

PERTEMUAN 14

DATA WAREHOUSE

Data Warehouse

Definisi :

- Data Warehouse adalah Pusat repositori informasi yang mampu memberikan database berorientasi subyek untuk informasi yang bersifat historis yang mendukung DSS (Decision Suport System) dan EIS (Executive Information System).
- Salinan dari transaksi data yang terstruktur secara spesifik pada query dan analisa.
- Salinan dari transaksi data yang terstruktur spesifik untuk query dan laporan

Tujuan :

- Meningkatkan kualitas dan akurasi informasi bisnis dan mengirimkan informasi ke pemakai dalam bentuk yang dimengerti dan dapat diakses dengan mudah.

4 KARAKTERISTIK DATA WAREHOUSE

1. Subject oriented

- Data yang disusun menurut subyek berisi hanya informasi yang penting bagi pemrosesan decision support.
- Data yang disusun menurut subyek berisi hanya informasi yang penting bagi pemrosesan decision support.
- Database yang semua informasi yang tersimpan dikelompokkan berdasarkan subyek tertentu misalnya : pelanggan, gudang, pasar, dsb.
- Semua informasi tersebut disimpan dalam suatu system data warehouse.
- Datadata disetiap subyek dirangkuman ke dalam dimensi, misalnya: periode waktu, produk, wilayah, dsb, sehingga dapat memberikan nilai sejarah untuk bahan analisa.

2. Integrated

Jika data terletak pada berbagai aplikasi yang terpisah dalam suatu lingkungan operasional, encoding data sering tidak seragam sehingga bila data dipindahkan ke data warehouse maka coding akan diasumsikan sama seperti lazimnya.

3. Time-variant

Data warehouse adalah tempat untuk storing data selama 5 sampai 10 tahun atau lebih, data digunakan untuk perbandingan atau perkiraan dan data ini tidak dapat diperbaharui.

4. Non volatile

Data tidak dapat diperbaharui atau dirubah tetapi hanya dapat ditambah dan dilihat.

MASALAH-MASALAH DALAM MENERAPKAN DATA WAREHOUSE

- Dokumentasi dan pengelolaan metadata dari data warehouse.
- Penentuan aturan dalam proses transformasi untuk menetapkan berbagai sumber legacy data yang akan dimasukkan ke dalam data warehouse.
- Pencapaian proses pengembangan yang handal, baik dalam membangun, mengimplementasikan, maupun memelihara data warehouse.

KEUNTUNGAN DATAWAREHOUSE

- Datawarehouse menyediakan model data yang bervariasi, dan tidak bergantung pada satu sumber data saja. Hal ini memudahkan pimpinan perusahaan/manager membuat laporan dan menganalisa.
- Saat me-load data ke dalam datawarehouse, data yang tidak konsisten akan diketahui dan secepatnya dirubah. Mendukung proses pembuatan laporan, agar keputusan yang diambil adalah keputusan yang benar sesuai data.
- Keamanan informasi didalam datawarehouse terjamin, karena datawarehouse selalu digunakan dan dimonitor oleh pengguna datawarehouse tersebut.
- Dalam membuat laporan tidak membuat proses transaksi yang ada menjadi lambat, karena datawarehouse terpisah dengan database operasional.
- Datawarehouse menyediakan berbagai macam bentuk laporan yang terbaru.

KERUGIAN DATAWAREHOUSE

- Datawarehouse tidak cocok untuk data yang tidak struktur.
- Data perlu di extract, diubah, dan di load ke datawarehouse, sehingga membutuhkan waktu (**delay**) kerja untuk datawarehouse yang belum terbentuk.
- Semakin lama masa hidup bisnis yang menggunakan datawarehouse, maka semakin banyak biaya yang dikeluarkan oleh perusahaan untuk memodifikasi teknologi datawarehouse atau perawatan berjalan datawarehouse.
- Jika data yang diambil lambat, maka data yang dimiliki di datawarehouse tidak berkualitas/ sehingga laporan tidak optimal.

MENCIPTAKAN SEBUAH DATA WAREHOUSE

Ada banyak tantangan dalam menciptakan dan memelihara sebuah data warehouse yang besar. Sebuah skema database yang baik harus dirancang untuk menahan koleksi terpadu dari data yang disalin dari berbagai sumber. Sebagai contoh, sebuah warehouse perusahaan mungkin termasuk persediaan dan database personil departemen ', bersama-sama dengan database penjualan dikelola oleh kantor di negara yang berbeda. Karena source database sering dibuat dan dipelihara oleh kelompok yang berbeda, ada sejumlah ketidaksesuaian semantik di database ini, seperti unit mata uang yang berbeda, nama yang berbeda untuk atribut yang sama, dan perbedaan dalam bagaimana tabel dinormalisasi atau terstruktur; perbedaan-perbedaan ini harus didamaikan ketika data dibawa ke warehouse. Setelah skema warehouse dirancang, warehouse harus diisi, dan dari waktu ke waktu, itu harus tetap konsisten dengan database sumber.

Ciri-ciri Data Warehouse

Terdapat 4 karakteristik data warehouse

1. Subject oriented

- Data yang disusun menurut subyek berisi hanya informasi yang penting bagi pemrosesan decision support.
- Database yang semua informasi yang tersimpan di kelompokkan berdasarkan subyek tertentu misalnya: pelanggan, gudang, pasar, dsb.
- Semua Informasi tersebut disimpan dalam suatu sistem *data warehouse*.
- *Data-data di setiap subyek dirangkum ke dalam dimensi, misalnya : periode waktu, produk, wilayah, dsb, sehingga dapat memberikan nilai sejarah untuk bahan analisa.*

Ciri-ciri Data Warehouse

2. Integrated

- Jika data terletak pada berbagai aplikasi yang terpisah dalam suatu lingkungan operasional, encoding data sering tidak seragam sehingga bila data dipindahkan ke data warehouse maka coding akan diasumsikan sama seperti lazimnya.

3. Time-variant

- Data warehouse adalah tempat untuk storing data selama 5 sampai 10 tahun atau lebih, data digunakan untuk perbandingan atau perkiraan dan data ini tidak dapat diperbaharui.

4. Non volatile

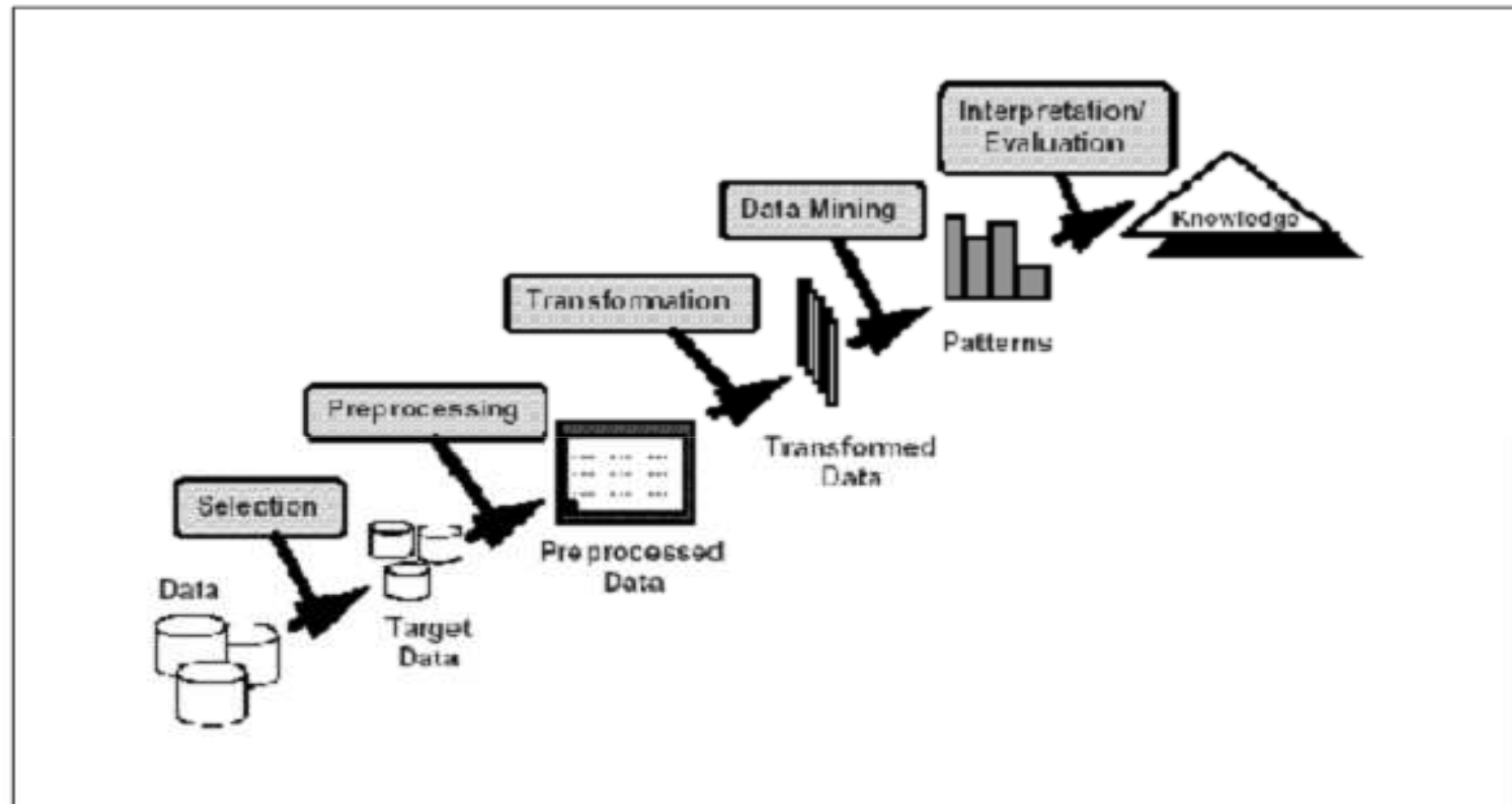
- Data tidak dapat diperbaharui atau dirubah tetapi hanya dapat ditambah dan dilihat.

Data Preprocessing

- Data preprocessing menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan proses prosedur yang lainnya.
- Dalam data mining menstrasformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan pemakai, contohnya Neural Network.
- Terdapat beberapa alat dan metode yang berbeda yang digunakan untuk preprocessing seperti :
 - *Sampling : menyeleksi subset representatif dari populasi data yang besar.*
 - *Transformation : memanipulasi data mentah untuk menghasilkan input tunggal.*
 - *Denoising : menghilangkan noise dari data*
 - *Normalization : mengorganisasi data untuk pengaksesan yang lebih spesifik*
 - *Feature extration : membuka spesifikasi data yang signifikan dalam konteks tertentu.*

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

- KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dari pola-pola sejumlah kumpulan data.
- *Knowledge discovery in databases (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.*



Gambar. 1. Tahapan KDD

TAHAPAN PROSES KDD

1. *Data Selection*

- Menciptakan himpunan data target , pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.
- Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, *disimpan dalam suatu berkas, terpisah dari basis data operasional*.

2. *Pre-processing/ Cleaning*

- Pemrosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan.
- Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan *proses cleaning* pada data yang menjadi fokus KDD.
- Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*).
- Dilakukan proses *enrichment*, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformation

- Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai.
- Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. *Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data*

4. *Data mining*

- Pemilihan tugas data mining; pemilihan goal dari proses KDD misalnya klasifikasi, regresi, clustering, dll.
- Pemilihan algoritma data mining untuk pencarian (searching)
- *Proses Data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.*

5. *Interpretation/ Evaluation*

- Penerjemahan pola-pola yang dihasilkan dari *data mining*.
- Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.
- Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.