CS 544 Exam 2 (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 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. What technique does HDFS use to DETECT DataNode failures?

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

Q2. 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 2019
2. Use `-p 3886:2019` in the `docker run ...` command
3. Use `-L localhost:4626:localhost:3886` when establishing the SSH tunnel
4. Enter `http://localhost:????/` in the browser

What should `????` be in step 4?
(A) 2019 (B) 3886 (C) 4626 (D) 5000 (E) 8888

Q3. Assuming 2x replication, which node(s) are responsible for row token 3, assuming the following token map?

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

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) n1+n2 (B) n1+n3 (C) n2+n3

Q4. What query language(s) support JOINs?

(A) just CQL (B) just SQL (C) both CQL and SQL

Q5. Say an HDFS client is writing data to 3 DataNodes in a pipeline. Which DataNode will do the LEAST network I/O?

(A) 1st DataNode (B) 2nd DataNode (C) 3rd DataNode (D) they all do the same amount of network I/O

Q6. What Linux tool can help you see what process is using a port?

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

Q7. 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

Q8. You start a container `bright-spark` 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 bright-spark`
(B) `docker logs bright-spark`
(C) `docker exec -it bright-spark`
(D) `docker exec bright-spark stdout`

Q9. 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

Q10. What is the signature of a map function in MapReduce?

(A) `f(key, value)` (B) `f(keys, value)` (C) `f(key, values)` (D) `f(keys, values)`

Q11. At what granularity can a user specify the block size in HDFS?

(A) per cluster (B) per keyspace (C) per directory (D) per file

Q12. What expression most closely resembles the calculation done during hash partitioning to assign a row to a partition? Assume there are N partitions.

(A) row.key (B) hash(row.key) (C) row.key % N (D) hash(row.key) % N (E) hash(row.key % N)

Q13. Which caching level will be BETTER in terms of load balance?

(A) MEMORY_ONLY (B) MEMORY_ONLY_2

Q14. You are joining a large table X with a smaller table Y using Spark. X is too big to fit in memory on a single machine, but Y can fit. Which join algorithm(s) could we use? NOTE: we're not asking which is faster, just what can run given memory limits.

(A) neither will work (B) only SMJ will work (C) on BHJ will work (D) either SMJ or BHJ will work

Q15. An HBase "compaction" primarily involves what operation?

(A) compressing a file (B) merge sorting small files (C) reallocating regions to RegionServers when machines are removed from a cluster

Q16. What does the "A" in OLAP stand for?

(A) action (B) access (C) aggregation (D) analytical

Q17. What is most likely to DECREASE how much memory a NameNode needs in an HDFS cluster?

(A) adding more DataNodes
(B) increasing the block size for the files
(C) increasing the replication factor for the files

Q18. Is the below data layout column oriented or row oriented?

Table:

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

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

(A) column oriented (B) row oriented

Q19. Cassandra Quorums: Given W=8 and RF=8, what should R be to make sure readers see successful writes? If multiple satisfy this, choose the smallest correct.

(A) 1 (B) 2 (C) 3 (D) 6

Q20. Assume x starts at "A" and y starts at 0. After running the following threads together, what is the biggest possible value for y? 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).

```python
# thread 1
if x == "B":
    y += 1
if x == "C":
    y += 2

# thread 2
x = "B"
x = "C"
```

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