CS 544 Exam 2 (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 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 60 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. For what kind of model does Spark use the PLANET algorithm for training?

(A) ALS
(B) LinearRegression
(C) NaiveBayes
(D) DecisionTree
(E) LogisticRegression

Q2. 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

Q3. Which system primarily provided the inspiration for HBase?

(A) Arrow (B) BigTable (C) Cassandra (D) Dynamo (E) HDFS

Q4. Consider this MapReduce program:

```
def map(key, value):
    # movie format: (title, year, genre)
    title, year, genre = value
    emit(year, 1)

def reduce(key, values):
    emit(key, sum(values))
```

What does this MapReduce job compute?
(A) The total number of movies across all years
(B) The number of movies released in each year
(C) The most popular year for movie releases
(D) The average number of movies per year
(E) The list of movie titles grouped by year

Q5. How does HBase assign data to RegionServers? Assume we are using 4x replication.

(A) each region will belong to one RegionServer
(B) each region will belong to 4 RegionServers
(C) each column will belong to 4 RegionServers
(D) each column will belong to one RegionServer

Q6. 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

Q7. True/False: in MapReduce, different rows with the same KEY will sometimes be passed to different `map` calls, running on different machines.

(A) True (B) False

Q8. Which join implementation is preferable?

You have a Spark cluster with 10 machines, each with 32 GB of memory. You need to join two tables. Smaller table: 2.1 GB. Bigger table: 227.5 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)

Q9. 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:

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

How many rows will be in the table?

(A) 0 (B) 1 (C) 2 (D) 4 (E) 5

Q10. True/False: when you add an additional DataNode to an HDFS cluster, the majority of physical blocks in the cluster must be reassigned to a different DataNode.

(A) True (B) False

Q11. Which line of the below Spark code will probably take longest to run?

```
df2 = df.filter(filter_fn).cache() # line 1
print(df2.mean())                  # line 2
print(df2.sum())                   # line 3
```

(A) line 1 (B) line 2 (C) line 3

Q12. Assuming 2x replication, which node(s) are responsible for row token -7, assuming the following token map?

`token(n1) = [-8, 3, -1], token(n2) = [2], token(n3) = [-5]`

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

Q13. What does the "which" command print out?

(A) the PATH variable (B) a path to a program (C) the Linux distro version (D) the shell being used

Q14. In HDFS, between what nodes are heartbeats sent?

(A) NameNode to DataNode (B) DataNode to NameNode
(C) NameNode to client (D) client to NameNode

Q15. If you want multiple Python threads to be able to run at the same time on different CPU cores, what kind of Python runtime do you need?

(A) with GIL (B) without GIL

Q16. The "hdfs dfs ..." commands use which NameNode interface?

(A) RPC calls (B) REST calls

Q17. For the below Spark SQL query, over which column(s) will hash values be calculated for hash partitioning?

SELECT T, SUM(R) FROM mytable GROUP BY K, T;

(A) R and T (B) T (C) K (D) R (E) K and T

Q18. What query language(s) support JOINs?

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

Q19. What type of Spark operation produces an RDD?

(A) action (B) transformation

Q20. In Cassandra, imagine your partitions are too big. If you want more partitions, each smaller, how might you adapt your schema?

(A) more columns in the partition key
(B) fewer columns in the partition key
(C) more cluster columns
(D) fewer cluster columns

Q21. What is NOT a feature built into gRPC?

(A) for a failed call, it will automatically retry to a different server
(B) for small integers, it will use variable length encoding to save space
(C) it allows clients and servers to be written in different programming languages
(D) it allows clients and servers to have different versions of a protocol in some cases

Q22. You write 150 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) 150 MB written, 300 MB read
(B) 150 MB written, 150 MB read
(C) 300 MB written, 300 MB read
(D) 300 MB written, 150 MB read

Q23. How many hits are there for a FIFO cache of size 3 for the following workload?
7, 3, 2, 5, 1, 5, 2, 3

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

Q24. Which of the following uses a gossip protocol for updating information about cluster membership?

(A) HDFS only (B) Spark only (C) Cassandra only (D) HDFS and Spark

Q25. Is the below data layout "column oriented" or "row oriented"?

Table:

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

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

(A) column oriented (B) row oriented