SRN

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	**UE20CS936**

February 2025: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER III

UE20CS936 - INTRODUCTION TO BIG DATA

Time: 3 Hrs	Answer All Questions	Max Marks: 100

Instructions

1. Answer all the questions.
2. Section A and B should be handwritten in the answer script provided and signed at the end of the same.
3. Section C contains programming questions which have to be answered in the system.
4. Follow the instructions for Section C which is available in question paper (jupyter notebook).

		Section A (20 marks)	
1	a)	Provide a diagram of Hive architecture and explain its main elements.	5
	b)	What is Big Data? Describe the three fundamental aspects that define it.	5
	c)	What are the key differences between a Data Lake and a Data Warehouse?	5
	d)	What are the different types of NoSQL databases? Write a brief note.	5
		Section B (40 Marks)	
2	a)	Write HDFS shell commands for the following- 1. Print contents of the directory by path, showing the names, permissions, owner, size and modification date for each entry? (2 marks) 2. How do you upload multiple files from the local system to a directory in HDFS?. (2 marks) 3. How do you display the contents of a file in HDFS line by line (paged view)? (2 marks) 4. Write command to remove a file or directory identified by path in HDFS and recursively delete any child entries. (2 marks) 5. Write command to copy the 'testfile' of the hadoop filesystem to the local file system (2marks) Note: Consider InputDir, OutputDir, XYZ, SampleDir, and file.txt are under the present working directory.	10
	b)	Considering **sc** as spark content object, and **rdd** as RDD object, write Spark commands to, 1. Create an RDD from the following list: **List(1, 2, 3, 4, 5,6,7,8,9,10).** (2 marks) 2. Display/Print first four elements of the RDD. (2 marks) 3. Display/Print the first element of the RDD. (2 marks) 4. Explain with example map,filter Apache Spark transformations.(4 marks)	10

SRN

	c)	Write below queries in Hive; 1. Write hive query to create databases name: **emp**. (2 Marks) 2. Write hive query to **CREATE EXTERNAL TABLE** in emp name it- **employee** with **emp_id,name,location, dep,designation** and **salary** as columns (4 Marks) 3. Write a hive query to perform an inner join on the Table1 and Table 2 on ‘id’ column (4 Marks)	10
	d)	Write commands / query in MongoDB 1. Create a collection named ‘product collection’. (2 mark) 2. Insert 5 documents in product collection based on name, rating, brand.(2 mark) 3. Write query to find those products which have received 5/5 rating.(3 mark) 4. Write a query to update those records where the product name is AC to "Air conditioner"and print it. (3 mark)	10
		Section C (40 marks)	
3		Using PySpark and Spark-SQL libraries process the given dataset in order to find out solutions of queries mentioned below..	
	a)	I. What’s the overall minimum, maximum and average salary from the dataset? (6 marks) II. How many female candidates are not placed ? (4 marks) III. Out of total male candidates placed, how many do not have any work experience ? (3 marks) IV. Remove the feature 'sl_no' and also remove null values from the DataFrame. (2 marks)	15
	b)	Using Spark ML libraries process the Dataframe as questioned below. I. Convert all string columns into numeric values using StringIndexer transformer and make sure now DataFrame does not have any string columns anymore. (5 marks) II. Using vectorAssembler combines all columns (except target column i.e., 'salary') of spark DataFrame into a single column (name as features). Make sure DataFrame now contains only two columns: features and salary. (5 marks) III. Split the vectorized dataframe into training and test sets with one fourth records being held for testing (3marks) IV. Build a LinearRegression model on train set use featuresCol="features" and 'salary'(6 marks) V. Perform prediction on the testing data and Print MSE value? (6 marks)	25