SRN | | | | | | | | | | | | |

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

July 2024: 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 programing questions which have to be answered in the system.
4. Use command ***sudo docker run -it -p 8888:8888 bda-esa-06july2024*** to access/open Jupyter notebook of Section C.
5. Do not use the command ***jupyter notebook*** to open notebook.

		Section A (20 marks)	
1	a)	Draw Spark architecture and explain its various components.	6
	b)	List any 4 differences between Data Lake and Data Warehouse.	4
	c)	Explain partitioning (in hive) with an example.	6
	e)	List any 2 differences between Coalesce and Repartition (in spark)	4
		Section B (40 Marks)	
2	a)	Write HDFS shell commands for the following- 1. To Copy **file1.txt** from folder **InputDir** to **OutputDir** as **file2.txt.** (2 marks) 2. To Delete an empty directory named as **XYZ.** (2 marks) 3. To List the files and directories under folder named **SampleDir.** (2 marks) 4. To Recursively list the files and directories exist under folder named **SampleDir**. (2 marks) 5. To change the Permission of file named **file.txt to** Read only (**444**) **(**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).** (2 marks) 2. Read/load a text file located at "/path/to/file.txt" into an RDD. (2 marks) 3. Filter out the even numbers from RDD. (2 marks) 4. Map each element in the RDD to its square. (2 marks) 5. Count the number of elements in the RDD. (2 marks)	10
	c)	Write below queries in Hive; 1. Write hive query to create databases name: **anotherDB**. (2 Marks)	10

SRN | | | | | | | | | | | | |

		2. Write hive query to **CREATE EXTERNAL TABLE** in anotherDB name it- **orders1** with **order_id**, **order_date**, **order_customer_id** and **order_status** as columns(4 Marks) 3. Write hive query to load data in **orders1** table using file which is available in local file system. (4 Marks)	
	d)	Write commands/query in MongoDB to, 1. Create a collection named **orders**. (2 mark) 2. Insert below two records in orders. (4 mark) {"order_id": 1, "order_customer_id": 11599, "order_status": "CLOSED" } {"order_id": 2, "order_customer_id": 11698, "order_status": "OPEN" } 3. Fetch orders with **order_status** as COMPLETE. (4 marks)	10
		Section C (40 marks)	
3		QS World University Rankings dataset is provided and loaded as Spark-DataFrame. Using Spark libraries execute the steps, as questioned below.	
	a)	Using PySpark Dataframe or Spark-SQL libraries process the Dataframe to find out solutions of queries mentioned below. i. How many Institutions are included in the dataset? (2 mark) ii. How many Institutions from 'India' are included in dataset? (3 marks) iii. Print the average "Citations per Faculty" for universities located in 'India'? (5 marks) iv. List Institutions where "International Students" percentage is 100 % along with their location ("Location Full"). (5 marks)	15
	b)	Using PySpark ML build a regression model, as questioned below. i. Recreate the Dataframe by Dropping all the rows where 'QS Overall Score' is mention as '-' and also convert it as float type. (2 marks) ii. Remove all the rows with any missing entry.(3 marks) iii. Convert all string columns into numeric values using StringIndexer transformer and make sure now DataFrame does not have any string columns anymore. (5 marks) iv. Using vectorAssembler combines all columns, except target column i.e. 'QS Overall Score', of spark DataFrame into single column (name it as features). Make sure DataFrame now contains only two columns, 'features' and 'QS Overall Score'. (5 marks) v. Split the vectorised Dataframe into training and test sets with one fifth records being held for testing. (2 marks) vi. Train default LinearRegression model with features as 'featuresCol' and 'QS Overall Score' as label on training set. (3 marks) vii. Perform prediction on the testing data and Print RMSE value. (5 marks)	25