Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

The WAMS Power Data Processing based on Hadoop

Zhaoyang Qu [1], Shilin Zhang [2] [+]

[1] School of Information Engineering, Northeast Dianli University, Jilin Jilin 132012, China

[2] School of Information Engineering, Northeast Dianli University, Jilin Jilin 132012, China

Abstract. For the mass data efficient processing power, Cloud computing platforms started to popularize in the world scope, which is mainly used to mass data processing and analysis, and it's better to save and use hardware resources. For massive WAMS data, this paper used the MapReduce to make parallel data ETL operations for several files, used MapReduce to to improve Apriori algorithm for improve the efficiency of data mining, and proposed the model of data mining of text log file based on Hadoop. According to this model and created the platform for mining of cascading failure power site based on Hadoop, Which digged out the relationship of power site when cascading failures occurred, and verify the efficiency of data mining on Hadoop. This platform is suitable for mass power grid files data mining by high performance local area network connection of computer cluster.

Keywords: Cloud computing; data mining; WAMS data; MapReduce; ETL

1. Introduction

WAMS (wide area measurement system) real-time dynamic power grid monitoring system[1], which in order to high-speedly and real-time achieve acquisit full power grid synchronization phase angle and each power site data, as an power grid dynamic monitoring platform, it is an important part of smart power's real-time monitoring platform. But in the current information system fault diagnosis and fault studies, WAMS-based platform [2] problems are mainly:

(1) the data redundancy, this redundancy exist inside of measurements unit, between different measurement devices and between adjacent sub-stations; (2) when the data acquisition , for the data lack of data process , data classification by application, and without the classification transport by application characteristic, under the situation of power grid size increases and the failure affected range growing, which lead to the information acquisition device provide increased exponentially data and upload a lot of useless data. (3) WAMS data processing and analysis platform is still using conventional methods of data storage and management, whose infrastructure are using expensive large-scale server, storage hardware using disk arrays, so the system scalability is poor and the cost is higher. According to above problems, the mass WAMS data mining algorithms can't high-efficiency run, so it's time to offer efficient data processing methods.

This paper research on the massive WAMS log dada processing based on the platform of Hadoop[3][4]. Loading the vast historical WAMS platform real-time data to Hadoop platform and complating the data distributed stored and processing, makingthe data classification and beckup.

2. Platform Structure

[+] Corresponding author. Tel.: + 8615944212906;
E-mail address: zhangshilin2008@163.com.

Hadoop is a distributed system architecture, users can develop distributed programs in order to use the cluster's high-speed, effective data processing ability and storage function, without knowing the underlying detail of this distributed structure. Combined with the basic methods of data processing and hierarchical thinking, this paper use the Hadoop's data storage ability and data processing function in the data process platform without the assistance of database. The system organization chart is as follows:

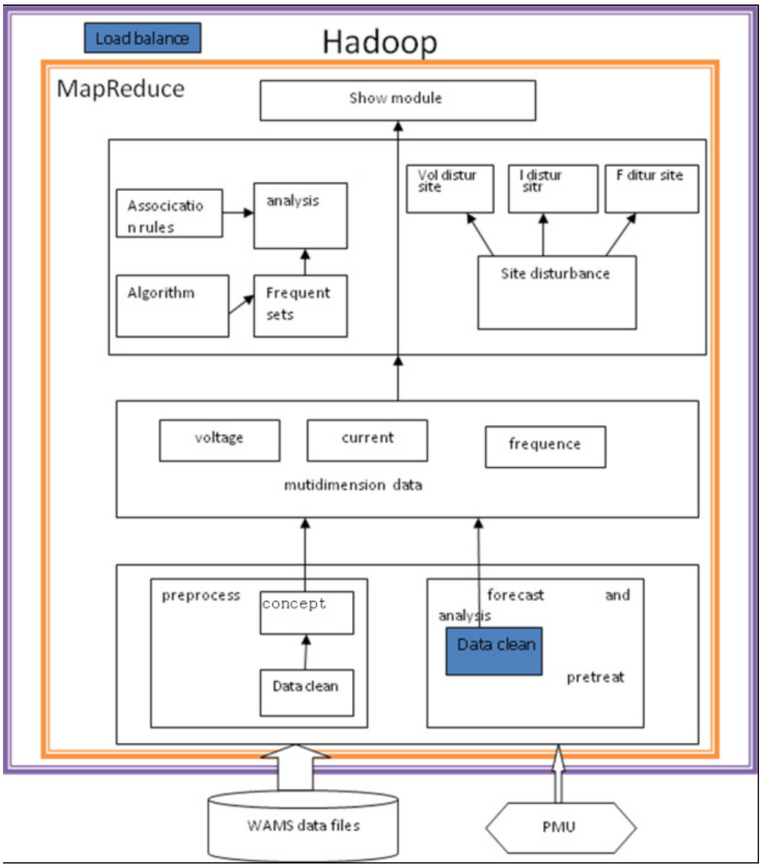


Fig.1 system organization chart

3. Platform Implementation

According to the functional structure of the mass log file data processing based on Hadoop and combined with the actual situation of WAMS data, the article would create the WAMS network data processing platform of Hadoop, excavate the main effected site at some grid sites' voltage mutation, verify the correlation, chain, the relevance, interaction among fault sites and determine the causality of fault sites. To realize the data loading module, the data ETL module, data mining algorithm module as well as result display module in the clouding computing environment is the main research achievement of this article.

3.1. Data Loading Module

Data are mainly from a network WAMS platform and China power grid frequency is 50HZ, that is to say, our network data is every 0.02s time the acquisition which records at each site every minute of it monitoring to the various parameters of the log journal. For example, the grid network log journal obtained from one site under WAMS grid data platform is made up by different aspects of contents such as file name :site name_time.log and a content recorded in the file :2005/03/29_09:30:450001112059845 0 542.05 84.79.The size of each file is 344K.

You should respectively arrange HDFS, MapReduce on each server if you need build Hadoop cloud computing platform on ubuntu9.10. You only need to run the load command such as Hadoop dfs-put when talking about inserting 2TB historical real-time log files into HDFS of Hadoop platform.2TB file can be completely composed by PutMerge method.

3.2. System Data ETL Module

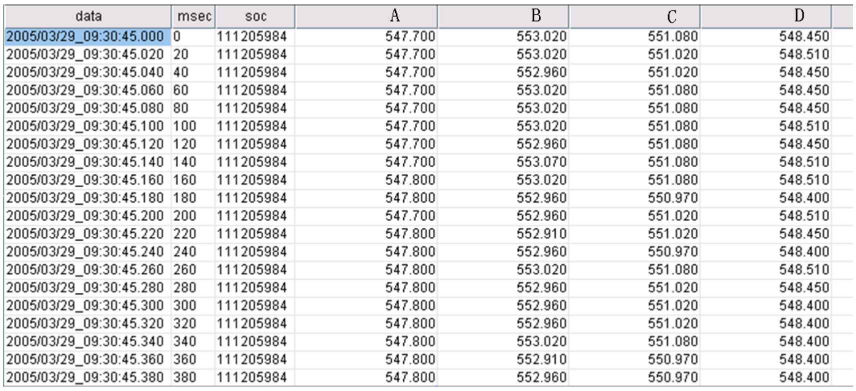
The main method for parallel ETL by MapReduce as follow:

1) Data in the log files does not involve access to the database, during the process of data load which equivalent to the data files read into the system by MapReduce.

2) Data convert and cleaning is operate and access every data, in order to remove, repair inconsistent data and dirty data in the data source, and complete the change of data type and data size. Pseudo code is as follows:

```
Map(String key, String value)
//key:the name of log file name
//value: every data in log file
For each data d in value:
    DataETL(d);
Reduce（String key, Iterator values）:
//key: a data
//value: the name of log file
For each v in value:
    Fputc(key, v)
Emit (AsFile (v));
```

In the absence of large-scale parallel databases, the MapReduce implementation of data ETL can improve the speed of parallel access to data, and reduce the system's operating costs and maintenance costs for large databases. For example, voltage results map for multiple sites data ETL as follow:



data	msec	soc	A	B	C	D
2005/03/29_09:30:45.000	0	111205984	547.700	553.020	551.080	548.450
2005/03/29_09:30:45.020	20	111205984	547.700	553.020	551.020	548.510
2005/03/29_09:30:45.040	40	111205984	547.700	552.960	551.020	548.450
2005/03/29_09:30:45.060	60	111205984	547.700	553.020	551.080	548.450
2005/03/29_09:30:45.080	80	111205984	547.700	553.020	551.080	548.450
2005/03/29_09:30:45.100	100	111205984	547.700	553.020	551.080	548.510
2005/03/29_09:30:45.120	120	111205984	547.700	552.960	551.080	548.450
2005/03/29_09:30:45.140	140	111205984	547.700	553.070	551.080	548.510
2005/03/29_09:30:45.160	160	111205984	547.800	553.020	551.080	548.510
2005/03/29_09:30:45.180	180	111205984	547.800	552.960	550.970	548.400
2005/03/29_09:30:45.200	200	111205984	547.700	552.960	551.020	548.510
2005/03/29_09:30:45.220	220	111205984	547.800	552.910	551.020	548.450
2005/03/29_09:30:45.240	240	111205984	547.800	552.960	550.970	548.400
2005/03/29_09:30:45.260	260	111205984	547.800	553.020	551.080	548.510
2005/03/29_09:30:45.280	280	111205984	547.800	552.960	551.020	548.450
2005/03/29_09:30:45.300	300	111205984	547.800	552.960	551.020	548.400
2005/03/29_09:30:45.320	320	111205984	547.800	552.960	551.020	548.400
2005/03/29_09:30:45.340	340	111205984	547.800	553.020	551.080	548.400
2005/03/29_09:30:45.360	360	111205984	547.800	552.910	550.970	548.400
2005/03/29_09:30:45.380	380	111205984	547.800	552.960	550.970	548.400

Fig.2 Voltage results map

Data ETL are data preprocessing, in order to improve efficiency of executing Apriori association rules algorithm by MapReduce in this experiment, the author also made the following data processing.

(1) The deal of vacancy value: using the value of the adjacent time data to fill or changing by the average of adjacent time periods data.

(2) The system is mainly found when a voltage disturbance, the site's change result in other sites' change, when the voltage of the intermediate data file or data processing, the key is to determine the data has been changed or not. Therefore, the initial data can be set to 0 and when the data changes set the site data as 1.

3.3. Data mining Algorithm Module

Apriori algorithm Improve by MapReduce algorithm[5]can overcome the bottleneck of scanning data source database frequently. That make to find frequent item sets parallel execution, when find frequent k-item sets' middle of the results and sent to the Reduce function, at the same time K +1- itemsets map tasks can be carried out, which make parallel execution of data operations, and improve the operation of the system efficiency. MapReduce and the framework of Apriori[6] algorithm with the following diagram:

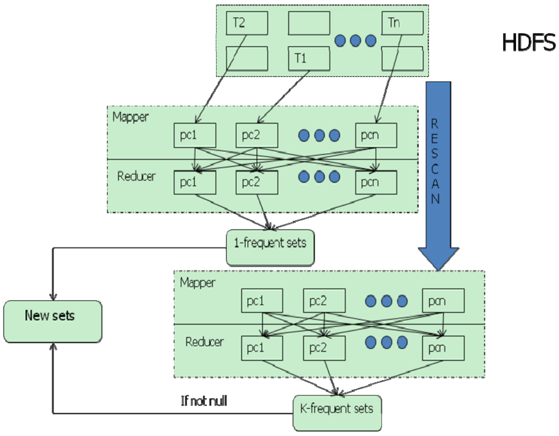


Fig.3 Voltage results map

Since the data processing phase system has got a simple 0, 1 data files, it's only need to use MapReduce to achieve the basic Apriori algorithm which can find frequent item sets, and get the appropriate disturbing sites and the disturbing effect sites.

4. Experiments and Results

This platform develops with 6 Dell's PowerEdge servers. Take about 20 days of a regional power system of WAMS data, the size about 1.5TB, to these historical log data for data processing and analysis of cascading failures. This platform was developed in Java on the Eclipse development environment based on component model, From A to F, this 7 sites Voltage related as Tab:1

Results clearly seen through above table, B and A have a better relation, but B with C or E has little relation, this result is meet with the real environment.

Excluded from the network element of chance, we conclude that: Hadoop cloud computing platform for data mining grid massive data processing has a better efficient than traditional data mining platform, but the efficiency of their data processing needs based on data mining algorithms, data mining the complexity of the physical cluster resources to deal with files and other specific factors.

Tab.1 7 sites Voltage related

Consequent	Antecedent	Support%	confidence
A	B	38.687	49.627
B	A	39.52	48.229
B	E	37.021	42.844
A	E	37.021	42.664
D	E	37.021	41.404
D	B	38.678	41.258
E	B	38.678	40.999
D	A	39.52	40.472
E	A	39.52	39.996
A	D	40.253	39.735
B	D	40.253	39.652
E	D	40.253	38.079

5. Summaries

The massive log file data process based on Hadoop platform is using the function of Hadoop's process the massive log file data mining, refer data ETL as the MapReduce's parallel I/O to files, using MapReduce improve the data mining algorithms and enhance the data process algorithms' parallel process data ability. Using this method on power grid system WAMS platform data mining, achieve Hadoop for processing massive power WAMS data. This experiment confirmed: Cloud computing can be very good to improve the efficiency of WAMS data processing for the future to provide further data mining and data mining-based platform basic framework.

6. Acknowledgements

This paper was supported by the National Natural Science Foundation of China (No.51077010), Province Natural Science Foundation of Jilin (No.20101517).

7. References

[1] ZHOU Ziguan, BAI Xiaomin, LI Wenfeng, et al. A novel smart on-line fault diagnosis and analysis approach of power gridbased on WAMS [J]. Proceedings of the CSEE, 2009，29(13): 1-7.

[2] WANG Xiaobo, FAN Jiyuan. Construction of common data platform in the power dispatcher center. Automation of ElectricPower Systems, 2006, 30(22): 89-92.

[3] CHEN Kang, ZHENG Wei. Cloud computing: system instances and current research. Journal of Software, 2009, 20(5): 1337-1348.

[4] Apache. Welcome toApacheHadoop [DB/OL]. http: //hadoop. apache. org, 2010-05-12.

[5] Shafer J,Agrawal R,Mehta M. A scalable parallel classifier for data mining[A] .Sprint Proc of the 22nd IntConf on Very Large Databases[C]. Mumbai (Bombay)India: SL IQ, 1996, :544-555 .

[6] Jeffrey Dean,Sanjay Ghemawat. MapReduce:Symplified Date Processing on Large Clusters[J] .2008,51, 51 (1) :107~113 .