Paper / Subject Code: 52705 / 5)Big Data Analytics

Time: 3 Hours **Total Marks: 80**

N.B. Question No: 1 is Compulsory
Attempt any three from the remaining
Assume suitable data wherever necessary

1 **a** Find Manhattan distance for the points X1= (1,2,2) , X2 = (2,5,3) **5**

b How finding plagiarism in documents is a nearest neighbor problem. **5**

c Draw and Explain Bow-tie structure of web. **5**

d How big data problems are handled by Hadoop system. **5**

2 **a** Explain how Hadoop goals are covered in hadoop distributed file system. **10**

b Write pseudo code for Matrix vector Multiplication by MapReduce. Illustrate with an example showing all the steps. **10**

3 **a** The snapshot of 10 transactions is given below for online shopping that generates big data. Threshold value = 4 and Hash function= (i*j) mod 10 **10**

T1 = {1, 2, 3} T2 = {2, 3, 4} T3 = {3, 4, 5}
T4 = {4, 5, 6} T5 = {1, 3, 5} T6 = {2, 4, 6}
T7 = {1, 3, 4} T8 = {2, 4, 5} T9 = {3, 4, 6} T10 = {1, 2, 4}

Find the frequent item sets purchased for such big data by using suitable algorithm. Analyse the memory requirements for it.

b Explain DGIM algorithm for counting ones in stream with example. **10**

4 **a** How recommendation is done based on properties of product? Explain with suitable example. **10**

b Explain how the CURE algorithm can be used to cluster big data sets. **10**

5 **a** What are the different architectural patterns in NoSQL? Explain Graph data store and Column Family Store patterns with relevant examples. **10**

b Explain Girvan-Newman algorithm to mine Social Graphs. **10**

6 **a** List down the steps in modified Page Rank Algorithm to avoid spider trap with one example. **10**

b Explain Park-Chen-Yu algorithm. How memory mapping is done in PCY. **10**

****

68858