Research on the data warehouse testing method in database design process based on the shared nothing frame

Keming Chen

School of Continuing Education, XinYu University,XinYu University, JiangXi, 338004, China

Corresponding author's e-mail: chenkemingxyu@163.com

Received 1 October 2013, www.cmnt.lv

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. After the infrastructure of database products was briefly introduced, the performance of cloud computing database under the workload of business type, testing technology standard of cloud computing database was especially analyzed, and the evaluate and assessment methods of capacity of cloud computing database was expatiated.

Keywords: data warehouse; design process; performance; benchmark.

1 Introduction

As the development of enterprise, traditional ERP, which is taking daily operating type handling as purpose, cannot 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.

Cloud computing, or in simpler shorthand just "the cloud", also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shred by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. This approach should maximize the use of computing power thus reducing environmental damage as well since less power, air conditioning, rack space, etc. are required for a variety of functions. With cloud computing, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications.

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 Multidimensional 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 do not 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 [5].

The problem that different information system may have heterogeneous or redundancy information may not be 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 [6-7].

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; in homogeneity of items in fact tables; the interchange ability of dimension tables and fact tables; diversity and so forth [8].

2 The basic framework of shared nothing and shared disk

Shared nothing and shared disk are two main technologies in the process of design the database or data 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 [9].

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 information. 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 [10].

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 will be very complicated. It seems impossible for the overall department to extract data directly from the distributed departments' online transaction database or file. So, we must develop distributed data warehouse to realize these needs. Hence, in information grid, wended to develop the distributed decision support system to analyze the distributed data.

2.1 THE FRAMEWORK OF SHARED NOTHING

Shared-nothing architecture is a distributed computing architecture and each node is independent, self-sufficient with the physical resources are not shared. The system even does not have a single point of competition. The most special is that all the nodes are not mutually shared memory and hard disk storage. The basic architecture is shown in figure 1. When large amounts of data and related application code concurrency are very hour, shared-nothing architecture is more suitable for the typical OLAP applications. Due to non-shared data, when one node fails, other nodes can take over the data node failure. Its main feature contains the large amount of data technology processing, low concurrency and low availability. This architecture is used currently in Teradata, Greenplum, Netezza, Vertical data warehouse and other products in the modern world.

Shared-nothing architecture can also be built from inexpensive ordinary PC and network hardware and Google, Amazon, Yahoo and MSN have proved this point. According to reports, Google's search cluster is supported by thousands of ordinary PC node acts as the shared-nothing nodes.

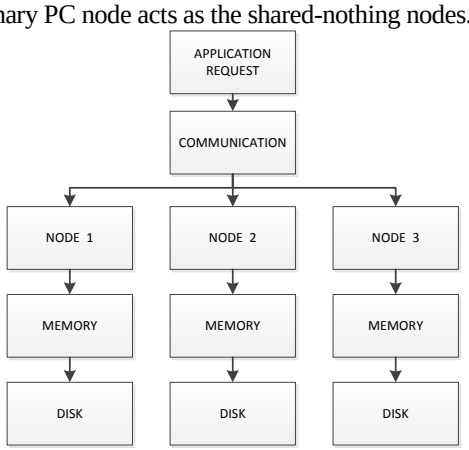

FIGURE 1 The basic architecture of shared nothing

2.2 THE FRAMEWORK OF SHARED DISK

Each individual node in shared-disk architecture is equipped with its own processor and memory. These nodes are accessed to the same physical storage sequence; the structure is shown in figure 2. When the application code size is large, high concurrency and the relevant data is relatively small, shared-disk architecture is more suitable for the typical OLTP applications. Due to the data sharing, when one node fails, you can transparently switch to another database node in order to run this job. Its main feature contains high concurrency and high availability. At present, a hybrid combination of EXA data database schema is prompted with the strengths of the shared-nothing and shared-disk architecture. This structure can effectively solve the conflict between the two and can absorb the two architectures to meet the high concurrency OLTP, and high availability. It can also deal with the large amount of data to meet the requirements of OLAP processing.

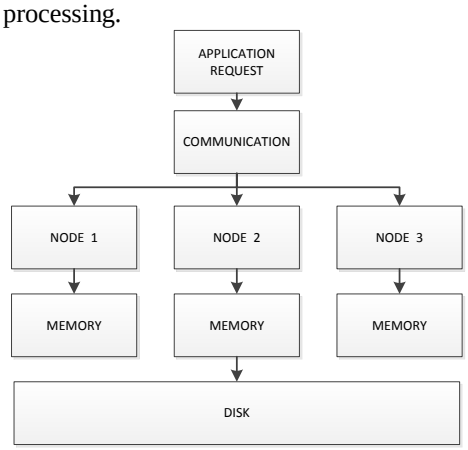

FIGURE 2 The basic architecture of shared disk.

3 Test design and assessment of test performance based on data 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 main board 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 data warehouse 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 will be very complicated. It seems impossible for the overall department to extract data directly from the distributed departments' online transaction database or file. So, we must develop distributed data warehouse to realize these needs. Hence, in information grid, we need 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 that 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.

Data warehouse in the application process is often associated with large table and related multi-table queries. But, it has to deal with multi-table split and the performance of the data warehouse is very important to pass this test. We can evaluate the multi-table database by the related query performance. Multidimensional association table and fact table is often used in data warehouse operation and the test case is used in multidimensional correlation of data warehouse between the table and the fact table or the speed of the connection. It provides multi-complicated list (on the order contains one hundred million or ten million) and some dimension tables is associated with the query and the main test case is consisted with the following three kinds:

Full table scan;

The big table is associated with a small table, then the whole table is again associated with a large table;

Large table is associated with many small tables.

Based on the existing data model, there is usually some new summary or query requirements, the database does not adjust logical, physical design related tables based on these new requirements, usually does not create a new database object to adapt to new demands. Database platform for such ad hoc query processing capability may cause a direct impact on users in the system evaluation.

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 efficiency, 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.

4 Experimental results

The test process can be described in following like this:

(1) Initial the value.

(2) Select a vector in random for calculation.

(3) Calculating the input source.

$$S^k = \sum_{i=1}^{3} \omega_{ij} X^k - \theta_j, \quad (1)$$

$$B^k = f\left(S^k\right), \quad j=1,2,3. \tag{2}$$

(4) Calculating the testing output.

$$L^k = \sum_{j=1}^{3} v_{jt} B^k - \gamma_t, \quad t=1,2, \tag{3}$$

$$C^k = f\left(L^k\right), \quad t=1,2. \tag{4}$$

Each performance item weight does not have a fixed set of criteria and we can test the performance of each month, accounting for project resource load situation set probably according to Telecom's data warehouse system. Resource accounting refers to the load items occupying time and processing capability. The server can handle the process by analysing the background.

The experimental result is shown in Table 1.

TABLE 1 The weight set for testing

Testing Item	Resources proportion	Weight
Loading	5%	6
Query	20%	15
Output	2%	2
Update	1%	2
…....	……	……

X86 architecture common to the test environment server cluster test environment (local server storage) 1, the test environment generic x86 architecture server cluster 2 (using SAN storage arrays) products and test environments dedicated database machine 3 is a product performance benchmarks test results are shown in Table 2.

TABLE 2 The Result for Performance Testing

Testing Item	Sample1	Sample2
Loading	32	35
Query	20	30
Output	2	3.3
Update	100	120
…....	……	……

Extending test results is shown in figure 3; Elongation test results is shown in figure 4 and data processing results is shown in figure 5.

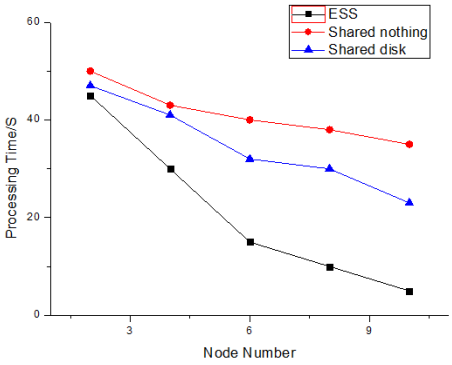

FIGURE 4 The elongation test results

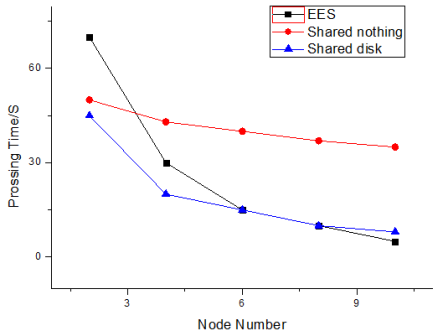

FIGURE 3 The extending test results

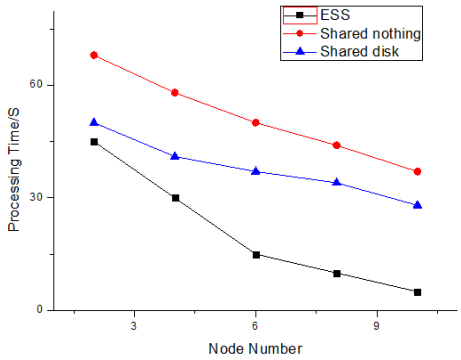

FIGURE 5 The data processing test results

5 Conclusions

It is can be seen from the experimental results, the use of shared-nothing architecture system can bring more excellent performance. The system only needs to add in the new common node query performance which can be achieved near-linear effective upgrade. This feature can help database system accommodate the massive data since the analysis and processing scenarios require rapid query performance improvement. The system only requires the addition of processing nodes based on performance to meet the performance requirements of the query. It should be noted that, due to limited capacity of the data loaded in the file server, its performance does not increase linearly.

References

[1] Wang L, Zhao S K 2008 The Data Warehouse Support the Research to Commercial Bank CRM *Information Science* **26**(3) 400-3

[2] Wang H 2004 The Design of Data Warehouse of the Civil Aviation Revenue Manage System *Computer Applications and Software* **21**(6) 49-50

[3] Anderson K L 2001 *Customer Relationship Management* McGraw-Hi11 237-9

[4] Inman W H 2013 *Building the Data Warehouse* John Wiley & Sons, Inc. 317-9

[5] Xu P J 2005 *Data Warehouse & Decision Support System* Beijing: Science Press 101-3

[6] Tai Jiang-zhe, Meleis W, Zhang Jue-min 2013 *Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads* 978-87 Northeastern University, Boston. USA

[7] Huberman B A, Adamic I A 2004 Information Dynamics in the Networked World *Lect. Notes Phys.* **650** 371-98

[8] Qin I, Yu J R, Chang I 2009 Keyword search in databases: the power of RDBMS *SIGMOD Conference* **1** 681-94

[9] Illenberger J, Kowald M, Axhausen K W 2011 Spatially embedded social network from insights into a large-scale snowball sample *The European Physical Journal B-Condensed Matter and Complex Systems* **2** 1-13

[10] Tirado J M, Higuero D, Isaila F 2011 Predictive Data Grouping and Placement Ivor Cloud-based Elastic Server Infrastructures *11th IEEE/ACM International Symposium on Cluster, Cloud and grid Computing. IEEE* 281-94

Author

Chen Keming, 1979-08-23, ShangRao City, JiangXi Province

Current position, grades: vice professorship
University studies: Computer science
Scientific interest: Database technology
Experience: From 2002 to 2014, working in XinYu University