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

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	**UE2**

March 2022: 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: 80

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 and uploaded in Olympus Login.
4. Write appropriate inferences.

Section A (20 marks)			
1	a)	List two primary differences between Hadoop version-2 and Hadoop version-3	4
	b)	What is Hive metastore? Can NoSQL Database -HBase can be configured as hive metastore? (3+1)	4
	c)	Using an example, depict how MapReduce computes word count.	4
	d)	Draw Spark architecture and explain its various components.	4
	e)	What is cap theorem ? Where does MongoDB stands in cap theorem? (3+1)	4
Section B (30 Marks)			
2	a)	Write HDFC shell commands for the following- i. To print **version** of installed Hadoop. (1 mark) ii. To Copy **file1.txt** from folder **InputDir** to **OutputDir** as **file2.txt.** (2 marks) iii. To Delete an empty directory named as **XYZ.** (2 marks) iv. To list the contents of folder named **SampleDir.** (2 marks) v. To fetch the usage instructions/details of **mkdir** command. (2 marks)	9

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

	b)	Write a Spark program pseudo-code to load a text file named as text.txt into spark RDD and compute its wordcounts.	7
	c)	Two hive tables are shown below. Write a hive query to perform an inner join on the Table1 and Table 2 on 'Id' column. Also, write the expected output of your inner join query (marks 4+3)	7
	d)	Write commands/query in MongoDB to, i. Create a collection named **orders**. (1 mark) ii. Insert below record in orders. (2 mark) {"order_id": 1, "order_date": '2013-07-25 00:00:00.0', "order_customer_id": 11599, "order_status": "CLOSED" } iii. Fetch orders with **order_status** as COMPLETE. (2 marks) iv. Compute count of orders with status COMPLETE and CLOSED. (2 marks)	7

Table1	
Name	Id
Joe	2
Hank	4
Ali	0

Table2	
Id	Name
2	Tie
4	Coat
3	Hat
1	Scarf

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

Section C (30 marks)

3		Bangalore Housing Dataset is provided and loaded as Spark-DataFrame. Using Spark libraries execute the steps, as questioned below.	
	a)	Using PySpark and Spark-SQL libraries process the dataset to find out solutions of queries mentioned below. i. Count the total number of housing-properties listed from 'HSR Layout' location. (2 marks) ii. How many '2 BHK' size housing-properties are listed from 'Whitefield' location? (3 marks) iii. What is the average price of '2 BHK' size housing-properties in 'HSR Layout' location? (5 Marks)	10
	b)	Using Spark ML execute the steps, as questioned below. i. Remove the features, having more than one third of their entries as missing/null. For the remaining missing values- remove the corresponding row entry from the DataFrame. (3 marks) ii. Convert all string columns into numeric values using StringIndexer transformer and make sure now DataFrame does not have any string columns anymore. (5 marks) iii. Using vectorAssembler combines all columns (except target column i.e., 'price') of spark DataFrame into single column (name as features). Make sure DataFrame now contains only two columns features and price. (5 marks) iv. Split the vectorized dataframe into training and test sets with one fourth records being held for testing (2 marks) v. Train default LinearRegression model with features as 'featuresCol' and 'price' as label. (5 marks)	20

This file is meant for personal use by rg.ravigupta91@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.