

DATA WAREHOUSE

DATA WAREHOUSE

Adalah kumpulan dari komponen-komponen perangkat keras dan perangkat lunak yang dapat digunakan untuk mendapatkan analisa yang lebih baik dari data yang berjumlah sangat besar sehingga dapat membuat keputusan yang baik.

Dengan kata lain sebagai gudang data.

Manfaat Data Warehouse

Data Warehouse biasanya digunakan untuk:

1. Memahami trend bisnis dan membuat perkiraan keputusan yang lebih baik.
2. Menganalisa informasi mengenai penjualan harian dan membuat keputusan yang cepat dalam mempengaruhi performance perusahaan.

Data Warehouse Customer Example

Salah satu pelanggan AS/400, menemukan kesulitan dalam hal informasi penjualan yang dibutuhkan untuk dapat memperluas bisnisnya. Tiap bulan laporan penjualan sederhana tidak tepat waktu atau tidak mendetail sehingga sangat tidak membantu. Untuk membantu membuat keputusan bisnis, perusahaan ini perlu dianalisa latar belakangnya sehingga dapat menemukan titik-titik trends bisnis.

Dengan AS/400 data warehouse-nya yang baru, para penjual dari perusahaan ini dapat membentuk strategi penjualan berdasarkan informasi yang diberikan warehouse. Informasi ini termasuk mengenai the success of previous promotions, regional trends, product profitability dan the effect of product packaging.

Perusahaan lain seperti departemen, seperti keuangan dan operasi, juga menggunakan warehouse untuk mengidentifikasi dan menganalisa produk yang berhasil menyeberangi daerah, penjualan dan waktu.

Data warehouse dapat sebagai kunci pembeda dalam suatu industri-industri yang berbeda.

Aplikasi Data Warehouse meliputi:

1. Sales and marketing analysis across all industries.
2. Inventory turn and product tracking in manufacturing.
3. Kategori manajemen, analisa penjualan, dan perbaikan analisa program pemasaran yang efektif.
4. Keuntungan dari jalan raya atau analisa resiko pengemudi dalam hal transportasi.
5. Analisa keuntungan atau resiko penetapan pajak atau mendenda dalam bank.
6. Analisa tuntutan dari deteksi penggelapan dalam asuransi.

Operational versus Informational data

Operational data adalah data yang digunakan untuk menjalankan bisnis. Data ini mempunyai ciri disimpan, diperoleh dan diupdate oleh system Online Transactional Processing (OLTP). Sebagai contoh, system pemesanan, aplikasi perhitungan atau an order entry application.

Operational data biasanya disimpan dalam relational database, tetapi mungkin disimpan dalam legacy hierarchical atau flat formats as well.

Karakteristik operasional data meliputi:

1. Sering diperbaharui dan transaksi yang bersifat online.
2. Non-historical data (tidak lebih dari tiga sampai dengan enam bulan lamanya).
3. Optimized for transactional processing.
4. Tingginya normalisasi dalam relasional database untuk memudahkan pembaharuan, pemeliharaan dan integritas.

Informational data biasanya disimpan dalam format yang membuat analisa lebih mudah. Analysis can be in the form of decision support (queries), report generation, executive information systems.

Informational data dibuat dari operational data kekayaan yang ada di dalam bisnis anda.

Informational data adalah apa yang membentuk sebuah data warehouse.

Ciri informational data:

1. Summarized operational data.
2. De-normalized and replicated data.
3. Infrequently updated from the operational data.
4. Optimized for decision support applications.
5. Possibly read-only (no updates allowed).
6. Stored on separate system to lessen impact on operational system.

Metadata

Informasi mengenai data warehouse dan data yang diisi ke dalam Data warehouse dibagi dua bagian. Yang pertama technical data the warehouse uses, dan yang kedua business data that is of use to the warehouse users. Semua data ini menunjukkan sebagai *metadata*, data about the data.

The technical data berisi penjelasan tentang operational database dan penjelasan dari data warehouse. Dari kedua penjelasan itu, atau skema, operasi pergerakan data dapat diimplementasikan. Data ini membantu administrasi data warehouse menjaga datanya dan mengetahui dari mana semua data berasal.

The business data membantu pemakai mencari informasi dalam data warehouse tanpa mengetahui implementasi databasenya. (This information is

presented in business terms, instead of the terms used by the programmers when the database was built)

The business data memberikan informasi kepada pemakai:

1. Pada saat data dipindahkan ke dalam warehouse (how current it is)
2. Dari mana data berasal. (which operational database).
3. Other information that lets the user know how reliable the data is.

Business Intelligence Software and Data Mining

Business intelligence software is a fairly new term referring to the tools that are used to analyze the data.

Software ini terdiri dari:

1. Decision Support System (DSS) tools
Dimana diijinkan untuk membangun ad hoc queries and generate reports.
2. Executive information system (EIS).
Which combine decision support with extended analysis capabilities and access to outside resources (such as Dow Jones News Services)
3. Data Mining tools.
Which allow automation of the analysis of your data to find patterns or rules that you can use to tailor business operations.

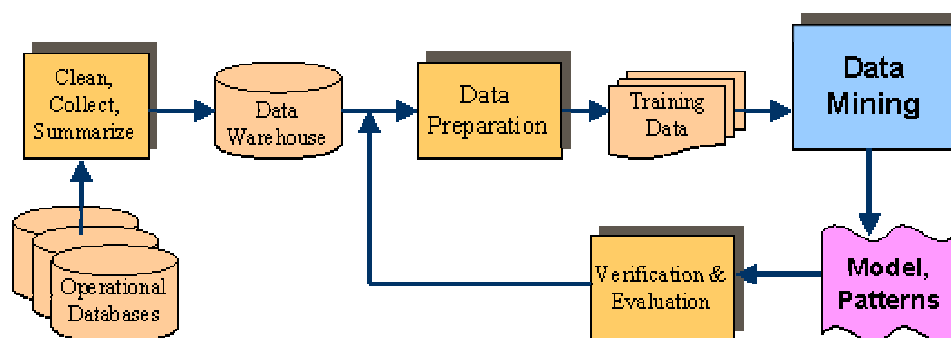
DATA MINING

Definisi Data Mining adalah

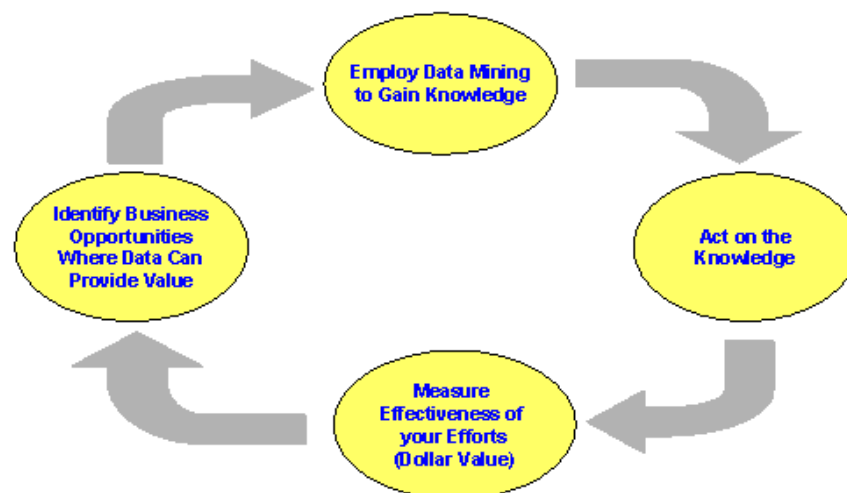
1. Mencari informasi yang berharga di dalam suatu data yang berjumlah besar.
2. Eksplorasi dan analisa secara otomatis atau semiotomatis dari suatu kuantitas data yang besar yang bertugas untuk mencari pola dan aturan yang berarti.

KDD Process

Knowledge Discovery in Databases (KDD) is a *non-trivial* process of *identifying valid, novel, potentially useful, and ultimately understandable* patterns in data.



BUSSINESS CYCLE OF DATA MINING



Alasan-alasan menggunakan Data Mining

1. Karena data dikumpulkan dan disimpan dengan kecepatan yang sangat besar (Gbyte/hour).
 - Sensor jarak jauh yang menggunakan satelit.
 - Telescope scanning the skies.
 - Micro arrays generating gene expression data.
 - Scientific simulations generating terabytes of data

2. teknik tradisional yang tidak layak lagi
3. Digunakan untuk mereduksi data atau data dibagi-bagi.
 - Catalog, klasifikasi, pembagian data.
 - Membantu para ahli sains dalam menghipotesa.

Asal mula Data Mining :

1. Penggambaran ide-ide dari Mesin Buatan atau Artificial Intelligent, pola, statistik, sistem database dan penggambaran data.
2. Tehnik tradisional mungkin tidak digunakan karena
 - Banyaknya data.
 - Tingginya dimensi dari suatu data.
 - Berbagai macam jenis data.

Tugas Data Mining dibagi menjadi dua metode yaitu:

1. Metoda prediksi
Menggunakan beberapa variable untuk memperkirakan suatu nilai yang tidak diketahui dari variable yang lain.
2. Metoda deskripsi
Mencari suatu pola yang dapat ditafsirkan manusia sehingga data dapat digambarkan atau diuraikan.

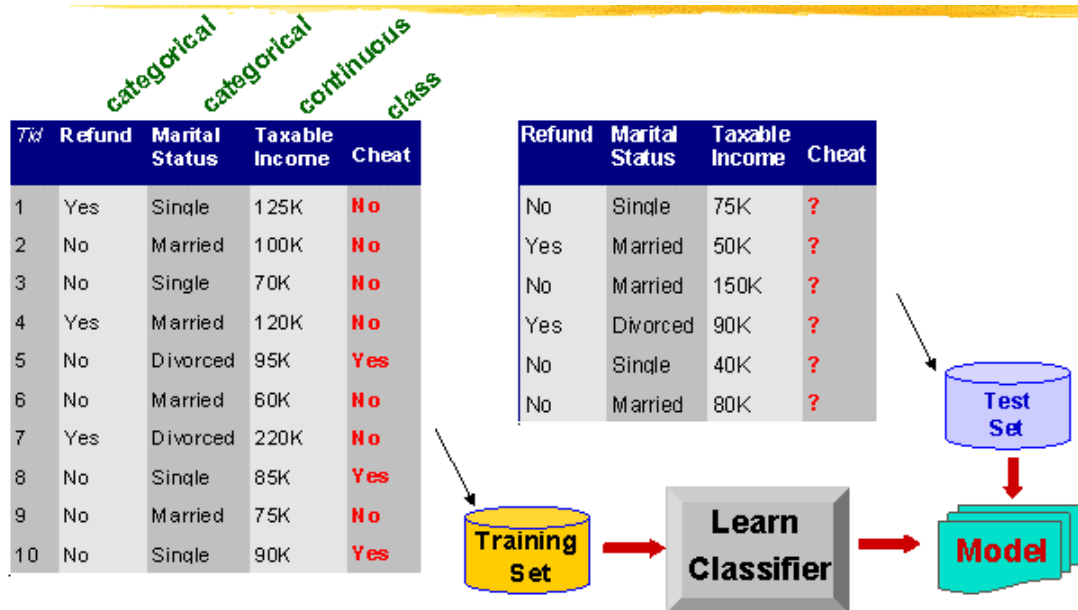
Jenis-jenis Tugas Data Mining

1. Classification [Predictive]
2. Clustering [Descriptive]
3. Association Rule Discovery [Descriptive]
4. Regression [Predictive]
5. Deviation Detection [Predictive]

Definisi Klasifikasi

1. Memberikan kumpulan record-record (training set)
 - Setiap record berisi sifat-sifat tertentu (attributes), salah satu dari attributes adalah kelas (class).
2. Mencari sebuah contoh atau model untuk class attribute sebagai fungsi dari suatu nilai dari attribute yang lain.
3. Tujuannya adalah record2 yang tidak kelihatan/ previously unseen record ditunjuk menjadi suatu class setepat mungkin.

CLASSIFICATION EXAMPLE



Klasifikasi pada Aplikasi 1:

Direct Marketing

Goal:

Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.

Pendekatannya adalah

- Menggunakan data untuk produk yang sama yang telah dikenalkan terlebih dahulu.
- Mengetahui pembeli mana yang memutuskan untuk membeli dan yang tidak. Keputusan ini (buy, don't buy) membentuk suatu class attributes.
- Mengumpulkan bermacam-macam demographic, gaya hidup dan interaksi perusahaan dan informasi yang berhubungan dengan para pelanggan.
Contohnya yaitu di mana mereka tinggal, berapa besar pendapatannya dan lain-lain.
- Use this information as input attributes to learn a classifier model.

Klasifikasi pada Aplikasi 2:

Fraud Detection

Tujuannya adalah untuk memprediksi atau memperkirakan kasus penggelapan transaksi credit card.

Pendekatannya adalah

- Dengan menggunakan informasi transaksi dan informasi dari kartu sebagai atributnya.
Contohnya

Kapan seorang pelanggan membeli, apa yang ia beli, seberapa sering ia membayar tepat waktu.

- Label past transactions. This forms the class attributes.
- Learn a model for the class of the transactions.
- Menggunakan model ini untuk mendeteksi penggelapan/fraud dengan mengobservasi/meninjau perhitungan transaksi credit card.

Klasifikasi pada Aplikasi 3:

Customer Attrition/Churn

Tujuannya adalah

To predict whether a customer is likely to be lost to a competitor.

Pendekatannya adalah

- Menggunakan record yang mendetail dari suatu transaksi dari tiap pelanggan untuk mencari atributnya.
- Memberi label pada pelanggan sebagai pelanggan setia atau yang bukan langganan.
- Find a model for loyalty.

Klasifikasi pada Aplikasi 4:

Sky Survey Cataloging

Tujuannya adalah

Memprediksi class (bintang atau galaksi) dari objek langit, khususnya menggambarkan yang lemah, berdasarkan gambar yang diambil menggunakan teleskop (dari Palomar Observatory).

Pendekatannya adalah

- Membagi gambar.
- Mengukur atribut gambar.
- Model dari suatu kelas berdasarkan dari penggambaran ini.

Definisi Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that.

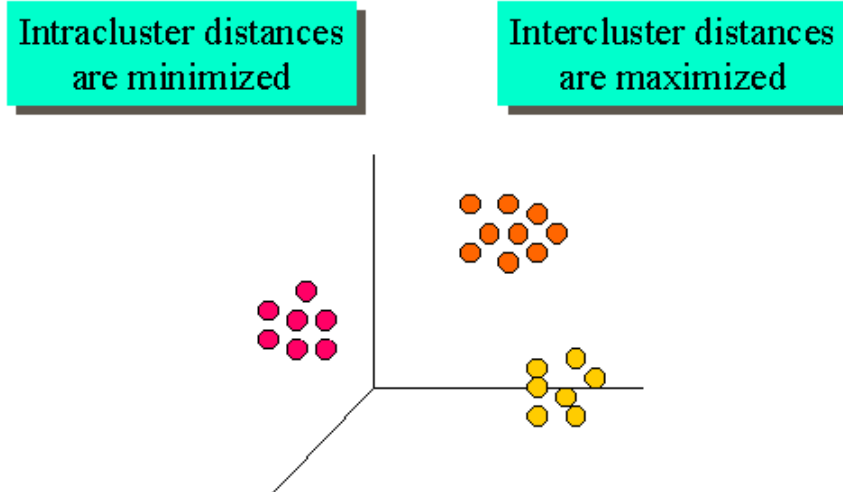
- Point-point data di dalam satu cluster hampir sama dengan yang lain.
- Point-point data di dalam cluster yang berbeda kurang mirip dengan yang lain.

Persamaan Ukuran

- Euclidean distance if attributes are continuous.
- Other problem-specific Measures.

ILLUSTRATING CLUSTERING

Euclidean Distance Based Clustering in 3-D space.



Clustering Application 1:

Market segmentation:

Tujuannya adalah

Subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Pendekatannya adalah

- Mengumpulkan attribute-attribut yang berbeda dari pelanggan berdasarkan informasi yang berhubungan dengan geographical dan gaya hidup pelanggan.
- Mencari cluster atau kumpulan dari pelanggan-pelanggan yang serupa.
- Mengukur kualitas clustering dengan memperhatikan atau mengamati pola pembelian dari para pelanggan di dalam cluster yang sama dengan cluster yang berbeda.

Clustering pada Aplikasi 2:

Document Clustering

Tujuannya adalah

Untuk mencari kelompok dari dokumen dimana kelompok-kelompok itu mirip satu dengan yang lain berdasarkan dari term yang ada.

Pendekatannya adalah

Untuk mengidentifikasi atau mempersamakan batas waktu di dalam tiap dokumen. Bentuk ukuran yang hampir sama berdasarkan frekuensi dari term yang berbeda digunakan sebagai cluster.

Untuk mendapatkannya

Information retrieval can utilize the clusters to relate a new document or search term to clustered documents.

ILLUSTRATING DOCUMENT CLUSTERING

Clustering Points: 3204 Articles of Los Angeles Times.
 Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

CLUSTERING OF S&P STOCK OF DATA

Observe Stock Movements every day.

Clustering points: Stock- $\{UP/DOWN\}$

Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.

- We used association rules to quantify a similarity measure.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Aroclor-Matl-DOWN, Bay-Network-Down, 3-COM-DOWN, Cabletron-Sys-DOWN, CECO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-Down, Telex-Int-Down, Natl-Semiconductor-DOWN, Oracle-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comm-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
3	Frank-Mae-DOWN, Fed Home Loan-DOWN, MBNA-Corp-DOWN, Monran-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Uthco-UP, Schlumberger-UP	Oil-UP

ASSOCIATION RULE
DISCOVERY: DEFINITION

Given a set of records each of which contain some number of items from given collection;

- Produce dependency rules which will predict occurrence of an item based on occurrences of their items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

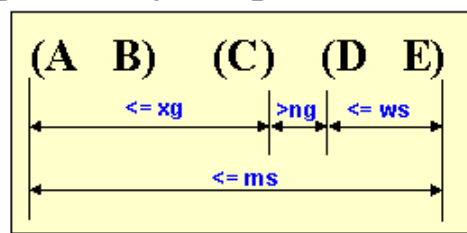
Rules Discovered:
 {Milk} --> {Coke}
 {Diaper, Milk} --> {Beer}

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong **sequential dependencies** among different events.

$$(A \ B) \ (C) \ \rightarrow \ (D \ E)$$

- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Regression

1. memperkirakan suatu nilai dari pemberian nilai variable yang berlanjut berdasarkan nilai dari variabel-variabel lain.
2. Pentingnya mempelajari statistik, neural network fields.
3. contohnya :

- Memperkirakan banyaknya penjualan dari suatu produk baru berdasarkan pemakaian advetising.
- memperkirakan kecepatan putaran sebagai kedudukan suhu, kelembaban, tekanan udara dan lain-lain.
- Perkiraan waktu dari persediaan barang di suatu pasar.

Deviation Detection

- menemukan perubahan yang paling banyak berarti pada data dari ukuran sebelumnya atau normative values.
- kategori atau golongan yang biasa terpisah dari tugas data mining yang lain
- perubahan klasifikasi, clustering, analisa time series dapat digunakan sebagai rata-rata untuk mencapai tujuan.
- outlier detection in statistics.