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

Autonomous Institute Affiliated to VTU

August 2024 Supplementary Examinations

Programme: B.E.   Semester: VI
Branch: Computer Science and Engineering   Duration: 3 hrs.
Course Code: 20CS6PEBDA   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.

<table>
<tr><td rowspan="20" style="writing-mode:vertical">**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.</td><td></td><td></td><td style="text-align:center">**UNIT - I**</td><td>**CO**</td><td>**PO**</td><td>**Marks**</td></tr>
<tr><td>1</td><td>a)</td><td>Discuss the techniques to find patterns in or interpret unstructured data.</td><td>CO1</td><td>PO1</td><td>**06**</td></tr>
<tr><td></td><td>b)</td><td>Differentiate between the following:
• Velocity and Variety
• SQL and NOSQL</td><td>CO2</td><td>PO2</td><td>**06**</td></tr>
<tr><td></td><td>c)</td><td>Demonstrate classification of analytics based of different schools of thought.</td><td>CO1</td><td>PO1</td><td>**08**</td></tr>
<tr><td></td><td></td><td style="text-align:center">**UNIT - II**</td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>a)</td><td>Cassandra works on the philosophy that it will always be available for writes. Justify</td><td>CO2</td><td>PO2</td><td>**06**</td></tr>
<tr><td></td><td>b)</td><td>i) Create a column family LIBRARY with the following fields :

Book_ID int,
Book_Name text,
Student_Name text,
Book_taken_count counter
 Demonstrate the usage of counters by adding necessary primary key into the table, insert the required values and display the student names who have taken a book more than once
 ii) Analyse which data model is better in terms of primary key selection among the following tables. Justify your answer.

Create table MusicPlaylist (SongId int, SongName text, Year int, Singer text, Primary key(SongId, SongName));

Create table MusicPlaylist   (SongId int, SongName text, Year int, Singer text, Primary key((SongId, Year), SongName)   );</td><td>CO2</td><td>PO2</td><td>**06**</td></tr>
</table>

		c)	Write Cassandra queries for the following.
			
			    i. Create a keyspace by name Employee and Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary, Dept_Name.
			    ii. Insert two values into the table in a batch
			   iii. Update Employee name and Department of Emp-Id 123
			   iv. Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.
			    v. Update the altered table to add project names
			   vi. Create a TTL of 15 seconds to display the values of Employees
				CO3	PO3	08

			UNIT - III			
3	a)		Customer relationship management (CRM) refers to the principles, practices, and guidelines that an organization follows when interacting with its customers. Collecting their information along the way and using that information to enhance their experience and foster long term relationships. Examine the benefits of CRM data analysis using Hadoop.	CO1	PO1	06
	b)		Demonstrate the importance of combiner in Hadoop with an example. Differentiate the same with Reducer.	CO2	PO2	06
	c)		Given below is the data regarding the electrical consumption of an organization.	CO3	PO3	08

	Jnn	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1991	22	38	21	20	31	33	25	35	34	32	21	22
1992	29	20	23	24	31	32	35	36	32	37	34	35
1993	32	28	32	34	21	22	24	37	28	23	25	33
1994	34	20	24	25	28	33	34	26	25	21	20	31
1995	37	30	38	33	36	32	23	25	23	27	34	24

Write the java mapper and reducer code to process the data and display the year with maximum average electrical consumption. Explain the significance of partitioner in map reduce process

			OR			
4	a)		Point out the two main daemons of map reduce programming model. Explain the steps carried out by these daemons during processing the data with Hadoop	CO1	PO1	06
	b)		Demonstrate the importance of HDFS 2 in Hadoop 2.0 over HDFS 1 introduced in Hadoop 1.0. with necessary diagram.	CO2	PO2	06
	c)		Given a set of Student records as shown below, each with a name and a set of test scores. Write a map reduce program to calculate the average test score for each student:	CO3	PO3	08

			Alice, 90, 85, 92 Bob, 75, 80, 80 Charlie, 90, 95, 85 David, 100, 100, 100			
			UNIT - IV			
5	a)		Consider a file log.txt which maintains errors and warnings. Create RDDs to count the number of lines that contained either error or warning. Show the corresponding lineage graph.	CO2	PO2	**06**
	b)		Write a Scala program to create a user define function to return largest number among two numbers	CO2	PO3	**06**
	c)		Point out the constraints applied while using fold() and reduce() functions in RDDs .Demonstrate how aggregate() operation is beneficial in overcoming the usage of this constraint with an example program	CO3	PO3	**08**
			OR			
6	a)		Demonstrate the three main capabilities provided by Spark SQL with a neat diagram.	CO1	PO1	**06**

	b)		Demonstrate how UDFs are handled in Spark SQL. Write Spark SQL queries for the following by making necessary imports and creating appropriate contexts.
			  i. Create a SchemaRDD by loading a JSON file which contains details of cars.
			  ii. Select car name, car_price in the increasing order of car_price for cars.
			  iii. Access the first column from the SchemaRDD created above
				CO2	PO2	**06**

	c)		Point out the difference between Map() and flatMap(). Consider two input RDD {Computer, Operations, Functions, API, DBMS} and {Web, Functions, RDBMS, Hadoop}. Write a Scala program to create above RDDs and perform set operation on the RDDs by calling necessary import and creating appropriate contexts.	CO3	PO3	**08**
			UNIT - V			
7	a)		Demonstrate how collaborative filtering could be applied in a recommendation system along with necessary calculations with an example.	CO2	PO2	**10**
	b)		Demonstrate the importance of user profiles and term frequency inverse document frequency in content based recommendation system.	CO3	PO2	**10**

******