Research and Application of Data Archiving based on Oracle Dual Database Structure

Cui Jin
Computer and Information Management Center of Tsinghua University, Beijing 100084, China
Email: jc@cic.tsinghua.edu.cn

Naijia Liu and Li Qi
Computer and Information Management Center of Tsinghua University, Beijing 100084, China
Email: lnj@cic.tsinghua.edu.cn, qili@cic.tsinghua.edu.cn

***Abstract*—For the universities' business system, to realize secure storage of historical data and long-term retention, saving high-end storage, online database backup pressure relief. A record-level data archiving method was proposed based on oracle dual database structure. This method included three aspects: related works, related technologies and strategy on data archive. For universities' business system, analysis of some problems need to concern and solution when the data archive, included four aspects: business analysis, storage analysis, data analysis and technical analysis. Finally, effectiveness of this method was validated by an application example in Tsinghua University.**

***Index Terms*—information flow, data management, data archiving, hierarchical storage management**

I. FORWARD

After twenty years' construction and development, Chinese universities' information system and application integration made great progress. With each passing day, every university's network and hardware improved a lot, and information system developed vigorously. Different data, such as teaching, research and management accumulates constantly. These data reflects university's daily activities, gradually becoming the core resources supporting university's teaching, research, management, service and cultural activities.

However, as time goes by, the data needed to be managed increases and accumulates continuously, and then many problems occurred. For example, numerous historical data of low access frequency and never-be-accessed brings low performance of data resources access; the pressure of data storage and backup becomes bigger, the cost of operation and maintenance increases constantly; data management becomes difficult and the security of the historical data cannot be ensured. Thus, historical data archiving is one key task of university information system construction.

II. DATA ARCHIVING DEMAND

A university relies on information flow and by this, it can support many activities such as teaching, research, management, services and cultural activities. The activities of a university, whatever it is teaching, research, social service or internal management, are the process of information's emergence, spread and acceptance. Moreover, every link of a university's teaching, research and management is passed mainly by information flow [1]. Thus, a university lays special emphasis on the effective organization and management of different data to realize standardization and process of university's management by speeding up information flow; to strengthen the management by promoting the closed-loop management of information flow; to advance the efficiency and the level of running a university by strengthening the development, share and using of information resources [2].

Data archiving is an important procedure of data lifecycle management, which supports information lifecycle. It stores historical data in a separate circumstance by a series of strategy flow and technology to reduce the pressure of product database and satisfy business demand. From the view of universities' data management, data archiving should meet these needs:

- Support the strategy of business policy is made by business departments and technical department. Business policy decides what kind of data should be archived, how much reserving time it needs and the classifying way of data.
- Support the variety and usability of the data. Database archiving need to consider the relationship among different information system, realizing data's centralized storage and management, meanwhile supporting the access and retrieval to archiving data.
- Ensure data's reality, integrity and security: business demands that the archived data shouldn't be revised or deleted to ensure the reality and integrity of the data, and stop the unauthorized access and tamper operation to secure the data.
- Make sure the consistency of the data: with the change of business, the meaning of the data may change a lot. The effective historical data may not exist and today's business or the new data and historical data may clash. Thus, when archive database we should not only ensure the integrity and efficiency but also solve the problem of consistency between current present data and historical data.

doi:10.4304/jsw.7.4.844-848

- Support the renewal of technology: With the development of technology, the promotion of database is inevitable. Data archiving should not only meet the need of current business but also consider these aspects: first how to move the data from low version database to high version database quickly; second, how to reduce the procedures of manual migration.

Universities' data has periodicity. Take management data of the students for example, the periodicity of universities' data include recruit students, educational administration, teaching, fee charging, and obtain employment. After graduation, the data produced from these procedures will become low-access frequency and never-be-accessed historical data. These historical data accumulates in product database, bringing great pressure to product database. Thus, the accessible speed of information system to data resources slows down. So we need to archive the historical data of product database according to the need of business. We should make sure the archiving range considering the need of business, such as objects, data scope, amount of data, the relationship of authorization in the level of data. Store the historical data in separate database to reserve long in storage level. Make sure the access from information system to the historical data in application level.

III. RELATED TECHNOLOGIES

The technology of data archiving is a way of keeping the scale of online database and providing the user with a stable database. Here is the principle (Finger 1): Move the seldom-used data to near-line device and archive the never-used data to file formats. Following the need of application, data moves among online, near-ling and file formats. For instance, if we need to access the historical data, it will move to online device automatically. For applications, this will not affect database application.

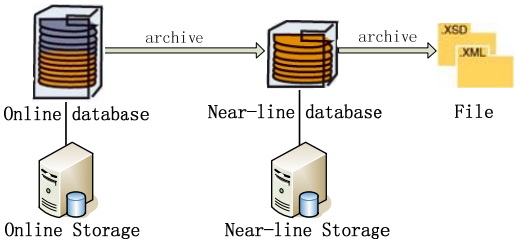

Figure 1. Principles chart of Database Archiving

Database archiving refers to copying some data to archiving database, and then deleting the data from online database. Generally, archiving is for analysis and audit, but not for recovery. The original file will be deleted after archiving, and must restore data manually. The technology of data archiving includes:

- Hierarchical storage means storing the data in terms of its level in different storage device at proper time based on data's importance, access frequency and amount of data. Hierarchical storage, including online storage, near-line storage and offline storage, as specified in Table I.

TABLE I.
HIERARCHICAL STORAGE

Storage	Data	Device
Online	Important and access to high-frequency	Reading speed, good performance, high cost
Near-line	Access frequency of low or no access	Capacity, relatively low performance and cost
Offline	The document is stored (such as XML)	Low-cost mass storage

- Data scanning means comparing the data of online database and near-line database, including the comparison between data structure and data content, counting of data as well.
- Data copy refers to copy the assigned data from online database to near-line database and keep the storage of these data in online database so as to ensure the integrity and consistency of data in online and near-line database.
- Data transfer refers to copy the assigned data from online database to near-line database, delete the archived data in online database and release the storage space in order to ensure the uniqueness of data and improve the performance of online database.
- ETL (Extract, Transform, Load) refers to extract data from source system and transform data format to destination system format. And then clean and process the error data or inconsistent data by business rule. The last load the cleaned data into destination system. ETL is used by building data warehouse generally and it is used by data archiving now.

Realize hierarchical storage of structured data using database archiving technique and storage historical data quickly of online database to near-line database via data scanning, data copy, data transfer and ETL. Thus, we can improve online database's performance and utilization of space, ensuring quick access form information system to the data of online database and supporting the retrieval of historical data.

IV. DATA ARCHIVING STRATEGY

Data archiving now refers to archiving structured data in database. It should have three stages: making archiving strategy, implement of archiving and the management of archiving data. First, making archiving strategy includes the analysis of business, storage and data, guarantee of technology as well. Then make out data archiving strategy eventually. Second, for implement of archiving, we should complete data archiving according to the existing archiving strategy. Third, in the stage of the management of archiving data, managing the archived data effectively and ensuring the accessibility and security of the data is important.

A. *Business analysis*

Considering from the aspect of time, the data stored in database is divided into two parts, that is, current detail data and historical detail data. Detail data is gathered by the system, reflecting source data of business. The analysis of business is based on the demand to data archiving. It means the analysis on system data flow and making sure the data area which need to be archived, considering the

relationship among the data. At present, bigger information systems support the management of many links. When the departments in charge manage its own business using information system, besides the data of its department, it may also need to use the data of other departments. So we say business analysis is the first task of making archiving strategy.

The data archiving model support the business can be described as a six-tuple formally.

$$A = \{B, I, R, T, D, M\}$$

B refers to the description of business systems, such as data flow and data structure, etc.

I refers to the interface description, is the business systems within and between business systems interface description, $I = \{I_i, I_o\}$, I_i describe the business system input data, I_o describe the business system output data.

R refers to the data archiving rules that is using natural language or formal language for data archiving implementation of norms and processes described.

T refers to the business requirements for data archiving, such as time threshold or time domain.

D refers to the business data sets, data sets associated with the business.

M refers to technology to support the implementation of the rules R method.

B. Storage analysis

From the angle of storage structure, Oracle database can be divided into physical structure and logical structure. Storage analysis refers to the counting of archiving data and estimating the logical storage space of database in advance, so that database administrator can prepare the storage space before data archiving. Meanwhile compare data of online database with near-line database to ensure the method of heterogeneous data archiving. The level of data archiving contains database, user, table, record, fields. Different data archiving level has different demands to data structure. Data archiving has two kinds, homogeneous database archiving and heterogeneous database archiving. Homogeneous means near-line database's structure and online database's structure is the same, otherwise, it is called heterogeneous. Database level archiving and user level archiving belong to homogeneous database archiving. Record level and fields level archiving belong to heterogeneous database archiving.

Data is stored in the table. Table of heterogeneous will directly affect the success to the archive data is stored near-line database. Table of heterogeneous is the near-line databases and online databases corresponding to the structure of the inconsistency. such as the number of fields, field names, field types, field length and other attributes in any one or more inconsistent. There are three treatment options for the field when the data archive.

- Field integration means based on business need, after the data is archived, some fields were added or deleted in some table of online databases. Data archiving, these tables in near-line database need to extension the corresponding field to ensure the consistency of the meaning of the field, needn't to consider the case of deletion, the field set of the on-line database can be a field subset of near-line database.
- Field mapping refers to online database is inconsistent with near-line database for the field name in the corresponding data table. Data archiving, the need to build the field mapping between the source field and the purpose field.
- Field Reset means that part of the field name of on-line database is modified after the data has been archived. Data archiving, the field name of near-line database must correspond to changes.

In addition, data archiving should also pay attention to a trigger for near-line database performance. If some tables of near-line database has triggers, the trigger is activated cause the database performance degradation when data archiving. However, for near-line databases, data integrity is not dependent on the trigger. Therefore, all triggers of near-line database are set to disable before the archiving of data. After completion of data archiving, and then all triggers are set to enable.

C. Data analysis

Thinking from application data contains encoded data, system data and business data. Encoded data is an important data combination of data, standing for one or a series of ordered symbol, and it should be easily dealt with by people or computer. In database many business data is stored in the form of code, so encoded data archiving is inevitable [3]. System data is one basic data supporting the running of information system and its archiving is to support query and counting from information system to archiving data. Business data is the main party of data archiving. The historical business data of online database are archived to near-line database, can improve the performance of online database. And then find the problem of data and solution and make sure the archiving flow.

- Encoded data

With the business management needs change, the meaning and content of the code may change, and even some of the historically valid code does not exist in the current business. Therefore, the data archiving requirements of code comparison in online database and near-line database, the results in three ways: equivalent code, compatible code, incompatible code.

1) Equivalent code refers to the structure and content of all the code table of near-line database are consistent with the online database. Therefore, there is no code deal with the problem in the data archive.
2) Compatible code refers to the code table of online database compared with the near-line database, some tables were added some fields or data, but does not affect the meaning and content of existing code. Therefore, only need to expand the structure or content of these code tables in the data archive.
3) Incompatible code refers to the meaning and content of the code have changed. Therefore, the

business of admin department needs to introduce a clear requirement handling code in the data archive.

- System data

System data is inserted into the database when the business system on the line. During system operation, the business administrator maintains data using the system function. In order to support the business systems query and statistical data in near-line database. The system data of online database is copied to near-line database when the data archive. The system data of near-line database ensure consistency with the online database.

- Business data

In the data archive, business data processing methods described below.

1) Code reduction refers to business data to the code stored in online database, when these data was moved to near-line database, the code is reduced to the name of the classification object in the business data table.
2) Code reset means that modify the code of near-line database for the code compatible, to maintain the business data tables and code tables in the same code.
3) ETL process is based on business requirements for online database data extraction, transformation and loading process to a near-line database.

D. Technology analysis

- Difference detection

Difference detection means that query some data dictionary views in oracle database, to check differences in data structure of near-line database and online databases. Frequently used data dictionary views include user_object, user_tables, user_lobs, user_tab_columns, user_constraints, user_tab_privs, etc.

- Calculate the amount of space

Row or column space S is calculated as follows. In the formula, getsize is a function used to obtain the field size, field is the field name in the data table. i is the column number or row number from 1 to n. Unit is byte.

$$S = \sum_{i=1}^{n} getsize(field[i])$$

LOB fields and non-LOB fields in the calculation of space to use a different SQL statement.

Example 1: LOB field calculation.

```
SELECT sum(dbms_lob.getlength(fieldname))
FROM table name
WHERE conditions;
```

Example 2: non-LOB field calculation.

```
SELECT sum(vsize(fieldname))
FROM table name
WHERE conditions;
```

The amount of space of a data table is equal the table size increases the size of LOB field in the table. Table size can be obtained through the data dictionary views, such as user_segments or dba_segments. For example:

```
SELECT sum(bytes)
FROM user_segments
WHERE segment_name = 'table name';
```

V. APPLICATION OF DATA ARCHIVING

Tsinghua university' educational system includes more than 20 business subsystem. There are more than 1,000 tables in database. Educational system' undergraduate thesis integrated management subsystem collected data nearly 30GB each year. Documents (DOC, DOCX, PDF format, text and abstracts of papers, etc.) are stored in four BLOB fields. In order to secure storage of these data and long-term retention, saving high-end storage, online database backup pressure relief. These data will be archive to the near-line database one-time after these students graduation. Data archiving technology involved data scanning, data classification storage, data replication, data migration, etc. The main work of data archiving are described below.

A. Preparation stage

- Check the difference of the structure and code. Data archiving involved more than 40 objects, including tables, synonyms, views, triggers and so on. Mapping (objects and fields) are created between the online databases and near-line databases. The results showed differences in 14 fields, including field name, field type and field length. The contents of these differences fields are queried to confirm the differences on the field did not affect data archiving. However, need field mapping and type conversion in the data archive. Code difference check results show that there are three kinds of incompatibilities. The first case is the different properties of the same code. The second case is some code to add in online database. The third case is some code of near-line database no longer exists in the online database. Using code compatibility mode to preserve the content and meaning of history code, business system is not affected to access the archive data.
- Calculated the amount of data and tablespace management. Calculated bytes field of the views USER_SEGMENTS using summation function to get the total amount of storage of archive data tables. Using getlength function in DBMS_LOB to get the total amount of storage of all BLOB fields in archive data tables. The corresponding SQL statement is as follows.

```
SELECT sum(BYTES)/(1024*1024) as "size(MB)"
FROM USER_SEGMENTS
WHERE SEGMENT_NAME in (T1', ' T2', …)

SELECT (sum(dbms_lob.getlength(FIELD1)) +
        sum(dbms_lob.getlength(FIELD2)) +
        ...) / power(1024,3) as "size(GB)"
FROM TABLE_NAME
```

Using above results to check and allocate storage of

tablespaces for near-line database, including free space and assigned to database users to use limited.

- Some programs written in SQL or PL/SQL.

B. Implementation stage

- Near-line database is set to 'write'. in daily operation, to avoid data corruption due to human factors, near-line database is set to 'read'.
- Disable all triggers in near-line database.
- Run the archive program. Run log is automatically written to TXT file. Log used to check the operation of SQL statements and data archiving progress. In the data archiving process, the database administrator monitors performance of the database.
- Check the archive results. Non-BLOB field using forward and reverse minus operation, the result is equal to 0, said the source field and target fields of the same content. Calculate the total capacity and compare the contents using ETL tool for the source BLOB field and the target BLOB field. Results are consistent that the source BLOB field and the target BLOB field are the same. In the business system, set various conditions, query and statistical data, check integrity and availability of the archive data.
- Enable all triggers in near-line database and near-line database is set to 'read'.
- Delete the source data in online database based on archiving requirements.

C. Ending stage

- For online database, free space of tablespaces check and adjustment.
- Complete the physical backup of the online database and near-line database.

In recent years, with the change and development of educational administration system, we replanned and constructed database storage environment, and set up two databases (online database and archiving database) structure based on oracle. So we can store product data and archived data in different databases to manage separately. Meanwhile, considering the characteristic of university's information system data and change of storage environment, we continuously adjusted and formed the archiving strategy and flow fitting for the characteristics of university's data. We achieved a lot in data archiving, such as

1) Introduced the concept of information life-cycle manage to data management. We can monitor, analyze and estimate in advance the amount of data and move the seldom-accessed data to near-line database.
2) Improved the performance of online database. Data archiving reduced the amount of online database, speeded up user's query, update and delete operation.
3) Made sure the security of data. Archiving database is a read-only database, and use can query but can't update and delete the archived data. If application wanted to edit the archived data, the archived data must copy to online database.
4) Increased efficiency of database backup. A lot of data in online database was copied to near-line database and deleted from online database. So database backup operations become more quickly and backup data was less.
5) Decreased storage's cost. Online database use high performance storage and near-line database didn't need high performance storage. So we decreased the online database storage cost.

VI. CONCLUSION

With the development of digital campus going deeper, the need for data archiving increases and it is the focus of university informatization at the stage of information integration. However, structured data directly reflects a university's business, having its own specificity and complexity, so it is hard to form a unitive and standard data archiving way and strategy. The research of structured data archiving is less than unstructured data archiving at home and abroad. So how to setup data archiving strategy fitting for its own characteristics in a university and how to reserve the historical data for a long time and manage it effectively? At present, we don't have a good way to solve these two problems and we hope to exchange ideas with the experts in this field.

REFERENCES

[1] Jiang Dongxing, Fu Xiaolong, Liu Qixin, Shen Liqiang, Wu Haiyan, Yuan Fang, "Digital campus construction tutorial in higher education", Beijing:Higher Education Press. 2008

[2] Yuan Fang, Fu Xiaolong, Jiang Dongxing. "Research and Practice of Data Management System in Digital Campus", Journal of DaLian Maritime University, Vol.35 supplement. 2009

[3] Kuang Kongwu, Wang Xiaomin. "Information systems analysis and design". Beijing:Tsinghua University Press. 2006

Cui Jin was born in Beijing, China in the year 1968. She is a engineer in Tsinghua University. She has Master degree. Her research interests include structured data management, data mining, information user service, etc.

Naijia Liu was born in Daqing, Heilongjiang Province, China in the year 1980. He is a engineer in Tsinghua University. He has Master degree. His research interests include database, network and information security, virtualization, etc.

Li Qi was born in Dalian, Liaoning Province, China in the year 1969. She is a senior engineer in Tsinghua University. She has Master degree. Her research interests include ITIL, educational informationization, etc.