Course Code	**19IT603**	**Regulations**	**2019**
Course Title	**BIG DATA ANALYTICS**		
Semester(s) & Programme(s)	**VI. Semester & B.E. Information Technology**	**AY**	**2024-25**

Unit I	**INTRODUCTION TO BIG DATA**		
Q. No.	**PART A (2 Mark)**	**CO**	**BT Level**
1.	How would you classify social media data into structured, semi-structured, or unstructured?	CO1	RE
2.	Compare and contrast the characteristics of Big Data using real-world examples.	CO1	UN
3.	Point out the need for Big Data analytics in decision-making.	CO1	RE
4.	Summarize the significance of Big Data analytics in real-world applications.	CO1	RE
5.	Illustrate how the "Variety" characteristic of Big Data affects data storage.	CO1	RE
6.	List the different types of digital data with an example for each.	CO1	RE
7.	Differentiate between structured and unstructured data with examples.	CO1	UN
8.	What is the role of a data analyst in a Big Data environment?	CO1	RE
9.	What are the challenges in designing the Big data environment/ ecosystem?	CO1	RE
10.	Investigate the open source Apache Hadoop distribution based on a centralized architecture.	CO1	AN
Q. No.	**PART B (Either Four 14 Marks Questions or six 7 Marks Question)**	**CO**	**BT Level**
1.	A financial institution needs to design a fault-tolerant, scalable storage system for transaction logs (5TB/day) using HDFS.	CO1	UN
2.	A multinational company wants to migrate its on-premise Hadoop infrastructure to a cloud-based solution for better scalability and cost efficiency.	CO1	UN
3.	A smart city project generates 10TB/day of sensor data (traffic, weather, pollution) that requires real-time processing, long-term storage, and analytics. Design a Hadoop ecosystem architecture for this use case.	CO1	AP
4.	Classify the types of digital data (structured, semi-structured, unstructured) and explain their significance in Big Data environments. Provide examples for each type.	CO1	AN
5.	Explain Big Data analytics and its significance in today’s data-driven world.	CO1	UN
6.	What is Big Data, and what are its key characteristics (such as volume, velocity, variety, veracity, and value)? How do these characteristics distinguish Big Data from traditional data in terms of storage, processing, and analysis?.	CO1	UN
7.	Explain the block storage mechanism and how it is used to manage and store large files within an institution. Discuss the advantages of block storage for handling sizable data compared to other storage methods.	CO1	AN
8.	What are cloud-based Hadoop solutions, and how do they differ from traditional on-premises Hadoop deployments? Discuss the leading Hadoop distributions and integrated systems available in the market, highlighting their key features. Additionally, explain the benefits of using cloud-based Hadoop for Big Data processing.	CO1	AN

Unit II	**BIG DATA TECHNOLOGY LANDSCAPE**		
Q. No.	**PART A (2 Mark)**	**CO**	**BT Level**
1.	Why HDFS is preferred over RDBMS? Justify.	CO2	RE
2.	Compare SQL and NOSQL.	CO2	UN
3.	Analyze why HDFS serves in distributed resource allocation. Also justify whether they overcome SPOF in HADOOP system	CO2	AN
4.	Analyze how big data be generated and how they are managed in the ecosystem projects for processing. Justify.	CO2	AN
5.	Differentiate between Hadoop and traditional RDBMS	CO2	UN
6.	What is the purpose of HDFS in Hadoop?	CO2	RE
7.	Distinguish between features of key Hadoop distributions (Cloudera vs Hortonworks).	CO2	UN
8.	Illustrate how a NoSQL database can handle unstructured data more efficiently than SQL.	CO2	AP
9.	Point out the importance of cloud-based Hadoop solutions.	CO2	RE
10.	What is the role of MapReduce in Hadoop?	CO2	RE
Q. No.	**PART B (Either Four 14 Marks Questions or six 7 Marks Question)**	**CO**	**BT Level**
1.	A data engineering team is setting up Apache Hive for analyzing large-scale e-commerce data stored in HDFS.	CO2	AN
2.	A healthcare application currently uses a relational database (SQL) to store patient records	CO2	UN
3.	A retail e-commerce company is facing challenges in managing rapidly growing volumes of unstructured customer data, including product reviews, images, and activity logs, using traditional relational databases. To address these limitations, evaluate the suitability of NoSQL databases for this scenario. Tasks: a) Identify the most appropriate type of NoSQL database for handling unstructured data such as reviews, images, and logs. b) Explain why a NoSQL approach is more appropriate than traditional SQL-based solutions in this context. c) Discuss the key advantages of NoSQL databases in managing unstructured data. d) Compare NoSQL and SQL databases in terms of scalability and flexibility, especially for modern e-commerce applications.	CO2	AP
4.	How do the core features of Hadoop make it well-suited for Big Data processing? Illustrate your explanation with practical scenarios. Additionally, explain how Hadoop ensures fault tolerance and scalability in distributed computing environments.	CO2	AN
	How do the NameNode and DataNodes in Hadoop Distributed File System (HDFS) manage and process high-volume data streams? Provide a detailed explanation accompanied by neat diagrams illustrating their roles and interactions.		
5.	Analyze the need for cloud-based Hadoop solutions. What advantages do they offer over traditional deployments?	CO2	UN

Unit III	HADOOP		
Q. No.	**PART A (2 Mark)**	**CO**	**BT Level**
1.	Distinguish between Alteryx tool and Normal Excel processing in data cleaning.	CO3	UN
2.	List out the various challenges in distributed computing.	CO3	RE
3.	Distinguish between Cloudera and Hortonworks in Hadoop Ecosystem maintenance.	CO3	UN
4.	Investigate the open source Apache Hadoop distribution based on a centralized architecture.	CO3	UN
5.	Point out the roles of the ResourceManager and NodeManager in YARN.	CO3	RE
6.	Depict the concept of data locality in Hadoop.	CO3	RE
7.	Recall the role of the NameNode and DataNode in HDFS.	CO3	RE
8.	Compare MapReduce with traditional batch processing.	CO3	UN
9.	List the differences between RDBMS and Hadoop	CO3	RE
10.	Name the four major components of the Hadoop ecosystem.	CO3	RE
Q. No.	**PART B (Either Four 14 Marks Questions or six 7 Marks Question)**	**CO**	**BT Level**
1.	How does Hadoop YARN manage multiple applications running simultaneously in a Big Data environment? Explain its architecture and components, and demonstrate with an example how YARN allocates resources and schedules tasks to ensure efficient application management.	CO3	UN
2.	How can businesses effectively handle unstructured data? Propose a comprehensive framework outlining each step of the process, and illustrate each step with relevant examples from a real-world business scenario	CO3	UN
3.	Explain the challenges faced in Big Data Storage and Processing.	CO3	AP
4.	How has Big Data evolved from traditional data processing systems to modern distributed frameworks? Explain the limitations of earlier systems and describe how new technologies and architectures address the challenges of volume, velocity, and variety in data.	CO3	UN
5.	Explain the process of resource allocation and job scheduling.		

Unit IV	NOSQL: MONGODB & CASSANDRA		
Q. No.	**PART A (2 Mark)**	**CO**	**BT Level**
1.	Investigate the open source visual data analysis techniques and tools	CO4	AN
2.	Analyze why ETL process is most important in Data cleaning.	CO4	AN
3.	How to run proxy & Running map reduce job.	CO4	RE
4.	What are the Important Hadoop daemon properties	CO4	RE
5.	Summarize the features that make Cassandra suitable for distributed applications.	CO4	RE
6.	List any four CQL data types in Cassandra.	CO4	RE
7.	What does TTL stand for in Cassandra?	CO4	RE
8.	How does Cassandra handle large-scale data distribution?	CO4	RE
9.	Distinguish between how MongoDB and Cassandra handle data consistency.	CO4	UN
10.	List out the use of TTL in Cassandra to manage session data.	CO4	RE

Q. No.	PART B (Either Four 14 Marks Questions or six 7 Marks Question)	CO	BT Level
1.	A weather analytics startup wants to migrate from a relational database to Apache Cassandra for handling high-velocity sensor data from multiple locations.	CO4	AP
2.	An online retail company wants to shift from an RDBMS to a Hadoop-based system. Explain how you would transition their large structured databases to a distributed storage and processing model.	CO4	UN
3.	How can NoSQL databases be classified into distinct types such as Document, Key-Value, Column-Family, and Graph databases? Provide real-world examples of each type, and explain specific scenarios or use cases where organizations would prefer NoSQL databases over traditional relational (SQL) databases.	CO4	AP
4.	How can Hive be used to manage and query employee data? Specifically, demonstrate how to: 1. Create a Hive table to store employee information (including fields for id, name, department, and salary) 2. Insert sample employee records into the table 3. Write a HiveQL query to retrieve all employees earning a salary greater than 50,000	CO4	UN
5.	How can MongoDB's data types—such as Date and ObjectId—be utilized to redesign a patient records schema? Compare these data types with those used in traditional relational databases, and explain how MongoDB's flexible schema design influences the selection and use of data types..	CO4	UN
6	How can TTL (Time To Live) be implemented in a database to automatically expire outdated data and help maintain optimal storage usage? Demonstrate its use with a practical example.	CO4	AP
7	Design a Cassandra Keyspace and table structure to store this data efficiently.	CO4	UN
8	How would you design a MongoDB schema for an e-commerce platform that includes collections for products, customers, and orders? Provide MongoDB queries to: a) Update the price of all products within a specific category. b) Retrieve orders placed by a particular customer in the last 30 days. c) Delete users marked as inactive.	CO4	UN
9	What are the various data types supported by MongoDB? Discuss the basic types such as String, Integer, Boolean, Double, and Date, as well as more complex types like ObjectId, Array, and Embedded Documents. Explain how these data types enable flexible and efficient data modeling in MongoDB.	CO4	AP
10	How would you design a Cassandra table schema to efficiently handle user activity data for a social media platform? The data includes user IDs, timestamps, and action types (e.g., login, post, like). Explain your design choices considering Cassandra's data modeling principles for scalability and query efficiency.	CO4	UN
11	What are the various data types supported by Cassandra Query Language (CQL)? Enumerate and briefly describe them. Additionally, explain the roles of keyspaces, tables, and partitions in Cassandra, providing examples to illustrate their functions. Finally, design a Cassandra table schema suitable for a weather monitoring system.	CO4	UN
12	Write a sample CQL query to add new data and update user activity.	CO4	UN

Unit V	HIVE, PIG, REPORTING TOOLS		
Q. No.	**PART A (2 Mark)**	**CO**	**BT Level**
1.	State Zookeeper and its features with applications	CO5	RE
2.	Distinguish between Pig and Hive	CO5	UN
3.	How would you implement RCFILE in Hive?	CO5	RE
4.	Differentiate between Cassandra and Hadoop	CO5	UN
5.	Compare Hive's architecture with traditional SQL engines in terms of query flow.	CO5	UN

6.	List any three data types used in Hive.	CO5	RE
7.	Write a simple Hive query to retrieve employee names earning above a certain salary.	CO5	AP
8.	Name two file formats supported by Hive.	CO5	RE
9.	Use a simple HiveQL query to retrieve all customer names from a table named customers where the city is ‘Chennai’.	CO5	RE
10.	Point out the role of PiggyBank in Pig and how it enhances Pig’s capabilities.	CO5	RE
Q. No.	**PART B (Either Four 14 Marks Questions or six 7 Marks Question)**	**CO**	**BT Level**
1.	How can a retail analytics company integrate MongoDB (a NoSQL database) with JasperReports to generate dynamic business reports using data stored in NoSQL data structures? .	CO5	AP
2.	How can we integrate a data visualization tool like Tableau into the Decision Support System (DSS) for the City Council dashboard, in order to enhance data-driven decision-making based on the scenario described above?	CO5	AP
3.	How can you design a Hive table schema for log analytics that incorporates both partitioning and bucketing? Provide example HQL queries to perform aggregation on the data. Additionally, explain the Hive architecture with a clear and concise diagram.	CO5	UN
4.	Write a command for the following in Hive Query Language: i)changing directory to HIVE_HOME ii) creating a database iii) creating a table iv) loading data in a table v) exiting the Hive CLI	CO5	AP
5.	Design a workflow hive architecture to efficiently processes structured and semi-structured data.	CO5	UN
6	How does JasperReports handle NoSQL data structures, particularly when working with databases like MongoDB? List the key features of JasperSoft Studio, and enumerate the steps required to connect JasperReports to a MongoDB database for generating reports.	CO5	UN
7	Select appropriate Hive data types and file formats supported. Explain the purpose of SERDE in Hive with an example code.	CO5	UN