Unstructured Data Migration and Dump Technology of Large-scale Enterprises

Shuo Chen[1,*], Shixin Fan[2], Zhao Li[1], Xinliu Wang[3], Shuangying Zhang[4], Zhuang Li[5]

[1]ICT Department, State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang 110006, China

[2]Personnel Director Department, State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang 110006, China

[3]Operation & Maintenance Center, State Grid Liaoning Information and Communication Company: Shenyang 110006 China

[4]Department of technology & economic, State Grid Liaoning Electric Power Company Limited Economic Research Institute, Shenyang 110015, China

[5]Fujian Yirong Information Technology Co., Ltd: Fuzhou, 350100 China

*Corresponding Author email: cs@ln.sgcc.com.cn.

Keywords: Enterprise-class unstructured data; Data migration; Data dump

Abstract: With the ceaseless development of enterprises, unstructured data processing between the original business system and the new data management system of enterprises is becoming one of the main problems. In view of that, taking into account the current situation of the information-based management of unstructured data, this paper mainly conducts detailed elaboration in terms of the technology of unstructured data migration as well as dump in large-scale enterprise from the perspective of platform design, migration function design, and dump function design.

1. Introduction

With the popularization of information-based office, the office management efficiency has been significantly improved. In this context, the video, reports, contracts and other unstructured data generated by enterprise are increasingly overwhelming, which has seriously affected the operational effectiveness of the original business system of the enterprise. Therefore, it is practically significant to analyze the unstructured data migration and dump technology of large-scale enterprise.

2. Current information management of unstructured data

The main problems existing in the current unstructured data information management are shown in table 1. From table 1, it can be seen that the existence of overall permission control problem, unstructured data integration problem and system query speed problem of the deep database have a bad effect on information management efficiency and management quality of the unstructured data.

Table 1 Main problems in information management of unstructured data

Type of Problem	Problem Description
Deep database lacks the overall authority control	For example, most data management systems cannot control the security of the used electronic files
The problem of unstructured data integration	The data management system has fewer mature interfaces, but unstructured data has more storage types; it is difficult to integrate different data.
System query speed problem	With the increasing of unstructured data, the system query speed becomes more and more slow.

  DOI: 10.25236/systca.18.065

3. Unstructured Data Migration and dump technology of large-scale enterprise

The research and analysis of unstructured data migration and dump technology of large-scale enterprise is mainly carried out from the following aspects:

3.1 Platform design

Platform design is the basic link of building an unstructured data platform of large-scale enterprise. In the process of platform design, the suitable platform design should be carried out according to the types, as well as the requirements of migration and dump of unstructured data. The composition of the Hadoop Platform is shown in Figure 1.

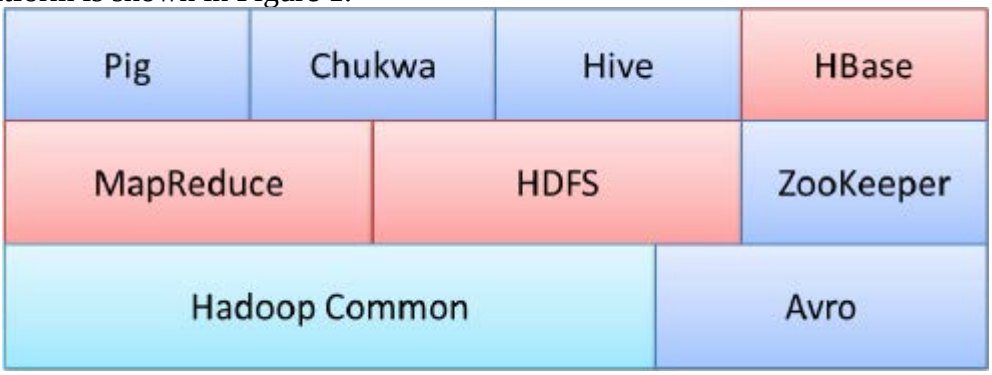

Figure 1 Composition of Hadoop Platform for unstructured data

3.2 Migration function design

In the process of realizing unstructured data migration technology of large-scale enterprise, the migration target of unstructured data of large-scale enterprise should be established. According to the data migration technology requirement, this research set up the migration target as: firstly, integrated management platform interface and business application system, whereby smoothing the migration of all the unstructured data in the original business system to the management platform; Secondly, to ensure the shortest maintenance time of unstructured data; thirdly, the data verification method that can process the management platform accurately and conveniently; fourth, ensure that the migration of unstructured data impost minimal disturbance for current system operations [1].

According to the above objectives and the requirements of large-scale enterprise data, the overall structure can be designed as follows: transfer is carried out from data source into adapter module (implemented by unified Interface); the module can transfer unstructured data to import processing module in a unified format. The transmission paths accessing Import processing module have three kinds: firstly, data transmission module. The import processing module will import data from the adapter module into the transmission module; secondly, data storage module; import the data into processing module, and transmit the data with unified format from adapter to data storage module, which is responsible for storing the corresponding data into the storage system and establishing a connection with the content management platform. Thirdly, import the log. The module's main parameters include the import start time, index start-up, log import speed, amount of import data, and whether it is successful. After the log is imported, the corresponding log file will be generated in the system, and the file contains both information log and deep log.

In order to ensure the successful migration and dump of large enterprise-level unstructured, with the case study of Hadoop structure, this paper analyzes the design of the enterprise-level unstructured data migration tool:

Judging from the composition of the data management system, Sqoop is often used as a migration tool for Hadoop architecture system, and the migration principle of the tool is shown in Figure 2.

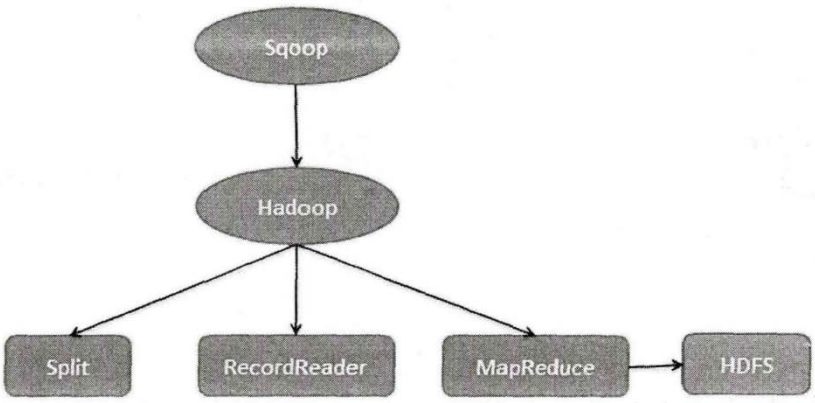

Fig. 2 Principle of Sqoop migration

However, the reliability of Sqoop data migration tool is evaluated according to the transfer requirement of large enterprise level unstructured data, and the analysis results are shown in table 2.

Table 2 Reliability evaluation of Sqoop migration tool for large enterprise-level unstructured data migration task

Reliability Impact factors	Factor description
Functional factors	In some application scenarios, data migration based on Sqoop tools can be completed by the supplementary means of complex data reprocessing.
Complexity factors of structure deployment	Sqoop architecture deployment is more complex, and requires a higher level of environmental configuration
Types and factors of supporting data	Sqoop Tool-supported large-scale enterprise unstructured data has less types and undesired practicality.

In this case, FTP can be used as a system migration tool, and the application advantages of this migration mode are shown in table 3.

Table 3 Advantages of the application of FTP in Hadoop structured data management system

Advantage Factor	Advantage Description
Factor of efficiency	The FTP tool is migrated in a multi-threaded manner, which shortens the migration time of large-scale enterprise unstructured data
Portability factors	FTP tool supports migration of entire folder data
Factors for rich supporting data-	FTP tools support a variety of large-scale enterprise structured data, thereby more practical

The migration process of enterprise-level unstructured data based on FTP is as follows: firstly, obtain the destination address of source file, the designated storage directory information from the configuration file; secondly, obtain the related file list information from the file list of the source file's directory; Thirdly, build FTP connection on the basis of data information and read the source file [2]; fourthly, transmit unstructured data through FTP connection; fifthly, complete the creation of blank documents under the specified directory, and input (write) related unstructured data in the

document[3].

3.3 The design of dump function

Conduct the analysis on the basis of the requirements of unstructured data dump of large enterprise-level. It can be considered that the dump function consists of 3 kinds of file archives: report summaries of unstructured data, table creation and storage of unstructured data, as well as unstructured file archive. Among them, the requirements for file content of unstructured file archives are shown in table 4.

Table 4 Requirements for file content of unstructured file archives

Content Elements	Specific Cntent
Source type tag	Facilitate the identification of a file directory or a single file
File size, file name, and relative path	Pre-verification of the related information in task files
Sequence codes	All documentation and data files correspond to unique sequence codes, which are listed in interface document

In the process of unstructured data dump, data query is undoubtedly the base of dump task. In order to ensure the smooth completion of the dump process, Hbase can be utilized to query. The query process based on Hbase is shown in Figure 3.

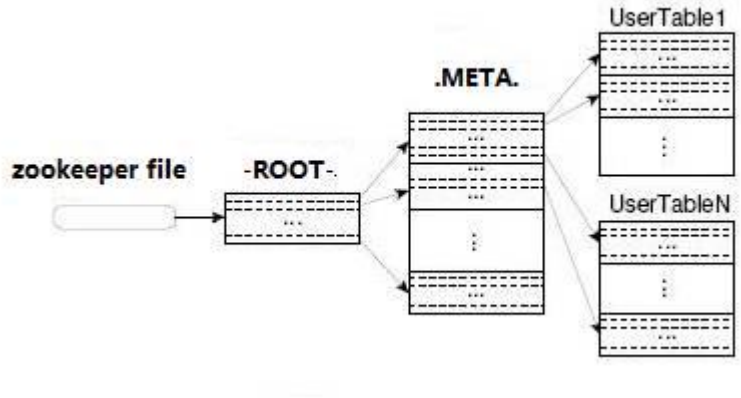

Figure 3 Unstructured data query based on Hbase

Based on the above, the dump process of the large enterprise-level unstructured data can be designed as follows: census the number of unstructured data entries in the system history; if the number exceeds 1000, unstructured data migration shall be carried out by adding to the queue. If the above requirements are not met, then conduct dump directly. This process needs to transform the values of the descriptive items in historical unstructured data; upon the conversion, upload the unstructured data attachment, and end the entire task after making sure the dump has been completed.

4. Conclusion

To sum up, unstructured data migration and storage of large enterprise-level are relatively difficult. In order to ensure the smooth completion of the migration and dump task, Hadoop structure can be applied, thereby designing the migration function and dump function that can meet the requirements of the unstructured data transfer of large enterprise level, so as to keep up with the growing requirements of the management of unstructured data, and to promote the benign development of the enterprises.

References

[1] Hwang J, Bai K, Tacci M, et al. Automation and orchestration framework for large-scale enterprise cloud migration [J]. Ibm Journal of Research & Development, 2016, 60(2-3):1:1-1:12.

[2] Barakos G, Vahdati M, Sayma A I, et al. A fully distributed unstructured Navier-Stokes solver for large-scale aeroelasticity computations [J]. Aeronautical Journal, 2016, 105(1050):419-426.

[3] Ibanez D A, Seol E S, Smith C W, et al. PUMI: Parallel Unstructured Mesh Infrastructure[J]. Acm Transactions on Mathematical Software, 2016, 42(3):1-28.