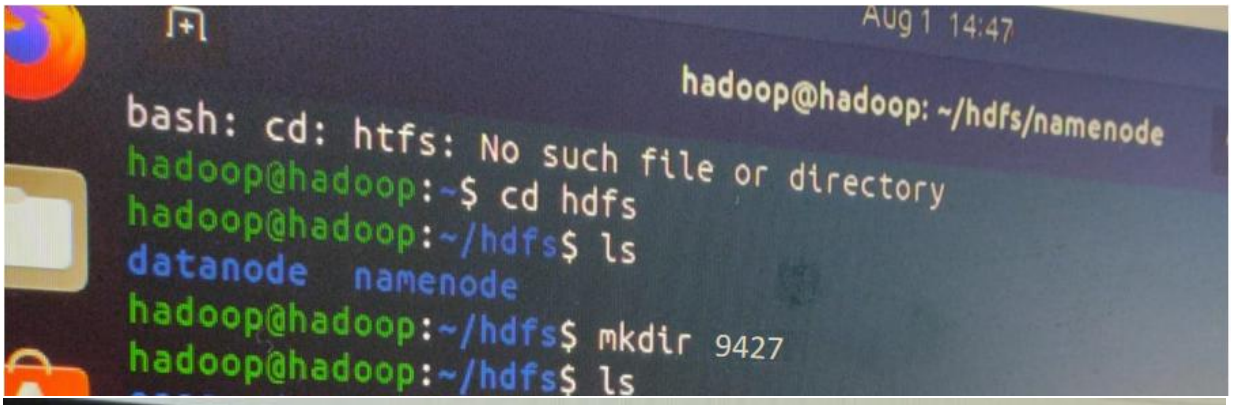
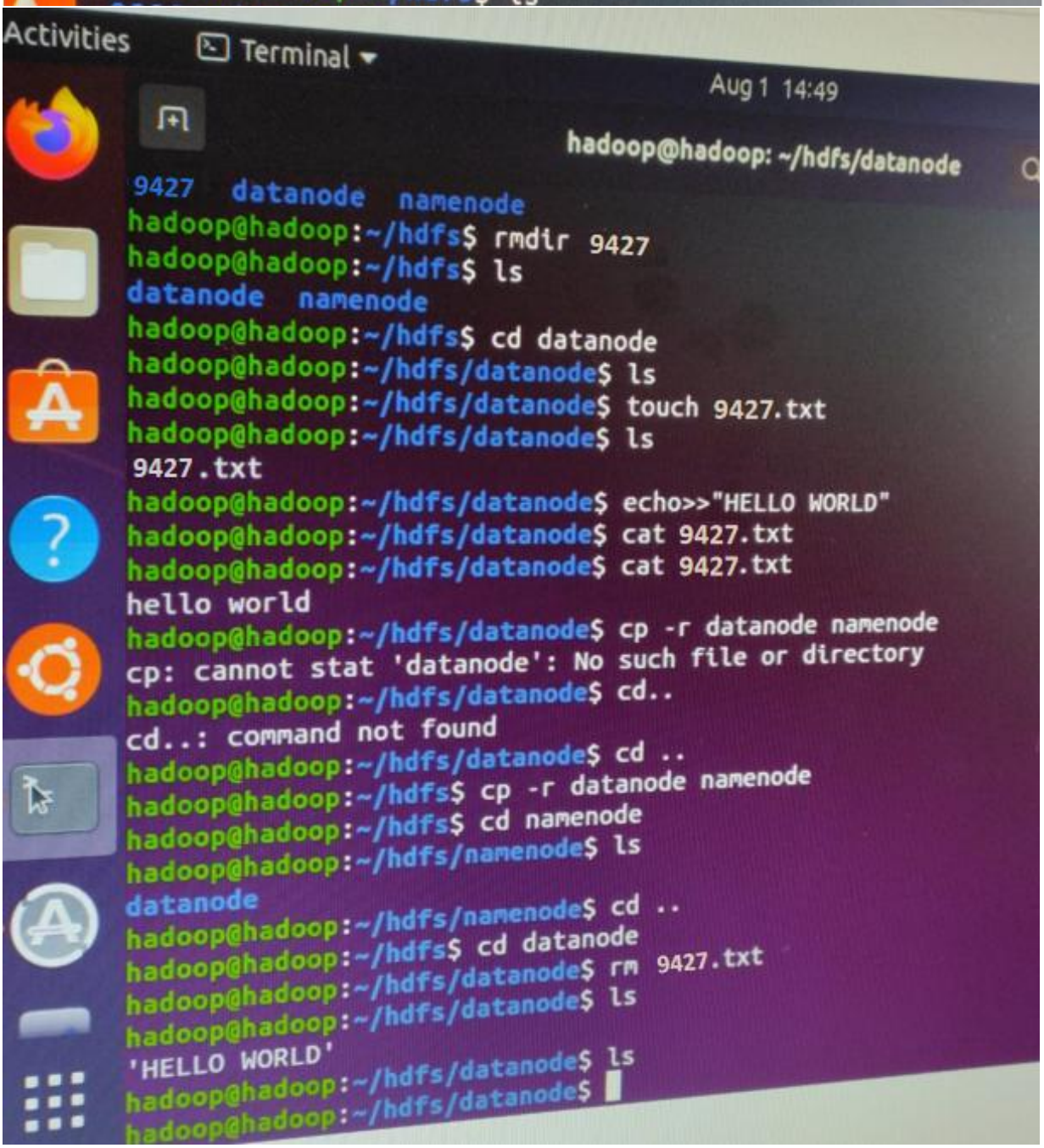
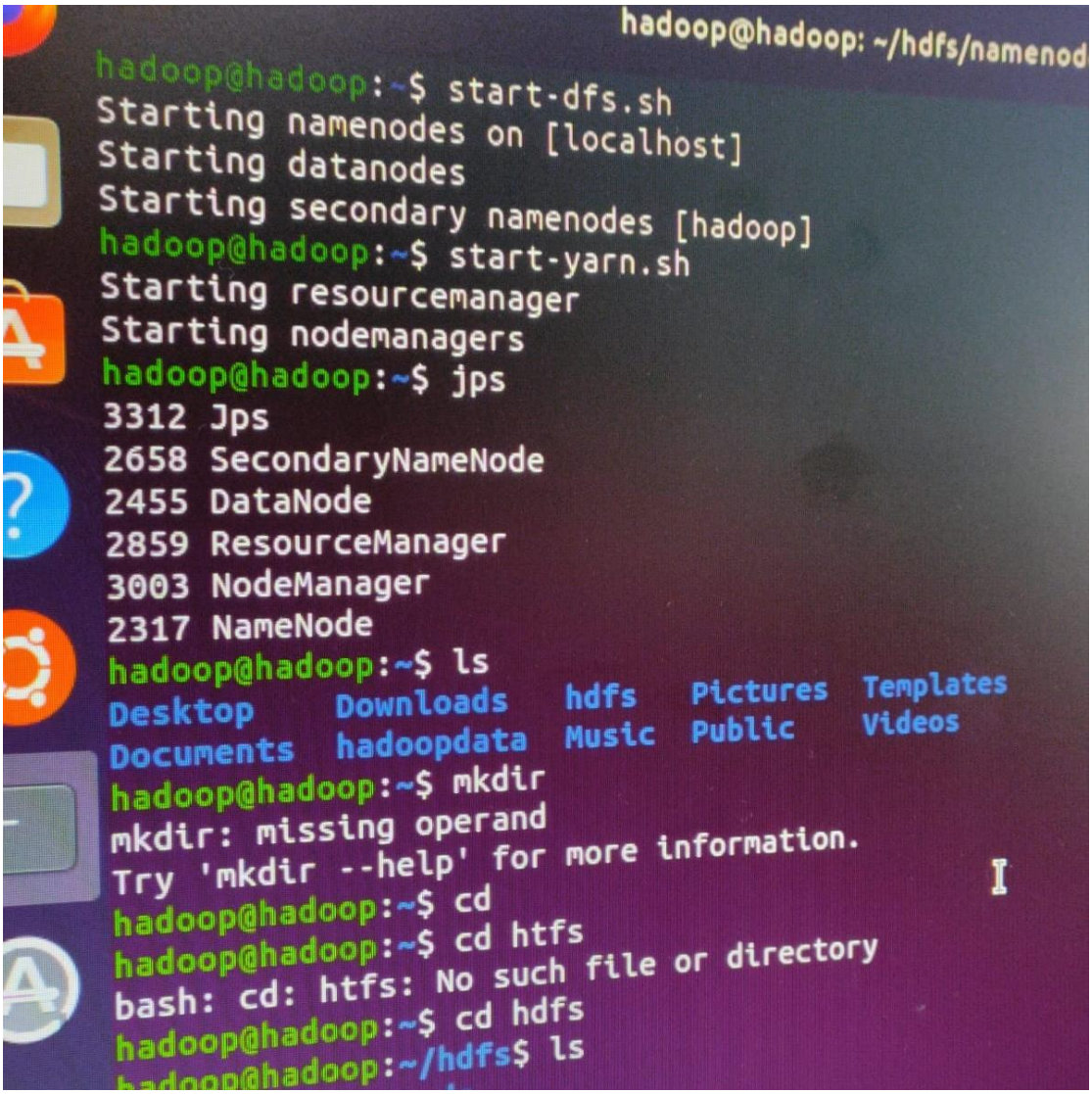
BDA - Exp - 2 - Hadoop HDFS Commands

Atharva Prashant Pawar - 9427







PostLab Questions:

Atharva Prashant Pawar (9427) [Batch - D]

BDA : Exp 2

Q1. What are the main components of a Hadoop Application?

⇒ A Hadoop application is typically composed of several key components that work together to process & analyze large volumes of data across distributed clusters of computers. These components collectively enable the efficient storage, processing & management of data.

A. Hadoop Distributed File System (HDFS):
HDFS is the primary storage layer in Hadoop. It divides data into blocks & replicates them across multiple machines in the cluster to ensure fault tolerance. It is optimized for handling large files.

B. MapReduce :
MapReduce is a programming model & processing framework for parallel data processing. It divides tasks into two phases- Map, which process data & produces key-value pairs, & Reduce which aggregates & processes those pairs.

C. YARN (Yet Another Resource Negotiator):
YARN is the resource management layer of Hadoop. It manages cluster resources & enables multiple application to share resources efficiently.

D. Hadoop Common:
This includes libraries & utilities used by other Hadoop modules. It provides the basic tools & libraries required by the Hadoop ecosystem, such as the Hadoop API, authentication & security.

②

E. Hive:

Hive is a data warehousing & SQL-like query lang. tool for Hadoop. It allows users to perform data analysis using a familiar SQL syntax, converting queries into MapReduce jobs.

F. Pig:

Pig is a high level platform for creating MapReduce programs using a scripting lang. called Pig Latin. It simplifies the process of writing complex MapReduce jobs by abstracting many of the low-level details.

G. HBase:

HBase is a distributed, scalable & consistent NoSQL database that can store & manage large amounts of sparse data. It is suited for random read & write operations.

Q2. Explain the difference b/w NameNode, Backup Node & Checkpoint NameNode.

A. NameNode:

NameNode is a crucial component in the Hadoop Distributed File System (HDFS)

It manages the metadata of the file sys. such as the hierarchy of files & directories, permissions & the mapping of blocks to data nodes. ~~The~~

The NameNode is a single point of failure in HDFS, as its loss can lead to data loss & service disruption.

It stores metadata in memory & maintains 2 persistent files:

fsimage (snapshot of metadata) &

edit logs (record of recent changes).

Responsible for handling client requests for data location, data retrieval & block management
If a cluster is large & has a lot of files & directories, the memory usage of the NameNode can become a bottleneck.

B. Backup Node :
The Backup Node is introduced to alleviate the single point of failure issue of the NameNode.
It maintains a copy of the NameNode's metadata, serving as a read-only & up-to-date backup
The Backup Node regularly fetches the fsimage & edit logs from the active NameNode to synchronize its metadata.
It helps reduce the recovery time in case of NameNode failure by providing a warm standby.
Clients can connect to the Backup Node for read-only operations, relieving some of the load from the primary NameNode.

C. Checkpoint Name Node :
The Checkpoint NameNode is another approach to mitigate the single point of failure in HDFS.
It periodically creates checkpoints by saving the current state of the metadata (fsimage & edit logs) from the active NameNode.
Checkpoints reduce the amount of edit logs that need to be processed during NameNode recovery, improving recovery time.
While a Checkpoint NameNode is running, it maintains a read-only, up to date version of the file system metadata.

(4)

Q3. Explain the use of cat, du, dus command

A. Cat Command:

the cat commd stands for "concatenate" & is used to display the contents of one or more files in the terminal.

It's commonly used to view the content of text files, display config. files, or concatenate & display multiple files.

cat can also be used to create new files or combine existing files & redirect the o/p to another file.

Usage:

cat file.txt , cat file1.txt file2.txt , cat > newfile.txt

B. du cmd:

the du cmd stands for 'disk usage' & is used to estimate the disk space used by files & directories.

It provides info. about the size of files & directories, including their subdirectories.

by default, du displays sizes in 'KB'

* Usage: du file.txt, du -h file.txt, du -s directory, du -sh *

C. dus cmd:

The dus cmd is not a standard Linux cmd but a ~~version~~ variation of du that provides a summary of disk usage.

It's often used with Hadoop clusters to calculate HDFS disk usage for specific users or directories.

The dus cmd is not available by default & might need to be installed separately on your sys.

* usage:

dus user username, dus directory /path/to/directory