B6

a) Explain and give an example of **each** of the following two machine learning processes that are used to filter data:

i) Collaborative filtering

(5 marks)

ii) Content-based filtering.

(5 marks)

b) Briefly explain the unsupervised machine learning technique of clustering and give an example of its use.

(5 marks)

c) Briefly describe the following two types of cluster algorithm:

i) Connectivity-based

(3 marks)

ii) Distribution-based.

(3 marks)

d) Explain the purpose of the following standard R function and state the meaning of the arguments that are passed to it:

i) kmeans (x, centers, iter.max = 10, nstart = 20)

(4 marks)

End of Examination

BCS THE CHARTERED INSTITUTE FOR IT

BCS HIGHER EDUCATION QUALIFICATIONS
BCS Level 5 Diploma in IT

BIG DATA MANAGEMENT

Wednesday 20th April 2022 - Afternoon

Answer any FOUR questions out of SIX. All questions carry equal marks.

Time: TWO hours

Answer any Section A questions you attempt in Answer Book A
Answer any Section B questions you attempt in Answer Book B

The marks given in brackets are **indicative** of the weight given to each part of the question.

Calculators are **NOT** allowed in this examination.

Section A
Answer Section A questions in Answer Book A

A1

a) Explain ways in which the following two characteristics of Big Data influence the choice of Big Data storage techniques:

i) Volume

(6 marks)

ii) Velocity.

(6 marks)

b) Explain the valence characteristic of Big Data and state **TWO** reasons why data with a high degree of valence is often regarded as a challenge for the analysis of Big Data.

(13 marks)

A2

a) Explain the following two tasks in the Map phase of a MapReduce job:

i) Map task

(6 marks)

ii) Combine task.

(6 marks)

b) Describe the basic components of the Spark framework and state **TWO** advantages of using Spark compared to using MapReduce in a Big Data processing task.

(13 marks)

A3

a) Explain and give an example of **each** of the following two categories of data analytics:

i) Diagnostic analytics

(6 marks)

ii) Predictive analytics.

(6 marks)

b) Briefly explain what a hash function is and state why a good hash function should be deterministic and uniform.

(6 marks)

c) Give an example of how simple hashing can be used to allow rapid look-up of entries in a key-value pairs database.

(7 marks)

Section B
Answer Section B questions in Answer Book B

B4

a) Explain the meaning of the term ‘cloud computing’.

(5 marks)

b) Explain **THREE** ways in which a cloud computing infrastructure can offer significant benefits to a Big Data initiative.

(12 marks)

c) Explain some of the consequences arising from the ethical and legal implications of using a cloud services provider for hosting a Big Data project that contains personal and financial data.

(8 marks)

B5

a) Describe **FOUR** ways in which a NoSQL document storage device differs from a NoSQL key-value pair storage device.

(5 marks)

b) Explain why a distributed database system is generally regarded as being unable to satisfy both the availability and consistency properties of Brewer’s CAP theorem.

(8 marks)

c) Briefly explain the term ‘eventual consistency’ when used to describe the behaviour of a distributed database system.

(4 marks)

d) Briefly describe the MongoDB document storage system and state how it can achieve both consistency and read availability.

(8 marks)

[Turn Over]