CS544 - Spring 2024

Instructor: Meenakshi Syamkumar

Exam 2 — 20%

(Last) Surname: ______________________ (First) Given name: ______________________

NetID (email): ____ @wisc.edu

Fill in these fields (left to right) on the scantron form (use #2 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 **F** of SPECIAL CODES, write **1** and fill in bubble **1**

If you miss step 3 above (or do it wrong), the system may not grade you against the correct answer key, and your grade will be no better than if you were to randomly guess on each question. So don't forget and double check it's correct!

You may only reference your note sheet. You may not use books, calculators, or other electronic devices during this exam. You may not sit near your friends or look at your neighbors during this exam. Please place your student ID face up on your desk. Turn off and put away portable electronics (including smart watches) now.

Use a #2 pencil to mark all answers. DO NOT USE PEN on the scantron.

When you're done, please hand in the exam and note sheet and your filled-in scantron form. The note sheet will not be returned.

(Blank Page)

1. For an LRU cache size of 4, what is the hit rate for the below access pattern? Assume that the cache is empty prior to the below access pattern.

```
P, Q, P, R, S, Q, T, P
```

A. 0 **B. 0.25** C. 0.375 D. 0.75 E. 1

2. Which of the following operations is not supported by Spark streaming?

   A. group by

   B. group by time intervals

   C. inner join

   D. watermarks

   E. pivots

3. Which of the following methods enables us to create a HIVE table?

   A. `createTempView`

   B. `createOrReplaceTempView`

   C. `createGlobalTempView`

   D. `saveAsTable`

4. Is the following function idempotent?

```
x = 10

def inverse_x():
    global x
    x = 1 / x
```

A. yes **B. no**

5. Suppose a Kafka producer is producing numbers to a particular topic "nums" and it has produced these numbers so far: 10, 20, 30. If a Spark streaming query, consuming the messages in the "nums" topic to calculate the product of the numbers, produces an output of 600, what semantics does this whole system seems to be providing?

A. at-most-once B. at-least-once C. exactly-once

6. Which SQL clause enables us to perform projection?

A. SELECT B. FROM C. WHERE D. GROUP BY E. ORDER BY

7. Suppose we want to use Spark to join two tables using a small number of worker machines. If one of those tables fits entirely into memory, which of the following join algorithms should we pick?

   A. Broadcast Hash Join B. Shuffle Sort Merge Join

8. A client is writing 5 MB of data to a 4x replicated HDFS file. Assuming pipelined writes, how much data does the client send over the network?

   A. 4 MB **B. 5 MB** C. 9 MB D. 20 MB

9. Which of the following is the fundamental data structure of Spark?

   A. DataFrame B. table C. view D. protocol buffer **E. RDD**

10. Which of the following techniques is used to avoid reading identical copies of the same data when Cassandra read quorum `R > 0`?

    A. pipelined reads B. caching **C. checksum** D. compression

11. Which Spark type is used for a decision tree model that has NOT been fit to the data yet?

    A. DecisionTreeRegressor B. DecisionTreeRegressionModel

12. What is the following an example of?

```
FROM ubuntu:23.10
RUN apt-get update && apt-get install -y unzip python3 python3-pip
RUN pip3 install pandas===2.1.0 --break-system-packages
```

    A. `yml file` B. `protocol buffer` **C.** `Dockerfile` D. `nodetool`

13. Consider the following Kafka messages. What can we guarantee about the order in which these messages will be consumed?

```
1. topic="sports", key="A", value="hello"
2. topic="sports", key="B", value="sports fans"
3. topic="weather", key="B", value="sunny"
4. topic="weather", key="B", value="and bright"
5. topic="international", key="A", value="overseas"
6. topic="international", value=10
```

   A. msg 1 before msg 5

   B. msg 2 before msg 3

   C. msg 3 before msg 4

   D. msg 4 before msg 3

   E. msg 5 before msg 6

14. In Kafka, both consumers and followers send fetch requests to the leader. Who can fetch uncommitted messages in Kafka?

   A. only consumers

   B. only followers

   C. both consumers and followers

   D. neither consumers nor followers

15. Suppose you have a column of data with the following values:

```
apple, apple, apple, banana, banana, orange, orange
```

If we decided to represent the data using the below format, what technique(s) are we using?

```
{3: 1, 2: 2, 2: 3}
{"apple": 1, "banana": 2, "orange": 3}
```

   A. only dictionary encoding

   B. only run-length encoding

   C. both dictionary encoding and run-length encoding

16. Suppose `banks_df` is a Spark DataFrame containing a column called “name”. Which of the following lambda expressions enables us to filter all rows that contain “First” as part of the bank names?

    A. `lambda banks_df:  "First" == banks_df["name"]`

    B. `lambda b:  "First" in b["name"]`

    C. `lambda banks_df:  "First" in banks_df["name"]`

    D. `lambda b:  "First" == b["name"]`

17. How do you make predictions using Spark ML implementations?

    A. invoke `transform` method on unfit model

    B. invoke `predict` method on unfit model

    C. invoke `fit` method on fitted model

    D. invoke `predict` method on fitted model

    E. invoke `transform` method on fitted model

18. Considering Kafka, suppose `RF=5`, `min.insync.replicas=3`. If there are currently 4 in-sync replicas and 1 out-of-sync aka lagging replica, what is the minimum number of replicas to which a message needs to be written for us to say it is committed?

    A. 1 B. 2 C. 3 **D. 4** E. 5

19. A portion of code we don’t want to be interrupted by another thread is called a ________?

    A. context switch **B. critical section** C. lock D. collision

20. In Cassandra, suppose you have a table named “customers” inside the below keyspace.

```
create keyspace banking
with
replication={'class': 'SimpleStrategy',  'replication_factor': 3};
```

Consider the below token map:

```
token(n1) = {5, 8}
token(n2) = {11, 15}
token(n3) = {-2, 12}
token(n4) = {-5, 20}
```

The partition key of the new row to be inserted into the “customers” table hashes to -1. What are the tokens of all the vnodes that are responsible for storing this new row?

A. n1

B. n1 and n3

C. n1 and n2

D. n1, n2, and n3

E. n1, n2, n3, and n4

21. Consider the below Cassandra token map:

```
token(n1) = {5, 8}
token(n2) = {11, 15}
token(n3) = {-2, 12}
token(n4) = {-5, 20}
```

Assume that node n5 joins the cluster with vnodes 9 and -4. Which existing node(s) will pass off some data to this new node?

A. n1 and n2

B. n2 and n3

C. n2 and n4

D. n1 and n4

E. n4 and n5

22. A 20x5 PyTorch tensor is storing double precision floats. How many bytes does the tensor consume (without counting overhead of the Python object)?

    A. 64

    B. 200

    C. 400

    D. 800

    E. 51200

23. Suppose you have created and trained a Bigquery machine learning model. Which of the following will enable you to determine the coefficients used to multiply features?

    A. `ML.PREDICT` B. `ML.OPTIONS` C. `ML.EVALUATE` **D.** `ML.WEIGHTS`

24. Which of the following is not a column-oriented format?

    A. Capacitor **B. CSV** C. ColumnIO D. Parquet

25. Consider the below Bigquery query and its corresponding output. Recall that `language` column contains `REPEATED RECORDS`.

```
SELECT repo_name, ARRAY_LENGTH(language) as total_languages
FROM `bigquery-public-data.github_repos.languages`
WHERE ARRAY_LENGTH(language) > 200
```

	repo_name	total_languages
0	polyrabbit/polyglot	216

Suppose we execute the below query, how many rows will be in `df`?

```
%%bigquery df
SELECT *
FROM `bigquery-public-data.github_repos.languages`, UNNEST(language)
WHERE repo_name = "polyrabbit/polyglot"
```

A. `0` B. `1` C. `2` D. `200` **E.** `216`

26. Which of the following is most similar to HDFS?

    A. Colossus B. BigTable C. BigQuery D. HBase E. Dynamo

27. In Bigquery, which of the following functions enables us to convert floating point longitude-latitude into geographic data?

    A. `ST_LATLONG` **B.** `ST_GEOGPOINT` C. `ST_CENTROID` D. `ST_MAKEPOINT`

28. Consider the below Cassandra queries:

```
CREATE TABLE sample(
    X INT,
    Y INT,
    Z TEXT,
    PRIMARY KEY ((X), Y)
);

INSERT INTO sample (X, Y, Z) VALUES (1, 1, 'a');
INSERT INTO sample (X, Y, Z) VALUES (2, 1, 'b');
INSERT INTO sample (X, Y, Z) VALUES (1, 1, 'c');
INSERT INTO sample (X, Y, Z) VALUES (2, 1, 'd');
INSERT INTO sample (X, Y, Z) VALUES (3, 1, 'e');
```

How unique values will be in the `Z` column of the “sample” table?

A. 1 B. 2 **C. 3** D. 4 E. 5

29. Which of the following systems is most similar to Spark?

A. Colossus B. BigTable **C. BigQuery** D. HBase E. Dynamo

30. Which of the following is NOT a Spark transformation operation?

A. `filter` B. `parallelize` **C.** `mean` D. `map`

31. Consider the below Cassandra token map:

```
token(n1) = {5, 8}
token(n2) = {11, 15}
token(n3) = {-2, 12}
token(n4) = {-5, 20}
```

What is the wrapping range?

A. `> 15 <= 20` **B.** `> 20` C. `>= 20` D. `<= -5` E. `< -5`

32. Consider Spark streaming. Is the following query stateless?

```
SELECT 1/x AS inverse FROM some_stream;
```

A. yes B. no

33. Considering Cassandra quorums, suppose `RF=10` and `R=5`, what should `W` be to make sure that we read the latest successful write?

A. 1 B. 2 C. 4 D. 5 **E. 6**

34. For a FIFO cache size of 4, what is the hit rate for the below access pattern? Assume that the cache is empty prior to the below access pattern.

    ```
    P, Q, P, R, S, Q, T, R
    ```

    A. 0   B. 0.25   **C. 0.375**   D. 0.75   E. 1

35. Suppose you want to search for “UN” inside `output.txt`, what should replace `???` in the following command?

    ```
    cat output.txt ??? grep "UN"
    ```

    A. `&`   B. `>`   C. `&>`   D. `>>`   **E.** `|`

(Blank Page)