Research on the Data Warehouse in the Design Process of the Database Application System

guojun Ma[1, a]
[1]Department of Computer Science，Gansu Normal University for Nationalities， Hezuo Gansu ,747000, China
[a]gnunjkxmgj@126.com

Keywords: Data Warehouse, Design Process, Optimization, Database Application System, Natural Processing

Abstract. This paper firstly introduces the recent research on data warehouse and describes the technology of data warehouse in process of design of database in detail. Data warehouse is a new technology in data management and information, and is mainly used to raise efficiency of data querying and to support decision. We use the theory of data warehouse to design database application system and to organize the database system in order to overcome the shortcomings of the database application system, such as low efficiency when there is a large number of data or in a new work, the data is difficult to transfer into useful information, and it can't satisfy the needs of long time analysis and prediction. According to the actual situation in a certain company, a concrete design of such a system is put forward in the paper.

Introduction

As the development of enterprise, traditional ERP, which is taking daily operating type handling as purpose, can't directly get the data needed by corporate executives, because data is extracted from different data sources. This phenomenon plays a restraining effect on enterprise management. As a result, enterprise needs an ERP system which is based on Data Warehouse and oriented analytical data to organize and present data as the demand of corporate executives [1]. This thesis introduces a typical case that a company applies the system based on Data Warehouse to the management.

Data Warehouse, based on the Database, is a new environment to support the decision analysis for satisfying management. Data Warehouse has some important properties: facing theme, data Integration, data long-period, therefore Data Warehouse is the core of Decision Support System [2]. ETL, the process of converting from different kinds of source data to the appropriate data type of Data Warehouse, is the beginning of Decision Support System. OLAP, based on the Multi-dimensional data type of Data Warehouse, can satisfy enterprise management by Slicing, Drilling (including rolling-up and drilling-down), and Rotating on Multi-dimensional data type. Thus, OLAP is the final result of Decision Support System. Then this thesis makes the development of Decision Support System example based on ERP system of a company. This section introduces and analyzes the demands of a Decision Support System. One is the demand of simple presentations of reports which don't need the multidimensional analyses; another one is the demand of multidimensional analyses of reports which don't need the drilling, last one is the demand of analyses from various years which don't need both multidimensional analyses and drilling. The author designs the data warehouse, ETL process and OLAP on Multi-dimensional data set, as the demands of users, to support the decision of corporate executives [3].

As defined by W. H. Inman, the Data Warehouse is a subject-oriented, integrated, non-volatile, time-accumulate data set, which is fit for decision. During the R&D in Statistic Data Warehouse, we apply the Data Warehouse to resolve the problem in Statistic System. During the research and development of Statistic Data Warehouse, we solve these OLTP related problem by using data warehouse technique and On Line Analysis Process (OLAP) application environment [4]. Also we design and implement a Visual Decision Support System based on data warehouse.

The problem that different information system may have heterogeneous or redundancy information is not conducive to the business process mining, assessment and optimization. Data warehouse can integrate the information in each data source, and makes it the basis of data analysis. A data warehouse built for the analysis of business process is called Process Warehouse. With the help of OLAP tools, it can implement information aggregating, analysis, and comparing, moreover, it can mining new process model and improve the quality of the existing process model.

Building a process warehouse will face lots of challenges: analyst may have different abstract level and data granularity; synchronization between process analysis and process automation; different information system has diverse life cycle, furthermore status number in life cycle is infinite; the relationship between dimension tables and fact tables; inhomogeneity of items in fact tables; the interchangeability of dimension tables and fact tables; diversity and so forth.

The Process Warehouse

In information systems, Process logs have detailed record of the execution of activities in process instance. It has a vital role as the data source of process warehouse. The process log in this paper adopts XES, which is based on XML, thus eliminating the need for evolution in process assessment model. The event here refers to the action defined in process model, and instance is the once execution of process.

In order to do process mining better in business intelligence environment, appropriate process mining models and assessment method for process mining results become key issues. Reference [5] proposes a process mining model used for event log mode, and it gives out a set of formulas for assessing process instance. Results of assessment are presented as numerical data, can be displayed directly for its significance. Through the setting of threshold, we can effectively select and classify the process instance, thus providing the basis for decision making for the follow-up process optimization.

Process warehouse based on the PAM stems from data warehouse, and here proposes a generic and process assessment-oriented process warehouse model (AOPWM). A single fact table is adopted in this paper, and its theme is the execution of process. To make it as granular as possible is suitable for a variety of data analysis.

In process instances, it can establish different tables according to different process types. The quality dimension makes it easy for evaluators to find the properties to assess, such as effrciency, customer satisfaction, cost and others. Here the quality dimension references attributes inefficiency dimension and cost dimension, and according to the importance of efficiency, customer satisfaction and cost, it can set a weight for them, calculates a numeric result of quality, which could intuitively represent the quality.

The efficiency of the process execution is defined as the number of nodes that are executed within unit time. Customer satisfaction is the percentage that the number of satisfied nodes accounts for the total in this process, and it is a very crucial performance indicator for enterprises, which can directly reflect the result of process to some extent. The cost dimension contains human, financial, material and other aspects of attributes, makes evaluators easily find the needed data. The time dimension has been throughout the entire fabric, plays a vital role, and can do a variety of statistical analysis using aggregate functions.

The process loaded into process warehouse during ETL is completed, here does not consider the uncompleted instance. Process belonging to different types may be executed more than once, and each execution will add one record in facts and the instance table at the same time. Due to the different data sources, it may lead to different integrity of each attribute in instance, in addition, the attributes and status of each instance are numerous, but in practical application, the properties enterprise concerned is limited, which makes our model feasible.

The Architecture of the Information Data Warehouse System

In information grid, there will be many databases or files store the history data. The data often includes: traffic information, business information, etc. Many departments will need to analyze its own data in the nodes, so we can design some node warehouse. The node warehouse will extract its data from the online transaction database or file. The overall data warehouse will extract data from every node warehouse by ETL tools. The ETL tools are used in data warehouse to extract and transform data from data source. The architecture of the information grid data warehouse system is given in Fig. 1.

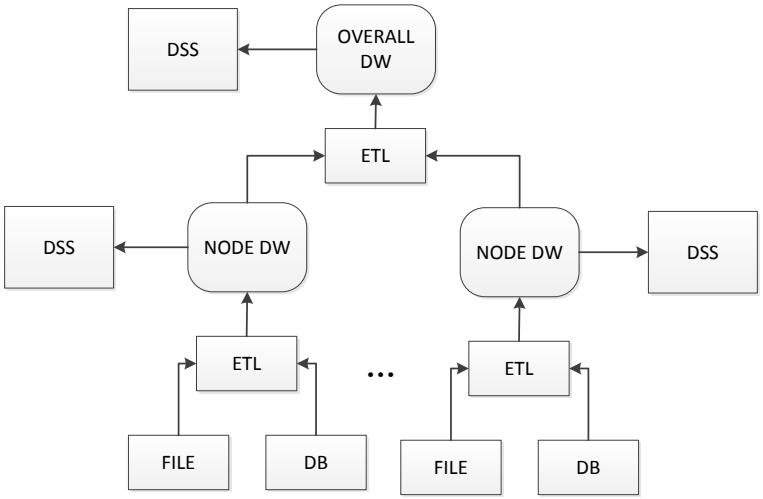

Fig. 1 The architecture of the information data warehouse system

In group company or large organization, there will be many departments inside of their organization. These departments often locate in different place of the world Each department will need to construct their own data warehouse called node warehouse. The overall department will need to realize the entire information of the group company or organization, so they need to construct the overall data warehouse. With the node warehouse, using DSS the direct department will analyze its data for decision; the overall department will analyze the overall information data for the overall decision.

The Difference and Relation between Database and the Node Warehouse

Data marts are usually smaller and focus on a particular subject or department. Some data marts, called department data marts, are subsets of larger data warehouses. Each data mart is used for a direct analysis, for instance; selling analysis, product analysis,etc. Compare with the node warehouse, the data marts and the node warehouse are two different concepts.

The node warehouse can contain some data marts and the overall data warehouse contains some data marts too. They are all subject oriented. They maybe contain the same subject. But in fact, the node warehouse' s data marts contain the node information and the overall data marts contain the overall information. The node warehouse usually is not subject oriented. For example, the node department is a sub company named company A, which is a sub company of a group company. So the node warehouse stores the sub company's information, the overall data warehouse store many sub company' s information The company A is a computer mainboard factory. This factory has a department of selling. So the company A's data warehouse is a node warehouse of the overall group company. The company A' s datawarehouse will at least contains two data marts: selling oriented and product oriented. The data marts will also contain in the overall group's data warehouse. It is the difference and relation of the data marts and node data warehouse.

Using distributed data warehouse, we can analyze the node data and overall database. This strategy can reduce the cost of development and maintenance. In a group company, if we only construct an

overall data warehouse to satisfy all the needs of each department, the management wiu be very complicated. It seems impossible for the overall department to extract data directly from the distributed departments'on-line transaction database or file. So, we must develop distributed data warehquse to realize these needs. Hence, in information grid, weneed to develop the distributed decision support system to analyze the distributed data.

The overall DSS can be disposed on the overall data warehouse. As discussed in the front of this paper, the overall data warehouse can extract data from the node warehouse using ETL tools. The overall data warehouse will contain the entirely data of all node warehouses. In the overall DSS, the data are from all node warehouses. So the overall data warehouse will lie a problem;how to reorganize the overall warehouse. To resolve this problem, we can do the follow steps; First, analyze the node warehouse and pick-up the public information; Second, redesign the model of the overall data warehouse;Third, extract data from the node warehouse or node data sources last, design and code the DSS' s analysis model After designing and loading the overall DSS, we can use ETL tools to extract data from the nodes.

The data structure in data warehouse is established based on the business system data structure. The data transformation in the system not only completes the simple task of converting the data format for the aims of untying the data format, but also integrates semantic differences between the two business systems, such as time characteristic and summary characteristic. The system should redefine data name, type, description and relationship including: unifying data type, adjusting data length and increasing time attribute.

Unifying data type. The same data with different data types must be unified as the same type. For example, as far as the date field is concerned, a system is defined as the date data, in other systems, it is defined as character data, and at last it should be unified as a character data.

Adjusting data length. If the same data own the inconsistent length, it should be adjusted for the unified length. If dealing with data's structure has the same structure with that of the data warehouse, data warehouse can load the data. After the data are loaded into the data warehouse, all records are ensured to be related to other table records, and verify each record in the fact table related to the record in the dimension table which is used by fact tables. All of these validations could be realized by the referential integrity between dimension tables and fact tables.

Conclusions

In this paper we have firstly given the grid information based data warehouse system architecture. Due to the fact that the distributed information grid data increases dramatically, and it is also becoming more and more difficult for us to obtain the useful knowledge, so developing one distributed DSS system is very meaningful and timely, it will also play an important role in using the historical grid information data of the northwest, China effectively and efficiently. We have designed the corresponding Data Warehouse model and DSS model under web environment and some security mechanism. So in our future work, we will focus on the following two tasks: Import the data mining functionality to the existing system, such as classification, clustering, such that it can help the analysts to make better decisions than the existing system to a larger extent. We will put the systems into more applications of real world. Such as the traffic field, enterprise management, etc.

References

[1] WANG L, ZHAO S K. The Data Warehouse Support the Research to Commercial Bank CRM. Information Science, Vo1.26, No.3, pp. 400-403, 2008.

[2] WANG H. The Design of Data Warehouse of the Civil Aviation Revenue Manage System. Computer Applications and Software, Vo1.21, No.6, pp. 49-50, 2004

[3] Kristin L. Anderson. Customer Relationship Management. McGraw-Hi11, 2001.

[4] W H Inman. Building the Data Warehouse. John Wiley & Sons, Inc., 2013

[5] XU P J, Data Warehouse & Decision Support System. Beijing: Science Press, 2005