Name: Nishant Sanjay Chavan erp: 180600283
class: BE/B Sub: BDA Rollno: 18

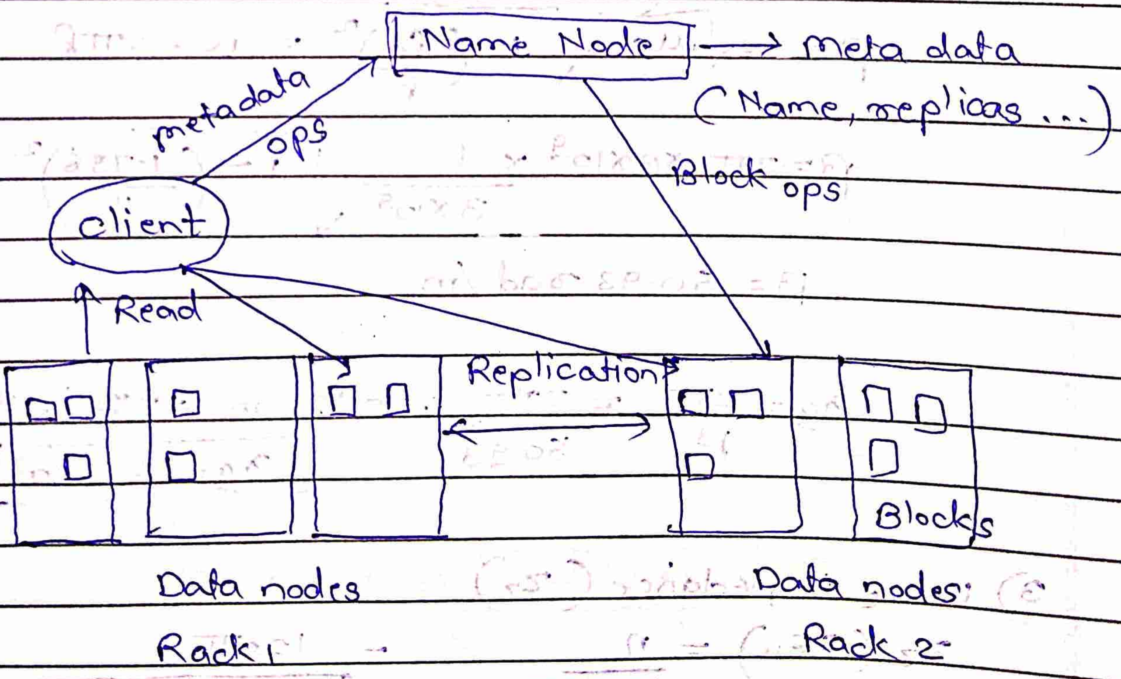
UT-1

Q.2) - Hadoop File system was developed using distributed file system. It is run on commodity hardware. Unlike other distributed systems. HDFS is highly faulttolerant and designed using low-cost hardware.

Features of HDFS:

1) It is suitable for the distributed storage and processing.
2) ~~Hd~~ Hadoop provides a command interface to interact with HDFS.
3) The built-in servers of namenode and datanode helps user to easily check status of clusters.
4) streaming access to file system data.

HDFS Architecture:



Name: Nishant Sanjay Chavan erp: 180600283
class: BE/B Sub: BDA Rollno: 18

UT-1

NameNode:
The Namenode is a commodity hardware that contains the GNU/Linux operating system and the ~~ma~~ namenode software. It is a software that can be run on commodity hardware. The system having the namenode acts as the master server and it does following tasks:-
1) managing the file system namespace.
2) Regulates client's access to file.
3) It also executes file system operations such as renaming, closing and opening files.

Data node:
The Data node is a ~~community~~ commodity hardware using GNU/Linux operating system. and datanode software. Every node in a cluster, there will be datanode. These nodes manage the data storage of their system.
1) Datanodes perform read-write operations on file system as per client request.
2) They also perform operations such as block creation, deletion and replication according to instructions of namenode.

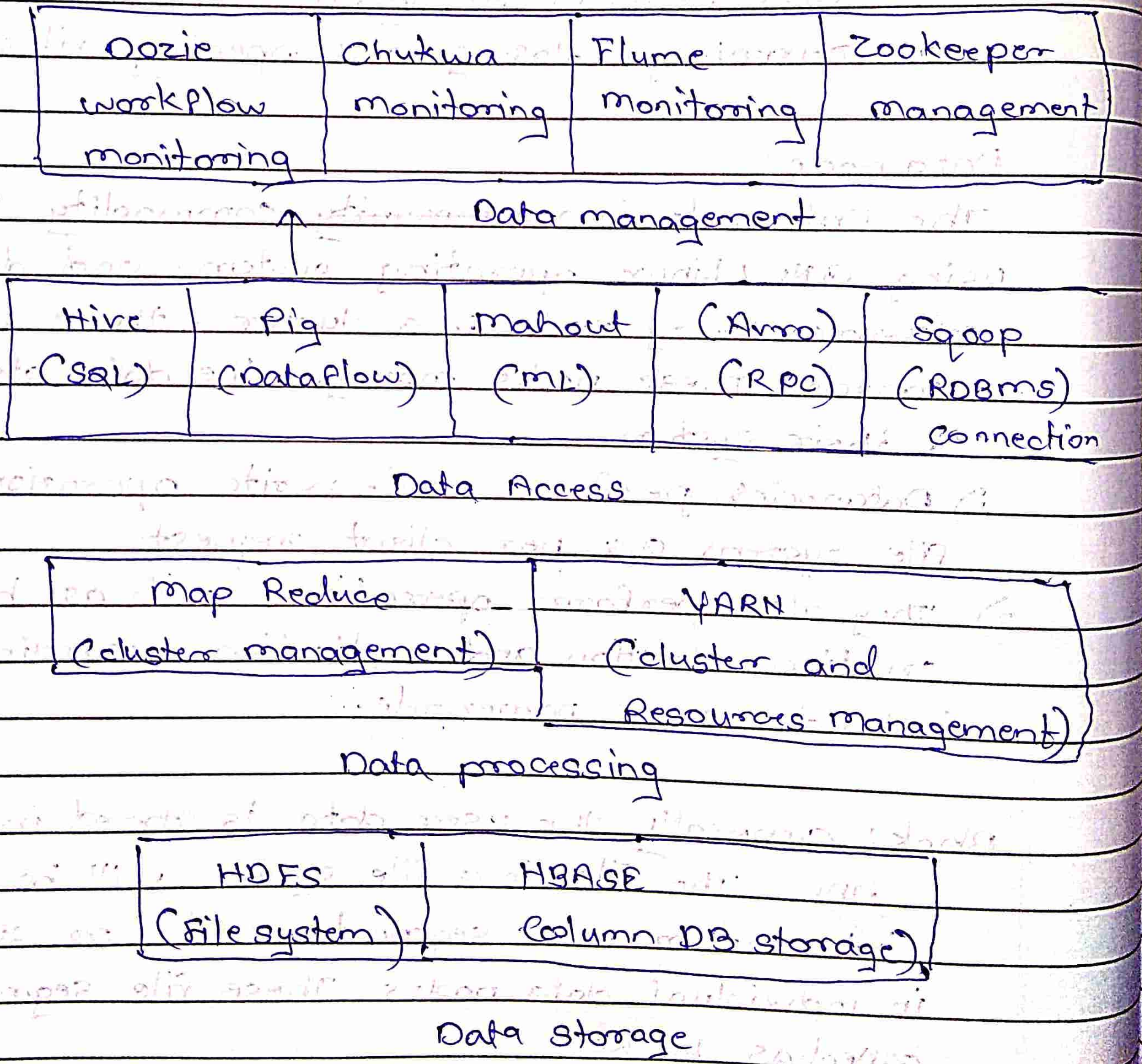
Block: Generally the user data is stored in files of HDFS. The file in a file system will be divided into one or more segments and / or stored in individual data nodes. These file segments are called as blocks.

@chavan

UT-1

Q.3) Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements in Hadoop i.e. Hadoop, MapReduce, YARN and Hadoop common.

Hadoop Ecosystem



Name:- Nishant Sanjay Chavan prn: 180600283
Class: BE/B Sub: BDA Roll no: 118

VT-1

1) HDFS: HDFS is the primary or major component of hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in form of log files.

HDFS consists of Name Node and Data node.

2) YARN

: Yet Another Resource Negotiator as the ~~ma~~ name replies, YARN is one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation.

It consists of :

1) Resource manager
2) Nodes manager
3) Application manager.

3) MapReduce:

By making the use of distributed and parallel algorithms. mapReduce makes it possible to carry over the processing logic and helps to write applications which transform big data sets into manageble one.

4) PIG: Pig was basically developed by Yahoo which works on pig latin language, which is Query language similar to SQL.

It is platform for structuring the data flow, processing and analyzing huge data sets.

Name: Nishant Sanjay Charan   erp: 1806002

Class: BE/B   Sub: BDA   Roll no: 18

UT-1

4) HIVE: With the help of SQL methadolology and interface, HIVE perform reading and writing large datasets.

5) Mahout: Mahout allows machine ~~lea~~ learning learnability to a system or application. Machine learning as name suggests helps the system to develop itself based on some patterns, user/enviromental interaction algorithms.

Acharan