GITAM (Deemed to be University)
[19ECS767]
M.Tech. Degree Examination

CSE & DS
II Semester

BIG DATA ANALYTICS
(Effective from the admitted batch 2022–23)

Time: 3 Hours   **Max.Marks: 60**

Instructions: All parts of the unit must be answered in one place only.
Figures in the right hand margin indicate marks allotted.

Section-A

1. **Answer all the questions:**   **(10×2=20)**

    a) Discuss the Industry application of big data where it has been successfully utilized and analyze their impact.

    b) Discuss the role of mobile business intelligence in leveraging big data for decision-making on mobile devices.

    c) Explain the differences between master-slave replication and peer-to-peer replication in distributed databases.

    d) Explain the use of version stamps in NoSQL databases and how they help to maintain consistency.

    e) Explain compression techniques used in Hadoop for efficient storage and processing of large-scale data.

    f) Compare and contrast Hadoop streaming and Hadoop pipes as mechanisms for processing data in Hadoop.

    g) How can you generate test data for local tests in MapReduce, and why is it important?

    h) Describe the anatomy of a MapReduce job run, including the key components and their functions.

    i) Provide examples of real-world use cases where HBase has been applied, and describe the practical applications of HBase in those scenarios.

    j) Describe the different data types and file formats supported by Hive, and how they can be used to define and manipulate data in Hive tables.

Section-B

Answer the following: (5×8=40)

UNIT-I

2. Analyze the role of web analytics in big data-driven decision-making for businesses. Discuss the different types of web analytics data and their applications in different industries. Provide examples of how web analytics can be used to optimize marketing and customer engagement.

OR

3. Analyze the advantages and limitations of big data technologies, including Hadoop and open source technologies. Discuss the different types of big data technologies and their applications in different industries. Provide examples of how big data technologies can be used to optimize business operations.

UNIT-II

4. Evaluate the strengths and limitations of graph databases as a NoSQL database model, including their use cases, benefits, and challenges. Describe how graph databases can be used to model complex relationships between data entities and provide examples of real-world applications.

OR

5. Evaluate the advantages and disadvantages of schemaless databases in NoSQL databases. Explain how schemaless databases can improve database flexibility and scalability, and provide examples of real-world use cases where schemaless databases would be preferred over other data modeling techniques.

UNIT-III

6. Evaluate strategies for ensuring data integrity in Hadoop, including data validation, error handling, and fault tolerance mechanisms. Discuss best practices for handling data errors, data corruptions, and hardware failures in Hadoop clusters.

OR

7. Compare and contrast Hadoop streaming and Hadoop pipes as mechanisms for processing data in Hadoop. Analyze their differences in terms of data ingestion, processing, and output.

UNIT-IV

8. Analyze common types of failures in classic MapReduce and YARN, including node failures, network failures, and task failures. Discuss best practices for fault tolerance and recovery in each environment.

OR

9. Discuss the execution of MapReduce tasks, including the use of mappers and reducers, and the interaction between the JobTracker and TaskTrackers. Analyze best practices for optimizing task execution and performance in MapReduce jobs.

UNIT-V

10. Provide detailed examples of real-world use cases where HBase has been successfully implemented, and analyze the challenges and solutions encountered in those scenarios, including considerations such as data modeling, data ingestion, and data retrieval.

OR

11. Describe the data model of Cassandra and how it differs from traditional relational databases. Include the concepts of keyspace, column family, column, and row, and the principles of distribution and replication in Cassandra.

[II S/223]