Ramakrishna Mission Vivekananda Educational and Research Institute
PO Belur Math, Howrah, West Bengal 711 202
School of Mathematical Sciences
Department of Computer Science

MSc BDA Semester 1 End Semester Exam
JH

*S*tudent Name (in block letters): Date: December 4, 2024
*S*tudent Roll No: *M*ax Marks: 100
*S*ignature: Time: 3hrs

Theory

Marks 40 (2x20)

Answer 2 questions from this section.

1. Please answer the following.
   (a) Why is the block size in HDFS typically much larger (e.g., 128 MB or 256 MB) than the block size of traditional file systems (e.g., 4 KB or 8 KB)?
   (b) Can increasing the block size indefinitely improve performance?
   (c) What are the impacts on the NameNode and DataNode when a large number of small files are stored in HDFS?

2. Please answer the following.
   (a) How do the NameNode and DataNode work together to ensure fault tolerance and high availability in the Hadoop Distributed File System?
   (b) Describe the block replication mechanism in HDFS.
   (c) How does HDFS manage DataNode failures?

3. Design a MapReduce system to perform common friend enumeration in a social media environment.

Practical

Marks: 60.

ONE of the following two questions need to be answered in this section.

1. Develop a MapReduce Application to analyze employee and department data stored in a Hadoop cluster. The objective is to combine employee details with department information, group the results by region, and write them to region-specific files.

EmployeeID	Name	DepartmentID	Salary
E001	Alice	D01	50000
E002	Bob	D02	60000
E003	Charlie	D01	55000
E004	David	D03	70000
E005	Eve	D02	65000
E006	Frank	D04	48000
E007	Grace	D01	53000
E008	Hannah	D03	72000
E009	Ian	D02	61000
E010	Jack	D04	45000

Table 1: Employee Details: employee1.csv

DepartmentID	DepartmentName	Region
D01	HR	North
D02	IT	East
D03	Finance	South
D04	Sales	West

Table 2: Department Details: department1.csv

EmployeeID	Name	DepartmentName	Salary
E001	Alice	HR	50000
E003	Charlie	HR	55000
E007	Grace	HR	53000

Table 3: Example output: north_employees.txt

EmployeeID	Name	DepartmentName	Salary
E002	Bob	IT	60000
E005	Eve	IT	65000
E009	Ian	IT	61000

Table 4: Example output: east_employees.txt

2. Develop a MapReduce Application to analyze e-commerce transaction data stored in a Hadoop cluster. The dataset contains the following fields, separated by commas. Generate reports summarizing the total sales for each ProductCategory, with the output files grouped by region.

TransactionID	**CustomerID**	**ProductCategory**	**Amount**	**Region**
T001	C123	Electronics	350	North
T002	C124	Books	200	East
T003	C123	Electronics	100	South
T004	C125	Clothing	400	North
T005	C126	Books	150	West
T006	C127	Electronics	500	North
T007	C128	Books	300	East
T008	C129	Clothing	250	South
T009	C130	Clothing	100	West
T010	C131	Electronics	600	East

Table 5: Transaction Data: sales2.csv

ProductCategory	**Amount**
Electronics	850
Clothing	400

Table 6: Example output: north_sales.txt

ProductCategory	**Amount**
Electronics	100
Clothing	250

Table 7: Example output: south_sales.txt