CS 544 Exam 2 (20%) - Fall 2023

Instructor: Tyler Caraza-Harter
First/Given Name: ______________________. Last/Surname: ______________________
Net ID: ____________________________________ @wisc.edu
Fill in these fields (left to right) on the scantron form (use pencil):

1. LAST NAME (surname) and FIRST NAME (given name), fill in bubbles
2. IDENTIFICATION NUMBER is your Campus ID number, fill in bubbles
3. Under A of SPECIAL CODES, tell us about the nearest person (if any) to your left. 0=no person to the left in your row, 1=somebody you do not know is there, 2=somebody you do know is there.
4. Under B of SPECIAL CODES, do the same as B, but for the person to your right
5. **Under C of SPECIAL CODES, write 3 and fill in bubble 3**. This is very important!

Make sure you fill all the special codes above accurately in order to get graded.

You have 2 hours to take the exam. Use a #2 pencil to mark all answers. When you're done, please hand in these sheets in addition to your filled-in scantron. You may not sit adjacent to your friends or other people you know in the class (having only one empty seat is considered "adjacent"). You may only reference your notesheet. You may not use books, your neighbors, calculators, or other electronic devices on this exam. Please turn off and put away portable electronics now.

If multiple answers are correct, choose the best answer.

(Blank Page for You to Do Scratch Work)

Q1. The single NameNode in an HDFS cluster is becoming a bottleneck. The cluster contains a small number of files, but each is extremely large. What is most likely to help alleviate load on the NameNode?

(A) add more DataNodes
(B) increase the block size
(C) decrease the block size
(D) split the few big files into many small files

Q2. The query engine for BigQuery is internally based on what system?

(A) GFS (B) Dremel (C) Spark (D) MapReduce

Q3. For which increase in K should we expect the biggest decrease in the time it takes the loop to execute?

Answer with respect to this loop we discussed during lecture:

```
for (int i = 0; i < arr.Length; i += K) arr[i] *= 3;
```

Assume `arr` contains 32-bit integers and the CPU has 64-byte cachelines. Increase K...

(A) from 1 to 2 (B) from 2 to 4 (C) from 8 to 16 (D) from 16 to 32

Q4. Assuming 2x replication, which node(s) are responsible for row token -5, assuming the following token map?

`token(n1) = [-6], token(n2) = [-5, 7], token(n3) = [4, -2]`

Feel free to annotate the following if it is helpful:

-8|-7|-6|-5|-4|-3|-2|-1| 0| 1| 2| 3| 4| 5| 6| 7

(A) n1+n2 (B) n1+n3 (C) n2+n3

Q5. For which one do you NOT usually need to write custom code when using Kafka?

(A) producers (B) brokers (C) consumers

Q6. What does COS stand for, in the context of Google's cloud?

(A) Container-Optimized OS (B) Cloud Operating System (C) Container Orchestration Service

Q7. You want to connect from a browser on your laptop to Jupyter running in a container on your VM. You take the following steps:

1. Write a command in the Dockerfile to launch Jupyter on port 32366
2. Use `-p 32357:32366` in the `docker run ...` command
3. Use `-L localhost:3923:localhost:32357` when establishing the SSH tunnel
4. Enter `http://localhost:????/` in the browser

What should `????` be in step 4?
(A) 8888 (B) 3923 (C) 32357 (D) 5000 (E) 32366

Q8. A single Spark task typically runs on __________ and operates on ___________.

(A) one core, one partition
(B) one core, many partitions
(C) multiple cores, one partition
(D) multiple cores, many partitions

Q9. Which SQL clause is responsible for projection?

(A) WHERE (B) HAVING (C) GROUP BY (D) SELECT (E) PROJECT

Q10. Is the following function idempotent?

```python
def set_abs():
    global x
    x = abs(x) # absolute value
```

(A) Yes (B) No

Q11. In Kafka, we can separately configure the replication factor and min in-sync replicas. When is a message considered committed?

(A) it has been writen to min in-sync replicas
(B) it has been written to all the in-sync replicas
(C) it has been written to a number of replicas equal to the replication factor

Q12. Cassandra Quorums: Given W=8 and RF=9, what should R be to make sure readers see successful writes? If multiple satisfy this, choose the smallest correct.

(A) 1 (B) 2 (C) 5 (D) 7

Q13. What is the primary reason HDFS does pipelined writes?

(A) to minimize CPU load on the NameNode
(B) to avoid assigning too much network load to a single machine
(C) so each node in the pipeline can be responsible for a single transformation
(D) so that progress can be checkpointed (that is, committed) after each stage of the pipeline

Q14. You want to do a streaming GROUP BY with Spark, using a Kafka topic as a source. However, the column you want to group by in Spark is different than the column used to set the key for the messages in the Kafka topic. Does Spark support such a query?

(A) Yes (B) No

Q15. True/False: when a thread is holding a lock during a critical section, the scheduler WILL NEVER context switch to another thread in the same process.

(A) True (B) False

Q16. At what granularity can a user specify the replication factor for HDFS?

(A) per cluster (B) per keyspace (C) per file (D) per block

Q17. During the first few epochs of optimization with PyTorch, your loss increases, before becoming Inf. What should you do?

(A) use a smaller learning rate (B) use a bigger learning rate

Q18. If you have little RAM but excess CPU resources, what Spark caching level is probably best? You are not concerned about load balance or fault tolerance. (NOTE: D was originally mislabeled as B)

(A) MEMORY_ONLY (B) MEMORY_ONLY_SER (C) MEMORY_ONLY_2 (D) MEMORY_ONLY_SER_2

Q19. For an LRU cache of size 4, how many hits are there for the following workload?

1, 1, 2, 3, 4, 1, 2, 3, 4

(A) 1 (B) 3 (C) 4 (D) 5 (E) 0.5

Q20. True/False. The PLANET algorithm (implemented by Spark) sometimes collects all the rows corresponding to a single node of a decision tree on a single machine.

(A) True (B) False

Q21. What is the following?

```
message {
        string key = 1;
        string name = 4;
        int32 age = 5;
}
```

(A) Python class (B) bytecode (C) C struct (D) protocol buffer (E) Kafka topic

Q22. HDFS is most similar to which proprietary system?

(A) Artifact Registry (B) BigQuery (C) Colossus (B) Dynamo

Q23. In BigQuery, both `ML.EVALUATE` and `ML.PREDICT` can take a query as the second argument. In which case must that query produce results with a label column?

(A) `ML.EVALUATE` (B) `ML.PREDICT`

Q24. You attempt a Cassandra INSERT with a primary key that is already used by one row that is already in the table (the table was created with a cluster key). What happens?

(A) the insert is ignored
(B) an error is raised
(C) previous row is updated
(D) there will be two rows with the same primary key

Q25. In the clause `FROM first, second`, what kind of JOIN is done between the tables?

(A) CROSS (B) INNER (C) LEFT (D) RIGHT

Q26. Consider the following Kafka messages. What can we guarantee about which messages will go to the same partition?

1. topic="z", key="z", value="X"
2. topic="X", key="X", value="z"
3. topic="z", key="X", value="X"

(A) 1 and 2 will go to the same partition
(B) 1 and 3 will go to the same partition
(C) 2 and 3 will go to the same partition
(D) We can't guarantee anything

Q27. In a Dockerfile, how do you specify the program that should launch (by default) when a container starts?

(A) `EXEC` (B) `RUN` (C) `CMD` (D) `DO`

Q28. What technique is used when updating multiple replicas of a Cassandra token ring data structure?

(A) gossip (B) quorums (C) pipelined writes (D) at-most-once semantics

Q29. Which BigQuery billing model uses "leftover" CPU and memory resources?

(A) capacity (B) on-demand (C) rollover (D) spare

Q30. Which storage device usually costs less, in terms of $/GB?

(A) HDD (B) SSD