CS 544 Exam 3 (19%) - Spring 2025

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 8 and fill in bubble 8**. 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. A Spark streaming query is maintaining a count for an interval starting at 1am. At what time could Spark reasonably discard the running count for events occuring in this interval?.

```
(animals.withWatermark("timestamp", "4 hours")
 .groupBy(window("timestamp", "2 hours"))
 .count())
```

(A) 2am (B) 3am (C) 5am (D) 7am (E) 8am

Q2. You are developing a library for generating unique IDs. You don't want to rely solely on random number generation because you don't want any chance (however small) of different computers using the same library to produce the same ID. What information about the machine where the code is running would be most helpful for generating a truly unique ID?

(A) IP address (B) MAC address (C) port number of current process (D) free disk space

Q3. You write 5 MB to a 2x replicated file in HDFS, then later read it back. How much data will be read and written to disks across the cluster?

(A) 10 MB written, 5 MB read
(B) 5 MB written, 5 MB read
(C) 10 MB written, 10 MB read
(D) 5 MB written, 10 MB read

Q4. The query engine for BigQuery is internally based on what system?

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

Q5. There are 2 Kafka groups, each with 1 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) 2 (C) 4 (D) 7 (E) 8

Q6. What kind of service is EC2?

(A) IaaS (B) PaaS

Q7. To connect to an HDFS cluster, what does a client need, at a minimum?

(A) address of any DataNode
(B) addresses of all the DataNodes
(C) address of the NameNode
(D) addresses of NameNode and all DataNodes

Q8. Which BigQuery billing model uses "leftover" CPU and memory resources?

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

Q9. A Kafka topic has a replication factor of 3. How will new data be written to the replicas?

(A) The client will write the message directly to the leader and both followers.
(B) The client will write the message to the leader, and the followers will later fetch it.
(C) The client will write the message to the leader, which will actively send it to both followers.
(D) The client will send the data to the leader, the leader will send it to the first follower, and the first follower will send it to the second follower.
(E) The client will send the data to the first follower, the first follower will send it to the second follower, and the second follower will send it to the leader, at which point it will be committed.

Q10. How many hits are there for a FIFO cache of size 3 for the following workload?
4, 2, 6, 4, 1, 6, 1, 5

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

Q11. In an HDFS cluster, load is poorly balanced across DataNodes. What is most likely to help improve balance?

(A) using smaller blocks (B) using bigger blocks

Q12. Which join algorithm uses hash partitioning to bring rows from each table that could potentially match with each other together on the same machine?

(A) SMJ (B) BHJ

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

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

(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. For Cassandra, R=4 and W=7. Readers are guaranteed to see previous writes. What can we infer about RF?

If multiple answers are correct, choose the answer that provides the tightest bound on RF.

(A) RF >= 12 (B) RF > 11 (C) RF < 11 (D) RF >= 11 (E) RF < 12

Q15. A Kafka partition leader fails, and there are three followers. Which are eligible to become the new leader?

- Follower 1: in-sync, and has all messages that the old leader had
- Follower 2: in-sync, but is missing 10 messages that the old leader had
- Follower 3: lagging, but is missing 1 message that the old leader had

(A) only 1 (B) 1 or 2 (C) 1 or 3 (D) 1, 2, or 3

Q16. Assuming 2x replication, which node(s) are responsible a new row being inserted?

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

Assume hash("red")=-5, hash("green")=5, hash("blue")=-7, hash(<"red", "green">)=6, and hash(<"red", "green", "blue">)=3.

Token map:

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

Q17. You have 4 billion floating point operations to do on a device capable of 8 GFLOPS. How many seconds will it take to do the operations?

(A) 0.002 (B) 0.5 (C) 1 (D) 2.0 (E) 2000.0

Q18. Which format inspired Parquet?

(A) Arrow (B) Capacitor (C) ColumnIO (D) Protocol Buffers

Q19. 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) 2 (B) 4 (C) 7 (D) 8 (E) 16

Q20. What are two systems that inspired the design of Cassandra?

(A) BigTable and Dynamo
(B) BigTable and MapReduce
(C) BigQuery and Dynamo
(D) BigQuery and MapReduce

Q21. How do "free tiers" usually work for cloud services?

(A) you are not charged for initial operations up to some limit
(B) you are not charged for additional operations after exceeding some limit

Q22. Which clause related to machine-learning does BigQuery add to SQL?

(A) TEST (B) TRAIN (C) TRANSFORM (D) TRANSPOSE

Q23. For RAM, what is the finest granularity at which every piece of memory has its own address?

(A) bit (B) byte (C) cacheline (D) page (E) block

Q24. What value(s) could possibly be printed?

```
x = 4
def task():
    global x
    x += 7
t = threading.Thread(target=task)
t.start()
t.join()
print(x)
```

(A) 4 or 7 (B) only 11 (C) only 4 (D) 4 or 11 (E) only 7

Q25. 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

Q26. In Docker, if you want a file/directory location on the host machine to be visible within a container, what flag should you pass to `run`?

(A) `-d` (B) `-f` (C) `-p` (D) `-v`

Q27. 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 2927
2. Use `-L localhost:4424:localhost:3219` to establish the SSH tunnel
3. Use `-p ????:2927` in the `docker run ...` command
4. Enter `http://localhost:4424/` in the browser

What should `????` be in step 3?
(A) 8888 (B) 4424 (C) 2927 (D) 5000 (E) 3219

Q28. Which Spark caching levels will use the JVM types to represent data?

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

Q29. For the below Spark SQL query, over which column(s) will hash values be calculated for hash partitioning?

SELECT MIN(P), K FROM mytable GROUP BY K, L;

(A) K (B) L (C) P and K (D) P (E) K and L

Q30. What Linux command provides documentation about how to use a given program?

(A) wget (B) which (C) cat (D) du (E) man