HDFS COMMANDS

Hadoop is bigdata tool that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

Hadoop has two main components;

1.HDFS

2. Hadoop MapReduce

HDFS (Hadoop Distributed File System)

HDFS splits files into blocks and sends them across various nodes. Also, in case of a node failure, the system operates and data transfer takes place between the nodes which are facilitated by HDFS.
HDFS comprising of 3 nodes:

1.Name Node (NN)

- Master node in HDFS
- NN stores metadata (Information about data)
- Only one NN under a HDFS

2.Data Node (DN)

- Data stored in DN
- Data stored in blocks of size 64 Mb in Hadoop and 128 Mb in case of yarn.
- Multiple number of DN will be there.

3. Secondary Name Node (SNN)

- Backup data is stored at SNN.

Hadoop MapReducing

MapReduce is an algorithm used for processing/analysing data in Hadoop. Mapreducing also is a master slave model. It has two components;

1.Job Tracker (JT)

- Assigns job to Task Tracker.
- Only single JT is present.
- Here, master is JT.

2.Task Tracker (TT)

- TT completes the task assigned by JT.
- There are multiple number of TT are present.

NN, DN, SNN, JT and TT are known as five daemons of Hadoop. Hadoop itself is primarily implemented in Java, but can use other tools and languages in the Hadoop ecosystem like Pig and Hive to process big data without needing extensive knowledge of Java.

HDFS Commands

HDFS is the primary or major component of the Hadoop ecosystem which is responsible for storing large data sets of structured or unstructured data across various nodes. Hadoop Distributed File System (HDFS) commands are used to interact with and manage files and directories stored in Hadoop's distributed file system.

1. To start HDFS Daemons

start-all.sh

2. To stop HDFS Daemons

stop-all.sh

3. To start processing Daemons only

start-yarn.sh

4. To stop Processing Daemons

stop-yarn.sh

5. To start HDFS Daemons only

start-dfs.sh

6. To stop HDF Daemons only

stop-dfs.sh

7. To verify whether the daemons are started or not

jps

jps: Java Processing Status tool

8. Individually Start Daemons

For HDFS:

hadoop-daemons.sh start <node>

For yarn:

y*arn-daemon.sh start <rm/nm>*

RM: Resource Manager

NM: Node Manager

9. To check Properties of Hadoop

hadoop fsck - /

fsck: File System Check

10. To list contents

hadoop fs -ls /

10. To create directory

hadoop fs -mkdir /directoryname

11. To create subdirectories inside a directory

hadoop fs -mkdir -p /dir.1/dir.2

12. To create file

hadoop fs -touchz /filename

13. To count directories, files and size

hadoop fs -count /

14. To delete a file

hadoop fs -rm /filename

15. To delete directory

hadoop -rm -r /directory

16. To copy file/directory

(i) Copy from HDFS to HDFS

hadoop fs -cp <source path> <Destination path>

(ii) Copy from local machine to HDFS

hadoop fs -copyFromLocal <source path> <Destination Path>

Or

hadoop fs -put <sourcepath> <Destination path>

(iii) Copy from HDFS to local

hadoop fs -copyToLocal <source path> <Destination Path>

17. To view data from HDFS file in terminal

hadoop fs -cat /file path

or

Hadoop fs -get <Source path> <Destination Path>

18. Tail in HDFS files

Hadoop fs -tail /file path