Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2024-2025

Mid-Semester Test
(EC2-Makeup)

Course No.	:	DSECLZG522
Course Title	:	Big Data Systems
Nature of Exam	:	Closed Book
Weightage	:	30%
Duration	:	2 Hours +20 Mins
Date of Exam	:	12-07-2025 (EN)

No. of Pages = 2
No. of Questions = 8

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. In a big data system with a memory hierarchy consisting of a cache and main memory, the cache access time (Tc) is 10 ns and the main memory access time (Tm) is 100 ns, calculate the average memory access time (Tavg) for a program with the given cache hit ratio (h). Assume that the program exhibits strong temporal locality of reference, resulting in a high cache hit ratio of 0.8. Show your calculations and explain the significance of the result in the context of big data system performance**.** **[** Tavg = h * Tc + (1-h) * (Tc + Tm)]

[Marks 3]

Q.2. In the following scenarios, refer to CAP theorem to briefly justify which type of database configuration is preferred among AP, CP, and CA.

(a) A large-scale event reservation system has less than 50% seats booked.

(b) An IPL cricket game is selling last minute tickets across multiple counters at the venue.

(c) A Bank runs its money transaction system and DB within a centralized data center.

(d) A large e-retailer running its backend systems across multiple data centers is giving a limited set of coupons to purchase an item on promotion.

[Marks 4]

Q.3. What are the key considerations when designing a scalable architecture using a NoSQL database like MongoDB or Cassandra, and how do factors such as primary key selection, data distribution, replication, and query optimization impact the overall performance and reliability of the system?

[Marks 4]

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 from the above 3 groups to form nodes of the Cassandra cluster for maximizing (1) Consistency (2) Availability, and (3) Partition Tolerance. Which consistency levels will be selected by you for (1) Write intensive applications, and (2) for Read intensive applications.

[Marks 4]

Q.5. A logistics company wants to optimize delivery routes of parcels within a city. Design a solution based on graph database. What are the properties to be assigned to the Nodes and Relationships of the graph database in your design. What are the advantages of your design when compared with a design based on RDBMS.

[Marks 3]

Q.6. What is the role of MapReduce programming paradigm in big data analytics, and how does it implement divide-and-conquer strategy and helps in designing scalable bigdata applications? What are the distinct types of parallelism used in MapReduce?

[Marks 4]

Q.7. 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 employees having the same age. You need to output Age and the total number of employees with this age. The number of output records should be equal to the number of distinct values of Age of employees.

[Marks 4]

Q.8. You have a 16384 MB file stored on HDFS as part of a Hadoop 3.x distribution. A data analytics program stores this file on the HDFS cluster with 3 data nodes and runs in parallel across the cluster nodes. The default values for HDFS block size and the replication factor is used in the configuration.

(a) Find out the total number of blocks of the data file including replicas that will be stored on one node of the cluster.

(b) The cluster has 48 cores to speed up the processing. If the program can at best achieve 60% parallelism in the code to exploit the multiple cores and the rest of it is sequential, what is the theoretical limit on speed-up you can expect with 48 cores compared to a sequential version of the same program running on one core with the same file? You can simplify the system to assume cluster nodes and cores mean the same and we can ignore the overheads of communication etc. depending on the specific cluster configuration, scheduling etc.

(c) Suppose you could use a more scalable algorithm with 80% parallelism and a larger file as you move to a 108-Core system. What would be the theoretical speed-up limit for 108 cores?

[Marks 4]

****