Power Big Data platform Based on Hadoop Technology

Jilin Chen[1, a,] Nana Liu[2, b], Yong Chen[2, c] and Weijiang Qiu[2, d]

[1]China Electric Power Research Institute, Beijing 100192, China;

[2] China Electric Power Research Institute, Beijing 100192, China.

[a]chenjilin@epri.sgcc.com.cn, [b]liunana@ epri.sgcc.com.cn

Keywords: Power big data, Hadoop, distributed storage.

Abstract. The utility industry has entered into big data era as the construction and development of smart grid. It becomes more critical to store and process the big data efficiently, and to make effective utilization when the power big data are massive complicate. This paper analyses the status of power big data. Then, it introduces the main architecture of power big data platform based on Hadoop technology. Several key technologies are analyzed including data storing and processing. Finally, discusses the application of power big data platform based on Hadoop technology

Introduction

The word is experiencing a data revolution. The total data people produced since human civilization to 2003 is five EB, but this is just the amount of data people produced in two year now, According to Internet Data Center in 2011, global data volume has reached 1.8 ZB, equivalent to more than 200 GB of data per person all over the world, The amount of available digital data at the global level grew from 150 EB in 2005 to 1200 EB in 2010. It is projected to increase by 40% annually in the next few years.

Big Data refers to datasets whose size is beyond the ability of typical software tools to capture, store, manage, and analyze. Since 2009, the "big data" has become the Internet IT industry buzzword. The global IT companies are aware of the coming of "Big data" era as data generates great value and becomes important assets for enterprises.

With the large-scale construction of smart grid and with the large-scale construction and development of smart grid, smart substation, electric vehicle charging station project put into operation as well as wind, photovoltaic, energy access, the power industry in the information age is in a crucial turning point, grid data will be more and more complex. The grid data size will be exponential growth every year, and the data contains a large number of semi-structured and unstructured information, at the same time, the smart grid requires the fast response to the grid fault, short-term load forecasting and real-time data processing, these key questions are very difficult to use traditional technology. On the one hand, the large scale of data leads to the original system is difficult to store and manage; on the other hand, the large data is responsible for the relevance of the traditional algorithm is failure. Therefore, it is necessary to accelerate the application of big data in the field of electric power. In this paper, with the new technology and the practical application of large data, the power industry is required to carry out a systematic study of the large data technology system, power big data technology structure, and the big data storage and large data processing and other core technologies are analyzed in detail.

The power big data and Hadoop technology

The power big data. According to McKinsey's description, the big data is not in a certain period of time with the traditional database software tools to capture, manage and deal with the data collection. The significant features of large data are data size (Volume), data type (Variety), processing speed (Velocity), low value density (Value), or 4V.

Electric power big data generally refers to the collection of information acquisition channels through the sensor, intelligent equipment, video surveillance equipment, audio communication equipment, mobile terminal and other information acquisition channels, which is massive, structured, semi-structured, unstructured, and related business data.

The power big data has 4V features: (1) Large data size. The data acquisition of power data includes transmission transformation, distribution, and sale parts. Only the sale the electricity side of collected data, there are 30 million users in one provincial network. Every collection of time, each year of electricity data has more than 100 billion. If the interval time is 15 minute, the amount of data will increase 96 times. (2) The types of data. The power data which is divided by structured includes the conventional structured data and system logs, table data, and other structured data, as well as documents, pictures, video and other unstructured data. From the data content, the power data includes measurement data, monitoring data, equipment ledger, log files, geographic information, meteorological data, etc... (3) The processing speed. The power system has strict time limit to the data acquisition, processing and analysis. Many services such as real-time query, rapid response to the network fault and so on. (4) Low value. The data in the power grid is generated every time, but the real useful data is not much. Such as equipment status monitoring data, the vast majority of normal data, while the abnormal data is the key to the status of equipment.

The Hadoop technology. The core of Hadoop is MapReduce, Google File System(GFS) and BigTable which are published by Google in 2003 to 2004. The MapReduce is a distributed computing framework. UFS is a distributed file system, and BigTable is a data storage system based on UFS. The three components constitute a new distributed computing model. Hadoop is built by Yahoo on those open sources and HDFS is built according the development of UFS, HBase is built according the BigTable. At the same time, many other open source projects such as Pig Hive and around Hadoop constitute a Hadoop ecosystem, as shown in Figure 1.

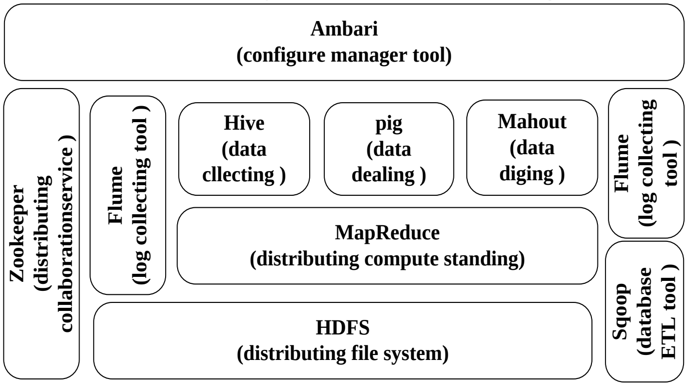


Figure 1：Hadoop ecosystem

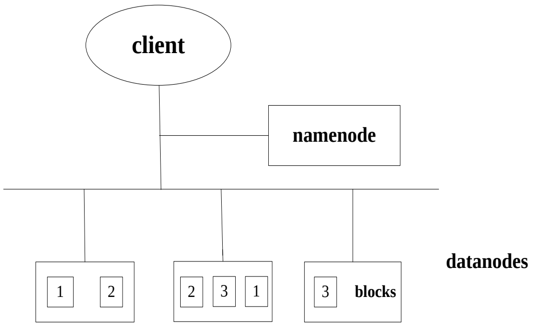


Figure 2：HDFS Architecture

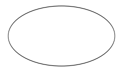
Hadoop has the advantages of high scalability and high fault tolerance. The high efficiency processing of massive heterogeneous data is realized by MapReduce and HDFS based on the distributed idea.

HDFS is a distributed file system, is a master / slave structure, a HDFS cluster is constituted by a name (namenode) nodes and data nodes (datanode), the name of the node is a management file namespace and adjusting client access to the file's main server and data nodes is often a machine and corresponding management node storage. HDFS develops the file namespace and allows the user to store the data in a file. Each file stored in the HDFS will be divided into one or more data blocks (block), each with multiple copies of each data block, each copy is stored in different data nodes, data blocks have multiple redundancy to address the problem of data loss caused by hardware.

MapReduce is a distributed computing software framework for parallel computing of large scale data sets. Through the high concurrent processing, and simultaneously manage multiple large scale computing process, the data processing ability is broken from the TB level to the PB level. The simple MapdReduce mainly consists of three parts: Map function, the main controller and reduce function. MapReduce uses parallel processing strategy for large scale data, and the large number of repeated data record processing procedure is summarized into two abstract operations of Map and Reduce, and provides a unified parallel computing framework. The implementation process is shown in figure 3. the large scale parallel computing is cost much more money in the past. With the application of the distributed file system and parallel computing, these calculations can be done by thousands of common computer clusters, and the cost is greatly reduced.

The architecture of the power big data platform

Big data is not a technology, but the integration of a variety of technologies. The power big data platform based on Hadoop technology includes data source, data integration, data storage, data processing, data show and security management, and other key technologies. Hadoop as the core of the open source products is the mainstream of large data open source solutions. The structural framework of power data platform based on Hadoop is shown in Figure 4.



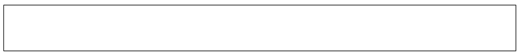


Figure 3: The process of MapReduce Figure 4: The architecture of power big data platform

Hadoop provides distributed file system and parallel computing is to solve large scale data storage and processing problems. On the basis of building up the upper application to achieve SQL, real-time computing, flow calculation, memory computing, data mining, data visualization and other functions.

The data source of power data. The data acquisition of power data includes transmission transformation, distribution, and sale parts, which can be divided into three categories according to the business: power grid operation, corporate operations and marketing services. Power grid operation data are mainly divided into the power grid operation and equipment state monitoring data:

(1)The power grid data is mainly from EMS (Energy Management System) in the D5000 Dispatching platform, Operational Management System, and SCADA (Supervisory Control and Data Acquisition). The data includes the line voltage current, power, and relay protection device information, protection of fault recording data, switch state, alarm information, new energy distribution, and reactive power compensation and so on. It will get a lot more accurate phase data with the vector measuring device deployment.

(2)The monitoring condition data source of equipment is from SCADA and Production Management System (PMS). The data mainly includes the equipment ledger, work ticket, defective family, poor working conditions, maintenance test, electrified test, on-line monitoring and other so on. At present, the value of these data has not been excavated; it needs a deep analysis to realize the comprehensive evaluation of the equipment status.

The data of marketing service mainly comes from the marketing business system, the information collection system, the measurement and management platform, the EMS, the 95598 platform, etc... It includes marketing equipment management data, user files and the use of electric information, the electric energy metering, network for load, power generation and 95598 businesses data and other data. These data can be used for load forecasting, power consumption forecasting, load characteristic analysis and economic situation analysis and so on.

The integration technology of big data. The big data integration technology needs to integrate traditional ETL technology and data connector, real-time message queue, platform service interface and other new technologies, the vast amounts of diverse data which are from data center; business platform, terminal and other external data source are imported into the big data storage system in accordance with the unified data specification standard processing.

(1) The big data connector is a kind of relational data collection technology, which is used to transfer data between the traditional data source and the distributed storage system.

(2) The service line is a kind of data acquisition and processing technology, which can be used to gather and store large data from normal traditional data source, structure data which cannot be effectively processed. The related products have Scribe, F1ume, etc...

The real-time message queue is a real-time data acquisition technology. Due to the large scale and fast changing data generated by sensors, the data acquisition and processing is needed. The distributed mass data acquisition technology is needed for the collection of real-time data. The related products have Kafka, etc...

The storage technology of big data. The storage technology of big data to meet the needs of all types of data storage and computing needs diversification. It realizes the data storage, low cost and NoSQL data access on the base of low storage devices and the distributed file system and a comprehensive distributed file system based on all kinds of database.

(1) The distributed file system. The file data is stored in the storage medium the dispersion of the low cost, providing consistent file access interface and fault tolerance and good security, for semi - structured, PB more than the size of the non structured data storage. The main products are HDFS, FastDFS, etc...

(2) The column store database. The data are storage by unit so that the data can be efficiently compressed size, providing fast retrieval of massive scale data and search function. It is for large quantities of data processing and real-time query. The main products are such as HBase.

(3) The distributed relational database. A large database, which is composed of distributed multiple nodes, is used to store and query large scale structured data. The main products are GeenPlun, etc.

(4) The key database. The non relational database model is stored by the key value, which is better read and write for high performance and semi structured data query. The main products are Redis, etc...

(5) The real time database. The database model, which is used to process the database model with time series characteristics, is used for the storage and query of real-time or quasi real time high frequency data acquisition. The main products are RealTimeBase and so on.

(6) The memory database. The structured data are put in the memory of the direct operation, which is for reading and writing fast, real-time querying and analyzing for high performance. The mainstream products are TimesTen, etc...

The big data processing technology. The processing technology of big data needs to meet the needs of mass data processing, and the core of the distributed computing, and other advanced computing models, which can be adapted to the computing framework of multiple computing scenarios.

(1) The distributed computing is for massive scale data, using MapReduce distributed computing framework, data processing capabilities to achieve the data breakthrough from TB level to PB level, for real-time requirements are not high for large quantities.

(2) The flow computing is a high real-time computing model for streaming data. the steady stream of massive audio and video streaming data generated by the flow computing are not long-term storage, directly put into the memory for real-time computing and extract valuable information. This is for real time computing for dynamic flow data. Mainstream products are SparkStreaming Storm, etc..

(3) The data storage and calculation exist in the main memory in memory computing, which is use of CPU and memory advantages, combined with parallel computing technology to achieve high performance computing. The memory computing is for real time statistics and interactive analysis. The main products are HANA SAP, etc..

The big data visualization technology. Most of the electricity data are less useful in directly .The high value information is extracted from the analyzing with appropriate mining algorithm. Traditional mining algorithms are: clustering analysis, correlation analysis, evolution analysis, text analysis, image and video analysis, etc... These algorithms have some limitations in the distributed data and distributed processing. At present, the data mining technology based on open source technology, such as R language, Mahout, can support data analysis and mining of large data, and its comprehensive analysis algorithms, development tools and visual controls. Large data visualization technology will display the large-scale, multi dimension, complex data results to the user in the form of intuitive graphical display, to help users quickly understand and make accurate judgments. Typical large data visualization techniques are: network diagram, sun chart, tree, chord diagrams, parallel coordinate chart, index map, table calendar, tag clouds, packing ring. The power big data has the characteristics of multi-source, heterogeneous, distributed widely, dynamic growth, and cross business, which is different from the traditional data management; the power big data are in a greatly increased security risk. Based on these features, The more care are focus on strengthening the authority management, privacy protection, storage security, access security and other security technology and the corresponding security management system, in order to achieve large data acquisition and application of the whole process of security monitoring.

The application of power big data

The power big data platform based on Hadoop technology has a wide range in the field of power, this paper analyzes several typical application with high development value.

（1）The new energy generation forecasting and management

The accurate forecasting of new energy generation capacity and optimization new energy dispatch management are based on the simulation of massive meteorological data, combined with the wind turbine and photovoltaic power generation output curve and historical power generation information, the relationship between the new energy output and wind speed, light, temperature and other meteorological factors is analyzed.

（2）The optimizations of strategy in fault equipment

The power transmission and transformation equipment fault are automatic recognized by real-time online monitoring data analysis. Through the associated equipment account running state records and other data analysis, it can find the cause of a failure. Then the comprehensive evaluation model is set up to realize the status of the equipment and the risk assessment and fault prediction.

（3）The forecasting of medium term load

The analysis of power transmission and transformation equipment load and prediction of long-term electricity demand distribution and change trend by the massive power data, GIS data, population information, regional planning and economic situation are guide for company's expansion of equipment replacement, distribution network upgrade transformation and transmission line planning.

（4）The real time data research and the analysis of electric energy user

The real time data acquisition and high speed processing capability are analyzed by using data mining and data analysis. The high frequency of the data is stored in high frequency, and the power consumption is improved. Based on the data of the user's history, analysis of the characteristics of the community or the large customer with the electric behavior, it can predict the short-term electricity demand of customers and realize the power management of the individual.

Summary

With the development of technology, the future of power grid will be more intelligent, safe, reliable, low cost, high efficiency, high reliability and large data technology will provide a solid technical support. In order to improve the security of power grid, to provide customers with high quality service, and actively promote the application of big data technology in the field of power is imperative.

The power data platform based on Hadoop meets the processing of massive data in the new situation, data analysis; the use of Hadoop open source framework has high scalability and high fault tolerance, providing a reliable platform for power data analysis.

References

[1] McKinley Global Institute. Big data; the next frontier for innovation, competition, and productivity. 2011.

Reference to a book:

[2] China Institute of Electrical Engineering Information Committee. Chinese power big data development since the book, 2013

[3] CHEN Ji-rong, YUE Jia-jin. Reviewing the big data solution based on Hadoop ecosystem. Computer Engineermg & Science,2013,35(10).

[4] Mladen Kezunovic, Le Xie, Santiago Urijalva. The Role of Big Data in Improving Power System Operation and Protection — 2013 IREP Symposium-Bulk Power Sys-tem Dynamics and Control-IX(IREP)，2013.

[5] Zhao Gang. The Guide for big data technology and Application.Beijing Electronic Industry Press 2013.

[6] Kanta M G Chi, Wang Xiaohai, Wu Zhigang,The data Mining: concept, model, method and algorithm.Beijing Tsinghua University Press,2013.

[7] Wang Xiulei，Liu Peng • Key big-data technologies[J].ZTE Technology Journal 2013，19(4) : 17-21

[8] Wigan M R，Clarke R. Big data’s big unintended consequences [J] . IEEE2013, 46(6): 46-53.

[9] Gong Xueqing，Jin Cheqing，Wang Xiaoling，etal. Data-intensive science and engineering: requirements and challenges[J]. Chinese Journal of Computers 2012 35(8): 1563-1578.

[10] Meng Xiaofeng，Ci Xiang • Big data management: concepts，techniques and challenges[J].Journal of Computer Research and Development,2013,50(1):146-169.

[11] Wu Xuzhong, How to use the big data in Electric industry, People’s Posts and Telecommucations,2013. 7. 1(3)

[12] YANG Fang,WANG Wendi, GE Xubo, et al Comprehensive assessment on development patterns of smart grid in china Electric Power, 2012,45(12):81-85

[13] Yu Lijun. Analysis on Intensive Management in Power Grid Enterprise Based in Theory of Company Value Chain. Energy Technology and Economics,2011,23(5):57-62