Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(An Autonomous Institute Affiliated to University of Mumbai)

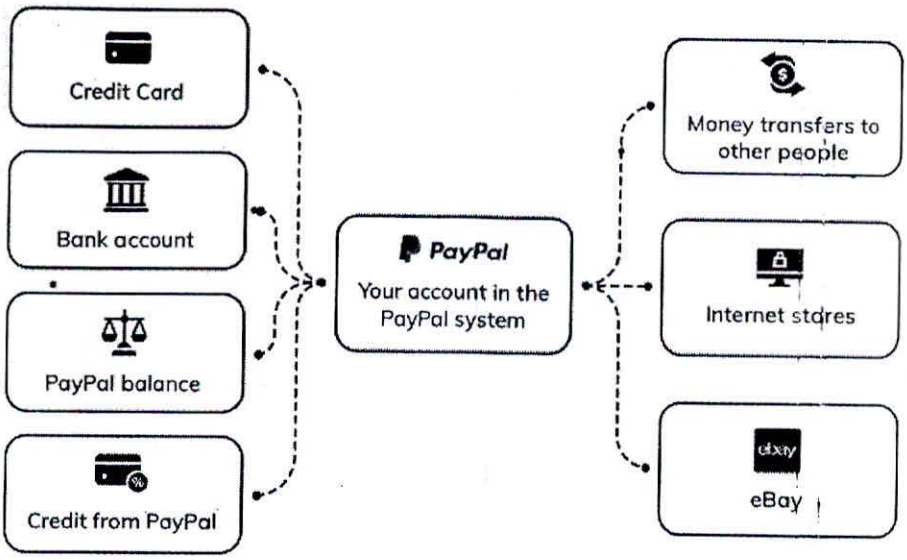
End Semester Examination

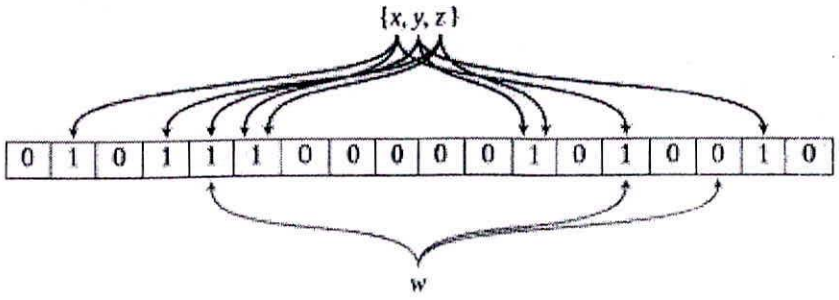
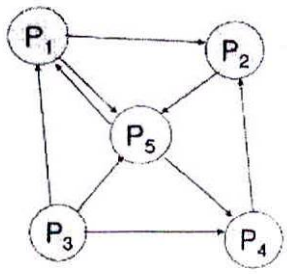
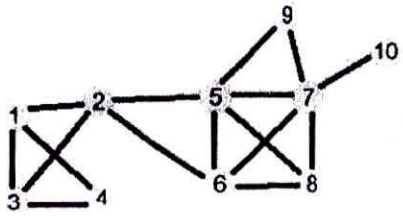
May 2023

Max. Marks: 100 | **Course Code:IT307A**
Class: TE IT | **Semester: VI**
Name of the Course: Big Data Analytics | **Duration: 3 Hours**

Instructions:
(1) All Questions are Compulsory
(2) Draw neat diagrams
(3) Assume suitable data if necessary

Question No.		Max. Marks	CO
Q1	Consider a sample sales data having columns like Order_id Product_id Product_category - (Furniture/Electronics/……..) Quantity CostOfEach - cost of one item Order_Date - DD/MM/YYYY Customer_name Customer_city Customer_state Deal_Size - (Small/Mediu,/Large) Shipment_mode - (air/land/sea) Write Hadoop map reduce code in java for following analysis.		CO3
Q1[A]	Find the most popular shipment mode in a given state. Take the state input from the user. **OR** Find the most frequent Deal_size in a given city. Take the city input from the user.	10M	
Q1[B]	Print top 3 product categories making maximum sale in each state.	10M	

Q2[A]	 PayPal was one of the first payment solutions to try and civilize online payments. PayPal handles a large volume of financial transactions every day, generating a large amount of data related to payments, transactions, user behavior, and more. Today, PayPal has over 377 million active accounts in 200 global markets. PayPal has also made significant acquisitions in the past two years. The transactions processed by PayPal occur in real-time, requiring fast data processing and analysis to keep up with the pace of data generation. 9.3 billion transactions happened on PayPal in 2021. Considering the above aspects of the case study, justify why PayPal can be considered as a big data case study.	10M	CO2
Q2[B]	Apply the Count min sketch algorithm to find the final sketch after inserting data stream A,C,A,D,B,A,C,B,D,A,D using 3 hash functions H1,H2 and H3 defined as follows. Draw a separate sketch of every iteration. Find the occurrence of D in the final sketch. Does it match with actual occurrence? Justify your answer. | | A | B | C | D | | H1 | 1 | 3 | 2 | 3 | | H2 | 2 | 3 | 2 | 2 | | H3 | 1 | 3 | 1 | 3 |	10M	CO2
Q2[C]	Following diagram shows a sketch of the bloom filter after inserting x,y,z and w in that order. With the help of this sketch show that Bloom filter does not support deletion.	10M	CO2

			
Q3 [A]	A pair which is not frequent can be counted as a frequent pair in the PCY algorithm. Do you agree? Illustrate this with the example. Discuss improved PCY algorithm to overcome this problem.	10M	CO4
Q3 [B]	Use the HITS algorithm to identify authority and hub pages in the following graph. Solve up to 5 iterations with normalization. 	10M	CO4
Q3 [C]	Consider the following graph as a part of a social network graph. Identify how many communities can be formed using the Clique Percolation Method for k=3. Clearly state the nodes in each community. Show all steps. 	10M	CO4
Q4 [A]	Find Eigenvalues and Eigenvectors of $\begin{bmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 20 \end{bmatrix}$	10M	CO1

<table>
<tr><td></td><td>Consider a page rank algorithm of the form r = Mr. Assume above matrix represents M of page rank algorithm. What is the significance of an eigen vector in finding the page rank of the given M.</td><td></td><td></td></tr>
<tr><td>Q4 [B]</td><td>Explain with example how Singular Value Decomposition can be used for Big Data Processing

OR

Given the following data use PCA to reduce the dimensions.

<table><tr><td>Feature</td><td>Ex1</td><td>Ex2</td><td>Ex3</td><td>Ex4</td></tr><tr><td>X</td><td>4</td><td>8</td><td>13</td><td>7</td></tr><tr><td>Y</td><td>11</td><td>4</td><td>5</td><td>14</td></tr></table></td><td>10M</td><td>CO1</td></tr>
</table>