

CS 544 Exam 3 (16%) - Fall 2024

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 1 and fill in bubble 1.** 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.

Do not communicate with anybody besides the teaching team about questions or answers until exam grades have been posted.

(Blank Page for You to Do Scratch Work)

Q1. What is something that capacity billing gives BigQuery users for free?

- (A) CPU (B) memory (C) Colossus I/O (D) Colossus Storage

Q2. Assuming 2x replication, which node(s) are responsible for row token 3, assuming the following token map?

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

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) only n1 (B) only n2 (C) n1+n2 (D) n1+n3 (E) n2+n3

Q3. 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

Q4. When is BHJ most beneficial?

- (A) both tables are small (B) one table is small and one is large (C) both tables are large

Q5. If you want to run BigQuery over data in the Capacitor format, how should you add tables to your dataset?

- (a) load job (b) external table

Q6. Assume the X variable starts at 5, and there is a global lock named `lock` used by two threads running the following code concurrently. What are the possible X values at the end?

```
# thread 1
with lock:
    X *= 2

# thread 2
with lock:
    X += 1
```

- (A) only 11 (B) only 12 (C) 11 or 12 (D) 6, 10, 11, or 12 (E) 5, 6, 10, 11, or 12

Q7. What does gRPC use to serialize messages?

- (A) ColumnIO (B) JSON (C) Parquet (D) Protocol Buffers

Q8. What is something that Kubernetes does that Compose does not do?

- (A) bin packing
- (B) use cgroups to isolate performance
- (C) deploy multiple replicas from the same Docker image

Q9. A single HDFS file is being read by many different clients, and HDFS is having trouble keeping up. What is most likely to help?

- (A) disable pipelines writes
- (B) increase the replication factor
- (C) decrease the replication factor

Q10. Is the following function idempotent?

```
def set_square():
    global x
    x = x ** 2
```

- (A) Yes
- (B) No

Q11. Which system uses pipelined writes to send data to all the workers that will store a new piece of data?

- (A) HDFS
- (B) Spark
- (C) Cassandra
- (D) Kafka

Q12. What generally costs more when deploying on a cloud? Option 1: one VM for 100 hours. Option 2: 100 VMs for 1 hour.

- (A) option 1 costs more
- (B) option 2 costs more
- (C) the costs are similar

Q13. Consider the following Kafka messages. What can we guarantee about which messages will go to the same partition?

1. topic="W", key="Z", value="X"
2. topic="W", key="X", value="X"
3. topic="X", key="Z", value="Z"

- (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

Q14. How does HBase assign data to RegionServers? Assume we are using 3x replication.

- (A) each column will belong to one RegionServer
- (B) each column will belong to three RegionServers
- (C) each region will belong to one RegionServer
- (D) each region will belong to three RegionServers

Q15. How many hits are there for a FIFO cache of size 3 for the following workload?

6, 4, 2, 3, 4, 3, 6, 6

- (A) 0 (B) 1 (C) 2 (D) 3 (E) 4

Q16. You have a high-volume Kafka topic. The brokers are able to keep up, but the consumers cannot keep up. What is most likely to help?

- (A) use more topic partitions
(B) use fewer topic partitions
(C) start more consumer groups
(D) start more consumers per consumer group

Q17. In Spark streaming, is the following stateless?

`SELECT MAX(x) AS total FROM mystream;`

- (A) yes (B) no

Q18. If you want to filter rows before they are grouped, what do you use in SQL?

- (A) HAVING (B) LIMIT (C) WHERE

Q19. Which of the following are immutable in Spark?

- (A) only Pipeline (B) only PipelineModel (C) both Pipeline and PipelineModel

Q20. Say you want to run a streaming Spark query over a Kafka topic. The topic is partitioned by column X, but the query is grouping by a different column, Y. What will happen?

- (A) Spark will refuse to run the query
(B) Spark will produce incorrect outputs
(C) Spark will be able to group correctly by column Y

Q21. Which of the following uses a gossip protocol for updating information about cluster membership?

- (A) HDFS (B) Spark (C) Cassandra (D) Kafka

Q22. Cassandra Quorums: Given W=4 and RF=9, what should R be to make sure readers see successful writes? If multiple satisfy this, choose the smallest correct.

- (A) 2 (B) 4 (C) 5 (D) 6

Q23. What Linux tool can help you see what process is using a port?

- (A) ls (B) ns (C) os (D) ps (E) ss

Q24. What technique does HDFS use to DETECT DataNode failures?

- (A) partitioning (B) replication (C) heartbeats (D) block maps (E) hashing

Q25. 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 2241
2. Use `-p 3752:2241` in the `docker run ...` command
3. Use `-L localhost:4432:localhost:3752` when establishing the SSH tunnel
4. Enter `http://localhost:????/` in the browser

What should `????` be in step 4?

- (A) 2241 (B) 5000 (C) 8888 (D) 4432 (E) 3752

Q26. You run `SELECT FUNC(geom) FROM geotable` in BigQuery. Which FUNC will generally result in more output rows?

- (A) ST_CENTROID (B) ST_CENTROID_AGG (C) ST_CENTROID and ST_CENTROID_AGG tie

Q27. If you do a correlated cross join between columns y and z (after unnesting each), how many rows will you get?

x, y, z
1, [2, 3], [4, 5]
6, [7], [8, 9, 10]

- (A) 0 (B) 2 (C) 4 (D) 7 (E) 15

Q28. What best describes cloud organization?

- (a) zones contain regions (b) regions contain zones (c) clusters contain regions

Q29. Which system(s) have a leader/follower approach to replication?

- (A) only HDFS (B) only Spark (C) only Cassandra (D) only Kafka (E) both HDFS and Cassandra

Q30. A Docker container `myapp` is running in detached mode, so you cannot immediately see what the process started by `CMD` is printing. How can you see that output?

- (A) `docker ps myapp`
(B) `docker logs myapp`
(C) `docker exec -it myapp`
(D) `docker exec myapp stdout`