International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

A Study of Data Warehousing

Paramjeet Kaur
Assistant Professor, Computer Science Guru Nanak College Ferozepur, Punjab, India

ABSTRACT

A data warehouse is a global repository that stores pre-processed queries on data which resides in multiple, possibly heterogeneous, operational or legacy sources. The information stored in the data warehouse can be easily and efficiently accessed for making effective decisions. It is collection of decision support technologies. It is a subject oriented, integrated, time verifying, non volatile collection of data that is used primarily in organizational decision making. Data warehouse can be built using a top-down approach, bottom – down approach or a combination of both. In this research paper we are discussing about the data warehouse and design process.

Keywords: Analysis, Data Warehousing, Data Warehouse Design, Process.

I. INTRODUCTION

A data warehouse is an integrated repository that stores information which may originate from multiple, possibly heterogeneous operational or legacy data sources. The contents of a data warehouse may be a replica of a part of data from a source or they may be the results of preprocessed queries or both. This way of storing data provides a powerful tool to business organizations for making business decisions. The architecture of a data warehousing system allows a number of ways to integrate and query information stored in it. This approach of integrating data from distributed data sources pays rich dividends when it translates into calculated decisions backed by sound analysis. Thus, the Data warehousing coupled with On-Line Analytical Processing (OLAP) enable business decision makers to creatively approach, analyze and understand business problems. The data warehouse system is used to provide solutions for business problems since it transforms operational data into strategic decision making information. The data warehouse stores summarized information over operational data. This summarized information is time-variant. As defined, any data warehouse (DW) should have the following characteristics:

- Subject - oriented: DW can be used to analyze any subject.
- integrated: DW integrates current and historical data from different sources.
- Time - variant: DW keeps historical data of different time.
- Non-volatile collection of data: content of DW should not be changed. It is historical data.

Unlike the modeling techniques used to design regular databases - Entity Relationship model, data warehousing is designed by using dimensional modeling techniques . Data warehousing modeling is complex. It needs: 1) knowledge of the business processes , 2) Understanding the structural and behavioral system's conceptual model, and 3) being familiar with data warehousing techniques .

The resulting data warehouse becomes the main source of information for report generation, analysis, and presentation through ad hoc reports, portals, and dashboards.

Data Warehousing System

The data warehousing system architecture is illustrated in Figure 1. Warehouse data is derived from the data contained in operational data systems. These operational data sources are connected to the wrappers/monitors, whose main function is to select, transform and clean data. It also monitors changes to the source data and propagates them to the integrator. The integrator's job is to combine data selected from different data sources. It resolves conflicts among data and brings data in consistent form. After integration, data is propagated into warehouse storage. Many commercial analysis and development tools are available for data selecting, altering, transforming, and loading.

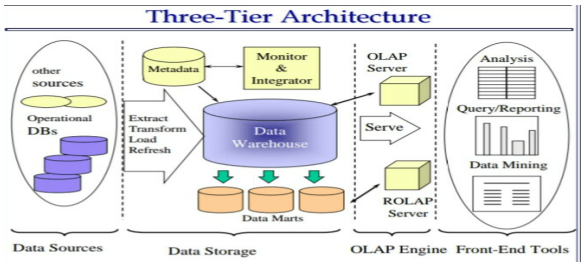

Figure 1. Architecture of Data Warehouse

There are two approaches to creating the warehouse data, namely, bottom up approach and top-down approach, respectively. In a bottom-up approach, the data is obtained from the primary sources based on the data warehouse applications and queries which are typically known in advance, and then data is selected, transformed, and integrated by data acquisition tools. In a top-down approach, the data is obtained from the primary sources whenever a query is posed. In this case, the warehouse system determines the primary data sources in order to answer the query. The bottom-up approach is used in data warehousing because user queries can be answered immediately and data analysis can be done efficiently since data will always be available in the warehouse. Hence, this approach is feasible and improves the performance of the system. Another approach is a hybrid approach, which combines aspects of the bottom-up and top-down approaches. In this approach, some data is stored in a warehouse, and other data can be obtained from the primary sources on demand. The metadata contains the informational data about the creation, management, and usage of the data warehouse. It serves as a bridge between the users of the warehouse and the data contained in it. The warehouse data is accessed by OLAP server to present the same in a multidimensional way to the front end tools (such as analytical tools, report writer, spreadsheets, data mining tools) for analysis and informational purposes. Basically, OLAP server interprets client queries (the client interact with front end tools and pass these queries to the OLAP server) and converts them into complex SQL queries (indeed extended SQL) required to access the warehouse data. It might also access the data from the primary sources if the client's queries need operational data. Finally, the OLAP server passes the multidimensional views of data to the front end tools, and these tools format the data according to the client's requirements.

II. DATA WAREHOUSE DESIGN PROCESS

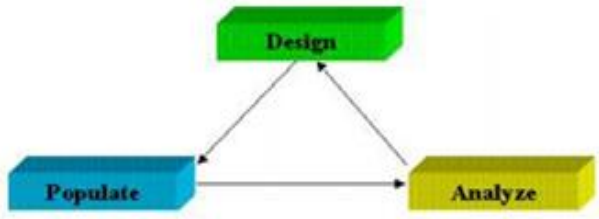

Figure 2

Here we discussed about various approaches to the data warehouse design process and the steps involved. A data warehouse can be built using a top-down approach, a bottom-up approach or a combination of both. The top – down approach starts with overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. The bottom -up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. And it also allowed an organization to move forward at considerable less expenses and

evaluate the technological advantages before making significant commitments. In the combined approach, an organization can be exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom – up approach. If we are thinking in from the software engineering point of view, the design and construction of a data analysis, warehouse design , data integration and testing, and finally deployment of the data warehouse. Large software systems can be developed by using one of the two technologies. The Waterfall method and The spiral method. So, here it is.

The Water Fall method performs a structured and systematic analysis at each step before proceeding to the next, which is like a water fall, falling form one step to the next. The Spiral Method involves the rapid generation of increasingly functional systems, with short intervals between successive releases.

This is always considered as a good choice for data warehouse development, especially for data marts, because the turnaround time is short, modifications can be done quickly, and new designs for the technologies and that can be adapted in a timely manner. So, here we are discussed about the warehouse design process. This includes various steps as follows:

Choose a Business Process to Model:-if the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.

Choose the business process gain, which is the fundamental, atomic level of data to be represented in the fact table for this process.

Choose the dimension that will apply to each and every fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transactions type, and status.

Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

Because the process of construction of data warehouse is a quite difficult and long-term task, its implementation scope should be clearly defined. The goals of an fundamental data warehouse implementation should be specific, achievable and measurable. This involves determining the time and budget allocations, the subset of the organization that is to be served. So, once a data warehouse is designed and constructed, the fundamental deployment of the warehouse includes the initial installations, roll – out planning, training, and orientations. And platform upgrades and maintenance must also be considered. So, the data warehouse administration includes data refreshment, data source synchronization, planning for disaster recovery, managing access control and security, managing data growth, managing data base performances and of course data warehouse enhancement and extension. Data warehouse development tools provide functions to define and edit metadata repository contents (i.e. schemas, scripts, or rules), answer queries, output reports, and ship metadata to and from relational database system catalogs. Planning and analysis tools study the impact of schema changes and of refresh performance when changing refresh rates or time windows.

III. DATA WAREHOUSE USAGE FOR INFORMATION PROCESSING

The proposed Meta model of data warehouse operational processes is capable of modeling complex activities, their interrelationships, and the relationship of activities with data sources and execution details. Moreover, the Meta model

complements the existing architecture and quality models in a coherent fashion, resulting in a full framework for quality oriented data warehouse management, capable of supporting the design, administration and especially evolution of a data warehouse. Data warehouse and data marts are used in a wide range of applications. Business executive use the data warehouses in data warehouses and data marts to perform data analysis and makes strategic decisions. In many firms, data warehouses are used as an integral part of a plan-execute-access" Closed-loop" feedback system for enterprise management. Data warehouses are used extensively in banking and financial services, consumer goods and retail distribution sectors, and controlled manufacturing such as demand-based production. Now, typically the longer a data warehouse has been in such a use, the more it will have evolved. This evolution should take place throughout a number of phases. Initially, the data warehouse is mainly used for generating reports and answering the predefined queries. Progressively, it is used to analyze, summarized and detailed data, where the results are presented in the form of reports and charts, later, the data warehouse is used for strategic purposes, performing multidimensional analysis and sophisticated slice-and-dice operations. So, at that stage we finally we reach the data warehouse may be employed for knowledge discovery and strategic decision making using data mining tools. In this context, the tools for data warehousing can be categorized into access and retrieval tools, database reporting tools, data analysis tools, and data mining tools. There are total three kinds of data warehousing applications: Information processing, Analytical processing, and data mining.

Information Processing supports querying, basic statistical querying, basic statistical analysis, and reporting using cross tabs, tables, charts or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.

Analytical Processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized and detailed forms. The major strength of online analytical processing over information processing is the multidimensional data analysis of data warehouse data.

Data Mining supports knowledge discovery by finding hidden pattern and association constructing analytical models, performing classification and prediction, and presenting the mining results using visualizations tools

So these are the three various data warehouse applications which will help to design and use of data warehouse

IV. FROM ONLINE ANALYTICAL PROCESSING TO MULITDIMENSIONAL DATA MINING

Multidimensional data mining integrates OLAP with data mining to uncover knowledge in multidimensional databases. Among the many different paradigms and architectures of data mining systems, multidimensional data mining is particularly important for the various reasons which are as follows:

High Quality of data in data warehouse: Most data mining tools need to work on integrated, consistent, and cleansed data, which requires costly data cleaning, data integration and data transformation as preprocessing steps. A data warehouse constructed by such preprocessing steps. While a data warehousing constructed by such preprocessing serves as a valuable source of high-quality data for OLAP as well as for data mining. Now, we notice that data mining may serves as a valuable tool for data cleaning and data integration as well

Available information processing infrastructure surrounding data warehouses: Comprehensive

information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which includes the accessing, integration, consolidation and transformation of multiple heterogeneous databases and OLAP analytical tools. It is prudent to make best use of the available infrastructure rather than constructing everything from scratch.

OLAP-Based exploration of multidimensional data: Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge in different forms. Multidimensional data mining provides facilities of pivoting filtering, dicing, and slicing on a data cube and intermediate data mining results.

Online Selection of data mining functions: Users may not always know the specific kinds of knowledge they want to mine. By integrating OLAP with various data mining functions, multidimensional data mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

So, these are the various multidimensional data mining resources for the data warehouse usage and designing. The data warehouse is concentrated on only few aspects. Here we are discussing about the data warehouse design and usage. Let's look at various approaches to the data ware house design and usage process and the steps involved. So, at the end of the research we are clearly said that data warehouse can be built using a top-down approach, bottom – down approach or a combination of both.

V. CONCLUSION

This paper brings together under one roof all the techniques of integrating data quickly to help effective decision making. The data warehousing phenomenon, which has grown from strength to strength in the last few years presents new challenges everyday as we find new applications where warehousing can play a crucial role.

VI. REFERENCES

[1]. Mayank Sharma, Navin Rajpal and B.V.R.Reddy (2010)," Physical Data Warehouse Design using Neural Network" "International Journal of Computer Applications 1(3):86–94, February 2010"Published By Foundation of Computer Science

[2]. "Sarkar, A., Choudhury, S., Chaki, N"& Bhattacharya, S, (2009) "Conceptual Level Design of Object Oriented Data Warehouse: Graph Semantic Based Model", INFOCOMP Journal of Computer Science, pp"60-70.

[3]. "Karen C"Davis Sandipto Banerjee (2007)," Teaching and Assessing a Data Warehouse Design Course", 24th British National Conference on Databases (BNCOD'07) 0-7695-2912-7/07 $20.00 © 2007 IEEE

[4]. "Kim W""On Optimizing a SQL-like Nested Query" ACM TODS, Sep 1982.

[5]. "Gang, K"and P"Yi"A Standard Process for Data Mining Based Software Debugging"in Networked Computing and Advanced Information Management, 2008"NCM '08"Fourth International Conference on"2008.

[6]. "Huifang, Z"and P"Ding"A knowledge discovery and data mining process model in E-marketing"in Intelligent Control and Automation (WCICA), 2010 8th World Congress on"2010

[7]. "Bora, S"Data mining and ware housing"in Electronics Computer Technology (ICECT), 2011 3rd International Conference on"2011"IEEE.

[8]. www.wikipedia.com

[9]. "ADVANCED DATABASE MANAGEMENT SYSTEM Rini Chakrabarti and Shilbhadra Dasgupta

[10]. "Database Management Systems: Understanding and Applying Database" Michael M"Gorman"