THE UNIVERSITY OF SYDNEY

Room Number _____

Seat Number _____

Student Number

ANONYMOUSLY MARKED

(Please do not write your name on this exam paper)

CONFIDENTIAL EXAM PAPER

This paper is not to be removed from the exam venue

Computer Science

SAMPLE EXAMINATION

Semester 2 - Final, 2023

COMP5339 Data Engineering

EXAM WRITING TIME: 2 hours

READING TIME: 10 minutes

EXAM CONDITIONS:
This is a RESTRICTED OPEN book test - specified materials permitted

MATERIALS PERMITTED IN THE EXAM VENUE:
(No electronic aids are permitted e.g. laptops, phones)

One A4 sheet of handwritten and/or typed notes double-sided is permitted.

MATERIALS TO BE SUPPLIED TO STUDENTS:
None

INSTRUCTIONS TO STUDENTS:

- The total number of marks is 60
- Answer all questions in the spaces provided on this question paper.
- Return this question paper and any additional answer booklets or materials.
- Write your final answers in black or blue ink, not pencil.
- Take care to write clearly and legibly.

Please tick the box to confirm that your examination paper is complete. ☐

Question 1 (4 marks):

Why is the Map-Reduce programming paradigm important in data processing?

The Map-Reduce programming paradigm breaks the problem into smaller parts that are distributed across multiple servers and then each part is processed in parallel with the results merged to find the answer. This produces a significant speedup in processing time.

Question 2 (4 marks):

Describe what we mean by a "Data Lake".

A data lake is a centralized repository for storing all types of data, regardless of its structure or format. This can include structured data, semi-structured data, and unstructured data. Data lakes are often used to store raw data that has not yet been processed or analyzed.

Question 3 (6 marks):

Describe the HTML format for web pages. Show the high level structure of a web page. Illustrate with examples.

- HTML is a text based format
- it is based on XML
- consists of a series of directives that are words in angle brackets.
- sections start with a directive and end with the same directive prefixed by a slash
- levels
    - begins with <HTML>, ends with </HTML>
    - includes a head section <HEAD>...</HEAD> and a body section <BODY>...</BODY
- Example:

```
<HTML>
<HEAD>
<TITLE>This is the title!</TITLE>
</HEAD>
<BODY>
<H1>First level heading</H1>
Some text
</BODY>
</HTML>
```

Question 4 (10 marks):

You are the data engineer of a data science team in the NSW government whose task is to build a dashboard to visualise the average petrol prices in the Sydney region as it relates to average income of the area. You have access to a real-time government database of fuel prices that includes the postcode of each petrol retailer. You also have access to average annual income for each postcode from the Australian Bureau of Statistics. You can download petrol price data using a Web API, while the average income data is only accessible by downloading a CSV file.

(a)(4 marks) Which data architecture would you suggest to build for this scenario?

Data Sources: fuel/postcode data via web API, income data via CSV download

Data Storage: a relational database such as Postgresql,
petrolprices (date, stationid, postcode, price)
income (postcode, income_amount)

Data management: a streaming data platform such as Apache Kafka
programs: gathering, cleaning, ingestion, visualisation

Visualisation: a web based dashboard program or use a tool such as powerBI

(c)(3 marks) What data ingestion issues do you expect? How do you plan to deal with them?

Data cleaning: the data might include errors or missing data. A data validation program would be used to extrapolate, correct or omit problem data. Serious errors would be flagged on the visualisation dashboard.

(d) (3 marks) What data security and privacy issues do you see that need to be addressed in this use case?

Security issues include: unauthorised access by developers, system breaches.

Privacy issues include: unauthorised release of data, use of personally identified data, data collection without permission

END OF EXAMINATION