THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

This exam paper must not be removed from the venue

Venue _____

Seat Number _____

Student Number |__|__|__|__|__|__|__|__|

Family Name _____

First Name _____

School of Information Technology and Electrical Engineering

EXAMINATION

Semester One Final Examinations, 2016

INFS3200/7907 Advanced Database Systems

This paper is for St Lucia Campus students.

Examination Duration: 120 minutes

Reading Time: 10 minutes

Exam Conditions:

This is a Central Examination

This is a Closed Book Examination - no materials permitted

During reading time - write only on the rough paper provided

This examination paper will be released to the Library

Materials Permitted In The Exam Venue:

(No electronic aids are permitted e.g. laptops, phones)

Calculators - No calculators permitted

Materials To Be Supplied To Students:

Instructions To Students:

Additional exam materials (eg. answer booklets, rough paper) will be provided upon request.

All questions are to be answered.
Write the answers in the space provided in the examination paper.
There are total 60 marks.

For Examiner Use Only

Question	Mark
1	
2	
3	
4	
5	
6	
7	
8	
9	

Total _____

Question 1. (6 marks)

Entity resolution plays an important role in database integration.

(1) (1 mark) What is entity resolution?

(2) (1 mark) Please give an example that entity resolution can be used to save money or improve customer satisfaction when integrating databases.

(3) (3 marks) Edit distance and Jaccard coefficient are two common string similarity measures used in entity resolution. Please define these two measures.

(4) (1 mark) Please give an example that using string similarity alone cannot solve the problem of entity resolution.

Question 2. (6 marks)

Semi-join can be used in a distributed system to improve query processing efficiency.

(1) (2 marks) What is semi-join?

(2) (2 marks) If local processing costs are negligible compared with inter-site data exchange costs, when is a semi-join strategy beneficial?

(3) (2 marks) What types of meta-data does a DDBMS need to maintain in its global data catalog in order to decide if semi-join strategy should be used?

Question 3. (6 marks)

(1) (2 marks) A data warehouse can often make use of materialized views (e.g., using materialized data cubes). Discuss advantages and disadvantages of building materialized views in data warehouses.

(2) (2 marks) It is not common for data warehousing systems to support update operations. List two reasons why supporting updates in data warehouses is not a good idea. Briefly justify your answer.

(3) (2 marks) Compare and contrast the following two data warehousing operations: SLICE and DICE.

Question 4. (8 marks)

GFS has one master and many chunk servers.

(1) (1 mark) What information is stored on Master in GFS?

(2) (3 marks) The one-master-only design of GFS can potentially make the master to be a performance bottleneck as there can be many chunk severs and many client applications. How is GFS designed to address this potential problem?

(3) (3 marks) Another problem for this one-master-only design of GFS is that the master could be the single point of failure for the entire system. How can GFS recover when the master fails?

(4) (1 mark) A chunk server could also fail. How does GFS detect and deal with failed chunk servers?

(Additional writing space for Question 4.)

Question 5. (7 marks)

With low-cost memory and multi-core parallel processing, in-memory database systems become an attractive alternative to traditional data warehousing systems.

(1) (3 marks) Compared with traditional disk-based data warehousing systems, list two main advantages of in-memory databases, with brief explanation.

(2) (4 marks) Tables in in-memory database systems are often stored with column as the basic storage unit. Compared with row-based storage, the structure of column-based storage is more suitable for what types of applications? Why?

Question 6. (5 marks)

A webpage is uniquely identified by its URL. An inverted index is a table of (keyword, URL-list) pairs that links each word to all webpages that contain it.

(1) (4 marks) One team needs to use MapReduce to build the inverted index for all webpages which are already stored in GFS. Please provide properly commented pseudo code for `map()` and `reduce()` functions to perform this task. You can use any programming language or just plain English.

(2) (1 mark) There are many nodes to execute the `map()` function and also many nodes to execute `reduce()` function. What does MapReduce system do between the map phase and reduce phase?

(Additional writing space for Question 6.)

Question 7. (6 marks)

ACID (Atomicity, Consistency, Isolation, Durability) is a set of properties that guarantee that database transactions are processed reliably. In both DDBMS and HDFS, data can be distributed in different machines and there can be multiple data replicas.

(1) (2 marks) In DDBMS, how data consistency among replicas at different sites is guaranteed? Please use one concrete strategy to answer this question.

(2) (2 marks) If HDFS also adopts the same strategy as you discussed above to maintain data consistency, what benefits and problems can be expected?

(3) (2 marks) How does HDFS address the problem of consistency?

Question 8. (9 marks)

Key-value store, document store, and column-family store are three main types of NoSQL data models.

(1) (6 marks) Discuss their main characteristics, their advantages and disadvantages.

(2) (3 marks) Please identify one type of suitable applications for each of these three data models, and explain why.

(Additional writing space for Question 8.)

Question 9. (7 marks)

A graph model consists of nodes connected by edges, with labels and possibly weights.

(1) (2 marks) How graph data can be stored in a standard relational database? Please give a simple relational table schema to store graph data to explain your answer.

(2) (3 marks) Use one concrete type of queries to illustrate the problems of using relational model to store graph data.

(3) (2 marks) Please describe an alternative data model to store graph data such that processing of the type of queries discussed above can be more efficiently supported. Explain why.

(Additional writing space for Question 9.)

END OF EXAMINATION