

Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First / Second Semester 2023-2024

Mid-Semester Test
(EC-2 Make-up)

Course No. : DSECLZG522
Course Title : Big Data Systems
Nature of Exam : Closed Book
Weightage : 30%
Duration : 2 Hours
Date of Exam : EN

No. of Pages	= 2
No. of Questions	= 6

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1. a) Explain how Amdahl's law and Gustafson's law is applied to parallel processing. [02Marks]
b) Why Amdahl's law appears to put a limit on parallel processing effectiveness. Explain how Gustafson's law can act as a counter argument to it. [04Marks]

Q.2. An online store is analyzing sales data to announce discounts on purchases made using the most popular Credit card. The data for the year 2023 is having a volume of 1 TB and the schema of the data is given below:

Transaction_date	Product_ID	Price	Card_Type	Name	City	State	Country	Account_created
------------------	------------	-------	-----------	------	------	-------	---------	-----------------

You need to find out (1) The type of the Card used for maximum total payments received in the year and (2) The highest amount paid through each type of the cards. Write pseudo code for a MapReduce program to extract these two parameters from the sales data. [06Marks]

Q.3. The employee data of a company with the following schema is given to you in the form of a CSV file. The file has 1 million records. The 2 instances given are examples of the data.

Employee_ID	Name	Age	Gender	Salary
1201	Gopal	45	Male	50000
1202	Manisha	40	Female	48000

Write pseudo code for a MapReduce program to find out the total number of Male and Female employees having the same age. You need to output Age, No. of Males with this age, No. of Females with this age. The number of output records should be equal to the number of distinct values of Age of employees. [05Marks]

Q.4. You are configuring a Cassandra NoSQL cluster with 7 nodes. The nodes of the cluster can be configured with servers (1) hosted within the same Rack, (2) hosted on different Racks within the same datacenter or (3) Servers hosted on remote datacenters. How do you select the right combination of servers for the nodes of the cluster for maximizing (1) Consistency (2) Availability, and (3)

Partition Tolerance. Which consistency level will be selected by you for (1) Write intensive applications, and (2) for Read intensive applications. [06Marks]

Q.5. “For an organization if the latency is more critical than data”. In this case whether to implement ACID or BASE data model? Justify your answer by explaining the importance and differences of each. [04Marks]

Q.6. Reading 1 MB data sequentially from a hard disk takes 16 milliseconds. What is the time taken by the Map Tasks running on a single node Hadoop 2.x cluster to read a file of size 600 Gb stored with the default block size on HDFS. A 4 node Hadoop 2.x cluster is configured with 3 Data-nodes using servers identical to that of the single node cluster. The same file with size of 600 Gb is stored on the HDFS filesystem of this cluster with the default block size. What will be the time taken by the Map tasks running on this cluster to read this file. Time mentioned in this question refers to wall clock time. By default, how many Map tasks will get deployed on one node of the cluster?

[03Marks]
