Hadoop notes:

3 versions :

Hadoop 1.x

==========

HDFS : (hadoop distributed file system) : Data Stroage

MapReduce : Data Processing + Resource management

Hadoop 2.x :

==========

HDFS : Data storage

MapReduce : Data Processing.

Yarn (Yet another resource negotiator) : Resource Manager

Hadoop(Cluster) Architecture : hadoop is a master and slave architecture.

==========

Cluster : more than 1 computer and it is ingreated with remaining all computers.

2 machines = 2 Node cluster.

4 Machines = 4 Node cluster.

10 Machines = 10 Node cluster.

100 Machines = 100 Node cluster.

Features of Hadoop :

==========

1. Reliabe : Handle failure.(Data Replication)

2. Flexible : Add more systems with out down time.

3. Economical : Commericial H/W used in cheap

4. Stable : Reliabe + consistency of the system. It will work with out any up expected error/failures.

HDFS :

==========

1. It is designed to store and manage large datasets/ files across cluster.

2. It is core component of the hadoop eco system.

3. It is responsible for providing reliable and fault tolarance storage for big data applications.

Components :

==========

Name Node : Name Node stores only metadata information

Data Node : Data Node stores actual data.

Secondary Name Node : In hadoop, the scondary name node is a helper node for the name node.

fsimage : It is file that contains a snapshot of metadata information stored in the namenode.

edit logs : After latest fsimage snapshot, Changes infromation is stored in edit logs.

Diffrence between data and metadata :

==========

image.jpg = 10 MB

Atual data size is 10 MB

Metadata : data about data.

file_name : image

file_type : JPG

file_size : 10mb

storage_location : /pictures/image.jpg

==========

blocks : Any kind of data is stored in block wise in HDFS. (hdfs-site.xml)

==========

Hard disk : 4KB

HDFS : 128 mb

Image.jpg : 10MB

Replication : property in configuration file : 3. It is possible to increase or decrease. (hdfs-site.xml)

==========

Heart beats :

==========

1. Data node sends heartbeats to NameNode every 3 seconds.Then NameNode knows that data nodes are available.

We can connect HDFS storage two ways :

---

1. using cli hdfs commands.
2. web browser.

HDFS Commands :

==========

ls

hdfs dfs -ls /user/cloudera/

mkdir honey

hdfs dfs -mkdir /user/cloudera/honey

rm -r honey

hdfs dfs -rm -r /user/cloudera/honey

chmod 777 honey

hdfs dfs -chmod 777 /user/cloudera/honey

Copy :

cp linux_source_location linux_target_location

hdfs dfs -put local_file_system(lfs) hadoop_distributed_file_system(hdfs)

hdfs dfs -copyFromLocal local_file_system(lfs) hadoop_distributed_file_system(hdfs)

hdfs dfs -get hdfs_location lfs_location

hdfs dfs -copyToLocal hdfs_location lfs_location

hdfs location -> hdfs another location.

hdfs dfs -cp /user/cloudera/test_hdfs.txt /user/

hdfs dfs -mv /user/test_hdfs.txt /user/hdfs/

file delete :

hdfs dfs -rm /user/cloudera/test_hdfs.txt

hdfs dfs -cat /user/cloudera/test_hdfs.txt

hdfs dfs -chgrp hadoop /user/cloudera/test_hdfs.txt

hdfs dfs -chown mapred /user/cloudera/test_hdfs.txt

Hadoop architecture :

hdfs architecture : hdfs commands

MapReduce architecture :

hdfs architecture : - Data storage

==========

Name Node : Metadata information.

Data Node : Actual information.

MapReduce Architecture : - Data Processing

==========

Job tracker :

==========

Job tracker is a key component of hadoop MapReduce Engine that manages and montiors the processing of jobs submitted to the HDFS.

It is responsible for accepting jobs from client,

scheduling jobs,

Monitoring task progress and

Managing the overall execution of jobs.

Task tracker :

==========

Task tracker is a hadoop MapReduce engine that run on individual slave nodes in hadoop cluster.

When job submitted to the hadoop cluster, Job tracker divides it into smaller tasks and assignes them to different task trackers.

Each task tracker is responsible for executing the tasks assgined to it and reporting the progress back to job tracker.

3 Node cluster :

1 - master node : job tracker -

2 - slave nodes : task tracker + task tracker

MapReduce :

==========

Map Phase and Reduce Phase

Map Phase : It will make Key-Value pairs

==========

-> Data is split into smaller chunks and process in parallely across multiple nodes in a cluster.

-> Each node applies a map function to the data, Which trasforms into Key-Value pairs.

Reduce Phase :

==========

-> The key-value pairs are grouped and procesed in parallel across multiple nodes in the cluster.

-> Each node applies a reduce function to the grouped data, Which aggregates the values of each key.

Features :

==========

Scalibility : We can add more systems with out downtime.

Cost effective : Commodity H/W is cheap.

Flexible : Java, Scala, Python and R. (txt, xls, jpg, video, audio)

fast : Prallel processing.

High availibility : Name Node fails ---> Secondary Name Node.