SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

COURSE NAME: Big Data Infrastructure **CLASS:** Third Year Beach

NAME: Falguni Parmar **BATCH:** C22

Experiment - 3

Theory:

DFS stands for the distributed file system, it is a concept of storing the file in multiple nodes in a distributed manner. DFS actually provides the Abstraction for a single large system whose storage is equal to the sum of storage of other nodes in a cluster.

Let's understand this with an example. Suppose you have a DFS comprises of 4 different machines each of size 10TB in that case you can store let say 30TB across this DFS as it

provides you a combined Machine of size 40TB. The 30TB data is distributed among these Nodes in form of Blocks.

Overview – HDFS

Now we think you become familiar with the term file system so let's begin with HDFS. HDFS (Hadoop Distributed File System) is utilized for storage permission is a Hadoop cluster. It mainly designed for working on commodity Hardware devices(devices that are inexpensive), working on a distributed file system design. HDFS is designed in such a way that it believes more in storing the data in a large chunk of blocks rather than storing small data blocks. HDFS in Hadoop provides Fault-tolerance and High availability to the storage layer and the other devices present in that Hadoop cluster.

HDFS is capable of handling larger size data with high volume velocity and variety makes Hadoop work more efficient and reliable with easy access to all its components. HDFS stores the data in the form of the block where the size of each data block is 128MB in size which is configurable means you can change it according to your requirement in hdfs-site.xml file in your Hadoop directory.

Some Important Features of HDFS (Hadoop Distributed File System)

- It's easy to access the files stored in HDFS.
- HDFS also provides high availability and fault tolerance.
- Provides scalability to scaleup or scaledown nodes as per our requirement.
- Data is stored in distributed manner i.e. various Datanodes are responsible for storing the data.
- HDFS provides Replication because of which no fear of Data Loss.
- HDFS Provides High Reliability as it can store data in a large range of Petabytes.
- HDFS has in-built servers in Name node and Data Node that helps them to easily retrieve the cluster information.

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

- Provides high throughput.

Code:

1. ls: this command is used to list all the files. Use ls for recursive approach. It is used when we want a hierarchy of folder.

```
[root@sandbox ~]# hdfs dfs -ls /
Found 12 items
drwxrwxrwx   - yarn   hadoop          0 2016-10-25 08:10 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:54 /apps
drwxr-xr-x   - yarn   hadoop          0 2016-10-25 07:48 /ats
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 08:01 /demo
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:48 /hdp
drwxr-xr-x   - mapred hdfs            0 2016-10-25 07:48 /mapred
drwxrwxrwx   - mapred hadoop          0 2016-10-25 07:48 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:47 /ranger
drwxrwxrwx   - spark  hadoop          0 2023-04-05 05:04 /spark-history
drwxrwxrwx   - spark  hadoop          0 2016-10-25 08:14 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2016-10-25 08:11 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 08:11 /user
[root@sandbox ~]#
```

2. mkdir: To create a directory. In Hadoop dfs there is no home directory by default.

```
[root@sandbox ~]# hdfs dfs -mkdir /god_user
[root@sandbox ~]# hdfs dfs -ls /
Found 13 items
drwxrwxrwx   - yarn   hadoop          0 2016-10-25 08:10 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:54 /apps
drwxr-xr-x   - yarn   hadoop          0 2016-10-25 07:48 /ats
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 08:01 /demo
drwxr-xr-x   - root   hdfs            0 2023-04-05 05:33 /god_user
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:48 /hdp
drwxr-xr-x   - mapred hdfs            0 2016-10-25 07:48 /mapred
drwxrwxrwx   - mapred hadoop          0 2016-10-25 07:48 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:47 /ranger
drwxrwxrwx   - spark  hadoop          0 2023-04-05 05:33 /spark-history
drwxrwxrwx   - spark  hadoop          0 2016-10-25 08:14 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2016-10-25 08:11 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 08:11 /user
[root@sandbox ~]#
```

3. touchz: It create an empty file.

```
[root@sandbox ~]# hdfs dfs -touchz /god_user/myfile.txt
[root@sandbox ~]# hdfs dfs -ls /god_user
Found 1 items
-rw-r--r--   1 root hdfs          0 2023-04-05 05:35 /god_user/myfile.txt
[root@sandbox ~]#
```

4. copyFromLocal (or) put: To copy files/folders form local files system to hdfs store. This is the most important command. Local filesystem means the files present on the OS.

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
[root@sandbox ~]# hdfs dfs -put Sample.txt /god_user
[root@sandbox ~]# hdfs dfs -ls /god_user
Found 2 items
-rw-r--r--   1 root hdfs          0 2023-04-05 05:48 /god_user/Sample.txt
-rw-r--r--   1 root hdfs          0 2023-04-05 05:35 /god_user/myfile.txt
[root@sandbox ~]#
```

5. cat: To print file contents.

```
[root@sandbox ~]# hdfs dfs -cat /god_user/Sample.txt
Y0000 LESG00000
```

6. copyToLocal (or) get: To copy files/folder from hdfs store to local file system.

```
[root@sandbox ~]# hdfs dfs -get /god_user/myfile.txt
[root@sandbox ~]# dir
anaconda-ks.cfg  blueprint.json  build.out  hdp  hello.txt  install.log  install.log.syslog  myfile.txt
[root@sandbox ~]#
```

7. cp: This command is used to copy files within hdfs.

```
[root@sandbox ~]# hdfs dfs -mkdir /god_user_copied
[root@sandbox ~]# hdfs dfs -cp /god_user /god_user_copied
[root@sandbox ~]# hdfs dfs -ls /god_user_copied
Found 1 items
drwxr-xr-x   - root hdfs          0 2023-04-05 06:00 /god_user_copied/god_user
```

8. mv: This command is used to move files within hdfs.

```
[root@sandbox ~]# hdfs dfs -ls /god_user
Found 2 items
-rw-r--r--   1 root hdfs         16 2023-04-05 05:53 /god_user/Sample.txt
-rw-r--r--   1 root hdfs          0 2023-04-05 05:35 /god_user/myfile.txt
[root@sandbox ~]# hdfs dfs -ls /god_user_copied
Found 1 items
drwxr-xr-x   - root hdfs          0 2023-04-05 06:00 /god_user_copied/god_user
[root@sandbox ~]# hdfs dfs -ls /god_user_copied/god_user
Found 2 items
-rw-r--r--   1 root hdfs         16 2023-04-05 06:00 /god_user_copied/god_user/Sample.txt
-rw-r--r--   1 root hdfs          0 2023-04-05 06:00 /god_user_copied/god_user/myfile.txt
[root@sandbox ~]# hdfs dfs -mv /god_user/myfile.txt /god_user_copied
[root@sandbox ~]# hdfs dfs -ls /god_user_copied
Found 2 items
drwxr-xr-x   - root hdfs          0 2023-04-05 06:00 /god_user_copied/god_user
-rw-r--r--   1 root hdfs          0 2023-04-05 05:35 /god_user_copied/myfile.txt
[root@sandbox ~]#
```

9. rmr: This command deletes a file from HDFS recursively. It is very useful command when you want to delete a non-empty directory.

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
[root@sandbox ~]# hdfs dfs -rm -r /god_user_copied
23/04/05 06:05:37 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox.hortonworks.com:8020/god_u
[root@sandbox ~]# hdfs dfs -ls /
Found 13 items
drwxrwxrwx   - yarn   hadoop          0 2016-10-25 08:10 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:54 /apps
drwxr-xr-x   - yarn   hadoop          0 2016-10-25 07:48 /ats
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 08:01 /demo
drwxr-xr-x   - root   hdfs            0 2023-04-05 06:03 /god_user
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:48 /hdp
drwxr-xr-x   - mapred hdfs            0 2016-10-25 07:48 /mapred
drwxrwxrwx   - mapred hadoop          0 2016-10-25 07:48 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2016-10-25 07:47 /ranger
drwxrwxrwx   - spark  hadoop          0 2023-04-05 06:05 /spark-history
drwxrwxrwx   - spark  hadoop          0 2016-10-25 08:14 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2016-10-25 08:11 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2023-04-05 05:52 /user
[root@sandbox ~]#
```

10. du: it will give the size of each file in directory.

```
[root@sandbox ~]# hdfs dfs -du /god_user
16  /god_user/Sample.txt
[root@sandbox ~]#
```

11. dus: This command will give the size of directory/file.

```
[root@sandbox ~]# hdfs dfs -dus /god_user
dus: DEPRECATED: Please use 'du -s' instead.
16  /god_user
[root@sandbox ~]# hdfs dfs -dus /user
dus: DEPRECATED: Please use 'du -s' instead.
688241023  /user
[root@sandbox ~]#
```

12. stat: It will give the last modified time of directory or path. In short it will give stats of the directory or file.

```
[root@sandbox ~]# hdfs dfs -stat /user
2023-04-05 05:52:47
[root@sandbox ~]#
```

13. setrep: This command is used to change the replication factor of a file/directory in HDFS. By default, it is 3 for anything which is stored in HDFS (as set in hdfs coresite.xml)

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
[root@sandbox ~]# hdfs dfs -ls /user
Found 14 items
drwxr-xr-x   - admin      hdfs          0 2016-10-25 08:11 /user/admin
drwxrwx---   - ambari-qa  hdfs          0 2016-10-25 07:47 /user/ambari-qa
drwxr-xr-x   - amy_ds     hdfs          0 2016-10-25 08:02 /user/amy_ds
drwxr-xr-x   - hbase      hdfs          0 2016-10-25 07:48 /user/hbase
drwxr-xr-x   - hcat       hdfs          0 2016-10-25 07:51 /user/hcat
drwxr-xr-x   - hive       hdfs          0 2016-10-25 08:10 /user/hive
drwxr-xr-x   - holger_gov hdfs          0 2016-10-25 08:03 /user/holger_gov
drwxrwxr-x   - livy       hdfs          0 2016-10-25 07:49 /user/livy
drwxr-xr-x   - maria_dev  hdfs          0 2016-10-25 07:58 /user/maria_dev
drwxrwxr-x   - oozie      hdfs          0 2016-10-25 07:52 /user/oozie
drwxr-xr-x   - raj_ops    hdfs          0 2016-10-25 08:04 /user/raj_ops
drwx------   - root       hdfs          0 2023-04-05 05:52 /user/root
drwxrwxr-x   - spark      hdfs          0 2016-10-25 07:48 /user/spark
drwxr-xr-x   - zeppelin   hdfs          0 2016-10-25 07:50 /user/zeppelin
[root@sandbox ~]#
```

Conclusion:

We successfully implemented 13 HDFS commands. We understood various commands and use cases along with its implementation.