U.S.N.										

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

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: VI**
Branch: Information Science and Engineering **Duration: 3 hrs.**
Course Code: 23IS6PEBDA / 22IS6PEBDA **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.

			UNIT - I	CO	PO	Marks
1	a)		Differentiate between Traditional Business Intelligence (Bi) Versus Big Data	CO1		5
	b)		Illustrate the challenges of Semi-Structured and Unstructured data with examples each	CO1		10
	c)		Mention the characteristics of data.	CO1		5
			OR			
2	a)		Heterogeneity and Security pose a challenge in distributed computing. Justify.	CO2	PO1	5
	b)		What are the challenges faced to store semi-structured data and how to extract the information from them?	CO1		10
	c)		What are the characteristics of Unstructured data?	CO1		5
			UNIT - II			
3	a)		Replication ensures the availability of the data in HDFS. Justify	CO2	PO1	5
	b)		Data storage nodes: Master node and Slave node plays an important role in HDFS. Justify.	CO1		10
	c)		Write a Mapper class, reducer class and the driver class with main function to find the number of occurrences of each word in a file in Hadoop framework.	CO3	PO2	5
			OR			
4	a)		Write Java program to **search** the word using Hadoop framework.	C03	PO2	10
	b)		Explain the tasks and responsibilities of Application manager and Application master.	CO1		5
	c)		Explain the goals of HDFS.	CO1		5

			UNIT - III			
	5	a)	Design a Table as 'employee' in **Cassandra** and the details are <table><tr><th>Emp_no</th><th>Name</th><th>Department</th><th>Salary</th></tr><tr><td>11</td><td>akash</td><td>sales</td><td>30000</td></tr><tr><td>22</td><td>anu</td><td>HR</td><td>40000</td></tr><tr><td>33</td><td>ajay</td><td>Support</td><td>35000</td></tr></table> Perform the following: 1. Create the keyspace Company 2. Create a Table as 'employee' 3. Insert the values shown above. 4. Alter the table by adding a column Email_ID 5. Update the table by replacing the Ram in the place of Ajay 6. Delete the first row from the table 7. Apply and Demonstrate the concept of index using a query. 8. Use a suitable collections concept for inserting a group of email_id for each student. 9. Truncate the table 10. Drop the table	***CO4***	***PO3***	**10**
		b)	Apply appropriate Hive file format required for the below requirement and explain them in detail. a) A Hive file format which offers high row level compression rates b) A file format provides a highly efficient way to store data in a Hive table. c) A file format which is column-oriented binary file format.	***CO2***	***PO1***	**10**
			OR			
	6	a)	CQL data types in collection facilitate storing multiple values in a single variable. Justify your answer using all the 3 collection types with create, insert and update example commands each.	***CO2***	***PO1***	**10**
		b)	Create a table Salses (s_id, name, name, pieces_sold, price) and Car (C_id,Brand,Colour, s_id) and perform the below queries in Hive framework. 1. Query to find the names and their prices of all the cars whose brand is BMW. 2. Query to find the names of the car, pieces sold and its brand with minimum price value. 3. Query to find the name of the car and its brand whose pieces sold is more than 50. 4. Drop a column "colour" from the Car_info table. 5. Drop both the tables.	***CO3***	***PO2***	**10**
			UNIT - IV			
	7	a)	Apache Spark is a light-fast cluster computing designed for fast computation. Justify with a neat spark execution model diagram	***CO1***		**6**
		b)	Create a table BOOK in mysql which has (book_id, book_name, Publisher). Design suitable queries/ commands to do the following in the Sqoop framework.	***CO4***	***PO3***	**10**

			a) Import the BOOK table from MySQL database server to HDFS. b) Verify the imported data in HDFS c) Import BOOK table data into '/queryresult' directory. d) The subset query to retrieve the book_ id and book_name, whose publisher is "McGraw Hill". e) Imagine a manager adds another row with data f) " 2220,The Alchemist, HarperCollins". Import this data in the BOOK table. g) Export the BOOK data from the HDFS to the BOOK table in MySql database server			
		c)	Identify any 4 features of data wrapper.	*CO1*		**4**
			OR			
	8	a)	Apply appropriate transformations () required for the below description and explain them with an example each: 1. Returns a new RDD, containing only the elements that meet a predicate 2. Data is shuffled according to the key value K in another RDD. 3. Returns a new dataset that contains the distinct elements of the source dataset 4. We get only the common element of both the RDD in the new RDD	*CO2*	*PO1*	**8**
		b)	Mention any 6 salient features of Sqoop.	*CO1*		**6**
		c)	A RAILWAYS table has 30 entries. Design a query to import from MySQL into HDFS. Later, a new entry was made "1119, Chennai express, Bangalore-Chennai". Write the appropriate import commands in order to import the new data into HDFS. Export the Student registration data from the HDFS to the student table in MySQL database server	*CO3*	*PO2*	**6**
			UNIT - V			
	9	a)	Explain how Writes in Zookeeper are done to prevent race conditions.	*CO1*		**4**
		b)	Flume Agent plays an important role in synchronization and aggregation of live stream data. Justify your answers with respect to each component of Flume agent	*CO2*	*PO1*	**6**
		c)	Explain any 10 features of apache flume	*CO1*		**10**
			OR			
	10	a)	Zookeeper plays an important role in the Hadoop ecosystem, and justify why is it important?	*CO1*		**6**
		b)	Explain applications of Apache flume	*CO1*		**6**
		c)	Explain the limitations of Apache Flume.	*CO3*	*PO2*	**8**

****