NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING

Semester 2 AY2021/2022

CS5425/CS4225 – Big Data Systems for Data Science

Mid-Term Test (Sample Answer)

March 2023 Time Allowed: 1 Hour and 15 Min

INSTRUCTIONS TO STUDENTS:

1. This assessment paper contains **TWO (2)** questions and comprises **TEN (10)** printed pages, including this page.

2. This is an **OPEN BOOK** test. You can any hard copy materials such as books and notes. **No electronic devices are allowed.**

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. **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: ______________________________

Venue: ______________________________

Seat Number: ______________________________

QUESTION 1: True or False. Each question weighs 1 mark. [10 marks]

(1) Sensor readings have different formats including video and audio etc. Such scenario shows Veracity of big data.

(A) True
(B) False

Answer: B. Veracity: uncertainty of data.

(2) In the Broadcast Join algorithm discussed in class, reducers are necessary in order to process the tuples emitted by the map functions.

(A) True
(B) False

Answer: B. No reducers are needed in the broadcast join.

(3) When a user requests to download a file from HDFS, the file will first be downloaded from the data node to the name node, and then sent to the user.

(A) True
(B) False

Answer: B False. When a user requests to download a file from HDFS, only the meta data of the file will be issued to the name node. The data will directly be downloaded from the data nodes.

(4) Current big data system designs mainly use scale-out architectures, rather than scale-up architectures.

(A) True
(B) False

Answer: A. Scale-out architectures are more cost efficient.

(5) Consider the task of near-duplicate document detection. Compute whether the following documents are candidate pairs, according to the MinHash algorithm. Use a shingle size of k=1 (word), and the following hash function h, defined as: h(`green`) = 2, h(`eggs`) = 3, h(`ham`) = 4.

Document 1: `green eggs`
Document 2: `green ham`

Is the following statement True or False?
S1. The min-hash signature of Document 1 is 2.

(A) True

(B) False

Answer: A.

(6) Following Question (5), is the following statement True or False?
S2. Document 1 and Document 2 are candidate pairs.

(A) True

(B) False

Answer: A.

Document 1: signature = min(2, 3) = 2
Document 2: signature = min(2, 4) = 2
Both documents have the same signature, so they are candidate pairs.

(7) HDFS has three replicas for each chunk. Thus, the system offers more flexibility of moving a task to the machine where a replica is stored.

(A) True

(B) False

Answer: A. Move data to processing.

(8) In GFS/HDFS, the chunk size is set to 64MB by default. What if we set this chunk size to be much larger (say, 1GB)? Is the following statement True or False?
S1. If the chunk size increases, the task parallelism of MapReduce will also improve.

(A) True

(B) False

Answer: B. For the same input size, the max numbers of map and reduce tasks reduce.

(9) The key function of using Combiner in Hadoop is to better spread out the load among different Reduce tasks.

(A) True

(B) False

Answer: B. That is the function of partitioner.

(10) The assignment of workers (i.e., machines) to map and reduce tasks in MapReduce is run within the Master node.

(A) True

(B) False

Answer: A. Task scheduling is done in the master node.

QUESTION 2: Choose the most appropriate option from the available options. Each question weighs 1.5 marks. [15 marks]

(11) Jim has a job 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 causes?

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

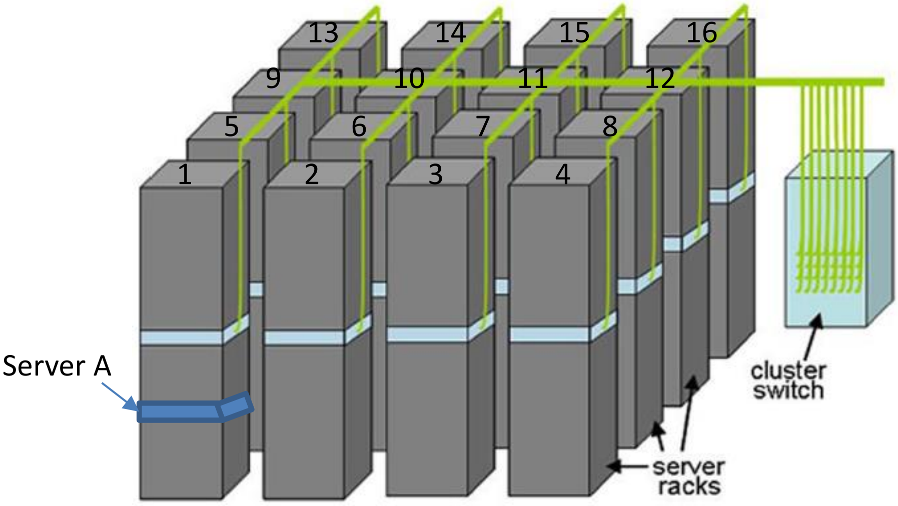
(E) None of the above answers

Answer: C.

(12) 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 jobs in the data center. The data read latency and bandwidth of different hardware components are given in Table 1.

Table 1. 1 ns= 1e-9 second, 1ms= 0.001 second, 1us = 1e-6 second.

	Latency	Bandwidth
Main memory (DRAM)	100ns	40GB/sec
Hard disk	10ms	200MB/sec
Rack switch (per port)	300us	200MB/sec
Cluster switch (per port)	500us	10MB/sec



Suppose a program P is running on Server A in Rack 1. Denote the time of P reading a chunk of 128MB from a file in the hard disk of Server A, reading a chunk of 128MB from a file in the hard disk of the other server in Rack 1, and reading a chunk of 128MB from a file in the hard disk of the other server in Rack 16 as L1, L2 and L3, respectively. Which of the following statements are True?

S1. L3 can be over ten times larger than L2 (L3>10 * L2).

S2. L2 can be over ten times larger than L1 (L2>10 * L1).

S3. L3 is within four times as L2 (L3< 4*L2).

S4. L2 is within four times as L1 (L2 <4*L1).

(A) Only S1 and S2 are True

(B) Only S3 and S4 are True

(C) Only S2 and S3 are True

(D) Only S1 and S4 are True

(E) None of the above answers

Answer: D. L3 is bounded by 10MB/sec, L2 is bounded by 200MB/sec, and L1 is bounded by 200MB/sec.

(13) Following Question (12). Consider Hadoop Distributed File System (HDFS) deployed in that cluster. Each data chunk is stored with three copies. Consider a data chunk with one copy stored in Sever A. Which servers are the possible locations of its other two copies stored in the data center?

(A) Server A and another server in Rack 2

(B) Two different servers in Rack 1 (except Server A)

(C) One server in Rack 2 for both copies

(D) Two different servers in Rack 2

(E) None of the above answers

Answer: D. Two replicas in the same rack, and one in the other rack.

(14) New main memory technologies have been developed recent years. One of the major trends is non-volatile main memory (NVMM). Compared with DRAM, NVMM provides latency in the range of that of DRAM and has high density and lower price in costs per GB. For example, Intel Optane DC Persistent Memory Modules (DCPMM) uses the 3D XPoint technology which scales to large capacity (up to 512GB per DIMM, Dual In-Line Memory Module) and offers read latency almost similar to DRAM. Despite various advantages in density, NVMM exhibits about 6-30 times higher write latency than DRAM.

NVMM will be integrated into next-generation data centers, and we need to redesign big data systems. Each **future server** will have a hybrid memory system consisting of 128-256 GB of DRAM as well as 1-2 TB (Tera Byte) of NVMM. This is a high increase from the current server design with only 64-128GB of DRAM per machine.

The following are the potential design performance considerations for adjusting the original Hadoop system to extend its current design to future data centers? Which of them are True?

S1. Allow the usage of IMC (In-mapper combiner) with larger memory working set in each mapper.

S2. Storing data chunks of HDFS (especially from the frequently accessed files) to NVMM can improve the read/write performance of 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.

(15) Consider the data center architecture in Question (12), with the data read latency and bandwidth of its hardware components shown in Table 1. Each server has a hybrid memory system consisting of 128-256 GB of DRAM as well as 1-2 TB (Tera Byte) of DCPMM. Suppose a program P is running on Server A in Rack 1. Denote the latency of P reading a byte in the DCPMM of Server A, reading a byte in the DCPMM of the other servers in Rack 1, and reading a byte in the DCPMM of the other server in Rack 16 as L1, L2 and L3, respectively. Which of the following statements are True?

S1. L3 can be over ten times larger than L2 (L3>10 * L2).

S2. L2 can be over ten times larger than L1 (L2>10 * L1).

S3. L3 is within ten times as L2 (L3< 10*L2).

S4. L2 is within ten times as L1 (L2 <10*L1).

(A) Only S1 and S2 are True

(B) Only S2 and S3 are True

(C) Only S1 and S3 are True

(D) Only S2 and S4 are True.

(E) None of the above answers

Answer: B. L1 is very short. S2 is True. L3 and L2 are comparable. S3 is True.

(16) Jim uses Hadoop to implement the following SQL query on a table named *students*. The *students* table consists of fields including *ID* (for student ID), *name* (for student name), *district* (for which district the student is from), *score* (for the score of the student), and other information such as *age* and *gender*. The students are from 100 districts, with ages between 18 to 30. The query computes the highest score in each district for the male student whose age is under 23.

SELECT *district*, MAX(*score*)

FROM *students*

WHERE *age*<23 and *gender*='M'

GROUP BY *district*

Jim implements the query with the following three Hadoop jobs.

Job 1:

Mapper: the mapper function performs the filter (*age*<23) on the HDFS file for *students*. Emit the records that satisfy the filter (in the form of key and value to be student ID and the record, respectively).

Reducer: null.

Job 2:

Mapper: the mapper function performs the filter (*gender*='M') on the output file of Job 1. Emit the records that satisfy the filter (in the form of key and value to be student ID and the record, respectively).

Reducer: null.

Job 3:

Mapper: For each key/value pair from the output of Job 2, emit a key/value pair with the key and value to be district and score, respectively.

Reducer: Compute the max for the values in the value list of each key, and emit the key/value pair with the key and value to be district and the max, respectively.

Jim runs the Hadoop jobs on a cluster with 10 machines. HDFS is running on that cluster with the default chunk size of 128MB. The *students* table is around 5GB size.

Which of the following configurations for the number of Map tasks for Job 1 is likely to achieve good task parallelism while without any idle Map tasks?

(A) 1

(B) 5

(C) 10

(D) 50

(E) 250

Answer: C. There are 10 machines, and 40 chunks in the HDFS file (5GB/128MB). More than 40 -> idle Map tasks.

(17) Following Question (16). Suppose Jim runs Job 3 with 5 mappers and 6 reducers. What are the possible smallest and largest values for the number of distinct copy operations will there be in the sort/shuffle phase (assuming there are no task failures or speculative task executions)? We denote the two values as min and max accordingly.

(A) min = 5, max = 6

(B) min = 0, max = 30

(C) min = 6, max = 30

(D) min = 0, max = 6

(E) None of the above answers

Answer: B. Min: 0 (there is no output from mappers), Max: 30

(18) Following Question (16). Perform I/O analysis as we learnt in the lecture to Job 1. Which of the following statements are True?

S1. The amount of input disk I/O for each Map task is 128MB.

S2. The amount of output disk I/O for each Map task is very small (almost zero bytes).

S3. As there is no reducer is needed, the Shuffling stage is not needed.

(A) All statements are False

(B) Only one statement is True

(C) S1 and S2 are True

(D) S1 and S3 are True

(E) S2 and S3 are True

Answer: D

(19) Following Question (16). Jim finds that the performance of his implementation is very bad. He decided to combine all the three jobs into a single job (see below). He finds that the performance becomes much better.

Job 4:

Mapper: The mapper function performs the filter (age<23 and gender='M') on the HDFS file for students, and emits a key/value pair with the key and value to be district and score of the tuple satisfying the filter, respectively.

Reducer: Compute the max for the values in the value list of each key, and emit the key/value pair with the key and value to be district and the max, respectively.

Which of the following statements are True?

S1. By combining Job 1, Job 2 and Job3 to Job4, the amount of input disk I/O of Job 4 is smaller than the total amount of input disk I/O for Job 1, Job 2 and Job 3.

S2. According to scalability analysis on the Map task, Job 4 has better task parallelism than other jobs (Job 1, Job 2 and Job 3).

S3. According to network I/O analysis in the shuffling, the Shuffling stage of Job 4 has a smaller amount of network I/O than that of Job 3.

(A) All statements are False

(B) Only one statement is True

(C) S1 and S2 are True

(D) S1 and S3 are True

(E) None of the above answers

Answer: B. Only S1 is True.

(20) 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) {(1, 1), (3, 7/3)}

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

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

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

(E) None of the above answers

Answer: A. Two clusters: {P1}, {P2, P3, P4}

BLANK PAGE
END OF ASSESSMENT