BCS THE CHARTERED INSTITUTE FOR IT

BCS HIGHER EDUCATION QUALIFICATIONS
BCS Level 5 Diploma in IT

BIG DATA MANAGEMENT

Wednesday 6th October 2021 - Afternoon

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

Time: TWO hours

Answer any <u>Section A</u> questions you attempt in <u>Answer Book A</u>
Answer any <u>Section B</u> questions you attempt in <u>Answer Book B</u>

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 why many organisations identify data storage as a challenge when considering the adoption of Big Data management techniques.

(7 marks)

b) Describe the stages involved in the following two Big Data processes:

   i)   Data acquisition and filtering;

(6 marks)

   ii) Data validation and cleansing.

(6 marks)

c) Explain why security issues for Big Data infrastructure are often different from the security issues of traditional enterprise data infrastructure.

(6 marks)

A2

a) Explain how a NoSQL key-value storage device stores key-value pair data.

(7 marks)

b)
   i)   Give **THREE** examples of when the use of a key-value pair storage device is appropriate.

(4 marks)

   ii) Give **THREE** examples of when the use of a key-value pair storage device is inappropriate.

(4 marks)

c) Briefly explain how a NoSQL document database stores document data.

(5 marks)

d) Describe **THREE** ways in which document storage devices differ from key-pair storage devices.

(5 marks)

A3

a) Explain the basic principles and use of the following two types of statistical analysis commonly used to analyse Big Data:

    i) Correlation;

(7 marks)

    ii) Regression.

(7 marks)

b) The cardat dataset below shows car name, fuel consumption (`mpg`), number of cylinders (`cyl`), displacement (`disp`), engine horsepower (`hp`) and time taken to travel a quarter of a mile (`qsec`).

Dataset `cardat`:

```
                mpg   cyl   disp    hp    qsec
Mazda RX4       21.0   6     160   110   16.46
Mazda RX4 Wag   21.0   6     160   110   17.02
Datsun 710      22.8   4     108    93   18.61
Hornet 4 Drive  21.4   6     258   110   19.44
Hornet Sport    18.7   8     360   175   17.02
Valiant         18.1   6     225   105   20.22
```

Using R core functions write R scripts to perform the following four statistical investigations:

    i) The Spearman correlation between `hp` and `mpg`.

(2 marks)

    ii) The Pearson correlation between **all** possible pairs of variables with the results rounded to three decimal places.

(2 marks)

    iii) The correlation test which tests the alternative hypothesis that the correlation between the variable `qsec` and `hp` is significantly different from zero.

(3 marks)

    iv) The multiple regression coefficients of a model with `mpg` as a response variable and `disp`, `hp`, and `qsec` as predictor variables.

(4 marks)

[Turn Over]

B4

a)

    i) Give **THREE** examples of Big Data sources that might be found internal to an organisation.

(4 marks)

    ii) Give **THREE** examples of Big Data sources that might come from an external data supplier.

(4 marks)

b) Explain the data quality factors that need to be considered in ensuring quality of service for a Big Data initiative.

(5 marks)

c) Briefly describe the following three types of data storage:

    i) Block storage;

(4 marks)

    ii) File storage;

(4 marks)

    iii) Object storage.

(4 marks)

B5

a) Describe **TWO** main goals of the Hadoop distributed file system (HDFS).

(5 marks)

b) Explain **THREE** key benefits of using the MapReduce framework for batch processing in a Big Data environment.

(12 marks)

c) Explain why MapReduce is considered by many as unsuitable for real-time Big Data processing tasks.

(8 marks)

B6

a) Explain the type of analytics carried out in each of the following two categories of Gartner's analytic ascendancy model:

    i) Predictive analytics; **(6 marks)**

    ii) Prescriptive analytics. **(6 marks)**

b) Explain what an artificial neural network is and state how it is trained.

(6 marks)

c) Describe and give an example of a supervised learning technique to perform data classification.

(7 marks)

End of Examination