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

Mid-Semester Test
(EC-2 Makeup)

Course No.	:	DSECLZG522
Course Title	:	Big Data Systems
Nature of Exam	:	Closed Book
Weightage	:	30%
Duration	:	2 Hours
Date of Exam	:	03-02-2024 (EN)

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.

Q1. When you enter an internet café, you find the following activities performed by users on different computers connected to the internet. What type of data is generated in each of these cases?
(a) A middle-aged person transferring money using online banking to the account of his daughter studying in a different city.
(b) A teen-aged person, doing video chat with his friend in US.
(c) A young man reading online newspaper and making his comments under the news items.
(d) Another person, visiting an online e-commerce site, browsing items of his choice and adding them to the shopping cart.

[2 Marks – 0.5 Marks for each correct answer]

Q2**.** 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
(b) MTTF of the database tier
(c) Availability of the database tier
(d) Overall availability of the 2-tier system, assuming MTTR of the application tier is negligible

[4 Marks]

Q3. In a distributed computing system consisting of multiple nodes, data is accessed by an application from other nodes. The data is cached in a memory of size 10MB on the local node, and it can be delivered only from this cache to the application. Calculate the average access time of a block of data consisting of 1MB fetched from one node and processed sequentially from memory in another node of the cluster if the hit rate is 80%. You may ignore the advantages in the access time contributed by the filesystem / network caching. Typical timing values are given below:

To read 1MB Sequentially from disk = 20,000 µs

To read 1MB sequentially from memory = 250 µs

To send 1MB over 1Gbps network = 100 µs

You need to create a suitable relationship between the parameters and calculate the average access time using the formula created by you.

[4 Marks]

Q4. You are configuring a 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 for the nodes of the cluster for maximizing (1) Consistency (2) Availability, and (3) Partition Tolerance. Which consistency level will be selected by you for (1) Write intensive applications, and (2) for Read intensive applications.

[4 Marks]

Q5. You have to design a microblogging application where a user can have various types of interactions - e.g., signup, connect to users as friends, read posts from friends and others, create a new post, reply to an existing post or reply to a comment on a post. The application needs to run at a large scale with users across the globe. Discuss your choice of database (e.g., RDBMS, NoSQL) and the type(s) of configuration you will choose as per CAP Theorem. Given the example user interactions listed above, it is possible that different interaction scenarios need unique configurations. [4 Marks]

Q6. 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.

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.

[4 Marks]

Q7. What are the basic design principles of Hadoop cluster?

[4 Marks]

Q8. In a Hadoop 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. The scheduler is configured to preempt containers from other queues if required. The weights allocated for each of the queues are given below:

Q0 - weight 0.00
Q1 - weight 0.50
Q2 - weight 0.30
Q3 - weight 0.20

What will be the percentage of total capacity allocated to the jobs at run time, in each of the following scenarios?

(a) Job1 is submitted to Q3 and it is the only job running on the cluster
(b) While Job1 was running in Q3, Job2 is submitted to Q2
(c) While Job1 was running in Q3 and Job2 was running in Q2, Job3 is submitted to Q1
(d) When Job3 was the only job running in Q1, Job4 is submitted to Q0

[4 Marks]

****