Research on Power Heterogeneous Data Storage Based on Distributed File

Zhao Li[1,*], Fengqiang Li[2], Nan Hu[3], ShiXin Fan[4], Nan Hu[5], Liang Yi[6]

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

[2]Office (Party Committee Office), State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang 110006, China

[3]Financial & Assets Department, State Grid Liaoning Electric Power Supply Co., Ltd: Shenyang 110006, China

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

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

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

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

Keywords: Distributed files; heterogeneous-configuration electric power data, data storage

Abstract: The expansion of the state power network coverage is ceaselessly forwarding, which directly product a large number of electrical power data, including video files, picture files and data files; therefore the storage of massive data has become an important factor limiting the construction of power management information system. Based on the distributed file storage system, this paper discusses the distributed file storage strategy that can meet the requirements of the heterogeneous data storage in state grid.

1. Introduction

Large data technology has been applied in more and more industries and scenes, and it has also played an important role in the process of collecting, analyzing and processing the power data of the state power grid. But at the present stage, the power information management system cannot meet the need of increasing scale of heterogeneous data storage. In order to support the further improvement of state power network intelligent management level and to meet the requirements of heterogeneous file storage in data explosion, new data storage technology is necessary.

2. Distributed file system

2.1 Hadoop distributed file system

Hadoop distributed file system can also be referred to as HDFS, this system is also the source implementation of GFS, which is widely used in PB-level data storage process. In recent years, Yahoo has achieved HDFS extension and expansion; the number of clusters of this system has reached 4000 nodes. When constructing HDFS, it can be accessed by streaming data, so as to facilitate the storage of ultra large file; it can realize write-once-read-many, so the data access is pretty efficient [1].

System has a strong fault tolerance, and can be compatible with poor equipment. The file model of it is enough to meet the need of write-once-read-many. In the process of reading, it can also be presented in the form of streaming data. The system architecture of HDFS is mainly a principal-and-subordinate form, which is composed with metadata nodes and other nodes [2].

  DOI: 10.25236/systca.18.066

Table 1 HDFS name nodes and data nodes

Name Nodes	Data Nodes
Storing Metadata	Storing File Contents
Metadata Saved in Memory	File Contents Saved in Disk
Save the Mapping Relationship Between File Block and Datanode	Maintained the Mapping Relationship Between the Block Id and the Datanote Local File

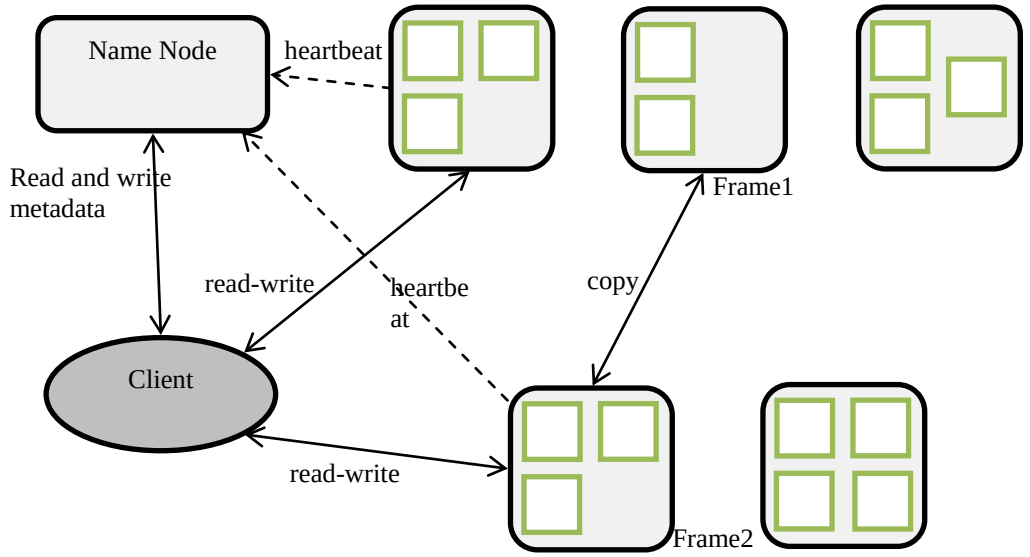

Figure 1 Diagram of HDFS system structure

In HDFS structure, Name node is the main node, undertaking the task of file name maintenance and management, and it is needed to manage the mapping relationship between data block and data node, including the namespace of file system. Save the metadata of all information and files, and a space image and modified log will be generated. Through the above channels, we can save file data Block, data Block distribution information, client access and other information.

Table 2 Advantages and disadvantages of HDFS system

Advantages of HDFS system		Disadvantages of HDFS system	
High fault tolerance	Data automatically saves multiple copies; Auto recover after any copy is lost	Low-latency data access	For example, millisecond level. Low latency and high throughput
Batch processing	Move calculations rather than move data Data location exposed to computing framework	Small file save and access	Consumes large amounts of memory of Name Node Seek time exceeds read time
Suitable for large data processing	GB, TB, or even PB-level data Number of files is over millions	Concurrent writes; random file modifications	A file can have only one writer Support only append

2.2 Taobao File Storage System

Taobao File Storage System, a heterogeneous data storage structure researched and developed by Taobao independently, is also known as TFS, which is used to store a large number of picture files; there are two problems needed to be solved in the application: first, a single server cannot meet the huge amount of metadata storage needs; second, a large number of small files cannot be accessed through one-time IO access.

In the organizational structure of TFS system, two domain name servers and multiple data servers are included. Domain name server detects data server running state through heartbeat; if the main domain name server fails, then its data storage services will be adjusted to standby domain name servers, and each data server will run multiple DSP process simultaneously; and each DSP process corresponds to a mount point which corresponds to an independent disk.

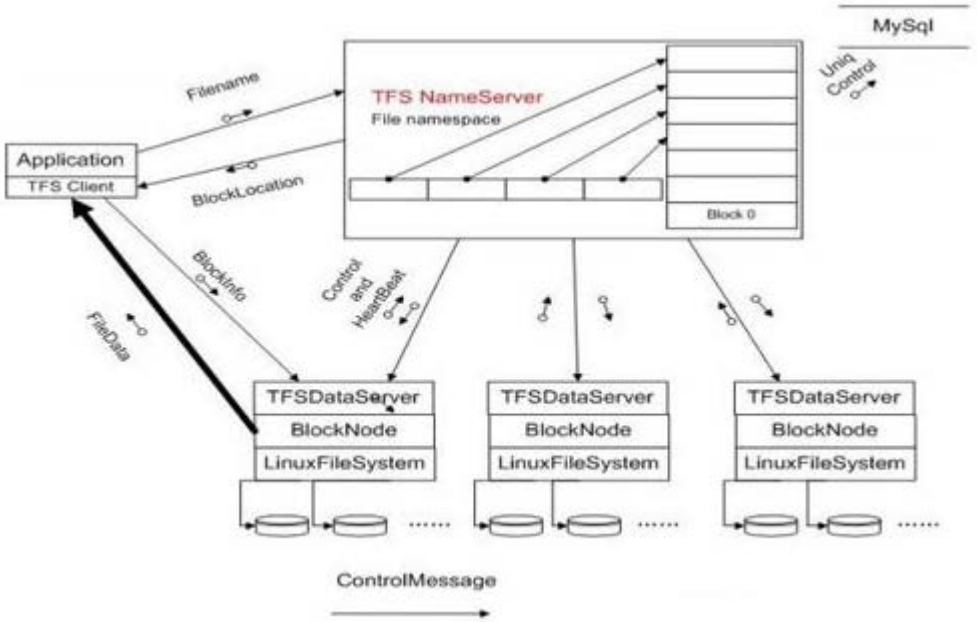

Figure 2 TFS system structure

In TFS system, a large file is composed of a considerable number of small files, which can be called Blocks, and each block has a corresponding number in the cluster that is corresponding to the corresponding file. The Block in TFS, whose actual data is stored in the data server, is typically around 64M, including 3 copies. The client application is the access interface that the system provides to the application, which can only store metadata. State Grid has a large number of small files, such as small pictures. Usually a CIM model is used to represent the system topology. There is a huge amount of SVG small pictures; TFS can properly meet its display needs.

Table 3 Advantages and disadvantages of TFS system

Advantages of TFS system	Disadvantages of TFS system
Storage for a large number of small files	Increase the number of large files, resulting in space waste
A physical file can handle multiple logical files	File storage system lacks flexibility

3. The strategy of distributed file's heterogeneous-configuration electric power data storage

3.1 Large file storage

As mentioned above, HDFS system can meet the compatible demand of low cost equipment, even if hardware appears error, it can still ensure the integrity of the data, and realize the rapid data migration and repair. In HDFS, files are usually in size of GB or TB; HDFS can support the need to write multiple reads at a time, so they can be applied as the best distributed

heterogeneous-configuration electric power data storage system. The electrical power data generated by provincial network companies is around billions of items, the bottom of which can be stored directly in HDFS system, and Hive can be built in the system; create the Hive with capitals on the surface to facilitate data retrieval work, The usual query speed is more than 10 seconds. Hive includes different tables in usage, respectively corresponding to different data storage directory. In actual application, a reasonable table needs to be selected based on the actual needs of data storage.

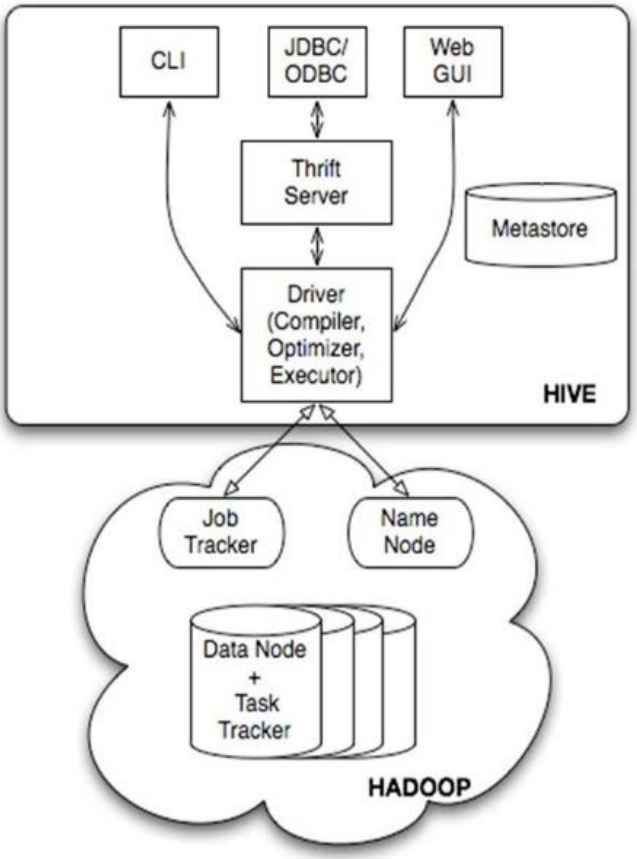

Figure 3 Hive structure

Table 4 different types of Hive tables

Type	Description
Table	Internal Table
Partition	Partition Table
External Table	External Table
Buck Table	Buck Table

3.2 Small file storage

State Grid Heterogeneous-Configuration Electric Power Data contains a large amount of small data, and the storage of which cannot be satisfied by HDFS system, while more options for the storage strategy and storage system are offered. So, in view of the requirements of state grid in terms of heterogeneous-configuration electric power data storage, the following strategies can be implemented:

First, apply TFS to solve small file storage problems. Usually, a small file's size does not exceed 1M; TFS system will not be in Linux device. In TFS, there is no concept of file, which stitches small files into "Block", thereby realizing data mapping. Therefore, the metadata contains a large number of "Blocks", so as to avoid HDFS system's storage problems; second, the HDFS system has some small file storage scheme, namely HAR file, which is stored by packing several small files into a large file, and can be operated by MapReduce Application System. The packaged files include index

portion and the storage section, where the index part contains the source file directory structure and file status. In practical application, HAR file does not have decent pruning effect, therefore is more suitable for regular archiving processing of small files. In HDFS, sequence files can meet the needs of small file storage, which consists of a large number of <key, value>; adjust the key, transfer it into a small file name, and take the value as file content, so as to consolidate small files together into relatively large files.

Since the past contains a large number of heterogeneous-configuration electric power data, the function of the data model is to describe the electrical running state and to describe the topological relationship of the device models. A set of CIM data models are presented with both XML and SVG files; a large number of small files can be used to solve the storage problem in TFS; the filing of electrical power data files is usually realized by the HAR file of HDFS.

4. Conclusion

Driven by the requirements of the development strategy of “Three Intensifications and Five Systems", the construction of state power grid needs to meet the needs of panoramic shooting and real-time monitoring; by then a large number of video and pictures in the form of file data will be produced, which needs more reasonable distributed file storage system to keep up with the demands of heterogeneous data storage. The rational application of distributed storage systems, such as HDFS and TFS, can support the need of massive data storage, data writing and reading, format consistency and data security.

References

[1] Morstyn T, Hredzak B, Agelidis V G. Cooperative Multi-Agent Control of Heterogeneous Storage Devices Distributed in a DC Microgrid [J]. IEEE Transactions on Power Systems, 2016, 31(4): 2974-2986.

[2] Bayati M. Power Management Policy for Heterogeneous Data Center Based on Histogram and Discrete-Time MDP [J]. Electronic Notes in Theoretical Computer Science, 2018, 337:5-22.