20IT701/PE3B

IV/IV B.Tech (Regular/Supplementary) DEGREE EXAMINATION

December, 2024 **IT**

Seventh Semester **Big Data Analytics**

Time: Three Hours **Maximum:**70 Marks

Answer question 1 compulsorily. **(14X1 = 14Marks)**

Answer one question from each unit. **(4X14=56Marks)**

			CO	BL	M
1	a)	List any three characteristics of Big Data.	CO1	L1	1
	b)	Mention one real-world example where Big Data Analytics is applied.	CO1	L1	1
	c)	What does HDFS stand for?	CO1	L1	1
	d)	Define the term "DataNode" in HDFS.	CO1	L1	1
	e)	What does YARN stand for?	CO2	L1	1
	f)	What are the input splits in MapReduce?	CO2	L1	1
	g)	What is scheduling in YARN?	CO2	L1	1
	h)	What happens during the **Shuffle** phase in MapReduce?	CO2	L1	1
	i)	What is Apache Spark?	CO3	L1	1
	j)	What does DAG stand for in Apache Spark?	CO3	L1	1
	k)	Define the term "NoSQL database."	CO3	L1	1
	l)	Define the term **CRUD operations** in MongoDB.	CO4	L1	1
	m)	What is transformation in spark	CO4	L1	1
	n)	What is an spark action	CO4	L1	1
		Unit-I			
2	a)	Describe the three V's of Big Data: Volume, Variety, and Velocity.	CO1	L2	7
	b)	Identify a Big Data application and explain how it solves a real-world problem.	CO1	L3	7
		(OR)			
3	a)	Differentiate between structured, semi-structured, and unstructured data as sources of Big Data.	CO1	L2	7
	b)	Develop a list of data sources for analyzing consumer behavior in e-commerce.	CO1	L3	7
		Unit-II			
4	a)	Explain the architecture of YARN and its key components.	CO2	L2	7
	b)	Compare the performance of an application running on YARN versus MapReduce 1	CO2	L3	7
		(OR)			
5	a)	What is the role of the **Map** and **Reduce** functions in MapReduce?	CO2	L2	7
	b)	Write a MapReduce job to calculate the word count of a given text dataset.	CO2	L3	7
		Unit-III			
6	a)	Explain what an RDD (Resilient Distributed Dataset) is in Apache Spark.	CO3	L2	7
	b)	Write a program to perform basic CRUD operations in MongoDB on a sample dataset.	CO3	L3	7
		(OR)			
7	a)	Explain the importance of Spark SQL for structured data processing.	CO3	L2	7
	b)	Demonstrate how to create and manipulate RDDs in Apache Spark for a simple dataset.	CO3	L3	7
		Unit-IV			
8	a)	What is data ingestion, and why is it important in Big Data processing?	CO4	L2	7
	b)	Implement an SQL-like query using Spark SQL to analyze structured data in a CSV file.	CO4	L3	7
		(OR)			
9	a)	Compare Flume and Kafka as data ingestion tools.	CO4	L2	7
	b)	Compare the performance of RDD-based and DataFrame-based approaches for analyzing large datasets.	CO4	L3	7

Scheme of Valuation

Q 1	Criteria for Evaluation	Marks
a)	- Any three correct characteristics of Big Data.	1
b)	- A relevant and real-world example where Big Data is applied.	1
c)	- Correct expansion of HDFS (Hadoop Distributed File System).	1
d)	- Clear and concise definition of "DataNode" in HDFS.	1
e)	- Correct expansion of YARN (Yet Another Resource Negotiator).	1
f)	- Explanation of input splits in MapReduce.	1
g)	- Definition or explanation of scheduling in YARN.	1
h)	- Accurate description of what happens during the Shuffle phase in MapReduce.	1
i)	- Clear definition of Apache Spark.	1
j)	- Correct expansion of DAG (Directed Acyclic Graph) in Apache Spark.	1
k)	- Definition of NoSQL database with relevant context.	1
l)	- Explanation of CRUD operations (Create, Read, Update, Delete) in MongoDB.	1
m)	- Definition of a transformation in Spark.	1
n)	- Explanation of an action in Spark.	1

Question	Criteria for Evaluation	Marks
2 a) Describe the three V's of Big Data: Volume, Variety, and Velocity.	- Clear explanation of **Volume**, **Variety**, and **Velocity** with examples. - Relevance and clarity in describing their significance in Big Data.	3 4
2 b) Identify a Big Data application and explain how it solves a real-world problem.	- Identification of a relevant Big Data application. - Explanation of how the application solves a real-world problem with examples or case studies.	2 5
3 a) Differentiate between structured, semi-structured, and unstructured data as sources of Big Data.	- Clear definitions of structured, semi-structured, and unstructured data. - Distinction with relevant examples for each category.	3 4

Question	Criteria for Evaluation	Marks
3 b) Develop a list of data sources for analyzing consumer behavior in e-commerce.	- Creation of a comprehensive list of data sources (e.g., transaction data, click stream data, social media, reviews, etc.). - Explanation of how these sources help analyze consumer behavior.	3 4
4 a) Explain the architecture of YARN and its key components.	- Description of YARN architecture, including ResourceManager, NodeManager, and ApplicationMaster. - Explanation of their roles and interaction in YARN.	4 3
4 b) Compare the performance of an application running on YARN versus MapReduce 1.	- Identification of key differences in architecture and resource management between YARN and MapReduce 1. - Comparison of performance improvements and scenarios where YARN is more efficient.	3 4
5 a) What is the role of the Map and Reduce functions in MapReduce?	- Explanation of the Map and Reduce functions with their purpose in distributed processing. - Examples or diagram to illustrate their roles in a typical workflow.	3 4
5 b) Write a MapReduce job to calculate the word count of a given text dataset.	- Outline of the program structure (Mapper, Reducer, and Driver code). - Functional explanation of the code logic for word count.	3 4
6 a) Explain what an RDD (Resilient Distributed Dataset) is in Apache Spark.	- Definition of RDD and its key features (e.g., fault tolerance, distributed computation). - Explanation of its significance in Spark and comparison with traditional data structures.	3 4

Question	Criteria for Evaluation	Marks
6 b) Write a program to perform basic CRUD operations in MongoDB on a sample dataset.	- Correct MongoDB syntax for Create, Read, Update, and Delete operations. - Explanation or demonstration using a relevant example dataset.	4 3
7 a) Explain the importance of Spark SQL for structured data processing.	- Description of Spark SQL and its capabilities for structured data. - Examples of use cases or benefits (e.g., integration with SQL queries, performance optimizations).	3 4
7 b) Demonstrate how to create and manipulate RDDs in Apache Spark for a simple dataset.	- Correct implementation of RDD creation and manipulation (e.g., map, filter, reduce). - Explanation of the operations applied with their significance.	4 3
8 a) What is data ingestion, and why is it important in Big Data processing?	- Definition of data ingestion and its role in Big Data workflows. - Examples of tools and challenges associated with data ingestion.	3 4
8 b) Implement an SQL-like query using Spark SQL to analyze structured data in a CSV file.	- Correct setup and loading of CSV data in Spark. - Implementation of SQL-like query with explanation of its purpose and results.	3 4
9 a) Compare Flume and Kafka as data ingestion tools.	- Overview of Flume and Kafka with key features. - Comparison of their use cases, strengths, and limitations.	3 4
9 b) Compare the performance of RDD-based and DataFrame-based approaches for analyzing large datasets.	- Explanation of RDD-based and DataFrame-based approaches. - Comparison in terms of performance, usability, and suitability for large datasets.	3 4

Detailed Scheme of Valuation

Q1 – 1 Mark

a) List any three characteristics of Big Data

Volume: Refers to the vast amount of data generated.

Variety: Refers to the diverse formats of data (structured, semi-structured, unstructured).

Velocity: Refers to the speed at which data is generated and processed.

b) Mention one real-world example where Big Data Analytics is applied.

Predictive maintenance in manufacturing industries

c) What does HDFS stand for?

Hadoop Distributed File System.

d) Define the term "DataNode" in HDFS.

A DataNode is a component in HDFS responsible for storing and retrieving data blocks as directed by the NameNode.

e) What does YARN stand for?

Yet Another Resource Negotiator.

f) What are the input splits in MapReduce?

Input splits are logical divisions of data used to distribute work among Map tasks in a MapReduce job.

g) What is scheduling in YARN?

Scheduling in YARN is the process of allocating resources to various applications based on policies and priorities.

h) What happens during the Shuffle phase in MapReduce?

During the Shuffle phase, intermediate outputs from the Map tasks are transferred to the Reducers for processing.

i) What is Apache Spark?

Apache Spark is an open-source distributed computing system designed for fast processing of large-scale data.

j) What does DAG stand for in Apache Spark?

Directed Acyclic Graph.

k) Define the term NoSQL database.

A NoSQL database is a non-relational database designed to handle a wide variety of data models, including key-value, document, column-family, and graph formats.

l) Define the term CRUD operations in MongoDB.

CRUD operations in MongoDB refer to Create, Read, Update, and Delete operations for managing data.

m) What is transformation in Spark?

A transformation in Spark is a function that produces a new RDD/DataFrame from an existing one, such as map or filter.

n) What is a Spark action?

A Spark action is an operation that triggers the execution of transformations and returns a result to the driver or writes to an external storage, such as collect or save.

2 a) Describe the three V's of Big Data: Volume, Variety, and Velocity. (CO1, L2, 7 Marks)

1. **Volume (2 Marks):**
   - Explain the concept of vast data sizes generated daily.
   - Mention examples like social media data, IoT devices, etc.
2. **Variety (2 Marks):**
   - Discuss the diversity in data formats (structured, semi-structured, unstructured).
   - Provide relevant examples like databases, XML files, and multimedia files.
3. **Velocity (2 Marks):**
   - Highlight the speed of data generation and processing.
   - Use examples such as stock market data or sensor data.
4. **Presentation (1 Mark):**
   - Clear explanation with relevant examples for all three characteristics.

2 b) Identify a Big Data application and explain how it solves a real-world problem. (CO1, L3, 7 Marks)

1. **Application Identification (2 Marks):**
   - Mention an application like fraud detection, predictive maintenance, or personalized marketing.
2. **Problem Description (2 Marks):**
   - Clearly describe the real-world problem addressed by the application.
3. **Solution Explanation (2 Marks):**
   - Detail how Big Data analytics solves the problem using techniques like pattern recognition, machine learning, or real-time processing.
4. **Clarity and Examples (1 Mark):**
   - Use appropriate examples to illustrate the solution.

3 a) Differentiate between structured, semi-structured, and unstructured data as sources of Big Data. (CO1, L2, 7 Marks)

1. **Structured Data (2 Marks):**
   - Define structured data and provide examples (e.g., relational databases).
2. **Semi-structured Data (2 Marks):**
   - Define semi-structured data and provide examples (e.g., XML, JSON).
3. **Unstructured Data (2 Marks):**
   - Define unstructured data and provide examples (e.g., images, videos).
4. **Tabular Representation (1 Mark):**
   - Bonus mark for presenting differences in a table format.

3 b) Develop a list of data sources for analyzing consumer behavior in e-commerce. (CO1, L3, 7 Marks)

1. **Identification of Data Sources (5 Marks):**
   - Transaction data, clickstream data, social media, customer feedback, IoT devices, etc.
   - Provide specific examples for each source.
2. **Clarity and Organization (2 Marks):**
   - Well-organized list with clear descriptions for each source.

4 a) Explain the architecture of YARN and its key components. (CO2, L2, 7 Marks)

1. **ResourceManager (2 Marks):**
   - Describe its role in allocating resources across the cluster.
2. **NodeManager (2 Marks):**
   - Explain its role in monitoring tasks and managing resources on nodes.
3. **ApplicationMaster (2 Marks):**
   - Highlight its role in managing application execution.
4. **Overall Clarity and Diagram (1 Mark):**
   - Bonus mark for a well-labeled diagram of the YARN architecture.

4 b) Compare the performance of an application running on YARN versus MapReduce 1. (CO2, L3, 7 Marks)

1. **Comparison Parameters (5 Marks):**
   - Discuss key differences in resource management, scalability, efficiency, and support for multiple frameworks.
2. **Clarity and Examples (2 Marks):**
   - Provide specific examples or scenarios where YARN outperforms MapReduce 1.

5 a) What is the role of the Map and Reduce functions in MapReduce? (CO2, L2, 7 Marks)

1. **Map Function (3 Marks):**
   - Define the purpose and functionality of the Map phase.
   - Provide an example of generating intermediate key-value pairs.
2. **Reduce Function (3 Marks):**
   - Define the purpose and functionality of the Reduce phase.
   - Explain how it aggregates key-value pairs.
3. **Clarity and Example (1 Mark):**
   - Bonus mark for including a clear example of a MapReduce task.

5 b) Write a MapReduce job to calculate the word count of a given text dataset. (CO2, L3, 7 Marks)

1. **Mapper Code (3 Marks):**
   - Evaluate correctness, clarity, and comments in the code.
2. **Reducer Code (3 Marks):**
   - Assess functionality, syntax, and logic for reducing word counts.
3. **Code Structure (1 Mark):**
   - Bonus mark for clear formatting and indentation.

6 a) Explain what an RDD (Resilient Distributed Dataset) is in Apache Spark. (CO3, L2, 7 Marks)

1. **Definition and Features (4 Marks):**
   - Fault tolerance, lazy evaluation, in-memory processing.
2. **Use Case Example (2 Marks):**
   - Provide an example to demonstrate RDD usage.
3. **Clarity (1 Mark):**
   - Well-organized explanation.

6 b) Write a program to perform basic CRUD operations in MongoDB on a sample dataset. (CO3, L3, 7 Marks)

1. **Create and Read Operations (3 Marks):**
   - Evaluate syntax and correctness for insertOne and find queries.
2. **Update and Delete Operations (3 Marks):**
   - Assess correctness for updateOne and deleteOne queries.
3. **Code Organization (1 Mark):**
   - Bonus mark for clear comments and well-structured code.

7 a) Explain the importance of Spark SQL for structured data processing. (CO3, L2, 7 Marks)

1. **Integration with SQL (3 Marks):**
   - Discuss ease of writing SQL queries for Big Data.
2. **Optimization (2 Marks):**
   - Mention the role of Catalyst Optimizer in query performance.
3. **Use Case Examples (2 Marks):**
   - Provide examples of structured data processing using Spark SQL.

7 b) Demonstrate how to create and manipulate RDDs in Apache Spark for a simple dataset. (CO3, L3, 7 Marks)

1. **RDD Creation (3 Marks):**
   - Assess correctness and clarity of the parallelize function.
2. **Transformation (2 Marks):**
   - Evaluate the use of transformation functions like map.
3. **Action (1 Mark):**
   - Verify correctness of the action function like collect.
4. **Code Organization (1 Mark):**
   - Bonus for clear structure and comments.

8 a) What is data ingestion, and why is it important in Big Data processing? (CO4, L2, 7 Marks)

1. **Definition (3 Marks):**
   - Explain the concept and role of data ingestion.
2. **Importance (3 Marks):**
   - Highlight the necessity of consistent data flow for analysis.
3. **Clarity (1 Mark):**
   - Well-structured explanation.

9 a) Compare Flume and Kafka as data ingestion tools. (CO4, L2, 7 Marks)

1. **Comparison Parameters (5 Marks):**
   - Discuss use cases, scalability, and persistence.
2. **Clarity and Examples (2 Marks):**
   - Provide relevant examples for each tool.

Faculty in Charge **Head of Department – IT**