Storage Optimization of Condition Monitoring Big Data of Transmission Based on Cloud Platform

Min Ji[1, a], Peng fei J[2, b]

[1] *Xijing University, Shaanxi Xi'an,710123,China*
[2] *China Railway 21, Bureau Group Second Engineering Co. Ltd., Shaanxi, xi 'an 710065 ,China*
[a]*E-mail:995010771@qq.com,* [b]*E-mail:573825625@qq.com*

***Keywords*:** big data; power transmission and transformation equipment; consistency Hash; cloud computing

***Abstract*:** Applying big data technology for improving the condition evaluation of power transmission and transforming equipment and solving its practical problems becomes a new challenge in power industry. For high reliable storage and rapid access of data, the data distribution strategy, data block size adjustment and the cluster network topology are studied based on hadoop. A multi-copy consistency Hash algorithm based on data correlation (CMCH) is proposed. The algorithm makes the relevant data gathering in the cluster and improves the data processing speed. Based on the CMCH algorithm and Map Reduce model, a multiple data sources map join query algorithm and multi-channel data fusion feature extraction algorithm are designed. The two algorithms are executed on our built clusters and the results show that the CMCH improves the efficiency of multiple data sources join query and multi-channel data fusion feature extraction, and the execution time is just 32% and 35% respectively comparing with standard Hadoop.

1. Introduction

With the rapid increase in the scale of power grids and the increasing complexity of power grid structures, power enterprises have increased the promotion and application of the status monitoring technology for power transmission and transformation equipment, and the number of smart power transmission and transformation equipment has continued to grow. The various types of data acquired and transmitted in the device are also experiencing geometric growth. These data not only include various types of signals that appear when the device is abnormal, the status information of various types of equipment that have been operated, but also contain a large amount of relevant data. Such as geographical information, weather, field temperature and humidity, and detection videos, images, and related experimental documents, gradually constitute large data for monitoring the state of power transmission and transformation equipment.

It is an important research topic that how to store the large data of state monitoring of power transmission equipment efficiently, reliably and quickly. Through centralized storage and management of big data for state monitoring, users can not only obtain the history and current status of the monitoring device directly in the data center, but also realize the perception of complex events and laws by analyzing the centrally stored group data. In addition, centralized management of state

DOI: 10.23977/iceccs.2018.006

monitoring data also makes it possible to dynamically evaluate load capacity, dynamically increase capacity, evaluate state of power transmission equipment, forecast failure, and intelligent scheduling based on state of equipment and system risk based on large data analysis.

2. Status monitoring data storage related work

This article based on Hadoop cloud computing platform to monitor the state of large data storage optimization of power transmission devices and based on Map Reduce

The parallel analysis of the processing of the study; Considering the correlation and spatiotemporal attributes of the data, the paper puts forward several pairs which take into account the correlation of the data

This consistent hash storage algorithm optimizes the data partition strategy, data block size adjustment and cluster network topology planning of Hadoop platform. On the basis of optimized storage, the multidata source parallel connection query and multichannel data fusion feature extraction parallel calculation based on the Map Reduce parallel framework are realized.

Data storage by enterprise relational database is widely used in the main station system of state monitoring of power transmission equipment. Since the relational database and the row storage mode used are mainly designed to support data recording and transaction processing(OLTP), the performance of mass data loading and query has dropped significantly. The requirements of quasi-real-time processing applications for large data can not be well monitored for adaptive status. At present, only a small amount of monitoring data is uploaded to the main station system, and a large amount of data that may have significant value is discarded, resulting in a great waste of data.

3. Application of cloud computing technology in power system

At present, the application of cloud computing technology in the domestic power industry is still in the exploration stage. The research content of related literature is mainly in the areas of system architecture design, system model and data processing platform design. However, there is also a small amount of literature to study the use of cloud computing technology to solve specific problems in power systems. Based on Hadoop cloud computing platform, a hierarchical voltage drop parallel computing method is designed to improve computing efficiency by parallel computing of multiple nodes in a cluster. For the massive data set of smart distribution network, the distributed lossless cluster compression of metered data is realized by using the Map/Reduce cloud computing engine.

According to the problem of redundancy and low processing efficiency in mass data processing of wide area measurement system(WAMS), the WAMS data processing platform based on Hadoop is designed and applied to the actual data processing of WAMS for regional power grids. According to the problem of data quantification and high-dimensionalization brought about by the intellectualization of power system, a parallel limit learning machine short-term power load prediction algorithm is designed using the Map Reduse framework, which improves the processing ability of massive high-dimensional data and the precision of power load prediction.

4. Data distribution strategy

In parallel query and processing of distributed data storage system, data distribution strategy is a key problem, which directly affects query efficiency. At present, researchers have done a lot of research on data distribution strategies and proposed many parallel data distribution methods. In load balancing systems and distributed real-time data systems, Hash algorithms are widely used.

5. Data distribution optimization

In order to minimize the comprehensive analysis of data layout factors based on the above, Hadoop's data layout is optimized. A multiple copies of this deterministic hash number based on data correlation is proposed according to the data correlations compared-trust method(CMCH). When the node fails or the node increases, the consistency hash algorithm is proposed and widely used in various distributed systems.

The data block size in HDFS is adjusted to 64 MB, which is much larger than the block size of the physical disk. The design goal of HDFS is to store large files. HDFS file access time mainly includes two parts: addressing time and data transmission time. Access performance is usually calculated using file transfer efficiency. File transfer efficiency effect can be calculated using equation(1).

$$\eta_{effect} = \frac{t_{trans}}{t_{trans} + t_{seek}} = 1 - \frac{t_{seek}}{\frac{s_{block}}{v} + t_{seek}}$$

In the formula: ttrans represents the transmission time and can be calculated using the formula sblock / V; V represents transmission speed; Sblock represents the size of the data block; TSeek represents the address time of the file system. As can be seen from equation(1), ηeffect<1. In the case of data layout and index , the file system addressing time and network transmission speed are usually determined values; Therefore, increasing the file transfer rate should increase the size of the data block. In HDFS, you can set it by setting the dfs.block. size parameter. However, the size of the data block is too large to cause the load balance to decline. This requires that the size of the data block be adjusted based on the data size of the access system, taking into account the transmission rate and load balance factors.

6. Hadoop Cluster Network Topology Programming

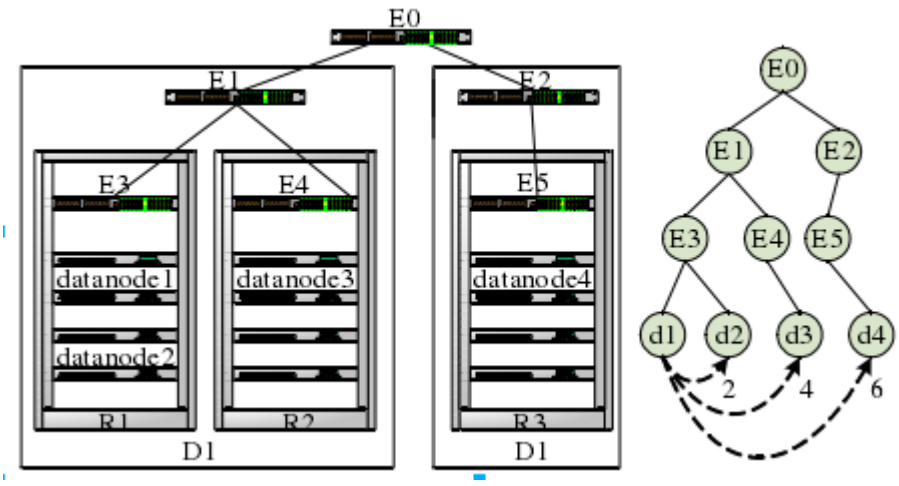

Figure 1. Diagram of Hadoop cluster and its tree structure

When reading data, the Name Node sorts multiple data nodes according to the distance between the data node and the client and returns them to the client for reading data from the nearest node. The network node tree structure in Hadoop. The root node of each subtree in the tree is usually the exchange node connecting the computer. The distance between the two nodes is defined as the number of jumps that one node passes to the other node. In Figure 1, datanode1 and datanode2 are

located in the same rack, connected by the exchange node E3, and the distance between them is 2; Datanode1 and datanode3 are located on adjacent racks in the same computer room. After a level 2 exchange, the distance between them is 4; Datanode1 and datanode4 are located in different computer rooms. After a three-level exchange, the distance between them is 6.

Hadoop's default configuration assumes that all nodes are in one rack, so the network topology of the cluster node needs to be passed to Hadoop according to the configuration of the actual cluster, so that the Hadoop scheduler can select reasonable data nodes for data reading and writing. The network extension structure can be transmitted to Hadoop using foot step code.

As the Hadoop cluster continues to grow in size and the number of copies and the size of data blocks can not be arbitrarily adjusted, the use of data distribution algorithms to aggregate relevant data and improve the local nature of data processing is an effective way to improve the performance of the algorithm.

7. Conclusion

In this paper, the optimization and parallel processing methods of large data storage based on Hadoop platform for power transmission monitoring are studied. The algorithm can make relevant data aggregate in the cluster according to the main attribute, time stamp and correlation coefficient of the device, thus accelerating the data processing speed.

References

[1] Song Yaqi, Zhou Guoliang, Zhu Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology, 2013 , 37(4): 927-935.

[2] Dean J, Ghemawat S. Map Reduce:simplified data processing on large clusters[J] .Communications of the ACM, Fonseca R, et al. Building a cloud for yahoo[J]. IEEE Data Eng. Bull, 2015,32(1): 36-43.

[3] Lars George.HBase: The definitive guide[M].Beijing: O'Reilly Media,2011:324-327.

[4] Mu Lianshun , Cui Lizhong , An Ning. Research andpractice of cloud computing center for power system[J].Power System Technology ,2017,35(6) : 170-175.

[5] Ahmed M U,Mandic D P.Multivariate multiscale entropy analysis[J].IEEE Signal Processing Letters, 2015, 19(2): 91-94.

[6] Cao L,Mees A,Judd K.Dynamics from multivariate timeseries[J].Physica D:Nonlinear Phenomena, 1998, 121(1): 75-88.

[8] Shang Pengju,Xiao Qiangju,Wang Jun.DRAW:a new data-grouping-aware data placement scheme for data intensive applications with interest locality[J]. IEEE Transactions on Magnetics 2015,49(6):2514-2520.