi sepenuhnya dikelola oleh aktivitas pemasaran internal. •
  + Merek produk dan reputasi perusahaan diciptakan bersama oleh perusahaan dan pelanggannya. •
  + Bisnis semakin tertarik untuk menggabungkan kumpulan data yang tersedia untuk umum dari media sosial dan sumber data eksternal lainnya.
* Hubungan Masyarakat dan Perangkat •
  + Cakupan Internet yang semakin luas dan pertumbuhan jaringan seluler dan Wi-Fi telah memungkinkan lebih banyak orang dan perangkat untuk terus aktif dalam komunitas virtual. •
  + Digabungkan dengan perkembangan sensor yang terhubung ke Internet, dasar-dasar Internet of Things (IoT), kumpulan besar perangkat cerdas yang terhubung ke Internet sedang dibentuk.
  + 
* Cloud Computing •
  + Kemajuan cloud computing telah mengarah pada penciptaan lingkungan yang mampu menyediakan sumber daya TI yang sangat skalabel dan sesuai permintaan yang dapat disewa melalui model bayar sesuai penggunaan. •
  + Bisnis memiliki peluang untuk memanfaatkan infrastruktur, penyimpanan, dan kemampuan pemrosesan yang disediakan oleh lingkungan ini untuk membangun solusi Big Data yang dapat diskalakan yang dapat melakukan tugas pemrosesan skala besar.
  + 
  + Contoh bagaimana lingkungan cloud dapat dimanfaatkan untuk kemampuan penskalaannya untuk melakukan tugas pemrosesan Big Data.
  + Bermigrasi ke cloud adalah logis bagi perusahaan yang berencana menjalankan analitik pada kumpulan data yang tersedia melalui pasar data, karena banyak pasar data membuat kumpulan data mereka tersedia di lingkungan cloud, seperti Amazon. •
  + Alasan perusahaan menggunakan cloud computing untuk Big Data: •
    - personel sudah memiliki keterampilan cloud computing yang dibutuhkan. •
    - data input sudah ada di cloud. •
  + Singkatnya, cloud computing dapat menyediakan tiga bahan penting yang diperlukan untuk solusi Big Data: •
    - kumpulan data eksternal •
    - kemampuan pemrosesan yang dapat diskalakan •
    - penyimpanan dalam jumlah besar
* Internet of Everything (IoE) •
  + Konvergensi kemajuan dalam teknologi informasi dan komunikasi, dinamika pasar, arsitektur bisnis dan manajemen proses bisnis semuanya berkontribusi pada peluang apa yang sekarang dikenal sebagai Internet of Everything atau IoE. •
  + IoE menggabungkan layanan yang disediakan oleh perangkat cerdas terhubung dari Internet of Things ke dalam proses bisnis yang bermakna yang memiliki kemampuan untuk memberikan proposisi nilai yang unik. •
  + IoE adalah platform untuk inovasi yang memungkinkan penciptaan produk dan layanan baru dan sumber pendapatan baru untuk bisnis. •
  + Big Data adalah jantung dari IoE. •
  + Masyarakat dan perangkat yang terhubung yang berjalan pada teknologi yang terjangkau dan perangkat keras komoditas mengalirkan data digital yang tunduk pada proses analitik yang dihosting di lingkungan komputasi awan yang elastis. •
  + Perusahaan khusus IoE dapat memanfaatkan Big Data untuk membangun dan mengoptimalkan alur kerja dan menawarkannya kepada pihak ketiga sebagai proses bisnis yang dialihdayakan.
  + Salah satu contoh area yang mendapat manfaat dari IoE adalah pertanian. •
  + Ketika digabungkan menjadi satu sistem, traktor yang dikendalikan GPS, sensor kelembaban dan pemupukan di lapangan, penyiraman sesuai permintaan, sistem aplikasi pestisida, dan peralatan pembibitan tingkat variabel dapat memaksimalkan produktivitas lapangan sekaligus meminimalkan biaya. •
  + Pertanian memungkinkan pendekatan “pertanian alternatif” yang menantang industry pertanian monokultur. •
  + Dengan bantuan IoE, pertanian yang lebih kecil dapat bersaing dengan memanfaatkan keragaman tanaman dan praktik yang peka terhadap lingkungan. •
  + Selain memiliki peralatan pertanian yang terhubung secara cerdas, analisis Big Data dari peralatan dan data sensor di lapangan dapat mendorong sistem pendukung keputusan yang dapat memandu petani dan mesin mereka

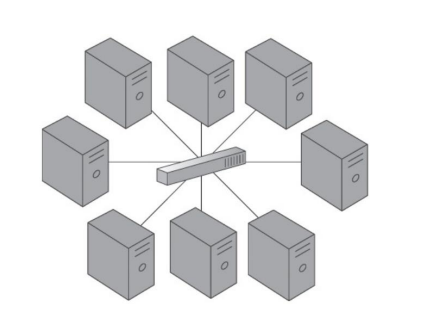
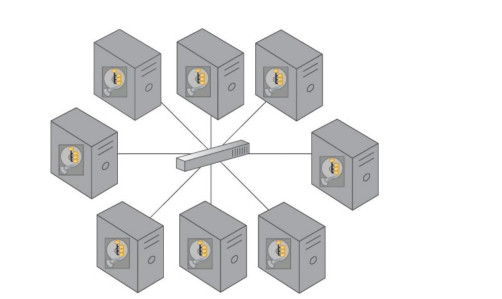
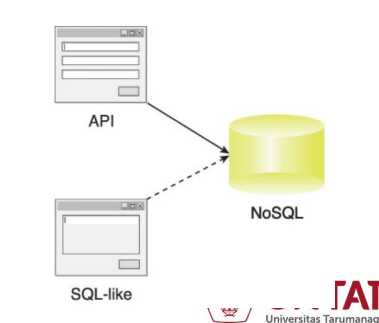
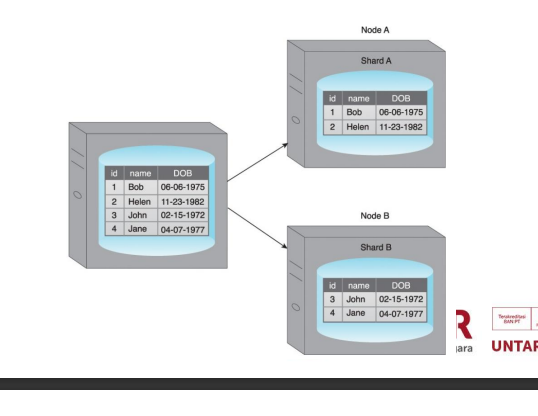
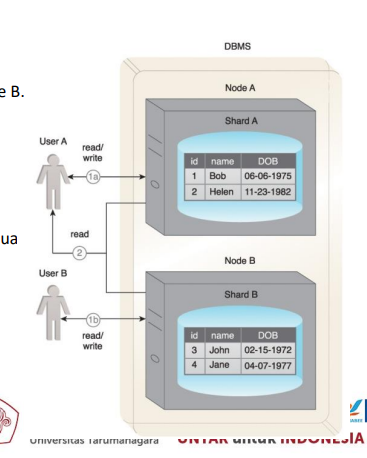
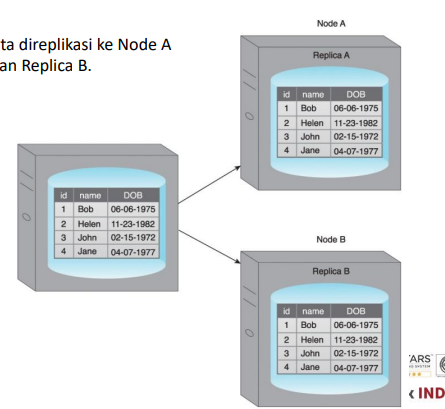
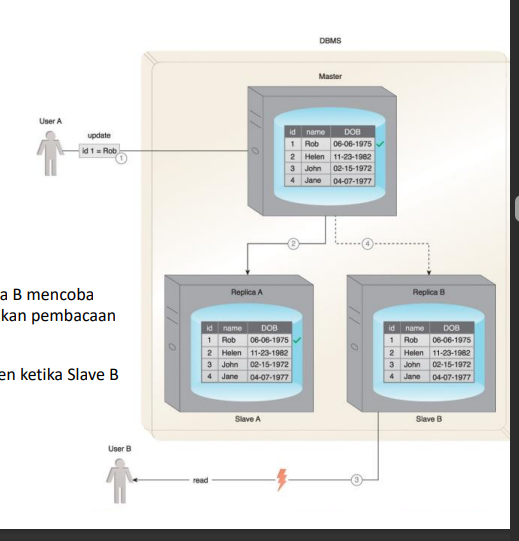
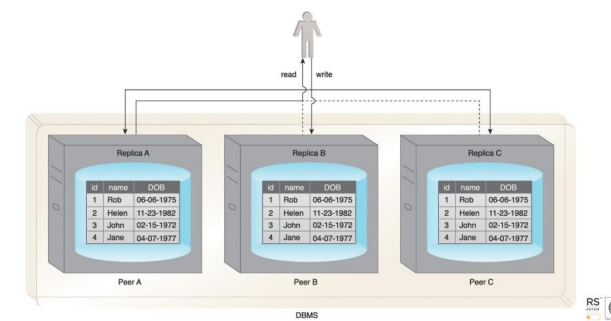
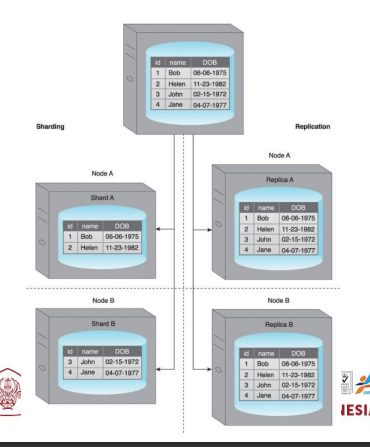
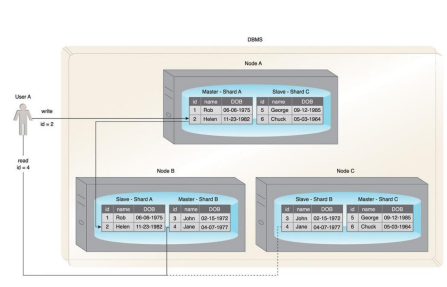
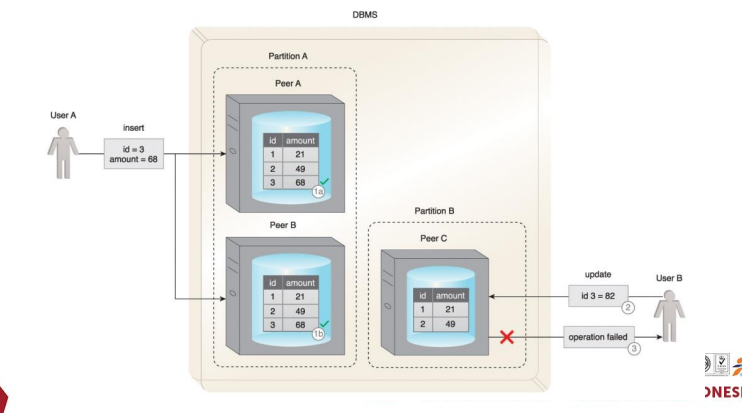
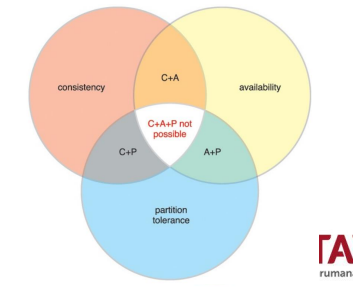
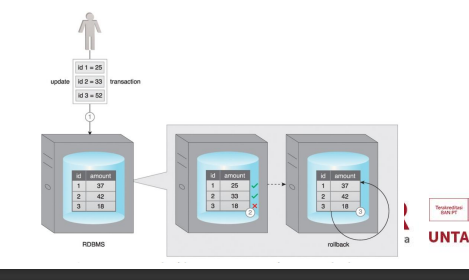
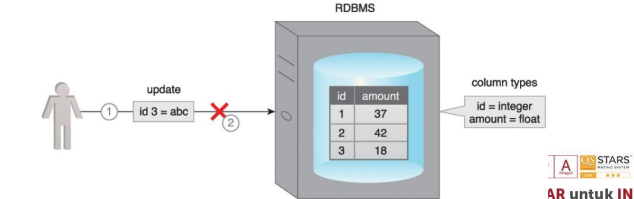
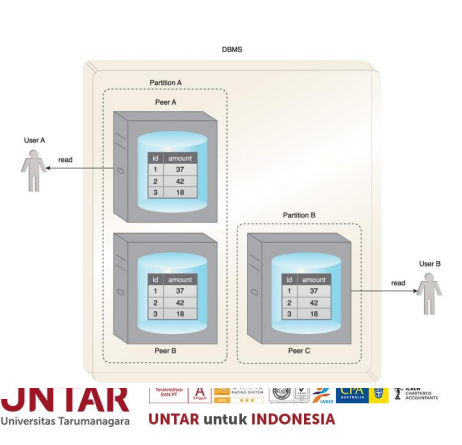
Big Data 4 - Adopsi Big Data dan Pertimbangan Perencanaan

* Prasyarat Organisasi •
  + Data yang kedaluwarsa, tidak valid, atau teridentifikasi dengan buruk akan menghasilkan input berkualitas rendah, terlepas dari seberapa bagus solusi Big Data, akan terus menghasilkan hasil berkualitas rendah. •
  + Kualitas data yang ditargetkan untuk diproses oleh solusi Big Data perlu dinilai. •
  + Roadmap perlu ditentukan untuk memastikan bahwa setiap perluasan atau penambahan lingkungan yang diperlukan direncanakan untuk tetap sinkron dengan persyaratan perusahaan.
* Pengadaan Data •
  + Akuisisi solusi Big Data dapat menjadi ekonomis, karena ketersediaan platform dan peralatan open-source serta peluang untuk memanfaatkan perangkat keras komoditas. •
  + Namun, anggaran yang besar mungkin masih diperlukan untuk mendapatkan data eksternal. •
  + Sifat bisnis dapat membuat data eksternal sangat berharga. •
  + Semakin besar volume dan variasi data yang dapat diberikan, semakin tinggi peluang untuk menemukan wawasan tersembunyi dari pola.
  + Sumber data eksternal termasuk sumber data pemerintah dan pasar data komersial. •
  + Data yang disediakan pemerintah, seperti data geospasial, mungkin gratis. •
  + Namun, sebagian besar data yang relevan secara komersial perlu dibeli dan mungkin melibatkan kelanjutan biaya berlangganan untuk memastikan pengiriman pembaruan ke kumpulan data yang diperoleh.
* Privacy •
  + Melakukan analitik pada kumpulan data dapat mengungkapkan informasi rahasia tentang organisasi atau individu. •
  + Menganalisis kumpulan data terpisah yang berisi data yang tampaknya tidak berbahaya dapat mengungkapkan informasi pribadi ketika kumpulan data dianalisis bersama. •
  + Hal ini dapat menyebabkan pelanggaran privasi yang disengaja atau tidak disengaja. •
  + Mengatasi masalah privasi ini memerlukan pemahaman tentang sifat data yang dikumpulkan dan peraturan privasi data yang relevan, serta teknik khusus untuk penandaan data dan anonimisasi. •
  + Misalnya, data telemetri, seperti log GPS mobil atau pembacaan data meteran pintar, yang dikumpulkan selama periode waktu yang lama dapat mengungkapkan lokasi dan perilaku seseorang.
  + 
  + Informasi yang dikumpulkan dari menjalankan analitik pada file gambar, data relasional, dan data tekstual digunakan untuk membuat profil John.
* Keamanan •
  + Beberapa komponen solusi Big Data tidak memiliki solusi ketahanan pada lingkungan perusahaan tradisional dalam hal kontrol akses dan keamanan data. •
  + Mengamankan Big Data memastikan bahwa jaringan data dan repositori cukup diamankan melalui mekanisme otentikasi dan otorisasi. •
  + Keamanan Big Data selanjutnya melibatkan penetapan tingkat akses data untuk berbagai kategori pengguna. •
  + Misalnya, tidak seperti sistem manajemen basis data relasional tradisional, basis data NoSQL umumnya tidak menyediakan mekanisme keamanan bawaan yang kuat. •
  + Mereka mengandalkan API berbasis HTTP sederhana di mana data dipertukarkan dalam plaintext, membuat data rentan terhadap serangan berbasis jaringan
  + 
  + Basis data NoSQL rentan terhadap serangan berbasis jaringan.
* Sumber (Provenance) •
  + Provenance mengacu pada informasi tentang sumber data dan bagaimana data itu diproses. •
  + Sumber informasi membantu menentukan keaslian dan kualitas data, dan dapat digunakan untuk tujuan audit. •
  + Mempertahankan sumbernya karena sejumlah besar data diperoleh, digabungkan, dan dimasukkan melalui beberapa tahap pemrosesan dapat menjadi tugas yang kompleks. •
  + Pada tahap yang berbeda dalam siklus hidup analitik, data akan berada dalam status yang berbeda karena faktanya mungkin sedang dikirim, diproses, atau dalam penyimpanan. •
  + Status-status ini sesuai dengan gagasan tentang data-dalam-gerak, data-dalam- penggunaan, dan data-diam. •
  + Yang penting, setiap kali Big Data berubah status, harus memicu penangkapan sumber informasi yang dicatat sebagai metadata.
  + Saat data memasuki lingkungan analitik, catatan asalnya (sumbernya) dapat diinisialisasi dengan perekaman informasi yang menangkap silsilah data. •
  + Pada akhirnya, tujuan menangkap sumbernya adalah untuk dapat mempertimbangkan hasil analitik yang dihasilkan dengan pengetahuan tentang asal data dan langkah atau algoritma apa yang digunakan untuk memproses data yang mengarah ke hasil. •
  + Sumber informasi sangat penting untuk dapat mewujudkan nilai hasil analitik. • Sama seperti penelitian ilmiah, jika hasil tidak dapat dibenarkan dan diulang, maka tidak memiliki kredibilitas. •
  + Ketika sumber informasi diperoleh, hasilnya dapat lebih mudah dipercaya, dengan demikian dapat digunakan dengan percaya diri.
  + 
  + Data perlu dianotasi dengan atribut kumpulan sumber data dan detail langkah pemrosesan saat melewati transformasi data.
* Dukungan Realtime Terbatas •
  + Dasbor dan aplikasi lain yang memerlukan streaming data dan peringatan sering kali menuntut transmisi data real-time atau nearrealtime. •
  + Banyak solusi dan alat Big Data open source berorientasi batch; namun, ada generasi baru alat open source yang mampu real-time yang memiliki dukungan untuk analisis data streaming. •
  + Banyak dari solusi analisis data real-time yang ada adalah paten. •
  + Pendekatan yang mencapai hasil near-real-time sering kali memproses data transaksional saat tiba dan menggabungkannya dengan data batch-processed yang diringkas sebelumnya.
* Tantangan Performa yang Berbeda •
  + Karena volume data yang harus diproses oleh beberapa solusi Big Data, performa sering kali menjadi perhatian. •
  + Misalnya, kumpulan data besar yang digabungkan dengan algoritma penelusuran yang kompleks dapat menyebabkan waktu kueri yang lama. •
  + Tantangan kinerja lainnya terkait dengan bandwidth jaringan. •
  + Dengan meningkatnya volume data, waktu untuk mentransfer satu unit data dapat melebihi waktu pemrosesan data yang sebenarnya.
  + 
  + Mentransfer 1 PB data melalui koneksi LAN 1-Gigabit memakan waktu sekitar 2.750 jam.
* Persyaratan Tata Kelola yang Berbeda •
  + Solusi Big Data mengakses data dan menghasilkan data, yang semuanya menjadi aset bisnis. •
  + Kerangka tata kelola diperlukan untuk memastikan bahwa data dan lingkungan solusi sendiri diatur, distandarisasi, dan dikembangkan secara terkendali
  + Contoh kerangka kerja tata kelola Big Data meliputi: •
    - standarisasi bagaimana data ditandai dan metadata yang digunakan untuk penandaan •
    - kebijakan yang mengatur jenis data eksternal yang mungkin diperoleh •
    - kebijakan mengenai pengelolaan privasi data dan anonimisasi data •
    - kebijakan untuk pengarsipan sumber data dan hasil analisis •
    - kebijakan yang menetapkan pedoman untuk pembersihan dan penyaringan data
* Metodologi yang Berbeda •
  + Metodologi akan diperlukan untuk mengontrol bagaimana data mengalir masuk dan keluar dari solusi Big Data. •
  + Perlu mempertimbangkan bagaimana loop umpan balik dapat dibuat untuk memungkinkan data yang diproses mengalami penyempurnaan berulang. •
  + Misalnya, pendekatan berulang dapat digunakan untuk memungkinkan personel bisnis memberikan umpan balik kepada personel TI secara berkala. •
  + Setiap siklus umpan balik memberikan peluang untuk penyempurnaan sistem dengan memodifikasi persiapan data atau langkah-langkah analisis data.
  + 
  + Setiap pengulangan dapat membantu menyempurnakan langkah pemrosesan, algoritma, dan model data untuk meningkatkan akurasi hasil dan memberikan nilai yang lebih besar.
* Clouds •
  + Cloud menyediakan lingkungan jarak jauh yang dapat menampung infrastruktur TI untuk penyimpanan dan pemrosesan skala besar. •
  + Terlepas dari apakah suatu organisasi sudah mendukung cloud, adopsi lingkungan Big Data mungkin mengharuskan beberapa atau semua lingkungan ditaruh di dalam cloud. •
  + Misalnya, perusahaan yang menjalankan sistem CRM-nya di cloud memutuskan untuk menambahkan solusi Big Data di lingkungan cloud yang sama untuk menjalankan analitik pada data CRM-nya. •
  + Data ini kemudian dapat dibagikan dengan lingkungan Big Data utamanya yang berada dalam batas-batas perusahaan.
  + Pembenaran umum untuk menggabungkan lingkungan cloud untuk mendukung solusi Big Data meliputi: •
    - sumber daya perangkat keras internal yang tidak memadai •
    - investasi modal dimuka untuk pengadaan sistem tidak tersedia •
    - proyek akan diisolasi dari bisnis lainnya sehingga proses bisnis yang ada tidak terpengaruh •
    - inisiatif Big Data adalah bukti konsep •
    - kumpulan data yang perlu diproses sudah menjadi penduduk cloud •
    - batas sumber daya komputasi dan penyimpanan yang tersedia yang digunakan oleh solusi Big Data internal telah tercapai
* Siklus Hidup Analisis Big Data •
  + Analisis Big Data berbeda dari analisis data tradisional terutama karena karakteristik volume, velocity (kecepatan), dan variety (variasi) dari data yang sedang diproses. •
  + Untuk mengatasi persyaratan berbeda untuk melakukan analisis pada Big Data, metodologi langkah demi langkah diperlukan untuk mengatur aktivitas dan tugas yang terkait dengan perolehan, pemrosesan, analisis, dan penggunaan kembali data.
  + 
  + Sembilan tahap siklus hidup analitik Big Data.
  + Evaluasi Kasus Bisnis (Business Case Evaluation) •
    - Setiap siklus hidup analitik Big Data harus dimulai dengan kasus bisnis yang terdefinisi dengan baik yang menyajikan pemahaman yang jelas tentang pembenaran, motivasi, dan tujuan pelaksanaan analisis. •
    - Tahap Evaluasi Kasus Bisnis, mengharuskan kasus bisnis dibuat, dinilai dan disetujui sebelum melanjutkan dengan tugas analisis langsung yang sebenarnya.
    - Evaluasi kasus bisnis analitik Big Data membantu pengambil keputusan memahami sumber daya bisnis yang perlu digunakan dan tantangan bisnis mana yang akan ditangani oleh analisis. •
    - Identifikasi KPI lebih lanjut selama tahap ini dapat membantu menentukan kriteria penilaian dan panduan untuk evaluasi hasil analitik.
    - Jika KPI tidak tersedia, upaya harus dilakukan untuk membuat tujuan proyek analisis SMART yang spesifik, terukur, dapat dicapai, relevan, dan tepat waktu. •
    - Untuk memenuhi syarat sebagai masalah Big Data, masalah bisnis harus terkait langsung dengan satu atau lebih karakteristik Big Data volume, kecepatan, atau variasi. •
    - Perhatikan juga bahwa hasil lain dari tahap ini adalah penentuan anggaran dasar yang diperlukan untuk melaksanakan proyek analisis
    - Setiap pembelian yang diperlukan, seperti peralatan, perangkat keras dan pelatihan, harus dipahami terlebih dahulu sehingga investasi yang diantisipasi dapat ditimbang terhadap manfaat yang diharapkan dari pencapaian tujuan. •
    - Iterasi awal dari siklus hidup analitik Big Data akan membutuhkan lebih banyak investasi di muka untuk teknologi, produk, dan pelatihan Big Data dibandingkan dengan iterasi selanjutnya di mana investasi sebelumnya ini dapat dimanfaatkan berulang kali.
  + Identifikasi Data (Data Identification) •
    - Tahap Identifikasi Data didedikasikan untuk mengidentifikasi kumpulan data yang diperlukan untuk proyek analisis dan sumbernya. •
    - Mengidentifikasi berbagai sumber data yang lebih luas dapat meningkatkan kemungkinan menemukan pola dan korelasi yang tersembunyi. •
    - Misalnya, untuk memberikan wawasan, akan bermanfaat untuk mengidentifikasi sebanyak mungkin jenis sumber data terkait, terutama jika tidak jelas apa yang harus dicari. •
    - Bergantung pada ruang lingkup bisnis proyek analisis dan sifat masalah bisnis yang ditangani, kumpulan data yang diperlukan dan sumbernya dapat berasal dari internal dan/atau eksternal perusahaan.
    - Dalam kasus kumpulan data internal, daftar kumpulan data yang tersedia dari sumber internal, seperti data mart dan sistem operasional, biasanya dikompilasi dan dicocokkan dengan spesifikasi kumpulan data yang telah ditentukan sebelumnya. •
    - Dalam hal kumpulan data eksternal, daftar kemungkinan penyedia data pihak ketiga, seperti pasar data dan kumpulan data yang tersedia untuk umum, dikompilasi. •
    - Beberapa bentuk data eksternal mungkin disematkan di dalam blog atau jenis situs web berbasis konten lainnya, dalam hal ini data tersebut mungkin perlu dikumpulkan melalui alat otomatis.
  + Akuisisi dan Penyaringan Data (Data Acquisition and Filtering) •
    - Selama tahap Akuisisi dan Penyaringan Data, data dikumpulkan dari semua sumber data yang diidentifikasi pada tahap sebelumnya. •
    - Data yang diperoleh kemudian difilter secara otomatis untuk menghilangkan data yang rusak atau data yang dianggap tidak memiliki nilai untuk tujuan analisis. •
    - Bergantung pada jenis sumber data, data mungkin datang sebagai kumpulan file, seperti data yang dibeli dari penyedia data pihak ketiga, atau mungkin memerlukan integrasi API, seperti dengan Twitter. •
    - Dalam banyak kasus, terutama jika menyangkut data eksternal yang tidak terstruktur, beberapa atau sebagian besar data yang diperoleh mungkin tidak relevan (noise) dan dapat dibuang sebagai bagian dari proses penyaringan.
    - Data yang diklasifikasikan sebagai "rusak/corrupt" dapat mencakup catatan dengan nilai yang hilang atau tidak masuk akal atau tipe data yang tidak valid. •
    - Data yang disaring untuk satu analisis mungkin berharga untuk jenis analisis yang berbeda. •
    - Disarankan untuk menyimpan salinan rinci dari dataset asli sebelum melanjutkan dengan penyaringan. •
    - Untuk meminimalkan ruang penyimpanan yang diperlukan, salinan dapat dikompresi.
    - Baik data internal maupun eksternal perlu dipertahankan setelah dihasilkan atau memasuki batas perusahaan. •
    - Untuk analitik batch, data ini disimpan ke disk sebelum dianalisis. •
    - Dalam kasus analitik real-time, data dianalisis terlebih dahulu dan kemudian disimpan ke disk.
    - 
    - Metadata ditambahkan ke data dari sumber internal dan eksternal
    - Metadata dapat ditambahkan melalui otomatisasi ke data dari sumber data internal dan eksternal untuk meningkatkan klasifikasi dan kueri. •
    - Contoh metadata yang ditambahkan mencakup ukuran dan struktur kumpulan data, sumber informasi, tanggal dan waktu pembuatan atau pengumpulan, dan informasi khusus bahasa. •
    - Sangat penting bahwa metadata dapat dibaca mesin dan diteruskan di sepanjang tahap analisis selanjutnya. •
    - Ini membantu menjaga asal data di seluruh siklus hidup analitik Big Data, yang membantu membangun dan menjaga akurasi dan kualitas data.
  + Ekstraksi Data (Data Extraction) •
    - Beberapa data yang diidentifikasi sebagai masukan untuk analisis mungkin dalam format yang tidak sesuai dengan solusi Big Data. •
    - Kebutuhan untuk mengatasi jenis data yang berbeda lebih mungkin dengan data dari sumber eksternal. •
    - Tahap siklus hidup Ekstraksi Data didedikasikan untuk mengekstraksi data yang berbeda dan mengubahnya menjadi format yang dapat digunakan oleh solusi Big Data yang mendasarinya untuk tujuan analisis data.
    - Tingkat ekstraksi dan transformasi yang diperlukan bergantung pada jenis analitik dan kemampuan solusi Big Data. •
    - Misalnya, mengekstrak bidang yang diperlukan dari data tekstual yang dibatasi, seperti dengan file log server web, mungkin tidak diperlukan jika solusi Big Data yang mendasarinya sudah dapat langsung memproses file tersebut. •
    - Mengekstrak teks untuk analitik teks, yang memerlukan pemindaian seluruh dokumen, disederhanakan jika solusi Big Data yang mendasarinya dapat langsung membaca dokumen dalam format aslinya.
    - 
    - Ilustrasi ekstraksi komentar dan ID pengguna yang disematkan di dalam dokumen XML tanpa perlu transformasi lebih lanjut.
    - 
    - ID pengguna dan koordinat pengguna diekstraksi dari satu JSON.
  + Validasi dan Pembersihan Data (Data Validation and Cleansing) •
    - Data yang tidak valid dapat mengacaukan dan memalsukan hasil analisis. •
    - Tidak seperti data perusahaan tradisional, di mana struktur data telah ditentukan sebelumnya dan data telah divalidasi sebelumnya, input data ke dalam analisis Big Data dapat tidak terstruktur tanpa indikasi validitas. •
    - Kompleksitasnya selanjutnya dapat mempersulit untuk sampai pada serangkaian batasan validasi yang sesuai. •
    - Tahap Validasi dan Pembersihan didedikasikan untuk menetapkan aturan validasi yang seringkali rumit dan menghapus data tidak valid yang diketahui. Validasi dan Pembersihan Data (Data Validation and Cleansing) •
    - Solusi Big Data sering kali menerima data yang berlebihan di seluruh kumpulan data yang berbeda. •
    - Redundansi ini dapat dimanfaatkan untuk mengeksplorasi kumpulan data yang saling berhubungan untuk mengumpulkan parameter validasi dan mengisi data valid yang hilang
    - Nilai pertama di Dataset B divalidasi terhadap nilai yang sesuai di Dataset A. •
    - Nilai kedua di Dataset B tidak divalidasi terhadap nilai yang sesuai di Dataset A. •
    - Jika ada nilai yang hilang, nilai tersebut disisipkan dari Dataset A.
    - 
    - Validasi data dapat digunakan untuk memeriksa kumpulan data yang saling berhubungan untuk mengisi data valid yang hilang.
    - Untuk analitik batch, validasi dan pembersihan data dapat dicapai melalui operasi ETL offline. •
    - Untuk analitik real-time, sistem dalam memori yang lebih kompleks diperlukan untuk memvalidasi dan membersihkan data saat tiba dari sumbernya. •
    - Sumber data dapat memainkan peran penting dalam menentukan keakuratan dan kualitas data yang dipertanyakan. •
    - Data yang tampaknya tidak valid mungkin masih berharga karena mungkin memiliki pola dan tren tersembunyi.
    - Adanya data yang tidak valid mengakibatkan lonjakan. •
    - Meskipun data tampak tidak normal, mungkin merupakan indikasi pola baru.
    - 
  + Agregasi dan Representasi Data (Data Aggregation and Representation) •
    - Data dapat tersebar di beberapa kumpulan data, yang mengharuskan kumpulan data digabungkan bersama melalui bidang umum, misalnya tanggal atau ID. •
    - Dalam kasus lain, bidang data yang sama mungkin muncul di beberapa kumpulan data, seperti tanggal lahir. •
    - Metode rekonsiliasi data diperlukan atau kumpulan data yang mewakili nilai yang benar perlu ditentukan. •
    - Tahap Agregasi dan Representasi Data didedikasikan untuk mengintegrasikan beberapa kumpulan data bersama-sama untuk sampai pada tampilan terpadu.
    - Melakukan tahap ini bisa menjadi rumit karena perbedaan dalam: •
      * Struktur Data – Meskipun format datanya mungkin sama, model datanya mungkin berbeda •
      * Semantik – Nilai yang diberi label berbeda dalam dua kumpulan data yang berbeda dapat berarti hal yang sama, misalnya “nama keluarga (surname)” dan “nama belakang (last name
    - Persyaratan analisis data di masa depan perlu dipertimbangkan selama tahap ini untuk membantu mendorong penggunaan kembali data. •
    - Apakah agregasi data diperlukan atau tidak, penting untuk dipahami bahwa data yang sama dapat disimpan dalam berbagai bentuk. •
    - Satu bentuk mungkin lebih cocok untuk jenis analisis tertentu daripada yang lain. •
    - Misalnya, data yang disimpan sebagai BLOB tidak akan banyak berguna jika analisis memerlukan akses ke bidang data individual.
    - Struktur data yang distandarisasi oleh solusi Big Data dapat bertindak sebagai penyebut umum yang dapat digunakan untuk berbagai teknik dan proyek analisis. •
    - Memerlukan pembuatan repositori analisis standar pusat, seperti database NoSQL. •
    - Contoh sederhana agregasi data di mana dua kumpulan data digabungkan bersama menggunakan bidang Id.
    - 
    - Dataset A berisi bagian data yang diinginkan, tetapi merupakan bagian dari BLOB yang tidak mudah diakses untuk query. •
    - Dataset B berisi bagian data yang sama yang diatur dalam penyimpanan berbasis kolom, memungkinkan setiap bidang untuk ditanyakan satu per satu. •
    - Dataset A dan B dapat digabungkan untuk membuat struktur data standar dengan solusi Big Data.
    - 
  + Analisis Data (Data Analysis) •
    - Tahap Analisis Data didedikasikan untuk melaksanakan tugas analisis yang sebenarnya, yang biasanya melibatkan satu atau lebih jenis analisis. •
    - Tahap ini dapat bersifat iteratif, terutama jika analisis data bersifat eksploratif, di mana analisis kasus diulang sampai pola atau korelasi yang sesuai ditemukan. •
    - Pendekatan analisis eksplorasi akan dijelaskan secara singkat, bersama dengan analisis konfirmatori. Analisis Data (Data Analysis) •
    - Bergantung pada jenis hasil analitik yang diperlukan, tahap ini bisa sesederhana menanyakan kumpulan data untuk menghitung agregasi untuk perbandingan. •
    - Di sisi lain, ini bisa sama menantangnya dengan menggabungkan penambangan data dan teknik analisis statistik yang kompleks untuk menemukan pola dan anomali atau untuk menghasilkan model statistik atau matematis untuk menggambarkan hubungan antar variabel. •
    - Analisis data dapat diklasifikasikan sebagai analisis konfirmasi atau analisis eksplorasi, yang terakhir terkait dengan penambangan data.
    - Analisis data dapat dilakukan sebagai analisis konfirmasi atau eksplorasi.
    - 
    - Analisis data konfirmatori adalah pendekatan deduktif di mana penyebab fenomena yang diselidiki diusulkan terlebih dahulu. •
    - Penyebab atau asumsi yang diajukan disebut hipotesis. •
    - Data tersebut kemudian dianalisis untuk membuktikan atau menyangkal hipotesis dan memberikan jawaban definitif untuk pertanyaan tertentu. •
    - Digunakan teknik pengambilan sampel data. •
    - Temuan tak terduga atau anomali biasanya diabaikan karena penyebab yang telah ditentukan diasumsikan.
    - Analisis data eksplorasi adalah pendekatan induktif yang terkait erat dengan data mining. •
    - Tidak ada hipotesis atau asumsi yang telah ditentukan sebelumnya yang dihasilkan. Sebaliknya, data dieksplorasi melalui analisis untuk mengembangkan pemahaman tentang penyebab fenomena tersebut. •
    - Meskipun mungkin tidak memberikan jawaban yang pasti, metode ini memberikan arahan umum yang dapat memfasilitasi penemuan pola atau anomali.
  + Visualisasi Data (Data Visualization) •
    - Kemampuan untuk menganalisis data dalam jumlah besar dan menemukan wawasan yang berguna hanya memiliki sedikit nilai jika satu-satunya yang dapat menginterpretasikan hasilnya adalah para analis. •
    - Tahap Visualisasi Data didedikasikan untuk menggunakan teknik dan alat visualisasi data untuk mengkomunikasikan hasil analisis secara grafis untuk interpretasi yang efektif oleh pengguna bisnis. •
    - Pengguna bisnis harus dapat memahami hasil untuk mendapatkan nilai dari analisis dan selanjutnya memiliki kemampuan untuk memberikan umpan balik, seperti yang ditunjukkan oleh garis putusputus yang mengarah dari tahap 8 kembali ke tahap 7.
    - 
    - Hasil menyelesaikan tahap Visualisasi Data memberi pengguna kemampuan untuk melakukan analisis visual, memungkinkan penemuan jawaban atas pertanyaan yang bahkan belum dirumuskan pengguna. •
    - Hasil yang sama dapat disajikan dalam beberapa cara yang berbeda, yang dapat mempengaruhi interpretasi hasil. Akibatnya, penting untuk menggunakan teknik visualisasi yang paling cocok dengan menjaga domain bisnis dalam konteks. •
    - Aspek lain yang perlu diingat adalah bahwa menyediakan metode pengeboran ke statistik yang relatif sederhana sangat penting, agar pengguna memahami bagaimana hasil gabungan atau gabungan dihasilkan.
  + Pemanfaatan Hasil Analisis (Utilization of Analysis Results) •
    - Setelah hasil analisis tersedia bagi pengguna bisnis untuk mendukung pengambilan keputusan bisnis, seperti melalui dasbor, mungkin ada peluang lebih lanjut untuk memanfaatkan hasil analisis. •
    - Tahap Pemanfaatan Hasil Analisis ditujukan untuk menentukan bagaimana dan di mana data analisis yang diproses dapat dimanfaatkan lebih lanjut. •
    - Tergantung pada sifat masalah analisis yang ditangani, hasil analisis mungkin menghasilkan "model" yang merangkum wawasan dan pemahaman baru tentang sifat pola dan hubungan yang ada dalam data yang dianalisis. •
    - Sebuah model mungkin terlihat seperti persamaan matematika atau seperangkat aturan. •
    - Model dapat digunakan untuk meningkatkan logika proses bisnis dan logika sistem aplikasi, dan dapat membentuk dasar sistem atau program perangkat lunak baru.
    - Area umum yang dieksplorasi selama tahap ini meliputi:
      * ⎫Input untuk Sistem Perusahaan
      * ⎫Optimasi Proses Bisnis
      * ⎫Peringatan
* Input untuk Sistem Perusahaan •
  + Hasil analisis data dapat dimasukkan secara otomatis atau manual langsung ke sistem perusahaan untuk meningkatkan dan mengoptimalkan perilaku dan kinerjanya. •
  + Misalnya, toko online dapat diumpankan ke hasil analisis terkait pelanggan yang diproses yang dapat memengaruhi cara toko tersebut menghasilkan rekomendasi produk. •
  + Model baru dapat digunakan untuk meningkatkan logika pemrograman dalam sistem perusahaan yang ada atau dapat membentuk dasar dari sistem baru.
* Optimasi Proses Bisnis •
  + Pola, korelasi, dan anomali yang teridentifikasi selama analisis data digunakan untuk menyempurnakan proses bisnis. •
  + Contohnya adalah mengkonsolidasikan rute transportasi sebagai bagian dari proses rantai pasokan. •
  + Model juga dapat mengarah pada peluang untuk meningkatkan logika proses bisnis.
* Peringatan •
  + Hasil analisis data dapat digunakan sebagai masukan untuk peringatan yang ada atau dapat menjadi dasar peringatan baru. •
  + Misalnya, peringatan dapat dibuat untuk memberi tahu pengguna melalui email atau teks SMS tentang suatu peristiwa yang mengharuskan mereka mengambil tindakan korektif

Bab 5 - Enterprise Technologies and Big Data Business Intelligence

* Online Transaction Processing (OLTP) •
  + OLTP adalah sistem perangkat lunak yang memproses data berorientasi transaksi. •
  + Istilah “transaksi online" mengacu pada penyelesaian suatu aktivitas secara real-time dan tidak diproses secara batch. •
  + Sistem OLTP menyimpan data operasional yang dinormalisasi. •
  + Data OLTP adalah sumber umum dari data terstruktur dan berfungsi sebagai masukan untuk banyak proses analisis. •
  + Hasil analisis Big Data dapat digunakan untuk menambah data OLTP yang disimpan dalam basis data relasional.
  + Sistem OLTP, misalnya sistem point of sale, menjalankan proses bisnis untuk mendukung operasi perusahaan. •
  + Sistem OLTP melakukan operasi basis data sederhana untuk menyediakan waktu respons subdetik. •
  + Kueri yang didukung oleh sistem OLTP terdiri dari operasi insert, delete dan update sederhana dengan waktu respons sub-detik. •
  + 
  + Contohnya termasuk sistem reservasi tiket, perbankan dan sistem point of sale.
* Online Analytical Processing (OLAP) •
  + Online Analytical Processing (OLAP) digunakan untuk memproses kueri analisis data. •
  + OLAP merupakan bagian integrasi dari business intelligence, penambangan data, dan proses pembelajaran mesin. •
  + Hal tersebut relevan dengan Big Data karena dapat berfungsi sebagai sumber data serta sink data yang mampu menerima data. •
  + OLAP digunakan dalam analisis diagnostik, prediktif dan preskriptif. •
  + Sistem OLAP melakukan kueri kompleks yang berjalan lama terhadap database multidimensi yang strukturnya dioptimalkan untuk melakukan analisis tingkat lanjut.
  + Sistem OLAP menyimpan data historis yang dikumpulkan dan didenormalisasi untuk mendukung kemampuan pelaporan yang cepat. •
  + Sistem OLAP menggunakan database yang menyimpan data historis dalam struktur multidimensi dan dapat menjawab pertanyaan kompleks berdasarkan hubungan antara berbagai aspek data. •
  + Gambar di bawah ini menunjukkan sistem OLAP menggunakan database multidimensi.
  + 
* Extract Transform Load (ETL) •
  + Extract Transform Load (ETL) adalah proses memuat data dari sistem sumber ke sistem target. •
  + Sumber dapat berupa database, flat file, atau aplikasi. •
  + Target sistem dapat berupa database atau sistem penyimpanan lainnya. •
  + ETL mewakili operasi utama di mana data warehouses diumpankan. •
  + Solusi Big Data mencakup seperangkat fitur ETL untuk mengonversi data dari berbagai jenis.
  + 
  + Data yang diperlukan diperoleh atau diekstraksi dari sumbernya, setelah itu dimodifikasi atau ditransformasikan dengan penerapan aturan. •
  + Data dimasukkan atau dimuat ke dalam sistem target. •
  + Proses ETL dapat mengekstrak data dari berbagai sumber dan mengubahnya untuk dimuat ke dalam sistem target tunggal.
* Gudang Data (Data Warehouses) •
  + Sebuah gudang data (data warehouse) adalah pusat, repositori perusahaan. •
  + Data warehouse banyak digunakan oleh BI untuk menjalankan berbagai kueri analitik, dan biasanya berinteraksi dengan sistem OLAP untuk mendukung kueri analitik multidimensi yang terdiri dari data historis dan data saat ini. •
  + Batch memuat data secara berkala ke dalam data warehouse dari sistem operasional seperti ERP, CRM, dan SCM.
  + 
  + Data yang berkaitan dengan beberapa entitas bisnis dari sistem operasional yang berbeda secara berkala diekstraksi, divalidasi, diubah dan dikonsolidasikan ke dalam satu database denormalisasi. •
  + Dengan impor data berkala dari seluruh perusahaan, jumlah data yang terkandung dalam data warehouse tertentu akan terus meningkat. •
  + Hal ini menyebabkan waktu respons kueri yang lebih lambat untuk tugas analisis data. •
  + Untuk mengatasi kekurangan ini, data warehouse biasanya berisi database yang dioptimalkan, yang disebut database analitis, untuk menangani tugas pelaporan dan analisis data. •
  + Database analitik dapat eksis sebagai DBMS terpisah, seperti dalam kasus database OLAP.
* Data Marts •
  + Data mart adalah bagian dari data yang disimpan di data warehouse yang biasanya dimiliki oleh departemen, divisi, atau lini bisnis tertentu. •
  + Data warehouses dapat memiliki beberapa data mart. •
  + Data seluruh perusahaan dikumpulkan dan entitas bisnis kemudian diekstraksi. •
  + Entitas khusus domain disimpan ke dalam data warehouse melalui proses ETL.
  + Versi tunggal "kebenaran“ data warehouse didasarkan pada data yang dibersihkan, yang merupakan prasyarat untuk laporan yang akurat dan bebas kesalahan, sesuai dengan output yang ditampilkan di sebelah kanan.
  + 
  + Enterprise Data Warehouse Customer Relationship Management Enterprise Resource Planning Supply Chain Management Extract Transform Load
* BI tradisional •
  + BI tradisional terutama menggunakan analisis deskriptif dan diagnostik untuk memberikan informasi tentang peristiwa historis dan terkini. •
  + Ini bukan "cerdas" karena hanya memberikan jawaban atas pertanyaan yang dirumuskan dengan benar. •
  + Merumuskan pertanyaan dengan benar membutuhkan pemahaman tentang masalah dan isu bisnis dan data itu sendiri. •
  + BI melaporkan KPI yang berbeda melalui:
    - ✓laporan ad-hoc
    - ✓dashboards
  + Pelaporan ad-hoc adalah proses yang melibatkan pemrosesan data secara manual untuk menghasilkan laporan yang dibuat khusus. •
  + Fokus laporan ad-hoc biasanya pada area bisnis tertentu, seperti pemasaran atau manajemen rantai pasokannya. •
  + Laporan ad-hoc yang dihasilkan terperinci dan sering bersifat tabular. •
  + Sumber data OLAP dan OLTP dapat digunakan oleh BI untuk pelaporan dan dasbor ad-hoc.
  + 
* Dashboards •
  + Dasbor memberikan pandangan menyeluruh tentang area bisnis utama. •
  + Informasi yang ditampilkan di dasbor dihasilkan secara berkala dalam real-time atau near real-time. •
  + Penyajian data di dasbor bersifat grafis, menggunakan bar charts, pie charts, dan gauges.
  + BI menggunakan OLAP dan OLTP untuk menampilkan informasi di dasbor. 
  + Data warehouses dan data mart berisi informasi yang terkonsolidasi dan tervalidasi tentang entitas bisnis di seluruh perusahaan. •
  + BI tradisional tidak dapat berfungsi secara efektif tanpa data mart karena berisi data yang dioptimalkan dan dipisahkan, yang diperlukan BI untuk tujuan pelaporan. •
  + Tanpa data mart, data perlu diekstraksi dari data warehouse melalui proses ETL secara ad-hoc setiap kali kueri perlu dijalankan. •
  + Hal ini meningkatkan waktu dan upaya untuk mengeksekusi kueri dan menghasilkan laporan.
  + BI tradisional menggunakan data warehouse dan data mart untuk pelaporan dan analisis data karena memungkinkan kueri analisis data yang kompleks dengan beberapa operasi join dan agregasi untuk diterbitkan. •
  + Gambar di bawah ini contoh BI tradisional.
  + 
* Big Data BI •
  + Big Data BI dibangun di atas BI tradisional dengan bertindak berdasarkan data seluruh perusahaan yang telah dibersihkan dan dikonsolidasikan di data warehouse dan menggabungkannya dengan sumber data semiterstruktur dan tidak terstruktur. •
  + Terdiri dari analisis prediktif dan preskriptif untuk memfasilitasi pengembangan pemahaman kinerja bisnis di seluruh perusahaan. •
  + Sementara analisis BI tradisional umumnya berfokus pada proses bisnis individu, analisis BI Big Data berfokus pada beberapa proses bisnis secara bersamaan. •
  + Membantu mengungkapkan pola dan anomali di seluruh lingkup yang lebih luas di dalam perusahaan. •
  + Mengarah pada penemuan data dengan mengidentifikasi wawasan dan informasi yang mungkin sebelumnya tidak ada atau tidak diketahui.
  + Big Data BI memerlukan analisis data tidak terstruktur, semi terstruktur, dan terstruktur yang berada di data warehouse perusahaan. •
  + Memerlukan “next-generation” data warehouse yang menggunakan fitur dan teknologi baru untuk menyimpan data bersih yang berasal dari berbagai sumber dalam satu format data seragam. •
  + Penggabungan data warehouse tradisional dengan teknologi baru ini menghasilkan data warehouse hibrid. •
  + Data warehouse ini bertindak sebagai gudang seragam dan pusat dari data terstruktur, semi-terstruktur, dan tidak terstruktur yang dapat menyediakan alat Big Data BI dengan semua data yang diperlukan. •
  + Ini menghilangkan kebutuhan alat Big Data BI untuk terhubung ke beberapa sumber data untuk mengambil atau mengakses data.
  + Next-generation data warehouse menetapkan lapisan akses data standar di berbagai sumber data.
  + 
* Traditional Data Visualization •
  + Visualisasi data adalah teknik di mana hasil analisis dikomunikasikan secara grafis menggunakan elemen seperti bagan (charts), peta (map), kisi data (data grids), infografis (infographics), dan peringatan (alerts). •
  + Merepresentasikan data secara grafis dapat mempermudah untuk memahami laporan, melihat tren, dan mengidentifikasi pola. •
  + Visualisasi data tradisional sebagian besar menyediakan bagan dan grafik statis dalam laporan dan dasbor, sedangkan alat visualisasi data kontemporer bersifat interaktif dan dapat memberikan tampilan data yang diringkas dan terperinci. •
  + Hal ini dirancang untuk membantu orang-orang yang tidak memiliki keterampilan statistik dan/atau matematika untuk lebih memahami hasil analisis tanpa harus menggunakan spreadsheet. •
  + Alat visualisasi data tradisional meminta data dari database relasional, sistem OLAP, gudang data, dan spreadsheet untuk menyajikan hasil analisis deskriptif dan diagnostik.
* Visualisasi Data untuk Big Data •
  + Solusi Big Data memerlukan alat visualisasi data yang dapat terhubung dengan mulus ke sumber data terstruktur, semi-terstruktur, dan tidak terstruktur dan selanjutnya mampu menangani jutaan catatan data. •
  + Alat visualisasi data untuk solusi Big Data umumnya menggunakan teknologi analitik dalam memori yang mengurangi latensi yang biasanya dikaitkan dengan alat visualisasi data tradisional berbasis disk. •
  + Alat visualisasi data canggih untuk solusi Big Data menggabungkan analisis data prediktif dan preskriptif serta fitur transformasi data. •
  + Alat-alat ini menghilangkan kebutuhan akan metode pra-pemrosesan data, seperti ETL.
  + Alat ini juga menyediakan kemampuan untuk terhubung langsung ke sumber data terstruktur, semi-terstruktur, dan tidak terstruktur. •
  + Sebagai bagian dari solusi Big Data, alat visualisasi data tingkat lanjut dapat menggabungkan data terstruktur dan tidak terstruktur yang disimpan dalam memori untuk akses data yang cepat. •
  + Kueri dan rumus statistik dapat diterapkan sebagai bagian dari berbagai tugas analisis data untuk menampilkan data dalam format yang mudah digunakan, seperti di dasbor.
  + Fitur umum alat visualisasi yang digunakan dalam Big Data:
    - Aggregation – memberikan pandangan holistik dan ringkasan data di berbagai konteks. •
    - Drill-down – memungkinkan tampilan mendetail dari data yang diinginkan dengan berfokus pada subset data dari tampilan ringkasan. •
    - Filtering – membantu fokus pada kumpulan data tertentu dengan menyaring data yang tidak menarik secara langsung. •
    - Roll-up – mengelompokkan data di beberapa kategori untuk menampilkan subtotal dan total. •
    - What-if analysis – memungkinkan beberapa hasil divisualisasikan dengan memungkinkan faktor-faktor terkait diubah secara dinamis.
* Konsep Pemrosesan Big Data BIG DATA – TK13025
* Pemrosesan Data Paralel •
  + Pemrosesan data paralel melibatkan eksekusi simultan dari beberapa sub-tugas yang secara kolektif terdiri dari tugas yang lebih besar. •
  + Tujuannya adalah untuk mengurangi waktu eksekusi dengan membagi satu tugas yang lebih besar menjadi beberapa tugas yang lebih kecil yang berjalan secara bersamaan. •
  + Meskipun pemrosesan data paralel dapat dicapai melalui beberapa mesin jaringan, hal ini lebih sering dicapai dalam batas-batas satu mesin dengan beberapa prosesor atau cores.
  + 
  + Sebuah tugas dapat dibagi menjadi tiga sub-tugas yang dieksekusi secara paralel pada tiga prosesor berbeda dalam mesin yang sama
  + Pemrosesan data terdistribusi terkait erat dengan pemrosesan data paralel di mana prinsip yang sama "divide-and-conquer" diterapkan. •
  + Pemrosesan data terdistribusi selalu dicapai melalui mesin yang terpisah secara fisik yang terhubung ke jaringan bersama sebagai sebuah cluster.
  + 
  + Contoh pemrosesan data terdistribusi
  + Tugas dibagi menjadi tiga sub-tugas yang kemudian dieksekusi pada tiga mesin berbeda yang berbagi satu sakelar fisik.
* Hadoop •
  + Hadoop adalah open-source framework untuk penyimpanan data skala besar dan pemrosesan data yang kompatibel dengan perangkat keras komoditas. • Hadoop framework telah memantapkan diri sebagai platform industri de facto untuk solusi Big Data kontemporer. •
  + Hadoop framework dapat digunakan sebagai mesin ETL atau sebagai mesin analitik untuk memproses sejumlah besar data terstruktur, semi-terstruktur, dan tidak terstruktur. •
  + Dari perspektif analisis, Hadoop mengimplementasikan kerangka kerja pemrosesan MapReduce.
  + Gambar di bawah ini mengilustrasikan beberapa fitur Hadoop.
  + 
  + Hadoop adalah kerangka kerja serbaguna yang menyediakan kemampuan pemrosesan dan penyimpanan.
* Memproses Beban Kerja (Workloads) •
  + Pemrosesan beban kerja (workloads) dalam Big Data didefinisikan sebagai jumlah dan sifat data yang diproses dalam jangka waktu tertentu. •
  + Beban kerja (workloads) dibagi menjadi dua jenis: •
    - Batch •
    - transactional
  + Batch •
    - Pemrosesan batch dikenal sebagai pemrosesan offline, melibatkan pemrosesan data dalam batch dan biasanya menimbulkan penundaan, yang menghasilkan respons latensi tinggi. •
    - Beban kerja batch biasanya melibatkan data dalam jumlah besar dengan baca/tulis berurutan dan terdiri dari grup kueri baca atau tulis. •
    - Kueri bisa rumit dan melibatkan banyak operasi join. •
    - Sistem OLAP memproses beban kerja dalam batch. •
    - BI dan analitik strategis berorientasi pada batch karena merupakan tugas yang sangat intensif membaca yang melibatkan volume data yang besar.
    - 
    - Beban kerja batch terdiri dari baca /tulis yang dikelompokkan yang memiliki jejak data besar dan mungkin berisi operasi join yang kompleks dan memberikan respons latensi tinggi .
    - Beban kerja batch dapat mencakup baca /tulis yang dikelompokkan sebagai INSERT, SELECT, UPDATE, dan DELETE.
  + Transaksional •
    - Pemrosesan transaksional juga dikenal sebagai pemrosesan online. •
    - Pemrosesan beban kerja transaksional mengikuti pendekatan di mana data diproses secara interaktif tanpa penundaan, menghasilkan respons latensi rendah. •
    - Beban kerja transaksi melibatkan sejumlah kecil data dengan pembacaan dan penulisan acak. •
    - OLTP dan sistem operasional, yang umumnya intensif menulis, termasuk dalam kategori ini. •
    - Meskipun beban kerja ini berisi campuran kueri baca/tulis, umumnya lebih intensif menulis daripada membaca. •
    - Beban kerja transaksional terdiri dari pembacaan/penulisan acak yang melibatkan lebih sedikit operasi join daripada intelijen bisnis dan beban kerja pelaporan.
    - 
    - Beban kerja transaksional memiliki sedikit operasi join dan respons latensi lebih rendah daripada beban kerja batch.
* Cluster •
  + Cluster memberikan dukungan yang diperlukan untuk membuat solusi penyimpanan yang dapat diskalakan secara distri, cluster juga menyediakan mekanisme untuk memungkinkan pemrosesan data terdistribusi dengan skalabilitas linier. •
  + Karena cluster sangat skalabel, cluster menyediakan lingkungan yang ideal untuk pemrosesan Big Data karena kumpulan data besar dapat dibagi menjadi kumpulan data yang lebih kecil dan kemudian diproses secara paralel terdistribusi. •
  + Saat memanfaatkan cluster, kumpulan data Big Data dapat diproses dalam mode batch atau mode real-time. •
  + Idealnya, sebuah cluster akan terdiri dari node komoditas berbiaya rendah yang secara kolektif memberikan peningkatan kapasitas pemrosesan.
  + Sebuah cluster dapat digunakan untuk mendukung pemrosesan batch data massal dan pemrosesan data streaming secara real-time.
  + 
  + Manfaat tambahan dari cluster adalah menyediakan redundansi yang melekat dan toleransi kesalahan, karena terdiri dari node yang terpisah secara fisik. •
  + Redundansi dan toleransi kesalahan memungkinkan pemrosesan dan analisis yang tangguh terjadi jika terjadi kegagalan jaringan atau node. •
  + Memanfaatkan layanan infrastruktur cloud-host atau lingkungan analitik yang siap pakai sebagai tulang punggung sebuah cluster diperbolehkan karena elastisitas dan model komputasi berbasis utilitas bayar-untuk-penggunaan.
* Memproses dalam Mode Batch •
  + Data diproses secara offline dalam batch dan waktu respons dapat bervariasi dari menit ke jam. •
  + Data harus disimpan ke disk sebelum dapat diproses. •
  + Mode batch umumnya melibatkan pemrosesan berbagai kumpulan data besar, sendiri-sendiri atau digabungkan, pada dasarnya menangani karakteristik volume dan variasi kumpulan data Big Data. •
  + Mayoritas pemrosesan Big Data terjadi dalam mode batch. •
  + Proses ini relatif sederhana, mudah diatur, dan berbiaya rendah dibandingkan dengan mode real-time. •
  + BI strategis, analisis prediktif dan preskriptif, serta operasi ETL umumnya berorientasi batch.
* Memproses dalam Mode Realtime •
  + Dalam mode real-time, data diproses dalam memori sebelum disimpan ke disk. •
  + Waktu respons umumnya berkisar dari sub-detik hingga di bawah satu menit. •
  + Mode real-time membahas karakteristik velocity kumpulan data Big Data. •
  + Pemrosesan real-time juga disebut pemrosesan event atau stream karena data tiba secara terus-menerus (stream) atau pada interval (event).
  + Datum event/stream individu umumnya berukuran kecil, tetapi sifat kontinunya menghasilkan kumpulan data yang sangat besar. •
  + Mode interaktif, termasuk dalam kategori real-time. •
  + Mode interaktif umumnya mengacu pada pemrosesan kueri secara real-time. •
  + BI/analitik operasional umumnya dilakukan dalam mode real-time. •
  + Prinsip dasar yang terkait dengan pemrosesan Big Data disebut prinsip Speed, Consistency, dan Volume (SCV).
  + Speed Consistency Volume (SCV) •
    - Speed (kecepatan) – Kecepatan mengacu pada seberapa cepat data dapat diproses setelah dihasilkan. •
    - Consistency (konsistensi) – Konsistensi mengacu pada akurasi dan ketepatan hasil. •
    - Volume – Volume mengacu pada jumlah data yang dapat diproses.
    - 
    - Jika kecepatan (S) dan konsistensi (C) diperlukan, tidak mungkin memproses data dalam jumlah besar (V) karena sejumlah besar data memperlambat pemrosesan data.
    - Jika konsistensi (C) dan pemrosesan data dengan volume tinggi (V) diperlukan, tidak mungkin untuk memproses data dengan kecepatan tinggi (S) karena untuk mencapai pemrosesan data kecepatan tinggi memerlukan volume data yang lebih kecil.
    - Jika pemrosesan data volume tinggi (V) dan pemrosesan data kecepatan tinggi (S) diperlukan, hasil yang diproses tidak akan konsisten (C) karena pemrosesan data dalam jumlah besar dengan kecepatan tinggi melibatkan pengambilan sampel data, yang dapat mengurangi konsistensi
    - Perlu dicatat bahwa pilihan dua dari tiga dimensi untuk mendukung tercapainya tujuan sepenuhnya tergantung pada persyaratan sistem lingkungan analisis

Bab 6 – Konsep Penyimpanan

* Clusters •
  + Cluster adalah kumpulan server, atau node yang digabungkan. •
  + Server ini biasanya memiliki spesifikasi perangkat keras yang sama dan terhubung bersama melalui jaringan untuk bekerja sebagai satu unit. •
  + Setiap node dalam cluster memiliki sumber daya tersendiri, seperti memori, prosesor, dan hard drive. •
  + Sebuah cluster dapat menjalankan tugas dengan membaginya menjadi potongan-potongan kecil dan mendistribusikan eksekusinya ke komputer yang berbeda milik cluster.
  + 
  + Simbol yang digunakan untuk mewakili sebuah cluster.
* Sistem File dan Sistem File Terdistribusi •
  + Sistem file adalah metode menyimpan dan mengatur data pada perangkat penyimpanan, seperti flash drive, DVD, dan hard drive. •
  + File adalah unit penyimpanan atom yang digunakan oleh sistem file untuk menyimpan data. •
  + Sistem file menyediakan tampilan logis dari data yang disimpan di perangkat penyimpanan dan menyajikannya sebagai struktur pohon direktori dan file. •
  + Sistem operasi menggunakan sistem file untuk menyimpan dan mengambil data atas nama aplikasi. •
  + Setiap sistem operasi menyediakan dukungan untuk satu atau lebih sistem file, misalnya NTFS di Microsoft Windows dan ext di Linux.
  + Sistem file terdistribusi adalah sistem file yang dapat menyimpan file besar yang tersebar di seluruh node cluster. •
  + Bagi klien, file tampak lokal; namun, ini hanya pandangan logis karena secara fisik file didistribusikan ke seluruh cluster. •
  + Tampilan lokal ini disajikan melalui sistem file terdistribusi dan memungkinkan file diakses dari berbagai lokasi. •
  + Contohnya termasuk Google File System (GFS) dan Hadoop Distributed File System (HDFS).
  + 
  + Simbol yang digunakan untuk mewakili sistem file terdistribusi.
* NoSQL •
  + Database Not-only SQL (NoSQL) adalah database non-relasional yang sangat skalabel, toleran terhadap kesalahan, dan dirancang khusus untuk menampung data semi terstruktur dan tidak terstruktur. •
  + Basis data NoSQL sering kali menyediakan antarmuka kueri berbasis API yang dapat dipanggil dari dalam aplikasi. •
  + Basis data NoSQL juga mendukung bahasa kueri selain Bahasa Kueri Terstruktur (SQL) karena SQL dirancang untuk kueri data terstruktur yang disimpan dalam basis data relasional. •
  + Database NoSQL yang dioptimalkan untuk menyimpan file XML akan sering menggunakan XQuery sebagai bahasa kueri. •
  + Ada beberapa database NoSQL yang juga menyediakan antarmuka kueri seperti SQL.
  + 
  + Basis data NoSQL dapat menyediakan antarmuka kueri seperti API atau SQL.
* Sharding •
  + Sharding adalah proses mempartisi secara horizontal kumpulan data besar menjadi kumpulan kumpulan data yang lebih kecil dan lebih mudah dikelola yang disebut shard.
  + Shard didistribusikan di beberapa node, di mana node adalah server atau mesin. •
  + Setiap shard disimpan pada node yang terpisah dan setiap node hanya bertanggung jawab atas data yang tersimpan di dalamnya. •
  + Setiap shard berbagi skema yang sama, dan semua shard secara kolektif mewakili kumpulan data lengkap.
  + 
  + Contoh sharding di mana dataset tersebar di Node A dan Node B, menghasilkan Shard A dan Shard B.
  + Sharding sering kali transparan bagi klien, tetapi ini bukan keharusan. •
  + Penskalaan horizontal adalah metode untuk meningkatkan kapasitas sistem dengan menambahkan sumber daya berkapasitas serupa atau lebih tinggi di samping sumber daya yang ada. •
  + Karena setiap node hanya bertanggung jawab untuk sebagian dari keseluruhan dataset, waktu baca/tulis sangat meningkat.
  + 
  + Contoh sharding di mana data diambil dari Node A dan Node B.
  + Setiap shard dapat secara mandiri melayani membaca dan menulis untuk subset data tertentu yang menjadi tanggung jawabnya.
  + Bergantung pada kueri, data mungkin perlu diambil dari kedua pecahan (shards).
  + Manfaat sharding adalah memberikan toleransi parsial terhadap kegagalan. •
  + Jika terjadi kegagalan node, hanya data yang disimpan di node tersebut yang terpengaruh. •
  + Berkenaan dengan partisi data, pola kueri perlu diperhitungkan sehingga shard itu sendiri tidak menjadi hambatan kinerja. •
  + Misalnya, kueri yang memerlukan data dari beberapa shard akan memberlakukan penalti kinerja. •
  + Lokalitas data membuat data yang biasa diakses berada di lokasi yang sama pada satu shard dan membantu mengatasi masalah kinerja tersebut.
* Replikasi (Replication) •
  + Replikasi menyimpan banyak salinan dari kumpulan data, yang dikenal sebagai replica. •
  + Replikasi menyediakan skalabilitas dan ketersediaan karena fakta bahwa data yang sama direplikasi pada berbagai node. •
  + Toleransi kesalahan juga tercapai karena redundansi data memastikan bahwa data tidak hilang ketika node individu gagal. •
  + Ada dua metode berbeda yang digunakan untuk mengimplementasikan replikasi:
    - ✓master-slave
    - ✓peer-to-peer
  + 
  + Contoh replikasi di mana kumpulan data direplikasi ke Node A dan Node B, menghasilkan Replica A dan Replica B.
  + Master-Slave •
    - Selama replikasi master-slave, node diatur dalam konfigurasi master-slave, dan semua data ditulis ke node master. Setelah disimpan, data direplikasi ke beberapa node slave. •
    - Semua permintaan tulis eksternal, termasuk insert, update, dan delete, terjadi pada node master, sedangkan permintaan select (baca) dapat dipenuhi oleh node slave mana pun. •
    - Contoh replikasi master-slave di mana Master A adalah titik kontak tunggal untuk semua penulisan, dan data dapat dibaca dari Slave A dan Slave B. •
    - Replikasi master-slave ideal untuk beban intensif baca daripada beban intensif tulis karena permintaan baca yang meningkat dapat dikelola dengan penskalaan horizontal untuk menambahkan lebih banyak node slave
    - Penulisan konsisten, karena semua penulisan dikoordinasikan oleh node master.
    - Implikasinya adalah kinerja penulisan akan menurun seiring dengan peningkatan jumlah penulisan. •
    - Jika master node gagal, pembacaan masih dapat dilakukan melalui salah satu node slave
    - Sebuah node slave dapat dikonfigurasi sebagai node cadangan untuk node master. •
    - Jika master node gagal, penulisan tidak didukung sampai master node dibuat kembali. •
    - Node master dibangkitkan dari cadangan node master, atau node master baru dipilih dari node slave.
    - Salah satu masalah dengan replikasi master-slave adalah inkonsistensi pembacaan, yang dapat menjadi masalah jika node slave dibaca sebelum pembaruan master disalin ke sana. •
    - Untuk memastikan konsistensi pembacaan, sistem pemungutan suara dapat diterapkan di mana pembacaan dinyatakan konsisten jika mayoritas slave berisi versi rekaman yang sama. •
    - Implementasi sistem pemungutan suara semacam itu membutuhkan mekanisme komunikasi yang andal dan cepat antara para slave.
    - 
    - Contoh replikasi master-slave di mana terjadi inkonsistensi baca.
    - Skenario di mana inkonsistensi baca terjadi:
      * 1. Pengguna A memperbarui data.
      * 2. Data disalin ke Slave A oleh Master.
      * 3. Sebelum data disalin ke Slave B, Pengguna B mencoba membaca data dari Slave B, yang menghasilkan pembacaan yang tidak konsisten.
      * 4. Data pada akhirnya akan menjadi konsisten ketika Slave B diperbarui oleh Master.
  + Peer-to-Peer •
    - Dengan replikasi peer-to-peer, semua node beroperasi pada level yang sama. •
    - Dengan kata lain, tidak ada hubungan master-slave antara node. •
    - Setiap node, yang dikenal sebagai peer, sama-sama mampu menangani membaca dan menulis.
    - Penulisan disalin ke peer A, B dan C secara bersamaan.
    - 
    - Data dibaca dari Peer A, tetapi juga dapat dibaca dari Peer B atau C.
    - Replikasi peer-to-peer cenderung inkonsistensi operasi tulis yang terjadi sebagai akibat dari pembaruan simultan dari data yang sama di beberapa peer. •
    - Ini dapat diatasi dengan menerapkan strategi konkurensi pesimis atau optimis. •
      * Konkurensi pesimis adalah strategi proaktif yang mencegah inkonsistensi. •
      * Konkurensi optimis adalah strategi reaktif yang tidak menggunakan penguncian.
  + Sharding dan Replication •
    - Sharding dan replikasi dapat digabungkan untuk meningkatkan toleransi kesalahan terbatas yang ditawarkan oleh sharding, mendapatkan manfaat tambahan dari peningkatan ketersediaan dan skalabilitas replikasi. •
    - Perbandingan sharding dan replikasi yang menunjukkan bagaimana dataset didistribusikan antara dua node dengan aplikasi yang berbeda
    - 
  + Menggabungkan Sharding dan Replikasi Master-Slave •
    - Saat sharding digabungkan dengan replikasi master-slave, beberapa shard menjadi slave dari satu master, dan master itu sendiri adalah shard. •
    - Meskipun menghasilkan banyak master, satu slave-shard hanya dapat dikelola oleh satu master-shard. •
    - Konsistensi penulisan dipertahankan oleh master shard. •
    - Jika master-shard menjadi non-operasional atau terjadi pemadaman jaringan, toleransi kesalahan terkait dengan operasi penulisan akan terpengaruh. •
    - Replika shard disimpan di beberapa node slave untuk memberikan skalabilitas dan toleransi kesalahan untuk operasi baca.
    - Setiap node bertindak sebagai master dan slave untuk shard yang berbeda. •
    - 
    - Menulis (id = 2) ke Shard A diatur oleh Node A, karena ini adalah master untuk Shard A.
    - A mereplikasi data (id = 2) ke Node B, yang merupakan slave untuk Shard A. •
    - Pembacaan (id = 4) dapat dilayani langsung oleh Node B atau Node C karena masing-masing berisi Shard B. •
    - Gambar di samping menunjukkan contoh yang menunjukkan kombinasi sharding dan replikasi master-slave.
  + Teorema CAP •
    - Teorema Consistency, Availability, and Partition tolerance (CAP), juga dikenal sebagai teorema Brewer, mengungkapkan tiga kendala yang terkait dengan sistem basis data terdistribusi. •
    - Ini menyatakan bahwa sistem database terdistribusi, berjalan di sebuah cluster, hanya dapat menyediakan dua dari tiga properti berikut: •
      * Konsistensi – Pembacaan dari node mana pun menghasilkan data yang sama di beberapa node. •
      * Ketersediaan – Permintaan baca/tulis akan selalu diakui dalam bentuk keberhasilan atau kegagalan.
      * Konsistensi: ketiga pengguna mendapatkan nilai yang sama untuk kolom jumlah meskipun tiga node berbeda melayani permintaan.
      * 
      * Ketersediaan dan toleransi partisi: jika terjadi kegagalan komunikasi, permintaan dari kedua pengguna masih dilayani (1, 2).
      * Namun, dengan Pengguna B, pembaruan gagal karena catatan dengan id = 3 belum disalin ke Peer C.
      * Pengguna diberi tahu (3) bahwa pembaruan telah gagal.
      * 
      * Skenario berikut menunjukkan mengapa hanya dua dari tiga sifat teorema CAP yang dapat didukung secara bersamaan. •
      * 
      * Diagram Venn yang merangkum teorema CAP.
      * Jika konsistensi (consistency - C) dan ketersediaan (availability - A) diperlukan, node yang tersedia perlu berkomunikasi untuk memastikan konsistensi (consistency - C). •
      * Oleh karena itu, toleransi partisi (partition tolerance - P) tidak dimungkinkan. •
      * Jika konsistensi (consistency - C) dan toleransi partisi (partition tolerance - P) diperlukan, node tidak dapat tetap tersedia (availability - A) karena node akan menjadi tidak tersedia saat mencapai keadaan konsistensi (consistency - C). •
      * Jika ketersediaan (availability - A) dan toleransi partisi (partition tolerance - P) diperlukan, maka konsistensi (consistency - C) tidak dimungkinkan karena kebutuhan komunikasi data antar node. •
      * Jadi, database dapat tetap tersedia (availability - A) tetapi dengan hasil yang tidak konsisten.
      * Dalam database terdistribusi, skalabilitas dan toleransi kesalahan dapat ditingkatkan melalui node tambahan, meskipun ini menantang consistency (C). •
      * Penambahan node juga dapat menyebabkan availability (A) terganggu karena latency yang disebabkan oleh peningkatan komunikasi antar node.
      * Sistem database terdistribusi tidak dapat 100% toleran terhadap partisi (P). •
      * Meskipun gangguan komunikasi jarang terjadi dan bersifat sementara, toleransi partisi (P) harus selalu didukung oleh database terdistribusi; oleh karena itu, CAP umumnya merupakan pilihan antara memilih C+P atau A+P. •
      * Persyaratan sistem akan menentukan mana yang dipilih.
  + ACID •
  + ACID adalah prinsip desain database yang terkait dengan manajemen transaksi. •
  + ACID adalah akronim yang berarti: •
    - Atomisitas (atomicity) •
    - Konsistensi (consistency) •
    - Isolasi (isolation) •
    - Daya tahan (durability) ACID •
  + ACID adalah gaya manajemen transaksi yang memanfaatkan kontrol konkurensi pesimis untuk memastikan konsistensi dipertahankan melalui penerapan kunci catatan. •
  + ACID adalah pendekatan tradisional untuk manajemen transaksi database karena dimanfaatkan oleh sistem manajemen database relasional. •
  + Atomicity memastikan bahwa semua operasi akan selalu berhasil atau gagal sepenuhnya. •
  + Tidak ada transaksi parsial.
  + Seorang pengguna mencoba untuk memperbarui tiga record sebagai bagian dari transaksi. •
  + Dua record berhasil diperbarui sebelum terjadinya kesalahan. •
  + Dua record berhasil sebelum terjadinya kesalahan.
  + 
  + Konsistensi memastikan bahwa database akan selalu tetap dalam keadaan konsisten dengan memastikan bahwa hanya data yang sesuai dengan batasan skema database yang dapat ditulis ke database. •
  + Dengan demikian database yang dalam keadaan konsisten akan tetap dalam keadaan konsisten setelah transaksi berhasil.
  + Pengguna mencoba memperbarui kolom jumlah tabel yang bertipe float dengan nilai varchar. •
  + Basis data menerapkan pemeriksaan validasi dan menolak pembaruan ini karena nilainya melanggar pemeriksaan batasan untuk kolom amount.
  + 
  + Pengguna A mencoba memperbarui dua record sebagai bagian dari transaksi. •
  + Basis data berhasil memperbarui record pertama. •
  + Namun, sebelum dapat memperbarui record kedua, Pengguna B mencoba memperbarui record yang sama. •
  + Basis data tidak mengizinkan pembaruan Pengguna B hingga pembaruan Pengguna A berhasil atau gagal sepenuhnya. •
  + Ini terjadi karena record dengan id 3 dikunci oleh database sampai transaksi selesai. ACID •
  + 
  + Seorang pengguna memperbarui record sebagai bagian dari transaksi. •
  + Basis data berhasil memperbarui record. •
  + Tepat setelah pembaruan ini, terjadi kegagalan daya. •
  + Basis data mempertahankan statusnya saat tidak ada daya. •
  + Listrik dihidupkan kembali. •
  + Basis data menyajikan record sesuai pembaruan terakhir saat diminta oleh pengguna.
  + Pengguna A mencoba memperbarui record sebagai bagian dari transaksi. •
  + Basis data memvalidasi nilai dan pembaruan berhasil diterapkan. •
  + Setelah transaksi berhasil diselesaikan, ketika Pengguna B dan C meminta record yang sama, database memberikan nilai yang diperbarui kepada kedua pengguna.
* BASE •
  + BASE adalah prinsip desain database berdasarkan teorema CAP dan dimanfaatkan oleh sistem database yang menggunakan teknologi terdistribusi. •
  + BASE adalah singkatan dari : •
    - basically available •
    - soft state •
    - eventual consistency
  + Ketika database mendukung BASE, database lebih menyukai ketersediaan daripada konsistensi. •
  + Dengan kata lain, database adalah A+P dari perspektif CAP. •
  + BASE memanfaatkan konkurensi optimis dengan melonggarkan batasan konsistensi kuat yang diamanatkan oleh properti ACID. •
  + Jika basis data “basically available”, basis data itu akan selalu mengakui permintaan klien, baik dalam bentuk data yang diminta atau pemberitahuan berhasil/gagal. •
  + Database pada dasarnya tersedia, meskipun telah dipartisi sebagai akibat dari kegagalan jaringan.
  + 
  + Basis data pada dasarnya tersedia, meskipun telah dipartisi sebagai akibat dari kegagalan jaringan. •
  + Pengguna A dan Pengguna B menerima data meskipun database dipartisi oleh kegagalan jaringan. •
  + Gambar di samping menunjukkan: •
    - Pengguna A memperbarui record pada peer A. •
    - Sebelum peer lain diperbarui, Pengguna B meminta record yang sama dari peer C. •
    - Basis data sekarang dalam keadaan soft state, dan data basi dikembalikan ke Pengguna B.
  + Soft state berarti bahwa database mungkin berada dalam keadaan tidak konsisten saat data dibaca; dengan demikian, hasilnya dapat berubah jika data yang sama diminta lagi. •
  + Ini karena data dapat diperbarui untuk konsistensi, meskipun tidak ada pengguna yang menulis ke database di antara dua pembacaan. •
  + Properti ini terkait erat dengan konsistensi akhir