A Review:Data Warehousing, Its Issues, Architecture and Tools

Sandeepak Bhandari
CT Institute Of Technology & Research Greater Kailash G.T Road, Maqsudan Jalandhar

Tarun Sharma
CT Institute Of Technology & Research Greater Kailash G.T Road, Maqsudan Jalandhar

Jagpreet Singh
CT Institute Of Technology & Research Greater Kailash G.T Road, Maqsudan Jalandhar

Sarabjit Kaur
CT Institute Of Technology & Research Greater Kailash G.T Road, Maqsudan Jalandhar

Abstract

Data Warehouse can be defines as a collection of pieces of data that are subject-oriented, time-variant, integrated, and non-volatile. These also support the process of decision-making performed by management.In this paper,we present what exactly the Data warehouse is,its architecture,models,tools and techniques and at its problems and issues which need to be analyzed for a successful data warehouse project.However lot of work has been done in the field regarding design and development of data warehouse,but still lot of areas which need special attention.

Keywords: Datawarehouse, Three.Tier Architecture, Models,Problems and Isues

I. INTRODUCTION

In computing, a data warehouse (DW, DWH), or an enterpris**e** data warehouse (EDW), is a system which can be used for reporting and data analysis. Data warehouse collects data from one or more disparate sources creates a central repository of data, a data warehouse (DW). Data warehouse use current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons and also for decision making.

The data stored in the warehouse is uploaded from the various operational systems (such as marketing, sales, etc).A data warehouse can be defind as 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. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources..In addition to relational database, a data warehouse also contain 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. Some of the characteristics of Data warehouse which separates it from the other repository system namely Transaction system and File System.These characteristics includes Subject Oriented,Integrated,Nonvolatile and Time Variant .The first characteristic Subject Oriented help you to analyze data For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales.The second characteristic integrated is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.The third characteristic include Non volatile,which means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.At last,the fourth characteristics is Time variant,in order to analyze and discover trends in business ,analyst need large amounts of data,It is very much contrast to Online Transaction Processing system,where performance requirements demand that historical data to be moved to an archive.Data Warehouse is a step which making the computer system able to analyse the trends and help intaking critical decision making in any organization. Sometimes we obtain 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.As there are various objectives of data warehousing, some of them includes:-

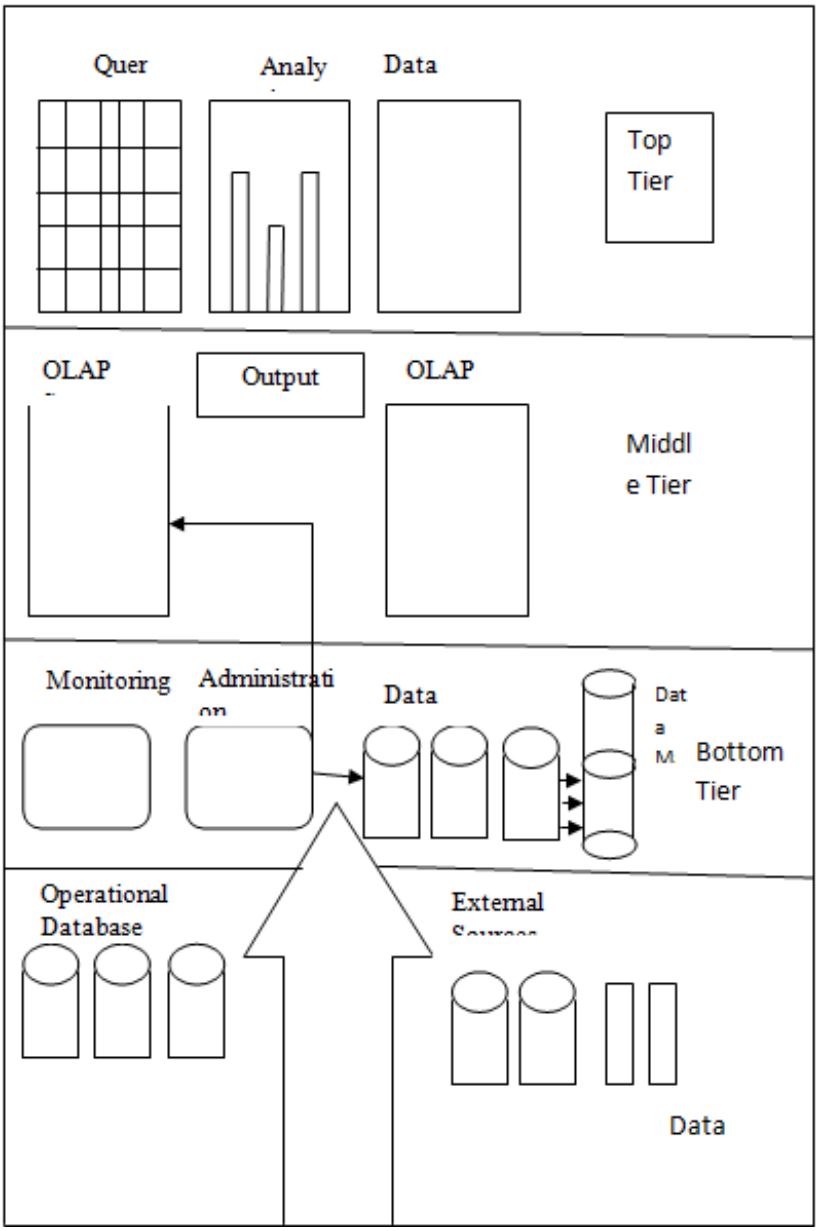
- Efficient distribution of information via the WEB.
- Minimize technicial involvement by enabling users to generate and maintain their
- own reports.
- Create a user friendly reporting environment.
- Provide easy access to data from different sources

– .Lay the foundation and develops plans for full data warehouse development and implementation.

II. ARCHITECTURE OF DATAWAREHOUSING

Generally,Data warehouse adopt three tier architecture.The three tier architecture of datawarehouse includes:-

(1) Bottom Tier
(2) Middle Tier
(3) Top Tier



A. Bottom Tier

The data warehouse database server is the bottom tier of the architecture. It is the relational database system, use the back end tools and utilities to feed data into bottom tier.these back end tools and utilities performs the Extract, Clean, Load, and refresh functions.

B. Middle Tier

In the middle tier we have OLAP Server. the OLAP Server can be implemented in either of the following ways.
By relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
By Multidimensional OLAP (MOLAP) model, which directly implements multidimensional data and operations.

C. Top-Tier

This tier is the front-end client layer. This layer hold the query tools and reporting tool, analysis tools and data mining tools.

III. DATA WAREHOUSE MODELS

From the perspective of data warehouse architecture we have the following data warehouse models:

(1) Virtual Warehouse.
(2) Data Mart.
(3) Enterprise Warehouse

A. VIRTUAL WAREHOUSE

The view over an operational data warehouse is known as virtual warehouse.It is easy to built the virtual warehouse.To build the virtual warehouse we need an excess capacity on operational database servers.

D. DATAMART

Data mart can be defined as subset of organization's wide data.This subset of data is valuable to specific group of an organization.Thus,we can say that data mart contains only that data which is specific to a particular group, For example the marketing data mart may contain only data related to item, customers and sales. The data mart are confined to subjects.

E. ENTERPRISE WAREHOUSE

An enterprise warehouse collects all information of all the subjects spanning the entire organization.It provide us the enterprise wide data integration.The data is integrated from operational systems and external information providers.This information mat be vary from a few Gigabytes to hundreds ofGigabytes,Terabytes or beyond.

1) *BACK END TOOLS AND UTILITIES:*

These tools are also known as Extraction,Transform,Load i.e ETL tool.These tools are used to perform the following operations:-

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

Some of the popular tools used in the market are Oracle warehouse Builder (OWB), Microsoft Integration Services (SSIS), Telnet Open Studio, IBM Cognos Manager, Open Text Integration Centre.

IV. CONCEPTUAL MODEL AND FRONT END TOOLS

Front-end tools are those tools that are available to transform data in a Data Warehouse into actionable business intelligence. The use of appropriate Data Warehousing tools can help ensure that the right information gets to the right person via the right channel at the right time.Front end tool are also known as OLAP tool. There are mainly three types OF OLAP.

(1) Multidimensional OLAP (MOLAP)
(2) Relational OLAP (ROLAP
(3) Hybrid OLAP (HOLAP). [20].

A. MOLAP

MOLAP stands for multidimensional OLAP.Its performance is fast due to optimized storage.multidimensional indexing and caching. Smaller on-disk size of data compared to data stored in relational database due to compression techniques, Automated computation of higher level aggregates of the data. It is very compact for low dimension data sets. Array models provide natural indexing.Effective data extraction achieved through the pre-structuring of aggregated data.

B. ROLAP

ROLAP sands for Relational OLAP.ROLAP is considered to be more suitable in handling large data volumes, especially models with dimensions with very high cardinality (i.e., millions of members).With a variety of data loading tools available, and the ability to fine tune the ETL code to the particular data model, load times are generally much shorter than with the automated MOLAP loads.The data are stored in a standard relational database and can be accessed by any SQL reporting tool (the tool does not have to be an OLAP tool).ROLAP tools are better at handling *non-aggregatable facts* (e.g., textual descriptions). MOLAP tools tend to suffer from slow performance when querying these elements. By decoupling the data storage from the multi-dimensional model, it is possible to successfully model data that would not otherwise fit into a strict dimensional model.

C. HOLAP

HOLAP stands for Hybrid OLAP.It is combination of MOLAP and ROLAP. HOLAP allows storing part of the data in a MOLAP store and another part of the data in a ROLAP store, allowing a tradeoff of the advantages of each.There are two modes of HOLAP.

D. Vertical partitioning

In this mode HOLAP stores aggregations in MOLAP for fast query performance, and detailed data in ROLAP to optimize time of cube processing.

E. Horizontal partitioning

In this mode HOLAP stores some slice of data, usually the more recent one (i.e. sliced by Time dimension) in MOLAP for fast query performance, and older data in ROLAP. Moreover, we can store some dices in MOLAP and others in ROLAP, leveraging the fact that in a large cuboid, there will be dense and sparse subregions.

V. PROBLEMS AND ISSUES

There are various issues surrounding data warehouses that companies need to be prepared for. A failure to prepare for these issues is one of the key reasons that's why many data warehouse projects are unsuccessful. One of the first and most important issues is companies need to confront is that they are going to spend a great deal of time loading and cleaning data.It takes about 80% of the total time of a datawarehouse project according to experts. While the percentage may or may not be as high as 80%, one thing that you must realize is most vendors will understate the amount of time you will have to spend doing it. While cleaning the data can be complicated, extracting it can be even more challenging.Second issue that companies will have to face in datawarehouse project is having problems with their systems placing information in the data warehouse.

When a company enters into this stage firstly, lots of problems appear suddenly.When these problems are seen by business manager,he/she will have to make the decision to fix the problems via the transaction processing system or a data warehouse which is read only.Company is also responsible for storing data that has not be collected and store by the existing system.This can be a headache for developers who run into the problem, and the only way to solve it is by storing data into the system. Many companies will also find that some of their data is not being validated via the transaction processing programs.
In this situation ,the data will need to be validated.when a data is stored in a warehouse,there will be various inconsistencies.One of the most common issues is when controls ere not placed under the names of customers.It will cause headaches for the warehouse user that will want the data warehouse to carry out an ad hoc query for selecting the name of a specific customer. The developer of the data warehouse may find themselves having to alter the transaction processing systems. In addition to this, they may also be required to purchase certain forms of technology.

One of the most critical problems that a company will face is that a transaction processing system that feeds data into the data warehouse with little detail.It may occur frequently in a data warehouse that is tailored towards products or customers. Some developers may refer to this as being a granular issue. Regardless, it is a problem you will want to avoid at all costs. It is important to make sure that the information that is placed in the data warehouse is rich in detail.

VI. CONCLUSIONS

Now-a-days,Data warehousing is the leading and most reliable technology used today by companies for planning, forecasting, and management for e.g. resource planning, financial forecasting and control etc. After the evolution of the concept of data warehousing during the early 90's it was thought that this technology will grow at a very rapid pace but unfortunately it'snot the reality. A lot has been done in this field regarding design anddevelopment of data warehouses and a lot still needs to be done.There are many areas which need more attention and lots of work can be done in these areas, but one area which needs special attention from research community is data warehouse maintenance.A major reason for data warehouse project failures is poor maintenance. Without proper maintenance desired results are nearly impossible to attain from a data warehouse.

REFERENCES

[1] Sachin Chaudhary, Devendra Prasad Mural2 and V. K. Srivastav, A Critical Review of Data Warehouse,published in 2011 Global Journal of Business Management and Information Technology. Volume 1, Number 2 (2011), pp. 95-103.
[2] 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.
[3] Greenfield, L.,"The Case Against Data Warehouseing" LGI Systems, Inc,http://www.dwinfocenter.org/gotchas.html, June 2001.
[4] Greenfield, L.," Data Warehouseing Gotchas" LGI Systems, Inc, ,http://www.dwinfocenter.org/gotchas.html, June 2001.
[5] Harinarayan V., Rajaraman A., Ullman J.D. " Implementing Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996.
[6] Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin, Vol. 18, No.2, June 1995.
[7] L. Chen, K. Soliman, E. Mao, and M. Frolik, Measuring user satisfaction with data warehouses: an exploratory study. Information & Management Apr 2000 Vol. 37 No. 3 pp103 – 110.

[8] B.cooper, H. Watson, B. Wixom and D. Goodhue, Data Warehousing Supports Corporate Strategy at First American Corporation (FAC). *MIS Quarterly* Vol 24, No. 4, Dec 2000, pp 547 – 567.
[9] Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems,Techniques, and Applications." Data Eng. Bulletin, Vol. 18, No. 2, June 1995.
[10] Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytica Processing) to User Analyst: An IT Mandate." Available from Arbor Software's websitehttp://www.arborsoft.com.
[11] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.
[12] 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.
[13] W.H. Inmon and C. Kelley. Rdb/VMS: Developing the Data Warehouse. QED Publishing Group, Boston, Massachussetts, 1993.
[14] 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}.