DHARMSINH DESAI UNIVERSITY, NADIAD.
FACULTY OF TECHNOLOGY
ONLINE SESSIONAL EXAMINATION

B.Tech (CE) sem → 7
subject : Big Data Analytics

Rollno. → 142
Signature → Prakruti

Date : 28/8/2020
Time : 10:00 to 11:15
Total pages : 10

Q.1

(a) responsibilities of namenode.

→ manages & maintains datanodes
→ receives block report & heartbeat from all datanodes.
→ records metadata about blocks

(b) checkpointing.

→ process involving merging of fsimage along with latest ~~et~~ editlog for creating new fsimage for namenode to have latest metadata
→ This task is performed by secondary name node or standby name node.

(c) components of resource manager

① components - scheduler
→ allocates resources to applications & doesnot

offers guarantees about restarting failed tasks.
→ it has pluggable policy like Capacity Scheduler & Fair schedulers

(2) application manager

→ accepts job-submissions
→ negotiates first container for executing application specific Application Master
→ provides service for restarting ApplicationMaster on failure.

(d) IoT deals with million of devices that collect & operate very large data. This is mainly unstructured data.
→ Big Data helps IoT to process this large data on real-time basis and storing it using different storage technologies.

(e) schema on read

→ This is a strategy for analysing the data in tools like hadoop in which data is applied to schema when it is pulled out from a storage location, rather than when it is stored.

→ Data is loaded as it is and can be read in different manner.

example : hive, impala

→ It is used when raw / atomic data is being stored or we need flexibility on consumption of data.

(f) sharding

→ process of storing data in multiple machines.
→ helps mongoDB meeting the demands of data growth
→ As data size increases, single machine may not handle it
→ sharding solves this issue.

policies :

① vertical scaling

→ refers to adding more resources to servers as on demand.

② horizontal scaling.

→ mongoDB supports horizontal scaling through sharding.

→ here dataset is divided & loaded over multiple servers.

Q.2

(b) Task:

→ execution of Mapper or Reducer on a slice of data
→ also called Task-In-Progress (TIP)
→ This means processing the data is in progress on mapper or reducer.

Task attempt:-

→ particular instance of attempt to execute task on a node.
→ it is possible that a machine can go down.
→ if it happens, task is rescheduled on some other node. But this cannot be done infinite time.
→ default value of task attempt is 4.
→ if a task fails 4 times, job is considered as failed job.
→ this value can be changed for huge or high priority job.

Failed task: →

→ A task that is failed more than no. of task attempt times.
decided

→ it is a task attempt that is completed but with an unexpected status value. i.e. task that generate errors.

killed task →

→ a task attempt which is a duplicate of a task attempt which was started as a part of speculative execution, & is killed by Hadoop because of below reasons.

→ it didn't report progress during timeout
→ scheduler needed slot for other pool or queue.
→ its result is not needed because it has completed on other place.

speculative execution :→

→ execution of same task is started in multiple boxes.

→ The first task to finish is joining & other ^ tasks are killed duplicate.

input split :→

→ it is created by inputformat
→ logically represents data processed by

individual mappers.

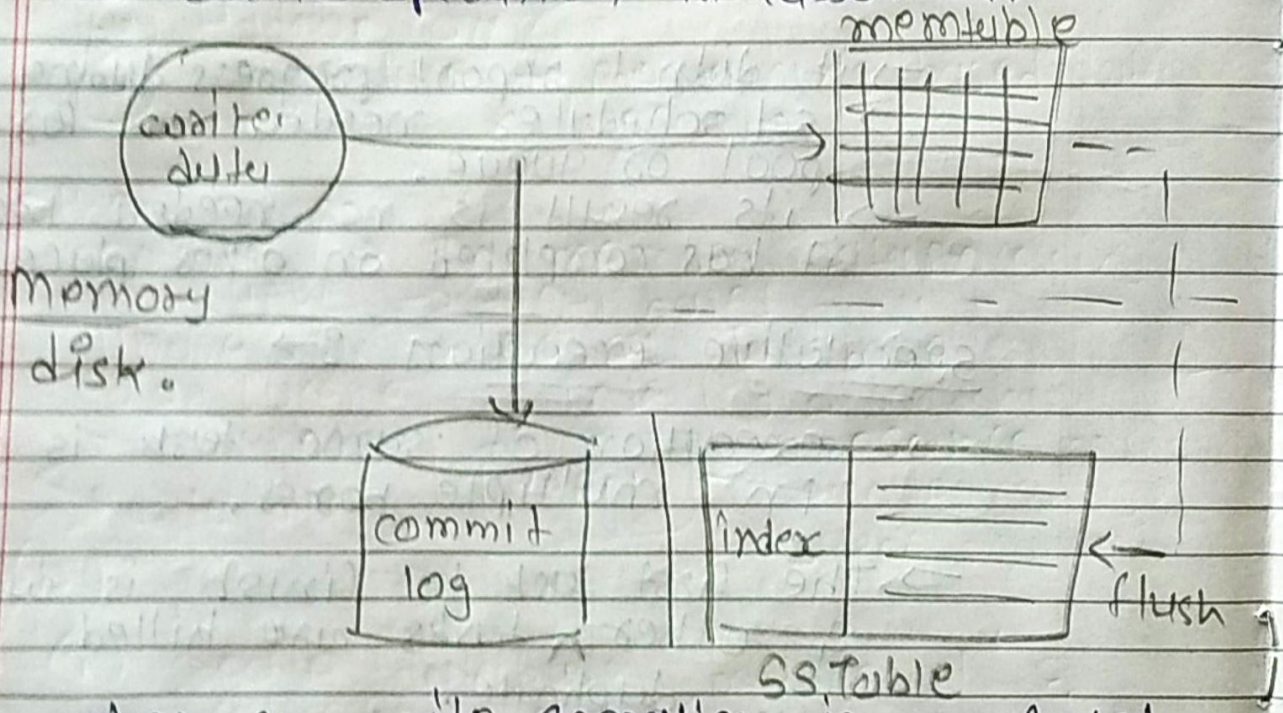
→ one map task created for each split.

→ split divided in records & each record is processed by mapper.

→ its length is measured in bytes.

→ size is approx. equal to block size.

(c) write operation in cassandra.



→ when a write operation is completed, it is firstly written in commit log.

→ Then, it is pushed in memory resident data structure called memtable having a predefined threshold value.

(2)

→ A node respons ~~suss~~ successful message to coordinator only when write is completed in memtable & commit log.

→ when number of objects reaches a threshold, at that time, the contents of memtable are flushed to a disk in a file called SSTable (Sorted String Table)

→ Flushing is a non-blocking operation.

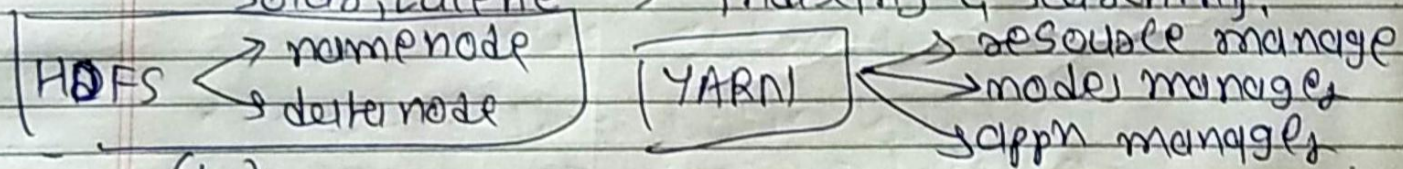
Q.3

(a) Hadoop eco system & related components.

→ Hadoop ecosystem is a platform providing various services to solve big data problems.
→ an open source framework.

→ There are 4 major components in Hadoop.
1. HDFS (Hadoop distributed file system)
2. ~~MapReduce~~ (yet another resource Negotiator) YARN →
3. Hadoop common.
4. MapReduce → programming based data processing.

other components.
Spark → In memory data processing
PIG, HIVE → query based "
HBase → NoSQL database
Solar, Lucene → indexing & searching.



(b)

(i) create database 'bdaspecial' & collection 'wordCloud'.

use bdaspecial to create database.

show dbs to check if database is created.

(9)

db.createCollection("wordCloud")
(to create collection)

Show collections
(to check if it is created)

(ii) import records from csv file

mongoimport -d bdaspecial -c coord cloud --type csv --file myfile.csv --headerline

-d → database name
-c → collection
--file → file address
--headerline → to use 1st line as field names.

(iii) find max salary per group of branches only for year 2020

(c) (2) steps

(1) map fn → takes, data & converts into another set of data where individual elements are broken in tuples
input → set of data (comma separated words)

output → another set of (key, value) pairs e.g. (car, 1) (bus, 1) ...

reduce fn takes map's o/p as i/p & combines tuples into smaller set of tuples

i/p = o/p of map fn

o/p = (key, value)

e.g. (car, 5)
(bus, 3)
⋮

map fn.

```
public void map( LongWritable key , Text
                 ,value, Context con)
        throws IOException.
{   String line= value.toString();
    String [] coords = line.split(",");

    for ( String coord : coords)
    {  Text okey = new Text (coord.toUpper
                  case().trim());
       IntWritable ovalue = new IntWritable(1)
       con.write (o_key, ovalue);
    }
}
```

end of answersheet