NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING
ASSESSMENT FOR
Semester 1 AY2022/2023

CS5425/CS4225 – Big Data Systems for Data Science
Final Test Paper – Sample Solutions
13 Apr 2023 **Time Allowed: 1 Hour 30 Mins**

INSTRUCTIONS TO STUDENTS:

1. This assessment paper contains **SIX(6)** questions and comprises **FOURTEEN (14)** printed pages, including this page.
2. Students are required to answer ALL the questions.
3. Write your answers within the space provided. Answers written on other parts of the answer script will not be graded unless you specify explicitly.
4. This is an **OPEN BOOK** examination, allowing any physical materials but no electronic devices allowed.
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.**
6. Please circle your class and write your matriculation number below.

Class: CS5425 CS4225

Matriculation Number: ______________________________

This portion is for examiner's use only

Question	Marks	Remarks
Q1 [4 marks]		
Q2 [6 marks]		
Q3 [3 marks]		
Q4 [5 marks]		
Q5 [4 marks]		
Q6 [3 marks]		
Total [25 marks]		

QUESTION 1: Indicate whether the following statements are true or false. Justify your answers. Each question weighs one mark. [4 marks]

(A) NoSQL databases often use denormalized views, also known as duplication, to improve query performance and overcome the limitations of not supporting joins.

Your Answer (True / False): True (0.5m)

Justification:
NoSQL databases often use denormalization to duplicate data across multiple tables or documents as storage is cheap. This makes it easier to query the data, as all the necessary information is available in a single table / document. (0.5m)

(B) Spark creates a Directed Acyclic Graph (DAG) to record the transformations into a few stages. We should try to avoid transformation across stages to improve performance.

Your Answer (True / False): True (0.5m)

Justification: Across stage transformation (with wide dependencies) needs data to be shuffled across different servers (i.e. network shuffling). This is a very expensive operation and should be avoided. (0.5m)

(C) We are using join() function from Spark DataFrame API to join two tables. We understand Spark DataFrame API supports multiple languages but using Java/Scala will achieve much better performance as Catalyst Optimizer is written in Scala.

Your Answer (True / False): False (0.5m)

Justification: In this case (i.e. join function), all the languages will achieve similar performance through DataFrame API. API / Query language only provides a logical plan, and Catalyst Optimizer will generate the optimized logical plan, physical plan and RDD codes, which is the same no matter which API language you are using. (0.5m)

(D) According to Pregel model, in each superstep, each worker loops through the vertices assigned to it and executes compute() function on each vertex. The compute() function can update the state of the vertex and send messages to the vertices on the same worker only.

Your Answer (True / False): False (0.5m)

Justification: messages from each vertex are sent to all the neighboring vertices which can be on the same worker or on a different worker. (0.5m)

QUESTION 2: NoSQL [6 marks]

(A) Mike is designing a NoSQL document store for a company to store the sensor data from IoT devices. Each document contains fields such as device ID, timestamp, and sensor readings. His colleague suggests choosing device ID as partition key to partition the data. Please state and explain:

1) the possible advantage of this choice
2) the possible disadvantage of this choice

[2 marks]

Your Answer:

Advantage: allow more efficient query and analysis on the data generated from the same sensor (i.e. the same device ID)
[1 mark]

Disadvantage: data may be skewed, e.g. if certain device generates too many data, one partition will contain too many records, leading to poor parallelization
[1 mark]

(B) Mike decided to use timestamp as the partition key to partition the data. But he cannot decide whether to use range partitioning or hash partitioning. Please help Mike to make decision by providing the advantages and disadvantages for both choices (i.e. range partitioning and hash partitioning).

[4 marks]

Your Answer:

Range Partitioning

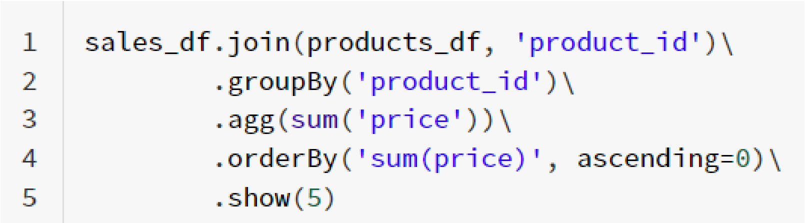
- Advantage: Allow more efficient query and analysis on data within a specific time range [1 mark]
- Disadvantage: Within a specific time period, all the data are written to a specific partition, leading to the bottleneck issue at this specific partition / server. [1 mark]

Hash Partitioning

- Advantage: spread out the data evenly among different partitions / servers [1 mark]
- Disadvantage: a query and analysis on data within a specific time range may need to be sent to all the partitions / servers [1 mark]

QUESTION 3: Spark [3 marks]

(A) Jim is working for an online retail company. On a large spark cluster, Jim creates a dataframe named **sales_df** to record last month's sales transactions (including customer_id, order_date, and product_id) and a dataframe named **products_df** to record all the products information (including product_id, product_name, and price). Jim wants to run the following Spark program to generate some results for analysis.



```
sales_df.join(products_df, 'product_id')\
        .groupBy('product_id')\
        .agg(sum('price'))\
        .orderBy('sum(price)', ascending=0)\
        .show(5)
```

Explain the results Jim will get by running the above codes. (Note: not the specific records but a description of the records)

[1 mark]

Your Answer:

The top five products sold in last month, in terms of sales revenue (i.e. price * quantity).

(B) Identify the potential performance bottlenecks in the above codes, indicate at which line(s) and provide your explanations accordingly.

[2 marks]

Your Answer:

At line 1, join two tables (sales_df and products_df). If both tables are super big (i.e. cannot fit into the memory) and are not already been partitioned using product_id, it will take a while to partition two super big tables across different machines / servers (i.e. network shuffling) and then do sort merge at each partition.

At line 3, if the grouped data after line 2 are highly skewed (e.g. many sales records for a certain product_id), line 3 will have task straggler issue, i.e. certain task takes much longer time than the other tasks to complete.

At line 4, depending on the number of unique product_id, the global sorting (a wide transformation requiring network shuffling) done by orderBy can also be a bottleneck.

(Get full marks, if correct on any two of the above points. But for each point you need to explain clearly why there is potential bottleneck).

QUESTION 4: Spark Streaming [5 marks]

(A) You are working for an online advertisement company and your CEO would like to track the number of user clicks for each advertisement in a certain period. You decide to use Spark Streaming to implement this.

First, you collect the user clicks data into a data stream called, UserClicks. This data stream contains a stream of records and each record has several fields, e.g. userID, adID (i.e. advertisement ID), clickTime(i.e. the time a user clicks this advertisement), and etc.

Then, you clarify with your CEO on what he wants exactly.

- He wants to have a real time dashboard showing the total user click counts for each advertisement in a period of one hour and the counts should be updated once every 30 minutes. In other words, the click counts should be in a period of 6pm-7pm, 630pm-730pm, 7pm-8pm and etc.
- He understands that some click records may be delayed before received or processed. But it is unlikely the click can be delayed by more than 20 minutes, so he is OK that we do not count a click in if it is delayed by more than 20 minutes.

Please fill up the blanks in the below Spark Streaming codes to properly process the input data stream (UserClicks) and meet your CEO's requirements.

[3 marks]

Your Answer:

```
UserClicks = (spark
.readStream.format("socket")
.option("host", "localhost")
.option("port", 9999)
.load())


   counts =
   (UserClicks
   .withWatermark("clickTime", "20 minutes")     ← Line 1: 1 mark
   .groupBy("adID", window("clickTime", "60 minutes", “ 30 minutes”))
   .count())                              ↑
           ↑                     Line 2: 1.5 marks (minus 0.5 mark for each error)
     Line 3: 0.5 mark


writer = counts.writeStream.format("console").outputMode("update")
```

(B) Assume there is an update / trigger happened at 8pm and the next update / trigger is at 830pm. Below table is showing a few latest records we received as of 8pm.

userID	adID	clickTime
1	3	8pm
2	2	8pm
3	5	8pm
3	3	7:59pm
4	4	7:59pm
1	6	7:59pm
3	7	7:58pm
6	4	7:58pm

Based on all the information you have:

1) Calculate the watermark for the period 8pm to 8:30pm
2) With this watermark, at the next trigger (8:30pm), indicate whether the below records will be updated or not.
    a. counts under period 6pm – 7pm
    b. counts under period 6:30pm – 7:30pm
    c. counts under period 7pm – ~~7:30pm~~ **8pm**
    d. counts under period 7:30pm – 8:30pm

[2 marks]

Your Answer:

Watermark = max clickTime – watermark delay = 8pm – 20 mins = 7:40pm (1 mark)

With this watermark (7:40pm), at the next trigger (8:30pm):

- counts under period 6pm – 7pm: not updated (0.25m)
- counts under period 6:30pm – 7:30pm: not updated (0.25m)
- counts under period 7pm – 8pm: updated (0.25m)
- counts under period 7:30pm – 8:30pm: updated (0.25m)

QUESTION 5: Graphs [4 marks]

(A) Rachel is analysing a simple social network with five users: A, B, C, D and E. The relationships between these users are listed per below:

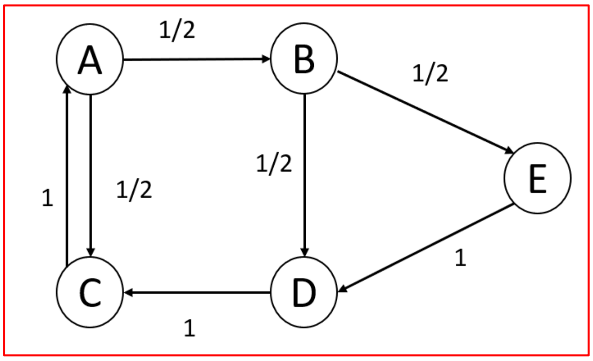
- A follows B and C
- B follows D and E
- C follows A only
- D follows C only
- E follows D only

Rachel wants to use PageRank algorithm to calculate the importance factor of each user, i.e. if a user has more followers and more important users are following him/her, then the user is more important.

Suppose we use PageRank algorithm with teleport (β=0.85, i.e. teleport probability = 1- β) to solve this problem. Write the equations to calculate the importance factors for users A, B, C, D, and E, denoted as I(A), I(B), I(C), I(D) and I(E) respectively.

[2 marks]

Your Answer:



(the graph drawing is not required in the answer)

I(A) = 0.85*I(C) + 0.15/5
I(B) = 0.85*(I(A)/2) + 0.15/5
I(C) = 0.85*(I(A)/2 + I(D)) + 0.15/5
I(D) = 0.85*(I(B)/2 + I(E)) + 0.15/5
I(E) = 0.85*(I(B)/2) + 0.15/5

(Get 2 marks if all five equations are correct, minus 0.5 mark for each incorrect equation)
(It is also OK to use Matrix representation, get full marks as long as the formulas / representations are correct, see details in the next page)

$$M = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 0 & 0 \end{bmatrix}, N = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

$$A = \beta M + (1-\beta)N = 0.85\, M + 0.15N = \begin{bmatrix} 0.03 & 0.03 & 0.88 & 0.03 & 0.03 \\ 0.455 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.455 & 0.03 & 0.03 & 0.88 & 0.03 \\ 0.03 & 0.455 & 0.03 & 0.03 & 0.88 \\ 0.03 & 0.455 & 0.03 & 0.03 & 0.03 \end{bmatrix}$$

I(A) = 0.03 I(A) + 0.03 I(B) + 0.88 I(C) +0.03 I(D) +0.03 I(E)
I(B) = 0.455 I(A) + 0.03 I(B) + 0.03 I(C) +0.03 I(D) +0.03 I(E)
I(C) = 0.455 I(A) + 0.03 I(B) + 0.03 I(C) +0.88 I(D) +0.03 I(E)
I(D) = 0.03 I(A) + 0.455 I(B) + 0.03 I(C) +0.03 I(D) +0.88 I(E)
I(E) = 0.03 I(A) + 0.455 I(B) + 0.03 I(C) +0.03 I(D) +0.03 I(E)

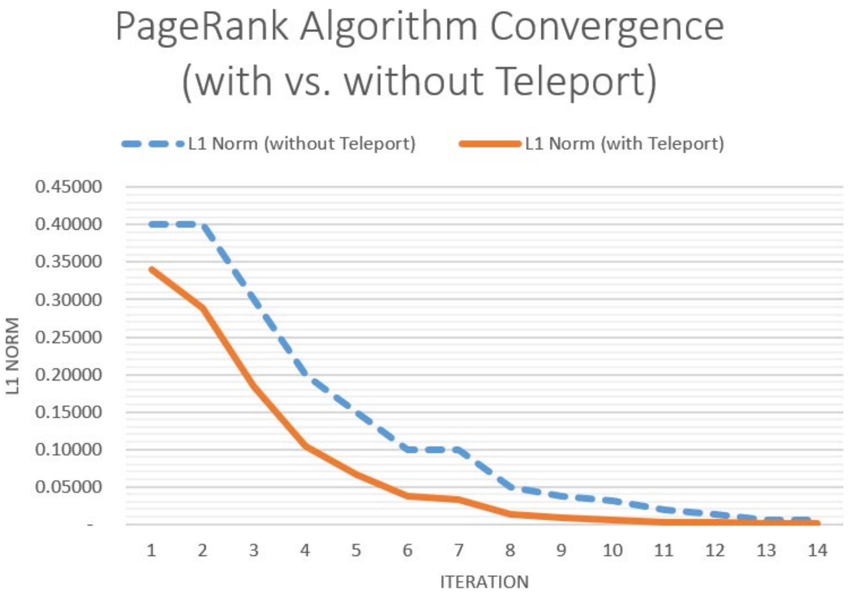
(B) Rachel tried below two algorithms to compute the importance factors:

1) PageRank with Teleport
2) PageRank without Teleport

She initialized all the important factors as 0.2 (= 1/5) and then monitors the algorithms convergence performance using L1 Norm, i.e. the sum of the absolute values of the importance factor change for each user, denoted as below equation:

$$\text{L1 Norm} = |I(A)^{(t+1)} – I(A)^{(t)}| + |I(B)^{(t+1)} – I(B)^{(t)}| +|I(C)^{(t+1)} – I(C)^{(t)}|+ |I(D)^{(t+1)} – I(D)^{(t)}|+ |I(E)^{(t+1)} – I(E)^{(t)}|$$

Rachel plots the L1 norm curves by iterations for both algorithms in the below figure. She noticed PageRank with Teleport converges faster than PageRank without Teleport. Please help Rachel to figure out the reasons behind this and provide your explanations.



[2 marks]

Your Answer:

Teleport is designed to solve the spider trap and dead ends problems in PageRank algorithm. For this specific problem, there is No spider trap and No dead ends. However, **Teleport still helps the random walker to occasionally jump to the less popular /important nodes so as to spread out the importance factors among all the nodes in the graph**. This helps the algorithm to converge faster.

(2 marks)

(Get 0 Mark if only focus on spider trap and dead ends
Get 2 Marks if clearly convey the idea that Teleport helps spread out the importance factors among all the nodes in the graph)

QUESTION 6: Big Data System [3 marks]

Suppose you work for a large agricultural commodity trading company that buys and sells crops such as wheat, corn, and soybeans around the world. Your company is looking to improve its decision-making by analysing a large amount of data, including weather patterns, commodity prices, transportation costs and etc. Therefore, your company is thinking to invest on big data systems to satisfy the below requirements:

- Store and query large volumes of unstructured data, such as weather data, satellite images, and social media feeds
- Analyse large volumes of collected data and predict the commodity prices, the market supply and demands and etc.
- Monitor crop prices and trigger automated trades when certain thresholds are met

Provide a solution of using one or multiple big data systems for this company based on what you have learned from this course.

[3 marks]

Your Answer:

A NoSQL data store (e.g. Mango DB) to store all the unstructured data; (1 mark)
Spark Batch processing to analyze and pre-process the collected data and Spark ML to build machine learning models to make predictions; (1 mark)
Spark Streaming to monitor the crop prices in real time and compare the price to a certain threshold so as to trigger automated trades. (1 mark)

(It is OK to use other systems, get full marks if the other systems are also reasonably satisfying the company requirements).

BLANK PAGE
END OF ASSESSMENT