CS 544 Exam 2 (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 5 and fill in bubble 5**. This is very important!

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

You have 40 minutes 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. Is the below data layout column oriented or row oriented?

Table:

```
3,5,1
4,2,6
```

Disk layout: 3,5,1,4,2,6

(A) column oriented (B) row oriented

Q2. You write 100 MB to a 3x replicated file in HDFS, then later read it back. How much data will be read and written to disks across the cluster?

(A) 100 MB written, 100 MB read
(B) 100 MB written, 300 MB read
(C) 300 MB written, 100 MB read
(D) 300 MB written, 300 MB read

Q3. What determines the number of output files generated by a MapReduce job?

(a) # of map tasks (B) # of map calls (C) # of reduce tasks (D) # of reduce calls

Q4. For a MapReduce job, you have 1000 input key/value pairs, 500 intermediate key/value pairs, and 30 output key/value pairs. Among the 1000 inputs, there are 8 unique keys. How many times will `map(...)` be called?

(A) 8 (B) 30 (C) 500 (D) 1000

Q5. What statement about the memory requirements for running the PLANET algorithm is correct?

(A) all training data must fit within the memory of a single machine
(B) all training data must fit in the cumulative memory available across all machines in the cluster
(C) it is OK if training data does not fit in memory across the cluster

Q6. Cassandra Quorums: Given W=6 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) 8

Q7. In Cassandra, what node(s) can server as a coordinator for a read to row X?

(A) only the boss
(B) only the first worker encountered when walking the ring starting from X's token
(C) only workers with a replica of X
(D) any worker

Q8. You have 3 billion floating point operations to do on a device capable of 6 MFLOPS. How many seconds will it take to do the operations?

(A) 0.5 (B) 1 (C) 2 (D) 500 (E) 2000

Q9. How many hits are there for a LRU cache of size 4 for the following workload?
4, 5, 5, 7, 2, 7, 6, 4

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

Q10. 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

Q11. You are using Spark to join two large tables that are roughly equal in size. A large number of worker machines will be involved. What join algorithm should you pick?

(A) Broadcast Hash Join (B) Shuffle Sort Merge Join

Q12. 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 2323
2. Use `-p 3400:2323` in the `docker run ...` command
3. Use `-L localhost:4566:localhost:3400` when establishing the SSH tunnel
4. Enter `http://localhost:????/` in the browser

What should `????` be in step 4?
(A) 2323 (B) 5000 (C) 4566 (D) 3400 (E) 8888

Q13. For the below Spark SQL query, over which column(s) will hash values be calculated for hash partitioning?

SELECT SUM(X), Y FROM mytable GROUP BY Y, Z;

(A) X (B) Y (C) Z (D) X,Y (E) Y,Z

Q14. 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

Q15. If you have lots of RAM, which caching level will generally be fastest?

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

Q16. 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`

Q17. Assuming 2x replication, which node(s) are responsible for row token -6, assuming the following token map?

```
token(n1) = [-2, 7, -3], token(n2) = [1, 4, -6], token(n3) = [3, -7]
```

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

Q18. What is `filter` in Spark?

(A) action (B) projection (C) transformation

Q19. What technique does HDFS use to DETECT DataNode failures?

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

Q20. 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