Unit 2 : Introduction to Data Warehousing

Data Warehouse

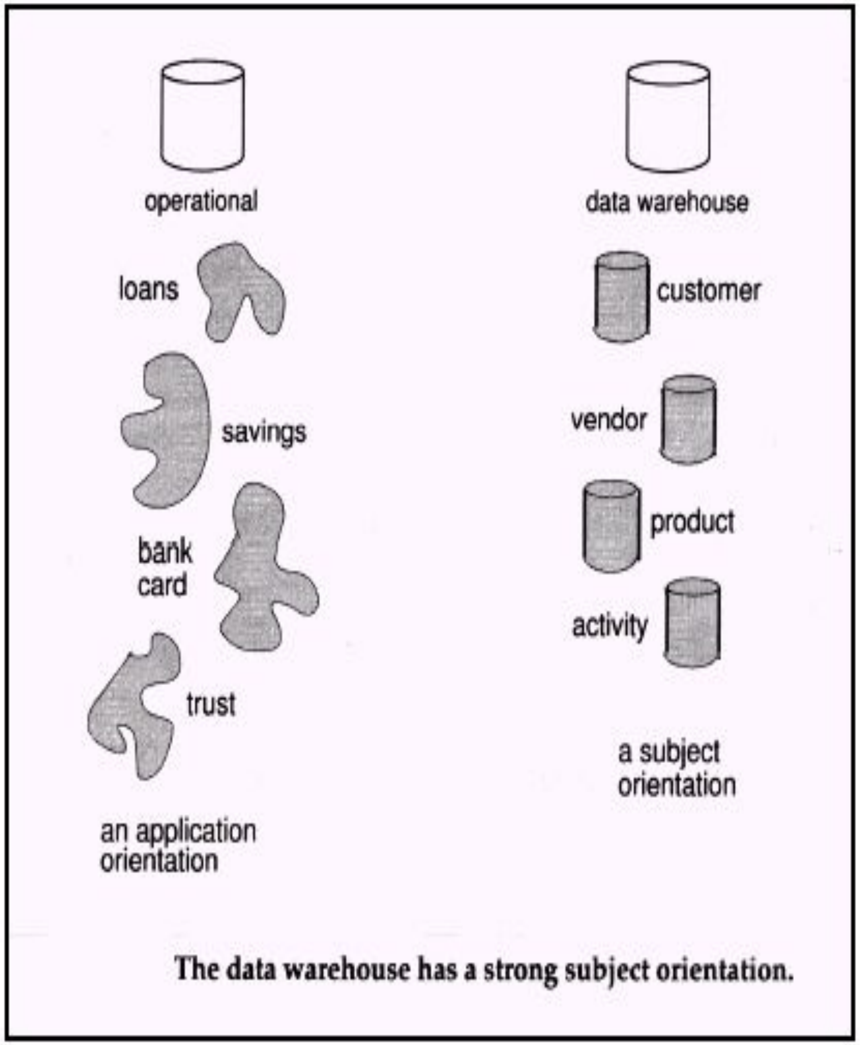
What is Data Warehouse?

- According to W. H. Inmon, a **data warehouse** is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
- “A data warehouse is a copy of transaction data specifically structured for querying and reporting” – Ralph Kimball
- **Data Warehousing** is the process of building a data warehouse for an organization.
- Data Warehousing is a process of transforming data into information and making it available to users in a timely enough manner to make a difference

Defining features:

- Four keywords:
  - subject-oriented,
  - integrated,
  - time-variant,
  - nonvolatile
- These keywords distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems and file systems.

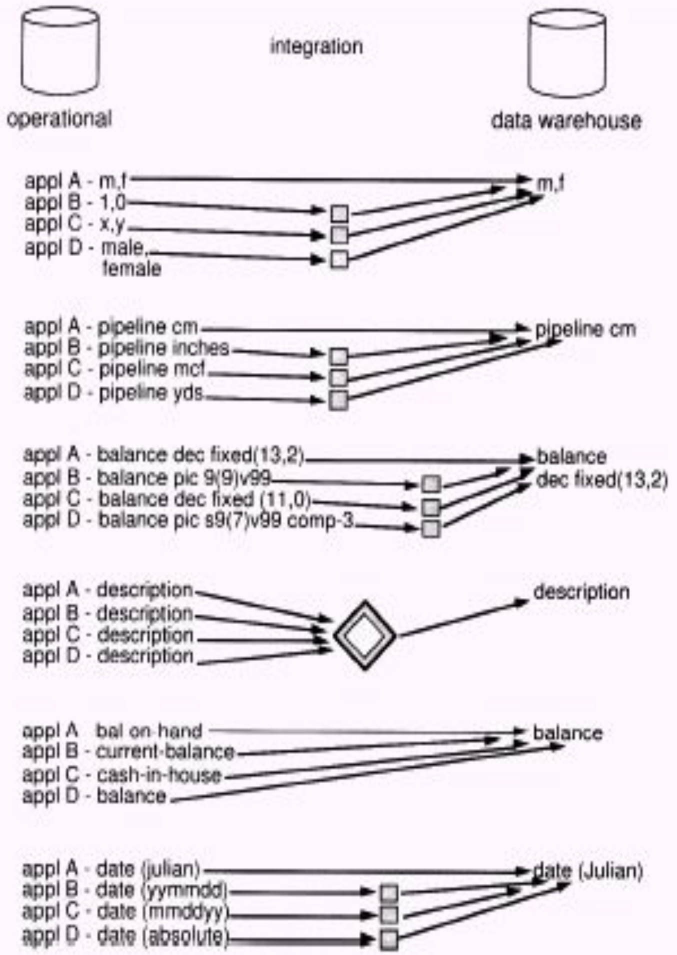
Subject Oriented



The data warehouse has a strong subject orientation.

- Focus is on Subject Areas rather than Applications
- 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.

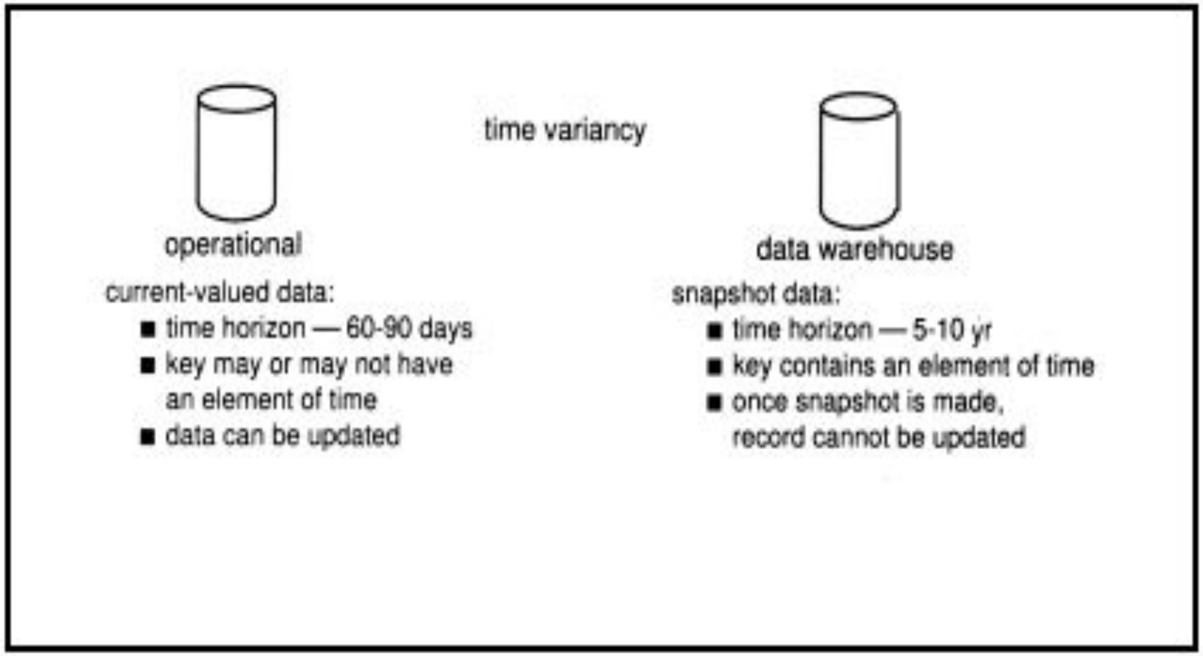
Integrated



When data is moved to the data warehouse from the application-oriented operational environment, the data is integrated before entering the warehouse.

- Constructed by integrating multiple, heterogeneous data sources
- Integration tasks handles naming conventions, physical attributes of data
- Must be made consistent.

Time Variant



current-valued data:

- time horizon — 60-90 days
- key may or may not have an element of time
- data can be updated

snapshot data:

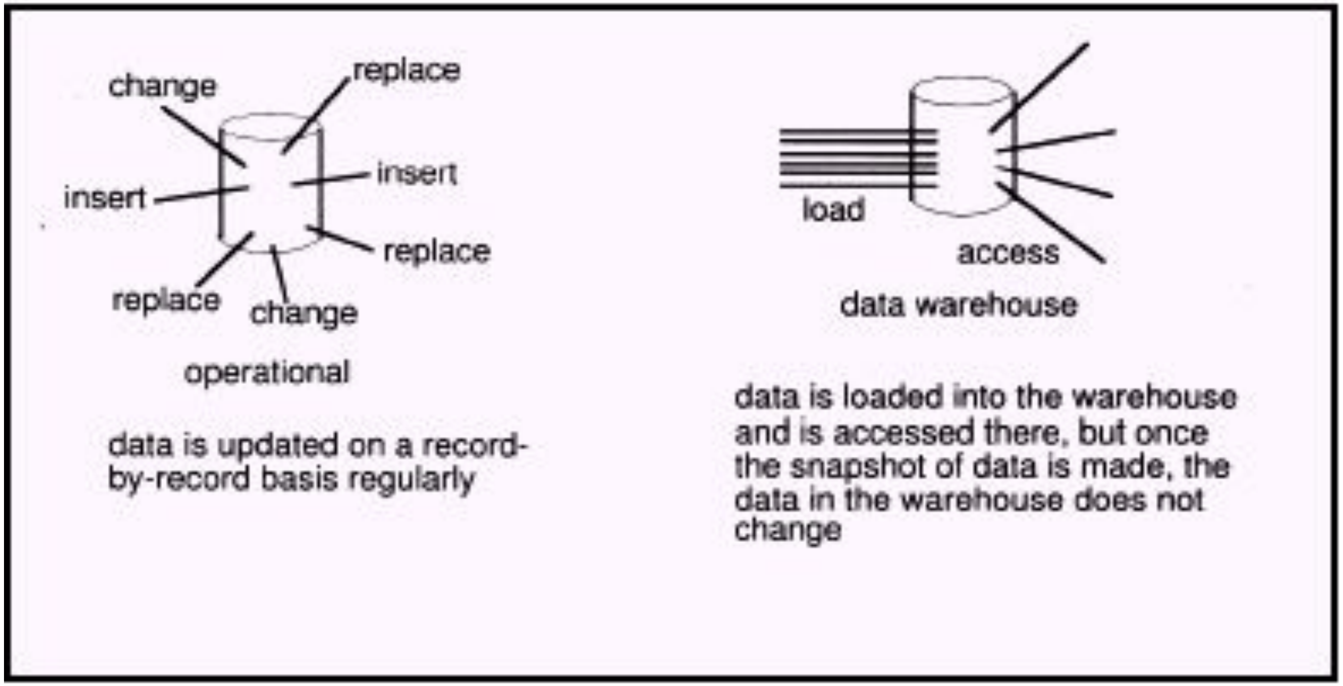
- time horizon — 5-10 yr
- key contains an element of time
- once snapshot is made, record cannot be updated

- Only accurate and valid at some point in time or over some time interval.
- The time horizon for the data warehouse is significantly longer than that of operational systems.

  Operational database provides current value data.

  Data warehouse data provide information from a historical perspective (e.g., past 5-10 years)

Non Volatile



data is updated on a record-by-record basis regularly

data is loaded into the warehouse and is accessed there, but once the snapshot of data is made, the data in the warehouse does not change

- Data Warehouse is relatively Static in nature.
- Not updated in real-time but data in the data warehouse is loaded and refreshed from operational systems, it is not updated by end users.

Data warehousing helps business managers to :

- Extract data from various source systems on different platforms
- Transform huge data volumes into meaningful information
- Analyze integrated data across multiple business dimensions
- Provide access of the analyzed information to the business users anytime anywhere

- **Data warehouse contains five types of data**
  - Older detail data,
  - Current detail data,
  - Lightly summarized data,
  - Highly summarized data, and
  - Meta data.
- **Goals of Data Warehousing**
  - To help reporting as well as analysis
  - Maintain organization's historical information
  - Be an adaptive and resilient source of information
  - Be the foundation for decision-making.

Need of Data Warehouse

- **Business user:** Business users require data warehouse to view summarized data from past. Since these people are non-technical, the data may be presented to them in a very simple form.
- **Store historical data:** Data warehouse is required to store the time variable data from past to be used for various purposes.
- **Make Strategic decisions:** Some strategies may be depending upon the data in data warehouse.
- **For data consistency and quality:** user can effectively undertake to bring the uniformity and consistency in data.
- **High response time:** Data warehouse has to be ready for fairly unexpected loads and types of queries, which demands a high degree of flexibility and quick response time.

BENEFITS OF IMPLEMENTING A DATA WAREHOUSE

- To provide a single version of truth about enterprise information.
- To speed up ad hoc reports and queries that involves aggregations across many attributes which are resource intensive.
- To provide a system in which managers that do not have a strong technical background are able to run complex queries.
- To provide a database that stores relatively clean data.
- To provide a database that stores historical data that may have been deleted from the OLTP systems.

BENEFITS OF IMPLEMENTING A DATA WAREHOUSE

- Improve data quality by providing consistent codes and descriptions, flagging or even fixing bad data.
- Provide the organization's information consistently.
- Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Add value to operational business applications, notably customer relationship manager (CRM) systems.
- Data warehouse helps to increase productivity and decrease computing costs.

- The benefits of the data warehouse can be sub-divided as
  - Tangible benefits
  - Intangible benefits
- **Tangible Benefits**
  - Cost of product comes down.
  - Better decisions in terms of cost and quality are taken.
  - Data warehouses have led to enhanced asset and liability management since it provides clear picture of enterprise wide purchasing and inventory patterns.
- **Intangible Benefits**
  - Improved productivity.
  - Enhanced customer relations.
  - Data warehouses enable re-engineering of business processes by providing useful insights into the work processes.

Benefits on successful implementation of Data Warehousing

- Queries do not impact Operational systems
- Provides quick response to queries for reporting
- Enables Subject Area Orientation
- Integrates data from multiple, diverse sources
- Enables multiple interpretations of same data by different users or groups
- Provides thorough analysis of data over a period of time
- Accuracy of Operational systems can be checked
- Provides analysis capabilities to decision makers

- Increase customer profitability
- Cost effective decision making
- Manage customer and business partner relationships
- Manage risk, assets and liabilities
- Integrate inventory, operations and manufacturing
- Reduction in time to locate, access, and analyze information (Link multiple locations and geographies)
- Identify developing trends and reduce time to market
- Strategic advantage over competitors

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers
- Provide reliable, High performance access
- Consistent view of Data: Same query, same data. All users should be warned if data load has not come in.
- Quality of data is a driver for business re-engineering.

Usage of Data Warehouse

- The traditional role of a data warehouse is to collect and organize historical business data so it can be analyzed to assist management in making business decisions.
- putting information technology to help the knowledge worker make faster and better decisions.
- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgements

Advantages of Data Warehousing

- Potential high Return on Investment (RoI)
- Competitive Advantage
- Increased productivity of corporate Decision Makers
- Standardizes data across an organization
- Smarter decisions for companies – moves towards fact-based decisions.
- Reduces costs- drops products that are not doing well
- Increases revenue – works on high selling products.

Problems in Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration

OLTP (Database) vs. Data Warehouse

- Online Transaction Processing (OLTP) systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
- OLTP applications normally automate clerical data processing tasks of an organization, like data entry and enquiry, transaction handling, etc. (access, read, update)
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  - e.g., *average amount spent on phone calls between 9AM-5PM in Kathmandu during the month of March, 2012*

- OLTP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User

- Data Warehouse
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)

- OLTP
  - Performance Sensitive
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB -100 GB

- Data Warehouse
  - Performance relaxed
  - Large volumes accessed at a time(millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes

- OLTP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety

- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

Difference between Operational System and Data Warehouse

Operational System	Data Warehouse
Holds current data	Holds historic data
Data is dynamic	Data is largely static
Read/Write accesses	Read only accesses
Repetitive processing	Ad hoc complex queries
Transaction driven	Analysis driven
Application oriented	Subject oriented
Used by clerical staffs for day-to-day operations	Used by top managers for analysis
Normalized data model (ER model)	De-normalized data model (Dimensional model)
Must be optimized for writes and small queries	Must be optimized for queries involving a large portion of the warehouse

Data Warehouse Applications

- Financial services
- Banking services
- Consumer goods Industry
- Retail sectors
- Controlled manufacturing
- Transportation Industry
- Telephone Industry
- Services Sector

Data Warehouse Applications

- The Retailers
- Manufacturing and Distribution Industry
- Insurance
- Hospitality Industry
- Healthcare
- Government and Education
- Biological data analysis
- Logistic and inventory management
- Trend analysis
- Agriculture

Warehouse Products

- Computer Associates -- CA-Ingres
- Hewlett-Packard -- Allbase/SQL
- Informix -- Informix, Informix XPS
- Microsoft -- SQL Server
- Oracle -- Oracle7, Oracle Parallel Server
- Red Brick -- Red Brick Warehouse
- SAS Institute -- SAS
- Software AG -- ADABAS
- Sybase -- SQL Server, IQ, MPP

Thank you !!!