University of Dundee
School of Science and Engineering
Examinations 2016

BSc Degrees in Computing

AC41011 Big Data Analysis

Time allowed: **TWO** hours

Instructions

There are **Five** questions in total.

Candidates must answer **FOUR** questions. All questions carry equal marks.

Diagrams should be drawn on the accompanying answer booklet

Approved calculators may be used in this examination.

Do not turn over this question paper until instructed to do so by the Senior Invigilator

Question 1

a) Contrast TWO noSQL database systems you have studied. Your answer should discuss both the data models and the physical configuration of their servers. Your answer should highlight any limitations the technologies exhibit and any particular strengths they have.

[10 marks]

b) A company has a data centre in New York running a large RDMBS appliance along with application servers. The company mirrors the data to centres in Los Angeles and Tokyo with application gateways in those centres as well . What problems does this produce? Suggest an architecture solution that addresses this problem.

[9 marks]

c) Given the following scenarios, suggest which of the three V's of big data apply to each and why :

Data from a room full of Mass Spectrometers is being collected and stored for many years to comply with EU data storage rules. Each file from the mass spectrometer is 10 Gigabytes or more and at least 100 files are generated a day.

A scientific experiment consists of 10,000 sensors with each sensor sampling it's reading at 100 times a second. The sensor has a resolution of 24 bits.

Player data is collected by a game company, the programmers have a system that sends events back to their server. The system is flexible and can be changed to add new data fields and new events and can easily send data of different lengths.

[6 marks]

Question 2

a) A proposed e-commerce platform consists of the following subsystems:

- Shopping cart and session data.
- Completed order records.
- Inventory and price data.
- Customer interaction data.

How would these systems be persisted if using Polyglot persistence ? Explain the reasons for your choices.

[8 marks]

b) Show (using the help of a diagram) the principal components of the Lambda architecture for dealing with big data. Suggest technologies for each of the components.

[10 marks]

c) How can immutable data be modelled with timestamps for use in Lambda architecture? Illustrate your answer with some simple examples.

[7 marks]

Question 3

a) Explain the roles of Mapper, Partitioner, Shuffle/Sort and reducer in map reduce.

[10 marks]

b) Briefly describe the role of the following technologies in the Hadoop architecture

HIVE
PIG
HCatalog
Mahout

[8 marks]

c) The Cassandra data model is accessed through the Cassandra Query language (CQL). Describe the major differences between this and SQL in relation to syntax, indexing and query restrictions.

[7 marks]

Question 4

a) What advantages does Spark bring over Hadoop? Your answer should address uses cases for the Spark ecosystem.

[5 marks]

b) Briefly explain the role of RDD's in the SPARK architecture.

[6 marks]

c) Spouts and bolts are important components of the Storm architecture, describe how they are used to implement a storm solution

[5 marks]

d) What features of Erlang make it particularly suited to distributed programming problems such as map reduce?

[9 marks]

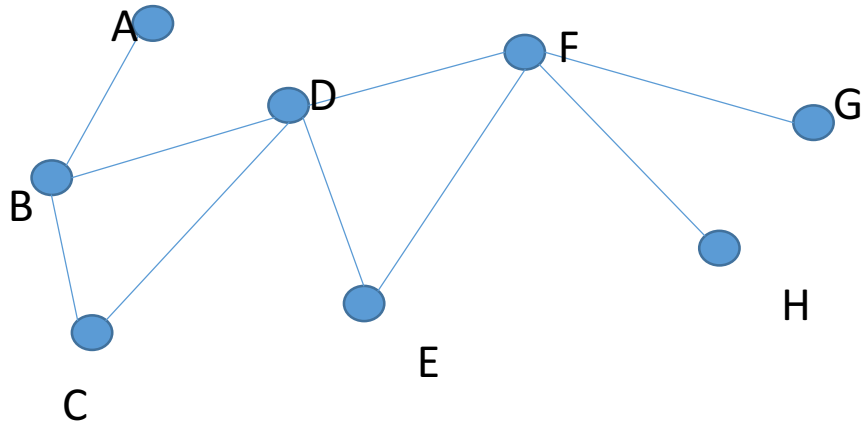
Question 5

a) Briefly suggest suitable big data technologies that you would employ in the following situation and describe implementation issues and the reasons for the technologies you have chosen:

A mobile phone operator has been keeping logs of SMS messages that its customers make. It wishes to analyse who are hubs on the network and who are outliers so it can make an effort to keep the hub users.

[8 marks]

b) In the graph below, what is the local clustering coefficient of D?.



[5 marks]

c) Explain how to calculate the overall clustering coefficient (network average clustering coefficient) for the graph.

[6 marks]

d) You have been exposed to a number of technologies for storing and analysing large amounts of data, often this could be data about people and their habits. Highlight the ethical issues and possible dangers of Big Data when used in this way.

[6 marks]

END OF PAPER