A Critical Review of Data Warehouse

Mohd Faizan Siddiqui, Rachit Saxena, Sarthak Gupta, Tanya

Department of Computer Science and Engineering, Inderpratha Engineering College
Ghaziabad-201010 [Uttar Pradesh] India.

Abstract

In today's modern world Data Warehousing and OLPA have emerged as an important assets for industries, they significantly increased their importance as an aid for the decision makers for any industry. Analyzing the historical data which is collected and stored in a warehouse and producing Analytical results is known as Data warehousing. Professionals uses this technologies as a helping hand, Industries like database industry are constantly using this technology to get best analytical results. In this paper, we cover the critical review on the Data warehousing, data modeling of the data warehouse, and different kind of architectures of the data warehousing. We then analyzed different problems and issues in data warehouse and we also discussed and identified several research areas in the field of Data warehousing.

Keywords: *Data Warehouse, Online Analytical Processing (OLAP).*

Introduction

Data warehouse is a emerging technology, which is used by industries as a repository containing historical data from heterogeneous sources. It is designed in a way that it is particularly used for query and analysis rather than for transaction processing. There are many useful tools and techniques available in the Data warehousing like transformation and loading, an OLAP engine, Extraction and client analysis tools and many other useful applications, which are used for data managing and data processing. It provides best supports to the knowledge workers or decision makers.

Wiliam H.Inmon who was known for his work in Data warehouse, in his theory told us that Data Warehouse is a subject-oriented non-volatile coactions of data which when analyzed can help in decision making in support of management process. And this can differentiate the data warehousing from many other Data repository system like transaction system and file system.

Data Warehousing a modern uprising technology which makes it place among the industries like database industries, the main purpose of trending is that it is able to analyze and helps in critical decision making. There are many events where significant historical data was extracted that can help for future planning. Data warehouse and OLPA are the technology which are meant to provide help to the decision makers unlike other database technologies. Though the technology has various developed phases in last two decades but there are many major areas where research and development is needed.

Foundation of Data Warehousing

The emergence of Data warehousing was between late 1980s and early 1990s. It came as a distinct type of computer database. Since normal operational database was old and not efficient enough to provide best analytical results therefore the arises of Data warehousing concept let to the fulfillment of the demand of higher management to get

best analytical results. Along with the higher user demand and improvement in technologies the concept of Data warehousing has gone through several fundamental stages namely

- Offline operational Database
- Offline Data warehouse
- Real time Data warehouse
- Integrated Data warehouse

Architecture of Data Warehousing:

The architecture of data warehousing depends on various process of business that an organization deals with, taking into account consolidation of Data with security across organization, the level of query requirement management of the Meta, modeling and organization of data, for utilization of optimum bandwidth the conduction of warehouse area planning and full technology implementation.

The warehouse architecture may include:

• Process Architecture

• Data Model architecture

• Technology Architecture

• Information Architecture

• Resource Architecture

Process architecture:

It refers to the process or steps followed in converting row Data into information. It mainly includes three sub process which are commonly referred as "ETL" process.

Extract: The process of extracting Data from different sources, it is accomplished with proper compression and encryption technique.

Transform: The process of Converting extracted Data from different sources into similar format.

Load: loading the transformed data into the data warehouse includes different specific stages.

Data model architecture:

There are five data modeling styles for warehouses as illustrated by Geogia University:

• Data mart bus architecture with conformed dimensions

• Centralized

• Independent Data Mart

• Hub and spoke

• Federated

Technology Architecture

Technology architecture refers to the technological structure of data warehousing.

It includes various aspects like Data base connectivity protocols (JDBC, ODBC, OLE DB etc.), middleware (based on RMI, ORB CCOM/DOM etc.), implementation of data base management standards, network protocols (LDAO, DNS etc.), and other technologies.

Information Architecture

In order to manage the storage, retrieval, modification and deletion of data, the information architecture structure provide the step by step conversion of information from one form to another in the Data warehouse.

Resource Architecture

Resources Architecture refers to the various resourses available for example software resources available for maintaining and managing data warehouse. The performance of Data warehouse system is directly influence by quality of resources architecture.

Typical model of Architecture of Data warehouse

Above mentioned classification gives a summary of the various quiet attributes that we should always confine our minds to create an architecture of a knowledge warehouse. But if we speak about the architecture of knowledge warehouses, it's usually multi-tiered architecture. A typical three-tier architecture is represented within the following image.

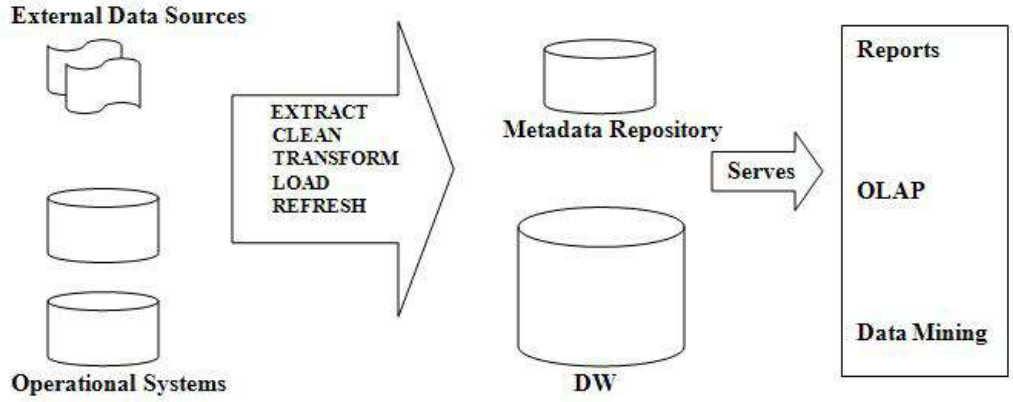

Figure 1: Architecture of Data warehouse

The bottom tier usually accommodates several database systems usually relational databases, face tools, and utilities to extract, clean, transform and feed data to the underside tier from different sources of databases.

The center tier is an OLAP server, it's going to either ROLAP, MOLAP or HOLAP server.

The top tier contains reporting tools, analysis tools, and data processing tools.

Multidimensional Data Model

Dimensional Modeling is a unique approach which uses concept of fact and dimension, for representing Data warehouse rather than entity relationship modeling which is used for normal operational Databases. Basically dimensional modeling technique is used for logical designing of data in a standard, intuitive framework. The high performance access composed of one table with a multipart key, called fact table,

and a set of smaller tables called dimension tables.

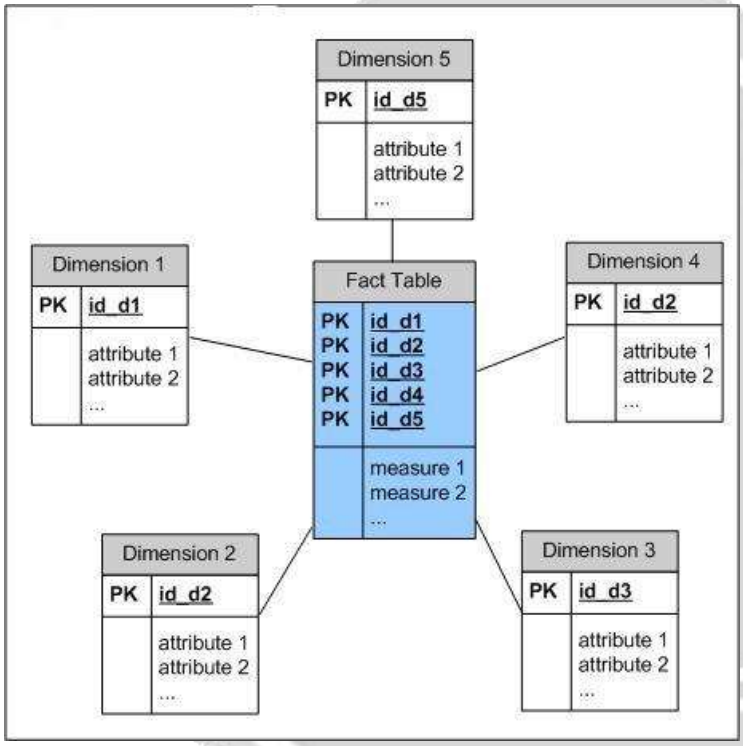

Figure 2: Multidimensional Data model

Fact table has two types of columns, one column contains fact and other column contain foreign key. Facts are numeric measures.

Dimension table is the table which contains the details of perspective or entities with respect to which an organization wants to keep record. It is also known as looked up reference table.

Data cube is the multidimensional view of the data which we get after combining the fact and dimensions. But this cube is n- dimensional not restricted to 3-D like the geometric cube. Using ER diagrams the multidimensional data modeling is advantageous as compared to the conventional relational data modeling technique.

Below figure shows the example of a data cube considering the sales volume as a function of product, month and region.

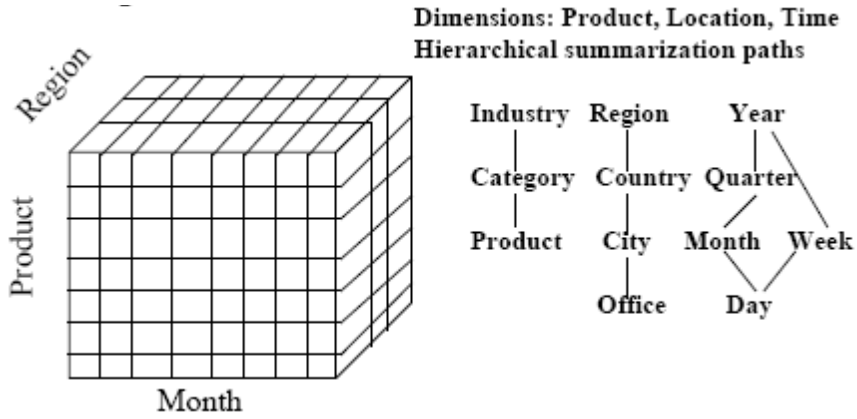

Figure 3: Data cube multidimensional model

Schemas of Multidimensional Model

There are three schemas in which multidimensional model can exit.

Star Schema: according to this schema, the data warehouse contains

(a) Large central table (Fact table) which contains bulk of data with nearly zero redundancy.

(b) Small dimension table one for each dimension. When star schema is represented on graph of schema represents as star in which fact table is surrounded by dimension tables which are arranged radically.

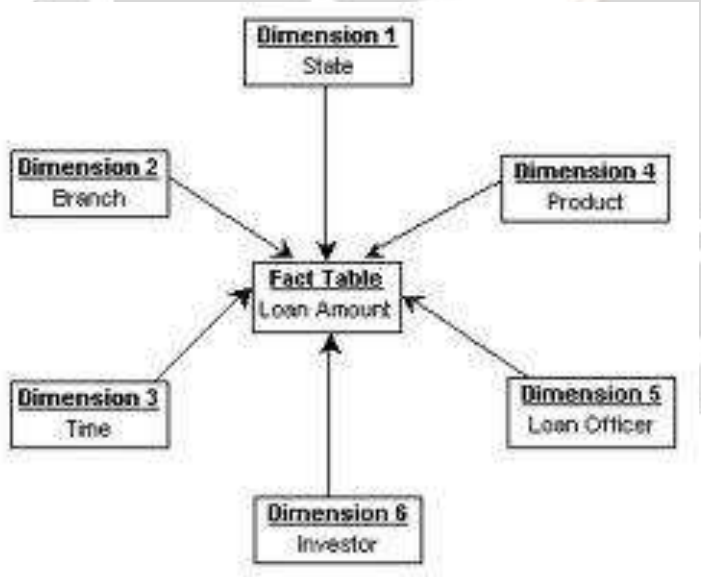

Figure 4: Example of Star Schema

Snowflake schema: Snowflake schema is also like Star schema, But in Snowflake schema we can normalize the dimension table to reduce the redundancy, which separates it from Star schema. This save storage space and is easy to maintain but can reduce the effectiveness of browsing, since more joins will be needed to execute the query.

Fact constellation (Galaxy schema): It has more complex structure which has multiple Fact tables which can share common dimension table.

Data Warehouse Models

Enterprise warehouse: It is large warehouse which contains data about subject spanning the entire organization. Data of all the subjects related to the entire organization is stored in this data warehouse. It is usually a huge data warehouse and requires detailed business modeling.

Data mart: It is the subset of the enterprise data warehouse which contains specific data that is of value to the specific group of users. The information stored in it is about specific subject only.

Virtual warehouse: It is a computer tools basically built fro desiccation-making. It is basically the set of views over operational database.

Tools and Techniques:

Data Warehousing Tools can be divided into the following categories.

Back End Tools and Utilities: These are tools which are generally known as ETL (Extraction, Transform, Load) tool, these tools are used to perform the following operation:

• Data extraction

• Data cleaning

• Data Transformation

• Load

• Refresh

The other tools which are used in the market and are important are Microsoft Integration Services (SSIS), IBM Information Server, IBM Cognos Manager, Open Text Integration Centre, ETL Solutions (ETI), Oracle warehouse Builder(OWB), Telnet Open Studio, Information Builders etc.

Conceptual Model and Front End Tools: They are also known as OLAP tool,

there are mainly three types Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP), Hybrid OLAP (HOLAP).

MOLAP: A cube is aggregated from relational data source. Since data is pre-aggregated within the cube the generation of report is fast.

ROLAP: There is no pre-aggregation of data into the cube just like MOLAP. The ROLAP engine may be consider as small SQL generator.

HOLAP: It is hybrid of both MOLAP and ROLAP. The tools that are available in HOLAP are Business objects, Microsoft, Micro Strategy, Cognos, Analysis service, Palo OLAP server.

Problems and Issues

Despite going through lot of development and improvements over the last decades Data warehouse have many areas to research and improve. Some of the predominant troubles to be tackled are as follows :-

1. Data extraction and cleansing are step one to construct a static warehouse of data. For any form of database the quality of data is the maximum essential element to get the preferred output efficiently. Today we have a wide variety of tools for Data extraction and Cleaning however they are not providing the preferred efficiency. For getting the quality result it is apparent that we need to have the quality data and therefore consequently extraction and cleansing of the data to get the best quality is one in every of eager studies region for data warehouse.

2. Data transformation and integration is another region to be researched further as data which is extracted from heterogeneous source is the building block of data warehouse therefore we should have tools that would be efficient enough to help in data warehouse development than the tools available at present. This is one of the important task in data warehousing as schema and format differs accordingly in databases, and converting them into similar format is essential before loading them into data warehousing. The transformation of information and data with the least blunders and lack of least errors continues to exits and getting rid of them is miles ahead.

3. Maintenance of a data warehouse is any other factor wherein we have a lot of possibilities to improve. For better and efficient managing the increasing size of data warehousing, we need to research for some better maintenance technologies along with the software and better hardware. Management of Mata data is also an important aspect to be researched further.

4. Efficient retrieval of the best result is the principle goal of any system. Though there are several technologies available for efficient query processing, there is still room for improvement. There are lot of thing to research in and query processing is one of them.

CONCLUSION

Data warehousing is the basis of an automated decision support system. Though there were a lot of research and development till now but there are lot of problems faced and tackled in present time and a lot of improvement is needed. The top and most important research issues are the Performance and management at present time. There are some of the latest tools that are identified for Data warehousing and some are classified in logical manner. The logical architecture as well as the typical model of the architecture of the Data warehousing is given. Some of the major research areas like Data cleaning, Data Transformation, maintenance and efficient query processing are also analyzed. We research through major areas in the data warehousing and the improvements that it needed in future to achieve the best out of our data warehousing.

References

- Inmon, W. (2002): Data Warehouse Building, edition number 3rd, New York, Wiley

- Stephen R. (1998) . Research in Building of Data Warehouse.

- SAS© (2002): Using SAS/Warehouse Administrator® the development of Data Warehouse, SAS Institute Inc., Cary, NC 27513, USA.

- Inmon, W.H., "What is a Data warehouse?" Research by 'Prisum solution,Inc'. http://www.cait.wustl.edu/cait/papers/prisum/voll_nol/,1995

- Sen A., and Sinha A. P., (2005): Data warehousing Methodologies direct comparison,

- Communications of the ACM, 41(9), 52-60 (September 1998).