















UNTAR untuk INDONESIA

Enterprise Technologies and Big Data Business Intelligence

BIG DATA – TK13025









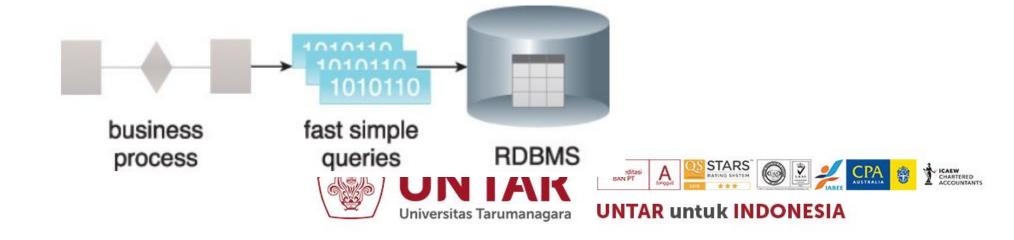
Online Transaction Processing (OLTP)

- 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.



Online Transaction Processing (OLTP)

- 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)

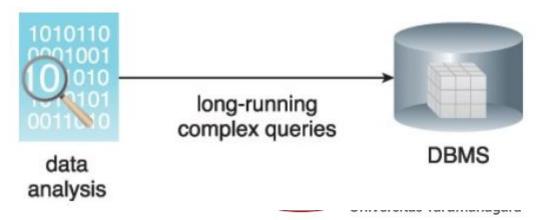
- Online Analytical Processing (OLAP) digunakan untuk memproses kueri analisis data.
- OLAP merupakan bagian integrasi dari business intelligence, penambangan data, dan proses pembelajaran mesin.
- Hal tersebut relevan dengan Big Data karena dapat berfungsi sebagai sumber data serta sink data yang mampu 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.





Online Analytical Processing (OLAP)

- 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)

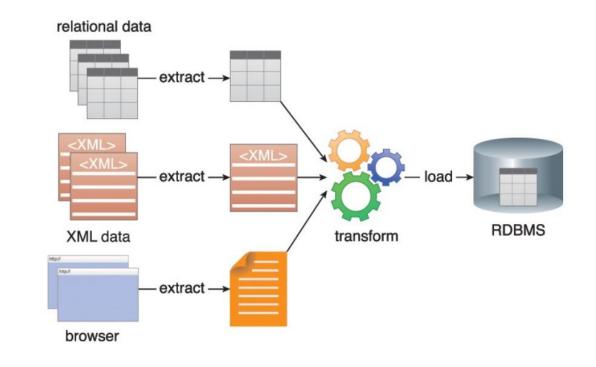
- Extract Transform Load (ETL) adalah proses memuat data dari sistem sumber ke sistem target.
- Sumber dapat berupa database, flat file, atau aplikasi.
- Target sistem dapat berupa database atau sistem penyimpanan lainnya.
- ETL mewakili operasi utama di mana data warehouses diumpankan.
- Solusi Big Data mencakup seperangkat fitur ETL untuk mengonversi data dari berbagai jenis.





Extract Transform Load (ETL)

- 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.







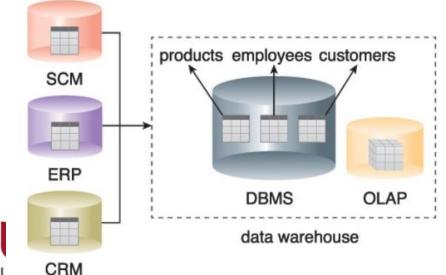
Gudang Data (Data Warehouses)

• 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.





Gudang Data (Data Warehouses)

- 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.

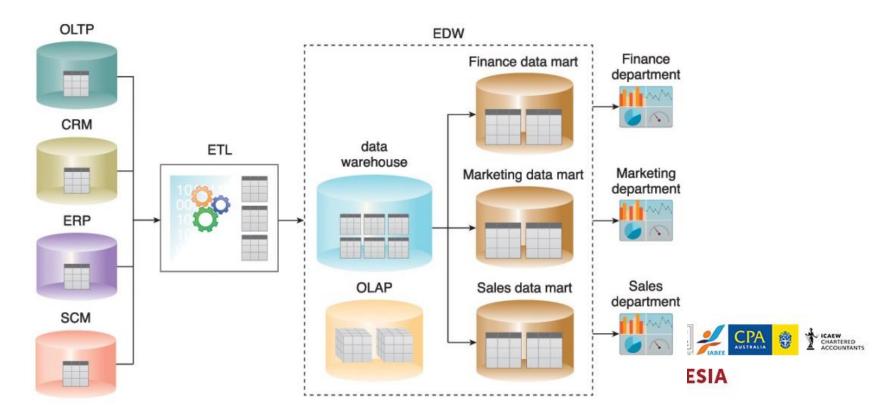




Data Marts

• 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

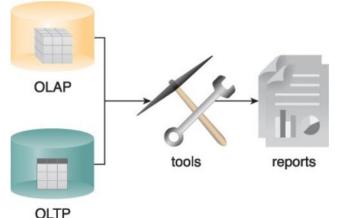
- BI tradisional terutama menggunakan analisis deskriptif dan diagnostik untuk memberikan informasi tentang peristiwa historis dan terkini.
- Ini bukan "cerdas" karena hanya memberikan jawaban atas pertanyaan yang dirumuskan dengan benar.
- Merumuskan pertanyaan dengan benar membutuhkan pemahaman tentang masalah dan isu bisnis dan data itu sendiri.
- BI melaporkan KPI yang berbeda melalui:
 - ✓ laporan ad-hoc
 - ✓ dashboards





Ad-hoc Reports

- 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.





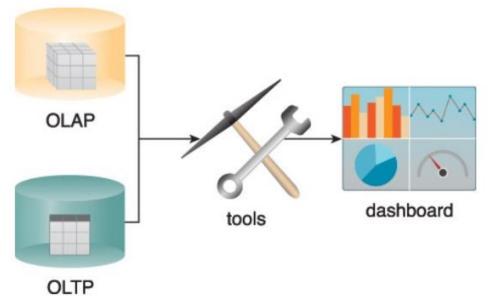


- Dasbor memberikan pandangan menyeluruh tentang area bisnis utama.
- Informasi yang ditampilkan di dasbor dihasilkan secara berkala dalam real-time atau near real-time.
- Penyajian data di dasbor bersifat grafis, menggunakan bar charts, pie charts, dan gauges.





 BI menggunakan OLAP dan OLTP untuk menampilkan informasi di dasbor.







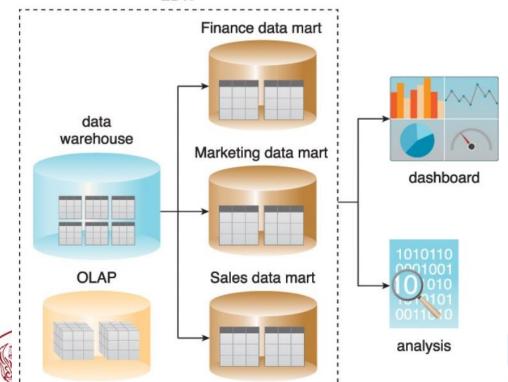
- Data warehouses dan data mart berisi informasi yang terkonsolidasi dan tervalidasi tentang entitas bisnis di seluruh perusahaan.
- BI tradisional tidak dapat berfungsi secara efektif tanpa data mart karena berisi data yang dioptimalkan dan dipisahkan, yang diperlukan BI untuk tujuan pelaporan.
- Tanpa 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 Bl

- 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.



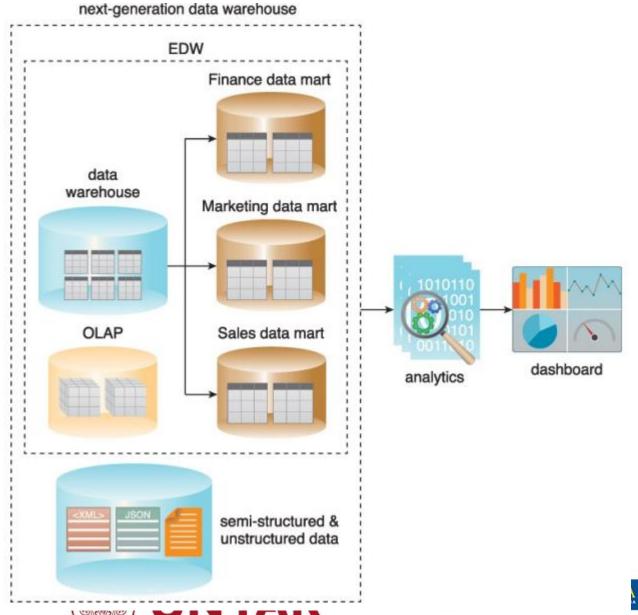
Big Data Bl

- 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.





Visualisasi Data untuk Big Data

- 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.





Visualisasi Data untuk Big Data

- 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

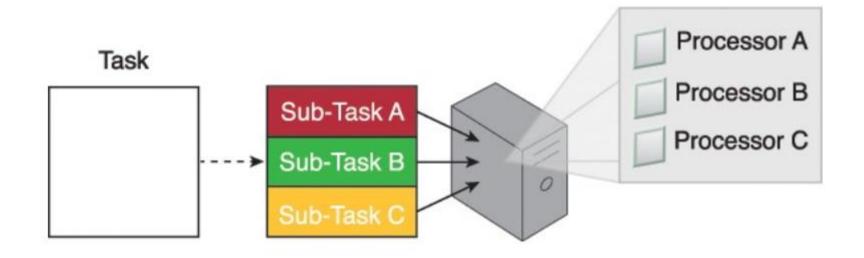




Pemrosesan Data Paralel

- 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.

Pemrosesan Data Paralel



Sebuah tugas dapat dibagi menjadi tiga sub-tugas yang dieksekusi secara paralel pada tiga prosesor berbeda dalam mesin yang sama.





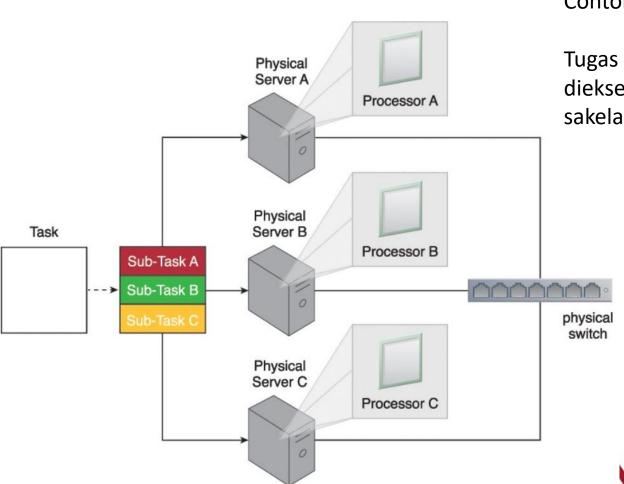
Pemrosesan Data Terdistribusi

- 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.





Pemrosesan Data Terdistribusi



Contoh pemrosesan data terdistribusi

Universitas Tarumanagara

Tugas dibagi menjadi tiga sub-tugas yang kemudian dieksekusi pada tiga mesin berbeda yang berbagi satu sakelar fisik.





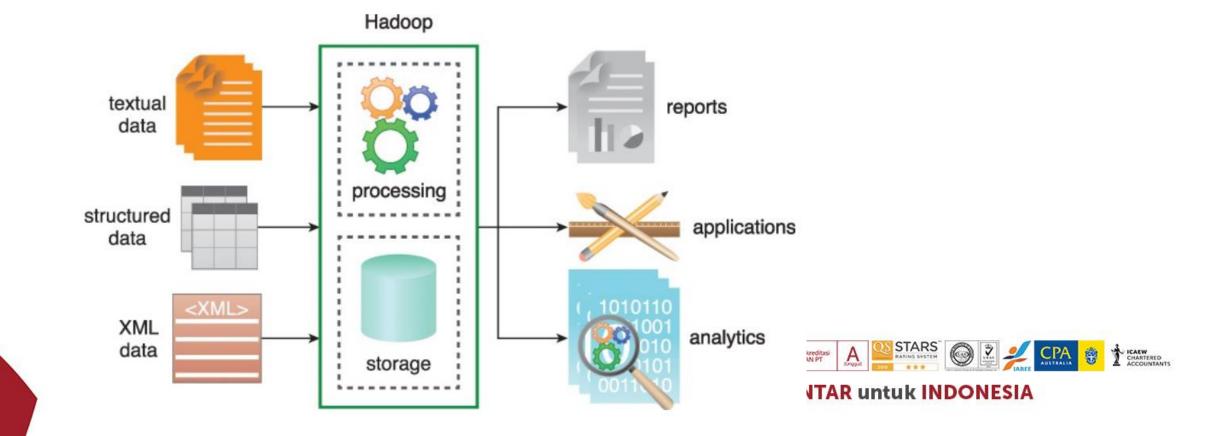
Hadoop

- 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.

Hadoop

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





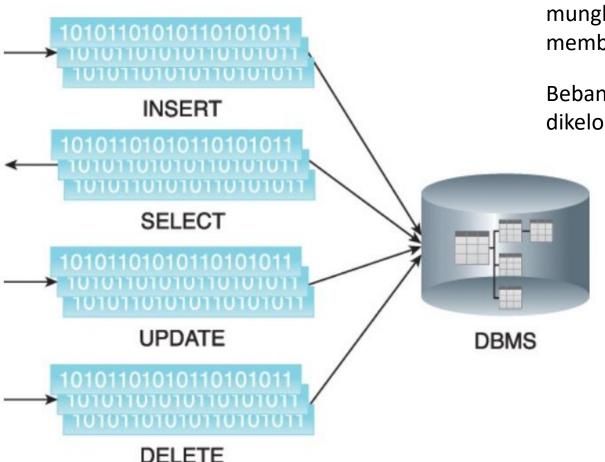
Batch

- 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.





Batch



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.















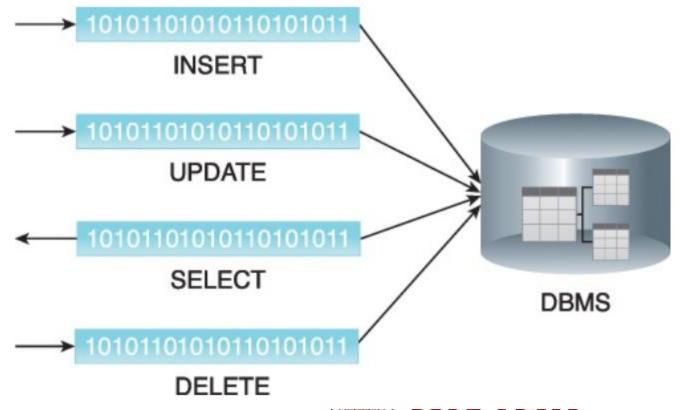


Transaksional

- 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.

Transaksional

Beban kerja transaksional memiliki sedikit operasi join dan respons latensi lebih rendah daripada beban kerja batch.





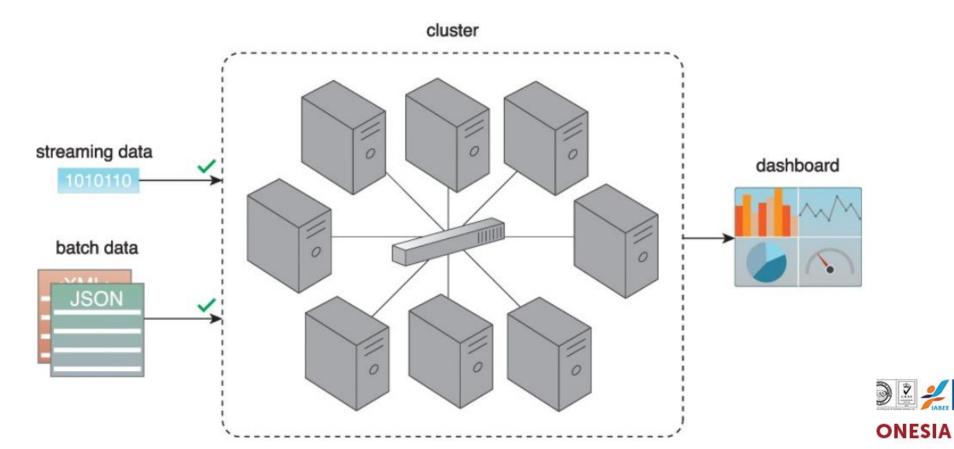


Cluster

- Cluster memberikan dukungan yang diperlukan untuk membuat solusi penyimpanan yang dapat diskalakan secara horizontal, 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.

Cluster

Sebuah cluster dapat digunakan untuk mendukung pemrosesan batch data massal dan pemrosesan data streaming secara real-time.



Cluster

- 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).





Memproses dalam Mode Realtime

- 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).





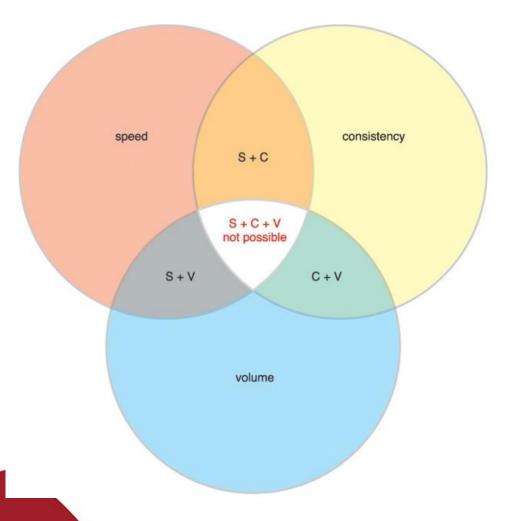
Speed Consistency Volume (SCV)

- 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.





Speed Consistency Volume (SCV)



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.

