Data Warehousing and OLAP Technology

(Data warehousing)

Priyanka Jaroli, Palak Masson
Computer Science Engineering Department
Banasthali University, Jaipur, India

Abstract – *Data warehousing and on-line analytical processing (OLAP)are essential elements, which has focus on the database industry. Many products and services are now available, and all the management concept is based on database management principle. Data warehousing is create using to approach (1)top down approach (2)bottom up approach . Decision support places database technology is used but in different- different ways. compared both traditional on-line transaction processing and modern on-line transaction processing applications. This paper provides an overview of data warehousing and OLAP technologies. We describe back end tools for extracting, cleaning and loading data into a data warehouse; multidimensional data models are typical to use in OLAP; Now describe front end client tools for querying and data analysis; server extensions for efficient query processing; and tools for metadata management and managing the warehouse. In this paper also identifies some research and database problem .this technology is presented at the VLDB Conference, 1996.*

Keyword: *ROLAP, MOLAP, redundant structures , rollup ,drill down, slice-dice, pivot, Snowflake.*

(I) INTRODUCTION

Online Analytical Processing Server (OLAP) is based on multidimensional data model. It allows the managers, analysts to get in sight the information through fast, consistent, interactive access to information. Data warehousing is a " Subject Oriented, Integrated, Time-Variant and Nonvolatile collection of data that support management's decision making process" aimed at enabling the *knowledge work* to make better and faster decisions technology .IN past three years have seen explosive growth, in the number of products and services. the also explosive growth the adoption of these technologies by industry. According to the *META GROUP* the data warehousing market, including hardware, database software, and tools,in their project & the project is $2 billion in 1995 and after three year it increase $8 billion. Data warehousing technologies have been successfully deployed in many industries: manufacturing, retail (for user profiling management), financial services (for claims analysis, risk analysis, credit card analysis, transportation, telecommunications, utilities (for power usage analysis), and healthcare. this paper focusing on the requirements that data warehousesplace on database management systems (DBMSs).

A data warehouse is "A Structured Repository of Historic data" that is used primarily in organizational decision making. It is developed in an Evolutionary Process By Integrating Data From Non-integrated Legacy Systems. The data warehouse is maintained from the organization's operational databases. There are many reasons for doing this. The on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

OLTP applications typically automate data processing tasks such as order entry and banking transactions that are the bread-and-butter day-to-day operations of an organization. These tasks are structured and repetitive, isolated transactions. The type of data required detailed, up-to-date data, and read or update a few records accessed typically on their primary keys. Operational databases required hundreds of megabytes to gigabytes in size. Consistency of the database are critical. Consequently, the database is designed to reflect the operational semantics of known applications to minimize concurrency conflicts. Historical and summarized data is more important than detailed, data records. Since data warehouses contain summarized data, perhaps from several operational databases. This summarized data is also useful and easy to maintain. The orders of magnitude larger than operational databases; data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are high and the cost of the data manage is very expensive ,use complex queries.

Query throughput and response times are more important for transaction and data searching. To facilitate complex analyses , the data in a warehouse is typically modeled is used *multidimensionally* example, in a sales data , time of sale, district, salesperson, and product. Often, these are hierarchical; sale may be organized as a day-month-quarter-year hierarchy, product as a product-category in industry hierarchy structure.

OLAP include four type of operations.

(1) rollup : (data is aggregated by ascending the location)

(2)drill down:(On drill-up the time dimension is descended from the level)

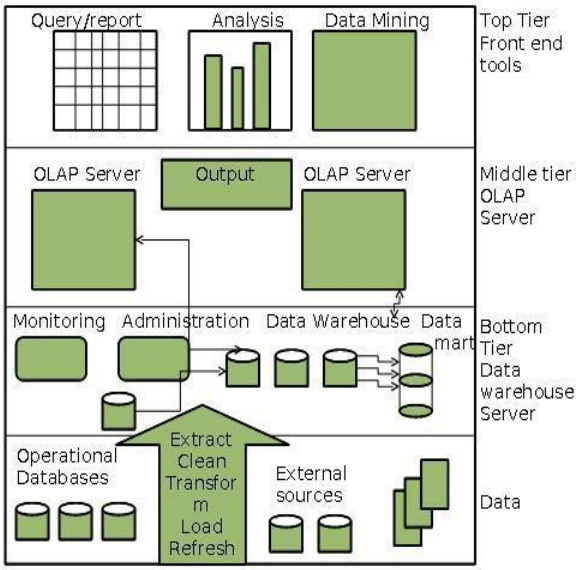
(3)slice-dice:(On drill-up the time dimension is descended from the level).

(4)pivot:(rotation)

OLTP workloads, trying to execute complex OLAP queries against the operational databases and result is unacceptable performance. In this reason decision support requires for data. predictions requires for store historical data, whereas operational databases store only current data. Decision support usually summarized data from many heterogeneous sources: these might include external sources like as stock market and several operational databases. The sources might contain data of varying quality or use inconsistent representation. Finally, the multidimensional data models and operations of OLAP requires special data organization, access methods, and implementation methods. The DBME not generally provided for OLTP. It is all these reasons that data warehouses are implemented separately from operational databases. The data warehouses might be implemented on standard relational DBMSs, called Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and they support to SQL queries and methods to efficiently implement the multidimensional data model and operations.

In multidimensional OLAP (MOLAP) servers are servers that directly store multidimensional data in data structures (e.g., arrays) and implement the OLAP operations. However, building an enterprise warehouse is a complex process and it take a lot of time. Some organizations are settling for *data marts.* which are focused on selected subjects (e.g., a marketing data include customer, product, and sales information).These data marts not work fast , since they do not require enterprise-wide consensus, but they may have complex integration problems and a complete business model is not developed.

(II) Data Warehousing Architecture and End-To-End process Communication:



It includes tools for extracting data from multiple databases and external sources is used for cleaning, transforming the data into the data warehouse and for periodically refreshing the warehouse and updates the database. In addition to the main warehouse, there may be several departmental data marts. Data marts is stored and managed warehouse servers, in the warehouse. multidimensional data is view and presentation is used variety of front end tools: query tool and data mining tools etc. Finally, there is storing 519 managing metadata, and tools for monitoring and administering the warehousing system. The warehouse may be distributed scalability, and higher availability. In a distributed architecture, the metadata is replicated with each fragment of the warehouse. Warehouse, is a federation of warehouses or data marts, each with its decentralized administration. Designing of the warehouse is a complex process.

Database is consisting of the following activities5.

(1) Define architecture, varies tool for capacity planning, select storage , database and OLAP servers.

(2) Combine to form a whole servers, storage, and client tools.

(3) Design the warehouse.

(4) Define the physical structure of the warehouse organization and access methods.

(5) Connect the gateways, ODBC drivers and other wrappers.

(6) Design and implement of the data store and retrieve.

(7)Design and implement and user applications.

(III) Back End Tools

Data warehousing systems use for a variety of data extraction ,storing, cleaning and for manage a proliferation. Data extraction mean "foreign". The sources is usually used for implemented uses gate ways and such interfaces tool (SQL, ODBC, Oracle Open Connect, Gateway).

Data Extraction:

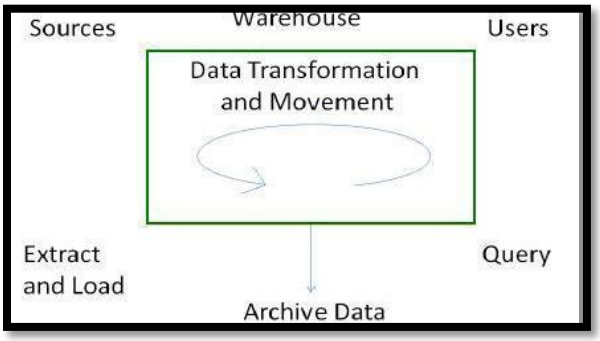
The data is store in the data warehousing system (disk, other devices)and when the data(information) is required it is extract from the data warehousing system(database).

Data Cleaning

A data warehouse is used as decision making, for that the data in the warehouse be correct. or not because, large volumes of data (proliferation) are available and store in warehousing and it is from multiple sources. the multiple sources are involved, a high probability of anomalies in the data.. The tools is used for that help to detect data anomalies and correct them can have a high payoff. Some
Examples of the data cleaning for necessary places: inconsistent field lengths, inconsistent descriptions, missing entries and violation of integrity constraints. *Data migration* tools allow simple transformation rules to be specified for *Data auditing* tools make it possible to discover rules and relationships 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 be required: checking integrity constraints, data sorting and other computation to build the tables stored in the warehouse . In addition to populating the warehouse, a load utility must allow the system administrator . The load utilities for data warehouses have to deal with much larger data volumes in the data base. Sequential loads
can take a very long time, e.g., loading a terabyte of data can take weeks and months.



Using checkpoints ensures that if a failure occurs during the load, the process can restart from the last checkpoint ,using parallelism, a full load may still take too
long . Only the updated tuples are insert. the load process now is harder to manage. The incremental load conflicts with ongoing queries, so it is treated as a sequence of shorter transactions (which commit periodically.

Refresh

Refreshing a warehouse propagating processupdates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two problem is created and we say that it is design issues *when* to refresh, and *how* to refresh.
In, the warehouse is refreshed periodically (e.g., daily or weekly). It necessary to update the every data. The warehousing administrator choose the polices for refresh according to the data and customer need and data traffic. Refresh techniques is also depend on the source and the database servers. When the data is extract from source file or database server it is expensive, but if we choose only legacy data
sources. In database server provide the replication (stand –by) server because if any fault is occur and data is lost the data is easily get from replication servers Such replication servers (stand –by) can be used for refresh a warehouse when the sources change. we are explain two techniques for replication.

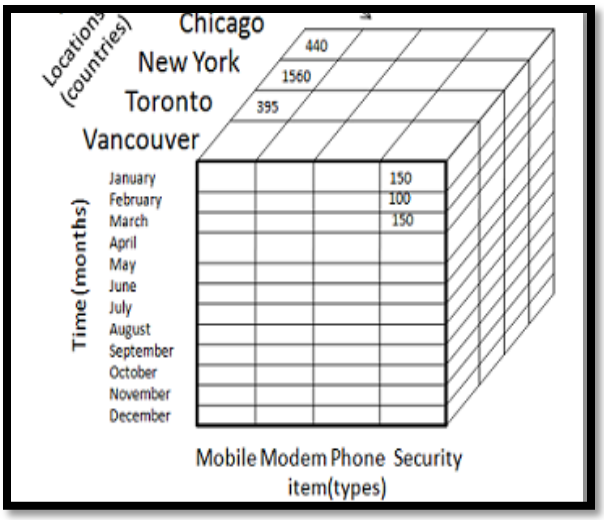
(1) data shipping and
(2)transaction shipping.

In data shipping used in the Oracle Replication Server Transactions hipping increase the workload on the operational source databases. it cannot always be used easily because there are no standard APIs for accessing the transaction log. data warehousing, the most significant classes of derived data are summary tables, single-table indices and join indices.

(IV)Front End Tool

The popular front-end tools, database design, and the query engines for OLAP is the *multidimensional data model* view of data in the warehouse. In this model there is a set of *numeric measures* that are the objects of analysis. Examples of numeric measures are budget, revenue. numeric measures depends on a
set of *dimensions.* Example: can be the city, product name etc. The dimensions are assumed to *uniquely* determine the measure. The multidimensional data views a measure as a value in the multidimensional space of dimensions. Example, the Product dimension consist of four attributes: (1)the category of the product (2) industry of the product, (3)year of

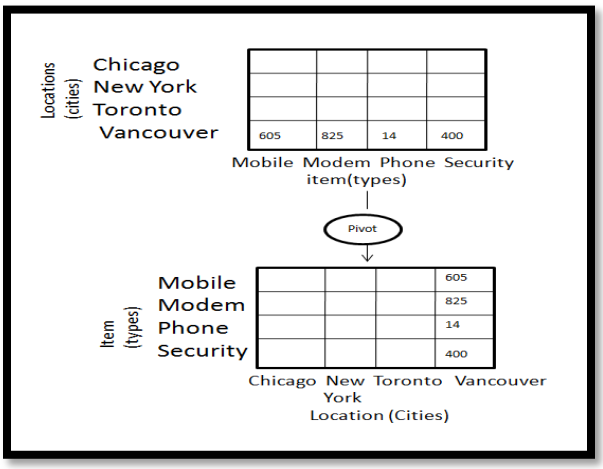
product introduction (4) average profit margin. For example, mobile ,the mobile industry, was introduced in 2005, and may have an average profit margin of 70%. The attributes of a dimension may be related via a hierarchy of relationships.



OLAP is its stress on *aggregation* of measures by one or more dimensions as one of the key operations computing and ranking the *total* sales by each county Other popular operations include *comparing* two measures aggregated by the same dimensions.

Front End Tools

The data model grew out of the view of business data popularized by PC spreadsheet. The spreadsheet is still the most compelling front-end application for OLAP. Microsoft Excel as the front-end tool for multidimensional engine. We shall briefly discuss some of the popular operations. One such operation is *pivoting*. Consider the example of pivot operation.



The simplest view of pivoting is that it selects two dimensions that are used to aggregate a measure. The aggregated values are often displayed in a grid where each value in the (x,y) coordinate corresponds to the aggregated value of the measure when the first dimension has the value x and the second dimension has the value y. Thus, in our example, if the selected dimensions are city and mobile phone security, then the x-axis may represent all values of city and the y-axis may represent the security. Other operators related to pivoting are *rollup* or *drill-down.*

Rollup corresponds to taking the current data object and doing a further group-by on one of the dimensions. Thus, it is possible to roll-up data, perhaps already aggregated. The drill-down operation is the converse of rollup. *Slice_and_dice* corresponds to reducing the dimensionality of the data. example, we can slice_and_dice sales data for a specific product to create a table that consists of the dimensions city and the month of security.

The other popular operators include *ranking* (sorting), *selections* and defining *computed* attributes.

(V) Database Design Methodology

The multidimensional data model described above is implemented directly by MOLAP servers.when a relational ROLAP server is used, the model and its operations have to be mapped into relations and SQL queries. entity Relationship diagrams techniques are used for database design in OLTP. the ER diagrams are inappropriate for decision support systems where efficiency in querying and in loading data are important. An entity is defined to be a person, place, thing, or event of interest to the business or the organization. An entity represents a class of objects, which are things in the real world that can be observed and classified by their properties and characteristics. In some books on IE, the term entity type is used to represent classes of objects and entity for an instance of an entity type. In the detailed ER model, defining a unique identifier of an entity is the most critical task. These unique identifiers are called candidate keys. From them we can select the key that is most commonly used to identify the entity. It is called the primary key. Chapter Most data warehouses use a *star schema* to represent the multidimensional data model. The database consists of a single table for each dimension. Each tuple is consist of a pointer for each of the dimensions that provide coordinates, and stores the numeric measures for those coordinates the values(data) is store in the database in the attribute.

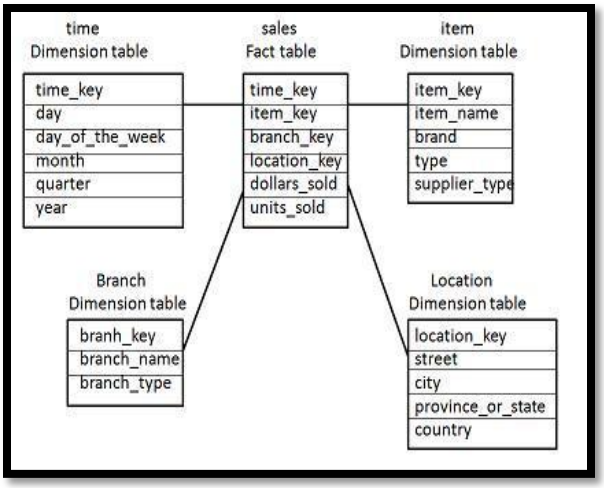


FIGURE: A STAT SCHEMA

In star schema each dimension is represented with only one dimension table. This dimension table contains the set of attributes. In the following diagram we have shown the sales data of a company with respect to the four dimensions namely, time, item, branch and location. There is a fact table at the centre. This fact table contains the keys to each of four dimensions. The fact table also contain the attributes namely, dollars sold and units sold.

Snowflake Schema:

In Snowflake schema some dimension tables are normalized. The normalization split up the data into additional tables. Unlike Star schema the dimensions table in snowflake schema is normalized for example the item dimension table in star schema is normalized and split into two dimension tables namely, item and supplier table.

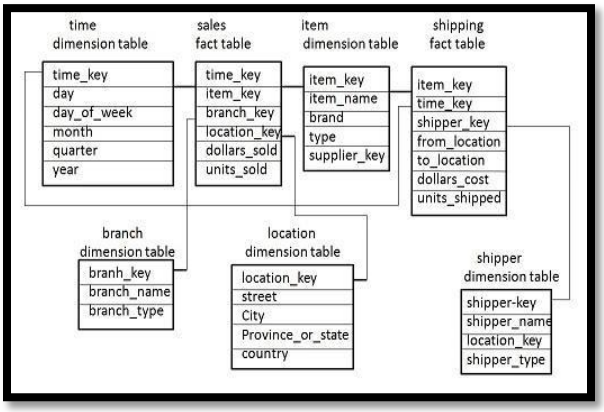


FIGURE: A SNOWFLAKE SCHEMA

(VI) Warehouse Servers:

Data warehouses may contain large volumes of data. Several issues arise. First, data warehouses use redundant structures such as indices and materialized views. Choosing which indices to build and which views to materialize is an important physical design problem. Second ,Optimization of complex queries is another important problem . Finally, parallelism needs to be exploited to reduce query response times.

Index Structures and their Usage

A number of query processing techniques that exploit indices

are useful. For instance, the selectivities of multiple conditions can be exploited through *index intersection.* . Consider a leaf page in an index structure corresponding to a domain value *d*. Such a leaf page traditionally contains a list of the record ids (RIDs) of records that contain the value *d*. bit map indices use an alternative representation of

the above RID list as a bit vector that has one bit for each record, which is set when the domain value for that record is *d*. In a sense, the bit map index is not a new index structure. Redundant array of inexpensive disks (RAID) technology has become common to the extent that almost all of today's data warehouses make good use of this technology. These disks are found on large servers. The arrays enable the server to continue operation even while they are recovering from the failure of any single disk. The underlying technique that gives the primary benefit of RAID breaks the data into parts and writes the parts to

multiple disks in a striping fashion. The technology can recover data when a disk fails and reconstruct the data. RAID is very fault-tolerant. The popularity of the bit map index is due to the fact that the bit vector representation of the RID lists can speed up index

intersection, union, join, and aggregation[11]. For example, if we have a query of the form column1 = d & column2 = d',

then we can identify the qualifying records by taking the AND of the two bit vectors .

(VII) Metadata and Warehouse Management:

A data warehouse use as a the business model of an Enterprise. An essential element of a data warehousing architecture is metadata management. Many type of data is store in warehouse and many technique for managed the data. *Administrative* metadata includes all of the information using a

Warehouse. the source databases is know as back-end and front-end tools define of the warehouse schema. The data, dimensions and hierarchies, predefined queries and reports as data mart locations and contents is physical organization such as data partitions, data extraction, cleaning, data refresh.

Business metadata includes business terms and condition, business definitions, ownership of the data, and charging policies. *Operational* metadata

includes information that is collected during the operation of the warehouse. the data is migrated and transformed data the currency of data in the warehouse and monitoring information such as usage statistics, error reports, and audit trails. the metadata is collection of data in the warehouse it store and manage all store data.

FUTURE Research Issues:

(1) Data cleaning
(2) Physical design of data warehouse(cost, efficiency, size)
(3) Management of data
(4) Data updating

Acknowledgment

This renew paper is made possible through the help and support from my parents, teachers, family. Especially, please allow me to dedicate my acknowledgment of gratitude toward the following significant advisors and contributors. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper. Finally, we sincerely thank to my parents, family. who provide the advice and financial support. The product of this renew paper would not be possible without all of them.

Conclusion

Data warehousing is an important field that is increasingly gaining attention as the internet. This OLAP technology is used in the data store the data in the database and specified how the data is managed in the database and how the large data is store this paper is solve the problem.

References:

[1] http://www.olapcouncil.org
[2] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.
[3] Data Modeling Techniques for Data Warehousing
Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell, Eunsaeng Kim, Ann Valencic
[4] Zhuge, Y., H. Garcia-Molina, J. Hammer, J. Widom, "View Maintenance in a Warehousing Environment, Proc. Of SIGMOD Conf., 1995
[5] Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Paulraj Ponniah
Copyright © 2001 John Wiley & Sons, Inc.
[6] Principle Partners, Inc. Info@PrinciplePartners
[7] Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin.
[8] O'Neil P., Quass D. "Improved Query Performance with Variant Indices".
[9] O'Neil P., Graefe G. "Multi-Table Joins through Bitmapped.
[10] Harinarayan V., Rajaraman A., Ullman J.D. " Implementing Data Cubes Efficiently" .
[11] Chaudhuri S., Krishnamurthy R., Potamianos S., Shim K. "Optimizing Queries with Materialized Views"
[12] Yang H.Z., Larson P.A. "Query Transformations for Queries"
[13] Widom, J. "Research Problems in Data Warehousing." [14] http://en.wikipedia.org/wiki/ olap design metholodegy
[15] Agrawal S. et.al. "On the Computation of MultidimensionalAggregates"
[16]Chatziantoniou D., Ross K. "Querying Multiple Features in Relational Databases"]
[17] Chaudhuri S., Shim K. "An Overview of Cost-based Optimization of Queries with Aggregates"