UNIVERSITY OF DUNDEE

BSc Examination

School of Science and Engineering (Computing)

2 hours

December 2015

AC41011

(Big Data Analysis)

This paper contains FIVE questions.

All questions carry equal marks.

Answer any **FOUR** questions in this paper.

Only calculators approved by the School for exam use may be used in this exam

This exam contains 5 questions, answer any 4.

Question 1

(a) Show a diagram that illustrates Brewer's CAP theorem and briefly describe its implications. Indicate where on the diagram Relational Databases and Cassandra lie.

[8 Marks]

(b) Compare and contrast the ACID database properties to the BASE model of database functionality.

[8 Marks]

(c) Big Data is often described in terms of the three Vs. State what these are and give examples of how they may materialise within different data scenarios.

[9 Marks]

Question 2

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

- Shopping cart and session data
- Completed order pdf records
- Inventory and price records
- Customer social graph

How would each of these subsystems be persisted if you are implementing Polyglot persistence? Explain the reasons for your choices.

[8 Marks]

(b) A company has a data centre in New York running a large RDMBS appliance and associated application servers. The same appliance and applications are also running in Los Angeles and Tokyo designed to share the same data to all users. What problems does this produce? Suggest an architecture solution that addresses this problem.

[7 Marks]

(c) Compare and contrast **TWO** NoSQL databases you have studied. Your answer should discuss both the data model and the physical configuration of the servers that each would require along with any other differences you wish to highlight.

[10 Marks]

Question 3

(a) What is a NameNode? How many instances of NameNode run on a Hadoop Cluster?

[4 marks]

(b) How does Hadoop implement the Map Reduce programming model? Your answer should explain how the map, reduce, shuffle and partitioner steps are distributed across Hadoop.

[9 marks]

(c) What advantages does Spark bring over Hadoop? Your answer should address uses cases for the Spark ecosystem and address the mechanisms that allow the ecosystem to implement its advantages.

[12 marks]

Question 4

(a) Describe Lambda architecture. You should include a diagram to show where data is stored in the architecture and the flow of data through it.

[10 Marks]

(b) Suggest suitable big data technologies that you would employ in the following situations and describe implementation issues and the reasons for the technologies you have chosen:

(i) A large number of sensors are producing time-stamped readings that need to be analysed at a later time. Although each sensor is producing data at a relatively slow rate, there are a large number of sensors producing data out of synch. Analysis of the readings can be done off line.

[5 Marks]

(ii) A On-Line betting firm wish to be alerted in real time when fraud is likely to be happening. The bets, along with user information, are fed into a stream of data and a relatively simple algorithm employs a sliding time window to detect the fraud. The data is not stored in this system, but an alert must be produced in a very small time window to shut down the fraudulent betting.

[5 Marks]

(c) Give a brief outline of the role of OLAP cubes and Business Intelligence in the era of Big Data

[5 Marks]

Question 5

(a) Explain how the functional programming model differs from the procedural or object oriented models.

[10 marks]

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

[10 marks]

(c) In the graph below, what is the local clustering coefficient of C?

[5 marks]