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

Mid-Semester Test
(EC-2 Regular)

Course No.	: DSECSZG522
Course Title	: Big Data Systems
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 18-01-2025 (FN)

No. of Pages = 3
No. of Questions = 10

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 each of the following scenarios, point out and give a brief reason for 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.

[3 Marks]

Q.2. Assuming a program consists of 50% non-parallelizable code.
(a) Compute the speed-up when using 2 and 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 2 and 4 processors.
(c) Explain why the speed-up in case (b) is higher than the speed-up in case (a).

[3 Marks]

Q.3. A healthcare data analytics application utilizes a Cloud-based PostgreSQL cluster. The analytics team conducts patient record analysis on the data stored on the Cloud instance. It takes approximately 30 seconds per user query on average. You are tasked to implement a Memcached based cache in the healthcare center's data center to expedite these queries, aiming for an average query latency of 10 seconds. The Memcached cache has an access latency of 5 seconds. When the cache is established, data can only be transferred from the cache to the application and not directly from the Cloud DB. What should be the projected hit rate in the cache to achieve this ambitious average query latency target? What type of locality of reference is used here? [Tavg = h * Tc + (1-h) * (Tc + Tm)]

[3 Marks]

Q.4. A 2-tier application uses a 3-Node application cluster and a 2 node DB cluster. The application works only when both tiers are available. The application tier is in an active-active load-balanced configuration with the given nodes. But the database tier is in a cold standby mode where it takes 12 hours to switch for bringing a passive node online. If an application node fails every 10 days and a DB node fails every 100 days, find the following:

(a) MTTF of the application tier and MTTF of the database tier

(b) Availability of the database tier

(c) Overall availability of the 2-tier system, assuming MTTR of the application tier is negligible

[3 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 for the given user. When you are designing your MongoDB document structure, make sure that it will give optimum query performance for your application. Select a suitable value for the field _id of the document.

[3 Marks]

Q.6. What is meant by Tunable Consistency in Cassandra NoSQL? In a 3 node Cassandra cluster, a KeySpace is created with a replication factor of 3. The Write and Read consistency levels are set in an application as given in the following scenarios:

Scenario 1: Write Level - Quorum, Read Level - One
Scenario 2: Write Level - All, Read Level - One
Scenario 3: Write Level - One, Read Level - All

Answer the following questions, in each of the scenarios mentioned above.

(1). Consistency (consistent or eventually consistent) of reads.

(2). How many nodes can be lost without data loss.

(3). What percentage of data is held in each of the nodes.

[3 Marks]

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 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.

[3 Marks]

Q.8. 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?

[3 Marks]

Q.9. A Hadoop cluster with 3 worker nodes is configured with container resources to run 4 Map tasks concurrently on each of the nodes. A file of size 12 GB is copied to HDFS with the default block size. Default number of Map tasks is selected for each block of data stored on HDFS. (a) How many Map tasks are to be executed to process the file? (b). The Map tasks get completed in how many waves of execution? (c). What are the options available to ensure that the Map tasks complete the execution in 1 wave without increasing the number of nodes of the cluster.

[3 Marks]

Q.10. In a Hadoop 2.0 cluster, jobs can be processed in 4 different queues. Fair share scheduler is plugged into YARN resource manager with queues Q0, Q1, Q2, and Q3. All these queues are allowed to preempt containers from other queues. The weights allocated for each of the queues are given below:

Q0 - weight 0.00
Q1 - weight 0.50
Q2 - weight 0.25
Q3 - weight 0.25

What percentage of total resources will be used by the jobs at run time, in each of the following scenarios?

(a) Job1 is submitted to Q1 and it is the only job running on the cluster
(b) While Job1 was running in Q1, Job2 is submitted to Q2
(c) While Job1 was running in Q1 and Job2 was running in Q2, Job3 is submitted to Q3

[3 Marks]

****