DATA WAREHOUSING

INTRODUCTION

Data warehouses usually contain historical data derived from transaction data, but it can include data from other sources. Also, it separates analysis work load from transaction workload and enables an organization to consolidate data from several sources.

Definition of Data Warehouse

A **Data Warehousing** (DW)-is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

It is a blend of technologies and components which aids the strategic use of data. It is also a electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

Other definitions of data warehouse include:

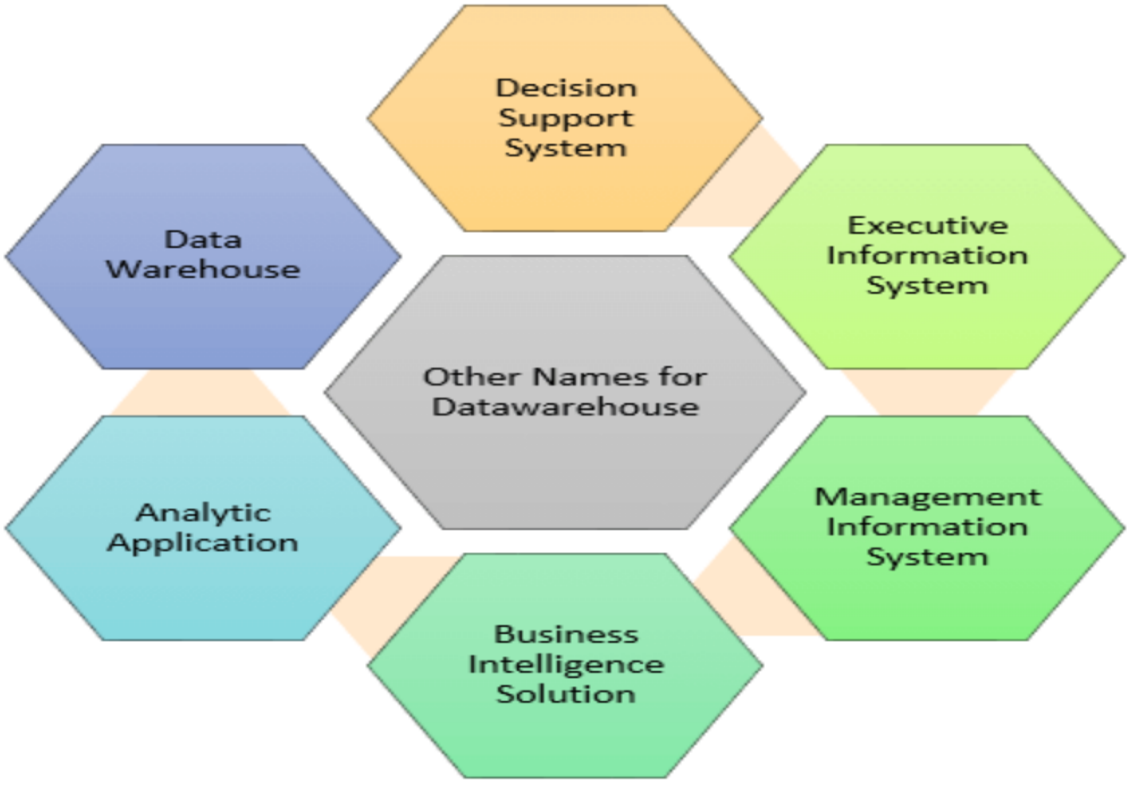
1. A data warehouse is a data structure that is optimized for distribution. It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts.

2. A data warehouse is that portion of an overall is architected data environment that serves as the single integrated source of data for processing information.

3. Data warehouse is a repository of an organization's electronically stored data designed to facilitate reporting and analysis.

Data warehouse system is also known by the following name:

- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse



How Datawarehouse works?

A Data Warehouse works as a central repository where information arrives from one or more data sources. Data flows into a data warehouse from the transactional system and other relational databases.

Data may be:

1. Structured
2. Semi-structured
3. Unstructured data

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.

Types of Data Warehouse

a) Enterprise Data Warehouse:

Enterprise Data Warehouse is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according

to those divisions

b) Operational Data Store:

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

c) Data Mart:

A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

General stages of Data Warehouse

Earlier, organizations started relatively simple use of data warehousing. However, over time, more sophisticated use of data warehousing begun.

The following are general stages of use of the data warehouse:

a. **Offline Operational Database:**

   In this stage, data is just copied from an operational system to another server. In this way, loading, processing, and reporting of the copied data do not impact the operational system's performance.

b. **Offline Data Warehouse:**

   Data in the Datawarehouse is regularly updated from the Operational Database. The data in Datawarehouse is mapped and transformed to meet the Datawarehouse objectives

c. **Real time Data Warehouse:**

   In this stage, Data warehouses are updated whenever any transaction takes place in operational database. For example, Airline or railway booking system.

d. **Integrated Data Warehouse:**

   In this stage, Data Warehouses are updated continuously when the operational system performs a transaction. The Datawarehouse then generates transactions which are passed back to the operational system.

Components of Data warehouse

Load manager: Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.

Warehouse Manager: Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to

ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving and baking-up data.

Query Manager: Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.

Steps to Implement Data Warehouse

The best way to address the business risk associated with a Datawarehouse implementation is to employ a three-prong strategy as below

1. **Enterprise strategy**: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.
2. **Phased delivery**: Datawarehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.
3. **Iterative Prototyping**: Rather than a big bang approach to implementation, the Datawarehouse should be developed and tested iteratively.

Here, are key steps in Datawarehouse implementation along with its deliverables

Step	Tasks	Deliverables
1	Need to define project scope	Scope Definition
2	Need to determine business needs	Logical Data Model
3	Define Operational Datastore requirements	Operational Data Store Model
4	Acquire or develop Extraction tools	Extract tools and Software
5	Define Data Warehouse Data requirements	Transition Data Model
6	Document missing data	To Do Project List
7	Maps Operational Data Store to Data Warehouse	D/W Data Integration Map
8	Develop Data Warehouse Database design	D/W Database Design
9	Extract Data from Operational Data Store	Integrated D/W Data Extracts
10	Load Data Warehouse	Initial Data Load
11	Maintain Data Warehouse	On-going Data Access and Subsequent Loads

Best practices to implement a Data Warehouse

1. Decide a plan to test the consistency, accuracy, and integrity of the data.
2. The data warehouse must be well integrated, well defined and time stamped.
3. While designing Datawarehouse make sure you use right tool, stick to life cycle, take care about data conflicts and ready to learn you're your mistakes.
4. Never replace operational systems and reports

5. Don't spend too much time on extracting, cleaning and loading data.
6. Ensure to involve all stakeholders including business personnel in Datawarehouse implementation process. Establish that Data warehousing is a joint/ team project. You don't want to create Data warehouse that is not useful to the end users.
7. Prepare a training plan for the end users.

Advantages of Data Warehouse:

1. Data warehouse allows business users to quickly access critical data from some sources all in one place.
2. Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
3. Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
4. Data warehouse helps to reduce total turnaround time for analysis and reporting.
5. Restructuring and Integration make it easier for the user to use for reporting and analysis.
6. Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
7. Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

Disadvantages of Data Warehouse:

1. Not an ideal option for unstructured data.
2. Creation and Implementation of Data Warehouse is surely time confusing affair.
3. Data Warehouse can be outdated relatively quickly
4. Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
5. The data warehouse may seem easy, but actually, it is too complex for the average users.
6. Despite best efforts at project management, data warehousing project scope will always increase.
7. Sometime warehouse users will develop different business rules.
8. Organizations need to spend lots of their resources for training and implementation purpose.

Goals of Data Warehouse

The major goals of data warehousing are stated as follows:

i. To facilitate reporting as well as analysis
ii. Maintain an organizations historical information
iii. Be an adaptive and resilient source of information
iv. Be the foundation for decision making.

Characteristics of Data Warehouse

The characteristics of a data warehouse as set forth by William Inmon are stated as follows:

- Subject–oriented
- Integrated

- Nonvolatile
- Time variant

i) Subject-Oriented

The main objective of storing data is to facilitate decision process of a company, and within any company data naturally concentrates around subject areas. This leads to the gathering of information around these subjects rather than around the applications or processes (Muhammad, A.S.)

ii) Integrated

The data in the data warehouses are scattered around different tables, databases or even servers. Data warehouses must put data from different sources into a consistent format.

They must resolve such problems as naming conflicts and inconsistencies among units of measure. When this is achieved, they are said to be integrated.

iii) Non-Volatile

Non-volatile means that information in the data warehouse does not change each time an operational process is executed. Information is consistent regardless of when and how the warehouse is accessed.

iv) Time-Variant

The value of operational data changes on the basis of time. The time based archival of data from operational systems to data warehouse makes the value of data in the data warehouses to be a function of time. As data warehouse gives accurate picture of operational data for some given time and the changes in the data in warehouse are based on time- based change in operational data, data in the data warehouse is called time-variant.

Other characteristics outside the definition of William Inmon are:

Accessibility: the primary purpose of a data warehouse is to provide readily accessible information to end-user.
Process-Oriented: data warehousing can be viewed as the process of delivering information; and the maintenance of a data warehouse is continuous and iterative in nature.

Data Warehouse Components

The major components of a data warehouse are:

1. Summarized data
2. Operational systems of record
3. Integration/Transformation programs
4. Current detail
5. Data warehouse architecture or metadata
6. Archives.

Approaches for Storing Data in a Warehouse

There are two leading approaches to storing data in a data warehouse. These are:

- The dimensional approach
- The normalized approach

i) **Dimensional Approach:** in dimensional approach, transaction data are partitioned into either –facts, which are generally numeric transaction data or –dimensions which are the reference information which gives context to the facts. For example, a sales transaction can be broken up into facts such as order date, customer's name, product number, order ship-to and bill-to locations and salesperson responsible for receiving the order.

Benefits of Dimensional Approach

1. This approach makes the data warehouse easier for the user to understand and to use.
2. The retrieval of data from the data warehouse tends to operate very quickly.

Disadvantages of Dimensional Approach

1. In order to maintain the integrity of facts and dimensions, loading the data warehouse with data from different operational systems is complicated.

2. It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business.

ii) **The Normalized Approach:** In this approach, the data in the data warehouse are stored following to a degree and database normalization rules. Tables are grouped together by subject areas that reflect general data categories e.g. data on customers, products, finances.

Benefits of Normalized Approach

The major benefits derived from this approach is that it is straight forward to add information into the database

Disadvantages of Normalized Approach

Because of the number of tables involved, it can be difficult for users to both join data from different sources into meaningful information and then access the information without a precise understanding of the sources of data and of the data structure of the data warehouse.

Data Warehouse Users

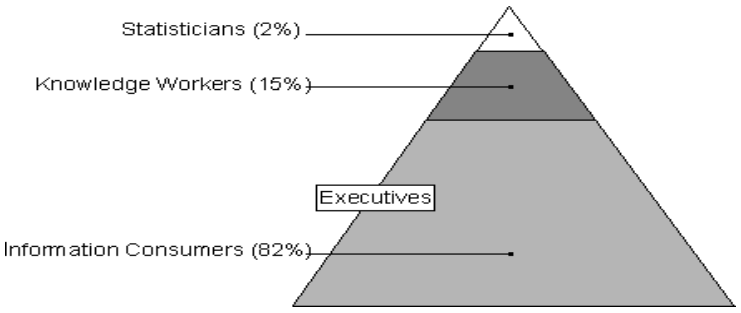
The successful implementation of a data warehouse is measured solely by its acceptance by users. Without users, historical data might as well be achieved by magnetic tape and stored in the basement. Successful data warehouse design starts with understanding the users and

their needs.

Data warehouse users can be divided into four categories:

1. Statisticians
2. Knowledge workers
3. Information consumers
4. Executives

Each makes up a portion of the user population as illustrated in this diagram



The User Pyramid
Source: Dave Browning & Joy Mundy, Dec. 2001

1. Statisticians

There are usually a handful of sophisticated analysts comprising of statisticians and operations research types in any organization. Though they are few in number but are best users of the data warehouse, those whose work can contribute to closed loop systems that deeply influence the operations and profitability of the company. It is vital that these users come to love the data warehouse. Generally, that is not difficult: these people are often very self-sufficient and need only to be pointed to the database and given some simple instruction about how to get to the data and what times of the day are best for performing large queries to retrieve data to analyze using their own sophisticated tools.

2. Knowledge Workers

A relatively small number of analysts perform the bulk of new queries and analysis against the data warehouse. These are the users who get the –designer or –analyst versions of user access tools. They figure out how to quantify a subject area. After a few iterations, their queries and reports typically get published for the benefit of the information consumers. Knowledge workers are often deeply engaged with the data warehouse design and place the greatest demands on the ongoing data warehouse operations team from training and support.

3. **Information Consumers**

Most users of the data warehouse are information consumers; they will probably never compose a true and ad-hoc query. They use static or simple interactive reports that others have developed. It is easy to forget about these users, because they usually interact with the data warehouse only through the work product of others. Do not neglect these users. This group includes a large number of people, and published reports are highly visible. Set up a great communication infrastructure for distributing information widely, and gather feedback from these users to improve the information sites over time.

4. **Executives**

Executives are a special case of the information customer group. Few executives actually issue their own queries, but an executive's slightest thought can generate an outbreak of activity among the other types of users. An intelligent data warehouse designer/implementer or owner will develop a very cool digital dashboard for executives, assuming it is easy and economical to do so. Usually this should follow other data warehouse work, but it never hurts to impress the bosses

How Users Query the Data Warehouse

Information for users can be extracted from the data warehouse relational database or from the output of analytical services such as OLAP or data mining. Direct queries to the data warehouse relational database should be limited to those that cannot be accomplished through existing tools, which are often more efficient than direct queries and impose less load on the relational database.

Reporting tools and custom applications often access the database directly. Statisticians extract data for use by special analytical tools. Analysts may write complex queries to extract and compile specific information not readily accessible through existing tools. Information consumers do not interact directly with the relational database but may receive e-mail reports or access web pages that expose data from the relational database. Executives use standard reports or ask others to create specialized reports for them. When using the Analysis Services Tools in SQL servers 2000, Statisticians will often perform data mining, analysts will write MDX queries against OLAP cubes and use data mining, and information consumers will use interactive reports designed by others.

Applications of Data Warehouse

Some of the areas where data warehousing can be applied are stated as follows:

1. Credit card churn analysis
2. Insurance fraud analysis
3. Call record analysis
4. Logistics management