Qn. Set Code-1

Semester: 5th
Programme: B.Tech
Branch: CSE, IT

AUTUMN END SEMESTER EXAMINATION-2023

5th Semester B.Tech (Deptt. Elective-II)

BIG DATA

CS 3032

(For 2022 (L.E), 2021 & Previous Admitted Batches)

Time: 3 Hours Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [1 × 10]

(a) Identify the unstructured, semi-structured, and structured data when creating big data applications for sports analytics.

(b) What are the benefits of using a hypervisor for an organization?

(c) Calculate the probability of false positives in Bloom filter of size 15, and 5 items are to be inserted.

(d) Consider a file of size 500 MB to be written into HDFS version 1.0. What is the minimum number of blocks required to complete the write operation?

(e) Draw the data model pyramid by establishing the relationship between audiences and purposes.

(f) Consider the dataset wherein 25 individuals are 20 years old, 40 are 50 years old, 30 are 60 years old, and 35 are 65 years old. Draw any data visualization depicting a univariate analysis.

(g) Why would developers design a big data application as a distributed and multi-threaded system?

(h) E-commerce uses session data to add items to a digital basket, a search engine to search for products, a recommendation engine to recommend customers who purchased this item also like it, and a payment platform. Identify two polyglot persistence technologies to facilitate its business operation.

(i) Draw the Euler diagram of the datasets: X = {1, 2, 5, 8}, Y = {1, 6, 9}, and Z = {4, 7, 8, 9}.

(j) Consider a Bloom filter of size 15 with three hash functions. During the insertion of stream elements, what is the probability that a slot will be hashed?

SECTION-B

2. (a) The central government needs to analyze the Aadhaar card dataset against different queries, for example, the total number of Aadhaar cards approved by state, rejected by state, the total number of Aadhaar cards by gender, and the total number of Aadhaar card applicants by age type. Explain how the big data architecture can be leveraged for different types of big data analytics. [4]

(b) A new company in the travel domain wants to start their business efficiently, i.e. high profit for low total cost of ownership. They want to analyse & find the most frequent & popular tourism destinations for their business. You have been tasked to analyse top tourism destinations that people frequently travel and top locations from where most of the tourism trips start. They also want you to analyze and find the destinations with costly tourism packages by leveraging big data. Identify the tasks/activities to be performed in each stage of the big data analytics life cycle. [4]

3. (a) An empty bloom filter of size 25 with the following hash functions: [4]

h1(x) = (5x+ 9) mod 6 mod 25

h2(x) = (7x+ 3) mod 2 mod 25

- Illustrate step-by-step insertion with the items: "John", and 67.
- Illustrate step-by-step membership test with "John".
- Illustrate step-by-step update of "John" with 679-811-629.

(b) Explain each step of the Flajolet-Martin probabilistic algorithm. Consider the data stream elements 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1, and the hash function (6x+1) mod 5. Find the number of distinct elements. [4]

SECTION-C

4. (a) Draw the MapReduce process to find the minimum temperature recorded for each year for the following data stored in tab-separated format in sample1.txt, sample2.txt and sample3.txt files. [4]

sample1.txt:

```
2023  23  28  29  43  24  25  26  22  26  20  25  26
2022  26  27  28  29  28  30  31  30  31  30  31  30
```

sample2.txt:

```
2021  31  32  30  32  33  34  35  36  38  34  33  34
2020  39  38  39  32  39  41  42  43  40  39  38  32
```

Sample3.txt:

```
2019  38  39  33  39  31  41  40  41  40  46  39  32
2018  28  29  23  29  21  31  30  31  30  31  29  22
```

(b) What is HIVE? Write a HIVE command for [4]

- Creating a database named Education
- Create a table Employee with the attributes: ID as integer, name as string, dept as string, and salary as float.
- Retrieve the rows from Employee who are working in IT department and drawing more than 30K.

5. (a) Explain the following terms by analyzing them in the context of big data with utmost importance on value (i.e., one of the Vs of big data in addition to volume, velocity, variety, and veracity): [4]

   - Traditional analytics architecture
   - Modern in-database analytics architecture
   - MPP database analytics architecture
   - In-memory computing

   (b) What is the need for virtualization in big data? Explain the different classes of hypervisors used in the virtualization. [4]

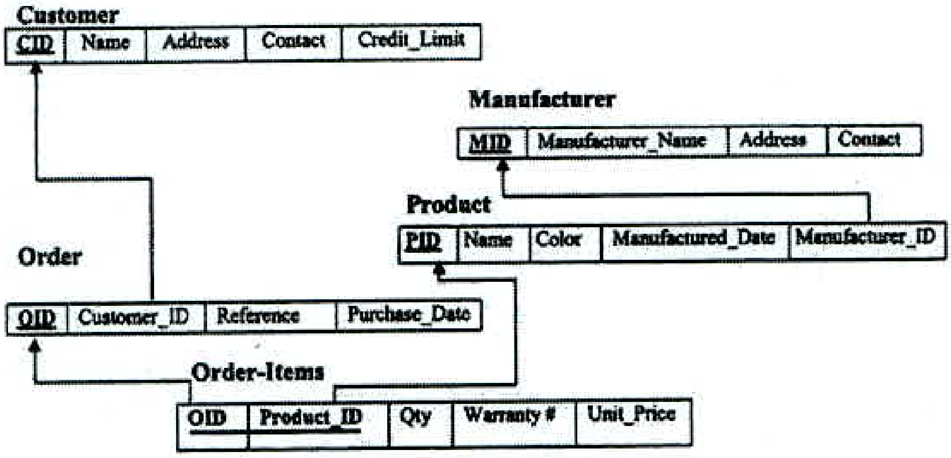
6. (a) List any three differences between a column-oriented and a row-oriented database. For the below diagram, draw its equivalent column-oriented and row-oriented database. [4]

ID	NAME	ADDRESS	AGE
1	Aarav	New Delhi	25
2	Arjun	Mumbai	29
3	Rishi	Chennai	29
4	Shankar	Cuttack	29
5	Nitin	Kolkata	32
6	Ishan	Ranchi	22

(b) What are the purposes of sharding? Given a dataset containing the details of the hostel boarder, by capturing the attributes such as name, roll, hostel name, and menu preference. Discuss the most suitable sharding strategy by choosing an attribute for fast querying and scalability. Subsequently, create each logical shard. [4]

SECTION-D

7. (a) Consider the ER model, with the entities and attributes as follows: [4]



Create the equivalent conceptual, logical and physical data model by mentioning facts and dimension tables.

(b) Consider the following multivariate data observed from 2018 to 2023. [4]

Year	Temp	Rain	Ice
2018	-3.9	23.65	18.5
2019	-4.5	25.75	13.5
2020	-5.5	20.75	16.5
2021	-6.5	22.85	14.5
2022	-4.8	24.65	17.5
2023	-5.3	25.45	12.5

- Draw a parallel coordinate plot where the minimum ice value is 13.
- Draw a chronological display of ice.
- Draw an ordinogram illustrating the relationship between rain and ice.
- Draw any data visualization depicting a univariate analysis by considering rain.

8. (a) Let A, B, and C represents a set of people who like oranges, pineapples, and strawberries, respectively. The number of people who likes oranges = 12 (O1 to O12), pineapples = 10 (P1 to P10), and strawberries = 15 (S1 to C15). Four people are such that they enjoy oranges, pineapples, and strawberries. Three of them like oranges and pineapples, and five people like pineapples and strawberries. Also, six people say that they like oranges and strawberries. Create the Venn diagram, illustrating: [4]

    - How many people like oranges only?
    - How many people like only one of the three?
    - How many people like all three?
    - How many people do not like any of the three?

   (b) Draw a cluster diagram for a big data environment by including a name node, a secondary name node, a standby name node, and data nodes arranged in 5 racks. Each rack has a different number of data nodes. [4]

*****