UNTAR
Universitas Tarumanagara

Terakreditasi BAN PT A (Unggul) | QS STARS RATING SYSTEM 2019 | JABEE | CPA AUSTRALIA | ICAEW CHARTERED ACCOUNTANTS

UNTAR untuk INDONESIA

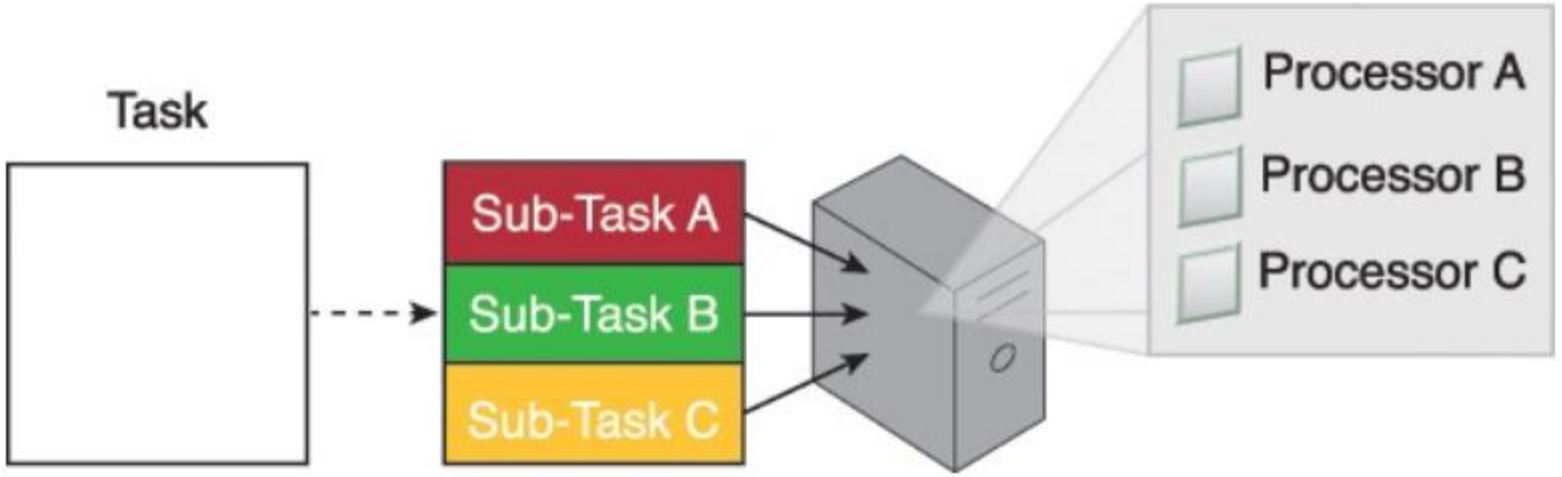
Konsep Pemrosesan Big Data

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

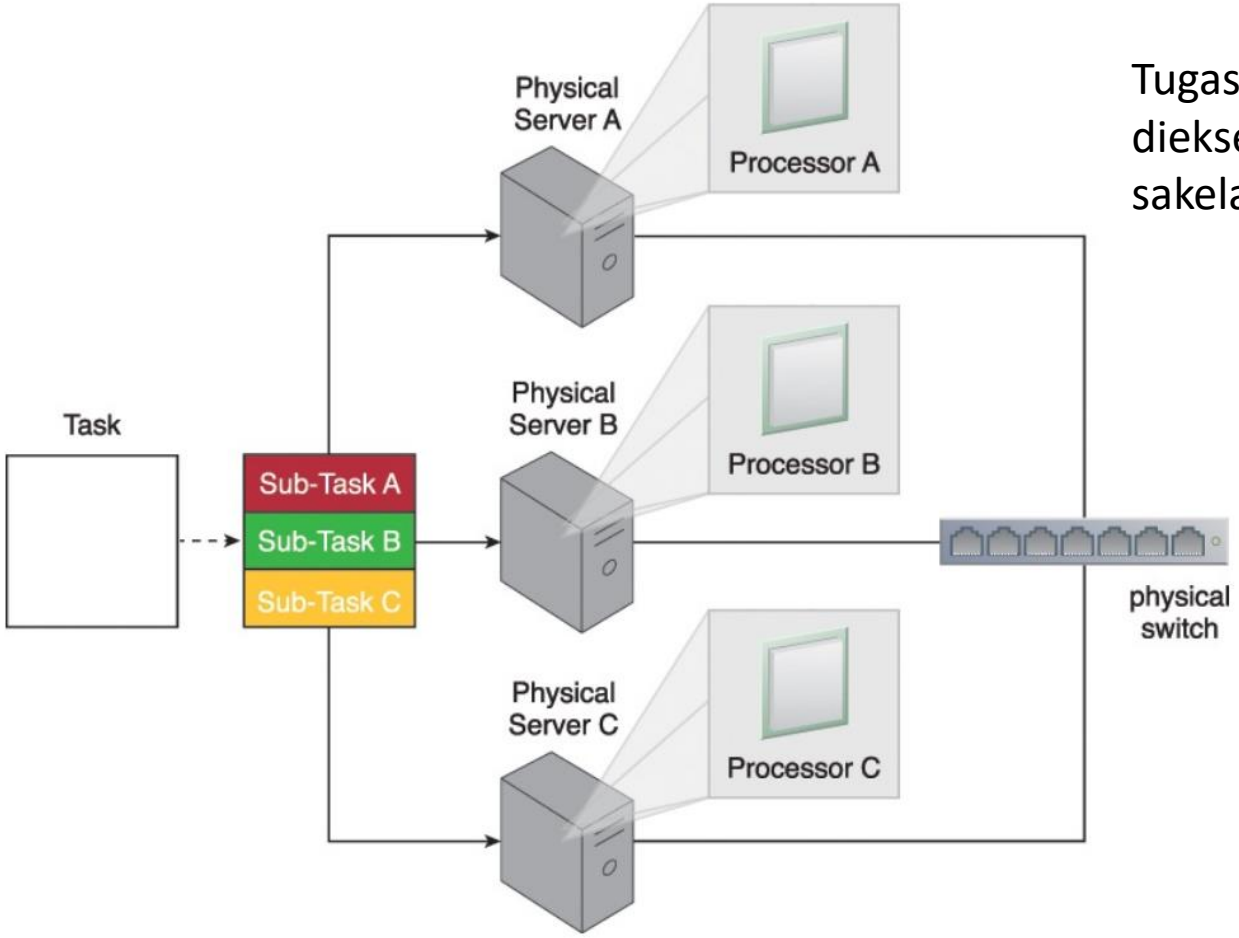


Sebuah tugas dapat dibagi menjadi tiga sub-tugas yang dieksekusi secara paralel pada tiga prosesor berbeda dalam mesin yang sama.

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

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

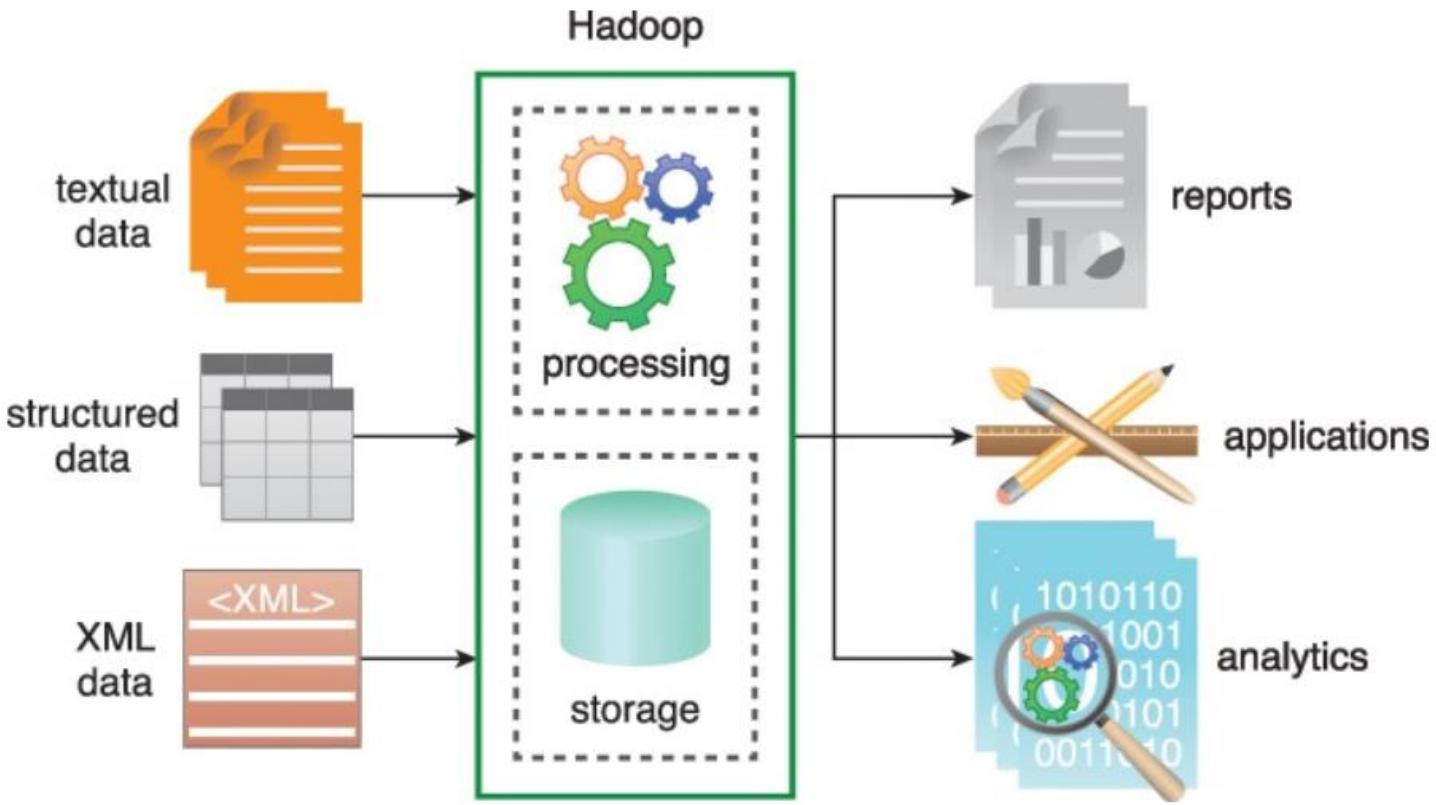
Hadoop

- 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.



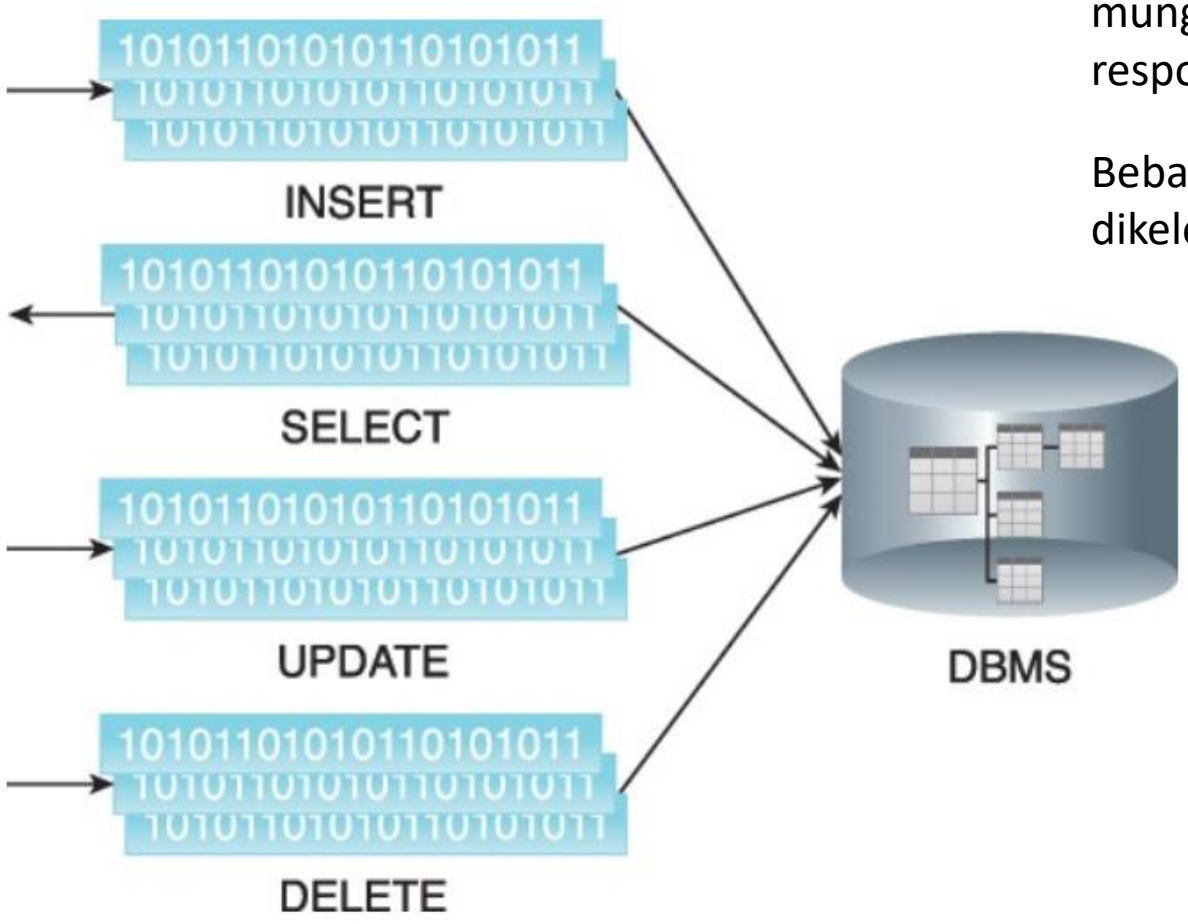
Memproses Beban Kerja (Workloads)

- 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 gabungan.
- 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 gabungan kompleks dan memberikan respons latensi tinggi.

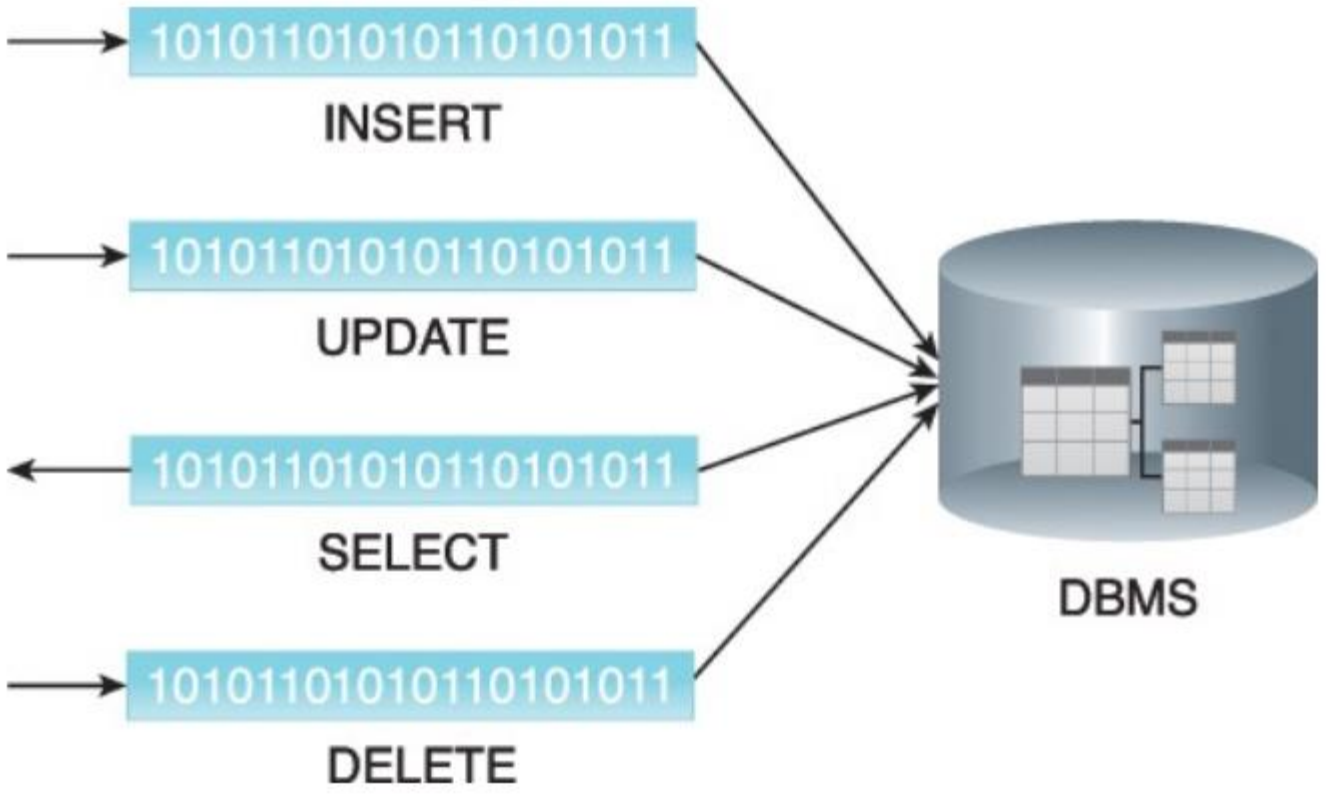
Beban kerja batch dapat mencakup baca/tulis yang dikelompokkan sebagai INSERT, SELECT, UPDATE, dan DELETE.

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 gabungan daripada intelijen bisnis dan beban kerja pelaporan.

Transaksional

Beban kerja transaksional memiliki sedikit gabungan (joins) dan respons latensi lebih rendah daripada beban kerja batch.

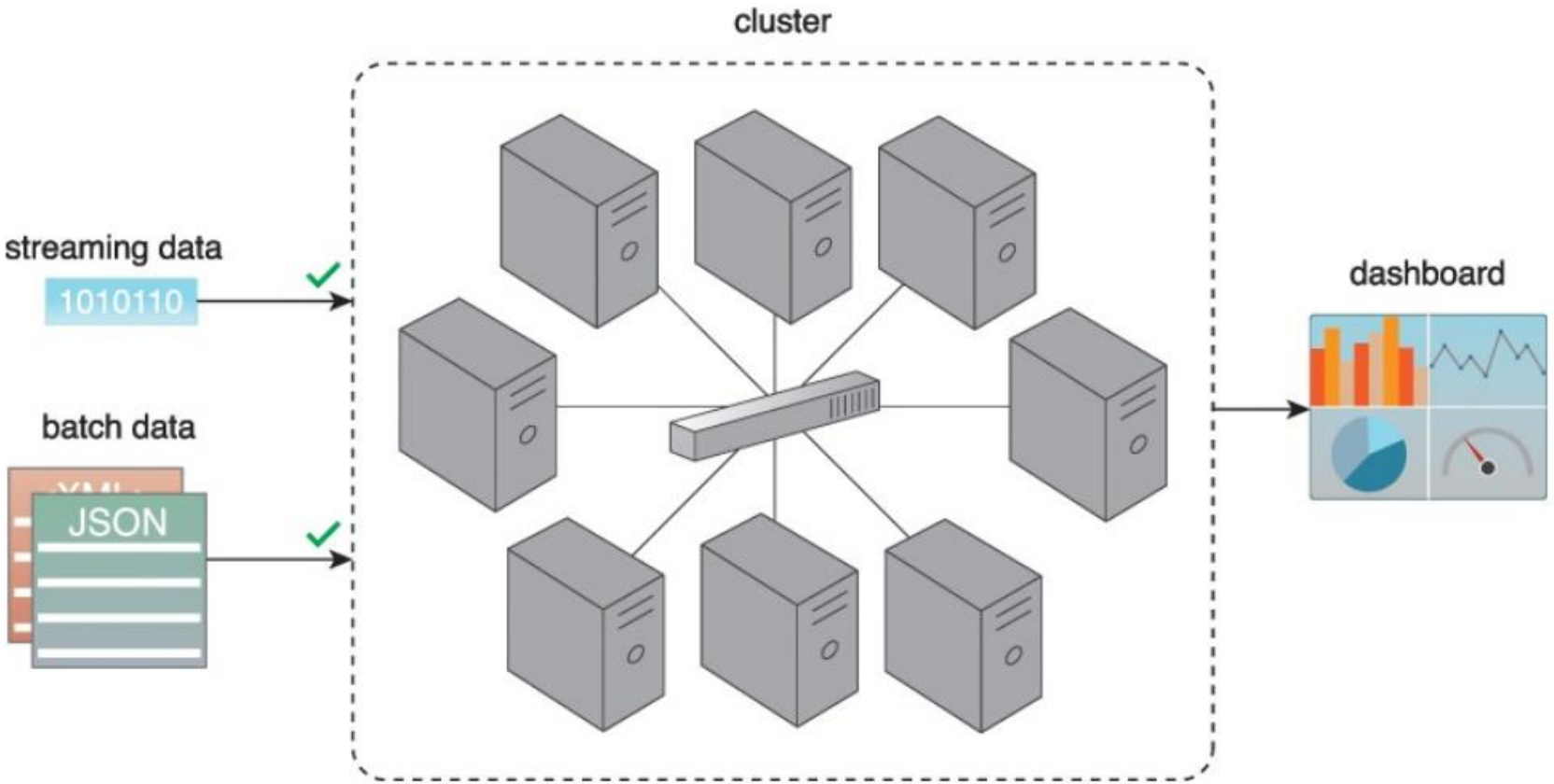


Cluster

- 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 kluster, 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 bahwa mereka menyediakan redundansi yang melekat dan toleransi kesalahan, karena mereka 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, analitik 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).

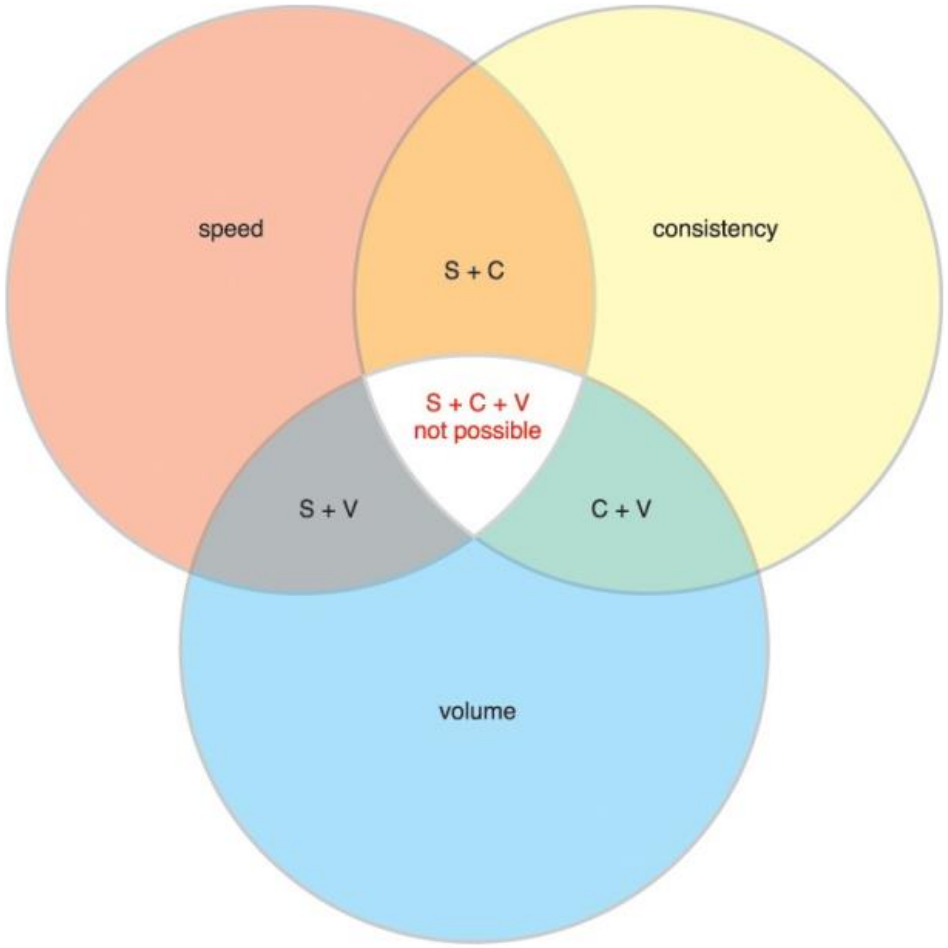
Memproses dalam Mode Realtime

- 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.

ICAEW CHARTERED ACCOUNTANTS