ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (667-671), Month: April - June 2015, Available at: www.researchpublish.com

Hadoop Big Data for Processing Data and Performing Workload

Girish T B[1], Shadik Mohammed Ghouse[2], Dr. B. R. Prasad Babu[3]

[1]M Tech Student, [2]Assosiate professor, [3]Professor & Head (PG), of Computer Science and Engineering Department, SEACET, Bangalore, India

Abstract: **Apache Hadoop is the Java based open source platform that makes processing of large data possible over thousands of the distributed nodes and hadoop development resulted from publication of the two Google authored white papers like Google File System and the Google MapReducing. There are numerous vendor-specific distributions available based on Apache Hadoop and from the companies such as Cloud era (CDH),the Horton works (1X/2X) and Map (M3/M5). From the addition, appliance-based solutions are offered by many vendors like IBM and Oracle.**

Hadoop distributed file system (HDFS) is the basic component of the Hadoop framework that manages the data stora and it stores the data in the form of the data blocks [default size is: 64M]) on hard disk and the input data size defines the block size. A block size is 128MB for the large file set is the good choice.

Keywords: **HDFS, Hive, Pig, Hbase, Mapreduce work.**

I. INTRODUCTION

Hadoop is an open source Apache framework and it is written in java that allows distributed processing of the large datasets across the clusters of computers using the simple programing models. In the Hadoop framework application works in the open environment that provides the distributed *storage* and the *computation* across the clusters of computers and hadoop designed is to scale up from single server to thousands of machines and each is offering local computation and the storage. Hadoop was created by Doug Cutting and the creator of the Apache Lucene, and are widely used text search library. Hadoop has its origins in Apache a Nutch is an open source web search engine, itself a part of the Lucene projects.

Fig. (a)

II. HADOOP BASE COMPONENTS

A. HDFS:

Hadoop distributed file system is the basic component of the apache hadoop framework and it manages the data storage and it stores data in the form of blocks of data on the hard disk and the input data size is defineing the block size to be used in cluster. A block size of 128MB for large file set is a good choice of the file system.

Keeping the large block sizes means that a many small number of blocks can be storing and, there by minimizing of the memory requirement the master node (also called Name Node) for storing the metadata information and the block size also the impacts the job execution time, and there its by cluster performance and many of the files can be stored in all HDFS with the different block size during the uploading process and many of the file sizes smaller than the block size and the hadoop clusters may not be performing optimally.

Hadoop framework includes the following two modules:-

Hadoop Common: These are the Java libraries and utilities required by the other hadoop modules.

Hadoop YARN: This is the framework for job scheduling,cluster resource management.

B. Map Reduce:

It is the second basic component of Hadoop the framework technology and it is managing the data processing part on the cluster of hadoop system. A cluster may runs in the Map Reduce jobs written in the Java or any other language via the streaming. A single job may consist of the multiple task and the total number of tasks that can run on the data node and it is governed by the amount of the memory is installed. It recommended to have the more memory installed on the each data node for better clusters performance. Based on the application of workload that going to run on clustering and a data node and it requires a very high compute power, and high disk input/output (or) its additional memory.

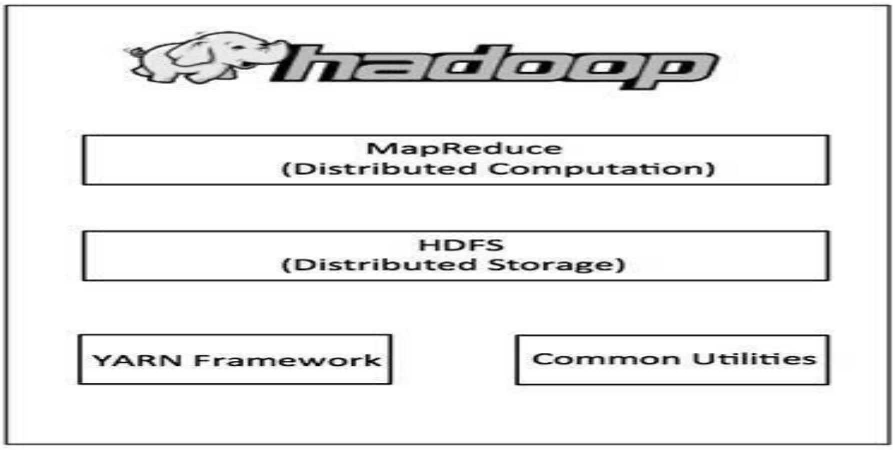

Fig. (b)

III. HADOOP ECOSYSTEM

Pig it provides the high level procedural language known as Pig Latin. Programs written in Pig Latin are automatically converted into the mapreduce jobs. This is very suitable in the situations where developers are not familiar with the high level languages such as c, java, but have the strong scripting knowledge and expertise and it will create their own customized functions in the system. It is pre dominantly recommended for processing the very large amounts of semi-structured data.

Hive provides the SQL like language and known as Hive QL and analysts can run the queries against the data stored in the HDFS using the Structured Query Language(SQL) like queries that automatically get converted into map-reduce jobs. It is recommended where we have large amounts of un-structured data and we want to fetch the particular details without writing the map reducing job and companies like Revolution Analytics can leverage the Hadoop clusters running Hive via

RHadoop connector. RHadoop provides the connector for the HDFS, Map Reduce and HBase connectivity of the hadoop system.

HBase is the another type of NOSQL database and offers scalable,distributed database system and however, it is not further optimized for the transactional applications and its relational analytics. HBase is not the replacement for the relational database. A cluster with the HBase installed and require an additional 24GB to 32GB of the memory of each of the data node.

IV. HDFS ARCHITECTURE

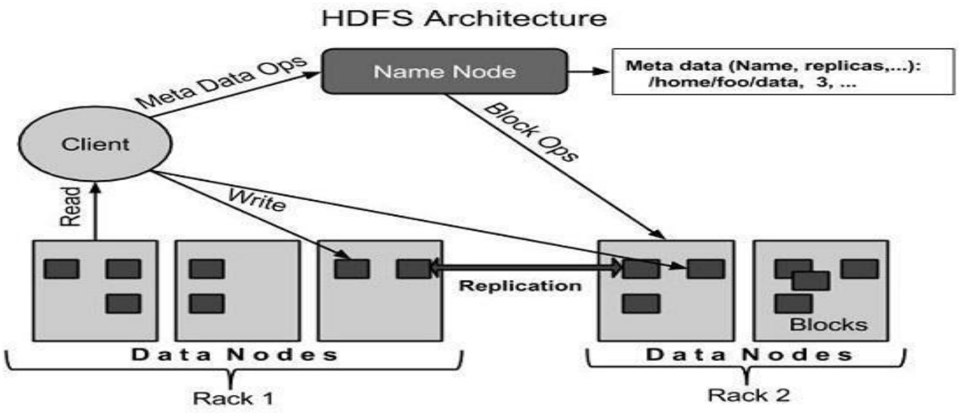

Fig. (c)

V. DESIGN OF HDFS

HDFS is the file system designing to storing large files with streaming data access patterns and running on the clusters of commodity hardware and it examine this statement.

In more detail, The files are given below

Very large files Very large is this the context means files that are hundreds of the megabytes, gigabytes, or terabytes in the size and there are many Hadoop clusters which are running today that stores in petabytes of data.

Streaming data access The HDFS is built around the idea that the most efficient data is processing pattern is a write only once and it reads many times in pattern design. A dataset is typically copies and generated from the source, then to various analyses are then performed on dataset over time and in hadoop each analysis will involve in a large proportion, and all of dataset to read, so the time is to read the whole dataset is most important than the latency of reading its first record and then the other records.

Commodity hardware It does not require the expensive, highly reliable hardware to run on and is mainly designed to run on the clusters of the commodity hardware (for multiple vendor) for which the chance of the node and failure across the cluster in the very high, at least for large clusters and the HDFS is designed to carry the working without the noticeable and an interruption to user in the face of the failure and it is also the worth examines the applications of which it is using HDFS system does not work and this may be change in the future and all these are areas where the HDFS is not good fit today and it works with the framework.

Low-latency data access The applications that are requires the low latency access to the data, in the tens of milliseconds range, it will not work well with the HDFS system. Remember, the HDFS file and it optimized for the delivery and the very high throughput of the data system and it may be at the expense of the latency of that data model.

Lots of small files Since the name node holds the file system metadata in the memory, the limit to the large number files in filesystem and it is governed by the amount of memory on the namenode and as the rule of the thumb, each file of the directory, block takes about to 150 bytes. So, for example we are having one million files, each taking one block,and you would need at least 300 MB of the memory and while storing the millions of all files is feasible, and billions of the capability of the current hardware of the file system and it is maintained the hadoop file system.

Multiple writers, arbitrary file modifications The Files in the HDFS may be written to by the single writer and always writes are always made at end of file. There is no support for the multiple writers or for the modifications at the arbitrary offsets in the file system.

VI. HADOOP CLUSTER AND HDFS

HDFS is a block-structured system.

HDFS does come out with its own utilities for the management.

HDFS system stores the metadata reliably: Name Node

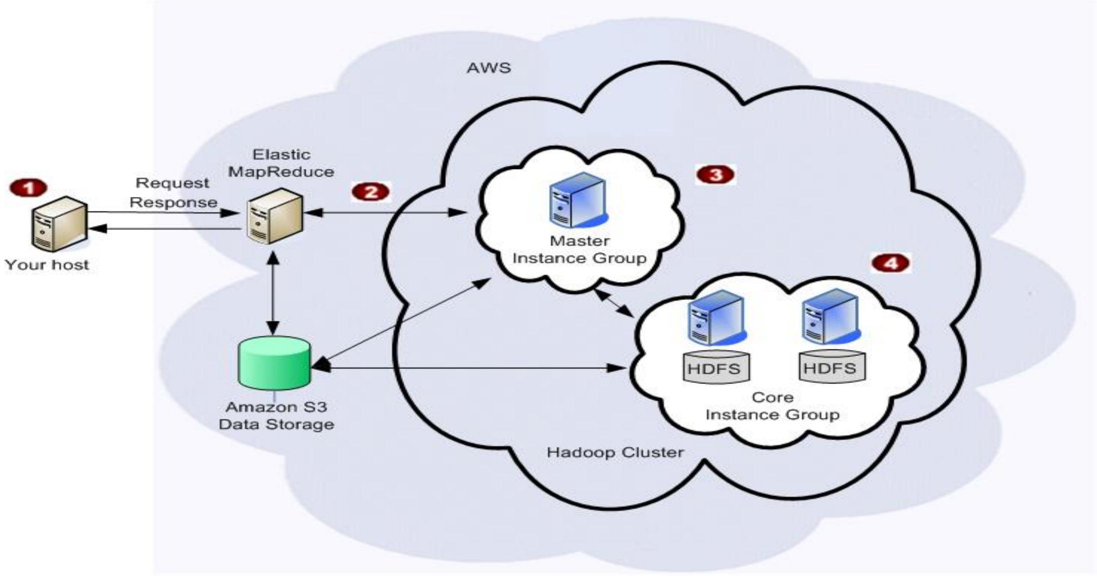

Fig (d)

Distribution of Hadoop:

Enterprise-grade platform for hadoop system.

Comprehensive management suite.

Industry-standard interfaces.

Combines with system open source packages and enterprising grade dependability.

For Higher performance.

Mount Hadoop with Direct Access NFS.

VII. FUTURE ENHANCEMENT

Hadoop is the strong framework for running the data intensive jobs it is still evolving to the ecosystem projects around it mature and expaninng the market demand is spurring new technological development. With these innovations come many various options are choosen from, the such as building new clusters and leveraging an appliance or making the use of Hadoop-as-a-Service(HAS) offerings.

Moreover, the market filling with the new companies and, which is creating the more competitive technologies to consider and options for the starters, the enterprises need to make the correct decision in choosing Hadoop distributions and enterprises can start the small by picking up any distribution and building 5 to 10 nodes of cluster to testing the features, functionality and its performance. The most reputable vendors to provide the free version of their distribution for an evaluation purposes and we are advise enterprises the test 2 to 3 vendors for distributions against business requirements and before making its any final decision of the hadoop system. However, IT decision makers may pay the close attention to the emerging competitors to their positions and to their organizations for impending technology enhancements.

REFERENCES

[1] The architecture of HDFS is described in "The Hadoop Distributed File System" by Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler (Proceedings of MSST2010, May 2010, http:// storageconference.org/ 2010/Papers/MSST/Shvachko.pdf).

[2] "Scaling Hadoop to 4000 nodes at Yahoo!," http://developer.yahoo.net/blogs/hadoop/2008/09/scaling _hadoop_to_4000_nodes_a.html.

[3] http://rtraining.comsysto.com/slides/slides.pdf.

[4] http://rtraining.comsysto.com/slides/slides.pdf.

[5] Oracle R Advanced Analytics for Hadooppackage:http://www.oracle.com/technetwork/bdc/big-data-connectors/ downloads/index.html

[6] IDC Report:Worldwide Big data technology & service 2012-2015 forecast.

[7] www.centurylink.com/business/rticles/pdf/resources/big-data-defining-the-digital-deluge.pdf

[8] http://en.wikipedia.org/wiki/Hadoop.

[9] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.

[10] http://pig.apache.org/;http://hive.apache.org/;http://hbase.apache.org"Comparing pig latin & sql for constructing data processing system.

[11] www.revolutionanalysis.com/products/resolution-enterprise.php.

[12] hpccsystems.com/why-HPCC/hpcc-vs- hadoop.