

JADAVPUR UNIVERSITY
DEPARTMENT OF INFORMATION TECHNOLOGY
B.E. INFORMATION TECHNOLOGY, THIRD YEAR
SECOND SEMESTER EXAM 2025
SUBJECT: BIG DATA (HONS.)

Full Marks :100

Time : 3Hrs

CO1 : (20) Answer any four questions 4 X 5 =20

- i) List the fundamental features that a good Big Data analytics platform should offer, and name three popular Big Data platforms.
- ii) What monitoring tools and metrics are available to track the status, progress, and performance of applications running on YARN, illustrate with an example.
- iii) How does YARN manage the lifecycle of an application — from submission to completion? Describe the steps with a diagram.
- iv) What data structures are kept on the master in Hadoop ? Which data structures are kept persistently? What are the pros and cons of keeping all of that meta-data in main memory?
- v) In Big Data how the protocol for record appends different than ordinary writes? How do applications deal with this model? Is it ever possible for a client to read a stale replica?

CO2 : (30) Answer any three questions 3X 10 =30

- i) Consider the following data:

```
From_Number|To_Number|CallStartTime|CallEndTime|STDFlag
9665128505|8983006310|2015-03-01 09:08:10|2015-03-01 10:12:15|1
9665128505|8983006310|2015-03-01 07:08:10|2015-03-01 08:12:15|0
9665128505|8983006310|2015-03-01 09:08:10|2015-03-01 09:12:15|1
9665128505|8983006310|2015-03-01 09:08:10|2015-03-01 10:12:15|0
9665128506|8983006310|2015-03-01 09:08:10|2015-03-01 10:12:15|1
9665128507|8983006310|2015-03-01 09:08:10|2015-03-01 10:12:15|1
```

Provide a map reduce solution to find the user(s) who speak more than 60 minutes total in STD. Provide pseudo code for the map and reduce functions along with necessary explanations. **5+5=10**

[Turn over

ii) Given the following data

TransactionID, Date, CustomerID, Product, Category, Amount

1001,2024-06-01,C001,Phone,Electronics,600
 1002,2024-06-01,C002,Laptop,Electronics,1200
 1003,2024-06-02,C001,Book,Books,30
 1004,2024-06-02,C003,Tablet,Electronics,400
 1005,2024-06-03,C004,Pen,Stationery,5

Write a MapReduce program to compute the total sales amount for each product category.

5+5=10

iii)

- a) How does the MapReduce framework handle data locality?
- b) Why is it important for users to explicitly define the number of reducers in a MapReduce job, and how does improper configuration affect resource utilization, data shuffling overhead, and job performance?
- c) How does MapReduce perform a relational join between two datasets.

3+4+3 = 10

iv) How does the MapReduce framework build an inverted index from a collection of tweet messages, and what roles do the Mapper and Reducer functions play in this process?

5+5 = 10

CO3 : (20)

Answer any two questions

10 X 2 =20

i)

- a) State a possible scenario where column family database is not suitable. Mention the suitable category of database that is applicable in such a scenario.
- b) Following the CAP theorem, explain the differences between MongoDB and Cassandra.
- c) What is the use of “_id” key in a MongoDB documents? Describe its format.
- d) How can downtime be reduced in NoSQL databases?

3+2+3+2=10

ii)

You are tasked to design a MongoDB data model for an e-commerce application that manages customers and their orders. Each customer can have multiple orders. Describe two possible ways to model this relationship in MongoDB using:

1. Embedded documents
2. Referenced (linked) documents

For each approach, explain the structure of the collections, mention the advantages and disadvantages, specify in which use-cases each approach would be preferable.

6+4=10

iii)

We have an HBase table called employee with: **Row Key: emp_id, Column Families: personal and professional, Columns: personal:name, personal:address, professional:department, professional:salary**

- a) How would you create the employee table in HBase with the above two column families?
- b) How would you insert the following record into the employee table? **emp_id: 101 , name: Alice, address: New York, department: HR, salary: 50000**
- c) How would you retrieve all columns for the employee with **emp_id 101**?
- d) How would you retrieve only the **professional:salary for emp_id 101**?
- e) Which command would you use in the HBase shell to list all existing tables in the database?
2+2+2+2+2=10

CO4 : (30)

Answer All questions

15 X 2 =30

i)

- a) Explain with an example the Hive data model (the way data are stored in Hive).
- b) What is the use of SerDes (Serializer&Deserializer) in HIVE and explain various types of SerDes'supported by HIVE.
- c) Explain various data types supported by HIVEQL with example.
- d) Consider the following table.

What will be the output if the following query is executed in Hive ?

```
INSERT INTO TABLE pageid_age_sum
SELECT pageid, age, count(1)
FROM pv_users
GROUP BY pageid, age;
```

4+4+4+3=15

ii) a) How does lazy evaluation contribute to the optimization of Spark jobs?

- b) How does Spark Streaming ensure fault tolerance and reliability in processing streaming data?
- c) Mention few Spark Control operations with their functionalities.

d) Write a Spark program to analyze temperature data from a given CSV file (temperatures.csv). Output the average temperature for each year to a text file (Average_temperatures.txt). Input (temperatures.csv): Year, Temperature

```
2010 25
2010 28
2011 30
2011 32
2012 22
2012 24
```

4+3+3+5=15