Introduction to Data Warehousing

Manish Bhardwaj
Student of Information Technology
Ch.Brahm Prakash Government Engineering College, New Delhi
E-mail:manishbhardwaj5885@gmail.com

ABSTRACT: Data Warehouses are provided for storage, functionality and responsiveness to queries beyond the capabilities transaction oriented databases. The use of Data Warehousing is to create front-end analytics that will support business executives and operational managers. Since the decisional process naturally requires an analysis of chronological trends, time and its management attain a huge significance. [1]

This paper describes the need & basic architecture of a data warehouse. The data warehouse is an informational environment that provides an integrated and total view of the enterprise and makes the enterprise`s current and historical information easily available for decision making. A data warehouse is not a single software or hardware product; rather it is a computing environment where users can find the strategic information. Software managers need to understand the rationale and importance of data warehouses because they may need to design and implement.[2][10]

KEYWORDS: Data warehouse, Knowledge discovery and data mining (KDD), Decision Support Systems (DSS) Online Analytical Processing Systems (OLAP).

I. INTRODUCTION

In today's information technology globe, there is an explosive expansion of information in diverse industrial sectors which led to run vast data warehouse for collecting numerous diverse type of pertinent information. This advance us in managing a massive amount of data and achieving success in enhancements of knowledge and understanding of business atmosphere at all level. A data warehouse is gathering of assorted production data, external data, archived data and internal data from unlike data sources. These sources are inculcated in the data warehouse and may amend their schema according to the user necessities. Such changes must be supported when they inhabit the data warehouse. [9]

A data warehouse is a subject-oriented, integrated, non volatile and time-variant collection of data in support of management's decisions. It means:

• Subject-Oriented: Stored data targets precise subjects. Example: It may store data on the subject of total Sales, Number of Customers, etc. and not universal data on daily basis operations.

• Integrated: Data may be dispersed across heterogeneous sources which have to be integrated. Example: Sales data may be on RDB, Customer information on Flat files, etc.

• Time Variant**:** Data stored may not be contemporary but varies with time and data have
an element of time. Example: Data of sales in last 5 years, etc.

• Non-Volatile: It is detach from the Enterprise Operational Database and hence is not subject to recurrent modification. It usually has only two operations performed on it: Loading of data and Access of data.

Features of a Warehouse**:**

• It is separate from Operational Database.
• Integrates data from assorted systems.
• Stores vast quantity of data, more historical than current data.
• Does not need data to be extremely accurate.
• Queries are usually compound.
• Objective is to carry out statistical queries and provide results which can influence decision making in favor of the Enterprise.
• These systems are thus called Online Analytical Processing Systems (OLAP). [3]

This paper has been divided into five sections. Preliminary being Introduction, Section II pertains to history or more specifically evolution of data warehousing. Section III regards to the need of data warehouse. Section IV describes the architecture of data warehouse. Lastly, Section V summarizes and tells about future scope for the same. Section VI is References.

I. History of data warehousing

A lot of computer users may have heard the term data warehouse to imply the central source of data which permits access to stored information effortlessly. They also come to comprehend that the term refers to a relational database and query system designed to assist them analyze data – a form of data mining. Data warehousing has develop into a trendy data management system. In contrast to databases, a data warehouse contains very huge amounts of data stored across a

number of organizational databases. The concept of data warehousing is not a new innovation.

In reality, even though the name appeared for the first time in a 1988 IBM Systems Journal article – "An Architecture for a Business Information System" –, Bill Inmon, the man who is considered the "Father of Data Warehousing," used a alike term way back in the 1970s while working as a data professional and becoming an expert in relational data modeling. His experience in database technology and in developing data warehouses lead to his first company, Prism Solutions in 1991. There he released the Prism Warehouse Manager product, which was one of the first examples of a product for creating and running a data warehouse.

During Inmon's career, he has written books, founded a new company where he came up with a methodology to achieve "data integration," and been a keynote speaker at industry conferences and trade shows. He even held seminars on developing data warehouses. His 1992 book "Building the Data Warehouse," is still exceedingly regarded by IT professionals fascinated in the database world.

Even though the original data warehouse concept was identified by Bill Inmon, the technology advanced as a result of Ralph Kimball's dimensional modeling concept for data warehouse design. His series of "Data Warehouse Toolkit" books, as well as the rising interest and significance of amorphous data and improvements in database technology that adjoin value to business operations, have also affected changes in data warehousing.

The 1990s presented operational business intelligence and shapeless content, as well as ordered data sources, to amplify pace of release and to permit less-structured decision-making processes and echo the requirements of the present day businesses.

These days, data warehousing is sprouting to meet the rising desires of professionals worldwide, but the ground work done by Bill Inmon and Ralph Kimball still influence today's' practices.

Inmon's work in support of centralized data warehouses of hefty size and Kindall's integrated systems of smaller data marts are still influencing today's architectures. As larger business may promote from Inmon's data warehouse approach, smaller businesses might profit from the Kindall's approach which normally requires a lesser budget to put into practice.

The development of today's data warehousing is also determined by users' need for real-time access to information on-the-go for research and decision-making purposes, as well as by the advances in the technology and enlargement of cloud computing. Significance is also given today to the governance and data quality but the key element still remains, regardless if using Inmon's or KIndall's approach, the capability to integrate the data warehouse with the existing business data architecture.[4]

II. NEED FOR A DATA WAREHOUSE

Data Warehousing is an vital component and in most cases the foundation of BI architecture. The data warehouse exists to answer questions users have about the business, the performance of the various operations, the business trends, and about what can be done to improve the business. Let me highlight what you need a data warehouse for:

A. Data Integration

Although you are a small Credit Union, still your enterprise data flows through and lives in a variety of in-house and peripheral systems. You want to ask questions that characterize those slices of key information (referred to as Key Performance Indicators or KPIs) such as - What is the member profitability or member value attrition? Oh, by the way, you want to be able to examine it across all products by location, time and channel. You understand that all the requisite data is possibly there but not integrated and organized in a way for you to get the answers simply.

Perhaps your IT staff has been providing the reports you necessitate each time through a series of manual and automated steps of stripping or extracted the data from one source, sorting / merging with data from other sources, manually scrubbing and enriching the data and then running reports against it. You surprise there ought to be a superior and trustworthy way of doing this. Data Warehouse serves not only as a repository for historical data but also as an exceptional data integration platform. The data in the data warehouse is integrated, subject oriented, time-variant and non-volatile to facilitate you to get a 360° view of your organization.

B. Advanced Reporting & Analysis

The data warehouse is intended specifically to prop up querying, reporting and analysis tasks. The data model is flattened (denormalized) and structured by subject areas to create it easier for users to acquire even intricate summarized information with a relatively easy query and carry out multi-dimensional analysis. This has two influential reimbursement – multilevel trend analysis and end-user empowerment.

Multi-level trend analysis provides the capability to analyze key trends at every level across numerous diverse dimensions, e.g., Organization, Product, Location, Channel and Time, and hierarchies within them. Most reporting, data analysis, and visualization tools seize benefit of the fundamental data model to give influential capabilities such as drill-down, roll-up, drill-across and diverse ways of slicing and dicing data.

The flattened data model makes it much easier for users to understand the data and write queries rather than work with potentially numerous hundreds of tables and write long queries with multifarious table joins and clauses.

C. Knowledge Discovery and DSS

Knowledge discovery and data mining (KDD) is the routine extraction of non-obvious veiled knowledge

from huge volumes of data. For example, Classification models could be used to classify members into low, medium and high lifetime value. As an alternative of coming up with a one-size-fits-all product, the membership can be divided into diverse clusters based on member profile using Clustering models, and products could be tailored for each cluster. Affinity groupings could be worn to recognize enhanced product bundling strategies.

These KDD applications employ diverse statistical and data mining techniques and rely on subject oriented, summarized, cleansed and "de-noised" data which a well designed data warehouse can readily offer.

The data warehouse also enables an Executive Information System (EIS). Executives naturally could not be anticipated to sift through numerous dissimilar reports trying to get a holistic picture of the organization's performance and construct decisions. They require the KPIs delivered to them.

Some of these KPIs may entail cross product or cross departmental analysis, which may be too manually demanding.

D. Performance

Lastly, the performance of transactional systems and query response time construct the case for a data warehouse. The transactional systems are meant to do just that – achieve transactions competently – and hence, are designed to optimize numerous database reads and writes. The data warehouse, on the other hand, is planned to optimize recurrent complex querying and analysis. Some of the improvised queries and interactive analysis, which could be performed in a small number of seconds to minutes on a data warehouse could take a weighty duty on the transactional systems and factually pull their performance down. Holding historical data in transactional systems for longer period of time could also obstruct with their performance. Hence, the chronological data requests to discover its place in the data warehouse. [5]

III. ARCHITECTURE

Diverse data warehousing systems have dissimilar structures. Some may have an ODS (operational data store), while some may have manifold data marts. Some may have a petite number of data sources, while some may have dozens of data sources. In general, all data warehouse systems have the following layers:

- Data Source Layer
- Data Extraction Layer
- Staging Area
- ETL Layer
- Data Storage Layer
- Data Logic Layer
- Data Presentation Layer
- Metadata Layer
- System Operations Layer

A. Data Source Layer:

It represents the unlike data sources that supply data into the data warehouse. The data source can be of any format — plain text file, relational database, other type of database, Excel file, etc., can all act as a data source. Many different types of data can be a data source: Operations — such as sales data, HR data, product data, inventory data, marketing data, and systems data. Web server logs with user browsing data. Internal market research data. All these data sources together form the Data Source Layer.

B. Data Extraction Layer:

Data gets pulled from the data source into the data warehouse system.
Staging Area:

This is where data is prior to being scrubbed and transformed into a data warehouse / data mart.

This is where data gains its logic applied to transform the data from a transactional nature to an analytical nature. This layer is also where data cleansing happens.

C. ETL Layer:

This is principally applicable to association marketing and profitability analysis. The data in data warehouse is already prepared and structured to sustain this kind of analysis.

D. Data Storage Layer:

This is where the transformed and cleansed data occurs. Based on scope and functionality, 3 types of entities can be found here: data warehouse, data mart, and operational data store (ODS). In any given system, you may have just one of the three, two of the three, or all three types.

E. Data Logic Layer:

This is where business rules are stored. Business rules stored here do not affect the underlying data transformation rules, but do affect what the report looks like.
Figure 1 below describes the Data Warehousing Architecture:

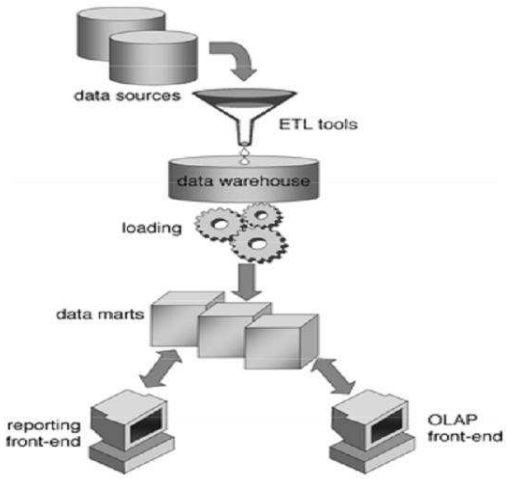

Fig.1. Data Warehousing Architecture [8]

F. Data Presentation Layer:

This refers to the information that reaches the users. This can be in a form of a tabular / graphical report in a browser, an emailed report that gets automatically generated and sent every day, or an alert that warns users of exceptions, among others. Usually an OLAP Tool is used in this layer.

G. Metadata Layer:

This is where information about the data stored in the data warehouse system is stored. A logical data model would be an example of something that's in the metadata layer. A metadata tool is often to used to supervise metadata.
The Figure 2 below shows the three-levels architecture for a data warehousing system:

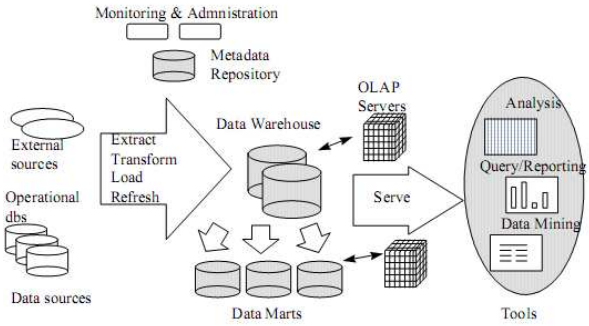

Fig.2: Three-level architecture for a data warehousing system [7]

H. System Operations Layer:

This layer includes information on how the data warehouse system operates, such as ETL job status, system performance, and user access history.[6]

IV. CONCLUSION:

Among the greatest benefits of a data warehouse is the ability to analyze and execute business decisions based on data from multiple sources. In this survey paper, we basically generate the need for a data warehouse and also discussed it`s architecture. We have analyzed that data warehouse is the inseparable part of many domains.

Data warehouse bears three tier architecture: Data sources, data marts and data access techniques. However, data warehouses are still an expensive and typically found in large firms. The development of a central warehouse is a huge undertaking and capital intensive with large, potentially unmanageable risks .

V. FUTURE SCOPE:

As future scope, we would like to discuss the various access tools meant for the data warehouse and also how we can enhance them.

REFERENCES

[1] Rajni Jindal and Shweta Taneja et al." Comparative study of data warehouse design approaches: a survey" International Journal of Database Management Systems (IJDMS) Vol.4, No.1, February 2012.

[2] Nirmal Sharma1, S.K. Gupta et al., "Design and implementation of access the contents in the data warehouse" December 2012, Volume 6, No. 1, pp. 61-64

[3] Jiawei Han , Micheline Kamber et al."Data Mining: Concepts and Techniques".

[4] Rabah Alshboul et al.," Data Warehouse Explorative Study "Applied Mathematical Sciences, Vol. 6, 2012, no. 61, 3015 - 3024

[5]Rajiv Maheshwari, Director Solutions, CS Solutions Inc. tml," The Need for Data Warehousing".

[6] Sohini Roy , Poulomi Ghosh,Student of Calcutta institute of Engineering and Management,"Data Warehousing".

[7] Matteo Golfarelli, Stefano Rizzi, DEIS - University of Bologna, Italy,et al, "Survey Article-A survey on temporal Data Warehousing" International Journal of Data Warehousing & Mining, 5(1), 1-17, January-March 2009 .

[8] Surajit Chaudhuri:Microsoft Research, Umeshwar Dayal : Hewlett-Packard Labs, Palo Alto," An Overview of Data Warehousing and OLAP Technology", (Appears in ACM Sigmod Record, March 1997)

[9] Meenakshi Arora, Anjana Gosain- University School of Information Technology, Guru Gobind Singh Indraprastha University Delhi, India, et al." Schema Evolution for Data Warehouse: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 22– No.6, May 2011

[10] "Data Warehousing Fundamentals" by Paulraj Ponniah