(6 pages)

Reg. No. : ....

Code No. : 7068 **Sub. Code : ZCAM 31**

M.C.A. (CBCS) DEGREE EXAMINATION,
NOVEMBER 2022.

Third Semester

Computer Application – Core

DATA SCIENCE AND ANALYTICS

(For those who joined in July 2021 onwards)

Time : Three hours Maximum : 75 marks

PART A — (10 × 1 = 10 marks)

Answer ALL the questions.

Choose the correct answer :

1. How do we perform Bayesian classification when some features are missing?

   (a) We integrate the posteriors probabilities over the missing features

   (b) We ignore the missing features

   (c) We assuming the missing values as the mean of all values

   (d) Drop the features completely

2. Data science is the process of diverse set of data through?

   (a) Organizing data (b) Processing data

   (c) Analysing data (d) All of the above

3. Which of the following is required by K-means clustering?

   (a) defined distance metric

   (b) number of clusters

   (c) initial guess as to cluster centroids

   (d) all of the mentioned

4. Which of the following methods do we use to best fit the data in Logistic Regression?

   (a) Least Square Error

   (b) Maximum Likelihood

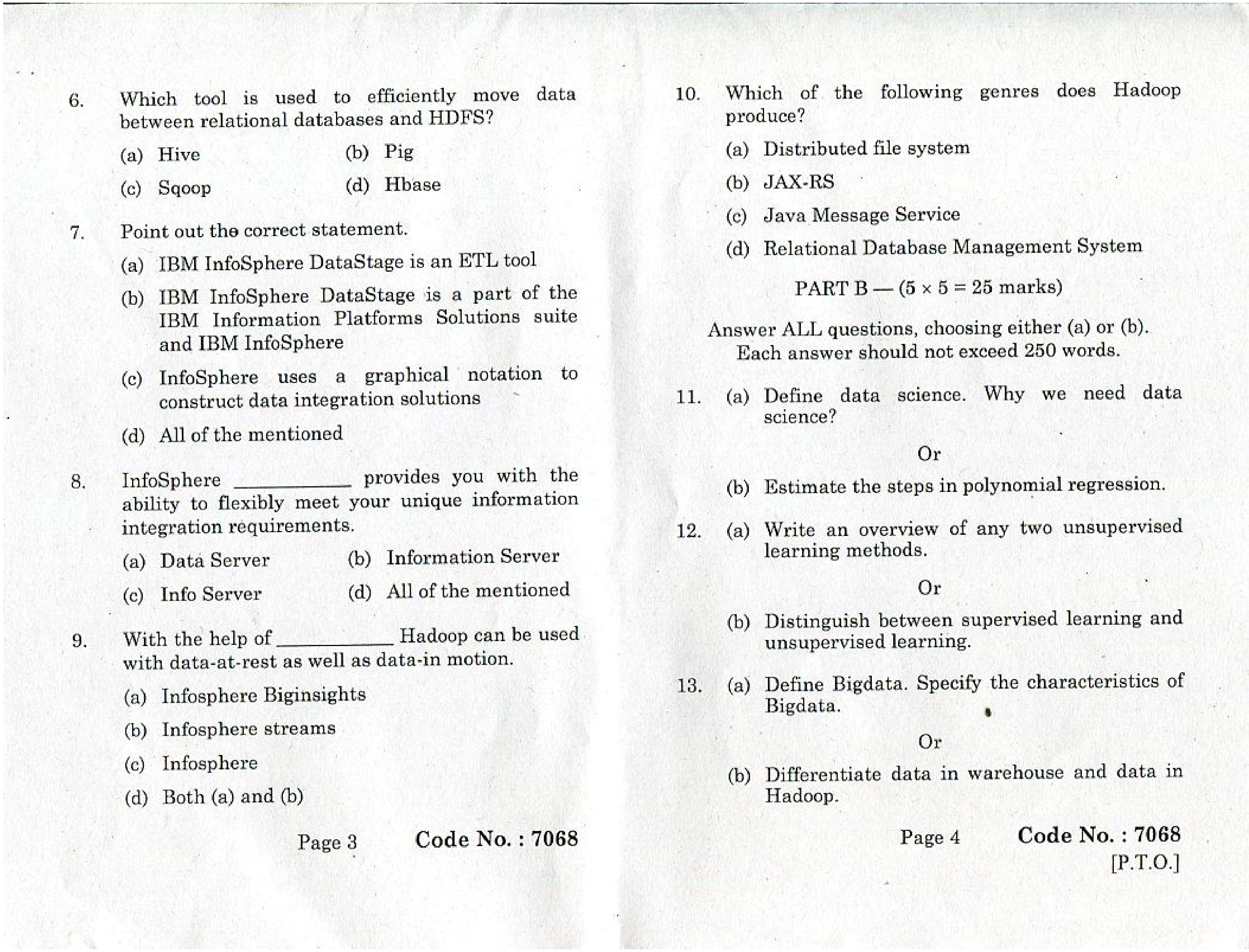
   (c) Jaccard distance

   (d) Both (a) and (b)

5. ____________ is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

   (a) Hive (b) MapReduce

   (c) Pig (d) Lucene



6. Which tool is used to efficiently move data between relational databases and HDFS?

(a) Hive (b) Pig

(c) Sqoop (d) Hbase

7. Point out the correct statement.

(a) IBM InfoSphere DataStage is an ETL tool

(b) IBM InfoSphere DataStage is a part of the IBM Information Platforms Solutions suite and IBM InfoSphere

(c) InfoSphere uses a graphical notation to construct data integration solutions

(d) All of the mentioned

8. InfoSphere __________ provides you with the ability to flexibly meet your unique information integration requirements.

(a) Data Server (b) Information Server

(c) Info Server (d) All of the mentioned

9. With the help of __________ Hadoop can be used with data-at-rest as well as data-in motion.

(a) Infosphere Biginsights

(b) Infosphere streams

(c) Infosphere

(d) Both (a) and (b)

10. Which of the following genres does Hadoop produce?

(a) Distributed file system

(b) JAX-RS

(c) Java Message Service

(d) Relational Database Management System

PART B — (5 × 5 = 25 marks)

Answer ALL questions, choosing either (a) or (b).
Each answer should not exceed 250 words.

11. (a) Define data science. Why we need data science?

Or

(b) Estimate the steps in polynomial regression.

12. (a) Write an overview of any two unsupervised learning methods.

Or

(b) Distinguish between supervised learning and unsupervised learning.

13. (a) Define Bigdata. Specify the characteristics of Bigdata.

Or

(b) Differentiate data in warehouse and data in Hadoop.

[P.T.O.]

14. (a) How to install Infoshphere BigInsights. Mention the components included in BigInsights 1.2.

Or

(b) Appraise Hadoop compression technique.

15. (a) Examine the Infosphere Stream basics.

Or

(b) Structure the Infosphere streams tool kits.

PART C — (5 × 8 = 40 marks)

Answer ALL questions, choosing either (a) or (b)
Each answer should not exceed 600 words.

16. (a) Speculate the Bayes rule supervised learning.

Or

(b) Intervene the prerequisite probability concepts for Bayes rule.

17. (a) Elucidate Naïve Bayes classifier.

Or

(b) Explain logistic regression and its different types.

18. (a) Paraphrase

(i) Importance of Bigdata

(ii) Bigdata use cases

Or

(b) Generalize the components of Hadoop.

19. (a) Generalize the Data Discovery and Visualization

Or

(b) Formulate the concepts behind General Parallel file System.

20. (a) Elucidate on industry use cases for InfoSphere Streams.

Or

(b) Elaborate on the Streams Processing Language.