Classification of Data Granularity in Data Warehouse

Haiyan LV
Naval Aeronautical and Astronautical University
Yantai,China
teacherlhy@163.com

Lijun Zhou, Yuan Zhao
Naval Aeronautical and Astronautical University
Yantai,China
xwc112329@126.com

Abstract: **To enhance the performance of the bank-data warehouse and reasonably determine the detail of the data and the data granularity degree of the warehouse, a sample of data granularity classification of bank-data warehouse is presented. Firstly the granularity model of data warehouse is analyzed, and then the strategy of granularity classification is introduced. Based on the strategy a method of estimating the data warehouse size is put forward. On the basis of knowing the principles of designing data granularity, a data granularity classification sample of bank is showed. Combined with the requirement of multiple granularity classification, design and use of granularity table is given to realize the effective management of multiple granularity. This divide method effectively solute one of the most important design questions faced by bank-data warehouse exploiter, and cause the other aspects of design and realization of data warehouse can be carried smoothly.**

Key Words: data warehouse; data granularity; classification strategy; overflow memorizer; Bank

I. INTRODUCTION

Classification of the data granularity is one of the most important design problems that data warehouse developers need to face. If the granularity of the data warehouse is determined to be reasonable, the rest of the design and implementation can be carried out very smoothly; conversely, if the is granularity determined to be unreasonable, it will make it difficult to carry out all the other data warehouse design. The data granularity is also important for data warehouse architecture designer because it will affect all the environments that need to obtain data from the data warehouse .The main problem for data granularity is to make it at an appropriate degree, and the degree is neither too high nor too low. Low data granularity degree will provide detailed data, but it will also lead the data to take up more storage space and require longer query times. High granularity degree can be quickly and easily queried, but cannot provide too much detail data. In the process of determining the appropriate data granularity degree, it is necessary to combine with the characteristics of the business, the type of analysis, the total storage space and so on. And the type of analysis is the most important factor.

II. DATA GRANULARITY MODEL IN DATA WAREHOUSE

Data granularity is the detailed or the comprehensive degree of the data unit in the data warehouse [1]. And it can be divided into two forms; the first form is the measure of the data comprehensive degree which will affect the data amount and the query types that can be satisfied by the data warehouse. The smaller the granularity, the higher the degree of detail, the lower the degree of comprehension, and the more query types can be satisfied. Conversely the larger the granularity, the lower the degree of detail, the higher the degree of comprehension, and the less query types can be satisfied. The another form is the granularity of the sample database, which is different from the above data granularity, and the granularity degree is not divided according to the degree of the data comprehensiveness, but divided by the sampling rate[2-4]. Sample database with different sampling granularity can have the same degree comprehensiveness. The sample database is typically a subset of extracted data from the detail file data or the mild composite data at a certain sampling rate. It is not a general purpose database, but a sample extracted from the data source according to certain demand or the importance of the data, and thus cannot satisfy detail query demand.

III. CLASSIFICATION OF THE DATA GRANULARITY

In practice, the above two forms of data granularity are both exist. In the traditional operational database system, the data processing and operation are carried out at the detailed degree, that is, at the lowest degree granularity [5-6]. As the data warehouse environment is mainly applied to do analysis procession, in order to facilitate the various analysis applications, the business data can be divided into four degrees by different data granularity: the current detail degree, the historical detail degree, light comprehensive degree and hign comprehensive degree. For the current detail degree data it usually kept with low degree granularity and the data have high detail. With the passage of time, according to the set time threshold and granularity threshold, the data are gradually aggregated, followed by the formation of mild comprehensive degree and highly comprehensive degree of data to save storage space and reduce system overhead. Data of different granularity degrees are used for different types of analysis.

A. ROUGH ESTIMATE

The first thing to do with granularity classification is to make a rough estimate of the number of rows in the data warehouse and the required number of DASDs (direct access storage devices). And often only an estimate of the magnitude. The method of estimating the number of rows and the occupancy of data warehouse is shown as following..

978-1-5386-3022-8/17 $31.00 © 2017 IEEE
DOI 10.1109/IHMSC.2017.142

The first step is to estimate the number of tables and the number of rows in each table to be created in the data warehouse, and then estimate the size of each row in the table. And it usually needs to estimate the upper and lower limits of the rows. Since the data access of the data warehouse is achieved by accessing the indexes, and the indexes are organized corresponded to the rows of the table, that is, there is always an index entry for each row in an index. Since the data access of the data warehouse is achieved by access, the index is organized in the rows of the corresponding table, that is, there is always an index entry for each row in an index. The size of the index is only related to the total number of rows in the table, regardless of the amount of data in the table. So the granularity classification is determined by the total number of rows rather than the total amount of data.

The second step is to estimate the minimum number and the maximum number of rows in the table for a year. And it is the biggest problem need the designer to solve. Such as: a customer table, it is necessary to estimate the current number of customers. If there is no business at present, it should be estimated as the produce of the total market business volume and the expected market share, and can use the competitors' estimate if the market share is unpredictable, and it need to use a reasonable estimate of the number of customers collected from one or more parties as the start value. Next is to estimate the number of one year's data units (estimated by the upper and lower limits), and then use the same method to estimate the five year's data units number. After the rough data estimate is completed, the space occupied by the index data is also estimated. Determine the length of the keyword or data element for each table and find out if there are keywords for each record in the table. The data storage space for each table can be represented by the sum of the table's storage space and the corresponding index occupied space.

B. DATA GRANULARITY CLASSIFICAT-ION STRATEGY

After a rough estimate of the size of the data warehouse is completed, the total number of rows in the data warehouse environment is compared with the table given in Table1. It is necessary to adopt different design, development and storage methods according to the number of total rows in the data warehouse environment. For example, if the total number of rows is less than 100 000 rows, then any design and implementation is actually feasible, no data need to be transferred to the overflow storage device to go. If the total number of rows is 1 000 000 lines or slightly less, then the design should be careful, but not necessarily the data transferred to the overflow memory. If the total number of rows exceeds 10 000 000 lines in a year, the design should not only be careful, but also some data to be transferred to the overflow memory. If there are more than 100 million rows in the total number of rows, there will be a lot of data to be transferred to the overflow memory, and the design should be very careful and careful.

TABLE I Data granularity classification strategy

one year's data		five years' data	
amount of data/number of rows	**classification strategy**	**amount of data/number of rows**	**classification strategy**
10 000 000	double granularity, carefully design strategy	20 000 000	double granularity, carefully design strategy
1 000 000	double granularity	10 000 000	double granularity
100 000	single granularity, carefully design strategy	1 000 000	single granularity, carefully design strategy
10 000	no need to consider the granularity	100 000	no need to consider the granularity

C. OVERFLOW MEMORY

The overflow memory that used to store the non-commonly used data is an important part of the data warehouse, and it has large effect on granularity. If there is no such memory, the designer must adjust the granularity degree to the capacity of the disk technology and the degree allowed by the budget [7]. But with the overflow memory, designers can let go to create the desired low-degree data granularity. The overflow memory can be built on any various storage media, commonly like the optical memory, tapes (sometimes referred to as "quasi-online memory") and inexpensive disks. Alternate form of mass storage is cheap and reliable, and can store much more massive data than the data stored by high performance disk devices, so the alternate form standby mass storage memory can be used as the data warehouse overflow memory.

However, if the user frequently accesses the data in the alternate form standby mass storage memory, the query is unpleasant, and in order to serve the query request, it will consume a lot of machine resources. Therefore, in the design of data warehouse, it needs to ensure that the data stored in the alternate form standby mass storage memory will not be frequently accessed. There are several ways to achieve this requirement. A simple way is to store the data into mass standby when the data reaches a certain age (e.g. 24 months). Another method is to store certain types of data in mass standby, while other types of data are stored in disk storage. Such as monthly customer record summary data can be stored in the disk storage, and the details data that produced the monthly summary data will be stored in the alternate form standby mass storage memory.

To allow the overflow memory environment to function properly, two software supports is required. One is a cross-media storage manager and another is a data activity monitor. Wherein the cross-media storage manager manages the data flow between the disk storage environment and the mass standby memory. The data

activity monitor is used to determine which data is being accessed, which is not accessed, and to provide the location information of the data storage (on disk storage or on mass standby storage).

IV. DETERMINE OF DATA GRANULARITY DEGREE

A. PRINCIPLES

When determine the data granularity, the following factors should be considered: the type of analysis to be accepted, the acceptable minimum granularity degree, and the amount of data that can be stored.

The type of analysis that can be satisfied in the data warehouse will directly affect the granularity classification of the data warehouse. The higher the degree of granularity, the less you can perform more detailed operations in the data warehouse. Such as: if the degree is defined as the month, it is impossible to use the data warehouse to do daily summary information analysis. And the data warehouses usually use multiple granularity in the same pattern, the granularity created this year and the previously year created both can be used, and this is based on the minimum granularity degree required in the data warehouse. Such as: you can use low degree data granularity to save the recent financial data and summary data, but to the far time data only reserve the high degree granularity summary data .And by this, you can not only take detail analysis for the recent financial situation, but also can use the summary data to analyze the financial trends.

Another important factor in defining the granularity of the data warehouse is the amount of space that multiple storage media can be used. If there are certain restrictions on storage resources, you can only use a higher degree data granularity classification strategy. And this classification strategy must be determined based on the user's understanding of the data needs and the information occupied storage size.

The classification of the data granularity is a compromise of the use of business decision analysis, hardware, and software and data warehouse. From the view of the demand analysis, it is hoped that the data can be saved in the most primitive (the detail state), and by this the conclusion of the analysis will be the most reliable. But too low granularity bring too large data size, and it will be increase the burden of the system CPU and I / O channel excessively and then reduce the efficiency of the system inevitably. So the classification of the data granularity must be combined with the characteristics of business data. The storage space of the system is another factor to consider. Too small granularity, very detailed data, will need great space requires cost, CPU and I / O pressure. And from this point of view high granularity degree is appropriate, but very general data often means the loss of the data detail and the reduced of the analysis reliability. In short, there is no strict stander for the classification of data granularity, and it will be determined by comprehensively consideration of analysis demand, system overhead, software capabilities and other factors on the basis of depth understanding of the business model. The Figure 1 shows the factors that need to be weighed to determine the granularity of the data.

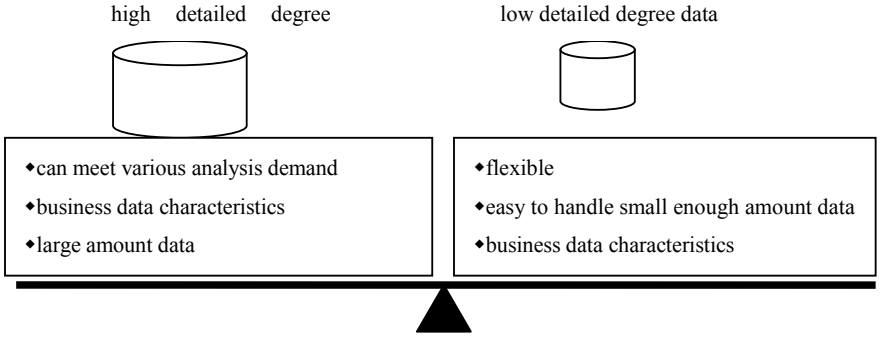

Figure1 Balance of data granularity classification

B. Multiple granularities

The data granularity will result the data amount and the ability of query answering for the data warehouse. [8]When the data warehouse uses a single granularity, all data is at the same granularity level. However, it is not possible to need the same granularity of all the data when the data warehouse provides decision support. Therefore, multiple granularities can be used to achieve a compromise between the amount of data and the ability of answering queries. Such as: a bank's data warehouse, we can use low-level granularity to save the recent customer transaction details data, and use high-level granularity to save the customer transactions summary data some time ago. By this we can not only do detail analysis for customer's recent transactions, and also can give guidance for customer's future transactions by use of historical summary data.Table II gives a comparison of single granularity and multiple granularities in terms of data storage volume and query answer ability.

TABLE II comparison of single granularity and multiple granularities

	single granularity		Multiple granularities
	low-level	high-level	
query answer ability	strong	weak	moderate
data storage volume	large	small	moderate

C. *ranularity table*

As described above, in order to meet the demand of data mount and query answer ability multiple granularities will be used more usually. And by this the problem of how to manage the granularity will be occurred, and it will influence the efficiency of data management.[9] And we propose to use granularity table to manage the granularities. Through the granularity table we can clearly know the granularity kind of each kind data, and it will improve the efficiency of the query. And when do regularly data update, the granularity table can provide the required granularity reference. Since the granularity table uses a relative time threshold, and it will not need to be updated when updating the data.

V. DATA GRANULARITY CLASSIFICATION IN BANK DATA WAREHOUSE

When data warehouse of a business or organization has a large amount of data, it is better to use double granularity (or multiple) degree to meet the detail need of the data. And in fact, it always needs multiple granularity degrees rather than one granularity degree. [11]

A. *Data granularities of bank data warehouse*

According to the principles and strategies described above, a example of multiple granularities classification of bank data warehouse is given in the paper, shown as Figure 2.At the operational degree is the operational data, that is, the details of the bank transaction data, the above is the current month trading data and the below is the synthesis data and mainly include the last month transaction records, and they are all stored in the operating environment. The right side of the operational data is highly integrated data totally include the last 5years the transaction records. The file degree data that is the overflow layer data mainly stores every detail of the record, and the data always saved in the alternate form standby mass storage memory. Not all of the fields are sent to the file degree, only those fields that need by law, information and other requirements required will be stored up.

B. *Design of the granularity table*

The granularity table usually consists of three elements: the relative time threshold, the subject name, and the granularity level. And because of the data warehouses are subject-oriented and only focus on data that can reflect the needs of managers for the macro analysis. For the data warehouse in the banking environment, the main storage data are the users' transaction records, that is, the subject can be considered only one: the transaction record. Therefore according to the above-mentioned bank data granularity, the granularity able can be defined as the following Table III

TaBLE III design of granularity table for rank data warehouse

time threshold	subject name	granularity level
the current month	product	low(detail)-level
	transaction	
the last month	product	light comprehensive-level
	transaction	
three years ago	product	high comprehensive-level
	transaction	

I.CONCLUSION

The granularity is essential for the reusability of the data, because it can be used by a large number of users in different ways. Such as: the data can meet the needs of the market, sale and finance departments at the same time. The data seen by the three departments are basically the same. The market department can understand the monthly sale of each region, the sale department can also understand the weekly sale of different sale staff, the finance department can understand the quarterly income of every production line. Another benefit of granularity is the history of the activities and events that encompass the entire enterprise. And another benefit of data granularity classification will let the data include the history of the activity and event of the entire enterprise, and it can meet the various reconstructions of the entire enterprise data based on various demand. Classification of data granularity is the most important problem of the design of data warehouse, and it will impact the entire architecture of the data warehouse. In this paper, we propose a method to estimate the size of data warehouse, and introduce the overflow memory related to data granularity. On the basis of deep understanding of the characteristics of banking data and the raised strategy and principles, the bank environment data granularity classification is given. The granularity classification of the sample is relatively simple, but it can better meet the need of banking industry analysis. And it can be taken as a template to give the other relevant data granularity classification of the banks..

REFERENCES

[1]John Wang. Encyclopedia of Data Warehousing and Mining[M] . Hershey, PA : Idea Group, July 8 2005.

[2]Wang Li-zhen,Zhou Li-hua.Principle And Application Of Data Warehouse And Data Miming[M].Bejing:SciencePress,2005.

[3]Peter C.Verhoef, Bas Donkers, predicting customer potential val ue an application in the insurance industry[J].2001, Decision Sup port Systems 32, 189-199.

[4]W.H.Inmon. Data Warehouse [M].Beijing : Machinery Industry Press,2003.

[5]Li Jing. Determination Principles of Data Granularity in Data Warehouse [J].Computer and Modernization,2007,02:57-59.

[6]Su Xin-ning,Yang Jian-lin,ect. Aata Warehouse And Data Mimin g[M].Beijing:Tsinghua Unicersity Press,2006,04.

[7]James Ang,Thompson S.H.Teo.Management issues in data warehousing:insights from the Housing and Development Board[J] .Decision Support Systems,2000.29:11-20

[8]Qi Li-na.Application of Data Granularity in Warehouse for E-co mmerce Website [J]. Computer and Modernization,2013,10:90-9

[9]Ren Chang-tao.Bank Customer Segmentation Model Based on D ata Mining[J]. Information Security and Technology,2013.05:84-8 7.

[10]Wu Hao,Yang Ji-shi.Analysis on Customer Behavior of Internet

Finance and Its Influence on Banking Innovation[J].Journal of Henan University(Social Science),2015,05:29-33.
[11]Zhang Jia.Commercial Bank Management Optimization UnderThe Big Data[J].Financial Management and Resear ch,2013,10:55-57.

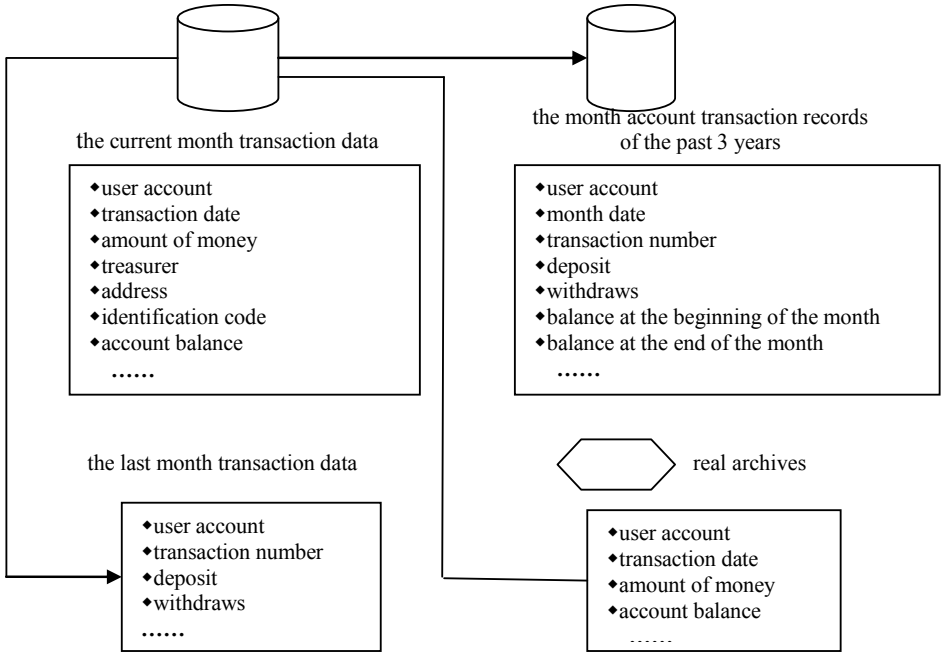

Figure 2 Multiple data granularity classification in bank data warehouse