Analysis of Optimize Approaches to work with massive amount Smaller Size File with respect to HDFS

Rajeshkumar R. Savaliya
Assistant Professor
Ambaba Commerce College, MIBM & DICA-Sabargam

ABSTRACT

Hadoop Distributed File System (HDFS) is a framework used with Big-Data. It developed for distributed data handling (DFS), data accessing and data storing. HDFS used to store the large size data files through the cluster of computer system. Hadoop is an open-source software implementations and framework which has two components MAPREDUCE and HDFS [1]. Hadoop Distributed File System became very popular to manage and handle large size files with very high performance of distributed system. But the working with the large volume of small size file becomes big challenged in HDFS because of large number of small files forces more burdens to creation and management of metadata on NameNode. This paper present on introducing a comparatively analysis for different corporate methodology such as Hadoop Archive(HAR) and MapFile are existing to enhance the performance of storing and processing large numbers of small size files in Hadoop.

KEYWORDS: HDFS, BIGDATA, HADOOP, MAPREDUCE, HAR, MAPFILE.

1. INTRODUCTION

Now a day, Big-Data becomes valuable Terminology for handle big size data file. Big-Data is generating using different digital statement or e-transaction and different available social media used for communication in the entire world of digitalization. Hadoop implement to manage as well as handle large size of data file. To storing as well as retrieving the large numbers of small size files with respect to Big-data that become current challenges in the field of IT. So that resolve the current challenges, we analyses the two different small size file handling mechanisms such as HAR and MapFile those are currently used with Hadoop. Both HAR and MapFile have the differences and similarities in the working with massive small size files. That design based on different parameters, its levels and different criteria to manage the huge volume of Small size data file. The main objective of HDFS with HAR and MapFile ware developed for maintain as well as handle the large amount of small size data files received from different terminals in a variety of formats. Both HAR and MapFile file system are developed to handle big-data in the form of large numbers of small size files. In this paper comparatively study and analysis done with respect to different parameter and features of both HAR and MapFile. HAR and MapFile file system were two different distributed file systems used to handle massive amount of smaller file. Hadoop is latest distributed cluster environment use to handle the large amount of big data.

2. HADOOP DISTRIBUTED FILE SYSTEM

Hadoop is the open source framework to manage as well as maintain big size data file. HDFS is a design by Apache. In current era, subsequently many network grounded applications implemented through Hadoop, most

popular social media like facebook, amazon as well aswhatsapp. HDFS is design to work with big size of data file. Hadoop file system has HDSF core component and HDFS has two components DataNode and NameNode [2]. HDFS architecture work as follows.

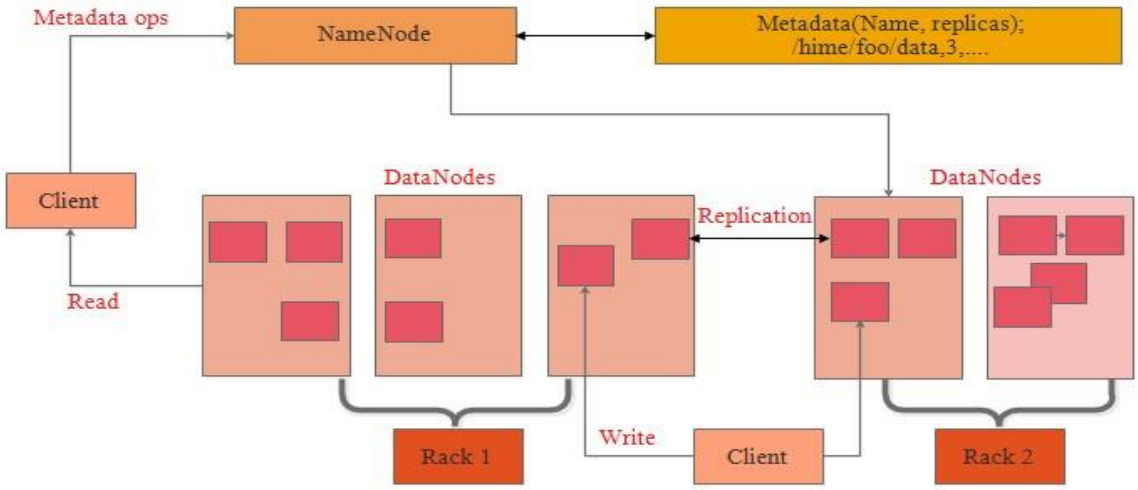

Figure 1: HDFS Architecture [6]

HDFS again divided into mainly two components such as NameNode and DataNode [7].

2.1 NameNode

NameNode component of the HDFS is use for handling as well as storing metadata of data file stored in DataNode. NameNode is also help for maintain heartbeats between NameNode and DataNode. NameNode. NameNode has the following impotence features as follows.

- The NameNode is used to maintain and executes the file system in HDFS.
- The NameNode saves a record of how the files are divided into blocks in HDFS.
- NameNode regularly receives a Heartbeat for DataNode.

2.2 DataNode

DataNode component is help to storage of the big size data files in the free available different data block in the bunch of HDFS. DataNode is also used to use store replica in available different data block around in the cluster of HDFS.

DataNode has the following impotence features as follows.

- DataNode is used to perform read as well as write operation call made from the clients.
- It is responsible to constructing data blocks as well as removing data blocks based on decisions taken by the NameNode.
- DataNode on a regular basis send details in formation for all blocks existing within cluster towards the Name Node.
- Data nodes send heartbeats at every 3 seconds to the NameNode to report the current condition of HDFS.

3. PROBLEM STATEMENT OF SMALL SIZE FILE HANDLING IN HDFS

Hadoop is the more popular Distributed File system for storing and processing as well as handle the large size data file in huge number of computer system in the cluster environment. The Hadoop operational capacity decrease when we work with handling, storing and processing large number of small size file in the HDFS Framework with

respect to Big-Data. In current eras' so many available systems produce and generate the number of smaller size data files within field of energy, biology, number of e-commerce small transactions, e-libraries, e-learning and currently available social media likes WhatsApp or Facebook [11]. It's become current challenges and problem to handle the storage, processing as well as accessing small size file in HDFS, because of NameNode generate the meta-data for each file so it become so critical ho handle the Metadata for all small size file. And replication is also becoming more difficult to handle in the HDFS open source Framework environment.

4. SOLUTION TO HANDLE LARGE NUMBER OF SMALL SCALE FILES

There are two majors optimize solution in Hadoop distributed processing file system for handle the large numbers of small scale file in the cluster base network, namely HRA and MapFile currently used with Hadoop open source Framework environment.

4.1 Hadoop Archive Files

Hadoop Archive (HAR) provide the archive facility that will combine the file in respected HDFS data block. The main roll of Hadoop Archive File is maintaining the two things metadata files as well as data files. HAR mainly used to growth the performance related to memory consumption for NameNode. HAR is used to resolve the current problem of storage great number of smaller data file. Hadoop Archive first merges the large number of small size files within single archive file. Hadoop Archive has the file extension.hrs. so that map reduces and NameNode requited to handle only archive file instead of handling numbers of small files, so performances will be improving for data processing [9]. Hadoop Archive (HAR) working structure as follows.

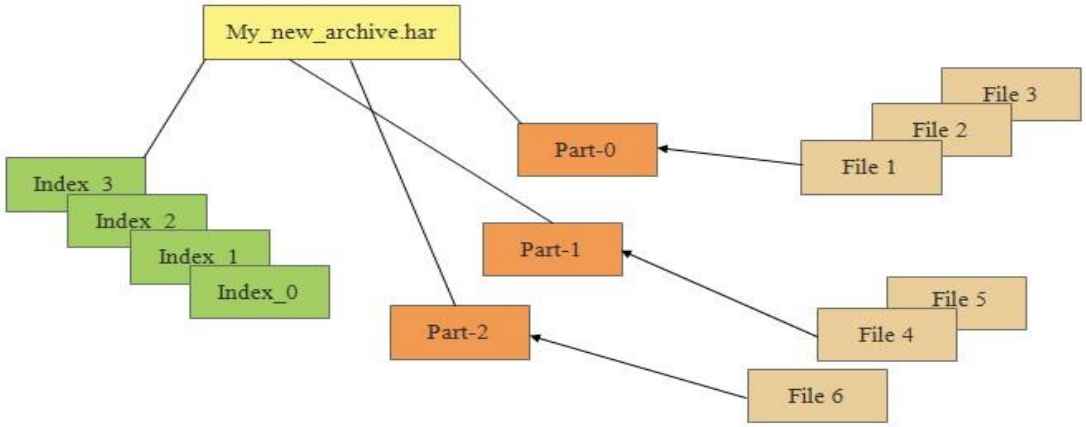

Figure 1: HDFS Architecture [5].

4.2 MapFile

MapFile is an extended scheme of Sequence file which is used to work with small size files. MapFile structure divided in two section, one is the Index section and second is the Data section [10]. In this approach first must have to sorting entire data element before storing then into HDFS. So that this procedure rises the extra overhead to work with large number of small size files. So we have analyses that Mapfile design has extra sorting over load than Hadoop archive. But, Mapfile is more flexible than Sequence file as well because of it index layout [8]. MapFile has following working structure as follows.

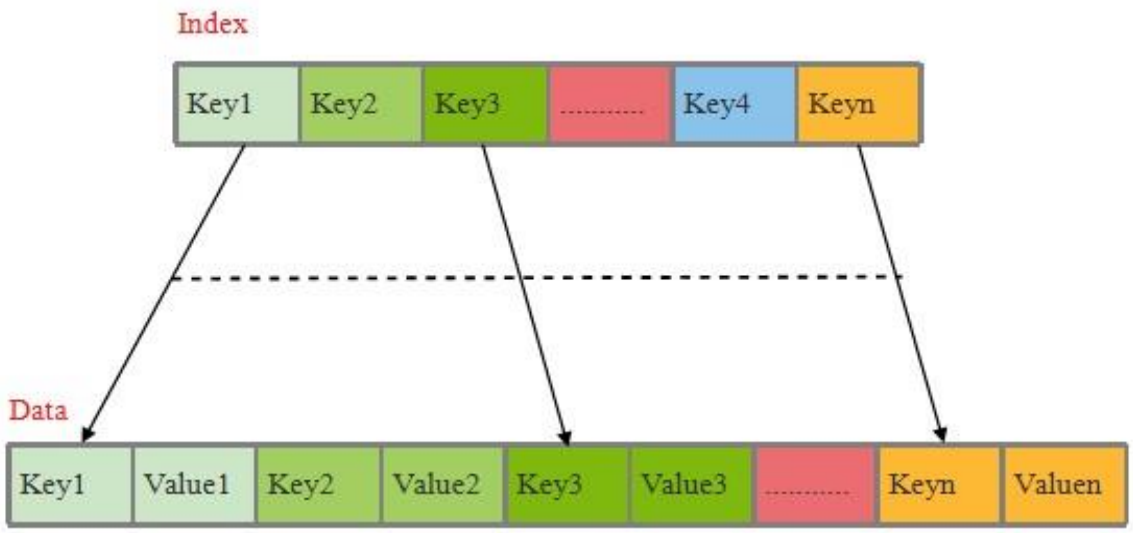

Figure 3: MapFile Layout [4].

4.3 ANALYSIS OFOPTIMIOZE APPROACHES FOR SMALL SIZE FILESHANDLING IN HDFS

Purpose	Hadoop Archive File	MapFile
Objective of file System	Hadoop archive file System was developed for work with great volume of smaller data file in the Hadoop.	MapFile System also develops for work with great volume of smaller data file in the Hadoop.
Consist of	HAR mechanism is comprises of Metadata file and Data file. Metadata file again contain Index and Master Index file.	MapFile mechanism is comprises of one Index file and one Data file.
Sorting	Files are not sorting before storing	All Files are sorting before storing
Memory Management	HAR will decrease the memory consumption for the NameNode.	MapFile will also efficiently reduce the memory consumption as compare as HAR for the NameNode.
Efficiency	Low	Middle

5. CONCLUSION

Hadoop Distributed file system(HDFS)open-source framework environment is used for better work with respect to Big-Data. HDFS mainly manage the storing as well as processing of very large size files in more resource fully fashion. But in opposite, HDFS is unproductive to handle bulky volume of small data file. Because HDFS has problem of metadata creation and metadata preservation for each small size file respectively. So it more problematic to store as well as process number of small size files in Hadoop. To overcome this problem in HDFS

for working with huge small files. In this paper we are mainly two importance mechanisms impose on Hadoop layer with specially focuses for provide solution of large number of small size files management mechanisms in HDFS.

6. REFERENCES

[01] http://hadoop.apache.org [Accessed: January. 10, 2018]
[02] http://en.wikipedia.org/wiki/Big_data. [Accessed: January. 12, 2018]
[03] http://www.cloudera.com/hadoop/ [Accessed: January. 13, 2018]
[04] https://www.researchgate.net/figure/The-way-of-sequencefile-HBase-is-used-tocomplete-data-writing-rather-than-write_fig2_335637597[Accessed: January. 13, 2018]
[05] https://www.waitingforcode.com/hdfs/handling-small-files-in-hdfs/read [Accessed: January. 15, 2018]
[06] https://bigdata2world.wordpress.com/2016/06/23/hdfs-architecture/[Accessed: January. 15, 2018]
[07] http://en.wikipedia.org/wiki/Big_data .[Accessed: January. 18, 2018]
[08] Bharti Guptu, Rajender Nath, Girdhar Gopal and Kartik, "An Efficient Approach for Storing and Accessing Small Files with Big-Data The quenology", International Journal of Computer Application, Vol. 146, Issue-1, PP. 0975-8887, July 2016.
[09] Deepika. "An Optimized Approach for Processing Small Files in HDFS", International Research Journal of Science and Research, PP.402-405, 2015.
[10] Divyashikha, Shalini Sheeoran, Huzur Saran "Optimize MapFile based Storage of Small Files Hadoop", 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, PP. 906-912, 2017.
[11] Yingchi Mao,Bicong Jia et.al, "Analyzing Optimization Scheme for Small Files Storage Baseds On Hadoop Distributed File System."International Journal of Database Theory and Application , PP. 241-254, 2015.