Huge Data Analysis and Processing Platform based on Hadoop

Yuanbin LI[1, a], Rong CHEN[2]

[1]Information Engineering College ,Sichuan Agricultural University, Yaan 625014,China

[2]Guangdong Songshan Polytechnic College, Shaoguan 512126,China

[a]liyuanbin@126.com

Keywords: Huge data analysis, Hadoop, HDFS, MapReduce

Abstract. SGC puts forward the major initiatives of the construction of strong smart grid and three integrates and five big systems, requiring the informationlization further improves to a level on the basis of the inheritance for fruit and deepening application, having even higher requirements especially for huge information data processing, data search grid. In response to meet the needs of the smart grid to the huge amounts of data processing, this article puts forward building dynamic scalable huge amounts of data processing platform on the basis of Hadoop in the virtualization resources management platform through in-depth analysis and research of the two core key technology, HDFS and MapReduce and gives its technical architecture, implementation plan and case analysis. Compared with traditional Hadoop distributed parallel computing system based on the physical machine deployment, to build Hadoop virtual server templates through virtualization platform not only can quickly complete the Hadoop deployment of distributed parallel computing system and can effectively make use of computing resources.

1. The brief introduction of Hadoop

With the rapid development of information technology, data is also in explosive growth, a lot of information began to show in the form of a semi-structured or unstructured text, but how to manage apply the large amounts of data stored in text mode has become a big problem in the field of data, also needs to be addressed. Then, the Hadoop platform technology becomes widely used distributed in text processing. The platform technology contains distributed file storage system (HDFS) as well as the calculation model (MapReduce). As an open-source platform for the large data processing, it has been widely applied in range of research and industry and academia. Company uses the platform to store the log of the data analysis and processing; Yahoo uses it for web search and is widely used in the field of advertising. Domestic Baidu is using the platform to handle with the log file of the daily web-pages, besides, and Taobao, Tencent and etc all use the platform. We can see that to study and use the Hadoop platform have become an important direction of many well-known companies, and created a lot of business value. Using Hadoop distributed technology to solve the problem of data under the text, not only solved the problem of the large space spending, also save a lot of time, from a large amount of data mining in a lot of use value, has the very vital significance.

Hadoop platform can provide a set of stable and reliable interface and data services, realize the algorithm of MapReduce to the text is divided into a number of small units, and each unit can be repeated. At the same time, the platform will be able to use the distributed processing system to store data, and can keep the data has a high throughput. Except, Hadoop platform can automatically handle node failure.

MapReduce, HDFS and HBase is existing throughout the platform, are the basis of the platform. The Hive is the language of higher level, but is based on the above three basic parts. The ecological system structure diagram of Hadoop platform is shown in FIG.1. HDFS and MapReduce are the two core system in the platform. HDFS can visit a large amount of data at higher rate and can be on a visit to the text data in the form of flow. MapReduce can carry on the decomposition of the data in the data storage nodes on the analysis of a large amount of data processing. In simple terms, Hadoop is the operation of the apache open source software foundation began to send on large-scale server for mass data storage, calculation, analysis of distributed storage and distributed computing framework. The main features include:

First is low cost: the Hadoop can run in general large cluster consisting of business machine and the machine configuration does not require high X86Linux server nowadays used.

Second is extensibility: Hadoop can almost linear increase the processing capacity of the cluster by extending the cluster nodes in order to deal with more data, the data stored can reach level.

Third is the higher availability: more cluster machines, the higher the probability of failures in hardware. Hadop can easily solve the problem of a hardware failure and loss of data backup.

It is because of the characteristics of Hadoop making it soon became the academic circles including the darling of the industry. Hardware cost is low, even college students can be easily deployed to accomplish her own cluster academic experiment, at the same time, scalability and high availability can be very good to undertake strict production tasks of the enterprises. Core service component of Hadoop includes two parts: the distributed file system HDFS MapReduce and distributed computing framework. Later we will detail. Hadoop family also includes the Hive, HBace, Zookeeper, Sqoop and other service components.

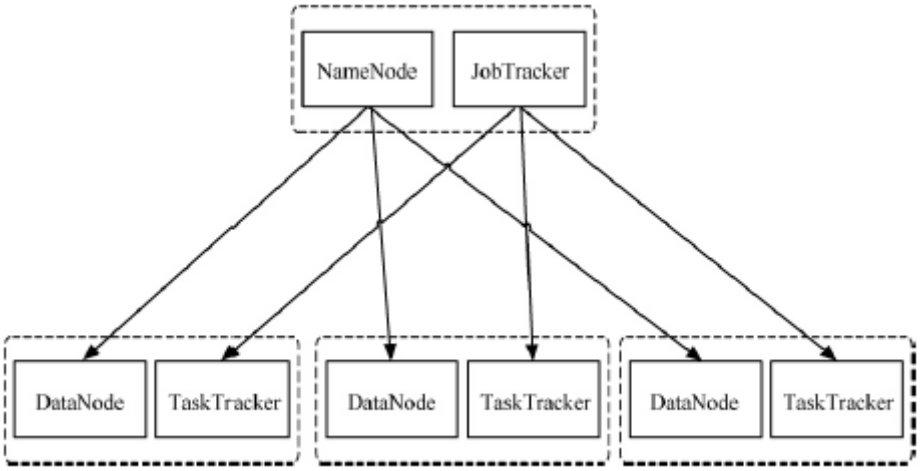

FIG.1 Basic structure of Hadoop

2.Structure of Hadoop and its key technologies

Hadoop is a distributed system infrastructure, developed by the Apache foundation. User can develop distributed application without understanding the distributed low-level details of the case, make full use of the power of the cluster in high speed computing and storage. Hadoop includes multiple components, mainly composed of distributed storage (HDFS), and distributed computing (MapReduce) these two basic parts, its typical basic deployment architecture is shown in FIG.1. Hadoop cluster is a typical structure of the master/slaves, the NameNode and JobTracker as the master, DataNodes and TaskTrackers as slaves. The NameNode and DataNodes complete the work of HDFS, the JobTracker and TaskTrackers are responsible for the work of MapReduce. Hadoop and HDFS are the open source implementation of Google GFS storage system, the main application scenario is the basis of MapReduce components, and is also the bottom of the distributed file system of BigTable (e.g., HBase、HyperTable). HDFS uses master/slave architecture, as shown in FIG.2. An HDFS cluster is consisted by a single NameNode, a certain number of the DataNode. The NameNode is a central server, is responsible for the management of the file system namespace and client access to the file. The DataNode in a cluster is commonly a node, is responsible for managing the nodes they attached storage. Internally, a file is divided into one or more of the block, the block is stored in the collection of the DataNode. The NameNode perform file system namespace operations, such as opening, closing, renaming files and directories, and at the same time decided to block to the specific DataNode node mapping. Hadoop MapReduce is a use simple software framework, the framework is shown in FIG.3. Based on its written application can run on large cluster composed of thousands of business machines, and parallel processing in the form of a

reliable fault-tolerant on TB level data sets. A MapReduce work usually the input data set segmentation for a number of independent data block, by the Map tasks (task) to completely parallel way of dealing with them. Framework to Map the output of the sorting first, then the result input to Reduce task. Usually work input and output will be stored in the file system. The entire framework is responsible for the task scheduling and monitoring, and to carry out the tasks of has failed.

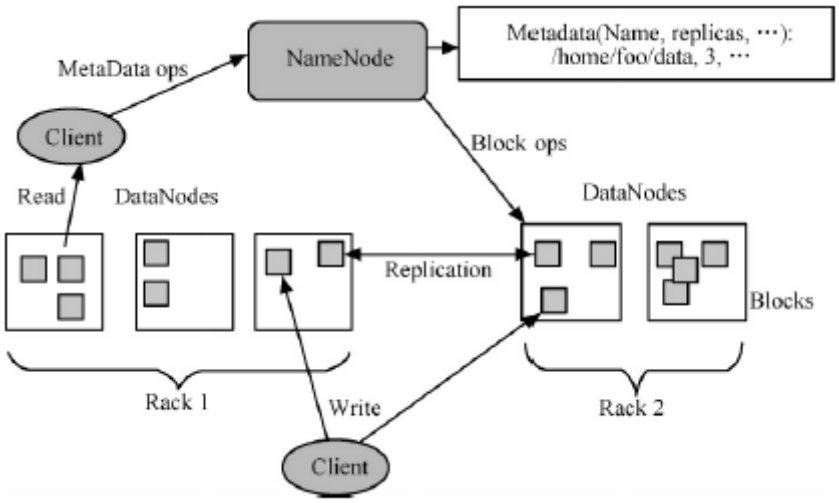

FIG.2 Structure of HDFS system

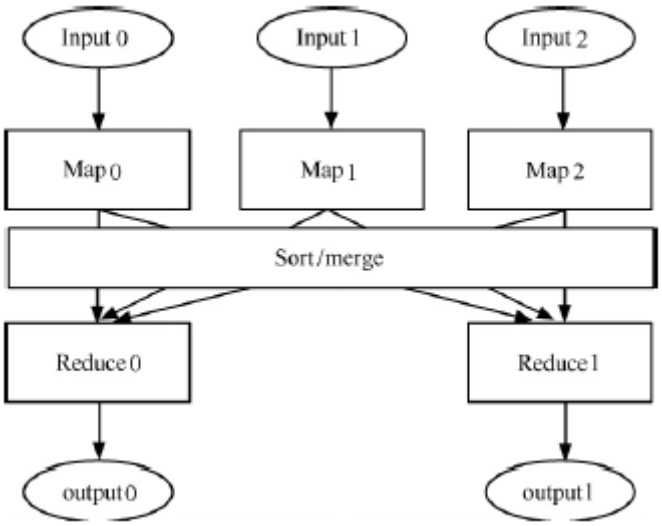

FIG. 3 Processing of MapReduce

3. Construction of huge data processing platform based on Hadoop

3.1 Technology architecture of mass data processing platform The prototype design of mass data processing platform including virtualization resources management platform and the Hadoop distributed parallel computing platform, the Hadoop distributed computing system deployment in resources management platform provides a virtual machine in the pool. Its characteristic is dynamically deploy Hadoop node server, rapid build Hadoop distributed computing system, its technical architecture.

Virtualization resources management platform is an experiment platform of our cloud computing laboratory based on XEN virtualization technology, the platform has 10 physical server nodes, 80 CPU, 160 GB memory, 10 TB storage resources, mainly including management system, resource management, security mechanism, intelligent scheduling and log review several core modules. System management includes virtual machine template management, performance monitoring and remote access management. Virtual machine template management is mainly used for rapid

customization system and installation business, namely through the existing production environment that exist in the physical machine or virtual machine intelligent template backup or test environment. Life cycle management, resource management includes the virtual machine physical machine life cycle management and cloud storage management module. Safety management includes user role management, unified management of authorization management and security auditing. Intelligent scheduling contains the resource balance transfer, power-saving mode transfer and elastic expansion of three modules.

3.2 Building experimental environment plan Based on the above analysis, the author deploy Hadoop servers in the laboratory management platform on the basis of the existing virtualization resources, prototype build huge amounts of data processing platform, the main implementation steps are as follows:

The server program Because the Hadoop distributed computing system can be deployed on the cheap hardware, so considering the limitation of the experimental environment and resources, only set 12 Hadoop deployment server nodes.

Create Hadoop virtual server templates To produce Hadoop virtual server templates is one of the key points to set up huge amounts of data processing prototype platform through the Hadoop virtual server template, which can be quickly installed in other Hadoop server nodes effectively saving the time of system installation, software environment configuration, quick to complete the deployment of the Hadoop distributed computing platform.

Clone Hadoop server templates According to the above production Hadoop virtual server templates, create other Hadoop virtual server nodes. Pay attention to when creating, machine name and IP address configuration are according to plan.

To start the Hadoop distributed computing system Before start the Hadoop distributed computing system, test environment first to check the network communication between the server and verify the SSH authentication normal or not. Then formatted file system, and finally to start the Hadoop distributed computing system. Can be activated after the success through a Web interface to control the Hadoop distributed computing system and management, and on a large scale of data processing.

One of the main advantages of MapReduce is fault tolerance. MapReduce realizes the fault tolerance through the monitoring of each node in the cluster. Each node to MapReduce on a regular basis and return to complete work and status updates, if a node in the silent time length exceeded expectations, the master node will be issued a circular, and redistribute the work to the other nodes. Hadoop running on the commercial independent service clusters can add or delete at any time in the Hadoop cluster server, Hadoop system will detect and compensate for any server hardware or system problem, namely, Hadoop is a self-healing system. In case of system changes or failure, it will still be able to run on a large scale high performance processing tasks, provide efficient data services.

Hadoop provides a low cost solution for the problem, it has no License limitations, so it is very loose can process mass data on 10, 50, or hundreds of machines, only need to develop a relatively simple mapping and simplifying the rules, they will be responsible for the distribution of these tasks to each machine and make sure all the tasks can be completed successfully. If there is any machine fault, they will redistribute the tasks on the machine to other normal machine.

4. Summary

With the rapid development of information technology, it has put forward the needs of building dynamic scalable huge amounts of data processing based on Hadoop prototype platform on virtualization resources management platform. After the analysis of the two core key technology of Hadoop, HDFS and MapReduce, it shows the urgent demand of the smart grid to huge amounts of data processing on the basis of in-depth analysis and research and gives its technical architecture, implementation plan and case analysis There are a lot of limitations and the insufficiency need more in-depth research , but Hadoop has more potential advantages in cost control and on the analysis of database as well as for analytical database scalability advantages.

References

[1]F. Wang, L.B.Hua. Model analysis of distributed file system in Hadoop. The research and development, 2010, (12)
[2]G.M.Hu, L.Zhou, L.X.Ke. Research on web log analysis system based on the Hadoop. Computer knowledge and technology, 2010, 6 (22)
[3] Dean J，Ghemawat S．MapReduce：simplified data processing on large clusters．In：Proc．of the 6th Sympon Operating System Design and Implementation． Berkeley：USENIX Association，2004
[4]J.Hu, R,J.Shuai, W.Hou. Huge amounts of data storage technology research based on cluster technology. Heaven and earth of software, 2010, 26 (61).
[5]X.Hu, J.Feng. Distributed search engine under the Hadoop. Computer system analysis, 2010, 19(7).