DataWarehousing

Past Business Scenario

- No Web for Business
  - Customers appear ”physically” in the store
  - Customers do not change to other stores more easily

Today's Business

- The Web came into existence
  - Customers do not appear ”physically” in the store
  - Customers can change to other stores more easily
- Thus Business Intelligence (BI) became more necessary
  - Can know customers using data and BI
  - Web logs makes is possible to analyze customer behavior in a more detailed than before (what was not bought?)
  - Combine web data with traditional customer data
- Wireless Internet adds further to this
  - Customers are always ”online”
  - Customer's position is known
  - Combine position and customer knowledge => very valuable

What Is Business Intelligence?

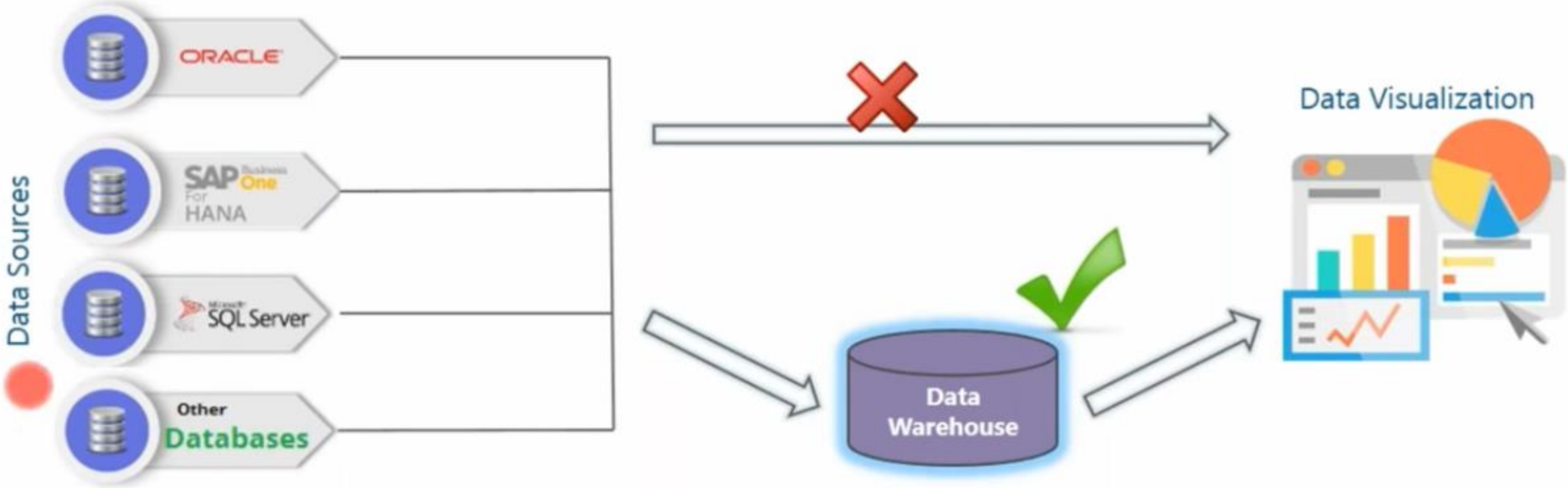
BI is the act of transforming raw/ operational data into useful information for business analysis.

How Does It Work?

1. BI based on Data Warehouse technology **extracts** information from a company's operational systems.
2. The data is **transformed** (*cleaned and integrated*), and **loaded** into Data Warehouses.
3. Since this data is credible, it is used for business insights.

Why Data Warehouse?

- Data collected from various sources & stored in various databases cannot be directly visualized.
- The data first needs to be **integrated** and then **processed** before visualization takes place.



Data Warehousing

- On October 11, 2000, find the 5 top-selling products for each product subcategory that contributes more than 20% of the sales within its product category?

- Regular database models and systems are not suitable for this type of queries

Why?

- Data is a valuable asset for ERP, etc.
- Data is available from every step of a sales pipeline, from
  - Internal Data Sources
    - ERP systems, Sales/Financials, Support/CRM, Marketing
  - External Data Sources
    - Social networks, Clickstreams, Websites, Supply chain/Logistics (Through customer support)

Why?

- Amount of data generated is very huge.
- Data comes from heterogeneous sources.
- It is difficult to build meaningful analytics with heterogeneous data sources

Why?

- Key Problems of Business

1) **Complex and unusable models**

Many DB models are difficult to understand

DB models do not focus on a single clear business purpose

2) **Same data found in many different systems**

Example: customer data in 14 systems

The same concept is defined differently

3) **Data is suited for operational systems**

Just for Accounting, billing, etc., Do not support analysis across business functions

4) **Data quality is bad**

Missing data, imprecise data, different use of systems

5) **Data are "volatile"**

Data deleted in operational systems (6 months), but Data change over time – no historical information

Enterprise Organizations Challenge

- Collect diverse kinds of data and maintain large databases from
  - Multiple, Heterogeneous and Distributed information sources
- Challenge
  - To integrate such data to provide easy and efficient access to it

To answer…

- Analyzing operations, to
  - Increase the customer focus
    - By the buying patterns of preference, time
  - Look for source of profit or to manage product portfolio
    - By comparing the performance of sales by quarter, year, geographic regions

Solution

- **New analysis environment**
- where data is

  Integrated (logically and physically)

  Subject oriented (versus function oriented)

  Supporting management decisions (different organization)

  Stable (data not deleted, several versions)

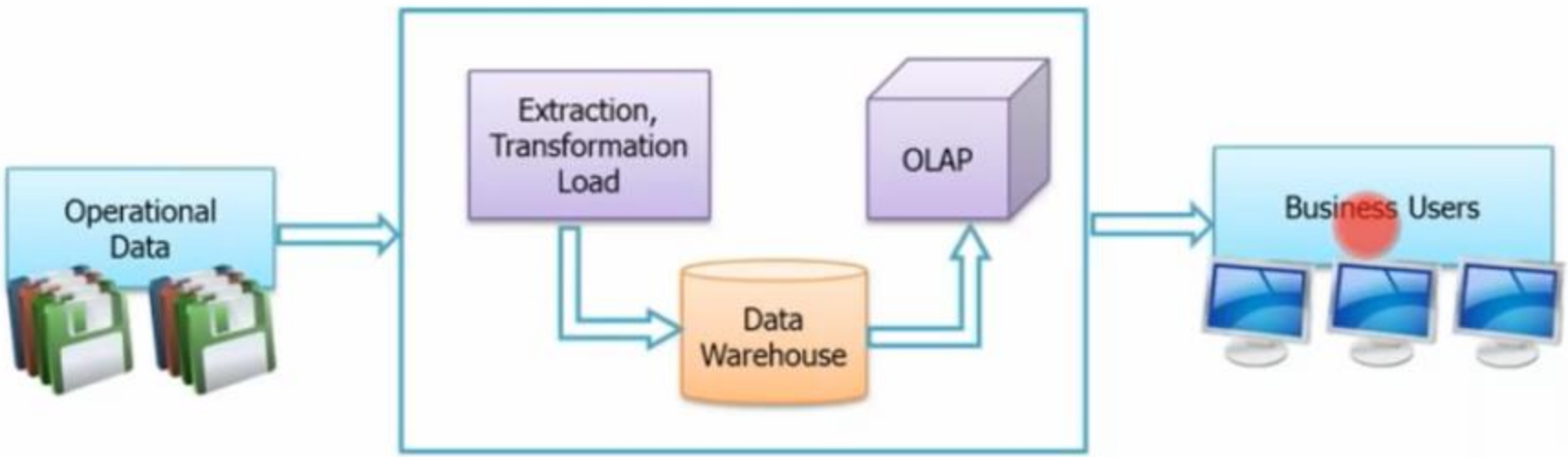
  Time variant (data can always be related to time)

Data Warehousing

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- From data warehousing to data mining

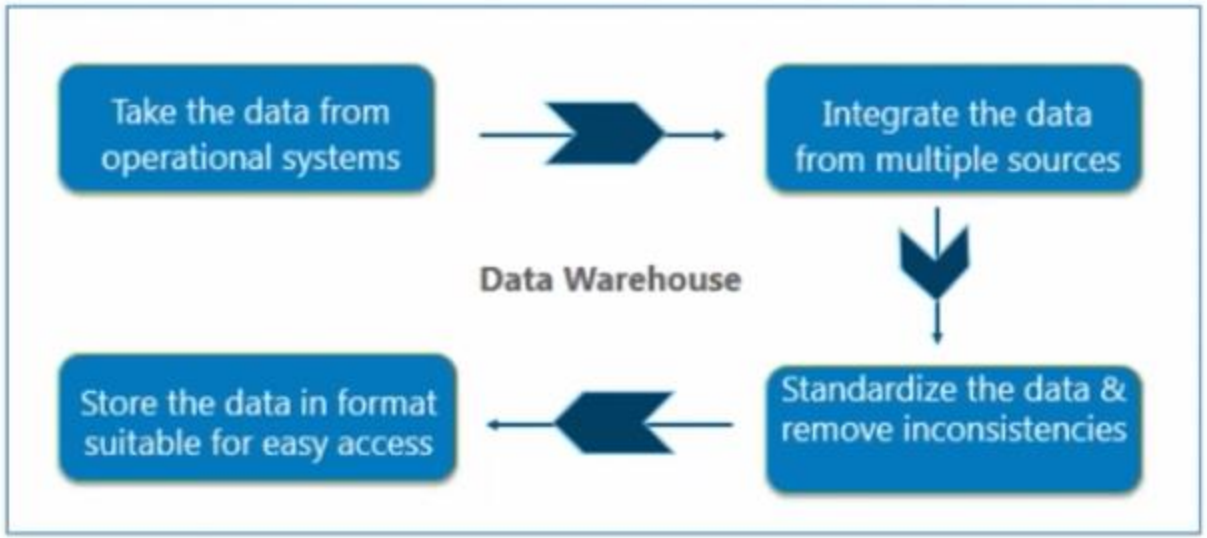
What Is A Data Warehouse?

- A central location where consolidated data from multiple locations (databases) are stored.
- DWH is maintained separately from an organization's operational database.
- End users access it whenever any information is needed.
- **Note**:- Data Warehouse is not loaded every time new data is added to database.



What Are The Advantages Of A Data Warehouse?

- Strategic questions can be answered by studying trends.
- Data Warehousing is faster and more accurate.
- **Note**:- Data Warehouse is not a product that a company can go and purchase, it needs to be designed & depends entirely on the company's requirement.



Characteristics of Data Warehouse

- Defined in many different ways
  - 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

Data Warehouse—Subject-Oriented

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

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources, like
  - Relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques applied, to ensure consistency in among different data sources
  - Naming conventions
  - Encoding structures
  - Attribute measures, etc.
    - 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

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- A data warehouse
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

Data Warehouse vs. Operational DBMS

- OLTP (On-Line Transaction Processing)
  - Major task of traditional relational DBMS
    - Day-to-day operations
      - Purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- Data Warehouse Systems
  - Known as OLAP (On-line Analytical Processing)

Data Warehouse vs. Operational DBMS

- User and system orientation
  - OLTP – Customer oriented
    - Transaction and query processed by clerks and clients
  - OLAP – Market oriented
    - Data analysis by knowledge workers – managers, executives and analysts
- Data contents for decision making
  - OLTP – Current data (Too detailed)
  - OLAP – Large amount of historical data
    - To provide summarization and aggregation
    - Stores and manages information at different level of granularity
- Database design
  - OLTP – ER Model and application oriented database design
  - OLAP – Star or Snowflake model and subject oriented database design

Data Warehouse vs. Operational DBMS

- View
  - OLTP – Current data within an enterprise or department
  - OLAP – Spans the multiple versions of a database schema due to evolutionary process of an organization and integrated from many data stores/ organizations.
    - Stored on multiple storage media
- Access patterns
  - OLTP – Consist of short, atomic transactions – Update
    - Need concurrency control and recovery
  - OLAP – Consist of MOSTLY read-only complex queries

OLTP vs. OLAP

	OLTP	OLAP
Users	Clerk, IT professional	Knowledge worker
Function	Day to day operations	Decision support
DB design	Application-oriented	Subject-oriented
Data	Current, up-to-date detailed, flat relational isolated	Historical, summarized, multidimensional integrated, consolidated
Usage	Repetitive	Ad-hoc
Access	Read/write index/hash on prim. key	Lots of scans
Unit of work	Short, simple transaction	Complex query
# Records accessed	Tens	Millions
#Users	Thousands	Hundreds
DB size	100MB-GB	100GB-TB
Metric	Transaction throughput	Query throughput, response