Data Warehouse and OLAP

Nan Huang
nhuang7@ur.rochester.edu

Jason Bergana
jbergana@ur.rochester.edu

ABSTRACT

In this paper, detailed descriptions of data warehouses and the OLAP system are provided. The features of a data warehouse that distinguish it from traditional relational databases are examined. The process of designing a data warehouse is discussed. The model behind the data warehouse and the OLAP are described. Lastly, queries and mining techniques in the OLAP system are also mentioned.

1. INTRODUCTION

In recent years, data warehouses have become a progressively important platform for data analysis. What is a data warehouse? How are data warehouses designed and built? How are OLAP systems (which have been designed from data warehouses) used to support decision making? In this paper these questions are going to be examined in detail.

2. DATA WAREHOUSE

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process"-William H. Inmon,. The words subject-oriented, integrated, time-variant, and nonvolatile can best describe how a data warehouse is distinguished from other data repository systems[1].

2.1 Key Features

2.1.1 Subject-oriented

Unlike a traditional database which is Object-oriented, data warehouse is subject-oriented. In a data warehouse, data is organized around subject instead of organized for the purpose of recording day to day transaction. A data warehouse is focused on modeling data for analysis purpose[1].

2.1.2 Integrated

A data warehouse is often includes data from multiple sources. These sources can be relational databases, flat files, or online transactions records. These sources are heterogenous. Hence, during the process of integrating the data, consistency needs to be ensured[1].

2.1.3 Time-variant

One purpose of a data warehouse is to store historical data for analysis. Since historical data is stored in a data warehouse, every structure in the data warehouse contains a time element[1].

2.1.4 Nonvolatile

Data warehouses are always separated from databases that aim to record daily transaction data. Therefore, recovery and concurrency control is not required for a data warehouse. Two things that are required for a data warehouse are the initial load of the data and the access of the data[1].

2.2 Why Use a Data Warehouse

Data warehousing provides the core function of decision support for an organization. This is distinguished from an operational database which is mainly used for clerical operations. Where the data warehouse is optimized for query throughput and response times the operational database is optimized for transaction throughput. Having a data warehouse that is isolated from the operational component is vital not only for functionality but also performance.

In terms of functionality, data warehousing is a collection of decision support technologies utilized across many industries[2]. These decision support technologies include the back end tools used for extracting, cleaning, and loading data, the front end tools used for querying and data analysis, and server extensions for efficient query processing[2]. Data warehouses provide tools for storing and managing multidimensional data and the supported OLAP technology utilizes multidimensional data models used for analysis[2]. All of this is aimed at enabling a knowledge worker to make better decisions.

Data warehouses are generally large in size designed to store hundreds of gigabytes - terabytes of data[2]. They also require a large investment of time and resources to build. The time required to build these data warehouses into an organization takes several years[2]. For this reason, an organization may opt for data marts instead of data warehouses. Data marts are departmental subsets focused on selected subsets of subjects.

2.3 Building a Data Warehouse

Data warehouses are complex, large, and tailorable (unique in respect to the organization or department). Building, planning, and deciding on certain features of a data warehouse is similarly complex for an organization. The degree of integration of the data warehouse is very much a factor in the complexity of the building process. Many organizations want an enterprise warehouse that spans the whole organization and collects information

about all subjects including customers, products, sales, assets, and personnel.

The designing, implementation, and testing of a data warehouse is a nonlinear process involving the coming-together of the back-end tools, the OLAP server, and the front-end client technologies.

Initially in the build process, the architecture of the data warehouse is defined[2]. The metadata repository, which will have metadata that is a reflection of the architecture of the data warehouse, is derived by thinking specifically about the business model of the enterprise. Different classes of metadata include: administrative (information required to set up and use a warehouse), business metadata (business terms and definitions, ownership of data, and charging policies), and operational metadata (information that is collected during the operation of the warehouse)[2]. Other initial design activities involve capacity planning, selecting storage servers, and selecting database and OLAP servers and tools[2]. This initial step in the process involves a lot of high-level decisions that have to do with a particular organization and the nature of relationship it will have with data warehousing. Continuing on, warehouse schema and views need to be designed[2]. The physical warehouse organization needs to be defined as well as data placement, partitioning and access methods[2]. On the front-end, applications need to be designed for eventual implementation. Implementation includes hardware integration. The servers, storage, and client tools need to be integrated[2]. Sources need to be connected including gateways, ODBC drivers and other wrappers[2]. Lastly, the repository needs to populated with schema and view definitions, scripts, and the rest of the metadata[2].

These activities are directly related to the functionality and substance of the data warehouse as being built. However, it is essential to mention the different classes of tools involved with facilitating the building process and the testing of the steps along the way. Development tools are used to design schemas, views, scripts, rules, queries, and reports as well as edit them[2]. Planning tools are used for capacity planning[2]. Analysis tools are used to analyze potential scenarios involving schema changes or refresh rates[2]. Lastly, the steps of the process are periodically checked and the scripts in the repository will be invoked given certain actions and events[2]. A workflow engine exist and is tasked with confirming the success of steps along the way as well as recording successful and unsuccessful trials[2]. This engine also provides failure recovery with partial rollback, retry, or roll forward[2].

3. OLAP

OLAP (online analytical processing) has been defined as the process of analyzing data from a data warehouse[1]. Data warehouses and OLAP tools are often developed on a multidimensional data model such as a data cube.

3.1 Data Cube

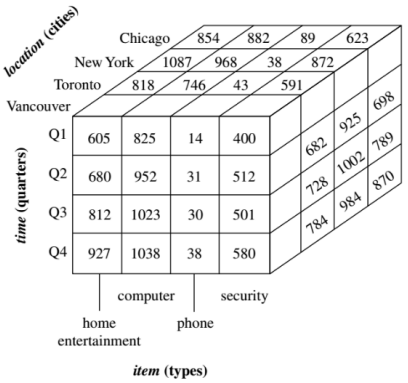

Figure 1: Data Cube (from [1] p.138)

Data is stored in multidimensional form visualized as a cube. Each dimension represents subsections of data that are comparable and aggregatable. The aggregation quality of the dimensions means that the dimensions have a hierarchy to them (e.g., months, quarters, years). Depending on the current state of display of the data cube, a particular level of the dimension will be displayed along that dimension of the cube. Other operations discussed below allow for other presentations of the data which contribute to the function of analyzing and supporting decisions.

3.2 Stars Schema

A data warehouse is subject-oriented, not object-oriented. The entity-relationship schemas which are used to design an object-oriented database are not appropriate to be used to design a data warehouse. Different schemas are required to design a multidimensional structure such as a data cube. Stars schema is the style of one of the most commonly used schemas for OLAP.

A stars schema contains a central fact table and several dimension tables. Each tuple in the fact table can be considered a recorded fact. The attributes of the tuples can be observed values or a pointer to the dimension tables. Inside dimension tables, tuples of attributes that are related to that dimension are recorded. Figure 2 shows an example of the stars schema. The graph in this example looks like a starburst in layout.

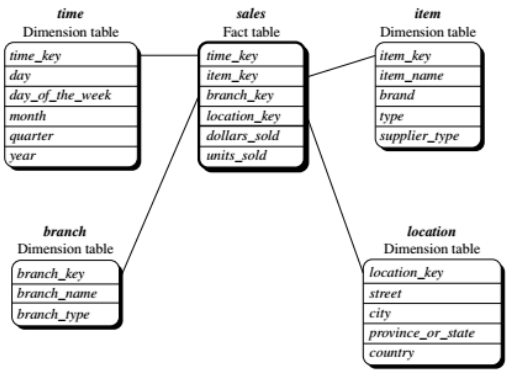

Figure 2: Star Schema (from [1] p.140)

Aside from a stars schema, there also other schemas that can be used to model multidimensional data such as the snowflake schema, a star schema variation. Fact constellation is another schema that has often been used. In the fact constellation schema multiple fact tables are allowed to share the same dimension table.

3.3 OLAP Operations

3.3.1 Roll-up and Drill-down

Roll-up and drill-down operations are performed along any given dimension. The roll-up operation will take the data cube from a lesser level of hierarchy to a more aggregated/higher level of hierarchy for a particular dimension. This will be shown with new labels (higher-level versions of the same dimension) displayed along that dimension of the data cube. The drill-down operation yields the opposite effect.

3.3.2 Slice and Dice

The slice operation takes data measures from one dimension and creates a subcube that visualizes only data from that dimension in relation to other dimensions. The dice operation performs is similar except that the subcube can involve two or more dimensions. Slicing and dicing are performed within the same hierarchical level that the data cube is currently in and displays certain subcubes of this data cube.

3.3.3 Pivot (rotate)

The pivot operation is an operation on the current data that is selected. This operation simply displays alternate views of the same data by shifting the data axes.

3.4 Query in OLAP

3.4.1 Starnet Query Model

OLAP is a multidimensional database, and queries in OLAP can be based on the starnet query model.

In a starnet model, radial lines emanate from a central point. Each radial line represents a dimension of the data, and the hierarchy of that dimension is represented along the line. Based on the starnet query model, the availability of OLAP operations such as Roll-up and Drill-down can be known.

Figure 3 shows an example of the starnet query model. The model has four dimensions which are time, location, customer, and item. To query the data, the data can be rolled up along the time dimension from day to month. The data can also be drilled down along the location dimension from country to city.

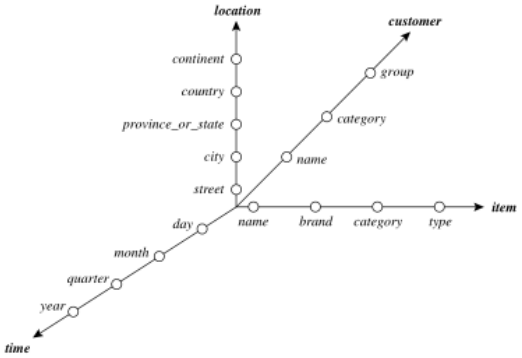

Figure 3: Starnet Query Model (from [1] p.149)

3.4.2 Indexing

Indexing can be used to assist processing a query in OLAP[1].

3.4.2.1 Bitmap indexing

Bitmap indexing is used to speed up searching in a data cube[1]. In a bitmap index, a bit vector is created for a given value of a given attribute. For a given row, 1 is stored if the attribute has the value, 0 is stored otherwise.

3.4.2.2 Join Indexing

In contrast to traditional indexing that maps values to rows that have a given value, a join index maps the joinable rows of two relations together.

3.4.3 SQL sever extentions

Serveral SQL extentions have been developed to process queries in OLAP[3].

3.4.3.1 Aggregated Function

Data warehousing is designed for analytical purposes. A variety of functions used for analysis such as mean, mode, and median are supported. The function rank and percentile are supported[3].

3.4.3.2 Reporting Features

For analytical purposes, aggregated data from a time window are needed. The SQL extensions makes it possible to query data that is similar to a moving average[3].

3.4.3.3 Multiple Group-By

Some queries in OLAP require multiple sets of attributes due to the fact that OLAP is a multidimensional database. Although this can be done by using a set of SQL queries, it is inefficient. SQL extensions have been developed to make roll-up in multiple dimensions efficient[3].

3.5 OLAP Mining

OLAP mining is a combination of on-line analytical processing and data mining. In OLAP mining, data mining is performed on different portions of the data warehouse and at different levels.

Because data mining can be performed in multi-dimensions and at multi-levels, OLAP mining is flexible and interesting knowledge can be discovered[2].

3.5.1 Cubing then Mining

In OLAP mining, the portion of the data that is going to be mined can be selected first. If this option is chosen, OLAP operations are performed first before applying data mining techniques. For example, a roll-up can be done on time in order to exam the data in a specific year[2].

3.5.2 Mining then Cubing

In OLAP mining, we can also mine the data first, then perform OLAP operations to fully understand the results. For example, a classification can be performed first on the whole data set, and then drill-down operations can be used to exam the result at each level[2].

.

3.5.3 Cubing while Mining

In OLAP mining operations, OLAP operations can also be performed during the process of mining. For example, while doing frequent pattern mining of the data, we can perform a drill down on time to find new patterns in a lower level , such as patterns in a different month[2].

4. Conclusion

A data warehouse is a collection of data from multiple sources. The data in a data warehouse is stored for analytical purposes. The structure of data in a data warehouse is multidimensional. Data warehouses support OLAP tools, and together they provide a valuable and fundamental function of data analysis and decision making.

5. REFERENCES

[1]Jiawei Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005

[2]Jiawei Han, "OLAP Mining: An Integration of OLAP with Data Mining", Proc. IFIP Conf. Data Semantics, pp. 1-11, 1997.

[3]Surajit Chaudhuri , Umeshwar Dayal, An overview of data warehousing and OLAP technology, ACM SIGMOD Record, v.26 n.1, p.65-74, March 1997