Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (W), Mumbai : 400058, India

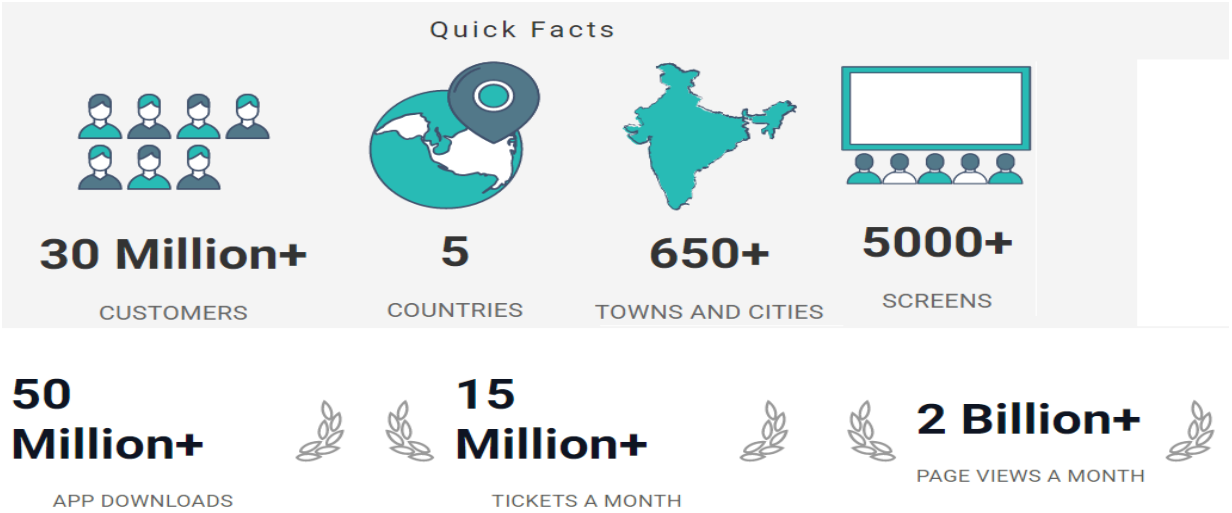
(Autonomous College of Affiliated to University of Mumbai)

End Semester Examination

December 2022

Maxi Marks : 100 | Duration : 3 hours
Class : BEIT/BE COMP | Semester : VII
Course code: CS414 | Branch : IT/Computer
Name of the course : Big Data Analytics and Information retrieval

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.



Volume - 1.5M
Velocity - 1.5M
Variety - 1.5M
Value - 1.5M
Veracity - 1.5M
Visualization - 1.5M
Presentation/flow - 1M

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.

List down the important columns/features which you think are important in the dataset. Describe your dataset in brief.
Dataset Description - ***1M***
Columns/features - ***2M***
Write 6 questions about this dataset which will give you useful insights. ***1/2M for each question***
What type of data visualization you will use to support these insights. Draw diagrams. ***3 Visualization dig - 1M each***
Presentation/flow - 1M

Consider the FIFA 20 dataset. Part of the dataset is shown below. Write Haddop map reduce code in java to answer following queries.

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
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	GK	Right	4
L. Suárez	Uruguay	FC Barcelona	ST, CF	Right	3

For each club, find the total number of players nation-wise.
Expected output will have 3 columns Club Name, Country name and Total player count

OR

For each club, find the total number of Left and Right foot preferred players .
Expected output will have 3 columns Club Name, Preferred Foot and Total player count
Mapper code - 3M, Reducer code - 3M

For each club, find the number of players for every player position.
Expected output will have 3 columns Club Name, player_position and Total player count
Mapper code - 3M, Reducer code - 3M

Find the average international reputation of every club.
Expected output will have 2 columns Club name, avg. international reputation.
Mapper code - 4M, Reducer code - 4M

(a) You decide to use the page rank algorithm r = Mr, on some web graph whose matrix M is as given below where column represents outlink and row represents inlink.. Numbers in each cell of the matrix represent edge weight.

$$\begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 20 \end{pmatrix}$$

Compute rank vector r, using Eigen vector formulation.

All correct eigen values - 4M
Eigen vector corresponding to eigen value = 1 - 4M
Justify why this eigen vector is final page rank vector - 2M

(b) Draw a web graph with spider trap and a dead node. Using this graph describe the technique that can be used to find page rank of web pages in a web graph that you have drawn.

web graph with spider trap and a dead node - 1M
what is spider trap and a dead node - 2M
Explain what is teleportation - 3 M
Apply teleportation on the graph drawn and show how it improves the page rank - 4M

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.
Derivation - 8M
Observation - 2M

HBase
Limitations of Hadoop - 2M

Properties of HBase - 2M
Advantages of HBase - 3M
Example with one query - 3M

Write a short note on YARN
Running an Application through YARN - 2M
Diagram - 2M
Scheduling in YARN with diagram - (3 schedules) - 3M
YARN components - 2M
Presentation/flow - 1M

Discuss Data-Stream-Management System. With example, explain the FM algorithm.

What is Data-Stream-2M
Why Data-Stream Management System is required - 2M
Diagram with its block explanation - 2M
Working of FM algorithm - 2M
Example of FM algorithm - 2M