Chapter 2
Introduction to Data Warehousing

Data Warehouse

Definitions:

According to W. H. Inmon, who is also known as "father of the data warehouse", a **data warehouse** is defined as "a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions".

According to Ralph Kimball - "A data warehouse is a copy of transaction data specifically structured for querying and reporting"

Data Warehouses a single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

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.

A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema.

Defining features:

In most of the organization, there occur large databases in operation for normal daily transactions called operational database. A data warehouse is a large database built from the operational database. The four keywords, subject-oriented, integrated, time-variant, nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems and file systems.

A data warehouse should be:

i. Subject Oriented

- Not all the information in the operational database is useful for a data warehouse.
- 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.
- A data warehouse should be designed especially for decision support and expert system with specific related data.

ii. Integrated

- In an operational data, many types of information being used with different names for same entity.
- In a data warehouse, all entities should be integrated and consistent i.e. only one name must exist to describe each individual entity.
- Constructed by integrating multiple, heterogeneous data sources
- Integration tasks handles naming conventions, physical attributes of data

iii. Time – variant

- There must be a connection between the information in the warehouse and the time when it was entered.
- One of the most important aspect of the warehouse as it relates to data mining, because information can then be sourced according to period.
- 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)

iv. Non-Volatile

- Data in a warehouse is not updated in real-time, but data in the data warehouse is loaded and refreshed from operational systems updated and used only for queries.
- End-users who want to update data must use operational database.
- A data warehouse will always be filled with historical data.
- Data Warehouse is relatively static in nature.

Data Warehouses are an important asset for organizations to maintain

- Efficiency,
- Profitability,
- Competitive advantages.

Data Warehouse can be viewed as an information system with the following attributes

- It is database designed for analytical tasks and investigative tasks, using data from multiple applications.
- It supports a relatively small number of users with relatively long interactions.
- It contains current and historical data to provide a historical perspective of information.
- Its contents are periodically updated.
- Its usage is read-intensive.
- It contains a few large tables.

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

Data warehouse is needed for the following reasons

1. **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.
2. **Store historical data:** Data warehouse is required to store the time variable data from past. This data is made to be used for various purposes.

3. **Make Strategic decisions:** Some strategies may be depending upon the data in data warehouse. So, data warehouses contribute in making strategic decisions.
4. **For data consistency and quality:** Bringing the data from different sources at a commonplace, user can effectively undertake to bring the uniformity and consistency in data.
5. **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 WAREHO USE

The benefits of implementing a data warehouse are as follows

1. To provide a single version of truth about enterprise information. This may appear rather obvious but it is not uncommon in an enterprise for two database systems to have two different versions of the truth.
2. To speed up ad hoc reports and queries that involves aggregations across many attributes which are resource intensive. The managers require trends, sums and aggregations that allow, for example, comparing this year's performance to last year's or preparation of forecasts for next year.
3. To provide a system in which managers that do not have a strong technical background are able to run complex queries. If the managers are able to access the information they require, it is likely to reduce the bureaucracy around the managers.
4. To provide a database that stores relatively clean data. By using a good ETL process, the data warehouse should have data of high quality. When errors are discovered it may be desirable to correct them directly in the data warehouse and then propagate the collections to the OLTP systems.
5. To provide a database that stores historical data that may have been deleted from the OLTP systems. To improve response time, historical data is usually not retained in OLTP systems other than that which is required to respond to customer queries. The data warehouse can then store the data that is eliminated from the OLTP systems.
6. Improve data quality by providing consistent codes and descriptions, flagging or even fixing bad data.
7. Provide the organization's information consistently.
8. Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
9. Add value to operational business applications, notably customer relationship manager (CRM) systems.
10. Data warehouse helps to increase productivity and decrease computing costs.

The benefits of the data warehouse can be sub-divided as

i. Tangible benefits
ii. Intangible benefits

Tangible Benefits

Successfully implemented data warehouse can realize same tangible benefits such as:

1. Cost of product comes down.
2. Better decisions in terms of cost and quality are taken.
3. Data warehouses have led to enhanced asset and liability management since it provides clear picture of enterprise wide purchasing and inventory patterns.

Intangible Benefits

1. Improved productivity.
2. Enhanced customer relations.
3. Data warehouses enable re-engineering of business processes by providing useful insights into the work processes.

The successful implementation of a data warehouse can bring major, benefits to an organization including:

- **Potential high returns on investment**
  Implementation of data warehousing by an organization requires a huge investment typically from Rs 10 lack to 50 lacks. However, a study by the International Data Corporation (IDC) in 1996 reported that average three-year returns on investment (RO I) in data warehousing reached 401%.
- **Competitive advantage**
  The huge returns on investment for those companies that have successfully implemented a data warehouse are evidence of the enormous competitive advantage that accompanies this technology. The competitive advantage is gained by allowing decision-makers access to data that can reveal previously unavailable, unknown, and untapped information on, for example, customers, trends, and demands.
- **Increased productivity of corporate decision-makers**
  Data warehousing improves the productivity of corporate decision-makers by creating an integrated database of consistent, subject-oriented, historical data. It integrates data from multiple incompatible systems into a form that provides one consistent view of the organization. By transforming data into meaningful information, a data warehouse allows business managers to perform more substantive, accurate, and consistent analysis.
- **More cost-effective decision-making**
  Data warehousing helps to reduce the overall cost of the· product· by reducing the number of channels.
- **Better enterprise intelligence.**
  It helps to provide better enterprise intelligence.
- **Enhanced customer service.**
  It is used to enhance customer service.
- **Provide quick response to queries**
- **Enables subject area orientation**
- **Integrates data from multiple and diverse source**
- **Enables multiple interpretations of same data by different users or groups**
- **Provides 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**
- **Reduction in time to locate access and analyze information.**
- **Identify developing trends and reduce time to market.**
- **Strategic advantage over competitors**
- **Potential high retains on investments.**

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 judgments

Advantages of Data Warehousing

- Potential high Return on Investment
- 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 with Data Warehousing

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

Problems of Data Warehousing

The problems associated with developing and managing a data warehousing are as follows:

Underestimation of resources of data loading

Sometimes we underestimate the time required to extract, clean, and load the data into the warehouse. It may take the significant proportion of the total development time, although some tools are there which are used to reduce the time and effort spent on this process.

Hidden problems with source systems

Sometimes hidden .problems associated with the source systems feeding the data warehouse may be identified after years of being undetected. For example, when entering the details of a new property, certain fields may allow nulls which may result in staff entering incomplete property data, even when available and applicable.

Required data not captured
In some cases the required data is not captured by the source systems which may be very important for the data warehouse purpose. For example the date of registration for the property may be not used in source system but it may be very important analysis purpose.

Increased end-user demands
After satisfying some of end-users queries, requests for support from staff may increase rather than decrease. This is caused by an increasing awareness of the users on the capabilities and value of the data warehouse. Another reason for increasing demands is that once a data warehouse is online, it is often the case that the number of users and queries increase together with requests for answers to more and more complex queries.

Data homogenization
The concept of data warehouse deals with similarity of data formats between different data sources. Thus, results in to lose of some important value of the data.

High demand for resources
The data warehouse requires large amounts of data.

Data ownership
Data warehousing may change the attitude of end-users to the ownership of data. Sensitive data that owned by one department has to be loaded in data warehouse for decision making purpose. But some time it results in to reluctance of that department because it may hesitate to share it with others.

High maintenance
Data warehouses are high maintenance systems. Any reorganization· of the business processes and the source systems may affect the data warehouse and it results high maintenance cost.

Long-duration projects
The building of a warehouse can take up to three years, which is why some organizations are reluctant in investigating in to data warehouse. Some only the historical data of a particular department is captured in the data warehouse resulting data marts. Data marts support only the requirements of a particular department and limited the functionality to that department or area only.

Complexity of integration
The most important area for the management of a data warehouse is the integration capabilities. An organization must spend a significant amount of time determining how well the various different data warehousing tools can be integrated into the overall solution that is needed. This can be a very difficult task, as there are a number of tools for every operation of the data warehouse.

OPERATIONAL SYSTEMS VS DATA WAREHOUSING SYSTENNS
The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems
Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in

order to accommodate the diverse needs of the different users. These systems are known as analytical processing (OLAP) systems.

Difference between Operational System and Data Warehouse

Operational System	Data Warehouse
1. Holds current data	1. Holds historic data
2. Data is dynamic	2. Data is largely static
3. Read/Write accesses	3. Read only accesses
4. Repetitive processing	4. Ad hoc complex queries
5. Transaction driven	5. Analysis driven
6. Application oriented	6. Subject oriented
7. Used by clerical staffs for day-to-day operations	7. Used by top managers for analysis
8. Normalized data model (ER model)	8. De-normalized data model (Dimensional model)
9. Must be optimized for writes and small queries	9. Must be optimized for queries involving a large portion of the warehouse

DBMS VS DATAWAREHOUSE

Database	Data Warehouse
1. Used for Online Transactional Processing (OLTP) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.	1. Used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.
2. The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.	2. The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
3. Entity – Relational modeling techniques are used for RDMS database design	3. Data – Modeling techniques are used for the Data Warehouse design.
4. Optimized for write operation.	4. Optimized for read operations.
5. Performance is low for analysis queries.	5. High performance for analytical queries.
6. Database is the place where the data is taken as a base and managed to get available fast and efficient access.	6. Data warehouse is the place where the application data is managed for analysis and reporting purpose.
7. Is designed to record	7. Is designed to analyze
8. Is an application-oriented collection of data	8. Is a subject-oriented collection of data
9. Normally limited to a single application	9. Stores data from any number of applications
10. Data is available real-time	10. Data is refreshed from source systems when needed
11. Is efficient in processing and storage	11. Is efficient in analytics
12. Size ranges from 100MB – 100GB	12. Size ranges from 100GB – few TB
13. Thousands of users	13. Hundreds of users
14. Specifies current data	14. Specifies both current and historical data
15. Data are isolated	15. Data are integrated

Data Warehouse Applications

Data warehouses are widely used in the following fields:

- Financial services
- Banking services
- Consumer goods Industry
- Retail sectors
- Controlled manufacturing
- Transportation Industry
- Telephone Industry
- Services Sector
- The Retailers
- Manufacturing and Distribution Industry
- Insurance
- Hospitality Industry
- Healthcare
- Government and Education
- Biological data analysis
- Logistic and inventory management
- Trend analysis
- Agriculture

Applications of Data Warehouse:

Data Warehouses owing to their potential have deep-rooted applications in every industry which use historical data for prediction, statistical analysis, and decision making. Listed below are the applications of Data warehouses across innumerable industry backgrounds. So here are the various applications of Data Warehouse.

- **Banking Industry:**

  In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.

  Most banks also use warehouses to manage the resources available on deck in an effective manner. Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.

  Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity. Apart from all these, there is also scope for co-branding.

- **Finance Industry**

  Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

- **Consumer Goods Industry**

  They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged business partners and clientele.

- **Government and Education**

  The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.

  The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers, and also their entire criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.

  Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management. The entire financial department of most universities depends on data warehouses, inclusive of the Financial Aid department.

- **Healthcare**

  One of the most important sectors which utilize data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

- **Hospitality Industry**

  A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

- **Insurance**

  As the saying goes in the insurance services sector, "Insurance can never be bought, it can be only be sold", the warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants. The design of tailor-made customer offers and promotions is also possible through warehouses.

- **Manufacturing and Distribution Industry**

  This industry is one of the most important sources of income for any state. A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.

  They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.

  For the distributions, the supply chain management of products operates through data warehouses.

- **The Retailers**

  Retailers serve as middlemen between producers and consumers. It is important for them to maintain records of both the parties to ensure their existence in the market.

They use warehouses to track items, their advertising promotions, and the consumers buying trends. They also analyze sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.

- **Services Sector**

  Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources.

- **Telephone Industry**

  The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.

  Apart from those operations, analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.

- **Transportation Industry**

  In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.

  The internal environment of the industry uses them to analyze customer feedback, performance, manage crews on board as well as analyze customer financial reports for pricing strategies.