Tutorial Quiz 2

● Graded

Student

Caresse Zhou

Total Points

8 / 10 pts

Question 1

Review MCQ Questions — 4 / 6 pts

1.1 **Q1a: CURRENT_TIME meaning** — 0 / 1 pt

- ✓ **+ 0 pts** Incorrect option selected

1.2 **Q1b: TF-IDF vs Word Embeddings** — 1 / 1 pt

- ✓ **+ 1 pt** Correct

1.3 **Q1c: Stateful Stream Operators** — 0 / 1 pt

- ✓ **+ 0 pts** Incorrect option selected

1.4 **Q1d: Data Partitioning** — 1 / 1 pt

- ✓ **+ 1 pt** Correct

1.5 **Q1e: Column Stores** — 1 / 1 pt

- ✓ **+ 1 pt** Correct

1.6 **Q1f: DataOps Lifecycle** — 1 / 1 pt

- ✓ **+ 1 pt** Correct

Question 2

CSV vs Parquet Files — 2 / 2 pts

- ✓ **+ 1 pt** Correct difference of storage models
- ✓ **+ 1 pt** Suitable use cases proposed

Question 3

Pipeline Design: Realtime Fraud Detection — 2 / 2 pts

- ✓ **+ 1 pt** Correct Pub/Sub approach chosen
- ✓ **+ 1 pt** Valid explanation given

THE UNIVERSITY OF SYDNEY

Tutorial Class	22
Student Name	Qianyou Zhou
Student Number	540930700

CONFIDENTIAL QUIZ PAPER

This paper is not to be removed from the quiz venue

COMP5339 Data Engineering

TUTORIAL QUIZ 2

Individual Quiz (5%)

QUIZ WRITING TIME: 30 minutes

INSTRUCTIONS TO STUDENTS:

Please make sure you fill in your tutorial and Student ID correctly.

This is a Closed Book test: electronic devices (including phones, laptops, tablets, etc.) are not allowed - Please place them in your bag on the floor.

This tutorial quiz consists of ***3 QUESTIONS*** with the first one consisting of 6 multiple-choice questions. All questions must be answered.

- Answer all questions within the spaces provided on this paper.
- Note that questions are of unequal value. The points for each question are shown at the start of the question.
  The total mark of this quiz paper is **10**.
- For short answer questions, take care to write legibly. Write your final answers in ink.
  Do not use a pencil or red ink.

Please hand the completed quiz paper to the tutor before you leave the room.

This paper is not to be removed from the quiz venue.

Question 1: Review Questions **[6 points]**

This question has six (6) parts, ((a)—(f)).Tick the box (or boxes) of each correct answer.

(a) **[1 point]** Which **one** of the following statements describes what the CURRENT_TIME keyword in SQL represents?

- □ The earliest valid time.
- □ The real-world event time.
- □ The database server's current system timestamp. ✓
- ☒ The database server's most recent start time.

(b) **[1 point]** Which **one** of the following statements best contrasts TF-IDF and word embeddings for ML?

- □ TF-IDF is dense and contextual; embeddings (Word2Vec/GloVe/BERT) are sparse counts.
- ☑ TF-IDF is sparse bag-of-words; embeddings (Word2Vec/GloVe/BERT) are dense vectors capturing semantics. ✓
- □ Both are sparse and ignore context.
- □ None of the above.

(c) **[1 point]** Which **one** of the following statements is true about stateful stream operators and how they differ from stateless stream processing?

- ☑ Stateful stream operators output results based on multiple events. ✓
- ☒ Stateless stream operators aggregate over individual stream windows.
- □ Stateful stream operators never retain more than one event.
- □ Stateful stream processing ignores event order.

(d) **[1 point]** Which **one** of the following statements is correct about the role of data partitioning?

- □ Data partitioning primarily improves fault tolerance by increasing data redundancy.
- ☑ Data partitioning improves performance primarily by reducing I/O contention and improving intra-query parallelism. ✓
- □ Data partitioning primarily simplifies schema-design.
- □ Data partitioning primarily improves performance by supporting nested data structures.

(e) **[1 point]** Which **one** of the following statements is correct about column stores?

- □ Column stores are particularly efficient for write-heavy transactional workloads.
- □ Column stores are particularly efficient for real-time OLTP systems.
- □ Column stores are particularly efficient for unstructured data storage.
- ☑ Column stores are particularly efficient for analytical queries that read a subset of columns. ✓

(f) **[1 point]** Which **one** of the following activities is NOT part of the DataOps Lifecycle?

- □ Planning
- □ Testing
- □ Development
- ☑ Procurement
- □ Deployment
- □ Monitoring

Question 2: CSV files versus Parquet files **[2 points]**

(a) **[2 points]** Briefly compare the CSV file and Parquet file formats. What is their primary storage model and give examples of their typical use cases.

CSV files are line-based storage model. The statistics in csv are generated by single lines, so it can be easily translated to JSON or other spot storaged files. It is used for data analysis based on samples. with structured data, like analyze the score of each student, or record the ifomation of the users.

Parquet is column-based storge model. It can be easily analyzed by columns. So the statistics needs to be structured, and prefer to analyze the whole than the single usage. For example, recording the different test results for a kind of car, or stream data.

Question 3: Pipeline Design: Realtime Fraud Detection **[2 points]**

Suppose that you are a data engineer who is designing a data pipeline for a financial services company to conduct fraud detection analysis on its financial transaction data. The goal is to provide real-time fraud detection, however you are given only limited server resources. Note also that the company has strict regulatory requirements for identifying any potential fraud.

(a) **[2 points]** You have the choice between two approaches: Using a Publish/Subscribe broker such as Apache Kafka or directly polling from operational databases with an ingestion script in your data pipeline.

Which approach would you recommend and why? Briefly explain your choice and why you consider it better suited than the alternative.

I will use a publish / Subscribe broker.

Because the server resources are limited, If I ~~use~~ directly polling with an ingestion script, the whole pipeline needs to use our server, it will be slow. Howerever, the goal requires a real-time system.

Secondly, for the fraud detection purpose. It is easier to apply the ~~companies~~ company's regulatory requirements to a real-time detection with the Publish / Subscribe broker. For we can set an alarm while scribing the transaction data, ~~and after~~ it is the most efficient way. And after that, only send the data without fraud to the following pipeline, those marked as fraud have been immedately send to the security department. If use a direct-polling pipeline, this will be slow.