

## Sardar Patel Institute of Technology Bhavan's Campus, Munshi Nagar, Andheri (W), Mumbai : 400058, India

(Autonomous College of Affiliated to University of Mumbai)

## **End Semester Examination**

December 2022

Maxi Marks: 100

Class: BEIT/BE COMP

Course code: CS414

Duration: 3 hours Semester: VII

Branch: IT/Computer

Name of the course : Big Data Analytics and Information retrieval

Q No		Max Marks	со	BL
1	Observe the following statistics about BookMyShow and justify why you think it is an example of Big Data. Ensure your answer is based on the details mentioned in the case study.	10M	CO1	4
	Quick Facts			
×	and Committee of the co			
	30 Million+ 5 650+ 5000+			
	CUSTOMERS COUNTRIES TOWNS AND CITIES SCREENS			
	50 15 & 2 Billion+ &			
	Million+  APP DOWNLOADS		9	
2	Consider a case study mentioned in question no 1. You are appointed as a data analyst in BookMyShow and assigned some task of data analyst.		CO1	5
	List down the important columns/features which you think are important in the dataset. Describe your dataset in brief.	03M		
	Write 6 questions about this dataset which will give you useful insights.	03M		
	What type of data visualization you will use to support these insights. Draw diagrams.	04M		
	Consider the FIFA 20 dataset. Part of the dataset is shown below. Write Haddop map reduce code in java to answer following queries.	:**	CO2	3

	Player Name	nationality	club	player_positions	Preferred Foot	International reputation			
	L. Messi	Argentina	FC Barcelona	CF	Left	5			- [
	Cristiano Ronaldo	Portugal	Real Madrid	LW, LM	Right	4		- 1	-
	A. Robben	Netherlands	FC Bayern München	RM, LM, RW	Left	3			
	Z. Ibrahimović	Sweden	Paris Saint-Germain	ST	Right	4		-	
	M. Neuer	Germany	FC Bayern München	annem (	Right	4			
	L. Suárez	Uruguay	FC Barcelona		Right	3			
3a	count  For each club, fi	t will have	al number of plays 3 columns Club N OR al number of Left 3 columns Club N	lame, Country r	name and	plavers .	6M		
3b	For each club, fir Expected output player count	nd the nun will have	nber of players for 3 columns Club N	r every player p ame, player_po	osition. esition and	d Total	6M		
3c	Find the average Expected output	internation will have :	nal reputation of 2 columns Club na	every club. ame, avg. intern	national re	eputation.	8M		
Q4	matrix M is as giver represents inlink.	2 1 3 2 3 3 3 20	s in each cell of th	presents outlinl e matrix repres	k and row		10M	CO5	****
	(b) Draw a web graph with spider trap and a dead node. Using this graph						10M	CO5	3
(5	Derive the equation for bloom filter to find Optimum number of hash functions required for a given size of data and size of bloom filter. Write your observation about this equation.					10M	CO4	3	
6	Draw and explain Spark Architecture  OR  Write short note on HBase			10M	соз	2			
7	Write a short note	on YARN		31			10M	соз	2
8	Discuss Data-Strea	m-Manag	ement System. W	ith example, ex	plain the	FM	10M	CO4	2

## For official use only (not for students)

CO Number	CO statement	Marks allotted
CO1	Identify challenges in big data management and inadequacy of existing technology to analyze big data problems.	20
CO2	Apply scalable algorithms based on Hadoop and Map Reduce to perform Big Data Analytics.	20
CO3	Use big data tools for information retrieval.	20
CO4	Use stream data model to provide real time analysis of big data.	20
CO5	Analyze data using graphs.	20

PI chart for percentage CO marks