B.E/B.TECH Degree Examination, December 2020

Seventh Semester

CS16004 - DATA ANALYTICS

(Regulation 2016)

Time: Three hours Maximum : 80 Marks

Answer **ALL** questions

PART A - (8 X 2 = 16 marks)

1. Real-time data stream is _____

   A) Sequence of data items that arrive in some order and may be seen only once.

   B) Sequence of data items that arrive in some order and may be seen twice.

   C) sequence of data items that arrive in same order

   D) sequence of data items that arrive in different order

2. Which of the classification algorithm uses a hyperplane which separates the data into classes

   A) SVM Classifier

   B) PCY Algorithm

   C) K-Nearest neighbor

   D) BFR Algorithm

3. Identify the property of frequent itemsets which is defined as follows 'If a set of items in a dataset is frequent , then so are all its subsets'

   A)Support

   B) Confidence

   C) Monotonicity

   D) Distinct

4. _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

   a) MapReduce

   b) Mahout

   c) Oozie

   d) All of the mentioned

5. How Data analysis is different from that of reporting?
6. When Principal Component Analysis in data modeling is preferred?
7. Justify how DGIM algorithm performs better in counting ones in a window of data stream.
8. How does visualization approaches support big data analysis?

PART B - (4 X16 = 64 marks)

09. (a) (i) Discuss the challenges involved in big data analytics and map the challenges to any real time application. **(6)**

(ii) Why is sampling required in Big data? Discuss the role of sampling distribution and resampling in data analysis. **(10)**

(OR)

(b) (i) Analyze the role of web data mining in online shopping. **(8)**

(ii) When is Internal statistical inference preferred over point estimation? Justify with an example. **(8)**

10. (a) (i) When is Analysis of Variance (ANOVA) used? Describe the steps involved in ANOVA process. **(8)**

(ii) Apply Genetic stochastic search technique to any application. Discuss the steps involved in the design. **(8)**

(OR)

(b) (i) Analyze the efficiency of learning in linear regression modeling **(8)**

(ii) When is fuzzy logic preferred? Discuss on how fuzzy logic is integrated with decision tree to model big data. **(8)**

11. (a) (i) Using AMS algorithm, compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. **(8)**

(ii) Discuss on how Real Time Analysis Platform (RTAP) is applied for Sentiment analysis. **(8)**

(OR)

(b) (i) Apply Flajolet-Martin algorithm to estimate the number of distinct elements in a data stream S = 1,3,2,1,2,3,4,3,1,2,3,1 using hash function h(x)=(6x+1) mod 5 **(8)**

(ii) When filtering is required on Data streams? Analyze the performance of Bloom filter in comparison with other approaches. **(8)**

12. (a) (i) Justify the need of Map reduce Architecture for Big data analytics. **(8)**

(ii) Discuss the working of Park-Chen-Yu algorithm in frequent itemset mining and analyze its performance. **(8)**

(OR)

(b) (i) Justify the need of NoSQL databases in real time applications. **(8)**

(ii) How is clustering performed using Bradley-Fayyad-Reina algorithm? Analyze its performance in comparison with k-means clustering. **(8)**