Roll No:

B TECH
(SEM-V) THEORY EXAMINATION 2020-21
DATA ANALYTICS

Time: 3 Hours   *Total Marks: 100*

Note: **1.** Attempt all Sections. If require any missing data; then choose suitably.

SECTION A

1. **Attempt *all* questions in brief.**   **2 x 10 = 20**

Q no.	Question	Marks	CO
a.	What are the different types of data?	2	1
b.	Explain decision tree.	2	1
c.	Give the full form of RTAP.	2	3
d.	List various phases of data analytics lifecycle.	2	1
e.	Explain the role of Name Node in Hadoop.	2	5
f.	Discuss heartbeat in HDFS.	2	5
g.	Differentiate between an RDBMS and Hadoop.	2	5
h.	Write names of two visualization tools.	2	4
i.	How can you deal with uncertainty?	2	3
j.	Data sampling is very crucial for data analytics. Justify the statement.	2	3

SECTION B

2. **Attempt any *three* of the following:**

Q no.	Question	Marks	CO
a.	Explain K-Means algorithms. When would you use k means? State weather the statement "K-Means has an assumption each cluster has a roughly equal number of observations" is true or false. Justify your answer	10	4
b.	Illustrate and explain the steps involved in Bayesian data analysis.	10	2
c.	Suppose that A, B, C, D, E and F are all items. For a particular support threshold, the maximal frequent item sets are {A, B, C} an {D, E}. What is the negative border?	10	1
d.	Discuss any two techniques used for multivariate analysis.	10	2
e.	Design and explain the architecture of data stream model.	10	3

SECTION C

3. **Attempt any *one* part of the following:**

Q no.	Question	Marks	CO
a.	Describe the architecture of HIVE with its features.	10	5
b.	Brief about the main components of MapReduce	10	5

4. **Attempt any *one* part of the following:**

Q no.	Question	Marks	CO
a.	Describe any two data sampling techniques.	10	1
b.	Explain any one algorithm to count number of distinct elements in a Data stream.	10	3

PAPER ID-310418

Roll No:

5. **Attempt any *one* part of the following:**

Q no.	Question	Marks	CO
a.	Brief about the working of CLIQUE algorithm.	10	4
b.	Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) Initial cluster centers are A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P (a, b) = |x2 – x1| + |y2 – y1| Use K-Means Algorithm to find the three cluster centers after the second iteration	10	4

6. **Attempt any *one* part of the following:**

Q no.	Question	Marks	CO
a.	What is prediction error? State and explain the prediction error in regression and classification with suitable example.	10	4
b.	Given data = {2, 3, 4, 5, 6, 7; 1, 5, 3, 6, 7, 8}. Compute the principal component using PCA Algorithm.	10	2

7. **Attempt any *one* part of the following:**

Q no.	Question	Marks	CO
a.	Develop and explain the data analytics life cycle	10	1
b.	Distinguish between supervised and unsupervised learning with example.	10	1

PAPER ID-410482

Roll No: | | | | | | | | | | | | | |

B. TECH/ M. TECH (INTEGRATED)
(SEM: V) THEORY EXAMINATION 2021-22
DATA ANALYTICS

Time: 3 Hours   *Total Marks: 100*

Note: 1. Attempt all Sections. If require any missing data; then choose suitably.

SECTION A

1.   Attempt *all* questions in brief.   2 x 10 = 20

Qno.	Question	Marks	CO
a.	Differentiate between Predictive and Prescriptive Data Analytics.	2	CO1
b.	Differentiate between Analysis and Reporting.	2	CO1
c.	What is Lasso Regression?	2	CO2
d.	Differentiate between Univariate & Multivariate Analysis.	2	CO2
e.	How is steam Processing Different from Traditional Data Processing?	2	CO3
f.	What is the role of Sliding Window in Analysis of Streaming Data?	2	CO3
g.	Explain the Principle behind Hierarchal clustering Technique.	2	CO4
h.	Define Lift in Association Data Mining.	2	CO4
i.	What is the basic description of a Box Plot In R	2	CO5
j.	List two Data Visualization Tool	2	CO5

SECTION B

2.   Attempt any *three* of the following:

Qno.	Question	Marks	CO
a.	Explain the Process Model and Computation Model of Big Data Platform	10	CO1
b.	Explain the Working of an Artificial Neural Network for Image Classification Task.	10	CO2
c.	Discuss the Publish/ Subscribe Model of Streaming Architecture.	10	CO3
d.	What are the advantages of PCY algorithm over Apriori Algorithm?	10	CO4
e.	What makes NoSQL Databases different from RDBMS	10	CO5

SECTION C

3.   Attempt any *one* part of the following:

Qno.	Question	Marks	CO
a.	Discusses the steps involved in Data Analysis Process.	10	CO1
b.	Compare and Contrast Traditional Analytics Structure to Modern Analytics Architecture.	10	CO1

4.   Attempt any *one* part of the following:

Qno.	Question	Marks	CO
a.	Discuss different types of Time Series Data Analysis along with its Major Application area.	10	CO2
b.	Differentiate different types of support vector and kernel methods of Data Analysis	10	CO2

Roll No:

5. **Attempt any *one* part of the following:**

Qno.	Question	Marks	CO
a.	Discuss the Components of a General Stream Processing Model. List few sources of Streaming Data	10	CO3
b.	Explain and apply Flajolet-Martin algorithm on the following stream of data to identify unique elements in the stream. S=1,3,2,1,2,3,4,3,1,2,3,1S=1,3,2,1,2,3,4,3,1,2,3,1 Given: h(x)=(6x+1) mod 5	10	CO3

6. **Attempt any *one* part of the following:**

Qno.	Question	Marks	CO
a.	Differentiate between CLIQUE and ProCLUS clustering	10	CO4
b.		10	CO4

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Find all the Association rule from the above given Transaction with Given minsup = 50%, minconf = 50 %.

7. **Attempt any *one* part of the following:**

Qno.	Question	Marks	CO
a.	Explain the working of Hadoop Distributed File Systems.	10	CO5
b.	List and explain five R function used in Descriptive Statistics	10	CO5

B.Tech.
(SEM V) THEORY EXAMINATION 2022-23
DATA ANALYTICS

Time: 3 Hours   *Total Marks: 100*

Note: Attempt all Sections. If require any missing data; then choose suitably.

SECTION A

1. **Attempt *all* questions in brief.**   **2 x 10 = 20**

    (a) What are the common open source tools for the model building phase?
    (b) What is decision tree?
    (c) What is learning rate?
    (d) What is rule induction?
    (e) How can you deal with uncertainty?
    (f) What is DSMS?
    (g) Write names of two visualization tools.
    (h) Explain the principle behind Hierarchical clustering technique.
    (i) Differentiate between Pig and SQL.
    (j) Write a short note on R graphical user interface.

SECTION B

2. **Attempt any *three* of the following:**   **10x3 = 30**

    (a) Compare and contrast traditional analytics structure to modern analytics architecture.
    (b) Explain multivariate analysis and Bayesian network.
    (c) Explain Datar-Gionis-Indyk-Motwani (DGIM) algorithm for counting oneness in a window.
    (d) Why PCY algorithm is preferred over Apriori algorithm?
    (e) How RDBS is different from NoSQL?

SECTION C

3. **Attempt any *one* part of the following:**   **10x1 = 10**

    (a) Explain Apache Hadoop , KNIME & Open refine in detail.
    (b) Explain various phases of Data Analytics Life Cycle.

4. **Attempt any *one* part of the following:**   **10x1 = 10**

    (a) Differentiate between Crisp logic and Fuzzy logic.
    (b) What are the different kernel methods of Data Analytics?

5.   **Attempt any *one* part of the following:**   **10x1 = 10**

(a) Explain Bernoulli sampling with its algorithm.
(b) What are the different components of a general stream processing model? List few sources of streaming data .

6.   **Attempt any *one* part of the following:**   **10x1 = 10**

(a) What is Prediction error ? With the help of suitable example explain prediction error in classification and regression.
(b) Explain SON algorithm to find all or most frequent item sets using at most two passes.

7.   **Attempt any *one* part of the following:**   **10x1 = 10**

(a) Draw and discuss the architecture of Hive in detail.
(b) What are the approaches to integrate the human in data exploration process to realize different types of approaches to visual data mining?

Data Analytics AKTU - Previous year paper 2023-2024

B.tech (Dr. A.P.J. Abdul Kalam Technical University)

Scan to open on Studocu

Roll No:

BTECH
(SEM VI) THEORY EXAMINATION 2023-24
DATA ANALYTICS

TIME: 3 HRS   M.MARKS: 100

5.	**Attempt any *one* part of the following:**	**10 x 1 = 10**
a.	Discuss a case study that demonstrates the use of stream processing in stock market prediction, including the techniques used and the results achieved.	10	3
b.	Explain Real-time Analytics Platforms (RTAP), and how do they enable real-time decision-making and analytics.	10	3
6.	**Attempt any *one* part of the following:**	**10 x 1 = 10**
a.	Describe the K-means clustering algorithm and the process of initializing centroids.	10	4
b.	Explain how ProCLUS (Projected Clustering) algorithm differs from CLIQUE in its approach to high-dimensional clustering.	10	4
7.	**Attempt any *one* part of the following:**	**10 x 1 = 10**
a.	Compare and contrast NoSQL databases with traditional SQL databases. List four main types of NoSQL databases.	10	5
b.	Describe Apache Hive and its use in querying and managing large datasets stored in Hadoop.	10	5

BTECH
(SEM VI) THEORY EXAMINATION 2023-24
DATA ANALYTICS

TIME: 3 HRS **M.MARKS: 100**

Note: 1. Attempt all Sections. If require any missing data; then choose suitably.

SECTION A

1. Attempt *all* questions in brief. 2 x 10 = 20

Qno.	Question	Marks	CO
a.	List the steps involved in the analytic process.	2	1
b.	Differentiate data analysis from data reporting.	2	1
c.	Explain the concept of Bayesian inference.	2	2
d.	Define the concept of the 'kernel trick'.	2	2
e.	List the uses of sliding windows in stream processing.	2	3
f.	Explain the concept of a decaying window.	2	3
g.	Mention the importance of frequent itemsets in market basket analysis.	2	4
h.	Explain the significance of support and confidence in the context of frequent itemsets.	2	4
i.	Define Hadoop Distributed File System.	2	5
j.	List some visual data analysis techniques.	2	5

SECTION B

2. Attempt any *three* of the following: 10 X 3 = 30

		Marks	CO
a.	Discuss the key characteristics of data that impact data analytics.	10	1
b.	Differentiate linear and logistic regression models, and in what scenarios would each be most appropriate.	10	2
c.	Discuss differences between traditional batch data processing and stream data processing, and the main challenges associated with stream processing.	10	3
d.	Mention the limitations of the Apriori algorithm, and how they are addressed through optimization techniques.	10	4
e.	Explain the MapReduce programming model and its role in processing large datasets.	10	5

SECTION C

3. Attempt any *one* part of the following: 10 x 1 = 10

		Marks	CO
a.	Discuss the various phases of the data analytics lifecycle and activities involved in each phase.	10	1
b.	Explain modern data analytics tools in detail.	10	1

4. Attempt any *one* part of the following: 10 x 1 = 10

		Marks	CO
a.	Describe the process of conducting a principal component analysis (PCA) and explain how it can be used to reduce the dimensionality of a dataset.	10	2
b.	Explain how support vector machines (SVMs) utilize kernel functions to handle non-linearly separable data.	10	2