Question Paper

Exam Date & Time: 31-May-2023 (10:00 AM - 01:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
Second Semester Master of Engineering - ME (Cloud Computing) Degree Examination - May 2023

Data Streaming and Visualization [CDC 5203]

Marks: 100 **Duration: 180 mins.**

Wednesday, May 29, 2023

Answer all the questions.

1) Discuss the factors leading to Big Data. Briefly explain the major source of Big Data. (CDC 5203.1 L2) (TLO 1.1) (5 + 5 marks) (10)

2) Discuss the problems faced by traditional database systems. How batch processing is different from realtime or stream process? (CDC 5203.1 L2) (TLO 1.2) (5 + 5 marks) (10)

3) Discuss the architecture of Hadoop distributed file system. Give the role and responsibilities of Name Node and Data Node in HDFS. (CDC 5203.2 L2) (TLO 2.1) (6 + 4 marks) (10)

4) Differentiate between Spark and Hadoop MapReduce. (CDC 5203.2 L2) (TLO 2.2) (10 mark) (10)

5) Write pyspark application using RDDs to solve the following. Assume that bank.txt dataset is provided with fields as Bank ID', 'Account Number', 'Transaction Date', 'Transaction Type' (credit or debit), 'Transaction Amount'. Date format is dd-mm-yyyy. (10)
a. Count unique number of customers. (CDC 5203.2 L2) (TLO 2.2) (3+3+4)
b. Number of transactions for given Account Number
c. Number of credit transactions for given Account Number in a given year.

6) Write python application to scrape Indian cities population data from https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population. Data need to be scraped are "Rank", "City", "Population (2011)", "Population (2001)" and "Sate or Union territory". Class_ name for table to scrape data is "wikitable sortable". Scrapped data need to be stored in pandas dataframe and converted into csv file. (CDC 5203.3 L2) (TLO 3.1) (10 marks) (10)

7) Write a python application to scrape data from https://realpython.github.io/fake-jobs/. Data need to be scraped are "Job Title", "Company", "Location" and link for applying job. "Job Title" is found in tag **h2**, "Company" in tag **h3** and location in tag **p**. (CDC 5203.3 L2) (TLO 3.1) (10 marks) (10)

8) List out the reasons why visualization is required. Explain Informative visualization, Persuasive visualization and Visual Art. (CDC 5203.4) (TLO 3.2) (4 + 6 marks) (10)

9) With examples, explain the three ingredients of successful Visualizations. (CDC 5203.4 L2) (TLO 3.2) (10 marks) (10)

10) Develop python application to generate line plot in python. X axis represents date, Y axis represents price. Line chart should represent share prices of different companies over time. Plot for at-least 3 companies. Plot should have proper labels, title and legend. (CDC 5203.4 L4) (TLO 3.2) (10 marks) (10)

-----End-----