Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2022-2023

Mid-Semester Test
(EC-2 Regular)

Course No.	: DSECLZG522
Course Title	: Big Data Systems
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 08/01/2023 (FN)

No. of Pages = 2
No. of Questions = 7

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.

1. A node in a cluster fails every 200 hours of operation while other nodes never fail. On failure of the node, the whole system needs to be shutdown. It takes 1 hour to replace the faulty node and to bring up the cluster. Then the application needs to be started again. This takes another 1 hour. The production loss is $100k per hour. [Marks: 4]
What is MTTF in this case?

1. What MTTR in this case?
2. What is the availability of the system?
3. What is the cost of downtime per year?

2. Discuss briefly 3 key issues that will impact the performance of a data parallel application and need careful optimization. [Marks: 3]

3. Name a system and explain how it utilizes the concepts of data and tree parallelism. [Marks: 3]

4. A travel review site stores (user, hotel, review) tuples in a data store. E.g. tuple is (“user1”, “hotel ABC”, “<review>”). The data analysis team wants to know which user has written the most reviews and the hotel that has been reviewed the most. Write MapReduce pseudo-code to answer this question. [Marks: 4]

5. An e-commerce site stores (user, product, rating) tuples for data analysis. E.g. tuple is (“user1”, “product_x”, 3), where rating is from 1-10 with 10 being the best. A user can rate many products and products can be rated by many users. Write MapReduce pseudo-code to find the range (min and max) of ratings received for each product. So each output record contains (<product>, <min rating> to <max rating>). [Marks: 4]

6. In the following application scenarios, point out what is most important - consistency or availability, when a system failure results in a network partition in the backend distributed DB. Explain briefly the reason behind your answer. [Marks: 4]
(a) A limited quantity discount offer on a product for 100 items at an online retail store is almost 98% claimed.
(b) An online survey application records inputs from millions of users across the globe.
(c) A travel reservation website is trying to sell rooms at a destination that is seeing very few bookings.
(d) A multi-player game with virtual avatars and users from all across the world needs a set of sequential steps between team members to progress across game milestones.

7. The CPU of a movie streaming server has L1 cache reference of 0.5 ns and main memory reference of 100 ns. The L1 cache hit during peak hours was found to be 23% of the total memory references. [Marks: 4]

(a) Calculate the cache hit ratio h.
(b) Find out the average time (Tavg) to access the memory.
(c) If the size of the cache memory is doubled, what will be the impact on h and Tavg.
(d) If there is a total failure of the cache memory, calculate h and Tavg.

8. A table named **world** is created in Hive using the following command:

```
create table world (country string, population int, GDP float, GINI float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

Data is loaded into this table from a CSV file. Population is in millions.
Write HiveQL queries to find the following:

1. List of countries in which the population is more than 10 million
2. The country with the highest GDP
3. The country with the lowest value for the GINI coefficient
4. Name of 5 countries with the highest population.

[Marks: 5]

****