CS 544 Exam 2 (20%) - Spring 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. 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

Q2. In Spark streaming, is the following stateless?

```
SELECT SUM(x+y) AS total FROM mystream;
```

(A) yes (B) no

Q3. Spark uses the PLANET algorithm to train decision trees (DTs). The type of job that runs on a set of DT nodes depends on whether those nodes have few enough rows to run the in-memory algorithm. For the in-memory case, the job uses hash partitioning. What does it partition rows on?

(A) the first column in the row
(B) the DT node to which the row belongs
(C) the value in the column on which we're splitting

Q4. 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

Q5. Cassandra uses consistent hashing. For what does Cassandra use a hash function to get a token on the token ring?

(A) only for data (B) only for nodes (C) for both data and nodes

Q6. In which machine learning platform(s) are models immutable objects?

(A) only PyTorch (B) only Spark (C) both PyTorch and Spark

Q7. Assume `count` starts at 6, and two threads are running concurrently. For simplicity, assume: there is a single CPU core, context switches only occur between lines of Python code, and code/instructions within a single thread are not re-orderded by any system (such as the compiler or CPU).

```
# thread 1
if count > 5:
    with lock:
        count -= 5

# thread 2
if count > 3:
    with lock:
        count -= 3
```

What is the smallest possible final count?

(A) -2 (B) 0 (C) 1 (D) 3 (E) 6

Q8. In Cassandra, let RF=2, R=1, and W=1. The token of row X is -7. What nodes are responsible for storing X?

```
      n1    n2    n3    n2    n1    n2    n3
-8|-7|-6|-5|-4|-3|-2|-1| 0| 1| 2| 3| 4| 5| 6| 7
```

(A) n1 (B) n3 (C) n1 and n2 (D) n1 and n3 (E) n2 and n3

Q9. Which Spark type is used for a pipeline model that has NOT been fit to the data yet?

(A) Pipeline (B) PipelineModel

Q10. Consider the following Kafka messages. What can we guarantee about which messages will go to the same partition?

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

Q11. Which of the following Spark SQL clauses does NOT use hash partitioning?

(A) JOIN (B) WHERE (C) GROUP BY

Q12. What is the following?

```
message {
        int32 x = 1;
        int32 y = 2;
}
```

(A) Python class (B) bytecode (C) C struct (D) protocol buffer (E) PyTorch array

Q13. 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:

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

How many rows will be in the table?

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

Q14. A new cloud customer just wants to create one VM. What will be the cheaper option?

(A) sole-tenant host (B) multi-tenant host

Q15. A Spark streaming query is maintaining a count for an interval starting at 6am. At what time could Spark reasonably discard the running count for events occuring in this interval? (Note, fixed typo "B" => "C" after exam).

```
(animals.withWatermark("timestamp", "4 hours")
 .groupBy(window("timestamp", "1 hours"))
 .count())
```

(A) 7am (B) 9am (C) 10am (D) 11 am

Q16. For which resource does Google charge under both BigQuery billing models (capacity and on-demand)?

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

Q17. You are running low on storage space. Which Linux utility will be most useful as you try to free up at least 3 GB of storage space?

(A) chmod (B) df (C) htop (D) kill (E) pkill

Q18. If you want to get a bash shell for debugging purposes inside a container that is already started, what Docker command can you use?

(A) exec (B) logs (C) ps (D) run (E) shell

Q19. If you have lots of RAM, which caching level will generally be fastest?

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

Q20. 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`

Q21. What kind of database is MySQL?

(A) OLTP (B) OLAP

Q22. Assume you have an LRU cache of size 3. How many hits will there be for the following workload?

W, X, Y, Z, Z, Y, X, W

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

Q23. 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 command in Dockerfile to launch Jupyter on port 1000
2. use `-p 2000:1000` in the `docker run ...` command
3. use `-L localhost:3000:localhost:2000` when establishing the SSH tunnel
4. enter `http://localhost:????/` in the browser

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

(A) 1000 (B) 2000 (C) 3000 (D) 5000 (E) 8888

Q24. Say your bloom filter uses 2 hash functions and 10 bits. The bits contain this:

`1000000001`

hash1(X)%10=0, hash2(X)%10=9

Was X previously inserted?

(A) yes (B) no (C) maybe

Q25. What is accurate about Spark?

(A) a DStream is a series of RDDs
(B) an RDD is a series of DStreams

Q26. True/False: when one thread is holding a lock L, another thread that calls `L.acquire()` will be blocked until the first thread releases the lock.

(A) True (B) False

Q27. What kind of service does AWS EC2 provide?

(A) VMs (B) DBs (C) file storage (D) streaming

Q28. You're housesitting for your friend. They are texting you directions. If you don't reply quickly enough, they obnoxiously keep repeating the same directions. What is an example of an idempotent text they could send you?

(A) feed the cat
(B) close any open windows
(C) flip the light switch in the kitchen

Q29. Which format is most unlike the other three in terms of its use cases?

(A) Capacitor (B) ColumnIO (C) Parquet (D) Protobuf

Q30. Which non-cloud platform is most similar to Google's BigQuery?

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