ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

An Overview of Data Warehousing and OLAP Technology

B. Prabadevi[1], P. Aswini[2]

Assistant Professor, Department of Computer Applications, Pioneer College of Arts & Science, Coimbatore, India[1]

Student, Department of Computer Science, Pioneer College of Arts & Science, Coimbatore, India[2]

Abstract: Data warehousing and on-line analytical processing (OLAP) are decision support, which has increasingly become a focus of the database industry. Many marketable products and services are now available, and all of the principal database management system vendor now have offerings in these areas. Decision support places some rather different requirements on database technology compare to traditional on-line transaction processing applications. This paper provides an overview of data warehousing and OLAP technologies, with an emphasis on their new requirements. It can be back end tools for extracting, cleaning and loading data into a data warehouse; multidimensional data models typical of OLAP; front end client tools for querying and data analysis; server extensions for proficient query processing; and tools for metadata management and for managing the warehouse.

I. INTRODUCTION

Data warehousing is a collection of decision based technologies, aimed at enabling the knowledge worker to make improved and faster decisions. The last few years have seen explosive growth, both in the number of products and services offered, and in the acceptance of these technologies by industry. According to the META Group, the data warehousing market, including hardware, database software, and tools. Data warehousing technologies have been successfully deployed in many industries: industrialized, retail, financial services, transportation, telecommunications, utilities, and healthcare. A data warehouse is a subject-based, integrated, time, non-volatile set of data that is used primarily in organizational decision making process.[1] Typically, the data warehouse is maintain discretely from the organization's operational databases. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. Data warehouses might be implemented on standard or extended relational DBMSs, called Relational OLAP (ROLAP) servers. ROLAP servers are placed between relational back-end server and client front-end tools. With multidimensional data stores, the storage utilization may be low if the data set is sparse.
The review related technologies for loading and refreshing data in a data warehouse, warehouse servers, front end tools. In each case, we point out what is different from traditional database technology, and we declare representative products.

II ARCHITECTURE & END-TO-END PROCESS

This tool for extracting data from multiple operational databases and external sources. And for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for from frequently refreshing the warehouse to reflect updates at the sources and to cleanse data from the warehouse. There may be several department for data marts. Data in the warehouse and data marts is stored one or more warehouse servers, available for variety of front end tools: query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing metadata.

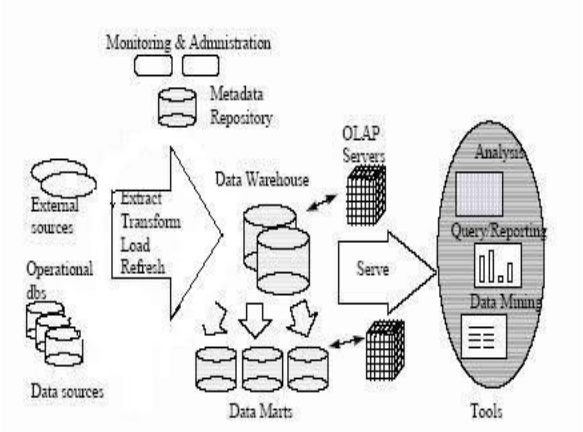

Fig. 1 Data Warehousing Architecture

The warehouse may be distributed for load balancing, scalability, and higher availability. An alternatively implemented for expediency when it may be expensive to construct a single logically integrated enterprise warehouse, is a union of warehouses or data marts.

III BACK END TOOLS AND UTILITIES

The warehousing systems used to a variety of data extraction and cleaning tools, and load and refresh utilities for warehouses. Data extraction from "foreign" sources is usually implemented via gateways and standard interfaces such as Information Builders EDA/SQL, ODBC, Oracle Open Connect, Sybase Enterprise Connect. A data warehouse is used for decision making, it is important that the data in the warehouse be correct. Large volumes of

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

data from multiple sources are involved, there is a high possibility of errors and anomaly in the data. Therefore, tools that help to detect data anomaly and correct them. There are three related data cleaning tools are Data migration tools allow simple transformation rules to be specified. Data scrubbing tools use domain-specific knowledge to do the scrubbing of data. Some tools make it possible to specify the "relative cleanliness" of sources. Tools such as Integrity and Trillium fall in this category. Data auditing tools make it possible to realize rules and relations by scanning data. Thus, such tools may be considered variants of data mining tools.

Load

After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints; sorting; summarization, aggregation and other computation to build the derived tables stored in the warehouse; building indices and other access paths; and partitioning to multiple target storage areas. In addition to populating the warehouse, a load utility must allow the system administrator to monitor status, to cancel, suspend and resume a load, and to restart after failure with no loss of data integrity.

The load utilities for data warehouses have to deal with much larger data volumes than for operational databases. There is only a small time window when the warehouse can be taken offline to refresh it. Sequential loads can take a very long time. Doing a full load has the advantage that it can be treated as a long batch transaction that builds up a new database. While it is in progress, the current database can still support queries.

Data Cleaning-Refresh

Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: when to refresh, and how to refresh. Usually, the warehouse is refreshed periodically. Only if some OLAP queries need current data, is it necessary to propagate every update. The refresh policy is set by the warehouse administrator, depending on user needs and traffic, and may be different for different sources.

Most modern database systems provide replication server that support incremental techniques for propagating updates from a primary database to one or more replicas. Such replication servers can be used to incrementally refresh a warehouse when the sources change. There are two basic replication techniques: data shipping and transaction shipping.

In transaction shipping the regular transaction log is used, instead of triggers and a special snapshot log table. At the source site, the transaction log is sniffed to detect updates on replicated tables, and those log records are transferred to a replication server, which packages up the corresponding transactions to update the replicas. Transaction shipping has the advantage that it does not require triggers, which can increase the workload on the operational source databases.

IV CONCEPTUAL MODEL AND FRONT END TOOLS

A popular conceptual model that influences the front-end tools, database design, and the query engines for OLAP is the multidimensional view of data in the warehouse. In a multidimensional data model, there is a set of numeric measures that are the objects of analysis.

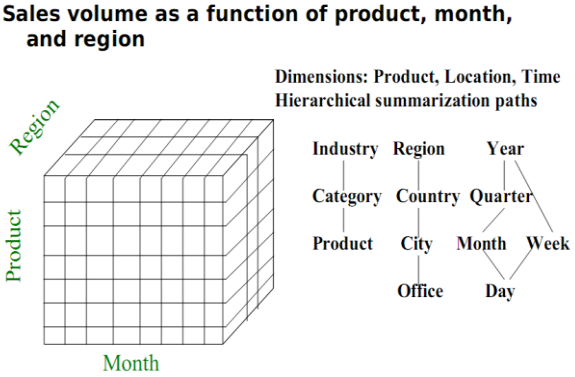

Fig. 2 Multidimensional data

Another distinctive feature of the conceptual model for OLAP is its stress on aggregation of measures by one or more dimensions as one of the key operations; e.g., computing and ranking the total sales by each county (or by each year). Other popular operations include comparing two measures aggregated by the same dimensions. Time is a dimension that is of particular significance to decision support.

Front End Tools

The multidimensional data model grew out of the view of business data popularized by PC spreadsheet programs that were expansively used by business analysts. The spreadsheet is still the most compelling front-end application for OLAP. The challenge in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operations efficiently over large multi-gigabyte databases.

V DATABASE DESIGN METHODOLOGY

The multidimensional data model described above is implement directly by MOLAP servers. However, when a relational ROLAP server is used, the multidimensional model and its operations have to be mapped into relations and SQL queries.

Entity Relationship diagrams and normalization techniques are popularly used for database design in OLTP environments. However, the database designs recommended by ER diagrams are inappropriate for decision support systems where efficiency in querying and in loading data is important. Most data warehouses use a

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

star schema to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. All the tuple in the fact table consists of a pointer (foreign key often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates.

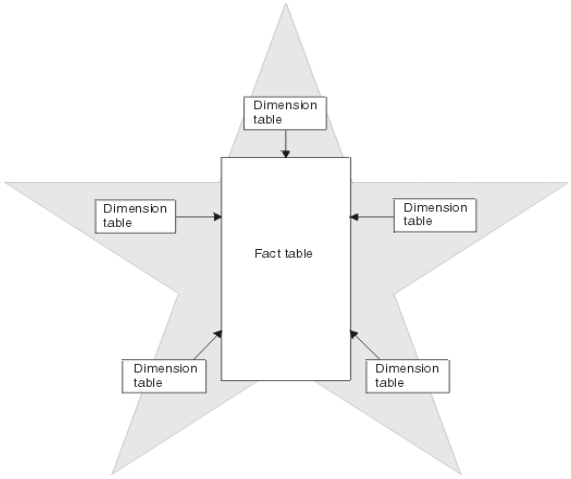

Fig. 3 A Star Schema.

Star schemas do not explicitly provide support for attribute hierarchies. Snowflake schemas provide a improvement of star schemas where the dimensional hierarchy is explicitly represented by normalizing the dimension tables, as shown in Figure 4. This leads to advantages in maintaining the dimension tables. However, the de-normalized structure of the dimensional tables in star schemas may be more apt for browsing the dimensions. Fact constellations are examples of more complex structures in which multiple fact tables share dimensional tables. For example, projected expense and the actual expense may form a fact constellation since they share many dimensions.

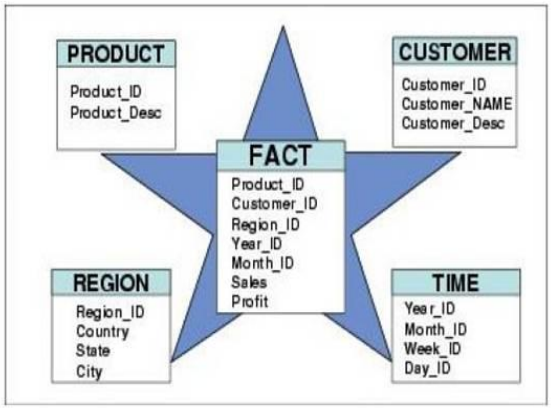

Fig. 4 Snow flake

In addition to the fact and dimension tables, data warehouses store selected summary tables containing pre-aggregated data. In the simplest cases, the pre-aggregated data corresponds to aggregating the fact table on one or more selected dimensions.

VI CONCLUSION

From the above work, Data warehouse do not contain the current information. However, data warehouse brings high performance to the integrated heterogeneous database system. It can be concluded that data warehouse have become base for OLAP tool for manage and retrieve information from large amount of database and implement technique on database that suitable for fast search of data within fraction of seconds. The construction of data warehouses involves data cleaning and data integration. Data cleaning attempt to fill in the missing values, smooth out the noise, while identifying the outliers and remove the inconsistencies in the data. As a result, data warehousing has become very popular in industry.

REFERENCES

[1] Inmon, W.H.,Building the Data Warehouse. John Wiley, 1992.
[2] http://www.olapcouncil.org
[3] Codd, E.F., S.B. Codd, C.T. Salley,"Providing OLAP (On-LineAnalytical Processing) to User Analyst: An IT Mandate."Available from Arbor Software's web site http://www.arborsoft.com/OLAP.html.
[4] Chatziantoniou D., Ross K. "Querying Multiple Features in Relational Databases" Proc. of VLDB Conf., 1996.
[5] Murlaikrishna, "Improved Unnesting Algorithms for Join Aggregate SQL Queries"Proc. VLDB Conf.,1992