Data Warehousing and Business analytics

Lecture-1

Introduction and Background

Data Warehouse (DWH)

- Data recording and storage is growing.
- History is excellent predictor of the future.
- Gives total view of the organization.
- Intelligent decision-support is required for decision-making.

Reason-1: Why a Data Warehouse?

- Size of Data Sets are going up ↑.
- Cost of data storage is coming down ↓.
  - The amount of data average business collects and stores is **doubling every year**

Reason-1: Why a Data Warehouse?

- A Few Examples
  - WalMart: 24 TB
  - France Telecom: ~ 100 TB
  - CERN: Up to 20 PB by 2006
  - Stanford Linear Accelerator Center (SLAC): 500TB

Caution!

A Warehouse of Data

is NOT a

Data Warehouse

Caution!

Size

is NOT

Everything

Reason-2: Why a Data Warehouse?

- Businesses demand Intelligence (BI).
  - Complex questions from integrated data.
  - "Intelligent Enterprise"

Reason-2: Why a Data Warehouse?

DBMS Approach

List of all items that were sold last month?
List of all items purchased by John?

The total sales of the last month grouped by branch?

How many sales transactions occurred during the month of January?

Reason-2: Why a Data Warehouse?

Intelligent Enterprise

Which items sell together? Which items to stock?

Where and how to place the items? What discounts to offer?

How best to target customers to increase sales at a branch?

Which customers are most likely to respond to my next promotional campaign, and why?

Reason-3: Why a Data Warehouse?

- Businesses want much more…
  - What happened?
  - Why it happened?
  - What will happen?
  - What is happening?
  - What do you want to happen?

Stages of Data Warehouse

What is a Data Warehouse?

A complete repository of historical corporate data extracted from transaction systems that is available for ad-hoc access by knowledge workers.

What is a Data Warehouse?

Transaction System

- Management Information System (MIS)
- Could be typed sheets (NOT transaction system)

Ad-Hoc access

- Does not have a certain access pattern.
- Queries not known in advance.
- Difficult to write SQL in advance.

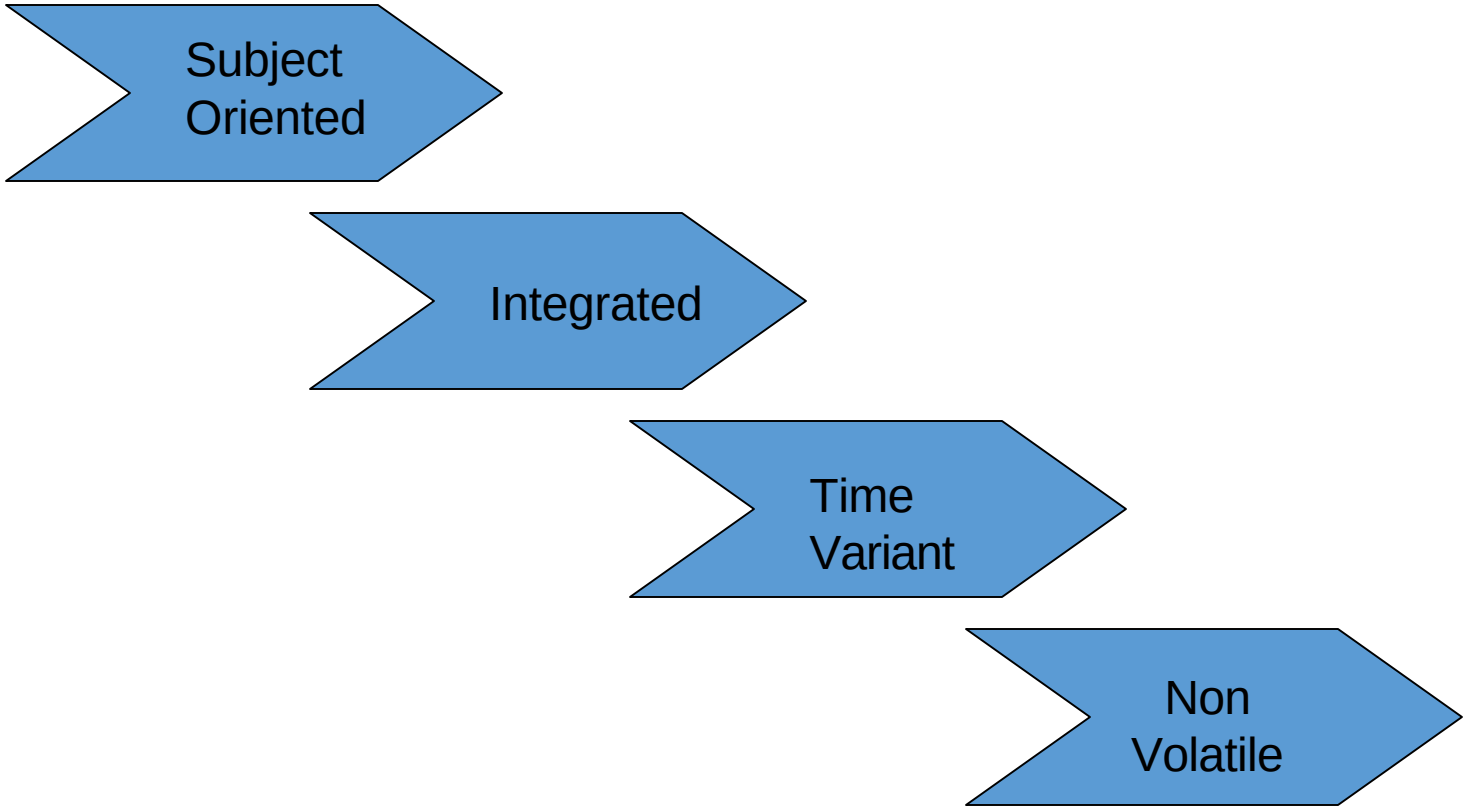
Knowledge workers

- Typically NOT IT literate (Executives, Analysts, Managers).
- NOT clerical workers.
- Decision makers.

What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision- making process."—W. H. Inmon

Another View of a DWH



Data Warehouse - Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.

Data Warehouse - Integrated

- Constructed by integrating multiple, heterogeneous data sources: relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.

When data is moved to the warehouse, it is converted.

Data Warehouse - Time Variant

- The time horizon for the data warehouse is significantly longer than that

  of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - However, the key of operational data may or may not contain “time element”.

Data Warehouse - Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not necessarily occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Often requires only two operations in data accessing: initial loading of data and access of data.

Another View of a DWH

A **data warehouse** is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.

The cost for historical data?

The factors affect to the cost of historical data:

- Industry.
- Cost of storing historical data.
- Economic value of historical data.

What is a Data Warehouse ?

It is a blend of many technologies, the basic concept being:

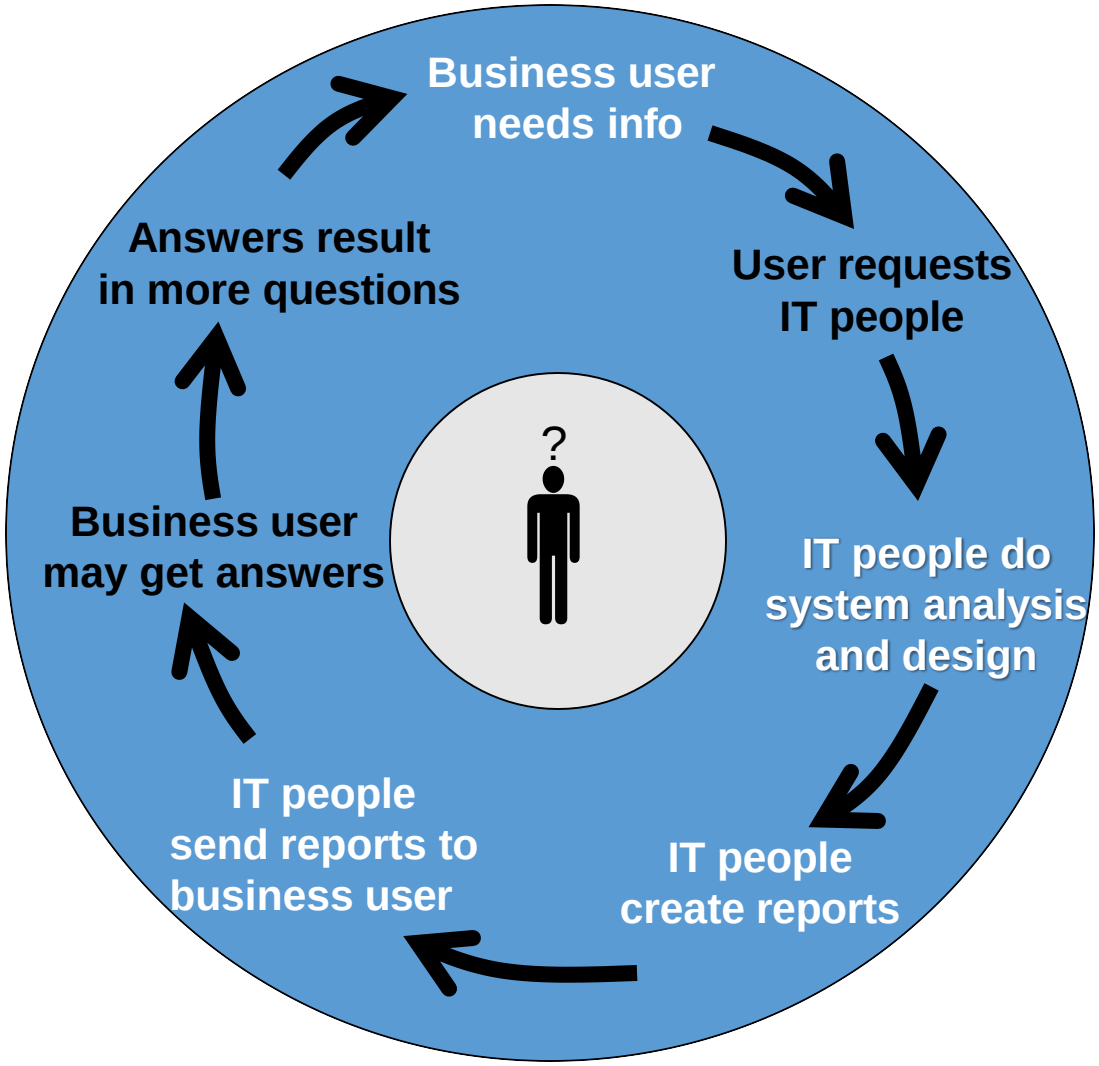
- Take all data from different operational systems.
- If necessary, add relevant data from industry.
- Transform all data and bring into a uniform format.
- Integrate all data as a single entity.

What is a Data Warehouse ? (Cont…)

It is a blend of many technologies, the basic concept being:

- Store data in a format supporting easy access for decision support.
- Create performance enhancing indices.
- Implement performance enhancement joins.
- Run ad-hoc queries with low selectivity.

- Fundamentally different



Data warehouse

- Combines operational and historical data.

  - Don’t do data entry into a DWH, OLTP or ERP are the source systems.
  - OLTP systems don’t keep history, cant get balance statement more than a year old.
  - DWH keep historical data, even of bygone customers. Why?
  - In the context of bank, want to know why the customer left?
  - What were the events that led to his/her leaving? Why?
  - Customer retention.

Rate of update depends on:

- volume of data,
- nature of business,
- cost of keeping historical data,
- benefit of keeping historical data.

Data Warehouse Vs. OLTP

OLTP (On Line Transaction Processing)	**DWH**
Primary key used	Primary key NOT used
No concept of Primary Index	Primary index used
Few rows returned	Many rows returned
May use a single table	Uses multiple tables
High selectivity of query	Low selectivity of query
Indexing on primary key (unique)	Indexing on primary index (non-unique)

Data Warehouse Vs. OLTP

OLTP: OnLine Transaction Processing (MIS or Database System)

	OLTP	Data Warehouse
Scope	* Application specific * Multiple databases with repetition * Off the shelf application * Runs the business	* Application –Neutral * Single source of "truth" * Evolves over time * How to improve business
Data Perspective	* Operational data * No summary * Fully normalized	* Historical, detailed data * Some summary * Lightly denormalized
Queries	* Based on PK * Number of results returned in hundreds	* Hardly uses PK * Number of results returned in thousands
Time factor	* Sub seconds to seconds * Typical availability 24x7	* Minutes to hours * Typical availability 6x12

Comparison of Response Times

- On-line analytical processing (OLAP) queries must be executed in a small number of seconds: Often requires denormalization and/or sampling.

- Complex query scripts and large list selections can generally be
executed in a small number of minutes.

- Sophisticated clustering algorithms (e.g., data mining) can generally be executed in a small number of hours (even for hundreds of thousands of customers).

Putting the pieces together