CMI/22/4/2023/Final Exam Apr 2023

Chennai Mathematical Institute

Distributed Computing and Big Data

DURATION: 3 HOURS MAX MARKS: 35.

Instructions

- Please remember to mention your name and roll number in your answer sheet.
- This is an individual task. Do not discuss with anyone.
- 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. Wherever heavy calculation is involved, you need not evaluate it to the final number unless it is explicitly asked for. For example, it is acceptable to leave the answer as $\frac{1}{1+\frac{5}{32}}$. You need not evaluate it to 0.865.
- 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. Answer in True/False.

Question 1. The name notwithstanding, there are most definitely servers in serverless computing. 'Serverless' describes the developer's experience with those servers—they are are invisible to the developer, who doesn't see them, manage them, or interact with them in any way. True

Question 2. Poorly maintained data lakes are often called Data Swamps. True

Question 3. Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications. True

Question 4. For the services that do not expose metrics, we can use the ladder based scaling strategy. Scaling ladders can be defined per million concurrent users on the platform (1M, 2M ..). This works well for predictable workloads. True

Question 5. A Content Delivery Network (CDN) is a distributed network of servers that are geographically distributed across the globe. True

Question 6. Load balancer is a device or software that distributes incoming network traffic across multiple servers. True

Question 7. Pods are the smallest deployable units of computing that you can create and manage in Kubernetes. True

Question 8. Neo4j is not ACID compliant. False

Question 9. Redis is a key-value store and MongoDB is a document Store. True

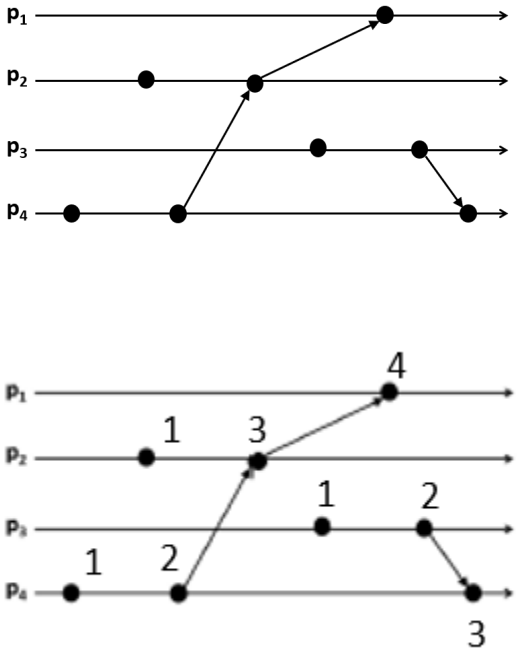
Question 10. On the CAP triangle, MongoDB falls on the AP side and Cassandra falls on the CP side. False

Question 11. General-purpose computing on graphics processing units is the use of a GPU, which typically handles computation only for computer graphics, to perform computation in applications traditionally handled by the CPU. True

Section 2: All questions carry two marks each.

Question 12. If only 10% computation can be executed in parallel, and if we have 10 processors, what is the best speed up achievable as per Amdahl's law?
R = 90%, P = 10. Best speedup = $\frac{1}{\frac{90}{100}+\frac{1-90/100}{10}} = \frac{100}{91} = 1.0989$

Question 13. Annotate the following space-time diagram with scalar time.



Question 14. For the same space-time diagram as in the previous question, annotate p_2 events with matrix time.

$$p_2^1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad p_2^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Question 15. What will be the output of the following pig script if the input file, file1, contains a single line "1,V Rao,40,Chennai"?

```
A = LOAD 'file1' USING PigStorage(',')
    AS (id:int, name:chararray, age:int,  city:chararray);
B = FOREACH A GENERATE TOKENIZE(name);
DUMP B;
```

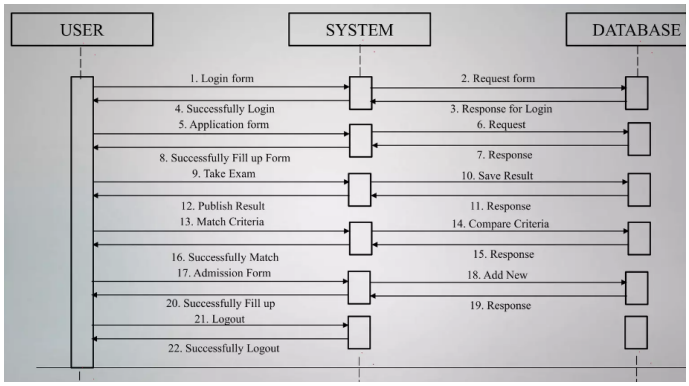
Solution: ((V),(Rao))

Question 16. What will be the output of the following pig script if the input file, file2, contains numbers 1 to 10, each in one line (i.e., 1 in first line, 2 in second line, and so on)?

```
A = Load 'file2' using PigStorage(',') as (num:int);
B = Foreach A generate 1 as gid, num;
C = Group B by gid;
D = Foreach C generate SUM(B.gid);
Dump D;
```

Solution: (10)

Question 17. Draw a sequence diagram to capture the CMI's admission process. Should have at least two objects, swim lanes, and life lines. A sample solution is here.



Question 18. Ramesh bought a hard disk with rotational delay of 3ms. With what RPM does the disk spin? If it had 20 sectors per track, what is its read time?

Solution: (10,000 RPM. Read time $= \frac{6}{20} = 0.3$ms)

Question 19. How many nodes are created when the following statement is executed by Neo4j?

```
create (p:Person {name:'Venkatesh'})-[:Teaches]->(c:Course {name:'BigData'})
```

Solution: 2 nodes

Section 3: All questions carry 4 marks each.

Question 20. The Chennai Public School wants to automate its system for grading students. Specifically, this system will allow creation, modification and deletion of exam marks and student grades from class V to class X. Design a RESTful web service for this scenario. You may scope your answer to three identified resources. Your answer must cover at least one idempotent method assignment and one non-idempotent method assignment.

Question 21. From a very large text file, we need to find the least five frequent words that contain only alphabets (i.e., no digits, no punctuations, etc). Describe a map-reduce design pattern to achieve the same.
