CBCS SCHEME

USN

17CS82

Eighth Semester B.E. Degree Examination, Feb./Mar. 2022
Big Data Analytics

Time: 3 hrs. Max. Marks: 100

Note: Answer any FIVE full questions, choosing ONE full question from each module.

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

Module-1

1 a. What is HDFS? With a neat diagram explain the components of HDFS (Hadoop Distributed File Systems) **(10 Marks)**

b. With a neat diagram, discuss the steps MapReduce parallel data flow with example of word count. **(10 Marks)**

OR

2 a. Explain Block replication in HDFS and its advantages. **(05 Marks)**

b. Explain the following roles in HDFS deployment with a diagram:
(i) High Availability (ii) Name Node Federation. **(10 Marks)**

c. With example, explain the following general HDFS commands:
(i) HDFS version (ii) List files (iii) Make directory
(iv) Copy files (v) Delete a file **(05 Marks)**

Module-2

3 a. What is the significance of Apache pig in Hadoop context? Describe the main components and the working of Apache pig with a simple example. **(10 Marks)**

b. Explain Apache squoop import and export method with neat diagrams. **(10 Marks)**

OR

4 a. With a neat diagram, explain Oozie DAG workflow and its types of nodes. **(10 Marks)**

b. Describe the various features of hadoop YARN administration. **(05 Marks)**

c. Discuss the three components of Apache frame. **(05 Marks)**

Module-3

5 a. Discuss how the data contributes to decision making in business intelligence. **(05 Marks)**

b. Justify the differences between datamart and data warehouse based on following :
(i) Scope (ii) Target organization (iii) Cost (iv) Approach (v) Time. **(10 Marks)**

c. Consider three dimensions of data warehouse:
Bank branch, time period, Loans and two measures accounts and Total balance, where total balance is outstanding loan amount from customers. Sketch star schema for above model. **(05 Marks)**

OR

6 a. Explain cross-industry standard process for data mining with a neat diagram. **(10 Marks)**

b. With a neat block diagram, describe the architecture of data warehouse. **(10 Marks)**

Module-4

7 a. Differentiate between Linear, Non-linear and Logistic Regression models. **(10 Marks)**

b. Employ decision tree learning (Total error based) for the following dataset where the objective is to predict the Class Category-Loan approved or not (C_0 & C_1). Find out class for

M	Luxury	Medium	?

Customer Id	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C_0
2	M	Sports	Medium	C_0
3	M	Sports	Medium	C_0
4	M	Sports	Large	C_0
5	M	Sports	Extra Large	C_0
6	M	Sports	Extra Large	C_0
7	F	Sports	Small	C_0
8	F	Sports	Small	C_0
9	F	Sports	Medium	C_0
10	F	Luxury	Large	C_0
11	M	Family	Large	C_1
12	M	Family	Extra Large	C_1
13	M	Family	Medium	C_1
14	M	Luxury	Extra Large	C_1
15	F	Luxury	Small	C_1
16	F	Luxury	Small	C_1
17	F	Luxury	Medium	C_1
18	F	Luxury	Medium	C_1
19	F	Luxury	Medium	C_1
20	F	Luxury	Large	C_1

(10 Marks)

OR

8 a. Explain the design principles of ANN by constructing a model for multilayer ANN. **(07 Marks)**

b. What is unsupervised learning? Describe 3 applications of cluster analysis. **(06 Marks)**

c. How does the Apriori algorithm for association rule mining works? Explain with example. **(07 Marks)**

Module-5

9 a. Discuss the importance of term document matrix in text mining with a neat diagram of Text Mining architecture. **(08 Marks)**

b. Explain the advantages and disadvantages of Naïve-Bayes classifier. **(04 Marks)**

c. What is support vector machine? Explain its model. **(08 Marks)**

OR

10 a. Discuss web structure mining and compute the rank values for the following network in Fig.Q10(a). Which is the highest ranked node?

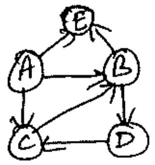

Fig.Q10(a) **(12 Marks)**

b. Discuss the application and practical consideration of social network analysis. **(08 Marks)**

* * * * *