CS 544 Exam 3 (15%) - Fall 2025

Instructors: Meena Syamkumar and 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 A, 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. When BigQuery computes query cost based on bytes of storage I/O, how does it round I/O?

(A) rounds up (B) rounds down

Q2. A Cassandra table has three columns: X (first column, a partition key), Y (second column, a cluster key), and Z (third column, regular column). You insert these rows:

- (2,2,4)
- (3,2,2)
- (3,2,4)
- (2,1,5)
- (2,1,2)

How many rows will be in the table?

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

Q3. In Spark, `.cache` is a convenience method that calls `.persist(...)` with what setting?

(A) MEMORY_ONLY (B) MEMORY_ONLY_SER (C) DISK_ONLY (D) DISK_ONLY_2

Q4. A client writes a 80 MB file to HDFS with 1x replication. The block size is 16 MB. How much data does the client send over the network?

(A) 5 MB (B) 16 MB (C) 80 MB (D) 112 MB (E) 400 MB

Q5. Which non-cloud platform is most similar to Google's BigQuery?

(A) Spark (B) Cassandra (C) Kafka (D) HBase (E) BigTable

Q6. What does `docker ps` show?

(A) what images Docker has locally
(B) what containers are running
(C) what processes are running within a container

Q7. The following Python library is complete, except that locking has only been used for `f` and `g`. Which other functions must acquire the lock to avoid race conditions?

```
lock = threading.Lock()
x = 2
y = 1
z = 2

def f():
    global y, x
    with lock:
        y += z
        x += 1

def g():
    global z
    with lock:
        z *= x

def A():
    print(x)

def B():
    print(y)

def C():
    print(z)
```

(A) A and C
(B) C only
(C) A and B
(D) A only
(E) A, B, and C

Q8. You need to run a once-a-day batch job that can wait if the VM is temporarily unavailable, and it is not critical if it gets interrupted. Which type of VM instance is most suitable?

(A) on-demand instances (B) spot instances

Q9. How many cache hits are there for the following workload?

D, D, A, D, B, A, A, C

Assume LRU eviction and cache size 3.

(A) 4 (B) 5 (C) 6 (D) 7 (E) 8

Q10. There are 4 Kafka groups, each with 5 consumer(s). All the groups are subscribed to the same topic, T. A new message in T will be consumed how many times?

(A) 1 (B) 3 (C) 4 (D) 5 (E) 20

Q11. How does a NameNode determine which DataNodes are live in the cluster?

(A) gossip (B) leader election (C) heartbeats (D) pipelines (E) checksums

Q12. When you power off a cloud VM, what do you usually still pay for while it is off?

(A) memory only (B) CPU and memory (C) CPU only (D) disk capacity

Q13. Cassandra Quorums: Given R=5 and RF=10, what should W be to make sure readers see successful writes? If multiple satisfy this, choose the smallest correct.

(A) 1 (B) 4 (C) 6 (D) 8

Q14. Which join implementation is preferable?

You have a Spark cluster with 50 machines, each with 64 GB of memory. You need to join two tables. Smaller table: 9.6 GB. Bigger table: 31655.5 GB. As long as you don't run out of memory, your goal should be to minimize network I/O.
(A) SMJ (Shuffle Sort Merge Join)
(B) BHJ (Broadcast Hash Join)

Q15. How do you redirect ONLY the standard output from program X to file Y?

(A) X > Y (B) X -> Y (C) X | Y (D) X & Y (E) X &> Y

Q16. Consider the following Kafka messages. What can we guarantee about which messages will go to the same partition?

1. topic="green", key="purple", value="red"
2. topic="red", key="blue", value="purple"
3. topic="green", key="green", value="red"

(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

Q17. Is `do_it` idempotent?

```
data = [3, 8, 2]

def do_it(val, d):
    data[d] = val
```

(A) Yes (B) No

Q18. The following code is about to run in a container on a VM. The VM has 4 GB of RAM currently available, the container was launched with `-m 2g`, and `big.file` is 3 GB. You do NOT have swap enabled.

```
with open("big.file", "rb") as f:
    data = f.read()
```

Will memory constraints PREVENT the code from running successfully?
(A) yes (B) no

Q19. What split points does PLANET consider?

(A) between any unique values
(B) thresholds between bins in an equi-width histogram
(C) thresholds between bins in an equi-depth histogram

Q20. Given the following Kafka partition state, is message E committed?

```
Leader: A, B, C, D, E
Follower 1 (lagging): A, B
Follower 2 (in-sync): A, B, C, D, E
Follower 3 (in-sync): A, B
```

(A) Yes (B) No

Q21. What characteristic does Cassandra's design prioritize?

(A) availability (B) atomicity (C) consistency (D) isolation

Q22. Assuming 2x replication, which node(s) are responsible for a new row being inserted?

Row: x="red", y="green", z="blue". The primary key is ("x", "y").

Assume hash("red")=2, hash("green")=-2, hash("blue")=-5, hash(<"red", "green">)=4, and hash(<"red", "green", "blue">)=-8.

Token map:

```
token(n1) = [4, -4], token(n2) = [-1, -5], token(n3) = [6]
```

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

Q23. Which I/O pattern is most challenging for SSDs?

(A) random reads (B) random writes (C) sequential reads (D) sequential writes

Q24. You have an HDFS file, F, that rarely changes, but it is very popular: clients read F so often that DataNodes storing blocks of the file cannot keep up with requests. It will not be catastrophic if F is lost, because you can just execute a MapReduce job to regenerate the contents of F as needed. What would be better, from a performance perspective?

(A) decrease replication factor for F (B) increase replication factor for F

Q25. You have a URL someprotocol://someaddr:someport/someresource. Which part will determine the specific running process on a machine that will receive the request?

(A) someprotocol (B) someaddr (C) someport (D) someresource

Q26. SQLX is an extension of SQL provided by which of the following?

(A) Arrow (B) BigQuery (C) Cassandra (D) Dataform (E) GCS

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
5 , [6, 7], [8, 9]
10, [11]  , [12, 13]
```

(A) 2 (B) 6 (C) 7 (D) 8 (E) 9

Q28. What usually costs more for cloud network I/O?

(A) ingress (B) egress

Q29. For which one do you NOT usually need to write custom code when using Kafka?

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

Q30. What value(s) could possibly be printed?

```
x = 4
lock = threading.Lock()

def task():
    global x
    with lock:
        x = x - 3

t = threading.Thread(target=task)
t.start()
with lock:
    x = x * 3
t.join()
print(x)
```

(A) only 9 (B) only 1 (C) only 3 (D) 3 or 9 (E) only 12