

# DATA WAREHOUSE ARCHITCTURE AND DATA ANALYSIS TECHNIQUES

Mrs. Vaishali V. Khachane  
 Dept of Information Technology  
 VPM's Polytechnic  
 Thane, Mumbai  
 Email: [vaishali0807@yahoo.co.in](mailto:vaishali0807@yahoo.co.in)

**Abstract**

*A data warehouse is an architectural construct of an information system that provides users with current and historical decision support information that is hard to access or present in traditional operational systems.*

*It is also an integrated set of products that enable the extraction and transformation of operational data to be loaded into a database for end-user analysis and reporting. It usually contains historical data derived from transaction data, but it includes data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. This paper presents concept of data warehousing, architecture of data warehouse and techniques of data analysis in data warehousing.*

*Query and reporting, multidimensional, analysis, and data mining run the spectrum of being analyst driven to analyst assisted to data driven. Because of this spectrum, each of the data analysis methods affects data modeling.*

**Index Terms: Data warehousing, operational data store vs. data warehouse, Data warehouse Architecture, Data Analysis techniques**

## I. INTRODUCTION

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

A data warehouse acts as a centralized repository of an organization's data. A data warehouse complements an existing operational system and is therefore designed and

Subsequently used quite differently. A data warehouse provides the base for the powerful data analysis techniques that are available today such as data mining and multidimensional analysis, as well as the more traditional query and reporting. Making use of these techniques along with data warehousing can result in easier access to the information you need for more informed decision making.

Data warehousing provides an excellent approach for transforming the vast amounts of data that exist in these organizations into useful and reliable information for getting answers to their questions and to support the decision making process. It is globally accepted that information is a very powerful asset that can provide significant benefits to any organization and a competitive advantage in the business world. Organizations have vast amounts of data but have found it increasingly difficult to access it and make use of it. This is because it is in many different formats, exists on many different platforms, and resides in many different file and database structures developed by different vendors. Thus organizations have had to write and maintain perhaps hundreds of programs that are used to extract, prepare, and consolidate data for use by many different applications for analysis and reporting. This would typically require modification of the extract programs or development of new ones. This process is costly, inefficient, and very time consuming. Data warehousing offers a better approach.

### Operational Data Store Vs Data Warehouse:

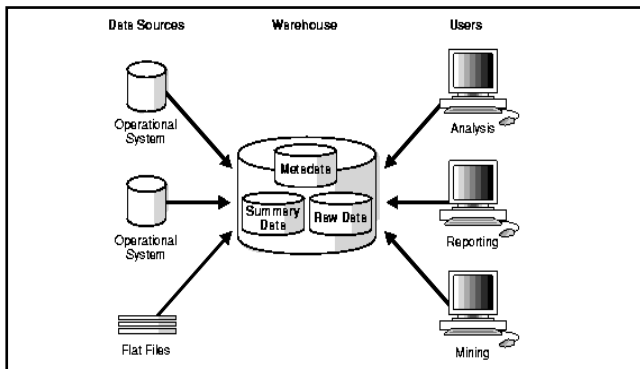
	Operational Data Store	Data Warehouse
Character istics	Process-oriented, Focused on specific business processes and task, Support high volume transaction Processing	Subject-oriented, Integrated, Non-volatile, Time-Variant, Support high volume OLAP
Age of the data	Current, Near Term (Today, Last Week)	Historic (Month, Quarter, Year)
Primary Use:	Day-To-Day Decisions Tactical Reporting Current Operational Results	Day-To-Day Decisions Tactical Reporting Current Operational Results
Frequenc y Of Load	Twice Daily , Daily, Weekly	Weekly, Monthly, Quarterly[2,3]

## II. DATA WAREHOUSE ARCHITECTURES:

Data warehouse Architecture is a design that encapsulates all the facets of data warehousing for an enterprise environment. Data warehousing is the creation of a central domain to store complex, decentralized enterprise data in a logical unit that enables data mining, business intelligence, and overall access to all relevant data within an organization. Data warehouse architecture is inclusive of all reporting requirements, data management, security requirements, bandwidth requirements, and storage requirements. There are three common types of data architecture which are as follows:

### A. DATA WAREHOUSE ARCHITECTURE - BASIC

Figure below shows a simple architecture for a data warehouse. End users directly access data derived from several source systems through the data warehouse.



In this figure the metadata and raw data of traditional OLTP system is present, as is an additional type of data, summary data. Summaries are very valuable in data warehouses because they pre-compute long operations in advance. For example, a typical data warehouse query is to retrieval something like August sales.

#### Operational Source Systems:

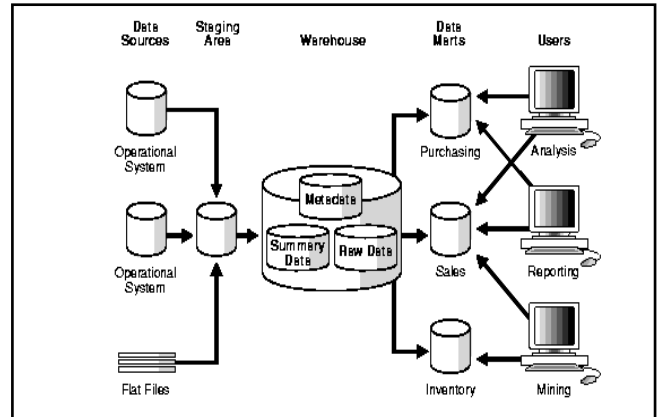
Operational source systems are developed to capture and process original business transactions. These Systems are designed for data entry, not for reporting, but it is from here the data in data warehouse gets populated.

#### Data Warehouse:

A data warehouse contains the data that is organized and stored specifically for direct user queries and reports. It differs from an OLTP database in the sense that it is designed primarily for reads not writes. Data warehouses and their architectures vary depending upon the specifics of an organization's situation.

### B. DATA WAREHOUSE ARCHITECTURE - WITH STAGING AREA

We need to clean and process the operational data before putting it into the warehouse. We can do this programmatically, although most data warehouses use a staging area instead. A staging area simplifies building summaries and general warehouse management.



#### Data Staging Area:

A place where data is processed before entering the warehouse. Data staging area is where the raw operational data is extracted, cleaned, transformed and combined so that it can be reported on and queried by users. This area lies between the operational source systems and the user database and is typically not accessible to users.

### C. DATA WAREHOUSE ARCHITECTURE - WITH A STAGING AREA AND DATA MARTS

We may want to customize your warehouse's architecture for different groups within your organization. You can do this by adding data marts, which are designed for a particular line of business. An example is where purchasing, sales, and inventories are separated. In this example, a financial analyst might to analyze historical data for purchases and sales.

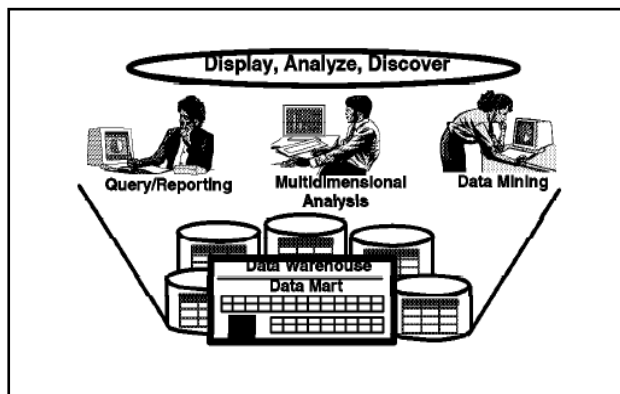
#### Data Mart:

Data mart is a logical subset of an enterprise-wide data warehouse. For example, a data warehouse for a retail chain is constructed incrementally from individual, conformed data marts dealing with separate subject areas such as product sales. Dimensional Data marts are organized by subject area such as sales, finance and marketing and coordinated data category such as

customer, product and location. These flexible information stores allows data structures to respond to business changes-product line additions, new staff, responsibilities, mergers, consolidations, and acquisitions[2].

### III. DATA ANALYSIS TECHNIQUES

A data warehouse is built to provide an easy to access source of high quality data. It is a means to an end, not the end itself. That end is typically the need to perform analysis and decision making through the use of that source of data. There are several techniques for data analysis that are in common use today. They are query and reporting, multidimensional analysis, and data mining . They are used to formulate and display query results, to analyze data content by viewing it from different perspectives, and to discover patterns and clustering attributes in the data that will provide further insight into the data content. The techniques of data analysis can impact the type of data model selected and its content.

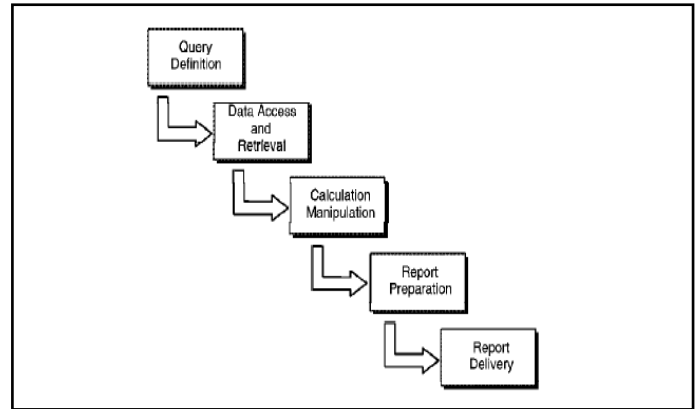


Following are the data analysis techniques:

#### A. Query and Reporting:

Query and reporting analysis is the process of posing a question to be answered, retrieving relevant data from the data warehouse, transforming it into the appropriate context, and displaying it in a readable format.

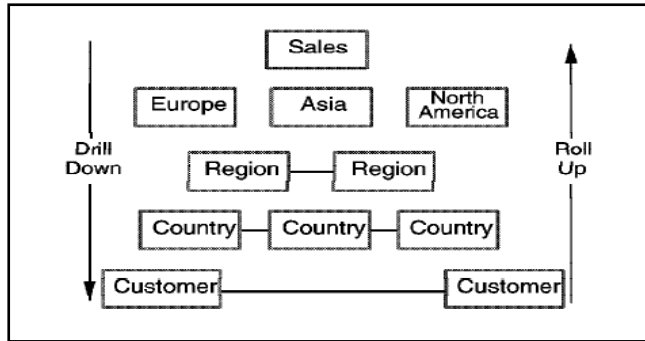
Query and reporting capability primarily consists of selecting associated data elements, perhaps summarizing them and grouping them by some category, and presenting the results. Executing this type of capability typically might lead to the use of more direct table scans. For this type of capability, perhaps an ER model with a normalized and/or denormalized data structure would be most appropriate.



Query definition is the process of taking a business question or hypothesis and translating it into a query format that can be used by a particular decision support tool. When the query is executed, the tool generates the appropriate language commands to access and retrieve the requested data, which is returned in what is typically called an answer set. The data analyst then performs the required calculations and manipulations on the answer set to achieve the desired results. Those results are then formatted to fit into a display or report template that has been selected for ease of understanding by the end user. This template could consist of combinations of text, graphic images, video, and audio. Finally, the report is delivered to the end user on the desired output medium, which could be printed on paper, visualized on a computer display device, or presented audibly.

#### B. Multidimensional Analysis

Multidimensional analysis has become a popular way to extend the capabilities of query and reporting. That is, rather than submitting multiple queries, data is structured to enable fast and easy access to answers to the questions that are typically asked. For example, How much of each of our products was sold on a particular day, by a particular sales person, in a particular store? Each separate part of that query is called a dimension. Many answers can be readily available because the results are not recalculated with each query; they are simply accessed and displayed. The data categorized by these different factors, or dimensions, makes it easier to understand, particularly by business-oriented users of the data. Dimensions can have individual entities or a hierarchy of entities, such as region, store, and department.



Multidimensional analysis enables users to look at a large number of interdependent factors involved in a business problem and to view the data in complex relationships. End users are interested in exploring the data at different levels of detail, which is determined dynamically. The complex relationships can be analyzed through an iterative process that includes drilling down to lower levels of detail or rolling up to higher levels of summarization and aggregation. Figure demonstrates that the user can start by viewing the total sales for the organization and drill down to view the sales by continent, region, country, and finally by customer. Or, the user could start at customer and roll up through the different levels to finally reach total sales. Pivoting in the data can also be used. This is a data analysis operation whereby the user takes a different viewpoint than is typical on the results of the analysis, changing the way the dimensions are arranged in the result. Like query and reporting, multidimensional analysis continues until no more drilling down or rolling up is performed.

### C. DATA MINING

Data mining is a relatively new data analysis technique. It is very different from query and reporting and multidimensional analysis in that it uses what is called a discovery technique.

Data mining is most typically used for statistical data analysis and knowledge discovery. Statistical data analysis detects unusual patterns in data and applies statistical and mathematical modelling techniques to explain the patterns. The models are then used to forecast and predict. Types of statistical data analysis techniques include linear and nonlinear analysis, regression analysis, multivariate analysis, and time series analysis. Knowledge discovery extracts implicit, previously unknown information from the data. This often results in uncovering unknown business facts.

Data mining is data driven. There is a high level of complexity in stored data and data interrelations in the data warehouse that are difficult to discover without data mining. Data mining offers new insights into the business that may not be discovered with query and reporting or multidimensional analysis. Data mining can help

discover new insights about the business by giving us answers to questions we might never have thought to ask[1,2].

### CONCLUSION

Over the next few years, the growth of data warehousing is going to be enormous with new products and technologies coming out frequently. The Data warehouse Architecture is simply a framework for understanding data warehousing and how the components of data warehousing fit together. Data warehouse architecture provides a useful way of determining if the organization is moving toward a reasonable data warehousing framework. The type of analysis that will be done with the data warehouse can determine the type of model and the model's contents.

### REFERENCES

1. *Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell ( Data Modeling techniques for Data Warehousing - redbooks.ibm.com)*
2. *Principle Partners Inc presentation(Data Warehouse concepts and Architecture)*
3. *Oracle Data warehousing guide – Release 2(9.2)*