DATA WAREHOUSE FOR STUDENT INFORMATION SYSTEM USING STAR AND SNOWFLAKE SCHEMA

Han Su Aung, Yin Ko Latt
University of Computer Studies, Magway
hansu.1986@gmail.com

Abstract

Data warehouses collect information from remote sources into a single database. Data Warehouse is created for speed up queries by reducing the volume of data to be scanned. Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. OLAP system is a methodology that support summarized report for knowledge worker (e.g. analyst, executive) in related field. And then develop OLAP tools to retrieve information from data warehouse. This paper proposes to develop a data warehouse for historical student information. It collects the data from different student data and creates a data warehouse. The purpose of this paper is intended to view the detail of historical student's information for Computer University of Magway. Finally, this paper provides not only OLAP operation by using star and snowflake schema but also compares the time that take for searching historical data by using roll-up and drill-down operation between star and snowflake schema.

Keywords: Data warehouse and OLAP, Star Schema and Snowflake Schema

1. Introduction

A data warehouse is a semantically consistent data store that serves as a physical implementation of decision support data model and stores the information on which an enterprise needs to make strategic decisions. An OLAP system manages large amounts of historical data.

A data warehouse is a "subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making." Typically, the data warehouse is maintained separately from the organization's operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema. Star schema is a relational database schema for representing multidimensional data. In this star schema, fact table is in normalized format and dimension tables are in de-normalized format. The snowflake schema is a more complex data warehouse model than a star schema. Snowflake schemas normalize dimensions to eliminate redundancy [4].

The Data Warehousing approach presents some advantages such as more cost effective decision making, better enterprise intelligence, enhanced customer service, business process re-engineering, and information system re-engineering over the traditional approach for the integration of multiple sources. The entity relationship data model is commonly used in the design of relational databases where a database schema consists of a set of entities and the relationship between them. A data warehouse, however, requires a concise, subject-oriented schema that facilitates on line data analysis [1].

2. Data Warehouse

Data warehouse is a relational database that is designed for query and analysis rather that for transaction processing. They support information processing by providing a solid platform of consolidated historical data for analysis [6]. Data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis.

This definition of the data warehouse focuses on data storage. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system.

Data warehouses are arranged around the corporate subject areas found in the corporate data model. Usually the data warehouse is built and owned by centrally coordinated organizations, such as the classic IT organization. The data warehouse represents a truly corporate effort.

The data warehouse data structure is an essentially normalized structure. The structure and the content of the data in the data warehouse do not reflect the bias of any particular department, but represent the corporation's needs for data. The volume of data found in the data warehouse is significantly different from the data found in the data mart. Because of the volume of data found in the data warehouse, the data warehouse is indexed very lightly. The data warehouse contains a robust amount of historical data. The data warehouse data is integrated from the many legacy sources [5].

A data warehouse can be viewed as an information system with the following attributes [2]:

1. It is a database design for analytical tasks using data multiple applications.
2. It supports a relatively number of users with relatively long interactions.
3. Its usage is read intensive.
4. Its content is periodically updated.
5. It contains current and historical data to provide an analytic perspective of information.
6. It contains large few tables.
7. Each query frequently results in a large result set and involves frequent full table scan and multiple joins.

A data warehouse system comprises the data warehouse and all components used for building and accessing the data warehouse [2].

2.1 Architecture of Data Warehouse

There are two major components of data warehouse: the operational systems and the data warehouse. In addition to these databases, every data warehouse requires two additional components: software programs that move data from the operational systems to the data warehouse, (for building data warehouse) and software that is used to develop queries and reports (for accessing data from data warehouse). The architecture of every data warehouse includes each of these fundamental components [3]. Figure 1 represents the architecture of data warehouse.

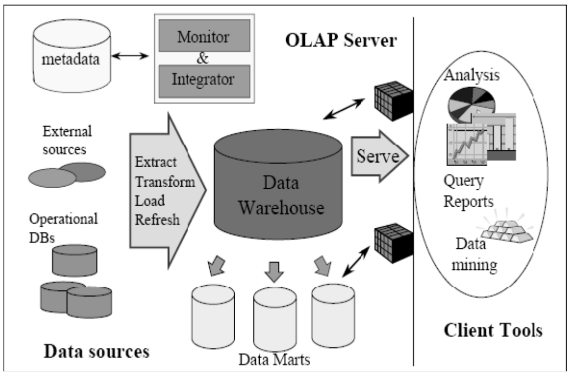

Figure. 1. Architecture of Data Warehouse

2.1.1 Operational System

An operational system is an application that supports the execution of a business process, recording business activity and serving as the system record. Operational systems may reside on a variety of database systems such as Access or SQL server database [2].

2.1.2 Dimensional Data Warehouse

The dimensional data warehouse is a database that supports the measurement of enterprise business processes. It stores a copy of operational data that has been organized for analytic purpose, according to the principles of dimensional modeling. Information is organized around a set of conformed dimensions, supporting enterprise-wide cross-process analysis. The dimensional data warehouse is usually implemented on a relational database management system (RDMS) [2].

3. Online Analytical Processing (OLAP)

OLAP stands for Online Analytical Processing. The Online Analytical Processing generally operates on historical data in both summarized and detailed forms. The major strength of on-line analytical processing is the multidimensional data analysis of data warehouse. OLAP system provides facilities for decision makers to access and analyze the data in the data warehouses. OLAP operation consists of drill-down, roll-up, dicing, slicing and pivoting. This system presents different data searching time between star and snowflake schema by using only two operations such as roll up and drill down.

3.1 Roll-Up

Roll-up which generalized one or a few dimensions and it performs appropriate aggregations in the corresponding measure(s). For non spatial measures, aggregation is implemented in the same way as in non spatial data cubes. However, for spatial measures, aggregation takes a collection of a spatial pointer in a map or map-overlay and performs certain spatial aggregation operation, such as region merge or map overlay. It is challenging to efficiently implement such operations since it is both time and space consuming to compute spatial merge or overlay and save the merged or overlaid spatial objects.

3.2 Drill-Down

Drill-Down which specialized one or a few dimensions and presents low-level objects, collection or aggregations. This can be viewed as a reverse operation of roll-up and can often be implemented by saving low-level cuboids, presenting it, or performing appropriate generalization from it when necessary.

4. Warehouse Schemas

A schema is a collection of database objects. The schema consists of star schema, snowflake schema and fact constellation schema. This system implements star schema and snowflake schema for student information.

4.1 Star Schema

An example of a star schema for student information system is shown in Figure 2.

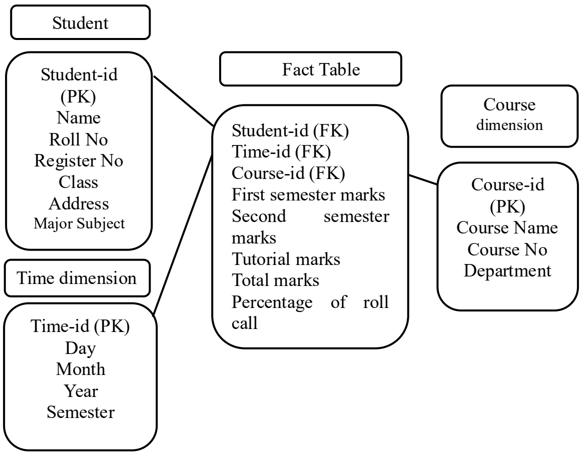

Figure 2. Star Schema of the Student Warehouse

Star schema contains a large central table (fact table) and dimension tables. The fact table contains the names of the facts, or measures as well as keys to each of the related dimension tables. This star schema is presented one fact table and dimension tables. Dimension tables are student dimension, course dimension and time dimension.

4.2 Snowflake Schema

An example of a snowflake schema for student information system is shown in Figure 3.

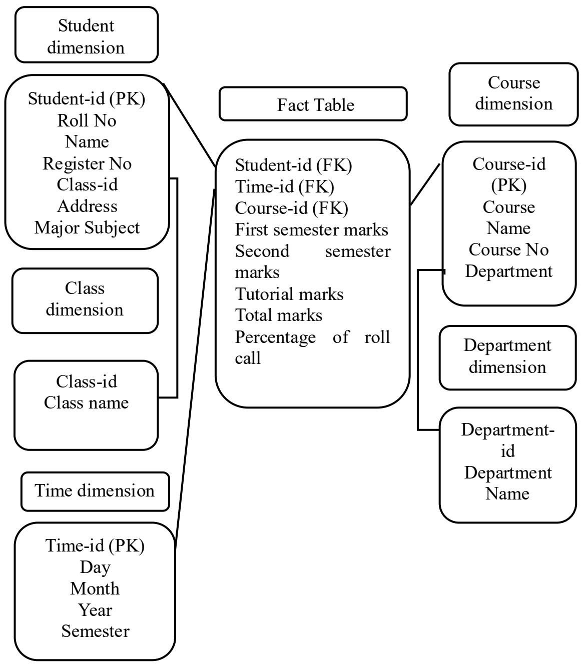

Figure 3. Snowflake Schema of the Student Warehouse

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, there by further splitting the data into additional tables.

Dimension tables are student dimension, course dimension, time dimension, class dimension and department dimension.

4.3 Differences between Star and Snowflake Schema

The differences between star and snowflake schema are as follows:

1. Star schema is de-normalized but snowflake schema is normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space because a large dimension table can become enormous when the dimensional structure is included as columns.
2. When star schema splits into additional dimension tables, it becomes a snowflake schema.
3. Performance will be more on star schema but performance less in snowflake schema due to joins.
4. Star schema is simple but snowflake is complex. Furthermore, the snowflake structure can reduce the effectiveness of browsing since more joins will be needed to execute a query. Hence the snowflake schema is not as popular as the star schema in data warehouse design.
5. Dimension table in star schema having no master table but dimension table in snowflake having so many master tables.
6. Star schema having one fact table surrounded by dimensional tables but in snowflake schema having more than one fact table surrounded by dimensional tables.

5. System Design

5.1 System Flow Diagram

In the system, the user can view the historical data of the student by using rollup or drilldown operation for both star and snowflake schema. And this system shows the difference of processing time between star and snowflake schema. Figure 4 shows System Flow Diagram.

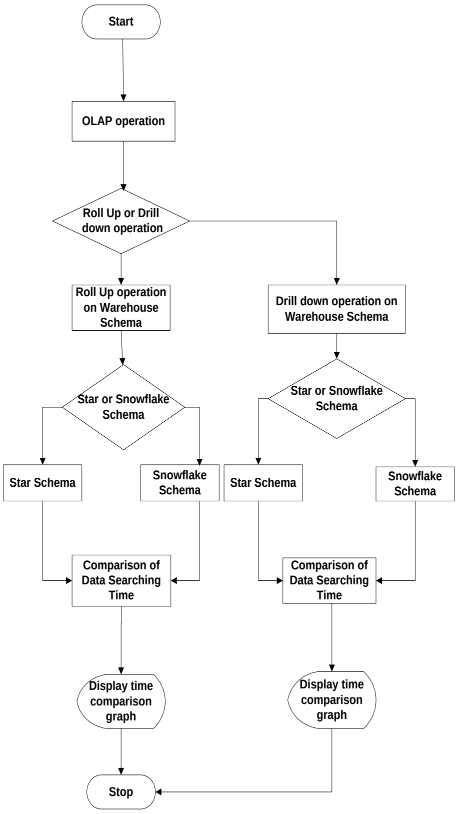

Figure. 4. System Flow Diagram

5.2 Use Case Diagram

The user can create and manage the data warehouse such as historical student data. The user can use the OLAP operation such as rollup and drilldown on the historical data about dimension. The user can view processing time comparison graph between star and snowflake schema. Figure 5 shows use case diagram of the system.

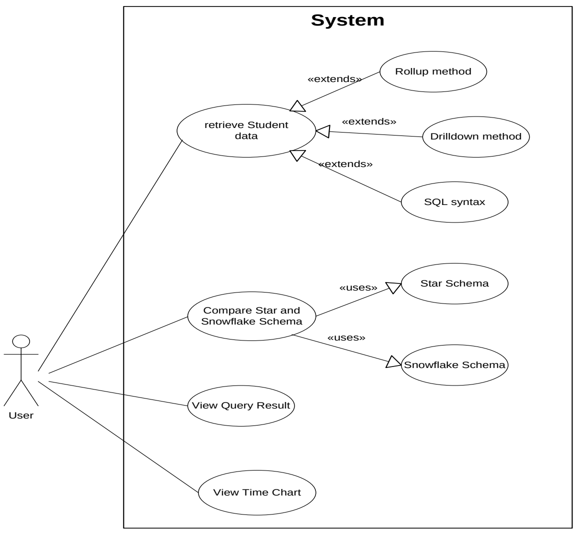

Figure. 5. Use Case Diagram of the System

6. System Implementation

In Roll-Up operation on Student and Department for Star schema, the user can view student historical data between 2000 and 2009 academic year. On Roll-Up operation, the user can view total first semester marks, total second semester marks, total tutorial marks, total marks and total percentage of roll call based on departments for each student. Figure (7) shows roll-up operation for star schema. Furthermore, the user can also view historical student data for both operations in star and snowflake schema.

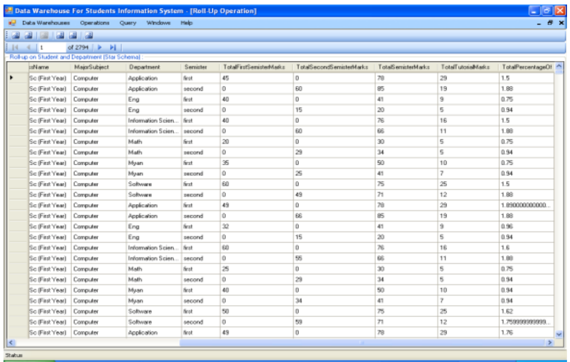

Figure. 7. Roll-Up Operation for Star Schema

The user can view the time that take for searching historical data in both star and snowflake schema by using roll-up and drill-down. When the user viewed comparison time graph for star and snowflake schema, the user can sharply known the star schema is better than snowflake schema. Because of star schema is compact and uncomplicated structure that facilitates the query responses on the data. Star schema provides highly optimized performance for typical star queries. It provides easy to understand and navigate between the tables due to less number of joins. This schema is simplest data warehouse schema and most suitable for query processing. The comparison time graph for roll-up operation is shown in figure (9).

Figure. 9. Comparison of Data Searching Time between Star and Snowflake Schema by using Roll-Up operation

7. Conclusion and Further Extension

A data warehouse is a subject- oriented, integrated, time- variant, non-volatile collection of data organized in support of management decision making. On-line analytical processing (OLAP) can be performed in data warehouses/ data marts using the multidimensional data model. A multidimensional data model is typically used for the design of corporate data warehouse and departmental data marts. Such a model can adopt a star schema, snowflake schema and fact constellation schema.

The construction of data warehouses based on a data cube model which consists of many dimensions. The development of data base technology for OLAP, the efficient storage of data cubes, to support for the user such as all teachers and staffs, and the historical student data interested user are interesting issues for research. So this system uses roll up and drill down operation for star and snowflake schema, the user can view the difference processing time between star and snowflake schema.

This system can be extended to other computer university in other divisions and states. This system is useful for all teachers and staffs from all universities and colleges in the whole Myanmar. It is also useful for other fields which business, finance, government and so on.

REFERENCES

[1] A. Perkiins, "Developing a Data Warehouse: The Enterprise Engineering Approach".

[2] A. Berson and S.J.Smith, "Data Warehousing, Data Mining and OLAP", Mc Graw Hill, England, 1999.

[3] C. Adamson, "Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance", http://www.amazon.com.

[4] C. Surajit and D. Umeshwar, "An Overview of Data Warehousing and OLAP Technology", pp. 2-3, Microsoft Research, Redmond, 2000.

[5] D. Murray and S. Anahory, "Data Warehousing in the Real World", Addision Wesley, England, 1997.

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Margon Kaufmann, USA, 2001.