USN										23ISPC402

MANGALORE INSTITUTE OF TECHNOLOGY & ENGINEERING
(A Unit of Rajalaxmi Education Trust®, Mangalore)
Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE, New Delhi
Accredited by NAAC with A+ Grade & ISO 9001:2015 Certified Institution

Model Question Paper

Big Data Analytics

Time: 3 Hours (180 Minutes) **Max. Marks: 100**

Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.
2. M: Marks, L: RBT (Revised Bloom's Taxonomy) level, C: Course outcomes.

		Module -1	M	L	C
Q1	a.	Define Big Data and explain its 3V's characteristics.	06	L2	CO1
	b.	Discuss the various challenges associated with Big Data.	06	L2	CO1
	c.	Explain the following: In memory analytics, In database processing, Symmetric Multiprocessor System, Massively Parallel Processing	08	L2	CO1
		OR			
Q2	a.	How do you classify data as structured, semi-structured, unstructured? Give an example for each.	06	L2	CO1
	b.	With a neat diagram, explain the structure of a typical Data Warehouse.	08	L2	CO1
	c.	What is a NoSQL database? List and explain various types of NoSQL databases.	06	L2	CO1
		Module- 2			
Q3	a.	Compare RDBMS with Hadoop, highlighting the key differences.	06	L3	CO2
	b.	Consider a file emp.txt of size 300 MB to be stored in HDFS with a default block size of 128 MB and a replication factor of 3. With a neat diagram, illustrate how the client interacts with the NameNode and DataNodes to Write this file, and how the Read operation retrieves it back.	08	L3	CO2
	c.	Consider the following two-line input document: Line 1: big data big data hadoop Line 2: hadoop spark big data With this example, illustrate the role of Mapper, Reducer, Combiner and Partitioner in computing word frequencies using the MapReduce programming model.	06	L3	CO3
		OR			
Q4	a.	With an example, illustrate the working of Hadoop YARN (Hadoop V2) architecture and compare its advantages over the Hadoop V1 architecture.	08	L3	CO2

	b.	Develop a Hadoop MapReduce Java program to implement the Word Count application. Explain how the Mapper and Reducer process the input data and produce the final word-frequency count.	06	L3	CO2
	c.	Consider a dataset of employee records with fields EmpID, Name, Department and Salary: 101, Asha, ISE, 65000 102, Ravi, CSE, 58000 103, Sneha, ISE, 72000 104, Kiran, ECE, 60000 Illustrate, with this dataset, how MapReduce is used for Searching (employees in ISE), Sorting (by Salary) and Compression of the output.	06	L3	CO3
Module - 3					
Q5	a.	Consider the following MongoDB document in a students collection: { "_id": 1, "name": "Anitha", "dept": "ISE", "marks": [78, 85, 90], "address": { "city": "Mangaluru", "pin": 575001 } } By considering this document, illustrate the key features of MongoDB.	06	L3	CO4
	b.	Consider a Student table in an RDBMS with columns USN, Name, Dept, Marks. Compare how the same data would be represented as a MongoDB document, and explain the various data types supported in MongoDB.	06	L3	CO4
	c.	Consider the following student collection in MongoDB: { "_id": 1, "name": "Anitha", "dept": "ISE", "marks": 88 } { "_id": 2, "name": "Ravi", "dept": "CSE", "marks": 72 } { "_id": 3, "name": "Sneha", "dept": "ISE", "marks": 91 } Using this collection, illustrate the MongoDB Query Language (MQL) to: insert a new document, find all ISE students, update Ravi's marks to 75, and delete the record with _id 2.	08	L3	CO4
OR					
Q6	a.	Consider a Cassandra keyspace hospital storing patient records across multiple nodes for a healthcare application handling millions of records. With this example, illustrate the key features of Cassandra.	06	L3	CO4
	b.	Consider a Cassandra table student_info with columns: usn (text), name (text), marks (list<int>), subjects (set<text>), address (map<text,text>). With this example, illustrate CQL data types and Collections used in Cassandra.	06	L3	CO4
	c.	Consider a Cassandra table student (usn text PRIMARY KEY, name text, dept text, marks int) with sample rows: ('4MT23IS001','SS1','ISE',88) ('4MT23IS002','SS2','CSE',72) With this example, illustrate CRUD operations, an Alter command to add a new column email, and Import/Export of this table's data.	08	L3	CO4
Module - 4					
Q7	a.	Consider the HiveQL query: SELECT dept, AVG(marks) FROM student GROUP BY dept;	06	L3	CO2

		With this example query, trace how it is processed through the Hive architecture (Driver, Compiler, Metastore and Execution Engine).			
	b.	Consider a student dataset with fields usn (string), name (string), marks (array of int) and dept (string). Illustrate the choice of appropriate Hive data types and an appropriate Hive file format for storing this dataset efficiently.	06	L3	CO2
	c.	Write a HiveQL statement to create a table named student with four columns as USN, Name, Dept, Phone and load the data by considering the following text file named "ss.txt" stored in directory /home/user1: 4MT23IS001, SS1, ISE, 8978987894 4MT23IS002, SS2, ISE, 8978987895 4MT23IS003, SS3, ISE, 8978987896 Also write a query to alter the table name to "ISEStudent" and the third column name to "Department".	08	L3	CO2
OR					
Q8	a.	Consider the following Pig Latin script: A = LOAD 'student.csv' USING PigStorage(',') AS (usn:chararray, name:chararray, dept:chararray, marks:int); B = FILTER A BY marks > 75; DUMP B; With this example script, demonstrate how Pig runs in Local mode versus MapReduce (Hadoop) execution mode.	06	L3	CO2
	b.	Consider the following student.csv file: USN, NAME, DEPT, MARKS 4MT23IS001, SS1, ISE, 88 4MT23IS002, SS2, CSE, 72 4MT23IS003, SS3, ISE, 91 4MT23IS004, SS4, ISE, 65 With this dataset, illustrate the working of any four Relational Operators of Pig Latin.	06	L3	CO2
	c.	Considering the following student.csv file: USN, NAME, DEPT, Phone 4MT23IS001, SS1, ISE, 8978987894 4MT23IS002, SS2, ISE, 8978987895 4MT23IS003, SS3, ISE, 8978987896 (a) Write and illustrate the Pig Latin script to load the student.csv file into relation "A". (b) Write and illustrate the Pig Latin "foreach" statement to dump the values of the first two columns. Also compare Pig with Hive.	08	L3	CO2
Module - 5					
Q9	a.	Consider an e-commerce company that wants to ingest live clickstream data, run SQL-based sales reports, and build a recommendation model using 5 years of purchase history. With this use case, illustrate how the components of the Apache Spark Software Stack work together.	06	L3	CO5

	b.	Consider the review text: "The battery life is excellent but the camera quality is average." With this example, illustrate how a Text Mining technique can extract keywords/sentiment from it, and how a Web Mining technique can analyze a webpage containing many such reviews.	07	L3	CO5
	c.	Consider a text file input.txt with the content: big data is big spark makes big data fast Write a Spark program for Word Count using PySpark to compute the frequency of each word in this file.	07	L3	CO5
		OR			
Q10	a.	Consider a sales dataset sales.csv with columns Region, Product, Quantity, Price: South, Laptop, 5, 55000 North, Mobile, 12, 18000 South, Mobile, 8, 18000 With a neat diagram and this dataset, illustrate how Data Analysis (e.g., computing total revenue per region) is carried out with Spark.	07	L3	CO5
	b.	Consider the following sample web server log entries: 192.168.1.10 – GET /home.html – 10:15:00 192.168.1.11 – GET /products.html – 10:15:05 192.168.1.10 – GET /products.html – 10:16:00 With this example, illustrate how Web Usage Analytics is applied to analyze the structure of the Web.	06	L3	CO5
	c.	Consider a small web graph with four pages A, B, C, D having the links: A→B, A→C, B→C, C→A, D→C. Explain the PageRank algorithm and illustrate, using this example, how the Web Graph is analyzed to compute page ranks.	07	L3	CO5

-***-**