Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058, India

(Autonomous College Affiliated to University of Mumbai)

Mid Semester Evaluation

(Synoptic)

Mar 2018

Max. Marks: 30 **Class:** B.E.

Course Code: CPE8035 **Name of the Course:** Elective-III Big Data Analytics

Semester: VIII **Branch:** Computer

1-A) Find the station wise highest temperature :

Santa, 2017, 38.3
Colaba, 2015, 39.9
Santa, 2017, 40.02
Colaba, 2017, 41.7
Colaba, 2015, 29.3
Santa, 2017, 42.7
Colaba, 2015, 38.1

Write a program using Mapreduce in java to find the station wise maximum temperature. Clearly state the output at mapper stage and reducer stage **CO-1** **[4M]**

Synoptic :

1. Correct Mapper Output code showing Following result :
   = 2 Marks
   Santa, List [temp]
   Colaba, List of [temp]
2. Correct Reducer Correct code showing Following result :
   = 2Marks
   Santa, Highest temp
   Colaba, Highest temp

1-B) During checkpointing contents of which two files are transferred from one node to another. Also give the name of these two nodes **CO-2** **[3M]**

Synoptic:

edits.log [1 M] fsimage [1M] NN to SNN [1 M] total = 3

OR

1-B) Write a command to merge three files a.txt, b.txt, c.txt from hadoop file system and copy it to local file system [3M]

Synoptic:

getmerge correct stepwisecommand 3 M

2-A) Differentiate between Big data and Traditional data. Give 10 parameters. **CO-1** **[5M]**

Synoptic:

0.5 Mark for each point

Parameters	Traditional Data	Big Data
Volume	GB	Constantly Updated(TB or PB currently)
Generated Rate	Per hour, per day…	More rapid(almost every second)
Structure	Structured	Semi-structured and unstructured
Data Source	centralized	Fully distributed
Data Integration	easy	Difficult
Data Store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real time
Update Scenarios	Repeated read and write	Write Once, Repeated Read
Data Structure	Static Schema	Dynamic Schema
Scaling Potential	Non-linear	Somewhat close to Linear

2-B) Write MongoDB query to perform following operations. **CO-1** **[3M]**

a) Command to create collection "students".
b) Add 2 students with property "name" in collection students.
c) Command to return all students.

Synoptic:

a) 1Mark for error free command. 0 Mark if incorrect

```
db.createCollection("students")
```

b) 1Mark for error free command. 0 Mark if incorrect

```
db.students.insert({"name":"sushil"})
```

c) 1Mark for error free command. 0 Mark if incorrect

```
db.students.find()
```

OR

2-B) Explain Sharding and Replication in NoSQL database. **CO-1** **[3M]**

Synoptic:

1.5 Mark for Sharding explanation.

1.5 Mark for Replication explanation

Sharding is the process of distributing data across multiple servers for storage. MongoDB uses sharding to manage massive data growth.It splits the data set and distributes them across multiple databases, or shards. Each shard serves as an independent database, and together, shards make a single logical database.A shard is a replica set or a single mongod instance that holds the data subset used in a sharded cluster. Shards hold the entire data set for a cluster. Each shard is a replica set that provides redundancy and high availability for the data it holds.

Replication is the process of synchronizing data across multiple servers. Replication provides redundancy and increases data availability with multiple copies of data on different database servers. Replication protects a database from the loss of a single server. Replication also allows you to recover from hardware failure and service interruptions.

2 types:
Master Slave replication
One node has the authoritative copy that handles writes. Slaves synchronize with master and handle reads.

Peer to peer replication
Allows write to any node, nodes coordinate between themselves to synchronize their copies of the data.

3-A) What is cardinality estimation problem? Explain the algorithm for counting distinct elements in stream **CO-4 [5M]**

Synoptic:

1. What is Count Distinct problem - 1M
2. Explain Flajolet-Martin Algorithm - 3M
3. Example - 1M

3-B) Explain the concept of Bloom Filter with the help of an example (5M) **CO-4 [5M]**

Synoptic:

1. What is Bloom Filter Data Structure - 1M
2. Bloom Filter Algorithm - 1M
3. Example - 2M
4. List doown applications of Bloom Filter -1M

OR

3-B) Explain how to perform Sampling on data in a stream with basic terms. What are the 2 types of Sampling algorithm. Explain any one Sampling algorithm (5M) **CO-4 [5M]**

Synoptic:

1. What is sampling with basic terms/concepts - 2M
2. 2 types of sampling - Probability and Non-probabilty Sampling - 1M
3. Explain any one sampling algorithm – SRS or Reservoir etc - 2M

3-C) What is Recommendation System? Explain Collaborative Filtering based recommendation system. How it is different from Content based recommendation system? **CO-5 [5M]**

Synoptic:

1. Recommendation System – 2M
2. Collaborative Filtering – 2M
3. Difference between Collaborative and Content based Recommendation System - 1M