U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

October 2024 Supplementary Examinations

Programme: B.E. **Semester: VI**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 24AM6PCBDA **Max Marks: 100**

Course: BIG DATA ANALYTICS

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

			CO	*PO*	Marks
		UNIT - I			
1	a)	Give a detailed analysis of how Walmart and Uber have effectively utilized Big Data Analytics to enhance operational efficiency and enhance user experience.	*CO1*	*PO2*	**8**
	b)	Provide the fundamental characteristics of Big Data that are essential for comprehending large-scale data environments.	*CO1*	*PO1*	**4**
	c)	Describe different types of Big Data Analytics that enhance decision-making in risk management within firms.	*CO2*	*PO3*	**8**
		UNIT - II			
2	a)	Highlight the performance characteristics that differentiate ORC, Avro, and Parquet file formats for real-time streaming data within a distributed data processing framework.	*CO2*	*PO2*	**6**
	b)	Explain two prominent file compression techniques used in modern computing. Provide real-world examples where each technique is suitable, emphasizing how they optimize storage and transmission efficiency.	*CO2*	*PO3*	**10**
	c)	Compare between lossless and lossy data compression techniques.	*CO2*	*PO1*	**4**
		UNIT - III			
3	a)	A company is planning to implement Hadoop for its large-scale data storage needs. Illustrate the role of Hadoop HDFS in managing big data by outlining the key modules of HDFS and their functionalities.	*CO3*	*PO2*	**10**
	b)	A startup is planning to build a data processing platform using Hadoop. Design a high-level architecture diagram incorporating HDFS, YARN, and MapReduce. Highlight the interactions between these components and explain how they collectively support scalable data processing.	*CO3*	*PO3*	**10**

<table>
<tr><td></td><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>4</td><td>a)</td><td>Outline the process of serialization and deserialization, highlighting the importance of these techniques to facilitate data storage, communication, and interoperability between different systems and languages.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Describe the Map-Reduce paradigm, detailing the essential stages of Mapper and Reducer. Provide a step-by-step account of how data flows through these phases that enables parallel data processing and aggregation in distributed computing environments.</td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - IV</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>5</td><td>a)</td><td>With a neat diagram, explain HIVE architecture and its services.</td><td><i>CO2</i></td><td><i>PO3</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Specify with an example the following HIVEQL commands-
i) Loading data into tables
ii) Sort by and order by clause
iii) Grouping and Aggregation
iv) Joining of 2 tables</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>05</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Consider the given tables:
<table>
<tr><th>Product_id</th><th>Product_name</th><th>Category_id</th></tr>
<tr><td>1</td><td>Laptop</td><td>1</td></tr>
<tr><td>2</td><td>Mouse</td><td>3</td></tr>
<tr><td>3</td><td>Keyboard</td><td>3</td></tr>
<tr><td>4</td><td>Monitor</td><td>1</td></tr>
<tr><td>5</td><td>Headphones</td><td>2</td></tr>
</table>
<table>
<tr><th>Order_id</th><th>Product_id</th><th>Quantity</th></tr>
<tr><td>1</td><td>1</td><td>5</td></tr>
<tr><td>2</td><td>2</td><td>10</td></tr>
<tr><td>3</td><td>3</td><td>8</td></tr>
<tr><td>4</td><td>1</td><td>3</td></tr>
<tr><td>5</td><td>4</td><td>6</td></tr>
</table>
Write a Hive query for the following:
i. Calculate the total quantity of each product ordered.
ii. Find the product that has the highest total quantity ordered.</td><td><i>CO3</i></td><td><i>PO4</i></td><td><b>05</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - V</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>6</td><td>a)</td><td>Illustrate how Apache PIG simplifies data processing tasks compared to traditional Map-Reduce programming in Apache Hadoop.</td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>10</b></td></tr>
</table>

			List the advantages of using PIG over directly coding in Map-Reduce for typical data analysis tasks.			
		b)	Depict with a diagram of how Zookeeper assists HBase in monitoring and coordinating cluster operations effectively.	*CO2*	*PO2*	**10**
			OR			
	7	a)	Illustrate the interaction of Spark components to process and analyze data efficiently with the help of diagrammatic representation. Describe how Spark manages tasks and data across the cluster to achieve high-performance data processing?	*CO3*	*PO5*	**10**
		b)	Outline the sequence of steps involved when a client application retrieves data from HBase, including how requests are processed by the region servers?	*CO3*	*PO5*	**10**

****