

CLOUD APPLICATION DEVELOPMENT (GROUP 1)

PHASE 1 : ASSIGNMENT NOTEBOOK SUBMISSION

**NAME : HEMANATHAN .M**

**EMAIL : chandrueasuki1@gmail.com**

**GitHub Repository URL : https://github.com/hemanathan457/Hemanathan.git**

**Title of the Project**

*PROJECT 7 : Data warehousing*

**ABSTRACT**

Data warehouse projects have special requirements for the physical architecture of the database system. These requirements distinguish data warehouse projects from operational data stores and are often underestimated. This chapter covers topics such as hardware optimization, optimization of the operating system, a “sales-pitch” for a dedicated data warehouse infrastructure (as opposed to adding the data warehouse to the existing, operational infrastructure), and some background information on hardware and database options. It also includes how to set up each individual layer of the data warehouse and the options available for each layer.

**History of Datawarehouse**

The Datawarehouse benefits users to understand and enhance their organization’s performance. The need to warehouse data evolved as computer systems became more complex and needed to handle increasing amounts of Information. However, Data Warehousing is a not a new thing.

Here are some key events in evolution of Data Warehouse-

1960- Dartmouth and General Mills in a joint research project, develop the terms dimensions and facts.

1970- A Nielsen and IRI introduces dimensional data marts for retail sales.

1983- Tera Data Corporation introduces a database management system which is specifically designed for decision support

Data warehousing started in the late 1980s when IBM worker Paul Murphy and Barry Devlin developed the Business Data Warehouse.

However, the real concept was given by Inmon Bill. He was considered as a father of data warehouse. He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory.

**Data warehouse Work**

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:**

Structured

Semi-structured

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.

**Benefits of data warehouse**

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

Integrate data from multiple sources into a single database and data model. More congregation of data to single database so a single query engine can be used to present data in an ODS.

Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long-running analysis queries in transaction processing databases.

Maintain data history, even if the source transaction systems do not.

Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.

Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.

Present the organization's information consistently.

Provide a single common data model for all data of interest regardless of the data's source.

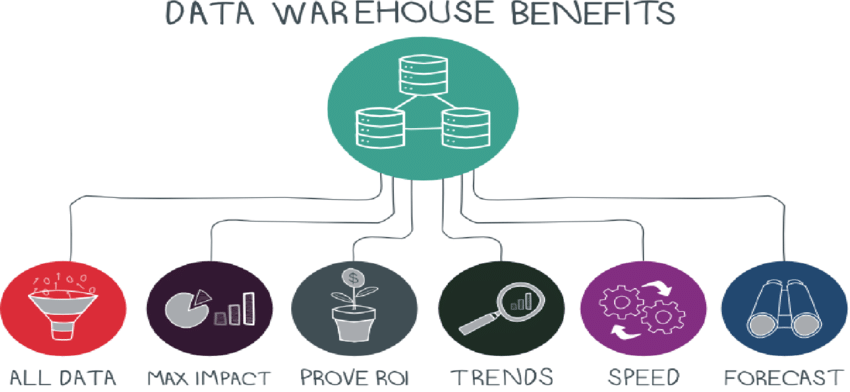
Restructure the data so that it makes sense to the business users.

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 management (CRM) systems.

Make decision–support queries easier to write.

Organize and disambiguate repetitive data.

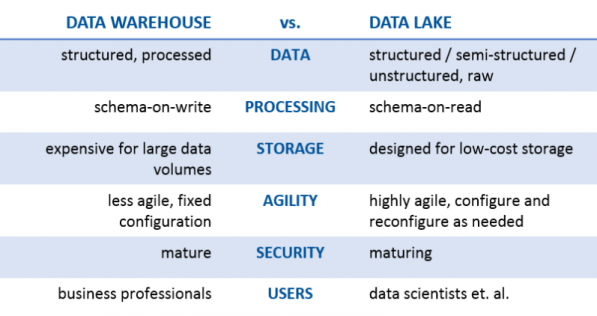


**Related systems (data mart, OLAP, OLTP, predictive analytics)**

A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), hence they draw data from a limited number of sources such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. The sources could be internal operational systems, a central data warehouse, or external data.[9] Denormalization is the norm for data modeling techniques in this system. Given that data marts generally cover only a subset of the data contained in a data warehouse, they are often easier and faster to implement.

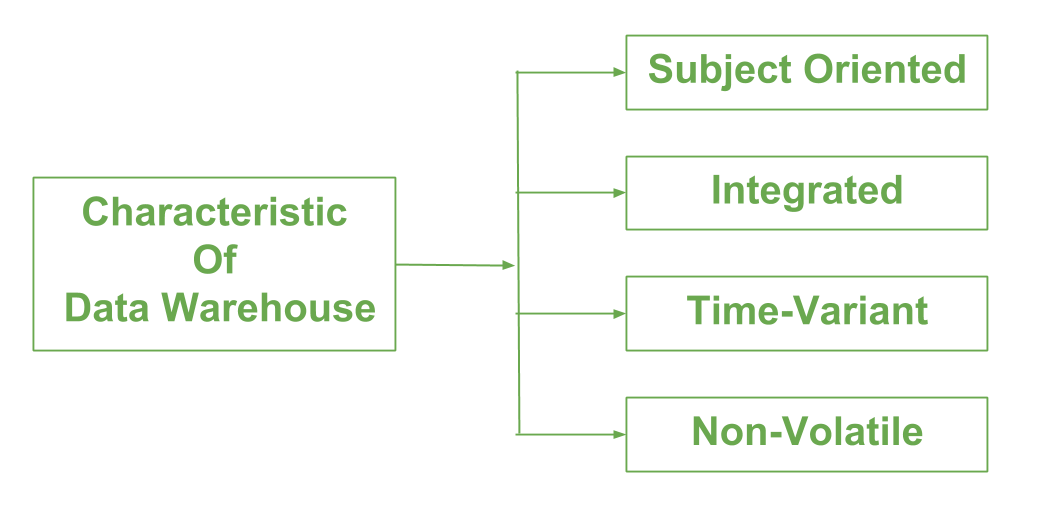
Types of data marts include dependent, independent, and hybrid data marts.[clarification needed]

Online analytical processing (OLAP) is characterized by a relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems, response time is an effective measure. OLAP applications are widely used by Data Mining techniques. OLAP databases store aggregated, historical data in multi-dimensional schemas (usually star schemas). OLAP systems typically have a data latency of a few hours, as opposed to data marts, where latency is expected to be closer to one day. The OLAP approach is used to analyze multidimensional data from multiple sources and perspectives. The three basic operations in OLAP are Roll-up (Consolidation), Drill-down, and Slicing & Dicing.



**Data warehouse characteristics**

There are basic features that define the data in the data warehouse that include subject orientation, data integration, time-variant, nonvolatile data, and data granularity.



**Subject-oriented**

Unlike the operational systems, the data in the data warehouse revolves around the subjects of the enterprise. Subject orientation is not database normalization. Subject orientation can be really useful for decision-making. Gathering the required objects is called subject-oriented.

**Integrated**

The data found within the data warehouse is integrated. Since it comes from several operational systems, all inconsistencies must be removed. Consistencies include naming conventions, measurement of variables, encoding structures, physical attributes of data, and so forth.

**Time-variant**

While operational systems reflect current values as they support day-to-day operations, data warehouse data represents a long time horizon (up to 10 years) which means it stores mostly historical data. It is mainly meant for data mining and forecasting. (E.g. if a user is searching for a buying pattern of a specific customer, the user needs to look at data on the current and past purchases.)

**Nonvolatile**

The data in the data warehouse is read-only, which means it cannot be updated, created, or deleted (unless there is a regulatory or statutory obligation to do so).

**Data warehouse options**

**Aggregation**

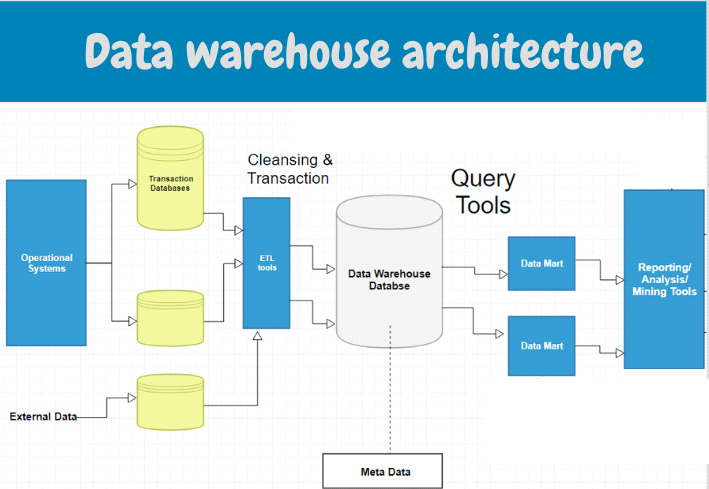
In the data warehouse process, data can be aggregated in data marts at different levels of abstraction. The user may start looking at the total sale units of a product in an entire region. Then the user looks at the states in that region. Finally, they may examine the individual stores in a certain state. Therefore, typically, the analysis starts at a higher level and drills down to lower levels of details.

**Virtualization**

With data virtualization, the data used remains in its original locations and real-time access is established to allow analytics across multiple sources creating a virtual data warehouse. This can aid in resolving some technical difficulties such as compatibility problems when combining data from various platforms, lowering the risk of error caused by faulty data, and guaranteeing that the newest data is used. Furthermore, avoiding the creation of a new database containing personal information can make it easier to comply with privacy regulations. However, with data virtualization, the connection to all necessary data sources must be operational as there is no local copy of the data, which is one of the main drawbacks of the approach.

Data warehouse architecture:-

The different methods used to construct/organize a data warehouse specified by an organization are numerous. The hardware utilized, software created and data resources specifically required for the correct functionality of a data warehouse are the main components of the data warehouse architecture. All data warehouses have multiple phases in which the requirements of the organization are modified and fine-tuned.



Advantages of Data Warehouse (DWH):

Data warehouse allows business users to quickly access critical data from some sources all in one place.

Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.

Data Warehouse helps to integrate many sources of data to reduce stress on the production system.

Data warehouse helps to reduce total turnaround time for analysis and reporting.

Restructuring and Integration make it easier for the user to use for reporting and analysis.

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.

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:**

Not an ideal option for unstructured data.

Creation and Implementation of Data Warehouse is surely time confusing affair.

Data Warehouse can be outdated relatively quickly

Difficult to make changes in data types and ranges, data source schema, indexes, and queries.

The data warehouse may seem easy, but actually, it is too complex for the average users.

Despite best efforts at project management, data warehousing project scope will always increase.

Sometime warehouse users will develop different business rules.

Organisations need to spend lots of their resources for training and Implementation purpose.

**The Future of Data Warehousing**

Change in Regulatory constrains may limit the ability to combine source of disparate data. These disparate sources may include unstructured data which is difficult to store.

As the size of the databases grows, the estimates of what constitutes a very large database continue to grow. It is complex to build and run data warehouse systems which are always increasing in size. The hardware and software resources are available today do not allow to keep a large amount of data online.

Multimedia data cannot be easily manipulated as text data, whereas textual information can be retrieved by the relational software available today. This could be a research subject.