NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING
Semester 2 AY2021/2022
CS5425/CS4225 – Big Data Systems for Data Science
Final Test
APR 2022 Time Allowed: 1 Hour 30 Minutes

INSTRUCTIONS TO STUDENTS:

1. This assessment paper contains **FIVE (5)** questions and comprises **SIXTEEN (16)** printed pages, including this page.

2. This is an **OPEN BOOK** test. You can use computer, pads, books, notes and any other supporting materials. **In case you use electronic devices, the internet connection must be turned off.**

3. Students are required to answer ALL the questions.

4. For MCQ Questions Q1 and Q2:
   (1) **Shade your answers on the OCR Answer sheet using a 2B pencil**. You need to hand in both the OCR sheet AND this paper at the end of the test.
   (2) We also suggest you to circle your answer in the test paper for later checking.

5. For other questions: Write your answers within the space provided. Answers written on other parts of the answer script will not be graded unless you specify explicitly.

6. **You should finish the test strictly on your own**. Do not discuss/share/copy with other students. **We have zero tolerance on plagiarism and cheating.**

Class: CS5425 CS4225
Matriculation Number: _____

This portion is for examiner's use only

Question	Marks	Remarks
Q1 [10 marks]		
Q2 [18 marks]		
Q3 [6 marks]		
Q4 [6 marks]		
Q5 [4 bonus marks]		
Total [40 marks + 4 bonus marks]		

The questions relevant to mid-term are highlighted.

QUESTION 1: Choose the most appropriate option from the available options. Each question weighs one mark. [10 marks]

(1) Sensor readings have precision issues or errors. Such scenario shows _____ of big data.
(A) Volume (B) Velocity (C) Veracity (D) Variety (E) None of the above answers Answer: C
(2) In GFS/HDFS, the chunk size is set to 64MB by default. What if we set this chunk size to be much smaller (say, 4KB)? Which of the following statements are True? S1. If the chunk size is too small, the disk I/O performance will degrade. S2. If the chunk size is too small, the task parallelism will degrade.
(A) Only S1 is True (B) Only S2 is True (C) Both S1 and S2 are True (D) Both S1 and S2 are False (E) None of the above answers Answer: A
(3) There are many development features in Spark that show the trend of convergence between relational databases and Spark. Which of the following features are True for this convergence? S1. The schema support S2. Query optimization S3. High level languages
(A) Only S1 and S2 are correct. (B) Only S2 and S3 are correct. (C) Only S1 and S3 are correct. (D) S1, S2 and S3 are all correct. (E) S1, S2 and S3 are all wrong. Answer: D

(4) Big data systems may have different mechanisms for reliability. Which of the following statements are True?

S1. In HDFS, each chunk is replicated for three times by default.

S2. In Spark, RDD uses lineage for reliability.

S3. Spark does not use replication for reliability, because replication would consume too much memory and the main memory is very costly.

(A) Only S1 and S2 are correct.

(B) Only S2 and S3 are correct.

(C) Only S1 and S3 are correct.

(D) S1, S2 and S3 are all correct.

(E) S1, S2 and S3 are all wrong.

Answer: D

(5) Which of the following statements are True?

S1. In topic-sensitive PageRank, the node with the highest topic-sensitive PageRank score is always in the teleport set.

S2. For the same input graph, personalized PageRank will generate the same ranking for any set of web pages.

(A) Only S1 is True

(B) Only S2 is True

(C) Both S1 and S2 are True

(D) Both S1 and S2 are False

(E) None of the above answers

Answer: D

(6) Which of the following are the responsibilities of Hadoop namenode?

S1. Managing the file system namespace

S2. Coordinating file operations

S3. Maintaining overall health

S4. Storing actual data chunks

(A) S1, S2, and S3

(B) S1, S2, and S4

(C) S1, S3, and S4

(D) S2, S3 and S4

(E) None of the above answers

Answer: A

(7) What are the advantages of column store compared to row store?

S1. Better read efficiency when reading many fields

S2. Better compression

S3. Vectorized processing

S4. Opportunities to operate directly on compressed data

(A) S1, S2, and S3

(B) S1, S2, and S4

(C) S1, S3, and S4

(D) S2, S3 and S4

(E) S1, S2, S3 and S4

Answer: D [If reading many fields, row store has better efficiency. Thus, S1 is false]

(8) Which of the following statements about RDD are True?

S1. RDD is main logical data unit in Spark.

S2. RDD partitions can be stored on the disk of different machines in a cluster.

S3. RDDs support transformations and actions.

S4. RDDs support both read and update operations.

(A) S1, S2, and S3

(B) S1, S2, and S4

(C) S1, S3, and S4

(D) S2, S3 and S4

(E) S1, S2, S3 and S4

Answer: A [RDD does not support update, is immutable. Thus S4 is false.]

(9) What are the benefits of Storm for real-time processing?

S1. High performance

S2. Good fault tolerance

S3. Reliability

S4. Scalability

(A) Only S1, S2, and S3

(B) Only S1, S2, and S4

(C) Only S1, S3, and S4

(D) Only S2, S3 and S4

(E) S1, S2, S3 and S4

Answer: E

(10) Which of the following about HDFS is NOT correct?

(A) Suitable for large batch of data read and write

(B) Suitable for concurrent writing of the same file

(C) Suitable for sequential reads and writes

(D) Not suitable for interactive applications due to high latency

(E) None of the above answers

Answer: B [HDFS is not suitable for concurrent writes from many processes, B is false]

QUESTION 2: Choose the most appropriate option from the available options. Each question weighs 2 marks. [18 marks]

(11) Consider the below data stream consisting of integers. What is the estimated cardinality of the stream by applying Flajolet-Martin (FM) Counter?

47, 34, 60, 42, 9, 30, 7, 5, …

Suppose we use the hash function h(a)= a % 128 in the FM counter (note, % is modulus).

(A) 2

(B) 4

(C) 8

(D) 16

(E) None of the above answers

Answer: B

R=2, thus, estimated count =2^2=4

47	101111
34	100010
60	111100
42	101010
9	1001
30	11110
7	111
5	101

(12) Jim has a task to cluster a large set of points which are stored in HDFS in a cluster of 32 servers. Jim has implemented the k-means algorithm in Hadoop. In his implementation, each iteration of k-means is implemented by one Hadoop job. Taking

a point as input, the mapper outputs the point's nearest cluster number as the key and the point itself as the value. The reducer computes the new centroid for each cluster, which is emitted as the value with the key being the cluster number.

If Jim runs the Hadoop program in a single server. He finds that the performance is much lower than a k-means implementation written in Java from scratch. Which of the following could be the cause?

S1. Hadoop is designed for distributed executions. It has the overhead of runtime scheduling and network oriented designs.

S2. Hadoop needs to repeatedly write HDFS.

(A) Only S1 is True

(B) Only S2 is True

(C) Both S1 and S2 are True

(D) Both S1 and S2 are False

(E) None of the above answers

Answer: C

(13) Assume that we are using MinHash for detecting duplicate documents as described in lecture. Which of the following statements are True?

S1. If two documents have the same MinHash signature, they must be exact duplicates.

S2. If two documents are exact duplicates, they must have the same MinHash signature.

S3. If we increase the length of each MinHash signature, we can reduce the number of candidate pairs to compare.

S4. In the LSH algorithm, candidate pairs are identified as those that hash to the same bucket for at least one band.

(A) S1, S2, and S3

(B) S1, S2, and S4

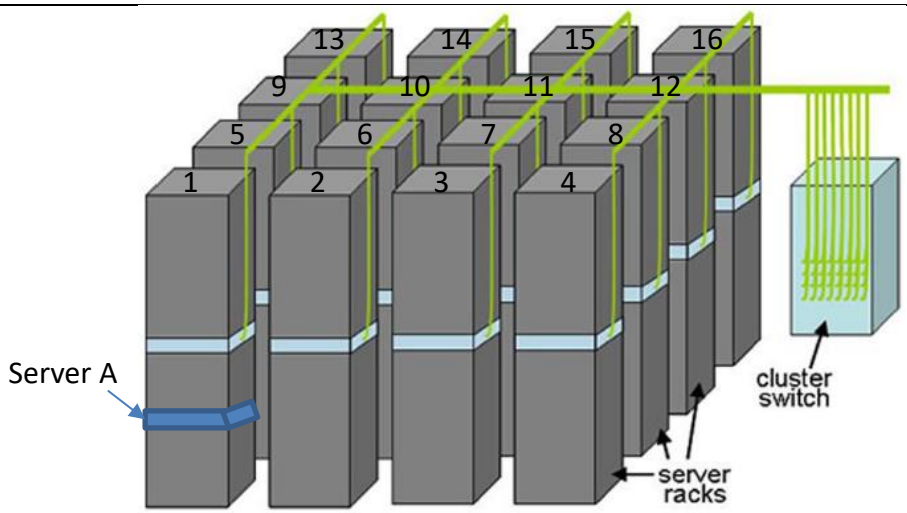
(C) S1, S3, and S4

(D) S2, S3 and S4

(E) None of the above answers

Answer: D

(14) The architecture of a commercial data center is illustrated in the below figure. The number on the top of each rack is the identifier of each rack. Users can run Hadoop or Spark jobs in the data center.



Suppose a program P is running on Server A in Rack 1. Denote the latency of P accessing a byte in the hard disk of Server A, a byte in the hard disk of the other servers in Rack 1, and a byte in the hard disk of the other server in Rack 16 is L1, L2 and L3, respectively. Which of the following statements are True?

S1. L3 can be ten times larger than L2.

S2. L2 can be ten times larger than L1.

S3. L3 is roughly the same as L2.

S4. L2 is roughly the same as L1.

(A) S1 and S2

(B) S3 and S4

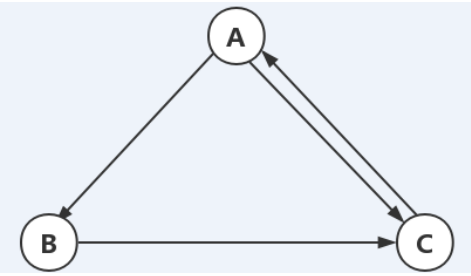
(C) S1 and S3

(D) S2 and S4

(E) None of the above answers

Answer: B

(15) By PageRank algorithm, what is the converged rank value of A, B, C, given their linkage in the following figure? PR(x) means the page rank value of x.



(A) PR(A) > PR(B) > PR(C)

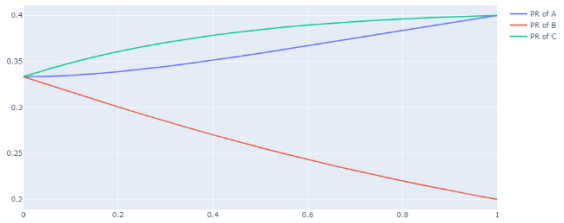
(B) PR(C) > PR(A) > PR(B)

(C) PR(A) > PR(C) > PR(B)

(D) PR(C) > PR(B) > PR(A)

(E) None of the above answers

Answer: B or E, depending on alpha value.



(16) Cluster the following four points (with (x, y) representing locations) into two clusters: P1(1, 1), P2(2, 2), P3(3, 2), P4(4, 3). Initial cluster centers are P1 and P2. The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as D(a, b) = |x2 – x1| + |y2 – y1|, i.e., Manhattan distance. Define cluster center as the mean of cluster members. Use K-Means Algorithm to find the two cluster centers after one iteration.

(A) {(3/2, 3/2), (7/2, 7/2)}

(B) {(2, 2), (3, 2)}

(C) {(1, 1), (3, 7/3)}

(D) {(2, 5/3), (4, 3)}

(E) None of the above answers

Answer: C, https://colab.research.google.com/drive/1hdPjFzzJP-LnHQLCLrci2DEB91yGjTgR?usp=sharing

(17) Compute the Jaccard similarity of the sets of 2-shingles for the following two strings: abcaba and abdbaba

(A) 1/2

(B) 1/3

(C) 1/4

(D) 1/5

(E) None of the above answers

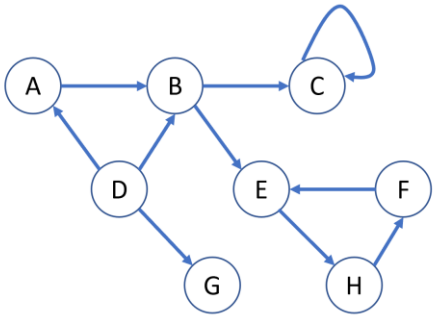
Answer: B

2-shingles for abcaba: {ab, bc, ca, ba}

2-shingles for abdbaba: {ab, bd, db, ba}

Jaccard sim. = 2/6=1/3.

(18) Given the following graph, denote X to be the number of dead ends are there in the graph, and Y to be the number of spider traps are there in the graph.



(A) X = 0, Y = 1

(B) X = 0, Y = 2

(C) X = 1, Y = 1

(D) X = 1, Y = 2

(E) None of the above answers

Answer: E. Dead ends: {G}, Spider traps: {C}, {E, F, H}, {B, C, E, F, H}, {A, B, C, E, F, H}

(19) According to the lecture of MinHashing, here is a matrix representing the signatures of seven columns, C1 through C7.

	C1	C2	C3	C4	C5	C6	C7
R1	1	2	1	1	2	1	4
R2	2	3	4	3	3	2	2
R3	3	1	2	3	1	3	2
R4	4	1	3	1	1	4	4

R5	5	2	5	1	1	5	1
R6	6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each (i.e., R1-R2, R3-R4, and R5-R6). Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Identify how many of the following pairs are candidate pairs.

S1. C1 and C3

S2. C1 and C6

S3. C2 and C5

S4. C2 and C6

S5. C3 and C4

S6. C3 and C7

(A) 1

(B) 2

(C) 3

(D) 4

(E) None of the above answers

Answer: C

S1, yes

S2, yes

S3, yes,

S4, no

S5, no,

S6, no.

QUESTION 3: Spark [6 marks]

On a certain large Spark cluster, Rose creates a data frame named **traceA**, and writes

```
Line 0: maxSql = spark.sql("""
Line 1: SELECT ip, sum(time) as access_total
Line 2: FROM traceA
Line 3: WHERE time>0.1
Line 4: GROUP BY ip
Line 5: ORDER BY sum(time) DESC
Line 6: """)
Line 7: maxSql.collect()
```

the shown program to process the trace. The **traceA** data frame keeps the logs, and each log entry represents the log information of one web page access, including various fields: **ip**, the IP address of the log entry and **time**, the amount of time of that access.

(A) For the operations from Line 1 to Line 5, what are the potential performance bottlenecks? Please identify which lines cause the bottleneck and justify your answer.

[2 marks]

Answer:
Depending on whether the data frame has been in the RAM, Line 3, the potential I/O cost by reading the trace
Depending on the size of data generated from Line 3, Line 4/5 can also be the bottleneck, the network I/O of data shuffling.

(B) Rose later runs the same program with the same hardware and software settings on another trace data frame TraceB. TraceB has the same schema as well as the same number of log entries as the one in the previous question. However, she finds that the operations in Line 5 run much more slowly. Give two potential reasons for this significant slow-down.

[2 marks]

Answer: there are many possible answers, and examples are: a) data skew causing task straggler, b) filter condition generates more results, etc

(C) Rose tries to run the same program on ten different input data sets (i.e., changing different data frames as inputs). Those data sets have the same schema as **traceA** but have different number of tuples. She finds that this program can successfully run in some cases, but fail in other cases. Please identify which line of the program may cause the issue, explain the reason and suggest an amendment to the program to fix this problem.

[2 marks]

Answer:

Line 7, may return too many results.
We can add LIMIT to the SQL query to limit the number of results.

QUESTION 4: Crazy Histograms [6 marks]

Jim wants to obtain the histogram of data sets in the following problems. For each problem, **your answer needs to have the following three parts**: (1) identify big data system platform(s) among those your learnt in this module and justify your choice, (2) give the **pseudo code** for solving the problem, and (3) analyze the potential performance bottleneck of your program (according to "Performance Guidelines for Basic Algorithmic Design", in lecture slide 2-mapreduce-basic.pptx). For example, if you suggest to use Hadoop/MapReduce, the pseudo code should describe the Map and Reduce functions. Your solution should be as efficient as possible.

(A) Given a very large collection of random strings, and Jim wants to obtain a histogram: the number of distinct strings in each string length (i.e., the number of distinct strings with the length of x, x=1, 2, 3, 4, …).

[3 marks]

Answer:

System: Mapreduce or spark. Batch processing.

API:

Map: Length as key, string as value.

Reduce: compare the strings in the value list for distinction and output the <length, count>

Performance:

Network and disk I/O in MapReduce

Data skews will cause memory consumption issue and parallelism issue.

(B) We have a database of research articles, and each article has a set of references (citing other articles in the database). Suppose we have extracted all the citation information in the form of tuples <*src*, *dst*>, where *src* and *dst* are the article identifiers (IDs), meaning the article *src* cites the article *dst* as its reference. All these tuples are stored in an HDFS file. Given the HDFS file and an input article ID *id0*, your task is to calculate the histogram based on the similarity between *id0* and other articles in the database. You need to define how to calculate the similarity score and normalize it to [0,1]. The histogram records the number of articles for each similarity level with an interval of 0.1, that is, [0,0.1], (0.1,0.2], (0.2,0.3]… (0.9,1.0].

[3 marks]

Answer:

Stage 1: calculate the similarity score:

System: Graph computation systems.

API:

Use personalized page rank.

Output the list of similarity score values.

Bottleneck:

Network and disk I/O in pagerank

Data skews

Stage 2: compute the histogram

System: Spark

API: compute a range partitioning

Bottleneck:

Network and I/O cost

Parallelism is limited by the 10 ranges of the histogram.

QUESTION 5: Imagine Future Big Data Systems [Bonus: 4 marks]

New hardware technologies have been developed recent years for future data centers. See the below scenario on the major advancement on network and CPU. For each of the two scenarios, we want to redesign Spark to run on such future data centers. Your task is to suggest two potential design performance considerations for adjusting the original Spark system to extend its current design to future data centers, and justify your answers.

(A) Scenario I: Suppose future data centers have very high-speed data center network with much higher bandwidth and lower latency.

[Bonus: 2 marks]

Answer:

Examples include:

Less aggressive IMC (In-mapper combiner).

More aggressive "move data to compute"

Cost plan adjustment in data partitioning.

Lower compression in data frame.

(B) Scenario II: Suppose future data centers have advanced CPUs with much more cores and higher clock frequency.

[Bonus: 2 marks]

Answer:

Examples include:

More aggressive compression.

More parallelism in a single node.

Query optimization by the more parallelism within a node.

BLANK PAGE
END OF ASSESSMENT