PAPER ID-310418

Subject Code: KCS051

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

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