Experiment 1

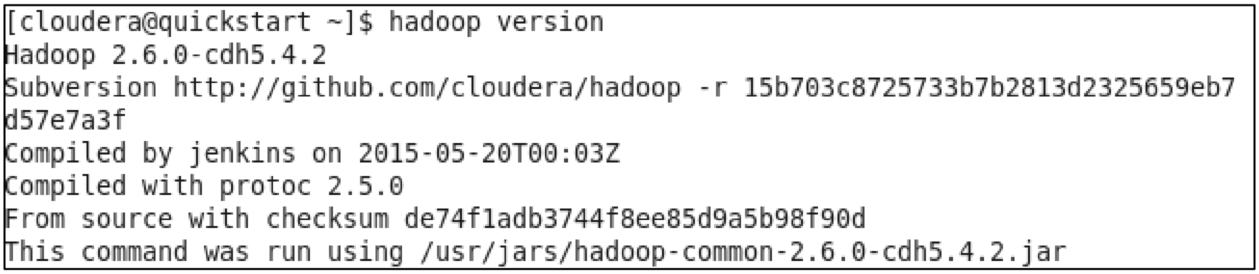
Aim:- To execute the HDFS Commands

Theory:- Write about the basics of HDFS.

The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.HDFS enables the rapid transfer of data between compute nodes. At its outset, it was closely coupled with MapReduce, a framework for data processing that filters and divides up work among the nodes in a cluster, and it organizes and condenses the results into a cohesive answer to a query. Similarly, when HDFS takes in data, it breaks the information down into separate blocks and distributes them to different nodes in a cluster. HDFS uses a master/slave architecture. The HDFS cluster's NameNode is the primary server that manages the file system namespace and controls client access to files. As the central component of the Hadoop Distributed File System, the NameNode maintains and manages the file system namespace and provides clients with the right access permissions. The system's DataNodes manage the storage that's attached to the nodes they run on.

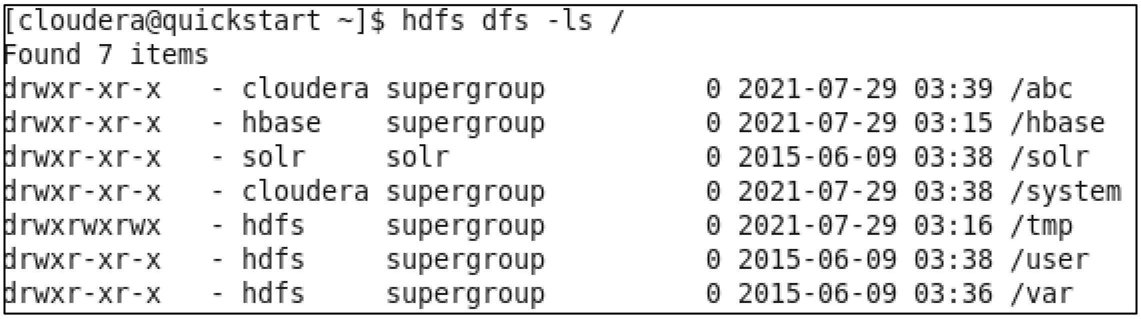
HDFS Commands

1. To display the version of hadoop use
   Hadoop version



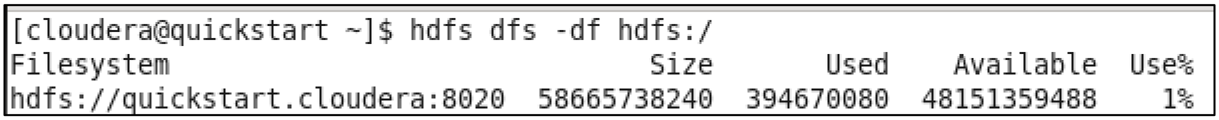
```
[cloudera@quickstart ~]$ hadoop version
Hadoop 2.6.0-cdh5.4.2
Subversion http://github.com/cloudera/hadoop -r 15b703c8725733b7b2813d2325659eb7
d57e7a3f
Compiled by jenkins on 2015-05-20T00:03Z
Compiled with protoc 2.5.0
From source with checksum de74f1adb3744f8ee85d9a5b98f90d
This command was run using /usr/jars/hadoop-common-2.6.0-cdh5.4.2.jar
```

2. List the contents of the root directory in HDFS
   hdfs dfs –ls /



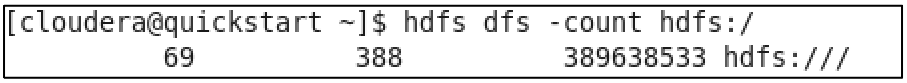
```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 03:39 /abc
drwxr-xr-x   - hbase    supergroup          0 2021-07-29 03:15 /hbase
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 03:38 /system
drwxrwxrwx   - hdfs     supergroup          0 2021-07-29 03:16 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
```

3. Report the amount of space used and available on currently mounted file system
   hdfs dfs -df hdfs:/



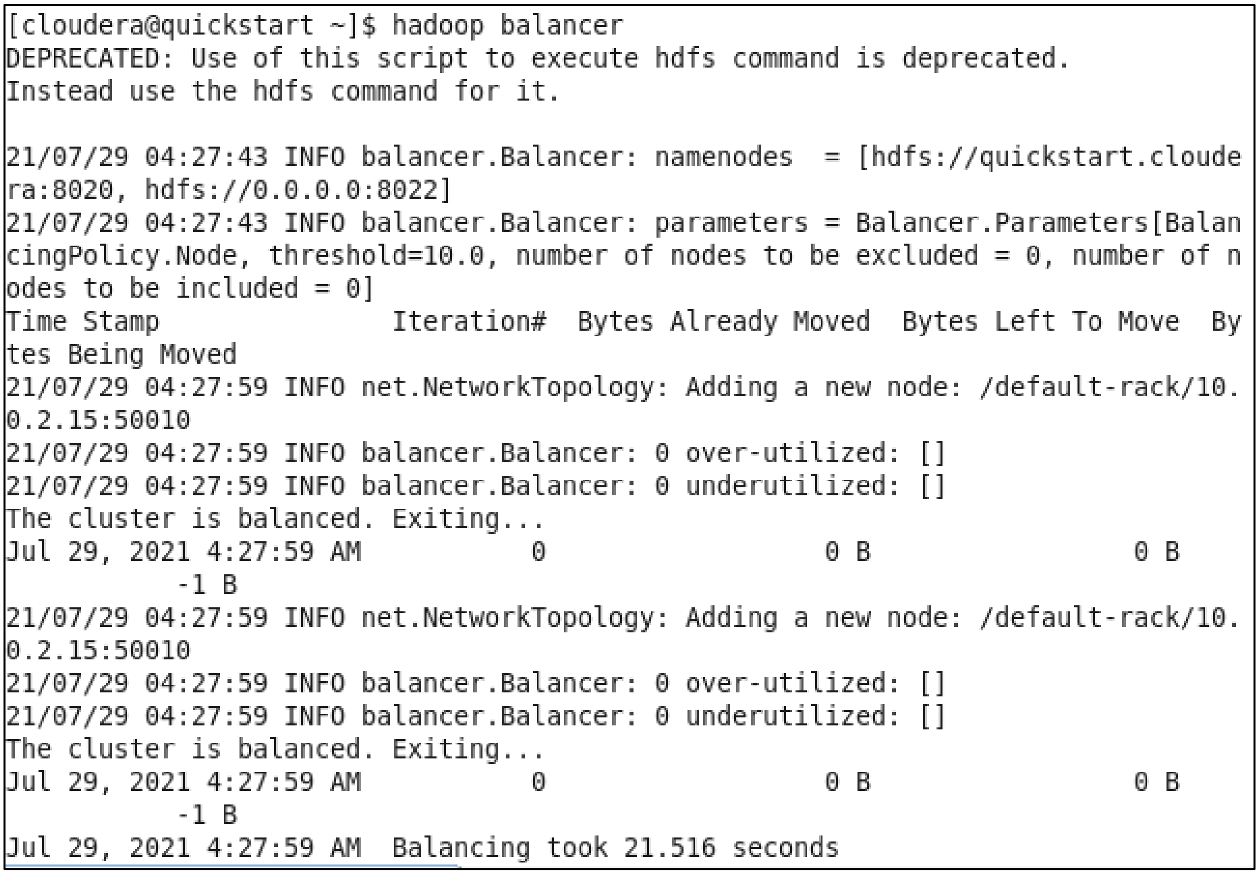
```
[cloudera@quickstart ~]$ hdfs dfs -df hdfs:/
Filesystem                              Size       Used    Available  Use%
hdfs://quickstart.cloudera:8020  58665738240  394670080  48151359488    1%
```

4. Count the number of directories, files and bytes under the paths that match the specified file pattern
   hdfs dfs -count hdfs:/



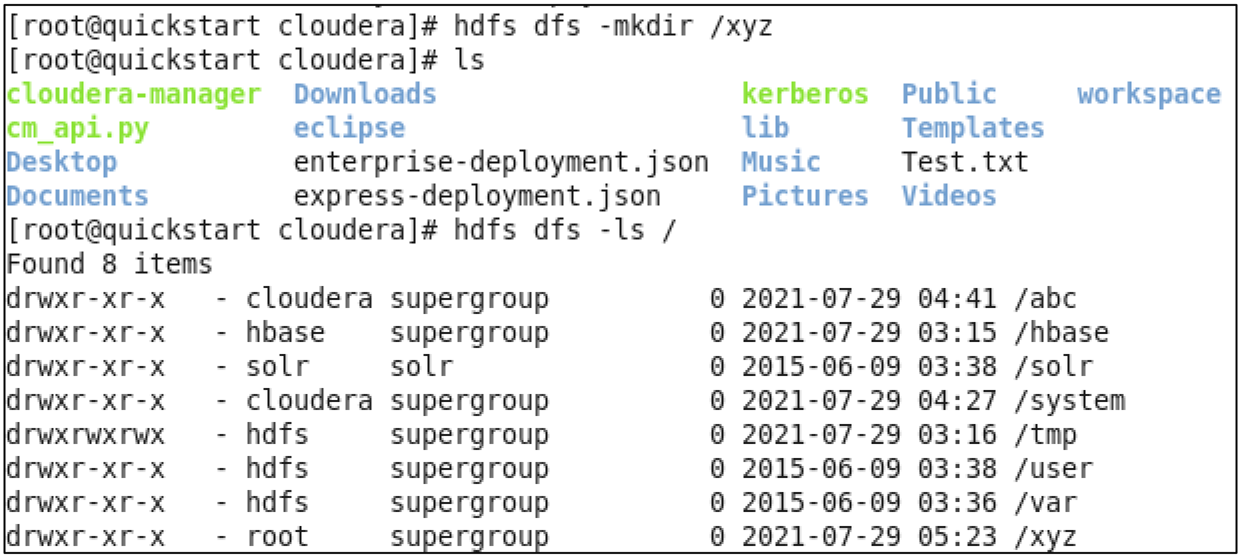
```
[cloudera@quickstart ~]$ hdfs dfs -count hdfs:/
          69          388          389638533 hdfs:///
```

5. Run a cluster balancing utility
   hadoop balancer



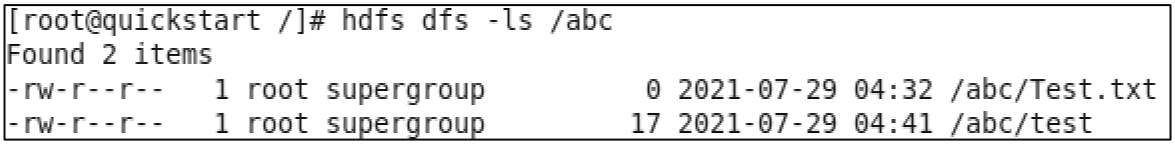
```
[cloudera@quickstart ~]$ hadoop balancer
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

21/07/29 04:27:43 INFO balancer.Balancer: namenodes  = [hdfs://quickstart.cloude
ra:8020, hdfs://0.0.0.0:8022]
21/07/29 04:27:43 INFO balancer.Balancer: parameters = Balancer.Parameters[Balan
cingPolicy.Node, threshold=10.0, number of nodes to be excluded = 0, number of n
odes to be included = 0]
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  By
tes Being Moved
21/07/29 04:27:59 INFO net.NetworkTopology: Adding a new node: /default-rack/10.
0.2.15:50010
21/07/29 04:27:59 INFO balancer.Balancer: 0 over-utilized: []
21/07/29 04:27:59 INFO balancer.Balancer: 0 underutilized: []
The cluster is balanced. Exiting...
Jul 29, 2021 4:27:59 AM           0                  0 B                 0 B
           -1 B
21/07/29 04:27:59 INFO net.NetworkTopology: Adding a new node: /default-rack/10.
0.2.15:50010
21/07/29 04:27:59 INFO balancer.Balancer: 0 over-utilized: []
21/07/29 04:27:59 INFO balancer.Balancer: 0 underutilized: []
The cluster is balanced. Exiting...
Jul 29, 2021 4:27:59 AM           0                  0 B                 0 B
           -1 B
Jul 29, 2021 4:27:59 AM  Balancing took 21.516 seconds
```

6. Creating a new directory
   hdfs dfs -mkdir /xyz



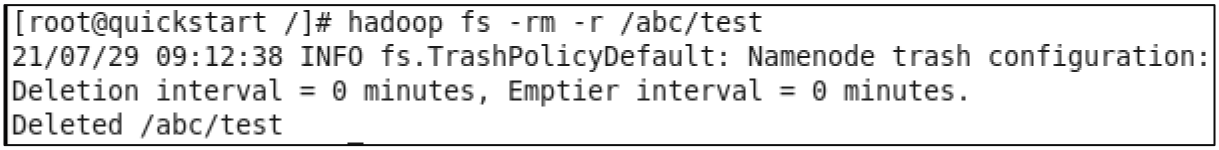
```
[root@quickstart cloudera]# hdfs dfs -mkdir /xyz
[root@quickstart cloudera]# ls
cloudera-manager  Downloads                   kerberos  Public     workspace
cm_api.py         eclipse                     lib       Templates
Desktop           enterprise-deployment.json  Music     Test.txt
Documents         express-deployment.json     Pictures  Videos
[root@quickstart cloudera]# hdfs dfs -ls /
Found 8 items
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 04:41 /abc
drwxr-xr-x   - hbase    supergroup          0 2021-07-29 03:15 /hbase
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 04:27 /system
drwxrwxrwx   - hdfs     supergroup          0 2021-07-29 03:16 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
drwxr-xr-x   - root     supergroup          0 2021-07-29 05:23 /xyz
```

7. To view the created directory
   root@quickstart:~$ hdfs dfs -ls abc



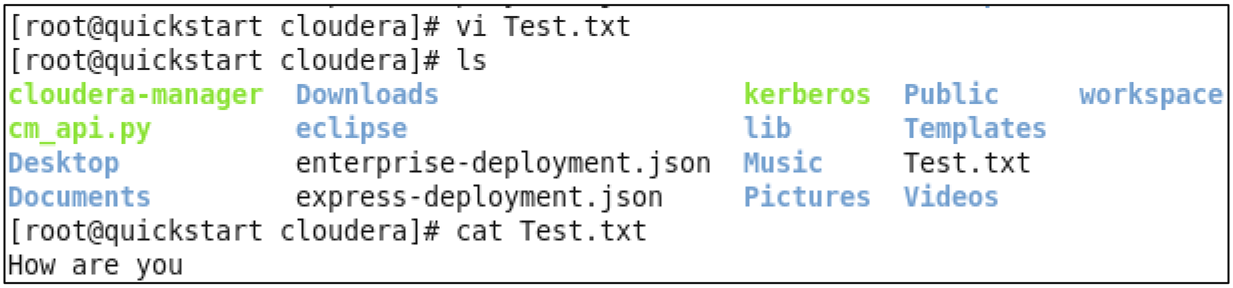
```
[root@quickstart /]# hdfs dfs -ls /abc
Found 2 items
-rw-r--r--   1 root supergroup          0 2021-07-29 04:32 /abc/Test.txt
-rw-r--r--   1 root supergroup         17 2021-07-29 04:41 /abc/test
```

8. To remove an already existing file use the command
   hadoop fs –rm - r abc/test



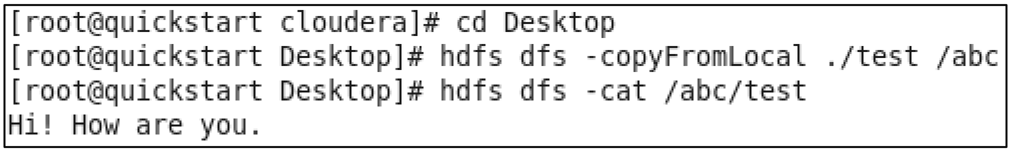
```
[root@quickstart /]# hadoop fs -rm -r /abc/test
21/07/29 09:12:38 INFO fs.TrashPolicyDefault: Namenode trash configuration:
Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /abc/test
```

9. To see the content of the file
   root@quickstart:~$ cat Test.txt



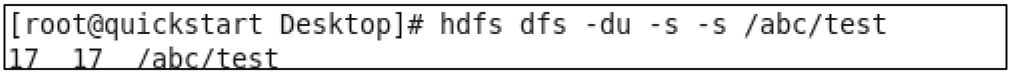
```
[root@quickstart cloudera]# vi Test.txt
[root@quickstart cloudera]# ls
cloudera-manager  Downloads                  kerberos  Public     workspace
cm_api.py         eclipse                    lib       Templates
Desktop           enterprise-deployment.json Music     Test.txt
Documents         express-deployment.json    Pictures  Videos
[root@quickstart cloudera]# cat Test.txt
How are you
```

10. To copy from the remote file of the local system to the hadoop distributed file system
    root@quickstart:~$ hdfs dfs -copyFromLocal ./test/abc



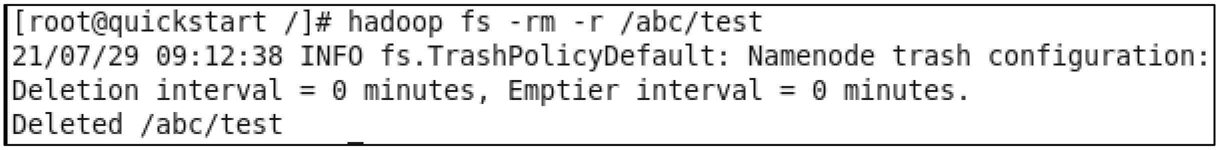
```
[root@quickstart cloudera]# cd Desktop
[root@quickstart Desktop]# hdfs dfs -copyFromLocal ./test /abc
[root@quickstart Desktop]# hdfs dfs -cat /abc/test
Hi! How are you.
```

11. To see how much space is occupied in HDFS.
    root@quickstart:~$ hdfs dfs -du -s -s abc/ABC



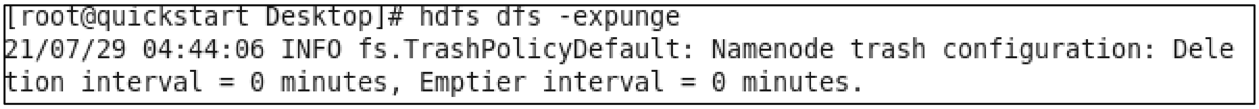
```
[root@quickstart Desktop]# hdfs dfs -du -s -s /abc/test
17  17  /abc/test
```

12. To delete a specific file in a specific folder
    root@quickstart:~$ hadoop fs -rm -r /abc/test



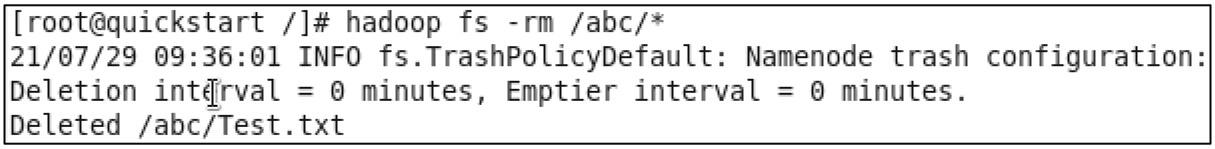
```
[root@quickstart /]# hadoop fs -rm -r /abc/test
21/07/29 09:12:38 INFO fs.TrashPolicyDefault: Namenode trash configuration:
Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /abc/test
```

13. To empty a trash the following command is used
root@quickstart:~$ hadoop dfs –expunge



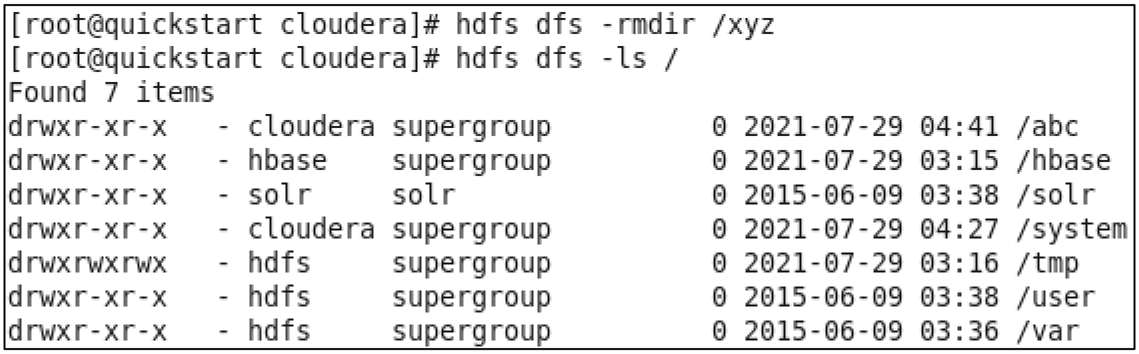
```
[root@quickstart Desktop]# hdfs dfs -expunge
21/07/29 04:44:06 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
```

14. To remove files within a directory
root@quickstart:~$ hadoop fs -rm /abc/*



```
[root@quickstart /]# hadoop fs -rm /abc/*
21/07/29 09:36:01 INFO fs.TrashPolicyDefault: Namenode trash configuration:
Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /abc/Test.txt
```

15. To remove a directory



```
[root@quickstart cloudera]# hdfs dfs -rmdir /xyz
[root@quickstart cloudera]# hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 04:41 /abc
drwxr-xr-x   - hbase    supergroup          0 2021-07-29 03:15 /hbase
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxr-xr-x   - cloudera supergroup          0 2021-07-29 04:27 /system
drwxrwxrwx   - hdfs     supergroup          0 2021-07-29 03:16 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
```

Conclusion:

Thus we learnt about HDFS and we have successfully downloaded and installed cloudera on a VM. We have also learn various commands like how to make or remove a directory, how to move and delete files etc and we have verified their outputs.