Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study of Hadoop Distributed File system in Cloud Computing

Radhika M. Kharode, Anuradha R. Deshmukh
Department of Computer Science and Engineering
Sant Gadge Baba Amravati University (SGBAU),
Amravati, Maharashtra, India

Abstract*— Hadoop is a free, open source framework that supports the processing of large data sets in a distributed computing environment. So Hadoop is applied widely as the most popular distributed platform. HDFS is comprised of interconnected clusters of nodes where files and directories are resided. HDFS, the Hadoop Distributed File System, is a distributed file system designed to hold very large amounts of data (terabytes or even petabytes), and provide high-throughput access to this information. Files are stored in a redundant fashion across multiple machines to ensure their durability to failure and high availability to very parallel applications. As Cloud computing is a super computing model which is based on Internet and participated by public. Cloud computing is a development of Distributed Computing, Parallel Computing and Grid computing by using this feature of cloud computing can make HDFS more powerful. This paper mainly presents a Hadoop platform computing model and the K-means to implement the effectiveness analysis and application of the cloud computing platform.

Keywords*— Hadoop, cloud computing, HDFS, Map/Reduce, K-means.

I. INTRODUCTION

Cloud computing technology is widely recognized for IT companies such as Google, IBM, Amazon and Microsoft which have been launched their own commercial products. hey also make the cloud computing technology as the priority strategy in the future development. But this will lead to a large number of data problems each user may have a huge amount of information. From now on, the transistor circuit has been gradually approaching its physical limits. Facing the massive information, how to manage and store the data are the important issues we should deal with this by using distributed file system based on Hadoop which is known as HDFS.

Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

Hadoop issued to build cloud computing platform which is designed by Apache open source projects. We use this framework to solve the problems and manage data conveniently. There are two major technologies: HDFS and Map/Reduce. HDFS is used to achieve the storage and fault-tolerant of huge document, Map/Reduce is used to compute the data by distributed computing.

II. RELEVENT CONCEPTIONS

A. *Cloud Computing*

Cloud computing is one of the outsourcing of computer services. It has a capacity to provide on demand networking resources. Cloud computing builds a virtual group of resources such as network, storage, central processing unit and memory. As cloud computing is the immerging technology which plays an important role to provide a network access and supports ongoing open source cloud services.

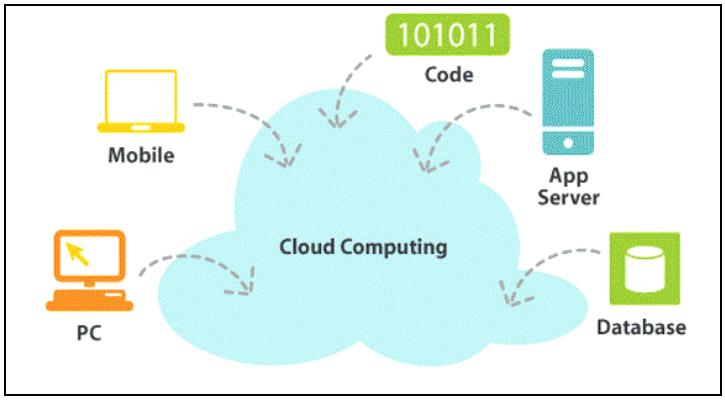

Fig 1: general structure of cloud computing

The goal of cloud computing is to provide easy, scalable access to computing resources and IT services which plays an important role to provide computing, communication and storage resources in a safe environment based on a service as fast as possible which is virtually provided via internet platform. It is also responsible for cost reduction, integration, and reusability of services. These services are mainly categorized into four parts that are as follows

- Software as a Service (SaaS) which offers renting application functionality from a service provider rather than buying, installing and running software by the user.
- Platform as Service (PaaS) which provides a platform in the cloud, up on which applications can be developed and executed.
- Infrastructure as a Service (IaaS) in which vendors offer computing power and storage space on demand.

In the bank system. We can store the data and use the application as conveniently as we can save and manage money in the bank. The user no longer need to use so many hardware as background supporting, the only thing is connecting to the cloud.

B. HDFS

HDFS is a distributed file system and uses the ordinary hardware, which has been achieved by the GFS structure from the Google paper. It adopts the master/slave model.

HDFS is comprised of interconnected clusters of nodes where files and directories reside. An HDFS cluster consists of a single node, known as a *NameNode and other is DataNodes. Name node* is manage the file system namespace and regulates client access to files. Name node is also called the master node which contains the whole information about the name system as well as the information the data blocks against each data node. In data nodes store data as blocks within files. Data nodes are the actual storage nodes where data need to be hold and upon request it can be read and write.

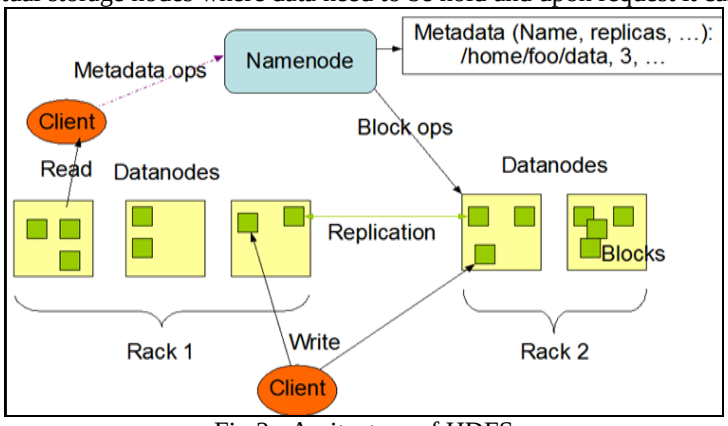

Fig 2 : Arcitecture of HDFS

Figure illustrate user applications access the file system using the HDFS client, a library that exports the HDFS file system interface.

Like most conventional file systems, HDFS supports operations to read, write and delete files, and operations to create and delete directories. The user references files and directories by paths in the namespace. The user application does not need to know that file system metadata and storage are on different servers, or that blocks have multiple replicas.

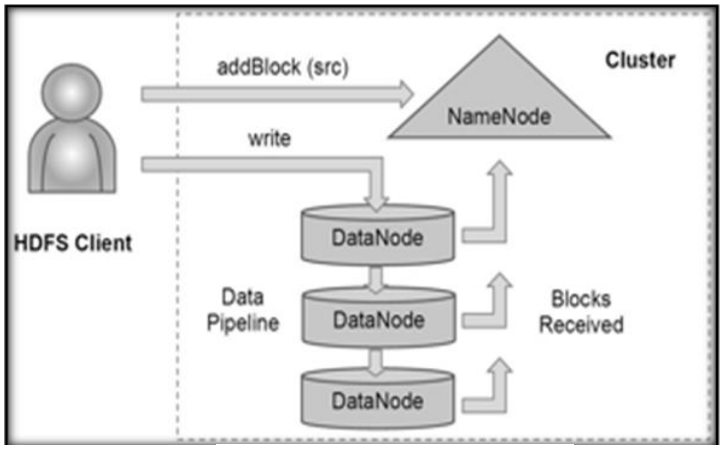

Fig 3: HDFS Client Creates a New File

When an application reads a file, the HDFS client first asks the NameNode for the list of DataNodes that host replicas of the blocks of the file. The list is sorted by the network topology distance from the client. The client contacts a DataNode directly and requests the transfer of the desired block. When a client writes, it first asks the NameNode to

choose DataNodes to host replicas of the first block of the file. The client organizes a pipeline from node-to-node and sends the data. When the first block is filled, the client requests new DataNodes to be chosen to host replicas of the next block. A new pipeline is organized, and the client sends the further bytes of the file. Choice of DataNodes for each block is likely to be different.HDFS also supports third-party file systems such as CloudStore and Amazon Simple Storage Service (S3).

C. Map/Reduce

Map/Reduce, presented by the Jeffery Dean and Sanjay Ghemawat, is a programming model in the massive data computing, which is developed by Google and meanwhile is a core technology of cloud computing. The model abstracts the common operation of large dataset as Map and Reduce steps to simplify the programmers' difficulty of distributed and parallel computing.

Firstly, dataset are split the several small parts by Map, then several parts are sent to a large number of computers to make the parallel computing, and which will produce a series of intermediate keys. Finally, Reduce will integrate the all results and output them. The structure is shown as Figure.

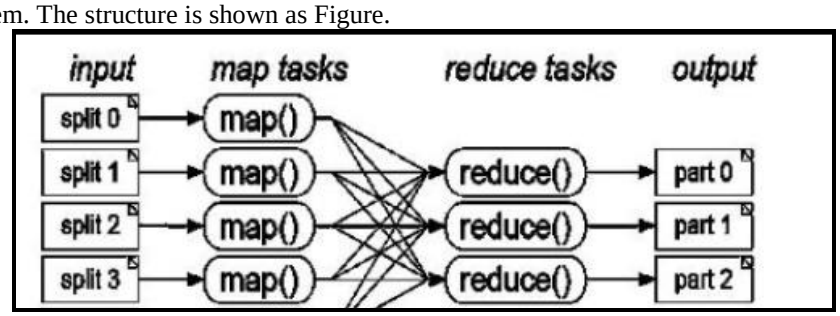

Fig 4 : The structure of Map/Reduce process

Map function is usually supported to the diffident needs of users according to the different business. It is used to deal with a pair of key-value and to get a new pair of key-value as intermediate results. The Map/Reduce function library will put the same value together and send to the Reduce function. Reduce function is the same as Map function defined by the user themselves. According to the transfer intermediate key, it will deal with the results and merge the same key to produce a smaller set of values. So, we can conclude the process of Map/Reduce operation:

$$\text{map}(\text{key}_{\text{in}}, \text{value}_{\text{in}}) \rightarrow \text{list}(\text{key}_{\text{out}}, \text{value}_{\text{intermediate}})$$

$$\text{reduce}(\text{key}_{\text{out}}, \text{list}(\text{value}_{\text{intermediate}})) \rightarrow \text{list}(\text{value}_{\text{out}})$$

III. K-MEANS ALGORITHM USED FOR HDFS IN CLOUD COMPUTING

Cluster algorithm is an important factor in data mining, especially in such a large system of data computing. The clustering is particularly vital in cloud technology. The division of data characteristics is the vital step in the storage and security of cloud computing. There are so many algorithms in cluster analysis such as combine of K-means and Map/Reduce .

K-means algorithm is an objective function which is aiming at optimizing the distance between the data point to the center point. That distance uses Euclidean distance similarity as measure. This function makes the clustering meet some rules: the similarity of objects in the same cluster is higher, and the difference between the cluster types has less similarity. It means the idea of "High cohesion, Low coupling" in software engineering.

HDFS and MapReduce are two principal parts of Hadoop.In Parallel *K-means* algorithm, Map and Reduce are performed separately in each iteration. In the Hadoop platform, HDFS is storage of mass data and manages these data file. In addition, it will record the data nodes where these data files are distributed to and then get from, while the initial center are recorded at the same time. The task of Map function is to work out the distance between each recorder and center site, and re-mark the type which it belongs to. The input should be all wait Clustering Data and the Clustering Center in the last round, also the Data Clustering<key,value>. On the basis of computed results from Map function, the task of Reduce function is to calculate new Clustering Center and send it to each node, and update the results in the HDFS before the next iteration, until convergence.Steps to work out Parallel *K-means* algorithm:

Step 1: Select K samples arbitrarily as initial center ID;

Step 2: Iterate and perform Map and Reduce;

- Each site receives cluster center from center site;
- Calculate sample size from every local cluster, and send it to center site;
- Calculate new Global Cluster Center, and send it to each site.

Step 3: Repeat until convergence.

Consequently, we can use K-means clustering method to separate the data which have similarity to each other. Combining with the Map/Reduce in the Hadoop, we can distribute storage and use the data in cloud computing. The structure is shown as

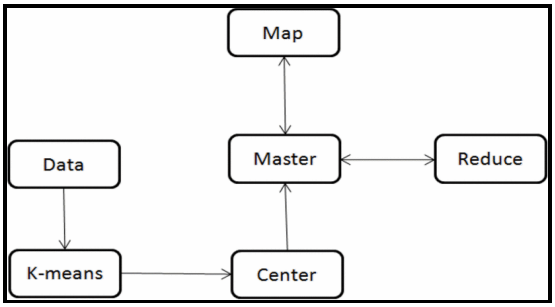

Fig 5 : structure of combing K-means with the Map/Reduce in Hadoop

According to the diagrams display, we can generally divide the process into five steps:

- In cloud computing, data will be distributed in so many blocks by using the K-means algorithm to ensure data in each block have high similarity.
- We distribute the block again by Centre controlled, and then we specify the corresponding pointer to the each distributed data to make the Map/Reduce operation become more convenient.
- Then we should notice to the Master Management the location of these small block data to facilitate the Master to assign the tasks.
- Master split the data point to Map to operate.
- Map return the intermediate value to Master then let Reduce operate.

This process, which compares to the previous algorithm without using the K-means, has more impact on the data distribution. Before the Map/Reduce, we use the K-means to classify the data types which is more effective to manage the data classifications.

IV. CONCLUSIONS

Data storage is an important element of cloud computing. This paper discussed working of HDFS and Map/Reduce in Hadoop framework. Combination of data mining and K-means clustering algorithm will make the data management is easier and quicker in cloud computing model. Cloud computing will develop towards the security and reliable directions. As cloud computing is the immerging technology which plays an important role to provide a network access and supports ongoing open source cloud services.HDFS is more robust by using cloud concept.

REFERENCES

[1] Apache Hadoop. http://hadoop.apache.org/
[2] J. Venner, Pro Hadoop. Apress, June 22, 2009.
[3] T. White, Hadoop: The Definitive Guide. O'Reilly Media, Yahoo! Press, June 5, 2009.
[4] HADOOP: Scalable, Flexible Data Storage and Analysis. http://www.cloudera.com/wpcontent/uploads/2010/05/Olson_IQT_Quarterly_Spring_2010.pdf
[5] L. Jiang, B. Li, M. Song, "THE optimization of HDFS based on small files", In 3rd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT2010), Beijing, 2010. pp.912-915.
[6] YangChenzhu. The Research of Data Mining Based on Hadoop. Chongqing Univerisity. 2010
[7] Twitter's Hadoop-LZO, http://github.com/twitter/hadoop-lzo
[8] Introduction to cloud computing by Shang Juh kao
[9] OpenCloudConsortium. Http://opencloudconsortium.org

