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

Mid-Semester Test
(EC-2 Regular)

Course No.	:	DSECLZG522
Course Title	:	Big Data Systems
Nature of Exam	:	Closed Book
Weightage	:	30%
Duration	:	2 Hours
Date of Exam	:	27-01-2024_(FN)

No. of Pages = 3
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. You are at the University library. You see few students browsing through the library catalog on a kiosk. You observe the librarians busy at work issuing and returning books. You see some new students filling out web-based forms for getting the library membership. You see a few students fill out the feedback form on the library website on the services offered by the library. The form has both fixed fields and columns for suggestions that can be entered in text form also. Quite a few students are learning using e-learning content. Think for a while and list down the different types of data that are being accessed and generated in this scenario. Give supporting arguments for your answer.

[2 Marks]

Q.2. In each of the following scenarios, point out and give a brief reason what type of multi-processor computer one would use as per Flynn's taxonomy, i.e., the choices are SIMD, SISD, MIMD or MISD.

(a) A scientific computing application does a f1(x) + f2(x) transformation for every data item x given f1 and f2 are specialized operations built into the hardware.

(b) A video is processed to extract each frame which can be either an anchor frame (full image) or a compressed frame (difference image wrt anchor). A compressed frame (C) is transformed using a function f, where each pixel is compared with the last anchor (A) to recreate the uncompressed image (B), i.e. B(i, j) = f(C(i, j), A(i,j)) for all pixels (i,j) in the input frames.

(c) A multi-machine Apache Hadoop cluster system for data analysis.

(d) A development system with multiple containers running JVMs and CouchDB nodes running on a single multi-core laptop.

[4 Marks]

Q.3. Assuming a program consists of 50% non-parallelizable code.
(a) Compute the speed-up when using 4 processors with fixed workload size.
(b) Now assume that the parallel workload size is fixed equally per processor. Compute the speed-up when using 4 processors.
(c) Explain why the speed-up in case (b) more than the speed-up in case (a).

[4 Marks]

Q.4. Consider the following two use cases carefully and suggest what is going to be your choice of a distributed database as per the design principles of CAP theorem, i.e., is it of type CA, CP or AP? Justify your design choice in each case.

(a) metaltrade.com is online commodities trading platform with users from across the globe. Their database is deployed across multiple regional datacenters, but trading is limited between users within a region. Users need to view the prices in real-time and trades are requested based on this real-time view. Users would never want their committed trades to be reversed. The database cluster has multiple nodes and failures cannot be ruled out.

(b) buymore.com is an online e-retailer. Every day early morning, the prices of various products (especially fresh produce) are updated in the database. However, the customers can continue shopping 24x7. Customer browsing uses the same database and customer churn is very sensitive to page access latency.

[4 Marks]

Q.5. Data stored on an SQL database is normalized into 3 tables (Users, Professions, Cars) as shown below.

Users

User_id	First_Name	Surname	Mobile	City	Location_x	Location_y
1	Paul	Miller	4085575051	London	45.123	47.232

Professions

ID	User_id	Profession
10	1	Banking
11	2	Finance
12	3	Trader

Cars

ID	User_id	Model	Year
20	1	Bentley	1993
21	2	Rolls Royce	1998
22	3	BMW	2005

Model the above data in MongoDB. When you are designing your MongoDB schema, make sure that the schema design will give optimum query performance for your application.

[4 Marks]

Q.6. 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.

[4 Marks]

Q.7. A worker node used in a Hadoop cluster has 128 GiB memory with 8 CPUs having 6 cores each (Total of 48 cores). We need to reserve 1 core each for the Operating system, HDFS DataNode and YARN NodeManager. 8 Gib memory is reserved for the Operating system and 2 Gib each for the DataNode and the NodeManager. For handling the task overheads, 4 Gib memory is to be reserved. One container with Application Master (AM) may get deployed on this node. The AM container will require 2 Gib memory and 1 Core. (Total of 4 cores reserved for OS, DN, NM and AM).

1. How many vCores will be available for deploying containers for (1) CPU intensive tasks and for (2) Standard I/O bound tasks.

2. If you are planning to use this node to launch containers for running standard I/O bound application tasks, give an estimate of the number of containers you will be deploying on this node. Describe a scheme for distributing the available resources (vCores and Memory) among these containers for achieving maximum parallelism.

[4 Marks]

Q8. Describe the most appropriate scheduling mechanism to be configured with Hadoop YARN in the following scenarios:

(a). A company wants to run jobs of equal size on a Hadoop cluster. The number of jobs can vary at any point of time.

(b). A company wants to run short jobs and long jobs on the same cluster and short jobs cannot always be starved by long jobs.

(c) A company wants to divide the Hadoop cluster resources in different proportions between the departments and does not care about wastage of unutilized resources.

(d) A company wants to run jobs of different nature concurrently on the same cluster and is worried about wastage of resources.

[4 Marks]

****