ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 4, April 2015

Study of Data Warehousing and Online Analytical Processing

Chandrakant Dewangan[1], Ankit Naik[2], Purushottam Patel[3]
Student,CSE, Kirodimal Institute of Technology, Raigarh,India[1]
Lecturer CSE, Kirodimal Institute of Technology, Raigarh,India[2]
HOD CSE, Kirodimal Institute of Technology, Raigarh,India[3]

Abstract: Data-driven decision support systems, such as data warehouses can serve the requirement of extraction of information from more than one subject area. Data warehouses standardize the data across the organization so as to have a single view of information. Data warehouses can provide the information required by the decision makers. Developing a data warehouse for educational institute is the less focused area since educational institutes are non-profit and service oriented organizations. In present day scenario where education has been privatized and cut throat competition is prevailing, institutes needs to be more organized and need to take better decisions. Institute's enrollments are increasing as a result of increase in the number of branches and intake. Now a day, any reputed Institute's enrollments count in to thousands. In view of these factors the challenges for the management are meeting the diverse needs of students and facing increased complexity in academic processes. The complexity of these challenges requires continual improvements in operational strategies based on accurate, timely and consistent information. The study may help decision makers of educational institutes across the globe for better decisions..

Keywords: Data warehouse architecture,Types of OLAP Servers,OLAP Operations, OLAP vs OLTP

I. INTRODUCTION

According to W. H. Inmon: "A data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of data in support of management's decision-making process." [1]

Subject Oriented - The Data warehouse is subject oriented because it provide us the information around a subject rather the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue etc. The data warehouse does not focus on the ongoing operations rather it focuses on modelling and analysis of data for decision making.

Integrated - Data Warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhance the effective analysis of data.

Time-Variant - The Data in Data Warehouse is identified with a particular time period. The data in data warehouse provide information from historical point of view.

Non Volatile - Non volatile means that the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database are not reflected in data warehouse.

Metadata - Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example the index of a book serves as metadata for the contents in the book.In other words we can say that metadata is the summarized data that lead us to the detailed data.

II. DATA WAREHOUSE ARCHITECTURE

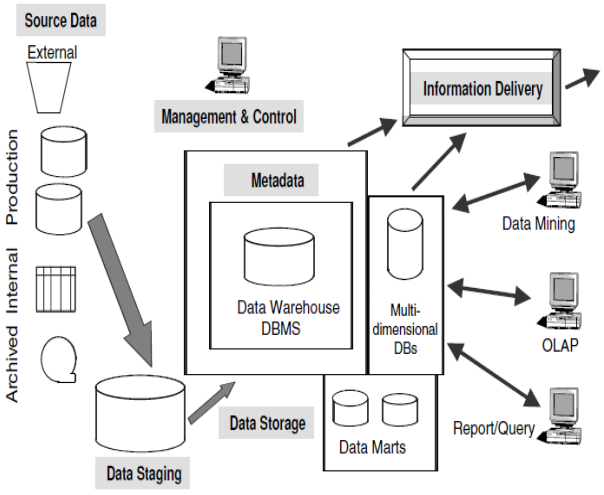

Fig 1 Data Warehouse Architecture

Architecture is the proper arrangement of the components. You build a data warehouse with software and hardware components. To suit the requirements of your organization you arrange these building blocks in a certain

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 4, April 2015

way for maximum benefit. We also want to review specific issues relating to each particular component.

A. Source Data Component

Source data coming into the data warehouse may be grouped into four broad categories, as discussed here.

(a) Production Data: This category of data comes from the various operational systems of the enterprise. Based on the information requirements in the data warehouse, you choose segments of data from the different operational systems. While dealing with this data, you come across many variations in the data formats. You also notice that the data resides on different hardware platforms. Further, the data is supported by different database systems and operating systems. This is data from many vertical applications and integrate the pieces into useful data for storage in the data warehouse.

(b) Internal Data: In every organization, users keep their "private" spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse. If your organization does business with the customers on a one-to-one basis and the contribution of each customer to the bottom line is significant, then detailed customer profiles with ample demographics are important in a data warehouse Although much of this data may be extracted from production systems, a lot of it is held by individuals and departments in their private files. You cannot ignore the internal data held in private files in your organization. It is a collective judgment call on how much of the internal data should be included in the data warehouse.

(c) Archived Data: Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. The circumstances in your organization dictate how often and which portions of the operational databases are archived for storage. Some data is archived after a year. Sometimes data is left in the operational system databases for as long as five years. Many different methods of archiving exist. There are staged archival methods. At the first stage, recent data is archived to a separate archival database that may still be online.

(d) External Data: Most executives depend on data from external sources for a high percentage of the information they use. They use statistics relating to their industry produced by external agencies. They use market share data of competitors. They use standard values of financial indicators for their business to check on their performance. For example, the data warehouse of a car rental company contains data on the current production schedules of the leading automobile manufacturers. This external data in the data warehouse helps the car rental company plan for their fleet management.

B. Data Staging Component

After you have extracted data from various operational systems and from external sources, you have to prepare the data for storing in the data warehouse.These three major functions of extraction, transformation, and preparation for loading take place in a staging area. The data staging component consists of a workbench for these functions. Data staging provides a place and an area with a set of functions to clean, change, combine, convert, reduplicate, and prepare source data for storage and use in the data warehouse.

(a)Data Extraction: This function has to deal with numerous data sources. You have to employ the appropriate technique for each data source. Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Some data may be on other legacy network and hierarchical data models

(b)Data Transformation: In every system implementation, data conversion is an important function. For example, when you implement an operational system such as a magazine subscription application, you have to initially populate your database with data from the prior system records. You may be converting over from a manual system. Or, you may be moving from a file-oriented system to a modern system supported with relational database tables. (c)Data Loading. Two distinct groups of tasks form the data loading function. When you complete the design and construction of the data warehouse and go live for the first time, you do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time. As the data warehouse starts functioning, you continue to extract the changes to the source data, transform the data revisions, and feed the incremental data revisions on an ongoing basis. Figure below in data storage illustrates the common types of data movements from the staging area to the data warehouse storage.

C. Data Storage Component

The data storage for the data warehouse is a separate repository. The operational systems of your enterprise support the day-to-day operations. These are online transaction processing applications. The data repositories for the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 4, April 2015

operational systems typically contain only the current data.

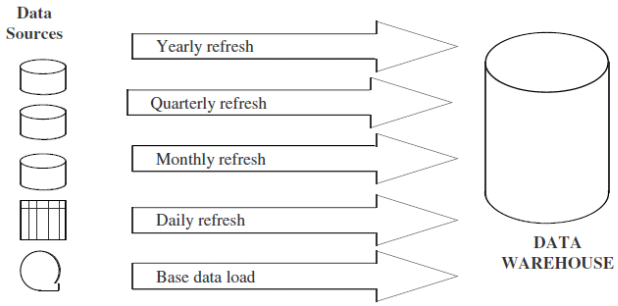

Fig 2 Data Storage Component

data, the data storage must not be in a state of continual updating. For this reason, the data warehouses are "read-only" data repositories. Generally, the database in your data warehouse must be open. Depending on your requirements, you are likely to use tools from multiple vendors. The data warehouse must be open to different tools. Most of the data warehouses employ relational database management systems. Many of the data warehouses also employ multidimensional database management systems. Data extracted from the data warehouse storage is aggregated in many ways and the summary data is kept in the multidimensional databases (MDDBs). Such multidimensional database systems are usually proprietary products.

D. Information Delivery Component

Who are the users that need information from the data warehouse? The range is fairly comprehensive. The novice user comes to the data warehouse with no training and, therefore, needs prefabricated reports and preset queries. The casual user needs information once in a while, not regularly. This type of user also needs prepackaged information.

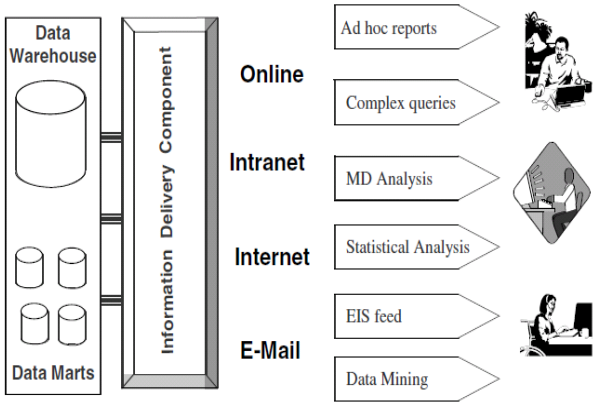

Fig 2 Information Delivery Component

The users will enter their requests online and will receive the results online. You may set up delivery of scheduled reports through e-mail or you may make adequate use of your organization's intranet for information delivery. Recently, information delivery over the Internet has been gaining ground.

E. Metadata Component

Metadata in a data warehouse is similar to the data dictionary or the data catalog in a database management system. In the data dictionary, you keep the information about the logical data structures, the information about the files and addresses, the information about the indexes, and so on. The data dictionary contains data about the data in the database. Similarly, the metadata component is the data about the data in the data warehouse. This definition is a commonly used definition. We need to elaborate on this definition. Metadata in a data warehouse is similar to a data dictionary, but much more than a data dictionary.

F. Management and Control Component

This component of the data warehouse architecture sits on top of all the other components. The management and control component coordinates the services and activities within the data warehouse. This component controls the data transformation and the data transfer into the data warehouse storage. On the other hand, it moderates the information delivery to the users. It works with the database management systems and enables data to be properly stored in the repositories.

III. ONLINE ANALYTICAL PROCESSING SERVER

Online Analytical Processing Server (OLAP) is based on multidimensional data model. It allows the managers, analysts to get insight the information through fast, consistent, interactive access to information. In this chapter we will discuss about types of OLAP, operations on OLAP, Difference between OLAP and Statistical Databases and OLTP. [2]

IV. TYPES OF OLAP SERVERS

We have four types of OLAP servers that are listed below.

A. Relational OLAP (ROLAP)

The Relational OLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data the Relational OLAP use relational or extended-relational DBMS. ROLAP includes the following. Implementation of aggregation navigation logic. Optimization for each DBMS back end. Additional tools and services.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 4, April 2015

B. Multidimensional OLAP (MOLAP)

Multidimensional OLAP (MOLAP) uses the array-based multidimensional storage engines for multidimensional views of data.With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore many MOLAP Server uses the two level of data storage representation to handle dense and sparse data sets.

C. Hybrid OLAP (HOLAP)

The hybrid OLAP technique combination of ROLAP and MOLAP both. It has both the higher scalability of ROLAP and faster computation of MOLAP. HOLAP server allows storing the large data volumes of detail data. the aggregations are stored separated in MOLAP store.

D. Specialized SQL Servers

Specialized SQL servers provides advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

V. OLAP OPERATIONS.

As we know that the OLAP server is based on the multidimensional view of data hence we will discuss the OLAP operations in multidimensional data. Here is the list of OLAP operations.

A. ROLL-UP

This operation performs aggregation on a data cube in any of the following way:
Consider the following diagram showing the roll-up operation.

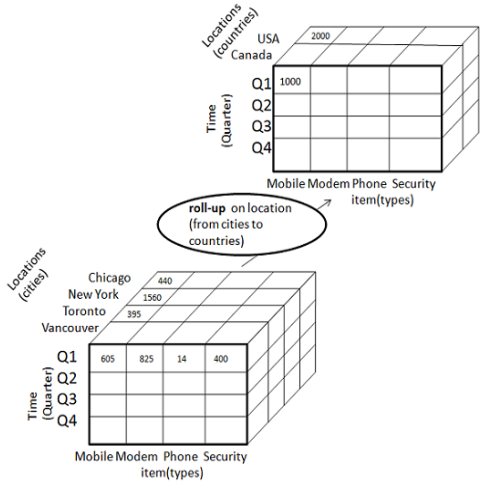

Fig 3 Rollup Operation

The roll-up operation is performed by climbing up a concept hierarchy for the dimension location. Initially the concept hierarchy was "street < city < province < country".

On rolling up the data is aggregated by ascending the location hierarchy from the level of city to level of country. The data is grouped into cities rather than countries. When roll-up operation is performed then one or more dimensions from the data cube are removed.

B. DRILL-DOWN

Drill-down operation is reverse of the roll-up. This operation is performed by either of the following way:
By stepping down a concept hierarchy for a dimension.
By introducing new dimension.
Consider the following diagram showing the drill-down operation:

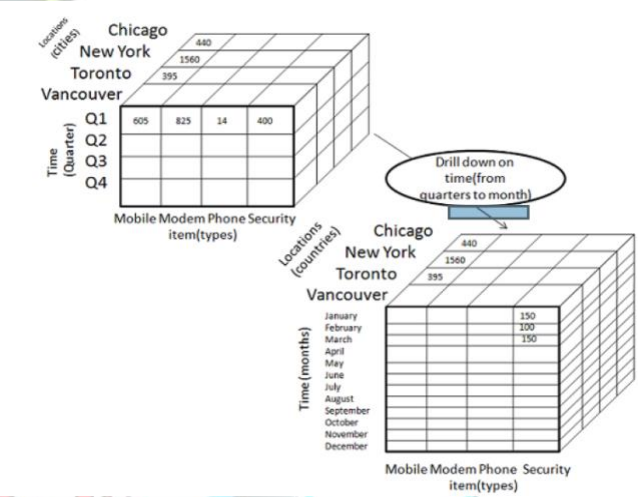

Fig 4 Drill Down Operation

The drill-down operation is performed by stepping down a concept hierarchy for the dimension time.

Initially the concept hierarchy was "day < month < quarter < year." On drill-up the time dimension is descended from the level quarter to the level of month. When drill-down operation is performed then one or more dimensions from the data cube are added. It navigates the data from less detailed data to highly detailed data.

C. SLICE

The slice operation performs selection of one dimension on a given cube and gives us a new sub cube. Consider the following diagram showing the slice operation. The Slice operation is performed for the dimension time using the criterion time ="Q1".
It will form a new sub cube by selecting one or more dimensions.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 4, April 2015

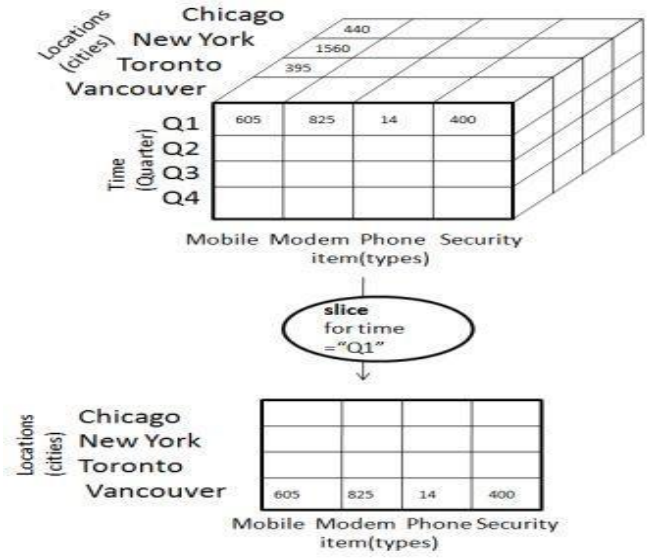

Fig 5 Slice Operation

D. DICE

The Dice operation performs selection of two or more dimension on a given cube and gives us a new subcube. Consider the following diagram showing the dice operation:

The dice operation on the cube based on the following selection criteria that involve three dimensions.
(location = "Toronto" or "Vancouver")
(time = "Q1" or "Q2")
(item =" Mobile" or "Modem").

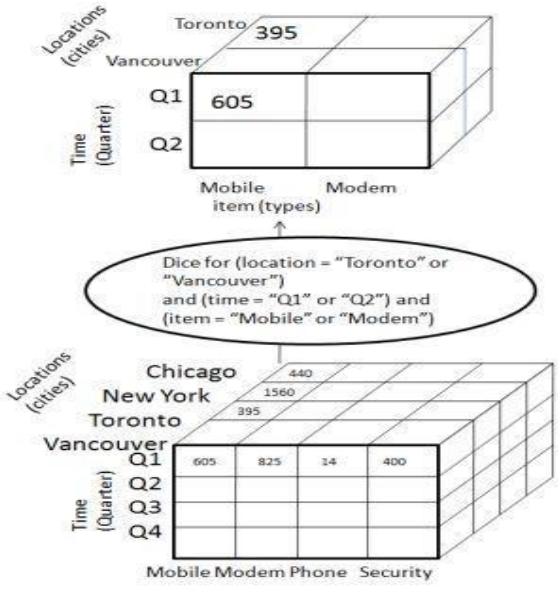

Fig 2 Dice Operation

VI. CONCLUSION

Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Many commercial products and services are now available, and all of the principal database management system vendors now have offerings in these areas. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications.

REFERENCES

[1]. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Paulraj Ponniah Copyright © 2001 John Wiley & Sons, Inc. ISBNs: 0-471-41254-6 (Hardback); 0-471-22162-7 (Electronic)

[2]. About tutorialspoint.com