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

Mid-Semester Test
(EC-2 Makeup)

Course No.	: DSECSZG522
Course Title	: Big Data Systems
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 07-02-2025 (FN)

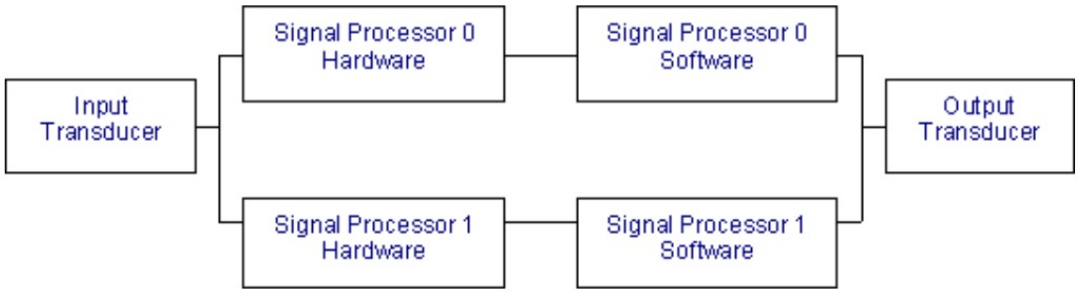
No. of Pages = 3
No. of Questions = 10

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. The block diagram of a signal processing system is shown below:



The MTBF and MTTR of each of the blocks of the system are show below:

Block	MTBF	MTTR
Input Transducer	100,000 hours	2 hours
Signal Processor Hardware	10,000 hours	2 hours
Signal Processor Software	2190 hours	5 minutes
Output Transducer	100,000 hours	2 hours

Calculate the total availability of the complete system.

[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. What are the desirable characteristics of Big Data Systems?

[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 salary of Male and Female employees. You need to output the total salary of Male and Female employees separately.

[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. You have a file of size 928 MB on HDFS as part of a Hadoop 2.x distribution. A data analytics program uses this file and runs in parallel across the cluster nodes.
(a) The cluster has 64 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 64 cores compared to a sequential version of the same program running on one core with the same file? How will this limit change if you doubled the compute power to 128 cores? 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.
(b) Suppose you could use a more scalable algorithm with 80% parallelism and the size of the file is doubled as you move to a 128-core cluster. What would be the theoretical speed-up limit for the 128 cores 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]

****