DATE:

Name: Prerna Sunil Jadhav
Sap ID: 60004220127
Batch : C22
Course: Big Data Infrastructure laboratory.
Course Code: DJ19CEEL6011

EXPERIMENT 03

AIM: Execute different HDFS commands.

THEORY:

→ The Hadoop Distributed file system (HDFS) is a distributed file system designed to run on commodity hardware.
→ It has many similarities with existing distributed file system. However, the differences from other distributed file systems are significant.
→ HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware.
→ HDFS provide high throughput access to application data and is suitable for the application that have large datasets.
→ HDFS relaxes a few POSIX requirements to enable streaming access to file system data
→ HDFS was originally built as infrastructure for the Apache Nutch web search engine project.

DATE:

COMMANDS:

1) ls: This command is used to list all the files. Use lsr for recursive approach. It is useful when we want a hierarchy of a folder.
⇒ hdfs dfs -ls /

2) mkdir:
To create a directory. In Hadoop dfs there is no home directory by default.
⇒ hdfs dfs -mkdir /user

3) touchz:
It creates an empty file.
⇒ hdfs dfs -touchz /user/myfile.txt

4) cat:
To print the file content
⇒ hdfs dfs -cat /user/sample.txt

5) cp:
This command is used to copy files within hdfs.
⇒ hdfs dfs -cp /user /user_copied.

6) mv:
This command is used to move files within hdfs
⇒ hdfs dfs -mv /user/myfile.txt /user_copied

7) rmr:

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

=> hdfs dfs -rmr /user-copied

8) du:

It will give the given size of each file in the directory

=> hdfs dfs -du /user

9) dus:

This command will give the total size of the directory / file

=> hdfs dfs -dus /user

10) stat:

It will give the last modified time of the directory or path.

=> hdfs dfs -stat /user

11) 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 core-site.xml).

=> hdfs dfs -setrep -R 4 /user.

DATE:

CONCLUSION:

Thus we have successfully executed different HDFS commands.

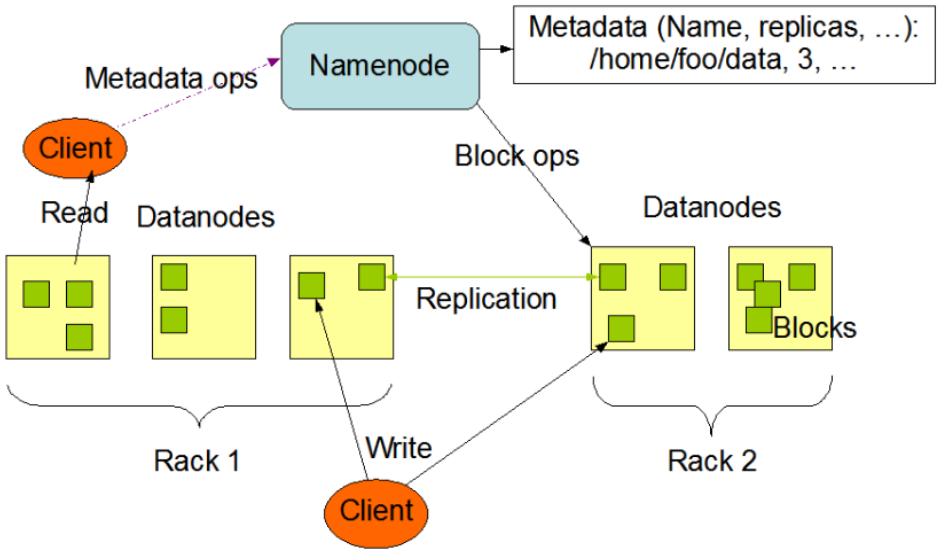
Name:	Prerna Sunil Jadhav
Sap Id:	60004220127
Class:	T. Y. B. Tech (Computer Engineering)
Course:	Big Data Infrastructure Laboratory
Course Code:	DJ19CEEL6011
Experiment No.:	03

AIM: Execute different HDFS Commands.

WHAT IS HDFS?

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project.

HDFS Architecture:

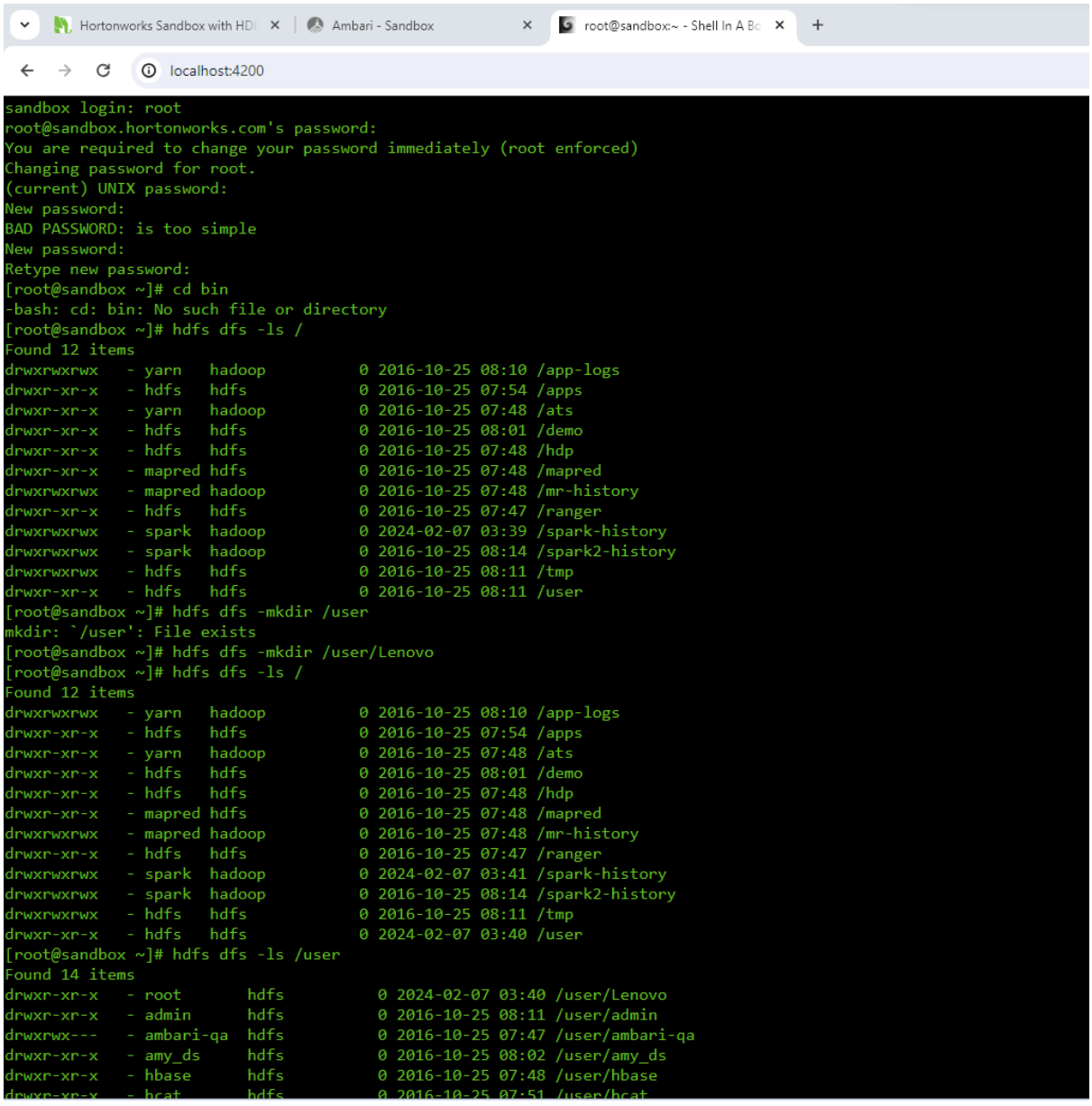


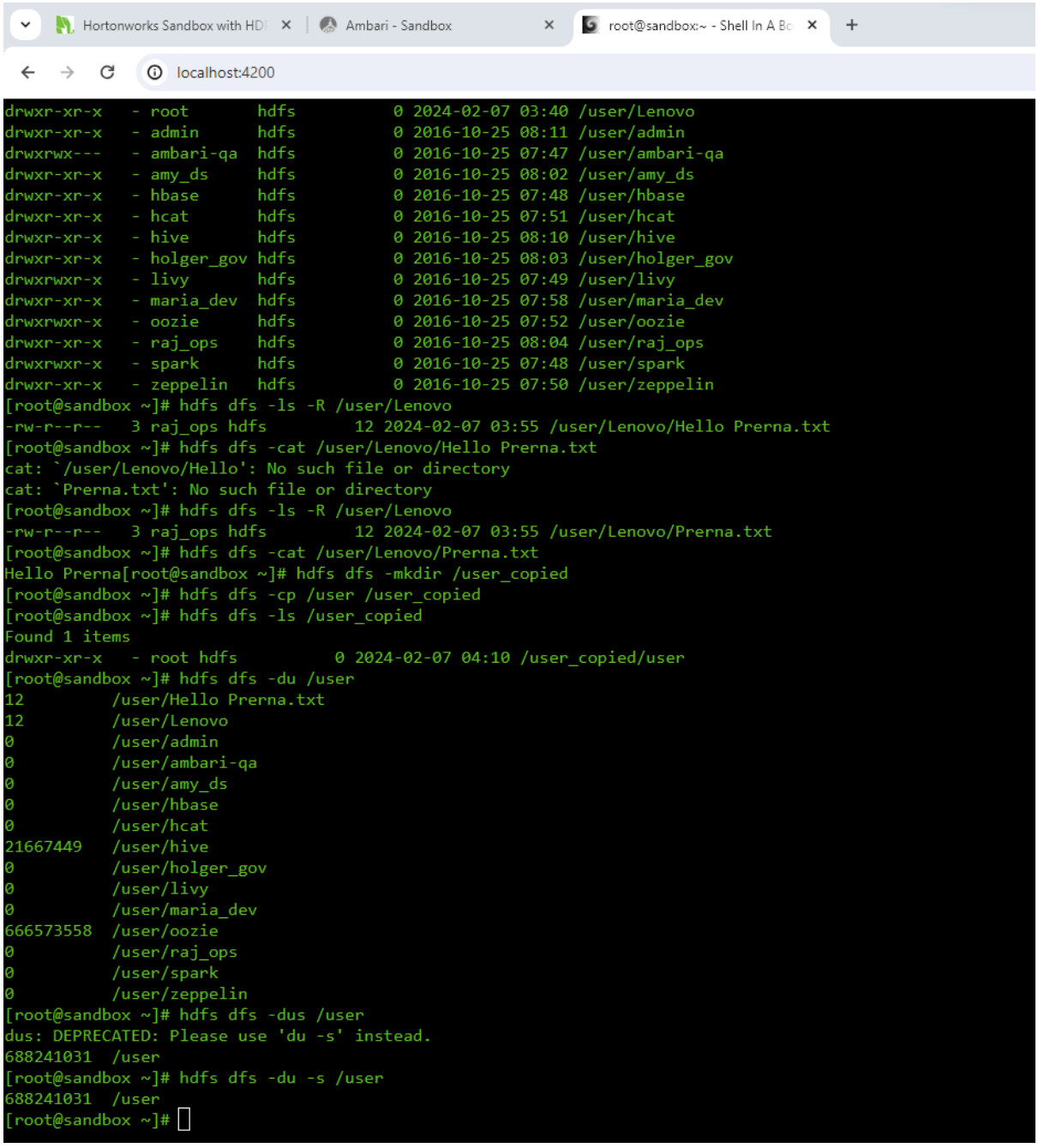
Commands:

- ***hadoop version***: The Hadoop fs shell command version prints the Hadoop version.
- ***hadoop fs –mkdir /path/directory_name:*** This command creates the directory in HDFS if it does not already exist. Use ***hadoop fs mkdir -p /path/directoryname***, so not to fail even if directory exists.
  Note: If the directory already exists in HDFS, then we will get an error message that file already exists.
- ***hadoop fs -ls /path:*** The Hadoop fs shell command ls displays a list of the contents of a directory specified in the path provided by the user. It shows the name, permissions, owner, size, and modification date for each file or directories in the specified directory.

- ***hadoop fs –cat /path_to_file_in_hdfs:*** The cat command reads the file in HDFS and displays the content of the file on console or stdout.
- ***hadoop fs -mv <src> <dest>:*** The HDFS mv command moves the files or directories from the source to a destination within HDFS.
- ***hadoop fs -cp <src> <dest>:*** The cp command copies a file from one directory to another directory within the HDFS.

Code & Output:





Advantages:

- Distributed data storage, High fault tolerance, Blocks reduce seek time.
- The data is highly available as the same block is present at multiple data-nodes.
- Even if multiple data-nodes are down we can still do our work, thus making it highly reliable.

Limitations: Though HDFS provide many features there are some areas where it doesn't work well.

- Low latency data access
- Small file problem