Hadoop Distributed File System: Introduction and usage on different workloads

Ekta Bhardwaj[1]
IEC Engineering College , Greater Noida

***Abstract*— Distributed file system does not able to hold the very large amount of data. It can be manageable by the help of Hadoop distributed file system (HDFS). Files are store in multiple locations (nodes) or stored on various servers.**

This paper contain step to step introduction of Distributed File System and Hardtop distributed file system(HDFS).In This paper Hadoop analysis is also done on different workloads. This is done on clusters (open cloud,M45,Web-Mining). Three clusters have different hardware and software configurations and range in size from 9 nodes (Web Mining), to 64 nodes (Open Cloud), and 400 nodes (M45). Hadoop clusters are used to improve interactivity, improve effectiveness of authoring and efficiency of workloads and automatic optimization.

Hadoop clusters are used to improve interactivity, improve effectiveness of authoring and efficiency
of workloads and automatic optimization.

Keywords— Distributed File System(DFS), Hadoop, Hadoop Distributed File System(HDFS).

I. SECTION: 1DISTRIBUTED FILE SYSTEM (DFS):

It is based on client-server architecture. Files are stored and access through this architecture.DFS use a mapping scheme to define the address(or to define where a file is located?) if user want to fetch the information then this request are send to server and server execute the request. It provides the distribution among several clients and provides a centralized system.

II. SECTION:2HADOOPDISTRIBUTED FILE SYSTEM

Hadoop is a software plateform .it implements the Map-Reduce
.Map-Reduce is implemented by HDFS.HDFS works in a Master –Slave fashion.In which it have one master node that is known as Name data(It manage the whole system),no of slave mode is known as Data nodes.
HDFS is fault tolerant and useful to design at low cost Hardware. It provide high throughput and useful where (large data set) we have a huge amount of data .HDFS is a part of Apache Hadoop core project.
An HDFS consist of 100 or 1000 of server Machine. There are huge no of component .The main goal of HDFS is to detect fault and recover them. The data can be gigabyte to terabyte in size. HDFS is easily portable from one platform to another. HDFS is built using the java language. Because of its portable nature it can be deployed on a wide range of machine.
HDFS supports a hierarchical file organization. It provides the same functioning as the other file system provide. such as creation, moving ,renaming etc. HDFS is Reliable because it reliably store very large file across machine in the form of clusters. Files are stored as a sequence of blocks. Replication increases the fault tolerance capacity. HDFS is robust in nature i.e. it store data reliably beside failure.
It provide data integrity i.e. the received data may be corrupted.HDFS provide checksum checking. Snapshots support storing a copy of data at a particular instant of time. The snapshot feature may be to roll back a corrupted HDFS instance.

Quantcast File System (QFS) is an efficient alternative to the Hadoop Distributed File System (HDFS). QFS iswritten in C++. It provides improvement in efficiency.As QFS works out of the box with Hadoop, it migrating data from HDFS to QFS.

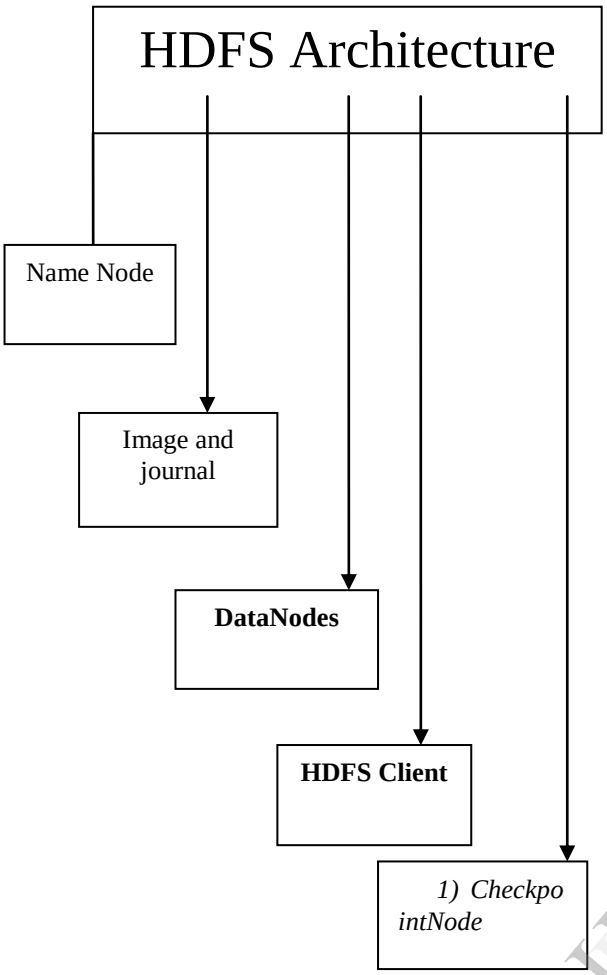
A. NameNode:

Namespace is used to make a hierarchy of file system. All related information is stored in RAM but periodically it goes to disk also.

B. DataNode:

The objective of data node is to store HDFS file data in local file system.

C. Architecture of HDFS:



DataNode plays following functionality:

(1) Replication of nodes
(2) Inform the nodes to shut down

D. HDFS client:

The objective of HDFS is to export the file system.Data is read by the help of Data Node.

III. HADOOP IN DIFFERENT WORKLOAD:

As we well know that hadoop is used for dataAnalyzer.Theycan writehadoopapplication, executethem and extract the information from their data.

The analyzer analyze following things:

(1) Application workload
(2) Tunings
(3) Resource usage and sharing

Hadoop cluster: It improves the effectiveness of authoring and efficiency of work loads,improve interactivity, better user education and automatic optimization.

Clusters are of 3 types:

- Open cloud
- M45
- Web-Mining

Open cloud: it supports 64 nodes. It is used by researcher.

M45: M45 are available by yahoo. It can have 400 nodes.

Web-Mining: web-mining can have 9 nodes i.e. it is small cluster owned.

The objective of HDFS is to export the file system.Data is read by the help of Data Node.
Hadoop are Batch oriented in nature.

When do user customize the execution of hadoop jobs?
Method: For each customization user supply an extra function to map and reduce.
Results: In open cloud 62% combiners are used while M45 used 43% ,Web-mining used 80% in map and reduce task.
Benefits: Job customization help in performance and also improves the correctness. Open cloud user use 2 other cluster.

Is hadoop is available to change the configuration of job?

Hadoop is do this job to improve performance .for doing this a no of user are performed for each type of turning at least once.

There can be following options:

If parameter is failed i.e if erroneous input given then how they will be controlled skiping in main quality.

JVM(Java Virtual Machine): it is a native hadoop map reduce interface is implemented in java.

Is it possible to handle big and small jobs?
Yes
Method: In this firstly aggregate input/output size of jobs submitted during the collection period.
Result:open cloud use small jobs comesthat are comes from twosources.Hadoop work as a scheduler here and datasets works as input.
Some short jobs are used for debugging purpose.massive scale data set is long duration jobs executed by hadoop. Large and long running job into smaller one as way to checkpoint intermediate result and ensure less work is lost when an entire jobs fails.This is done in web-mining.

IV. FUTURE WORK:

The main point in this paper is to big data analysis. Basically Map-Reduced designed for single job large input, long duration task.

Big data analysis is also a cumbersome task.In Hadoop system requires tools and novel debugging tools.

Another goal is optimization that can be used for disk block allocation and defragmentation algorithm,

V. CONCLUSION:

Hadoop represent an increasingly important approach for data-intensive computing. This paper explore through the components of the Hadoop system, HDFS . It also successfully pointed out the architecture of HDFS, distribution of data across the cluster based on the client applications. In this paper we have study the workload on clusters. But we find that application are highly diverse in style and structure.

REFERENCES :

1) Shvachko K, Kuang H, Radia S and Chansler R. The Hadoop Distributed File System in Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies, 2010.
2) Shafer J, Rixner S, Cox AL. The Hadoop Distributed Filesystem: Balancing Portability and Performance, in Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2010), White Plains, NY, 2010.
3) Feng Wang et al. Hadoop High Availability through Metadata Replication, IBM China Research Laboratory, ACM, 2009.