Global Journal of Business Management and Information Technology.
Volume 1, Number 2 (2011), pp. 95-103

http://www.ripublication.com

A Critical Review of Data Warehouse

Sachin Chaudhary[1], Devendra Prasad Murala[2] and V. K. Srivastav[3]

Department of Master of Science,
Asia Pacific Institute of Information Technology, SD, India
Panipat-132103 [Haryana] India.
E-mail: [1]dynamic.chaudhary@gmail.com, [2]murala7@gmail.com,
[3]virendra@apiit.edu.in

Abstract

Data warehousing and OLAP have become the most important aid for the decision makers of any industry. Basically Data warehousing refers to collecting and storing historical data into single repository, which is known as Data warehouse and using that warehouse to produce Analytical results. Being the helping hand for the top level professional, it is continuously under the focus of Database industry and posing new challenges to the database industry day by day. In this paper we present the critical review of the Data warehousing along with different kind of architectures and the data modelling of the data warehouse. We described some of the current tools and techniques available at present for data warehousing in terms of the front end and backend tools. We further analysed problems and issues and identified some of the research areas in the field of data warehousing.

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

Introduction

Data warehouse is a Data repository containing historical data from heterogeneous sources. It is designed for query and analysis rather than for transaction processing. In addition to this Data warehousing concept consists of the tools and techniques available for Extraction, Transformation and loading, an OLAP engine, client analysis tools and other applications that are used to manage and process the data to provide decision support to the knowledge workers or decision makers. (Managers, analyst etc.)

According to William H.Inmon, a well known Data warehouse architect, “A Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection

of data in support of management's decision making process. [6]. This definition separates the Data warehouse from the other Data repository system for example relational database system, Transaction system and File system.

Data Warehouse is a step towards making the computer system able to analyse the trends and help in critical decision making in any organization. Sometimes we get very interesting and useful trend from the historical data that we can use for the future planning. The normal operational databases were meant to provide a help in the clerical operations of the organization but data warehouse and OLAP technologies are meant to provide help to the decisions makers(e.g. Managers, Analyst etc.) of any organization. Therefore new challenges are arising everyday in the field of data warehousing and OLAP to satisfy the demands of the higher professionals.

From last two decades the field of data warehousing has gone through lots of research and changes. From offline operational database to integrated data ware house, it was a long journey, but we still have a long distance to cover. At present we have several areas to improve some of them are identified in this paper. The failure rate of data warehousing projects is still high and if successful the time it is taking is usually more than expected. Therefore we still have to work a lot to achieve a highly efficient data warehousing and OLAP technologies.

In this paper we present a critical review of the data warehousing technology. We described different kind of architectures and the data modelling of the data warehouse. We further analysed the tools and techniques available at present for data warehousing. Some of the major research issues are also identified.

Foundation of Data Warehousing

Data warehousing came into picture as a distinct type of computer database during the late 1980 and early 1990s. The concept of Data warehousing arises to fulfil the demand of the higher management to get analytical results which normal operational database was not providing efficiently. With the improvement in technologies and higher demand from the user 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 warehouse depends on the Business process of any organization taking into the account Data consolidation across the organization with security, the level of query requirement management of the Meta, Data modelling and organization, warehouse staging area planning for optimum bandwidth utilization and full technology implementation.

The warehouse architecture may include:
[18]

- 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 include three sub process which are commonly referred as "ETL" process

Extract: Extracting Data from different sources with proper compression and encryption technique.

***Transform*:** Conversion of the extracted Data from different sources into similar format.

***Load*:** The stages include loading the transformed data into the data warehouse.

***Data model architecture*:**
It is Dimensional Data model, According Georgia University, there are five Data modelling styles for warehouses:

- Independent Data Mart
- Data mart bus architecture with conformed dimensions
- Hub and spoke
- Centralized
- Federated

Technology Architecture
It refers to technological structure of data warehouse. It includes Data base connectivity protocols (ODBC, JDBC, OLE DB etc.), implementation standards in data base management, middleware (based on ORB, RMI, CCOM/DOM etc), network protocols (DNS, LDAO etc), and related technologies.

Information Architecture
It is the structure for step-by-step conversion of the information from one form to another to manage the storage, retrieval, modification and deletion of data in the Data warehouse.

Resource Architecture
It refers to the various resources available for example software resources to maintain and manage data warehouse. The quality of the resource architecture is directly proportional to the performance of the data warehouse system.

Typical model of Architecture of Data warehouse

Above mentioned classification gives an overview of the different kind of attribute that we should keep in our mind to build architecture of a data warehouse. But if we talk about the overall architecture of data warehouse, it is usually multi-tiered architecture. A typical three tier architecture is represented in the following image.

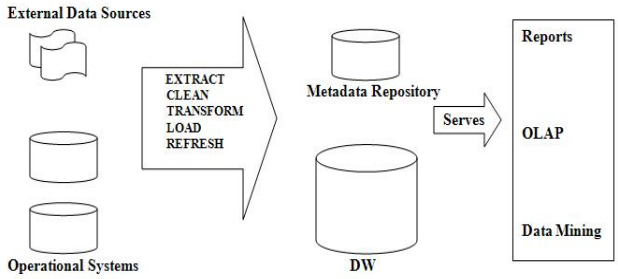

Figure 1: Architecture of Data warehouse

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

The middle tier is an OLAP server, it may either ROLAP, MOLAP or HOLAP server [7.2].

Top tier contains reporting tool, analysis tool, data mining tool.

Multidimensional Data Model

We are very much aware with entity relationship modeling for normal operational Databases but we use different approach known as dimensional modeling for representing the Data warehouse, using the concept of fact and dimension. Basically dimensional modeling is a technique for logical designing of data in a standard, intuitive framework for high performance access composed of one table with a multi-part key, called fact table, and a set of smaller tables called dimension tables.

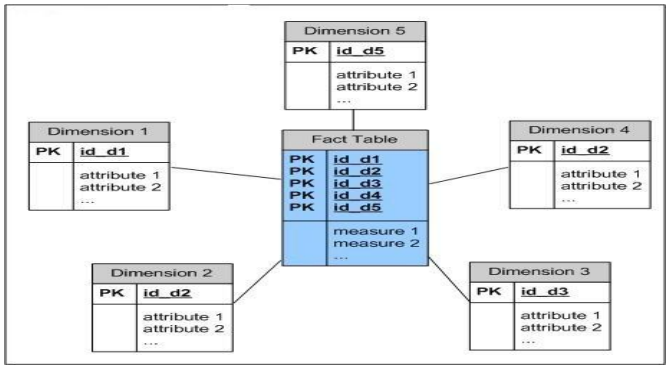

Figure 2: Multidimensional Data model

Fact table has two types of columns one containing fact and other containing foreign key. Facts are numeric measures.

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

Combining the facts and dimensions we get a multidimensional view of the data which is known as data cube. But this cube is n- dimensional not restricted to 3-D like the geometric cube. The multidimensional data modelling has several advantages compare to the conventional relational data modelling technique using ER diagrams. The figure shows the example of a data cube considering the sales volume as a function of product, month and region. [9]

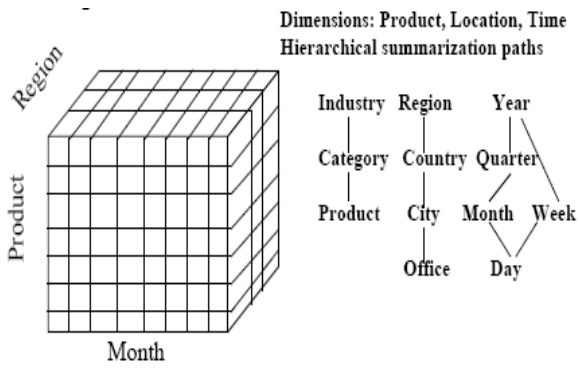

Figure 3: Data cube multidimensional model [9]

Schemas of Multidimensional Model

The multidimensional model can exit in the three schemas

***Star schema*:** According to this schema, the data warehouse contains (a).Large central table (Fact table) containing bulk of data with no redundancy. (b). some called dimension table one for each dimension. When represented on the graph of schema represents as star in which dimension tables are radially arranged around the fact table.[20].

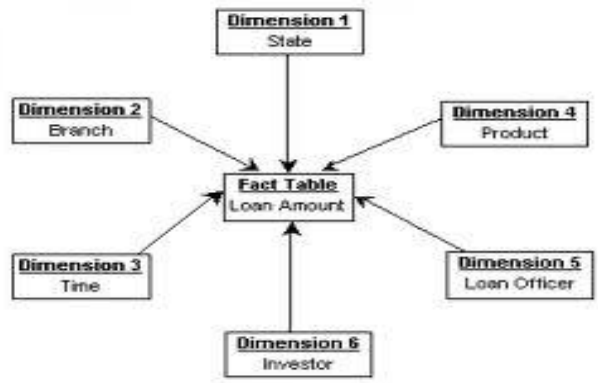

Figure 4: Example of Star Schema

Snowflake schema: It is also like Star schema but the main difference is that in snowflake schema we can normalize the dimension table to reduce the redundancy. This is easy to maintain and save storage space but can reduce the effectiveness of browsing, since more joins will be needed to execute the query.

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

Meta Data

Meta data is Data about data. In terms of Data warehouse it defines the objects of warehouse. Meta data is created to explain the following. [2]

- Description of the Data warehouse structure
- Operational metadata
- The algorithm used for summarization
- Mapping from the operational environment to the data warehouse
- Data related to system performance
- Business metadata.

Data Warehouse Models

Enterprise warehouse: It is a warehouse containing data about subject spanning the entire organization. It is usually a huge data warehouse and requires detailed business modelling. It is a data warehouse containing the data of all the subjects related to the entire organization.[15].

Data mart: It is the subset of the enterprise data warehouse containing the data about specific subject that of value to the specific group of users. They contain information about specific subject only.

Virtual warehouse: It is built over the operational databases as a set of views. 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 tools are also Generally Known has ETL (Extraction, Transform, Load) tool, these tools are used to perform the following operations:

- Data extraction
- Data cleaning
- Data Transformation
- Load
- Refresh

Some of the most important tools used in the market are Oracle warehouse Builder (OWB), Microsoft Integration Services (SSIS), Telnet Open Studio, IBM Information Server, IBM Cognos Manager, Open Text Integration Centre, Information Builders, ETL Solutions (ETI) etc. [19]

Conceptual Model and Front End Tools: Front end tool are also known as OLAP tool, there are mainly three types Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP), Hybrid OLAP (HOLAP). [20].

MOLAP: A cube is aggregated from relational data source. It is faster in generating report as data is pre-aggregated within the cube.

ROLAP: Unlike MOLAP there is no pre-aggregation of Data into the cube. The ROLAP engine may be consider as small SQL generator.

HOLAP: It is the Hybrid of both MOLAP and ROLAP. Some of the Tools available are Business objects, Cognos, Microsoft, Analysis service, micro Strategy, Palo OLAP server.

Problems and Issues

In spite of going through huge amount research during the last decade Data warehouse still have several areas to research and improve. Some of the major issues to be tackled are as follows

1. Data extraction and cleaning are the first step to build a data warehouse. For any kind of database the quality of data is the most important aspect to get the desired output efficiently. Today we have number of tools available for Data extraction and Cleaning but they are not providing the desired efficiency. For

getting the quality result it is obvious that we should have the quality data therefore extraction and cleaning of the data to get the quality data is one of keen research area for data warehouse.

2. Data transformation and integration is another area to be researched further as data warehouse is build up using data from heterogeneous sources therefore we should have efficient tools then available at present. This is one of the most important tasks in data warehousing as different databases have different schemas and format and it's a prerequisite to convert them to similar format before loading into the data warehouse. The transformation of data with least error and least loss of information is still to go miles ahead.
3. Maintenance of a data warehouse is another aspect in which we have lot of chances to improve. We should look for some better maintenance technologies along with the software and better hardware to efficiently manage the increasing size of the data warehouse. Management of Meta data should also be researched further.
4. Efficient retrieval of the result is the main aim of any system. In data warehouse we have several technologies available for efficient query processing but still they have to be improved a lot to achieve the required efficiency. Query processing needs to be researched further.

Conclusion

Data warehousing is the basis of automated decision support system. It has been researched a lot in the past decade but still there are many issues to be tackled in future. Performance and management are among the top research issues at present. We have identified some of the latest tools available for data warehousing and classified the tools in logical manner. The architecture of the data warehouse is also divided logically as well as a typical model of the architecture is also given. We further analysed some of the major research areas like data cleaning, data transformation, maintenance and efficient query processing. We identified major research areas in the data warehousing and the things to be done in future to achieve the best out of our data warehousing.

References

[1] Stolba, N., Banek, M. and Tjoa, A.M. (2006): The Security Issue of Federated Data Warehouses in the Area of Evidence- Based Medicine. Proc. of the First International Conference on Availability, Reliability and Security (ARES'06, IEEE), 20-22 April, 2006.

[2] Inmon, W. (2002): Building the Data Warehouse, 3rd edition, Wiley-New York.

[3] SAS© (2002): Building a Data Warehouse Using SAS/Warehouse Administrator®, Software Course Notes (Book code58787). SAS Institute Inc., Cary, NC 27513, USA.

[4] Sen, A. and Sinha, A. P. (2005): A Comparison of Datawarehousing Methodologies, Communication of the ACM, 48(3), 79-84.
[5] Stephen R. (1998) .Building the Data Warehouse.,
[6] Communications of the ACM, 41(9), 52-60 (September 1998).
[7] Inmon, W.H., "What is a Data warehouse?" Prisum solution, Inc, http://www.cait.wustl.edu/cait/papers/prisum/voll_nol/,1995
[8] Greenfield, L.,"The Case Against Data Warehouseing" LGI Systems, Inc, http://www.dwinfocenter.org/gotchas.html, June 2001
[9] Greenfield, L.," Data Warehouseing Gotchas" LGI Systems, Inc, , http://www.dwinfocenter.org/gotchas.html, June 2001
[10] Harinarayan V., Rajaraman A., Ullman J.D. " Implementing Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996.
[11] Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin, Vol. 18, No.2, June 1995.
[12] O'Neil P., Quass D. "Improved Query Performance with Variant Indices", To appear in Proc. of SIGMOD Conf., 1997.
[13] Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems, Techniques, and Applications." Data Eng. Bulletin, Vol. 18, No. 2, June 1995.
[14] Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate." Available from Arbor Software's web sitehttp://www.arborsoft.com/OLAP.html.
[15] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.
[16] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio, and Y. Zhuge. The Stanford Data Warehousing Project. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Ware housing, 18(2):41{48, June 1995.
[17] W.H. Inmon and C. Kelley. Rdb/VMS: Developing the Data Warehouse. QED Publishing Group, Boston, Massachussetts, 1993.
[18] A. Gupta and I.S. Mumick. Maintenance of materialized views: Problems, techniques, and applications. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing, 18(2):3{18, June 1995}.
[19] Providing Architecture of the Data warehouse.http://it.toolbox.com/wiki/index.php/Data_warehouse_Architecture. Providing ETL Back end tools. [http://etltool.com]
[20] Jiawei Han, Micheline kamber, Jian pei "Data mining Concepts and Techniques" Third edition.