Nirma University

Institute of Technology

Supplementary Examination (SPE), March - 2023

B. Tech. in Computer Science and Engineering, Semester-VII

2CS702 Big Data Analytics

Roll / Exam No. | Supervisor's Initial with Date

Time: 3 Hours | Max Marks: 100

Instructions:
1. Attempt all questions
2. Figures to right indicate full marks
3. Assume necessary data.
4. Use section-wise separate answer book.
5. Draw neat sketches wherever necessary.

SECTION-I

Q:1 **Answer the following questions (4 X 4)** **[16]**

CO1,BL2

1 How evolution of technology, use of internet of things and social media usage play a role in big data generation? Write your justification to the point. [4]

2 What are the algorithm level requirements to choose platform for big data application? Discuss in detail. [4]

3 What is an impact of the data streamer in the HDFS data write pipeline? Provide a thorough explanation. [4]

4 Describe any two types of input format in map-reduce programming with its appropriate application. [4]

Q:2 **Answer the following questions (8 X 2)** **[16]**

CO2,BL4

1 *Consider NCDC Weather data. Raw data provided by this agency is in form of log file as follows:* [8]

0100243519702349999N89+010+9999N89
0100243519712349999A71-008+9999A71
0100243519722349999A81+020+9999A89
....
0100243520222349999B11+022+9999B11

Underlined and highlighted data shows the year and temperature. Write down the logic of **mapper and reducer** to find out the maximum of temperature in each year. Also highlight the input and output datatype of key, value pair in the pseudocode.

OR

1 *Consider NCDC Weather data. Raw data provided by this agency is in form of log file as follows:* [8]

0100243519702349999N89+010+9999N89
0100243519712349999A71-008+9999A71
0100243519722349999A81+020+9999A89
....
0100243520222349999B11+022+9999B11

Underlined and highlighted data shows the year and temperature. Write down the logic of **combiner and partitioner** to find out the minimum of temperature in each year. Also highlight the input and output datatype of key, value pair in the pseudocode.

2 Consider employee id, employee name, engineering branch, phone, city and salary for employee dataset. Write following queries in **Cassandra**/MongoDB. [8]

1) Create table and insert five entries into it.
2) Update phone number of employees to 9999999999 for employee belong to Ahmedabad city.

3) Display only employee name from the table having salary more than 10K and belong to Ahmedabad city
4) Remove employee entries from dataset where phone of computer engineering branch.

OR

2 Consider student id, student name, engineering branch, phone, city and placement package for student dataset. Write following queries in **Cassandra/** MongoDB. [8]

1) Create table and insert five entries into it.
2) List out student branch details whose placement package in more than 5 LPA.
3) Display sorting order of student details based on city.
4) Update student branch from CE to CSE if student belongs to CE branch and student package range from 10 LPA to 15 LPA

Q:3 **Answer the following questions (6 X 3)** **[18]**

CO3,BL4

1 Explain the role and responsibilities of following XML files: [6]

1. HDFS-site.xml
2. Core-site.xml
3. Yarn-site.xml

2 Consider "Volume" of big data to answer following question. Fill up all the entries of table. [6]

Class	Size	Manage With: Platform	How it Fits: Memory requirement	Example
Small				
Medium				
Big				

3 Define Fault Tolerance. At what extent fault tolerance is supported in horizontal scaling platforms and vertical scaling platforms? Also justify your answer. [6]

SECTION-II

Q:4 **Answer the following questions [4 X 4]** **[16]**

CO1,BL2

1 Define "Veracity" and "Value" of big data. Elaborate and explain the terms by taking suitable example of Ecommerce application. [4]

2 You are at university library. You see few students browsing through the library catalog on kiosk. You see the working of librarians and other staff to issue/return books, magazines, and journals. Few students are using the e-library service, too. [4]

Which type of data is generated in this scenario? Support your answer by considering big data. Each category needs a minimum of five different types of data.

3 Describe the purpose of following HDFS command with Syntax. [4]

a) copyToLocal
b) rm
c) cp
d) chmod

4 Draw and explain Hadoop Eco-system. [4]

Q:5 **Answer the following questions [8 X 2]** **[16]**

CO4,BL3

1 Explain YARN architecture in detail with working of all the key components. [8]

OR

1 Explain CAP theorem and describe the features, applications, pros and cons of databases following CAP theorem. [8]

2 Discuss any one application in detail which present the limitation of MapReduce programming. How Apache Spark overcome this limitation? Explain the features and advantages of using Spark. [8]

Q:6 **Answer the following questions [6 X 3]** **[18]**

CO3,BL5

1 Describe the detail functionality of following [Draw Diagram if applicable] [6]
   a) Name Node
   b) Data Node
   c) Secondary Name Node

2 What is the need to use Partitioner in Map reduce programming? How does it work? Which scenarios present the poor partitioning in the program? [6]

3 Describe the applications for which Apache HIVE is most appropriate. Also describe the merits and demerits of each. [6]