ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 134

2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

Fast query of Big Data Based on rich Network Distribution Monitoring Information Flow

Zhijian Qu, Liang Zhao and Mingming Fan

East China Jiaotong University, Nanchang 330013, Jiangxi Province, China

***Abstract*—In view of the inefficient query efficiency of mass monitoring data in distribution network, a new method for fast querying large data applications with rich network is proposed. Using the MPP query engine, the distribution data is embedded into the network monitoring interface, and the asynchronous query of the cross platform query interface is realized by using Ajax asynchronous interaction. The test results show that the cluster query, MPP combined with asynchronous callback mechanism to interactive server query update to hundreds of MS, cluster CPU use rate of about 11%, a reasonable amount can improve the clustering performance, improve the response ability of mass data, but much less than the expansion to enhance the ability of response to mass data cluster scale.**

Keywords-rich network applications; MPP query

I. INTRODUCTION

Compared with the traditional distribution network, based on intelligent distribution network is a distributed data transmission, calculation and control technology, and a plurality of power unit data and control command transmission processing, especially the substation equipment condition monitoring, electricity information collection, distribution automation projects such as the promotion[1-2], however, affected by the distribution network measuring device synchronization, at the same time may be collected at different times, in order to obtain accurate real-time status information of operation equipment, There are more and more equipment in the station, and the conventional station automation system contains hundreds of thousands of collection points, equipped with electric data center will reach one million or even ten million class, taking into account the needs of intelligent distribution network Panorama data acquisition time, voltage, current and sequence acquisition equipment such as quasi real time information has become the support of the distribution network reliability and intelligent number According to the foundation[3], how to query and deal with the massive quasi real time data of distribution network quickly has become a bottleneck problem of intelligent distribution network decision-making with high efficiency. [4].

A large number of parallel data processing (MPP, massively parallel processing) query engine is completely free, shared data structure is composed of a plurality of single server nodes through network connection, each node only access local resources on its own theory of unlimited expansion capability[5]. MPP can scatter query tasks to multiple service nodes in parallel, complete their queries and summarize the results, thus obtaining the final results. Impala is a real-time interactive SQL Cloudera data by Google Dremel inspired by the development of search engine, and Hive+MapReduce batch query technologies compared to the successful implementation of the column storage of nested data[6-7].

According to the distribution network data query, support Impala cluster data processing platform fast query, we propose a distributed parallel architecture for fast query data rich Internet application method, finally, the experimental platform is built, have been tested and verified

II. MONITORING BIG DATA OF DISTRIBUTION NETWORK

Intelligent distribution network makes the device more high frequency, in equipment condition monitoring system, in order to evaluate the transient power quality diagnosis and status for the ZYNQ high performance control terminal, signal sampling frequency can reach 200kHz, for an intelligent distribution network equipment monitoring platform, each equipped with several sensors connect the sensor monitoring device, through appropriate channels of communication, by the data collection server substation of distributed storage system in accordance with the unified communication standard to upload scheduling in a workstation.

Control unit in accordance with the acquisition timing of monitoring objects, monitoring data collected through the pretreatment of front-end computer, monitoring object PD mapping table collection value summary to the dispatch station, and stored in PF memory mapping the corresponding columns in the table, after Map/Reduce treatment, the equipment information into distributed monitoring data nodes in the data file.

A. Monitoring Data Cache

A relational database is different from the traditional scheduling monitoring system, distributed storage system has powerful data processing capability, consisting of distributed file system HDFS and distributed database HBase, in the face of the dispatching station write once mass monitoring data to multiple observation, analysis and calculation, can guarantee the high fault tolerance. When the dispatcher needs to be monitoring data of PF table storage to HDFS when the request is sent to the leader node, the node will return the monitoring information on each node load capacity to the dispatching station according to the monitoring of file information, to facilitate the monitoring of file is divided into a plurality of data blocks, the address information in order to each node.

III. RICH NETWORK FAST QUERY BIG DATA

A. *Impala Cluster Quick Query*

Dispatch personnel through the network query command sent to a cluster node daemon, the node will work as the query coordinator node, the coordinator node through its own coordinator call planning is parsed SQL query, data communication and telemetry Hive master node on the meta database, to obtain the required data and information after the return, the planner remote execution plan tree, and then interact with the HDFS, the specific location where the file data block node, process as shown in figure I.

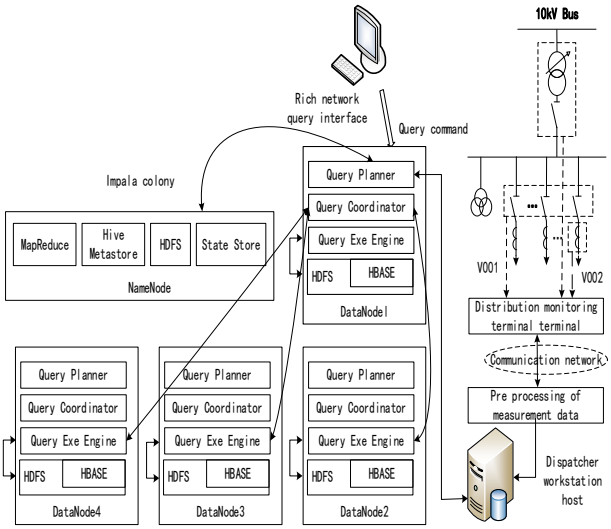

FIGURE I. F BIG DATA CLUSTER QUERY FRAMEWORK FOR POWER DISTRIBUTION MONITORING

To query the equipment monitoring data, monitoring data of dispatching personnel through the network monitoring command is transmitted to a monitoring node daemon, the monitoring node as the node coordinate data monitoring, monitoring for coordinated monitoring and coordination among nodes by JNI (Java Native Interface) to call itself SQL query planning analysis statements, data communication and telemetry data through the Hive database on the master node, then returned to the monitoring coordinator, and the coordinator monitoring according to the execution plan again with the monitoring data of the master node is connected, and interact with HDFS, the specific location of access to monitoring data block where the file work node, in order to facilitate the distribution of monitoring data query plan.

The coordinator in the coordination node is assigned the corresponding sub query in the planning tree to the corresponding work node execution engine to obtain the distribution monitoring data on the HDFS. After the Join connection and Group packet aggregation are performed on the coordinator of each node, the engine is sent to the coordinator of the coordination node, and then returned to the dispatcher after the final summarization and arrangement.

B. *Rich Network Applications*

Monitoring data query engine where the application server can satisfy the query task of the scheduling monitoring system. But at present, the traditional scheduling client server system is running on a variety of server-side scripting engine, the response speed is slow, when facing a large number of queries, prone to flash screen phenomenon, process as shown in figure II.

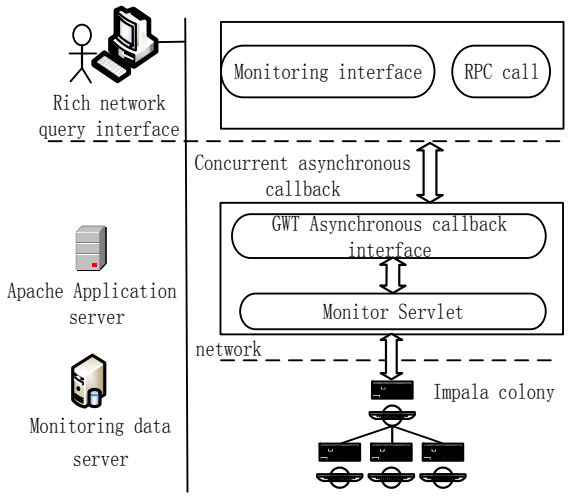

FIGURE II. MULTI LAYER FRAMEWORK FOR RICH NETWORK MONITORING INTERFACE

The application of AJAX GWT will not only object oriented engineering and rich UI interface component library into Java, and can complete the interaction with the telemetry data in the application server in the overloaded interface, greatly reduce the bandwidth. In order to make the rich network query interface and Impala cluster application, the server can carry out telemetry data interactively under different platforms.

JDBC remote access Impala is essentially in the Impala cluster to fast query data and loaded into the compiled script interface program, namely the telemetry data back to the client through the Impala page scheduling for the massive remote sensing data full table scan efficiency, without the use of traditional relational database, can avoid storing small scale and low throughput. And then solve the query speed between the query page of GWT and massive telemetry data. Such as: GWT scheduling monitoring query interface controlled station a state at the time of electrical equipment in the query, can access the Impala cluster through the JDBC cross platform, the dispatch station telemetry data information return and incoming GWT asynchronous interactive object, and then display the distribution of the electrical equipment of a time telemetry data information.

C. *Asynchronous Callback Process*

GWT is able to build apps running in the browser desktop like, through the object-oriented method in the interface and the server are respectively encapsulated Ajax asynchronous request and RPC remote service call and callback services, through asynchronous callback asynchronous scheduling telemetry information interactive processing. The query monitor interface program is compiled by the GWT compiler and converted into a script application by the compiler. The

network is sent back to the query interface of the dispatching station, and the specific query process is shown in figure III.

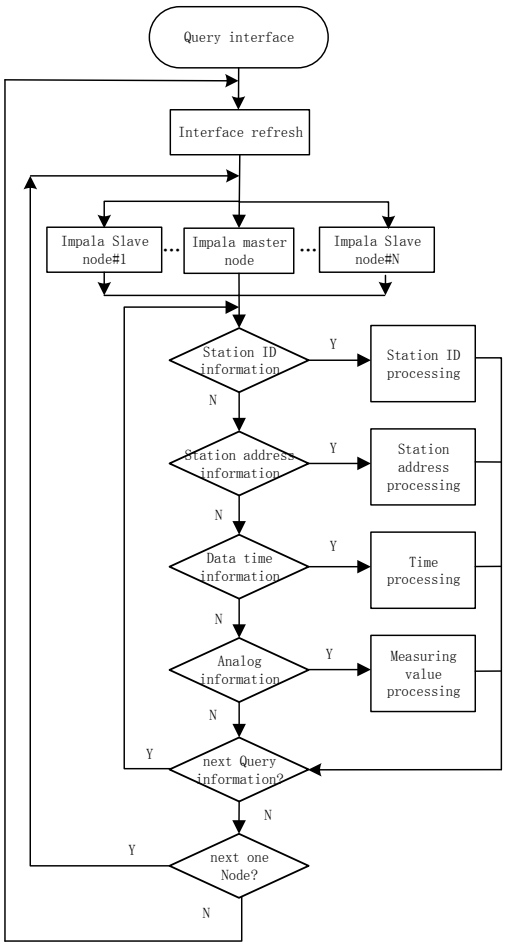

FIGURE III. QUERY INTERFACE QUERY FLOW CHART

The Impala telemetry cluster servers sequentially cycle each request to each node in turn, and each telemetry data from each telemetry node is aggregated to form a complete telemetry data form. In order to quickly load and refresh the telemetry data form, asynchronous query processing of telemetry data is realized by GWT-RPC asynchronous callback.

IV. CLUSTER QUERY TEST

A. *Experimental Environment*

Take railway power 10kV distribution monitoring system of historical database a month telemetry table as test data, the distribution network monitoring master station, on the Centos6.8 operating system, the establishment of four cluster distributed query is composed of main node and the 3 node word, each cluster node and test workstation server configuration as shown in table 1.

TABLE I DISTRIBUTED QUERY CLUSTER DEPLOYMENT

Impala colony	Computer configuration	operating system	scale	Local dependency software
Master node	4 core CPU, Intel, Core, i5-4590, main frequency 3.3GHz, Memory 16GB 8GB	64 bit Centos6.8	1 (platform)	The Hadoop version is 2.6.0, the Hive version is 1.1.0, and the Impala version is 2.5.0
Slave node	4 core CPU, Intel, Core, i5-4590, main frequency 3.3GHz, Memory 8GB	64 bit Centos6.8	3 (platform)	
Query server	Core 2 Duo CPU core 2 Duo CPU, clocked at 2.33GHz, Memory 2GB	32 bit Window XP	1 (platform)	Java1.7 version, Eclipse3.7

Under the station in time, station address, ID, analog as fast query processing in data form in power system, and load test workstation browser, query the monitoring data in the Impala cluster under the influence of different scale, performance and query density cluster on monitoring data.

B. *Query Times Are Clustered By Performance Tests*

Set three different cluster working state {(4,10s); (4,20s); (3,10s)} 4 clusters, that is, 4 complete clusters, each 10s query / time, 20s query / time, closes one node per 10s query / time.. Interval 1min cycle sampling mode. Continuous sampling of 10min statistics Impala cluster query at the end of the cluster CPU utilization, the results shown in figure IV.

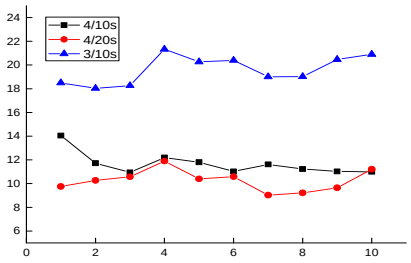

FIGURE IV. IMPACT OF CPU USAGE ON DIFFERENT CLUSTER WORK STATES

As shown in Figure 4, the memory performance and the working density of a single node have little influence on the cluster query time, and the size of cluster affects the query time significantly.Effect of delayed response to the query interface of smaller cluster size and performance, interactive query interface can be updated to hundreds of MS, in the face of a large number of information monitoring data can satisfy the query in the query, the number of certain conditions, a reasonable increase in cluster performance, to improve the response ability is not obvious telemetry data, far less than the expansion ability in response to the cluster size for large data.

V. SUMMARY

The integrated use of Impala cluster components and rich Internet application skills, fast query of mass monitoring data, and return it to the rich network interface end, through the test, the results show that the face of massive data can satisfy the query in the query, the number of certain conditions, a reasonable increase in cluster performance, big data to improve the response ability is not obvious. The cluster size is

far lower than the ability to respond to large data upgrade, this method can meet the requirements of monitoring system for large data processing interface end.

ACKNOWLEDGMENTS

Project Supported by the National Natural Science Foundation of China (51567008); Foundation Plan for Distinguished Young Scholars in Jiangxi Province (20162BCB23045); Natural Science Foundation of Jiangxi Province (20161BAB206156); Science and Technology Research Project of Jiangxi Provincial Education Department (GJJ160471).

REFERENCES

[1] ZHAO Lin, WANG Lili, LIU Yan, SUN Pai, ZHANG Liang.Research and Analysis on Visualization Technology for Power Grid Real-Time Monitoring[J]. Power System Technology,2014,38(2):538-543(in Chinese) JI Yang, AI Qian, XIE Da. Research on Co-Developmental Trend of Distributed Generation and Smart Grid[J]. Power System Technology, 2010(12): 15-23 (in Chinese).

[2] JI Yang, AI Qian, XIE Da. Research on Co-Developmental Trend of Distributed Generation and Smart Grid[J]. Power System Technology, 2010(12): 15-23 (in Chinese).

[3] SONG Yaqi1, ZHOU Guoliang1, ZHU Yongli2.Present Status and Challenges of Big Data Processing in Smart Grid[J]. Power System Technology, 2013, 37(4):927-935 (in Chinese).

[4] ZHANG weichao, CHEN Jiandian, JIANG Fangyu.Application of quasi real time data platform in distribution network operation management[J]. Science and technology innovation and Application, 2014(28):193-193 (in Chinese).

[5] Qu Zhijian, Liu Mingguang, Liu Jing, Wang Jian, Yang Gang, Wang Lin. A method for batch processing of distribution network monitoring information Based on belief-desire-intention agents[J]. Power System Technology, 2012,36(3): 2-8 (in Chinese).

[6] CHEN Shu-yong, SONG Shu-fang, LI Lan-x, SHEN Jie[J]. Power System Technology, 2009, 33(8):1-7 (in Chinese).

[7] ZHOU Xiaoxin1, LU Zongxiang2, LIU Yingmei1, CHEN Shuyong1. Development Models and Key Technologies of Future Grid in China[J]. Proceedings of the CSEE, 2014, 34(29):4999-5008 (in Chinese).