**HDFS and HDFS Commands**

What is HDFS?

HDFS, or Hadoop Distributed File System, is a distributed file system designed to store large volumes of data reliably and efficiently across multiple machines in a Hadoop cluster. In simple terms, it's like a huge virtual storage space spread across many computers.

Here's how it works in basic terms:

1. **Distributed Storage**: HDFS breaks large files into smaller blocks and stores these blocks across multiple machines (or nodes) in a cluster. This distribution ensures that if one machine fails, the data can still be accessed from other machines.
2. **Replication**: HDFS replicates each block of data multiple times across different nodes in the cluster to ensure fault tolerance. By default, it replicates each block three times, but this can be configured based on the needs of the system.
3. **Data Processing**: HDFS works closely with Hadoop's MapReduce framework, allowing data to be processed in parallel across the cluster. This means that large-scale data analysis tasks can be divided into smaller tasks and executed concurrently on different nodes, significantly speeding up the processing time.

**HDFS Commands**

**1: It shows version of hdfs**

**hdfs version**

2: Create a hdfs directory

hdfs dfs -mkdir proj1

3: Create a file sample1.txt and write some content and save it

in the current directory in your local file system. Copy the

sample1.txt to the proj1 directory of hdfs

hdfs dfs -put sample1.txt proj1

4: List the contents of proj1 directory

hdfs dfs -ls proj1

5: Check the content of file sample1.txt from proj1 directory.

hdfs dfs -cat proj1/sample1.txt

6: Copy the hdfs file(sample1.txt) from proj1 directory to proj2 directory in hdfs

hdfs dfs -cp proj1/sample1.txt proj2

7: Returns all the available files and recursively lists all the

subdirectories under HDFS home directory

hdfs dfs -ls -R <HDFS Home dir>

8: Delete the sample1.txt file from proj1 directory of hdfs

hdfs dfs –rm proj1/sample1.txt

9: Move the sample1.txt file from proj2 directory of hdfs to

proj1 of hdfs

hdfs dfs -mv proj2/sample1.txt proj1

10: Change the permissions of the file1.txt in proj1 directory to give complete access

hdfs dfs -chmod 777 proj1/sample1.txt

11: Create sample file in hdfs and put it to local machine

hdfs dfs -get proj1/sample1.txt data/sample-local.txt

12: Remove the proj1 directory

hdfs dfs -rm -r proj1

13: Create multiple directories with a single command in hdfs

hdfs dfs -mkdir maindir/subdir

hdfs dfs -mkdir -p maindir/subdir

hdfs dfs -mkdir -p proj1/module1

14: Change the permissions of the subdir directory to 755

hdfs dfs -chmod 755 proj1/module1

15: Displays the disk usage for all the files available under a

given directory.

hdfs dfs -du /

16: Displays disk usage of current hadoop distributed file

system.

hdfs dfs -df

17: This hadoop command will show the last records of file.

hdfs dfs -tail <hdfs Path>

18: Create an empty file at the specified location.

hdfs dfs -touchz <hdfs Path>/newfile.txt

19: This hadoop command is used to set the replication for a

specific file.

hdfs dfs -setrep -w 1 /user/bigdata/Sample1.txt

20: This hadoop command is basically used to change the group name

hdfs dfs -chgrp -R hadoop /tmp

21: This hadoop command is used to change the owner

hdfs dfs -chown -R hadoop /tmp

22: Returns the checksum information of a file.

hdfs dfs -checksum output/part-r-00000

23: This command used to check the health of file

hdfs fsck proj1/sample1.txt

24: Count the number of directories, files, and bytes under the

paths that match the specified file pattern.

hdfs dfs -count /user/ubuntu

25: Finds all files that match the specified expression with -name.

hdfs dfs -find [path to search]-name <file to search>