Name: Tejas Redkar
PRN: 10322 10937
Panel - BDT - 2
Batch - Batch - 1
Roll No: BDT - 22

Big Data Technologies
Assignment - 04

* Problem Statement

Install Hadoop & perform basic Hadoop commands on it.

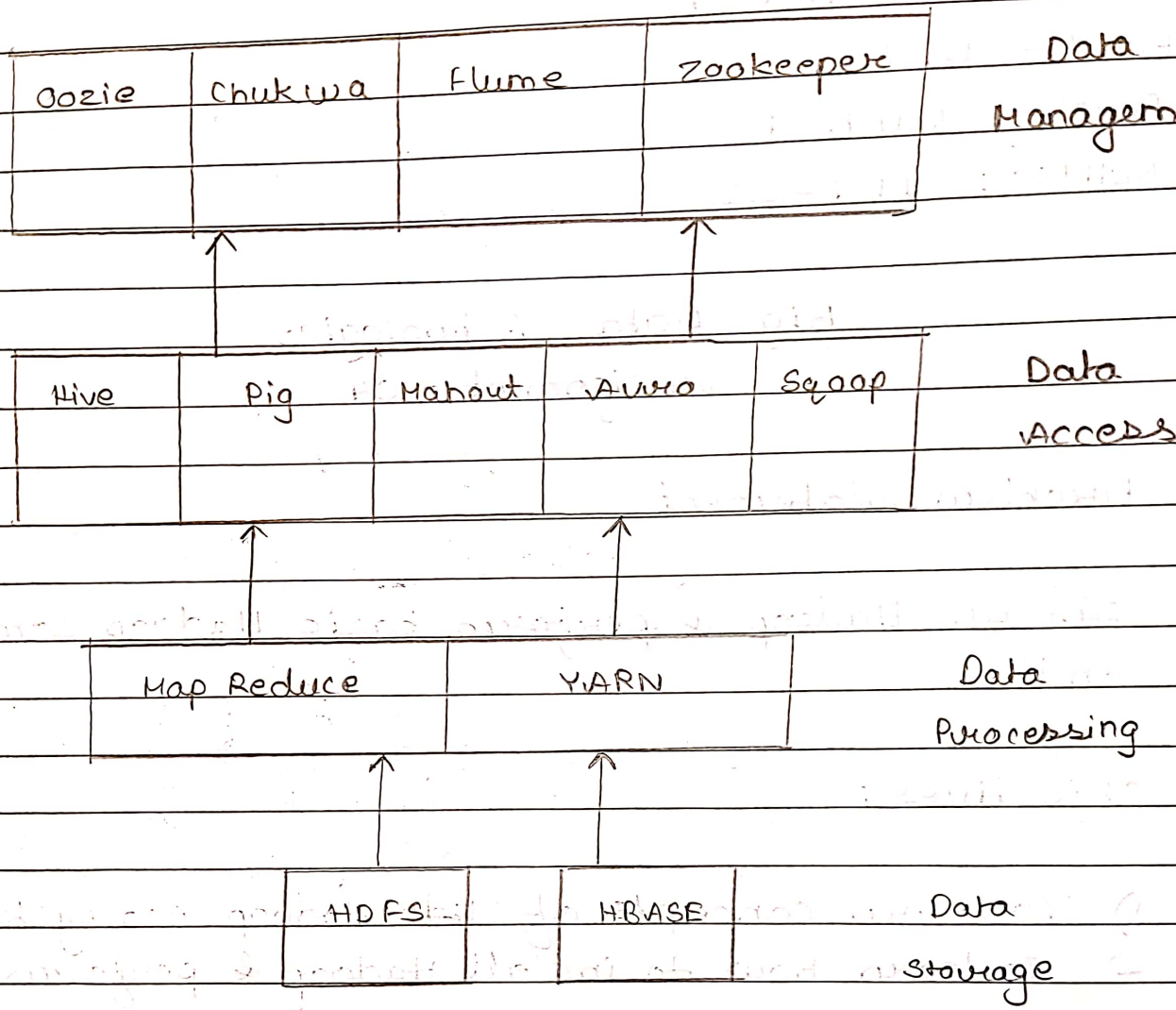
* Objectives:

1) To learn concepts of ~~Hdo~~ Hadoop ecosystem
2) To learn how to install Hadoop & perform basic Hadoop commands.

* Theory:

Draw Hadoop Ecosystem diagram

Hadoop Ecosystem



* History of Hadoop

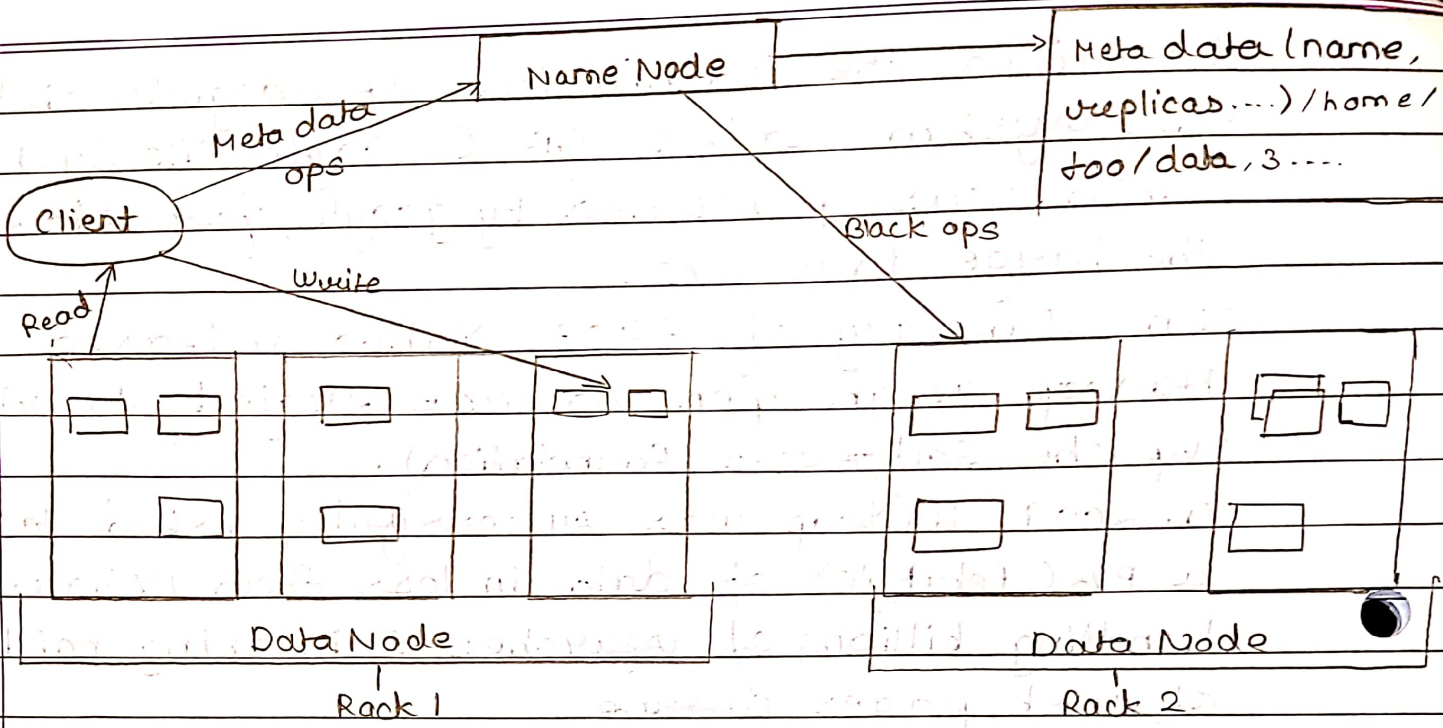
Hadoop was started with Doug Cutting & Mike Cafarella in the year 2002 when they both started to work on Apache Nutch project. Apache Nutch project was the process of building a search engine system that can index 1 billion pages.

In 2003, they came across a paper that

described the architecture of Google's distributed file system, called GFS (Google File System) which was published by google, for storing the large data sets.
In January of 2008, Yahoo released Hadoop as an open source project to ASF (Apache Software Foundation).
In 2009, Hadoop was successfully tested to sort a PB (PetaByte) of data in less than 17 hours of handling billions of searches & indexing millions of web pages ~~during~~.

Explain HDFS architecture in detail with suitable diagram

HDFS is an Open source component of the Apache Software Foundation that manages data. HDFS has scalability, availability, & replication as key features. Name nodes, secondary name nodes, data nodes, checkpoint nodes, backup nodes, & blocks all make up the architecture of HDFS. HDFS is a fault-tolerant & is replicated. Files are distributed across the cluster systems using the Name node & Data nodes. The primary difference between Hadoop & Apache Hbase is that Apache HBase is a non-relational database & Apache Hadoop is a non-relational data store.



* Platform : 64-bit open source Linux / Windows

* Conclusion

Hence, I learned to install Hadoop & perform basic Hadoop commands on it.

* FAQ's

Q1) Explain with syntax & example of any 10 basic Hadoop commands.

Ans 1) Version

example - hadoop version

The hadoop fs shell version command version prints hadoop version.

2) mkdir

example - hadoop fs - mkdir /newDataFlair

This command creates the directory in HDFS if it does not already exist.

3) LS

example - hadoop fs - ls /

The Hadoop fs shell command ls displays a list of all the contents of a directory specified in the path provided by user. It shows the name, permissions, owner, size, & modification date for each file or directories in the specified directory

4) put

example- hadoop fs - put ~/localfile1 / fromlocal
file

The hadoop fs shell command put is similare to the copyFromLocal, which copies files or directory from the local filesystem to the destination in the Hadoop filesystem.

5) copyFromLocal

example - hadoop fs - copyFromLocal ~/test1 /newData Flair /copytest

This command copies the file from the local file system to HDFS.

6) get

example - hadoop fs -get /littletestfile ~/copy from hadoop

The hadoop fs shell command get copies the file or directory from the Hadoop file system to the local file system.

7) copyToLocal

example - hadoop fs - copyToLocal /newDataFlair/sample ~/copysample

copyToLocal command copies the file from HDFS to the local file system.

8) cat

example - hadoop fs - cat /newDataFlair/sample

The cat command reads the files in HDFS & displays the content of the file on console or stdout.

9) mv

example - hadoop fs -mv /DRI/ DataFlair

The HDFS command moves the files or directories from the source to a destination within HDFS

10) cp

example - hadoop fs - cp/newDataFlair /file1/ data flair

The cp command copies a file from one directory within the HDFS.

Q2) State the use of Name node & Data node

Ans Namenode works as Master in Hadoop cluster.

1) It stores the MetaData of actual data.
2) Manages the file system namespace
3) Regulates client access request for actual file data file.
4) Assigns work to Slaves (DataNode)
5) Executes file system name space operation like opening / closing files, renaming files & directories.

DataNode works as Slave in Hadoop cluster

1) Actually stores Business Data
2) This is actual worker node where Read / Write/ Data

processing is handled.

3) Upon instruction from Master, it performs creation / replication / deletion of data blocks.

4) As all the Business data is stored on DataNode the huge amount of storage is required for its operation. Commodity hardware can be used for hosting Datanode.

Q3) State the different applications of Hadoop.

Ans 1) Finance sectores
2) Security & law enforcement
3) Companies use Hadoop for understanding customer requirements
4) Retail Industry
5) Real-time analysis of customers data
6) Government sectores
7) Advertisements
8) Sentiment Analysis
9) Financial trading & forecasting
10) Improving Personal Quantification
11) Health care sectores
12) Optimizing machine performance.

@11/10/23