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

Mid-Semester Test
(EC-2 Makeup)

Course No.	: DSECLZG522
Course Title	: Big Data Systems
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2.0 Hours
Date of Exam	: 05/Aug/2023 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. 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.

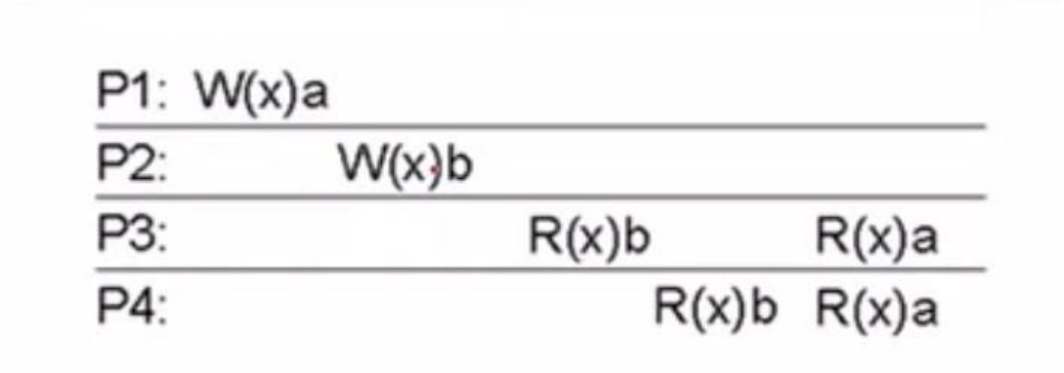
(c) A multi-machine Apache Hadoop system for data analysis.

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

(4 Marks)

Q.2. What are the 4 different levels of consistency?

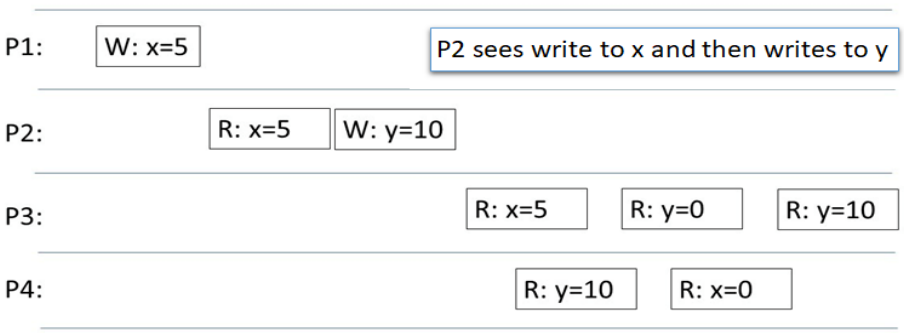
(a) Which data centric consistency level is satisfied in the following schedule?



```
P1:  W(x)a
P2:          W(x)b
P3:                  R(x)b       R(x)a
P4:                        R(x)b R(x)a
```

(b) In the following schedule, P3 adheres to a level of consistency and P4 violates it. Which is that consistency?

(Initial value of X and Y are 0)



(3 Marks)

Q.3**.** 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)

Q.4. 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 more than the speed-up in case (a).

(3 Marks)

Q.5. You have access to an expansive, nationwide dataset on farming yields. Each record in the dataset is a CSV file containing the following fields: Geographical location id (LOCID), Crop type (CTYPE), Crop yield indicator (CYI). An example record could be: L1234, Corn, 75.

Your task is to develop a list of LOCIDs where the average CYI > 90 (high yield) and another list where the CYI < 30 (low yield). The average CYI for a location is calculated over different crop types from the same location. You are required to construct a Map-Reduce program in Hadoop with either one or two iterations as necessary.

Write the map and the reduce functions, and clearly annotate your code to explain the logical steps involved. The code should be as close to Java or Python syntax for clarity, but language-specific syntax won't be penalized.

(4 Marks)

Q.6. Make a data store model using a 3 node HDFS cluster for SGP, SGPAs and CGPA of each student in 50 UG and 10 PG offerred at the University with a total of 6000 students intake capacity in each year. Each student information can extend upto 128 MB. If default replication factor is used, what is the total volume of the data in Gb occupied by the student data. If a MapReduce program is run to agregate the CGPA of all student, how many Map tasks will get executed on one node of the cluster. In resources on one node is configured to run 20 map tasks concurrently, how many waves/stages will be needed to run the Map tasks on 1 node of the cluster.

(4 Marks)

Q.7. You are given access to 10 Terabyte of data with details of daily sales transactions belonging to an e-commerce company. The transactional data files are stored in a folder on HDFS with a structure as shown below:

Transaction No. (Unique)	Invoice	Stock Code	Description	Quantity	Invoice Date	Price	Customer ID	Country

It is not required to modify the data. You are not allowed to copy or move the data into any other storage location. Select a suitable analytics tool from the Hadoop ecosystem family to run queries on this data. Give justifications for your selection. Using the tool selected by you, write a query to retrieve the total number of sales transactions in each country.

(4 Marks)

Q.8. 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. 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
(d) While Job1 was running in Q1 and Job2 was running in Q2, Job0 is submitted to Q0

(4 Marks)

****