



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

Class: TE IT

Name of the Course: Big Data Analytics

Course Code: IT307A

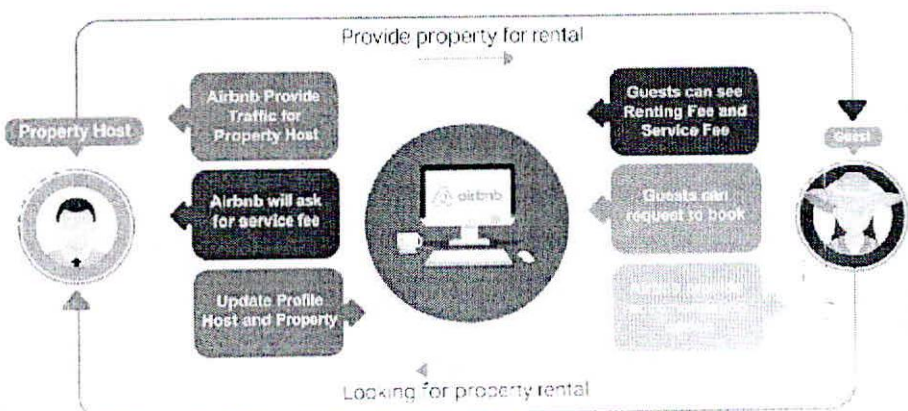
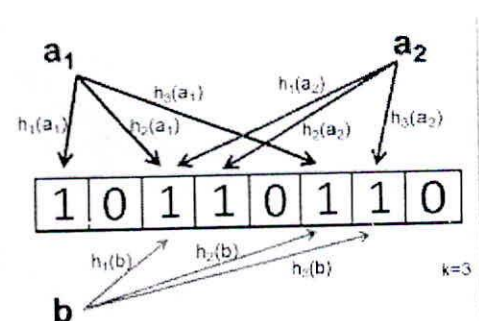
Semester: VI

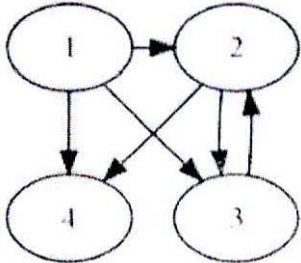
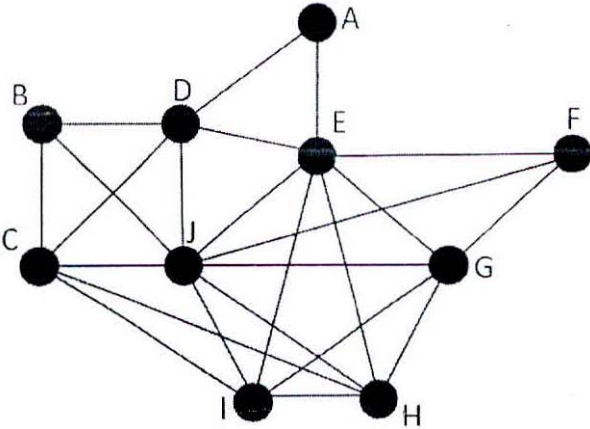
Duration: 3 Hours

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 car sales data having columns like year_of_manufacturer car_price manufacturer model condition cylinders Sales_in_thousand Engine_size Customer_name Customer_city Customer_state Write Hadoop map reduce code in java for following analysis.		CO3
Q1[A]	Find the most popular model in a given state. Take the state input from the user. OR Find the most popular Engine_size in a given city. Take the city input from the user.	10M	
Q1[B]	Print top 3 manufacturers making maximum sales in each state.	10M	

Q2[A]	<p style="text-align: center;">AIRBNB Booking System</p> <p style="text-align: center;">• • • • •</p>  <p>Airbnb is a platform that allows individuals to rent out their properties or rooms to travelers. As of 2021, Airbnb had over 7 million listings in over 100,000 cities, Airbnb has quickly become a major player in the travel industry. This industry spreads over 220 countries and regions. In the year 2020, the company's revenue was approximately \$2.5 billion.</p> <p>Considering the above aspects of the case study, justify why airbnb can be considered as a big data case study.</p>	10M	CO2
Q2[B]	<p>a_1, a_2 and b are inserted in the bloom filter as shown below. With the help of this sketch show that Bloom filter does not support deletion.</p> 	10M	CO2
Q2[C]	<p>Consider a data stream received and identify an item which is frequent using the majority algorithm. Show all steps.</p> <p>Pen, Pencil, Eraser, Pen, Pen, Eraser, Pencil, Pen, Pencil, Pen, Pen, Pen</p>	10M	CO2

Q3 [A]	<p>Using HITS algorithm identify Hubs and Authority pages in the following graph. Calculate hub and authority score for each node up to 3 iterations.</p> 	10M	CO4
Q3 [B]	<p>Apply PCY algorithm on the following dataset and show its disadvantage. $(1,2), (3,3), (1,2), (1,2), (2,4), (1,2), (1,2), (4,3), (1,2), (1,3), (1,2), (3,2), (1,2), (1,3), (2,4)$ Hash function $(i+j)\%3$</p>	10M	CO4
Q3 [C]	<p>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.</p> 	10M	CO4
Q4 [A]	<p>Explain the role of Singular Value Decomposition in Big data with example.</p> <p style="text-align: center;">OR</p> <p>Discuss with example the role of Eigenvalue in the calculation of page rank algorithm</p>	10M	CO1
Q4 [B]	<p>Reduce the dimensions of the given data set using principal component analysis.</p>	10M	CO1

	$\begin{bmatrix} 2 & 4 \\ 3 & 6 \\ 4 & 8 \\ 5 & 10 \end{bmatrix}$		
--	---	--	--