

CS 544 (CS 639-4): Intro to Big Data Systems

Spring 2023 Final Practice

Note: the final is cumulative, but this only focuses on material since the midterm.

Q1. You want to run a "docker build . -t ????" command to prepare for bringing up compose services defined in the following. What should "???" be?

```
services:
  main:
    image: myimg
    ports:
      - "127.0.0.1:5000:5000"
      - "127.0.0.1:4040:4040"
    volumes:
      - "./nb:/notebooks"
      - "./main.sh:/start.sh"
```

(A) nb (B) myimg (C) main (D) main.sh (E) image (F) services

Q2. What is the biggest limitation of MapReduce?

- (A) operations (like JOIN) involving two tables cannot be done with a single MapReduce job because mappers can only take in one file
- (B) it cannot scale to a large number of machines
- (C) it only works when the data fits in the cumulative memory of all the machines in the cluster
- (D) saving intermediate data after each job is inefficient

Q3. You have 10 Spark executors, each with 20 CPU cores and 64 GB of RAM. How many tasks can execute simultaneously?

(A) 30 (B) 10 (C) 20 (D) 200 (E) 640 (F) 64 (G) 2

Q4. You try caching your DataFrame using MEMORY_ONLY and find you use X GB of RAM. Then you try using DISK_ONLY and find you use Y GB of disk space. Which is bigger?

(A) X (B) Y

Q5. How does a broadcast hash join work in Spark?

- (A) both tables are sent to every executor
- (B) only the small table is sent to every executor
- (C) only the big table is sent to every executor

Q6. There are 5 partitions, numbered 0 to 4. Spark is hash partitioning some data across all these for the purpose of a GROUP BY on column X. The hash of the X value in a particular row is 10. What partition will this row go to?

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

Q7. How does HBase assign data to RegionServers?

- (A) each column will belong to one RegionServer
- (B) each row will belong to three RegionServers
- (C) each row will belong to one RegionServer
- (D) each column will belong to three RegionServers

Q8. 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)
- (1,2,3)
- (1,2,4)

How many rows will be in the table?

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

Q9. What operations does Cassandra support?

- (A) some JOINS and no GROUP BYs
- (B) all GROUP BYs and some JOINS
- (C) all JOINS and some GROUP BYs
- (D) some GROUP BYs and no JOINS

Q10. In Cassandra's use of consistent hashing, we use a hash function to choose tokens for which of the following?

(A) tables (B) columns (C) vnodes (D) nodes (E) rows

Q11. If $RF=4$ and $W=2$, what should R be in Cassandra if you want reads to see the latest successful write?

(A) 3 (B) 1 (C) 6 (D) 5 (E) 4 (F) 2

Q12. During normal operation, a Cassandra worker needs to look up the value for a key in its LSM tree. The key appears in the memtable, twice in the commit log, and once in an SSTable. Where will it find the value?

(A) memtable
(B) commit log (newer version)
(C) commit log (older version)
(D) SSTable

Q13. In which system do read quorum configurations often require the involvement of multiple replicas to serve a read operation?

(A) HDFS (B) Spark (C) Cassandra (D) Kafka

Q14. What can a consumer call when it wants partitions automatically assigned by Kafka?

(A) subscribe (B) list_topics (C) seek (D) assign

Q15. In Kafka, both consumers and followers send fetch requests to the leader. Who can read uncommitted messages?

(A) only followers
(B) only consumers
(C) neither consumers nor followers
(D) both consumers and followers

Q16. Kafka: suppose RF=5 and min in-sync replicas is 3. There are currently 4 in-sync replicas, with 1 lagging. A message is written to 3 replicas (the leader and two followers). Is it committed?

(A) no (B) yes

Q17. When is the `spark.sql.shuffle.partitions` configuration most important?

(A) batching (B) streaming

Q18. Numbers are being written to a Kafka topic that is read by a streaming Spark query. The Spark query is adding the numbers and showing the output. Three numbers have been produced so far: 100, 3, 4. The latest output from the Spark query is 207.

What semantics does the system (as a whole) seem to be providing?

(A) exactly once (B) at least once (C) at most once

Q19. The data owned by a node in a decision tree has 10 rows and 5 columns of feature data. No value appears more than once in the table.

How many different ways are there to split this node?

(A) 40 (B) 5 (C) 50 (D) 36 (E) 45 (F) 4 (G) 9 (H) 10

Q20. You are considering two ways to produce train and test data in Spark:

1. `train, test = df.randomSplit([0.50, 0.50])`
2. `train, test = df.randomSplit([0.50, 0.50], seed=544)`

Which way will always produce a deterministic split?

(A) only 1 (B) only 2 (C) both ways (D) neither way

Q21. On what kind of hosts did we deploy our VMs this semester?

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

Q22. Your table has 100 rows. You want to calculate statistics related to each unique value of column A in the table. Column A contains 20 unique values. The result set of your query contains 100 rows. What kind of operations seem to have been performed?

- (A) the rows were grouped by A, then an aggregate function was applied to each group
- (B) the rows were partitioned by A, then a window function was applied to each partition

Q23. You have lots of memory, but are tight on CPU resources. Which caching level is probably best for you?

- (A) MEMORY_ONLY
- (B) MEMORY_ONLY_SER

Q24. Within a static column of a Cassandra table, there is at most one value corresponding to each _____.

- (A) cluster key
- (B) partition key
- (C) primary key

Q25. To what is the space usage of the token map proportional?

- (A) Only the number of nodes in a Cassandra ring
- (B) Only the number of rows in Cassandra tables
- (C) To both number of nodes and number of rows

Q26. The wrapping range of a Cassandra ring consists of tokens that are...

- (A) $<$ smallest vnode token
- (B) \leq smallest vnode token
- (C) $>$ biggest vnode token
- (D) \geq biggest vnode token

Q27. Kafka: suppose $RF=4$ and min in-sync replicas is 2. There are currently 3 in-sync replicas and one lagging.

A message is written to two replicas (leader and one follower). Is it committed?

- (A) yes
- (B) no

Q28. Say your bloom filter uses 3 hash functions and 10 bits. The bits contain this:
0111000100

$\text{hash1}(X)\%10=2$, $\text{hash2}(X)\%10=9$, and $\text{hash3}(X)\%10 = 2$

If you ask whether X was inserted into the bloom filter, what result will you get?

(A) False (B) Maybe (C) True

Q29. Six messages are produced in the following order:

1. topic=A, key=X, value=8
2. topic=A, key=Y, value=8
3. topic=B, key=Y, value=8
4. topic=B, key=Y, value=7
5. topic=C, value=9
6. topic=C, value=10

What can we say about the order in which these will be consumed?

- (A) msg 1 before msg 2
- (B) msg 3 before msg 4
- (C) msg 4 before msg 3
- (D) msg 2 before msg 3
- (E) msg 5 before msg 6

Q30. 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 dog
- (B) set the thermostat to 62
- (C) water the house plants

Q31. Is it stateless? (in the context of Spark streaming).

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

- (A) yes (B) no

Q32. Spark is maintaining a count for an interval starting at 2pm. At what time could Spark reasonably discard the running count for events occurring in this window?

```
(animals.withWatermark("timestamp", "1 hours")  
.groupBy(window("timestamp", "2 hours"))  
.count())
```

(A) 3pm (B) 4pm (C) 5pm (D) 6pm

Q33. What is a method you WILL NOT be using when working with Spark models?

(A) fit (B) predict (C) transform

Q34. A feature column has these numbers, in the range of 0 to 100:

1, 2, 8, 9, 15, 16, 80, 90

What histogram would a Spark decision tree preferably compute for 4 bins?

(A) 0-25, 25-50, 50-75, 75-100

(B) 0-5, 5-10, 10-20, 20-100

Q35. You have a column of data with these numbers: 7,8,8,8,7,7,9. You decide on a new way to represent this series of numbers: 7{1}, 8{3}, 7{2}, 9{1}. What technique(s) are you using here?

(A) only dictionary encoding

(B) only run-length encoding

(C) both dictionary encoding and run-length encoding

Q36. Say you have a table like this:

```
x, y  
1, ["x"]  
2, ["y", "z"]
```

If you have a correlated cross join between the table and the unnesting of column y, how many rows of output will there be?

(A) 2 (B) 3 (C) 4 (D) 5 (E) 6

ANSWER KEY

Q1: B
Q2: D
Q3: D
Q4: A
Q5: B
Q6: A
Q7: C
Q8: B
Q9: D
Q10: E
Q11: A
Q12: A
Q13: C
Q14: A
Q15: A
Q16: A
Q17: B
Q18: B
Q19: E
Q20: D
Q21: A
Q22: B
Q23: A
Q24: B
Q25: A
Q26: C
Q27: B
Q28: A
Q29: B
Q30: B
Q31: A
Q32: C
Q33: B
Q34: B
Q35: B
Q36: B