Rakamin Academy

HOME CREDIT

Virtual Internship Experience

Big Data

Big Data Introduction
Data Warehouse (Fundamentals)

Big Data

Big Data Introduction

Big Data adalah himpunan data yang sangat besar yang tidak mampu diolah dengan menggunakan sistem komputer dan perangkat konvensional. Konsep Big Data mulai ada sejak tahun 2000-an dan didefinisikan oleh *Doug Laney* dimana Big data memiliki 3 karakteristik yakni *Volume,Variety*, dan *Velocity*.

a. Volume
   Kapasitas data yang akan diolah dalam jumlah yang besar.
b. Variety
   Sumber data yang sangat bervariasi membuat kesulitan dalam pengolahan data. Misalkan data yang berjenis video, text, gambar, email, audio, sinyal, dll.
c. Velocity
   Kecepatan transfer data yang sangat mempengaruhi efektivitas proses pengiriman data dan stabilitas.

   Selain itu, terdapat 2 bagian karakteristik big data yang lain,
d. Variability
   Bentuk pembenaran suatu data, jika data berasal dari berbagai sumber perlu mengkorelasikan beberapa hubungan data yang saling menguatkan keterkaitan data tersebut.
e. Value
   Nilai dari aliran data yang tidak teratur dan konsisten dalam beberapa kondisi dan periode.

Kapasitas data dan sumber data yang sangat banyak , tentunya akan sangat memberatkan pengolahan data menjadi insight bisnis. Lalu bagaimana bisa komputer mengolah data tersebut. Hal ini bisa dilakukan oleh super komputer yang memiliki sistem HPS(**High Performance Computing).** HPC adalah sistem komputer yang dibangun agar mampu membantu beban komputasi yang sangat berat dalam waktu yang efektif. Sebuah HPC akan terdiri dari ratusan/ribuan CPU

yang saling terkoneksi untuk menyelesaikan komputasi secara paralel. Pengolahan Big Data ini biasanya juga menggunakan beberapa sistem framework seperti gambar berikut ini,

Mengapa Big Data itu Penting?

Pentingnya Big Data tidak ada kaitannya dengan seberapa banyak data yang kita punya, tetapi bagaimana kita bisa menggunakan data tersebut. Dengan Big Data ini kita dapat menyelesaikan permasalahan bisnis seperti,

1. Menentukan penyebab suatu masalah atau kegagalan secara *near-real time*.
2. Menemukan anomali/keanehan pada data secara cepat dan akurat.
3. Mengurangi biaya, waktu dan meningkatkan performa dari sistem teknologi.

Aplikasi Big Data dalam kehidupan sehari-hari

Salah satu contoh implementasi Big Data yang bisa kita temukan dalam kehidupan kita adalah **IoT (Internet of Things)** . **IoT(Internet of Things)** secara sederhana dapat diartikan dengan benda-benda yang ada disekitar kita bisa saling terhubung dan memberikan informasi melalui

sebuah jaringan yang disebut internet. Contohnya data informasi lokasi kemacetan, detak jantung, transaksi perbankan bahkan hingga status pada social media.

Data Warehouse- Fundamental

Data Warehouse merupakan jenis sistem management data yang di dirancang untuk mendukung kebutuhan BI (Business Intelligence).
Data warehouse digunakan untuk *query* data dan analisa data yang besar dan merupakan data historis. Data Warehouse bersumber dari berbagai sumber data seperti file log aplikasi, transaksi aplikasi,dll. Hal ini bisa membantu dalam :

- Memelihara data transaksi secara Historis.
- Menganalisa data untuk mendapatkan pemahaman yang lebih baik tentang bisnis.

Selain membantu dalam relational data, Data Warehouse juga bisa dimanfaatkan untuk *Extraction, Transformation and Loading (ETL)* , analisa statistik, reporting, mining data dan aplikasi lain yang mengelola

proses pengumpulan data, mengubahnya menjadi informasi yang lebih berguna.
Dengan data yang dikumpulkan dari berbagai sumber akan sangat membantu dalam menentukan analisa bisnis. Sumber data tersebut dapat berasal dari sistem yang dikembangkan secara internal, aplikasi yang dibeli, dan sumber lainnya. Biasanya data ini melibatkan data transaksi, produksi, pemasaran,pelanggan, dll.

Data Warehouse berbeda dengan Online Transaction Processing (OLTP). Data Warehouse dapat melakukan proses analisa data secara langsung tanpa mempengaruhi sistem transaksi. Artinya Data Warehouse adalah sistem yang berorientasikan pada pembacaan data historis.

Data Warehouse dapat menyimpan data selama berbulan-bulan atau bertahun-tahun. Data pada Data Warehouse biasanya dimuat melalui proses transaksi, Transformasi, dan Loading (ETL) dari berbagai sumber. Pendefinisian proses ETL adalah bagian yang paling besar dalam pembentukan Data Warehouse. Demikian pula, kecepatan dan optimalisasi proses ETL adalah dasar dari Data Warehouse setelah aktif dan berjalan.

Pengguna Data Warehouse melakukan analisa data yang seringkali berhubungan dengan waktu. Contohnya, Angka penjualan tahun lalu, analisis persediaan, dan laba per produk dan per pelanggan. Selain itu, pengguna Data Warehouse juga bisa melakukan penarikan data kapan saja karena Data Warehouse dirancang dengan fleksibel untuk melakukan hal tersebut. Analisa yang lebih canggih mencakup analisa trend dan mining data, dimana menggunakan data untuk memprediksi masa depan. Data Warehouse dapat bertindak sebagai dasar data yang digunakan oleh bisnis intelijen dalam menyajikan laporan, dashboard, dan antar muka lainnya.

Data Mart

Meskipun pembahasan di atas telah difokuskan pada istilah "data warehouse", ada dua istilah penting lainnya yang perlu disebutkan. Ini adalah data mart dan penyimpanan data operasi (ODS).

Data mart memiliki peran yang sama dengan data warehouse, tetapi cakupannya sengaja dibatasi. Ini dapat melayani satu departemen atau lini bisnis tertentu. Keuntungan dari data mart dibandingkan data warehouse adalah akses data lebih cepat karena cakupannya yang terbatas. Namun, data mart juga menimbulkan masalah dengan inkonsistensi. Dibutuhkan disiplin yang ketat untuk menjaga konsistensi definisi data dan perhitungan di seluruh data Mart. Masalah ini telah diakui secara luas, sehingga data mart ada dalam dua sistem.

a. Data mart independen adalah mereka yang diberi makan langsung dari sumber data. Mereka dapat berubah menjadi pulau informasi yang tidak konsisten.
b. Data mart dependent menyerupai data warehouse yang ada. Data mart dependen dapat menghindari masalah inkonsistensi, tetapi mereka memerlukan data warehouse tingkat perusahaan yang sudah ada.

Penyimpanan data operasional ada untuk mendukung operasi sehari-hari. Data ODS dibersihkan dan divalidasi, tetapi tidak mendalam secara historis: mungkin hanya data untuk hari ini. Untuk mendukung query data historis yang dapat ditangani oleh Data Warehouse, ODS memberi Warehouse sebagai tempat untuk mendapatkan akses ke data terkini, yang belum dimuat ke dalam Data Warehouse. Karena teknik pemasukan data Warehouse lebih maju, Data Warehouse mungkin kurang membutuhkan ODS sebagai sumber untuk memuat data.

Karakteristik Utama Data Warehouse

Karakteristik utama dari data warehouse adalah sebagai berikut:

- Data disusun untuk kemudahan akses dan kinerja query yang lebih efisien.
- Penggunaan data dapat direspons dengan cepat.
- Data yang ada sebagian besar termasuk data historikal.

- Query sering kali mengambil data dalam jumlah besar, mungkin ribuan baris.
- Beban data melibatkan banyak sumber dan transformasi.

Secara umum, kinerja Query yang cepat dengan sumber data yang banyak adalah kunci keberhasilan data warehouse.

Membandingkan Lingkungan OLTP dan Data Warehousing

Ada perbedaan penting antara sistem OLTP dan data warehouse. Perbedaan utama bahwa Data Warehouse tidak secara eksklusif seperti Third Normal Form (3NF). Hal ini merupakan jenis normalisasi data yang umum di lingkungan OLTP yang berperan dalam mengurangi data duplikat, meningkatkan integriti data, menghindari anomali dan lainnya. Data Warehouse dan sistem OLTP memiliki persyaratan yang sangat berbeda. Berikut adalah beberapa contoh perbedaan antara data warehouse biasa dan sistem OLTP:

- Beban Kerja
  Data Warehouse dirancang untuk mengakomodasi *ad-hoc query* . Anda mungkin tidak mengetahui beban kerja data warehouse anda sebelumnya, jadi data warehouse harus dioptimalkan agar bekerja dengan baik untuk berbagai kemungkinan operasi *query* dan analisa, sedangkan sistem OLTP hanya mendukung operasi yang telah ditentukan sebelumnya.
- Modifikasi data.
  Data warehouse akan diperbarui secara teratur oleh proses ETL (berjalan setiap periode waktu tertentu) menggunakan teknik modifikasi data massal. Pengguna data warehouse tidak secara langsung memperbarui data warehouse kecuali saat menganalisis, seperti *mining* data atau membuat prediksi dengan probabilitas. Dalam sistem OLTP, pengguna harus secara rutin mengeluarkan pernyataan modifikasi data individual ke database. Basis data OLTP sangat update, dan mencerminkan keadaan terkini dari setiap transaksi bisnis.
- Desain skema
  Data warehouse sering menggunakan skema yang

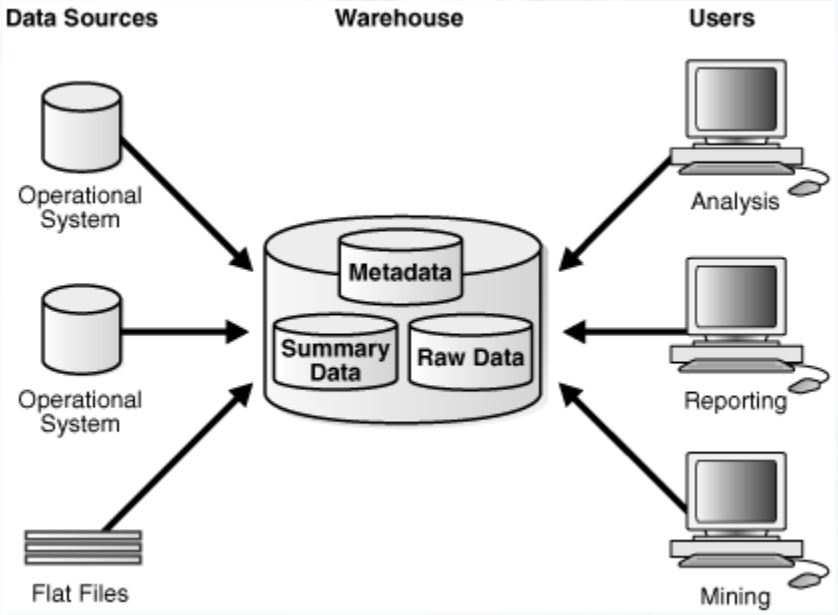
didenormalisasi sebagian untuk mengoptimalkan kinerja query dan analitis. Sementara, sistem OLTP sering menggunakan skema yang sepenuhnya dinormalisasi untuk mengoptimalkan kinerja pembaruan/penyisipan/penghapusan, dan untuk menjamin konsistensi data.

- Operasi
  Umum Permintaan data warehouse biasa memindai ribuan atau jutaan baris. Misalnya, "Temukan total penjualan untuk semua pelanggan bulan lalu."
  Operasi OLTP tipikal hanya mengakses segelintir record. Misalnya, "Ambil pesanan saat ini untuk pelanggan ini."
- Data Historis
  Data warehouse biasanya menyimpan data selama berbulan-bulan atau bertahun-tahun. Ini untuk mendukung analisis dan pelaporan historis. Sedangkan, sistem OLTP biasanya menyimpan data hanya dari beberapa minggu atau bulan. Sistem OLTP hanya menyimpan data historis yang diperlukan agar berhasil memenuhi persyaratan transaksi saat ini.

Arsitektur Data Warehouse: Umum

Gambar berikut ini menunjukkan arsitektur sederhana dari data warehouse. Pengguna dapat secara langsung mengakses data yang berasal dari beberapa sistem sumber melalui data warehouse.

Gambar 1-1 Arsitektur Data Warehouse



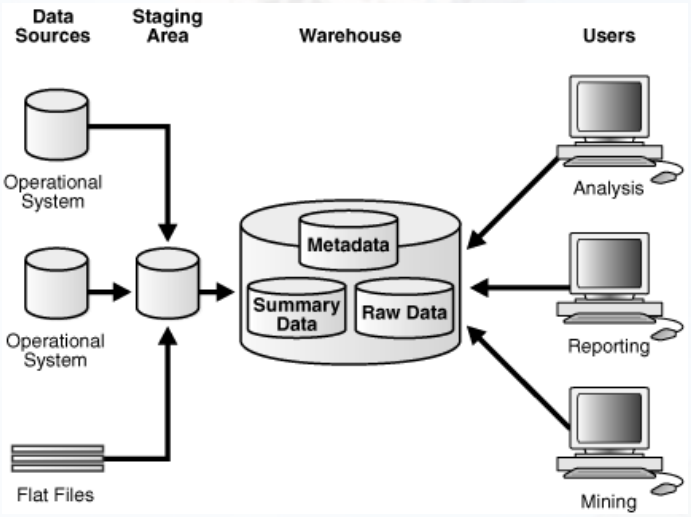
Pada gambar 1-1, metadata dan data mentah dari sistem OLTP yang ada, seperti tipe data tambahan, data ringkasan. Ringkasan merupakan mekanisme untuk melakukan pra-perhitungan operasi umum yang mahal dan berjalan lama untuk pengambilan data sub-detik. Misalnya, query data warehouse akan mengambil sesuatu seperti penjualan di bulan agustus dalam skala waktu menit.

Penyimpanan dari gabungan data mentah sebagai pusat arsitektur Data warehouse anda sering disebut sebagai Enterprise Data Warehouse (EDW). EDW memberikan pandangan 360 derajat ke dalam bisnis organisasi dengan menyimpan semua informasi bisnis yang relevan dalam format yang paling rinci.

Arsitektur Data Warehouse: Sistem Staging

Anda harus membersihkan dan memproses data operasional Anda sebelum memasukkannya ke dalam Data warehouse, seperti yang ditunjukkan pada gambar 1-2. Anda dapat melakukan ini secara sistematis, meskipun sebagian besar data warehouse menggunakan sistem staging. Sistem staging berperan dalam pembersihan dan konsolidasi data untuk data operasional yang berasal dari berbagai sistem sumber, terutama untuk data warehouse perusahaan di mana semua informasi yang relevan dari suatu perusahaan dikonsolidasikan.

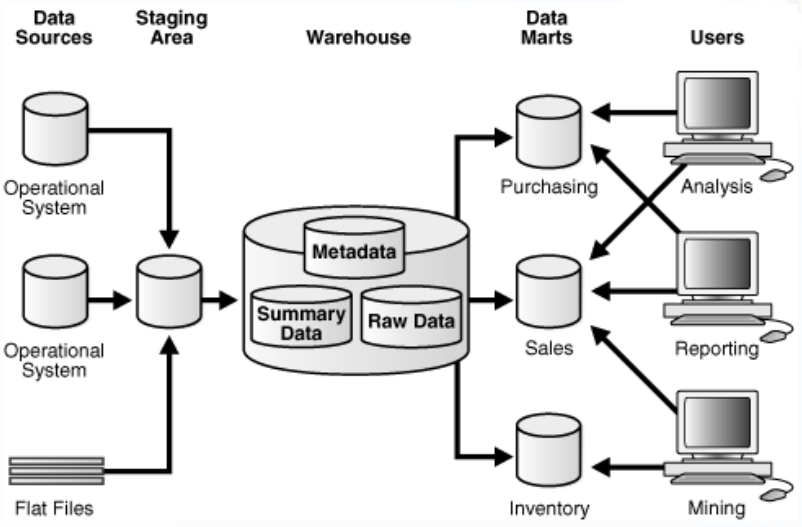
Gambar 1-2 Arsitektur Data Warehouse dengan Staging Area



Arsitektur Data Warehouse: Sistem Staging dan Data Mart

Meskipun arsitektur pada gambar 1-2 cukup umum, Anda mungkin ingin menyesuaikan arsitektur data warehouse anda untuk grup yang berbeda dalam organisasi anda. Anda dapat melakukan ini dengan menambahkan data mart, yang merupakan sistem yang dirancang untuk lini bisnis tertentu.Gambar 1-3 dibawah ini mengilustrasikan contoh dimana pembelian, penjualan, dan persediaan dipisahkan. Dalam contoh ini, seorang analis keuangan mungkin ingin menganalisis data historis untuk pembelian dan penjualan untuk membuat prediksi tentang perilaku pelanggan.

Gambar 1-3 Arsitektur Data Warehouse dengan Staging Area dan Data Mart



Daftar Referensi :

- SAS Institute Inc, Big Data: What it is and Why it matters ?, from https://www.sas.com/en_id/insights/big-data/what-is-big-data.html , [access: 23/-02/2022]
- ORACLE, Introduction to Data Warehousing Concepts, from https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG9288 , [access: 23/02/2022]