CMI/01/05/2025/Final Exam May 2025

Chennai Mathematical Institute

Distributed Computing and Big Data

DURATION: 3 HOURS MAX MARKS: 70.

Instructions

- This is a closed book exam. You are not allowed to carry books or cheatsheets.
- No electronic devices (calculators, laptops, etc) are allowed in the exam hall.
- First section has negative marks. No negative marks for the rest of the sections.

Section 1: All questions carry one mark each. -0.5 for wrong answers. Fill in the blanks. [11 * 1 = 11 Marks]

Question 1. In the vsfs, the disk is divided into ____________ .

Question 2. Based on ____________ law, we cannot expect a linear increase in speed-up for a specific job as we increase the number of processors.

Question 3. ____________ is designed to fit in a sweet spot between the declarative style of SQL, and the low-level, procedural style of map-reduce.

Question 4. While rate limiting, in ____________ approach, every request is served in a fixed time slot.

Question 5. ____________ storage for database tables is an important factor in optimizing analytic query performance, because it drastically reduces the overall disk I/O requirements. It reduces the amount of data you need to load from disk.

Question 6. In BASE, __________ means that the system allows temporary inconsistencies before eventually achieving consistency automatically over time.

Question 7. ____________ is a binary serialization format used to store documents in MongoDB.

Question 8. The HTTP method used to create a resource while designing a RESTful service is ____________.

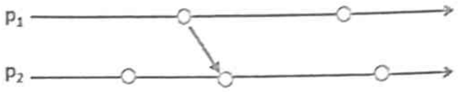
Question 9. ____________ is designed to efficiently transfer bulk data between RDBMS and Hadoop.

Question 10. Storm topology is made of sprouts and ____________.

Question 11. Native graph storage and processing uses ____________ so that it does not have to move through any other type of data structures to find links between the nodes.

Section 2: All questions carry 2 marks each. [2 * 5 = 10 Marks]

Consider the following space-time execution diagram while answering the questions in this Section.



Question 12. List all the happens-before relationships.

Question 13. Annotate the events using scalar time.

Question 14. Annotate the events using vector time.

Question 15. Annotate the events using matrix time.

Question 16. Identify an inconsistent cut.

Section 3: All questions carry 3 marks each. [8 * 3 = 24 Marks]

Question 17. Consider the following Pig script.

```
Lines = LOAD ‘file1’ USING PigStorage() as (line:chararray);
Words = FOREACH Lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
Groups = GROUP Words BY word;
Counts = FOREACH Groups GENERATE group, COUNT(Words) as Cnt;
Results = ORDER Counts BY Cnt DESC;
Dump Results;
```

Assume that the input file ‘file1’ contains the following two lines:

```
cmi is the best
the best in chennai is cmi
```

What does the pig script output?

Question 18. The following pig script was written to find the most expensive phone in each type. However, it has errors. Identify the errors and correct them.

Pig Script:

```
A = LOAD ‘file2’ USING PigStorage(‘,’) AS (year:int,type:chararray,cost:int);
B = GROUP A BY $2;
C = FOREACH B GENERATE MIN(A.cost);
DUMP C;
```

Input File (‘file2’ contains year,product,cost):

```
2022, iphone, 50000
2023, iphone, 65000
2024, iphone, 72000
```

Expected output is 72000.

Question 19. Change the pig script given above to output the price of the oldest model phone in each type known i.e., for the same input given above, the expected output is 50000.

Question 20. How many nodes and how many relationships are created when the following statements are executed by Neo4j? Draw the graph generated by the above Neo4j commands.

```
1. CREATE (p:Person{name:'Venkatesh'})-[:Teaches]->(c:Course{name:'BigData'})
2. CREATE (p:Person {name:'Raj'})-[:StudentOf]->(o:Org{name:'CMI'})
3. CREATE (p:Person {name:'Raj'})
4. MATCH (a:Person),(b:Org) WHERE a.name = 'Venkatesh' AND b.name = 'CMI'
   CREATE (a)-[:FacultyAt]->(b)
```

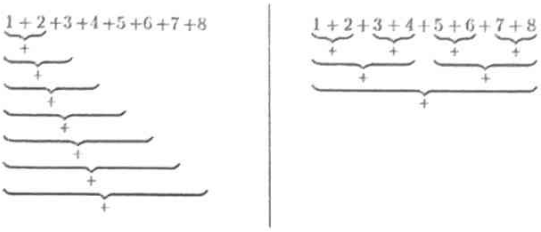
Question 21. In the muddy children puzzle, as discussed in the class, what would the children say if n=3 and k=2? i.e., there are three children, and two of them have muddy forehead and to start with, they are told that at least one of them have muddy forehead. Assume that each child can only say 'yes', 'no' or 'dont know'. Use the following format to provide your answer.

```
Round1: [<1st child response>,<2nd ...>,<3rd ...>]
Round2: [<1st child response>,<2nd ...>,<3rd ...>]
...
Last Round:  [<1st child response>,<2nd ...>,<3rd ...>]
```

Question 22. A file system is configured to have a block size of 4 MB. Given that a inode of a file holds pointers to 8 direct data blocks, and a pointer to a single indirect block. Further, assume that the single indirect block can hold pointers to 4 other data blocks. What is the maximum file size that can be supported by such an inode design?

Question 23. Chunnalal configured a HDFS instance to have 1 MB sized blocks and three replicas. Pannalal configured a HDFS instance to have 256 KB sized blocks and four replicas. Considering only these factors, which instance is better tuned to save a file of size 500 KB? Explain Why.

Question 24. Our ability to write parallelizable programs decides the speed up we can achieve through scaling. Consider two programs to add large list of numbers. The first program adds the first two numbers, remembers the result, and adds that result to the next number. It continues doing this until the end of the list is reached. The second program adds two numbers at a time in parallel. It recursively does so until the final results are arrived at. The following figure explains their logic with an example of eight numbers.



Assume that the list is large and the numbers may be unordered. As per Amdahl's law, assuming each addition is an operation, how much speedup (approximately) can these programs achieve if there are four processors?

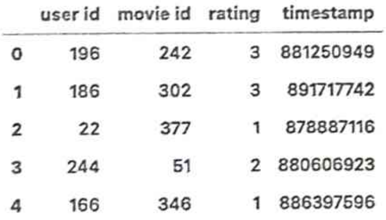
Section 4: All questions carry 5 marks each. [5 * 5 = 25 Marks]

Question 25. With a sequence diagram, explain how oAuth 2.0 works.

Question 26. Design a RESTful web service for storing and searching for books in a library. Include at least three resources in your object model.

Question 27. Design a twitter-like system. You may make reasonable assumptions on the scope.

Question 28. You are provided with the ML-100K dataset. It has the following data: user id, movie id, rating and timestamp. See few example rows in the picture below. Describe a map-reduce design to generate a sorted list of the most highly rated movies.



	user id	movie id	rating	timestamp
0	196	242	3	881250949
1	186	302	3	891717742
2	22	377	1	878887116
3	244	51	2	880606923
4	166	346	1	886397596

Question 29. Assume that bookmyshow wants to store the movie reservation information in MongoDB. Provide a database design along with at least two queries as example to indicate how you would query the data.