Big Data Analytics

- SCSA1603

Name: Mohnish Devaraj

Reg NO: 39110636

Assignment-II

MCQs/Fill ups

1. a) MapReduce
2. a) MapReduce, Hive and HBase
3. c) Task Tracker
4. a) Maptask
5. a) Reduce

PART-B

1. Functions of Namenode are:

a) To store all the metadata (data about data) of all the slave nodes in a Hadoop cluster.

b) Keep track of all the slave nodes (whether they are alive or dead). This is done using the heartbeat methodology.

c) Replication (provides High availability, reliability and Fault tolerance). Namenode replicates the data on slavenode to various other slavenodes based on the configured Replication Factor.

d) Balancing: Namenode balances data replication, i.e, blocks of data should not be under or over replicated. This needs to be manually configured.

2. Jobtracker is a master which creates and run the job. Jobtracker which can run on the Namenode allocates the job to task tracker. It is tracking resource availability and task life cycle management, tracking its progress, fault tolerance, etc.

Tasktracker ran the tasks and report the status of task to jobtracker. Tasktracker ran on Data Nodes. It has function of following the orders of the job tracker and updating the job tracker with its progress status periodically.

3. There are three types of Mode of operations.

a) Standard Mode
b) Pseudo Distributed Mode
c) Fully distributed Mode

4. Types of data supported by Hadoop

→ Numeric data

BIGint	float
Boolean	int
Decimal	Smallint
Double	Tinyint

→ String data

Binary String

charn Varcharn

→ Complex data

Array Struct

Map

→ Date and time data

Date Timestamp

Interval

5. Components of the Hadoop Ecosystem:

a) HDFS (Hadoop Distributed File System)

b) MapReduce

c) YARN

d) HBase

e) Pig

f) Hive

g) Sqoop

h) Flume

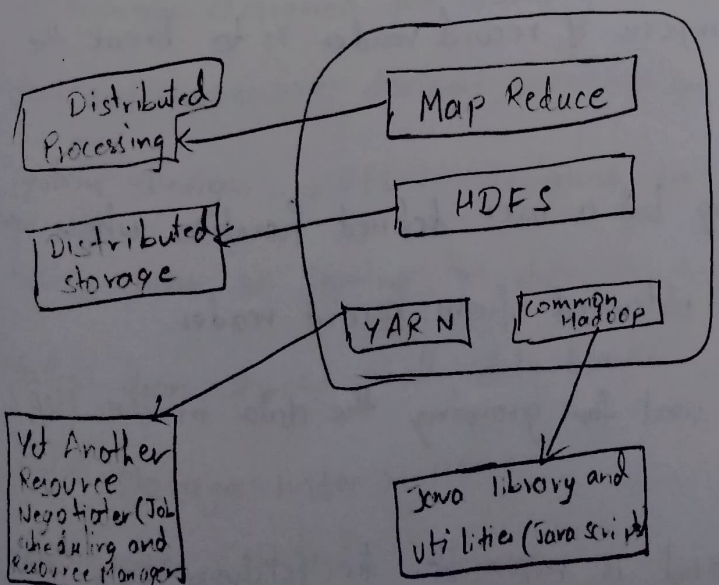
i, kafka

j, zookeeper

k) Spark

PART-C

1. Hadoop Architecture



The Hadoop ~~Consists for~~ Architecture mainly consists of 4 components:

-> Map Reduce
-> HDFS (Hadoop Distributed file system)
-> YARN (Yet another Resource Negotiater)
->Hadoop Common

1) Map Reduce

MapReduce nothing but just like an algorithm or a data structure that is based on the YARN framework. The major feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster which makes Hadoop working so fast. MapReduce has mainly 2 tasks which are divided phase-wise:

In first phase, Map is utilized and in next phase Reduce is utilized

Map Tasks

- Record Reader: The purpose of record reader is to break the records.
- Map: A map is nothing but a user-defined function whose work is to process the Tuples obtained from record reader.
- Combiner: Combiner is used for grouping the data in the Map workflow.
- ~~Partion~~ Partitionor: Partitional is responsible for fetching key-value pairs generated in the Mapper Phases.

Reduce Task:

- shuffle & sort : The task of Reducer starts with this step, the process in which the Mapper generates the intermediate key-value and transfers them to the Reducer task is known as shuffling.

- Reduce : the main function or task of the Reduce is to gather the tuple generated from Map and then perform some sorting and aggregation sort of process on those key-value depending on its key element.

- Output Format : Once all the operations are performed, the key-value pairs are written into the file with the help of record writer, each record in a new line, and the key and value in a space-seperated manner.

2) HDFS

HDFS is utilized for storage permission is a Hadoop cluster. It mainly designed for working on commidity Hardware devices (inexpensive devices), working on a distributed file system Design. HDFS is designed in such a way that it belives more in storing the data in a large chunk of blocks rather than storing small data blocks.

Data storage nodes in HDFS:

Namenode (Master)

Data node (Slave)

- Name Node: Name Node works as a Master in a Hadoop clusters that guides the Datanode (slaves). Name node is mainly used for storing the Metadata i.e., the data about the data. Meta Data can be transaction logs that keep track of the user's activity in a Hadoop cluster.

- Data Node: Data Nodes works as a slave Data Nodes are mainly utilized for storing the data in a hadoop cluster, the number of Data nodes can be from 1 to 500 or even more than that. The more number of Data Node, the Hadoop cluster will be able to store more data. So it is advised that the Datanode should have High storing capacity to store a large number of file blocks.

3) YARN (Yet Another Resource Negotiater)

YARN is a frame work on which Map Reduce works. YARN performs 2 operations that are Job scheduling and Resource Management. The purpose of Job schedular is to divide a big task into small jobs so that each job can be assigned to various slaves in a Hadoop cluster and Preprocessing can be Maximized. Job Scheduler also keeps: tracks of which job is important, which job has more priority, dependies between the jobs and all the other information like job timing, etc.

And the use of Resource Manager is to manage all the resources that are made available for running a Hadoop cluster.

4) Hadoop Common

Hadoop Common are nothing but the java library and java files or we can say the java scripts that we need for all the other components present in a Hadoop cluster. these utilities are used by HDFS, YARN, and MapReduce for running the cluster. ~~is common so it needs to be solved~~ Hadoop Common verify that Hardware failure in a Hadoop cluster is common so it needs to be solved automatically in software by Hadoop framework.