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?

Question 2 (4 marks):

Describe what we mean by a "Data Lake".

Question 3 (6 marks):

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

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?

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

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

END OF EXAMINATION