Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058, India

(Autonomous College Affiliated to University of Mumbai)

End Semester Examination
May 2022

Course Name/Code: Big Data Analytics/ IT307A

Max. Marks: 60

Class: T.Y.BTech

Semester: VI

Branch: IT

Duration:2 Hrs

Instruction:
(1) All questions are compulsory
(2) Draw neat diagrams
(3) Assume suitable data if necessary
(4) Mention the question number clearly while writing the answer

QNo	Question	Marks	CO
1 a	Consider the following Case Study. Chocolate Marketing company with large number of installed Automatic Chocolate Vending Machines (ACVM). Each ACVM sells five flavours (FL1, FL2, FL3,FL4 and FL5) KitKat, Milk, Fruit and Nut, Nougat and Oreo. Data get generated at various machine such as sales of chocolates, reports of unfilled or filled machine transaction data. Social network and web data on feedback and personalized messages based on interactions and human generated data on facial recognition of the buyers. The company uses this data for efficient and optimum planning of fill services for chocolates, sentiment analysis of buyer for specific flavours. i. Identify Different Big Data Types from the above case study. ii. Identify the challenges faced from large growth in volume of data.	5 5	CO1
b	Compare traditional versus Big data analytics with example.	5	CO1
2 a	Describe the component of SPARK Ecosystem	5	CO3

	OR Draw and explain Architecture of Apache Pig.		
b	Consider a .txt file of size 1024 MB. will be stored on HDFS. With the help of diagram demonstrate the block replication and rack awareness in HDFS.	10	CO3
3 a	Consider the pass consist of 1, 2, 3, 1, 2, 3, 4, 1, 2, 4. Hash function is 6X+1 mod 5. How many numbers of distinct elements presents in the pass? **OR** Apply the Misra Gries algorithm on the data stream consist of 2,3,9,5,7,9,9. With K=2	5	CO4
b	Identify best HUB and Authority for the given adjacency matrix. Calculate the HUB and Authority score using HITS algorithm for K=2 $A = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}$	10	CO4

4 a	Apply PCY algorithm on the following transaction to find the candidate sets (frequent sets). Use buckets and concepts of Mapreduce to solve the above problem.
	Consider Hash Function =(i*j) mod 10 and Threshold value =3
	T1= {6,5,4} T6= {5,3,1}
	T2= {5,4,3} T7= {6,4,3}
	T3= {4,3,2} T8= {5,3,2}
	T4= {3,2,1} T9= {4,2,1}
	T5= {6,4,2} T10= {6,5,3}
		10	CO2

b	Consider two matrices in the format M= [m11, m12, m21, m22] N= [n11, n12,n21,n22] where m11 can be interpreted as value in the first row and first column of matrix M and so on. The actual matrices are: M= [1,9,5,4] and N= [4,3,6,7] In case of multiplication of M and N using 2 phase MapReduce, what will be the output of the second map phase?	5	CO2