Qn. Set Code-3

Semester: 6th
Programme: B.Tech
Branch: CSSE

SPRING END SEMESTER EXAMINATION-2023

6th Semester B.Tech

BIG DATA

CS 3032

(For 2021 (L.E), 2020 & Previous Admitted Batches)

Time: 3 Hours | Full Marks: 50

Answer any SIX questions.
Question paper consists of three SECTIONS i.e. A, B and, C.
Section A is compulsory.
Attempt minimum ALL questions from SECTION-B and any TWO questions from SECTION-C.
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) List the uses of Big data management and data mining.

(b) Who coined the term Big data?

(c) Define the Maptask.

(d) Which country proposed the 'Social Media (Basic Expectations and Defamation) Bill 2021'?

(e) How should big data analytics programs be managed according to the best practices?

(f) What are probabilistic data structures? Explain with examples of their applications.

(g) Discuss about the task that used to be addressed by Clustering.

(h) What is the ideal number of hash functions needed for a Bloom filter with a size of 15 and 3 input elements?

(i) Name the programming tool useful to design Hadoop-based applications that can process massive amounts of data.

(j) Explain the role of ZooKeeper in HBase.

SECTION-B

2. (a) Explain the concept of MapReduce and provide an example of how it can be used to process large amounts of data. [4]

(b) Discuss the trade-offs between consistency and availability in distributed systems, and how these trade-offs are reflected in the design of MapReduce frameworks along with the real-world scenarios. [4]

OR

(a) How does Facebook's big data system for user enrolment tracking work? What technologies and platforms are involved in this system? [4]

(b) Illustrate the design choices that might impact the performance and efficiency of a MapReduce job for counting words. For example, how should the data be partitioned and sorted, and how many reducers should be used? [4]

3. (a) What is the function of YARN in Data Processing, and could you provide a well-organized diagram that shows the various components of YARN? [4]

(b) Enumerate some common use cases for big data and cloud computing, and how have organizations successfully leveraged these technologies to achieve their goals? Provide examples from a variety of industries, such as healthcare, finance, and retail. [4]

OR

(a) Describe the architecture of HIVE with a neat sketch and present the relative merits and demerits. [4]

(b) With the help of neat sketches explain the various components of a Pig Latin script, and provide an example of the script that demonstrates the use of these components. [4]

4. (a) What are the key differences between column-oriented and row-oriented databases, and how do these differences impact data visualization? [4]

(b) Develop an example for executing the data visualization that demonstrates the advantages of column-oriented and row-oriented databases. [4]

OR

(a) Explicate the concept of polyglot persistence, and how does it differ from traditional approaches to database management visualization through a single window? [4]

(b) Discuss the example of a use case where polyglot persistence is necessary, and discuss the challenges and benefits of implementing a polyglot persistence strategy in real-time application. [4]

SECTION-C

5. (a) Formulate the discussion on big data help companies to improve customer experience and engagement. [4]

(b) Explain the ingestion layer fit into a larger data architecture, such as a data lake or data warehouse with the help of various stages involved with the implementation. [4]

6. (a) Develop and explain the anatomy of File Read and File Writing in HDFS With Pictographic Representation. [4]

(b) Compare and contradict the various characteristics of Data warehouse, Hadoop, and Stream computing. [4]

7. (a) Evaluate the use of Bloom filters for optimizing database queries. Explain the key design considerations involved in implementing a Bloom filter. [4]

(b) Given a data stream with elements {10, 8, 7, 5, 3, 10, 9}, and a hash function $h(x) = (2x+3) \bmod 8$ of size 12, what is the count of distinct elements in the stream? [4]

*****