UNTAR
Universitas Tarumanagara

Terakreditasi BAN PT A (Unggul)

UNTAR untuk INDONESIA

Big Data Konsep dan Alat

BIG DATA – TK13025 & SI23016

Teny Handhayani

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, yang menerima liputan pers yang substansial dan menarik minat yang meningkat baik dari pengguna bisnis maupun profesional TI.
- Hasilnya adalah 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.

Definisi Big Data

- Dari mana Big Data berasal?
- Jawaban sederhananya adalah “di mana-mana”.
- Big Data dapat berasal dari berbagai sumber:

- ✓ log Web
- ✓ jaringan sosial
- ✓ astronomi
- ✓ fisika nuklir
- ✓ pengawasan militer
- ✓ radio-frequency identification(RFID)
- ✓ dokumen teks berbasis Internet
- ✓ ilmu atmosfer
- ✓ eksperimen biokimia
- ✓ arsip fotografi
- ✓ global positioning systems (GPS)
- ✓ indeks pencarian Internet
- ✓ biologi
- ✓ catatan medis
- ✓ arsip video
- ✓ jaringan sensor
- ✓ catatan panggilan detail
- ✓ genomik
- ✓ penelitian ilmiah
- ✓ praktik e-commerce skala besar

Definisi Big Data

- 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 gudang data di awal 1990-an.
- Dulu terabyte identik dengan gudang 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 hingga 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 real-time.
- *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 peluang 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.

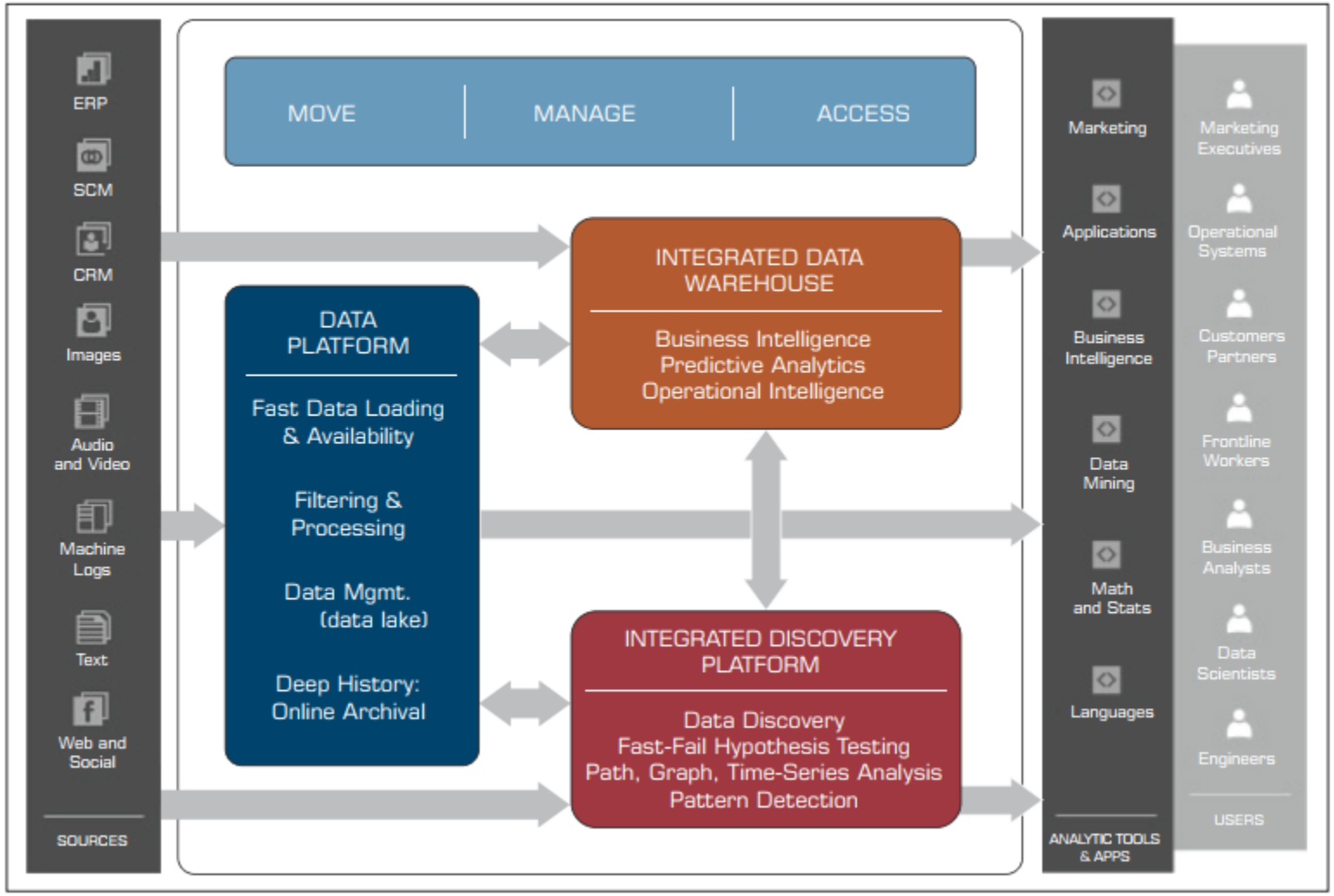
VARIABILITY

- 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.

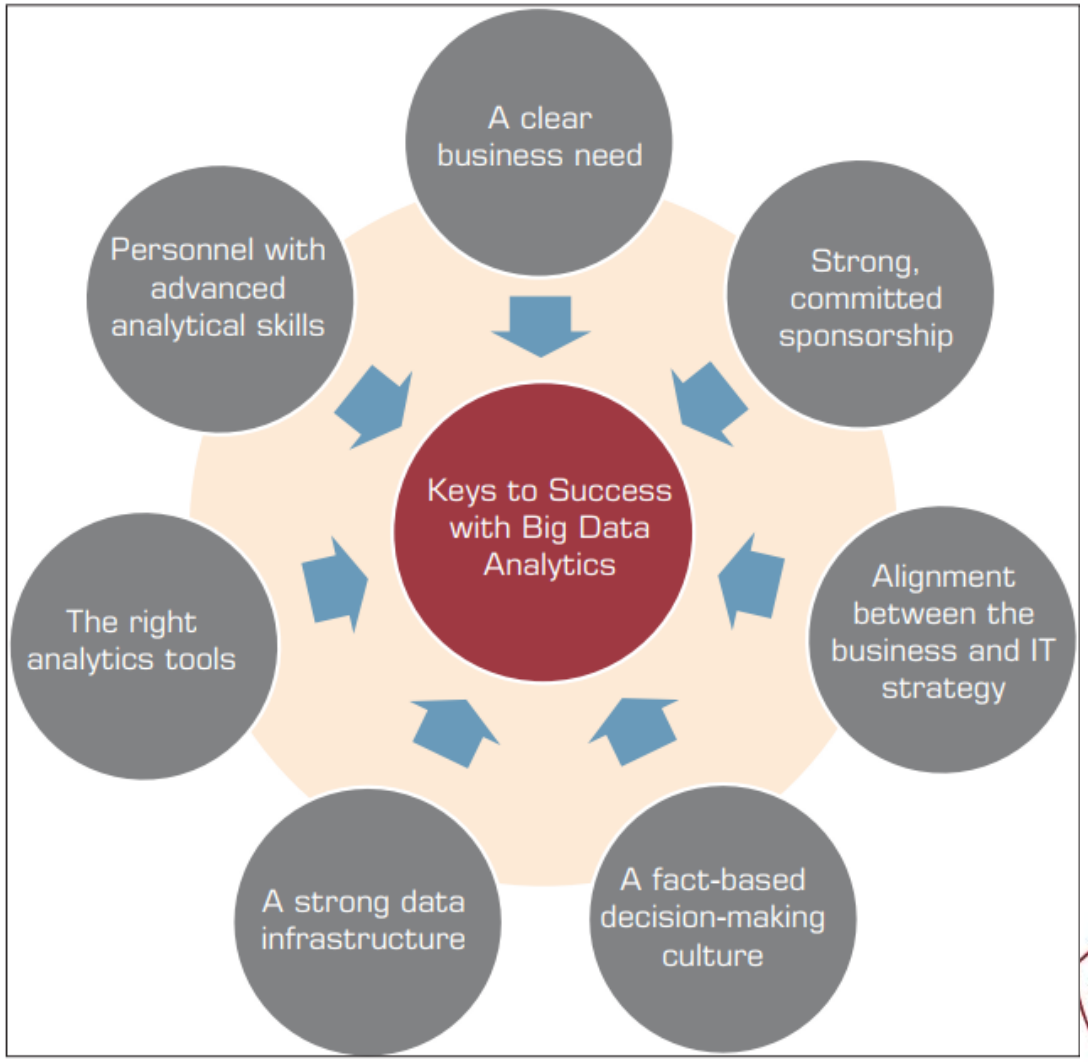
VALUE PROPOSITION

- 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



Faktor Keberhasilan Penting untuk Analisis 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 (jika 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 daripada 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

- Gudang data telah menyediakan infrastruktur data untuk analitik. Infrastruktur ini berubah dan ditingkatkan di era Big Data dengan teknologi baru.
- Sukses membutuhkan mengawinkan 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 untuk wawasan 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 di kumpulan sumber daya TI 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.

Tantangan Menerapkan Analitik Big Data

- 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 nilai paling banyak.
- 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.

Tantangan Menerapkan Analitik Big Data

- Skills availability:
  - Big Data sedang dimanfaatkan dengan alat-alat baru dan sedang dilihat dengan cara yang berbeda.
  - Ilmuwan data dengan keterampilan untuk melakukan pekerjaan itu.
- Skills availability:
  - 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
- Manajemen merek (*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

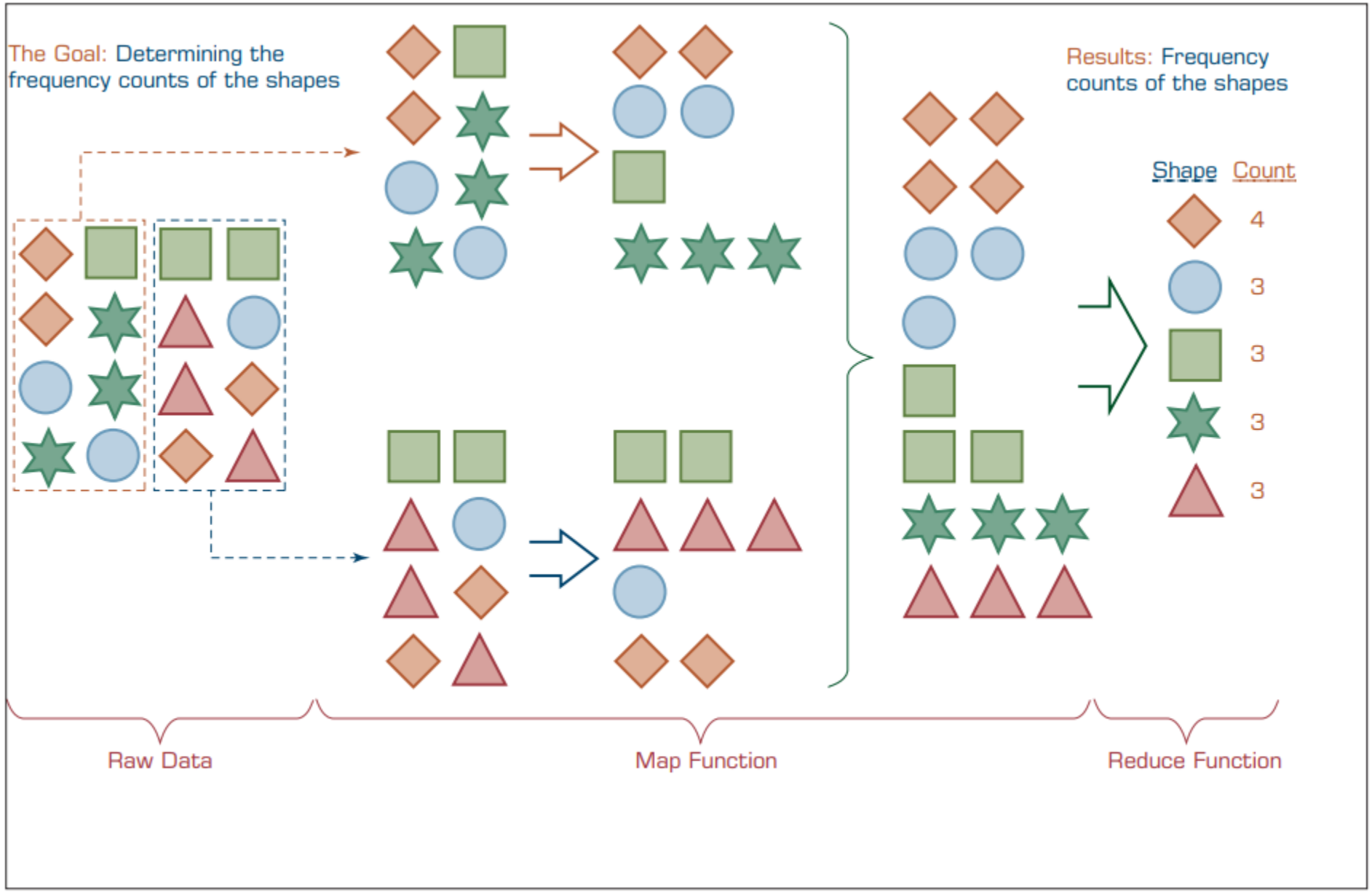
MapReduce

- MapReduce adalah teknik yang dipopulerkan oleh Google yang mendistribusikan pemrosesan file data multi-struktur yang sangat besar di sekelompok besar mesin.
- 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.


The Goal: Determining the frequency counts of the shapes
Results: Frequency counts of the shapes
Shape
Count
4
3
3
3
3
Raw Data
Map Function
Reduce Function

MapReduce

- Tujuannya adalah untuk menghitung jumlah kotak setiap warna.
- Programmer dalam bertanggung jawab untuk mengkodekan peta 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 peta yang berjalan secara paralel pada node cluster.
- Peran masing-masing program peta dalam hal ini adalah mengelompokkan data secara terpisah berdasarkan warna.
- Sistem MapReduce kemudian mengambil output dari setiap program peta 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.
- Ini juga memiliki keuntungan bahwa pengembang tidak perlu khawatir dengan penerapan komputasi paralel—ini 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

Komponen Teknis Hadoop

- Hadoop Distributed File System (HDFS):
  - Lapisan penyimpanan default di setiap cluster Hadoop tertentu.
- Name Node:
  - Node dalam kluster Hadoop yang menyediakan informasi klien tentang di mana dalam kluster 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:

Komponen Teknis Hadoop

- 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.
- Ini 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.

Bagaimana Hadoop Bekerja?

- *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 klaster data tersebut berada.
- Setelah ditentukan, Pelacak Pekerjaan 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.

Bagaimana Hadoop Bekerja?

- 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.

Bagaimana Hadoop Bekerja?

- 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 TI tradisional lainnya untuk analisis lebih lanjut dan/atau untuk mendukung pemrosesan transaksional.

Subprojek yang paling sering dirujuk untuk Hadoop

- ✓ HIVE
- ✓ PIG
- ✓ HBASE
- ✓ FLUME
- ✓ OOZIE
- ✓ AMBARI
- ✓ AVRO
- ✓ MAHOUT
- ✓ SQOOP
- ✓ 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 sebagian besar (meskipun ada beberapa pengecualian penting), 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.

Kasus untuk 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, hari 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.

Kasus untuk 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

TABLE 7.1 When to Use Which Platform—Hadoop versus DW

Requirement	Data Warehouse	Hadoop
Low latency, interactive reports, and OLAP	☑	
ANSI 2003 SQL compliance is required	☑	☑
Preprocessing or exploration of raw unstructured data		☑
Online archives alternative to tape		☑
High-quality cleansed and consistent data	☑	☑
100s to 1,000s of concurrent users	☑	☑
Discover unknown relationships in the data		☑
Parallel complex process logic	☑	☑
CPU intense analysis	☑	
System, users, and data governance		☑
Many flexible programming languages running in parallel		☑
Unrestricted, ungoverned sandbox explorations		☑
Analysis of provisional data	☑	
Extensive security and regulatory compliance	☑	☑

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 dari 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.

Vendor dan Platform Big Data

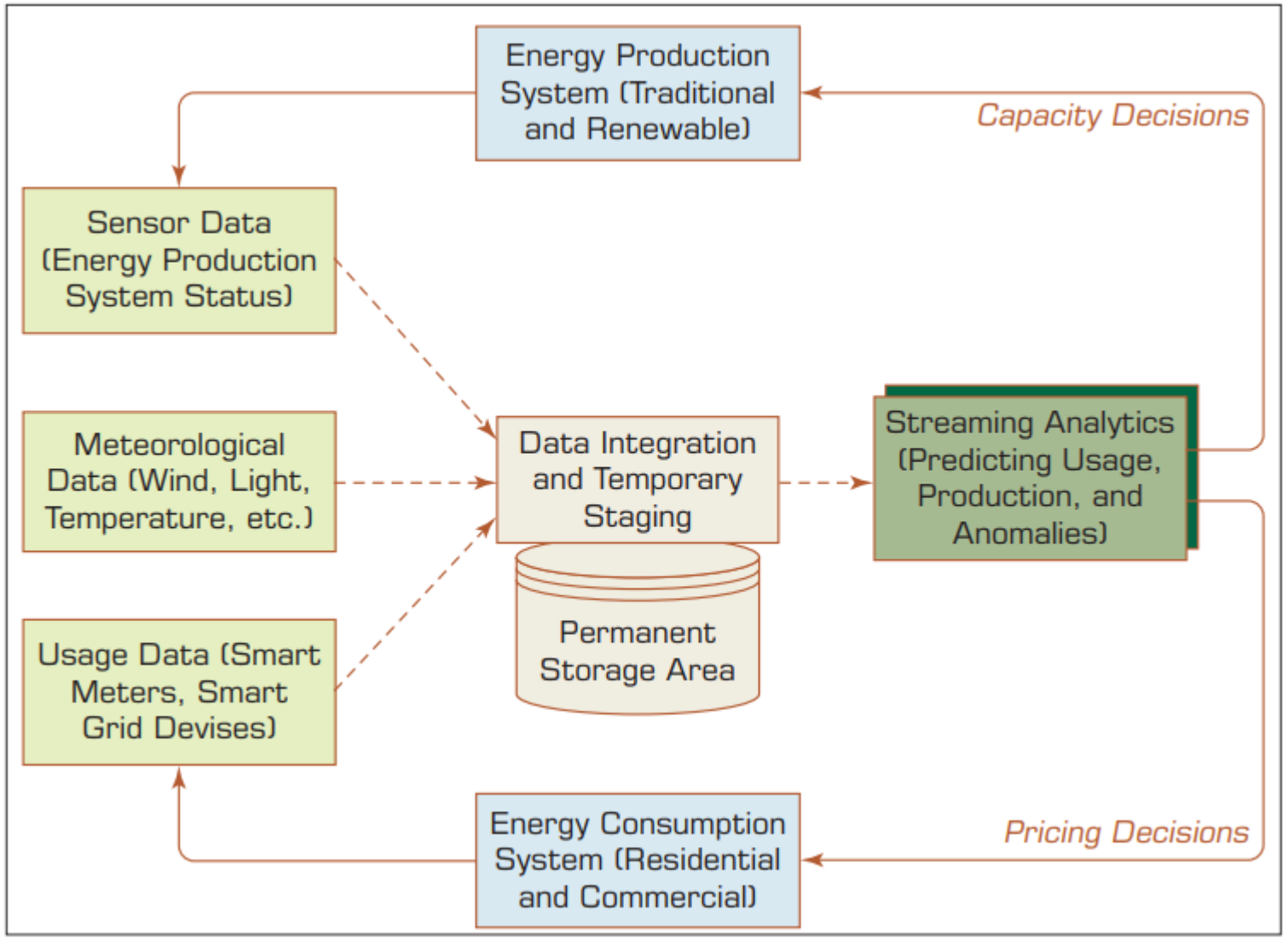
- IBM InfoSphere BigInsights
  - InfoSphere BigInsights IBM adalah platform berbasis open sourceProyek Apache Hadoop untuk menganalisis data terstruktur tradisional yang ditemukan dalam database lama bersama dengan data semi dan tidak terstruktur seperti teks, video, audio, gambar, media sosial, log Web, dan aliran klik.
- 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 Analisis Aliran

- Analisis aliran (juga disebut analitik data-dalam-gerak dan analitik data real-time, antara lain) 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 aliran 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
- Pemerintah