

	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: 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 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 (40 Marks)			
2	a)	Write HDFS shell commands for the following- <ul style="list-style-type: none"> i. To print version of installed Hadoop. (2 mark) ii. To Copy file1.txt from folder InputDir to OutputDir as file2.txt. (3 marks) iii. To Delete an empty directory named as XYZ. (3 marks) iv. To list the contents of folder named SampleDir. (3 marks) v. To fetch the usage instructions/details of mkdir command. (3 marks) 	14

b)	Write a Spark program pseudo-code to load a text file named as text.txt into spark RDD and compute its wordcounts.	8																						
c)	<p>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+4)</p> <table><tr><th colspan="2">Table1</th></tr><tr><th>Name</th><th>Id</th></tr><tr><td>Joe</td><td>2</td></tr><tr><td>Hank</td><td>4</td></tr><tr><td>Ali</td><td>0</td></tr></table> <table><tr><th colspan="2">Table2</th></tr><tr><th>Id</th><th>Name</th></tr><tr><td>2</td><td>Tie</td></tr><tr><td>4</td><td>Coat</td></tr><tr><td>3</td><td>Hat</td></tr><tr><td>1</td><td>Scarf</td></tr></table>	Table1		Name	Id	Joe	2	Hank	4	Ali	0	Table2		Id	Name	2	Tie	4	Coat	3	Hat	1	Scarf	8
Table1																								
Name	Id																							
Joe	2																							
Hank	4																							
Ali	0																							
Table2																								
Id	Name																							
2	Tie																							
4	Coat																							
3	Hat																							
1	Scarf																							
d)	<p>Write commands/query in MongoDB to,</p> <p>i. Create a collection named orders. (1 mark)</p> <p>ii. Insert below record in orders. (3 mark)</p> <pre>{"order_id": 1, "order_date": '2013-07-25 00:00:00.0', "order_customer_id": 11599, "order_status": "CLOSED" }</pre> <p>iii. Fetch orders with order_status as COMPLETE. (3 marks)</p> <p>iv. Compute count of orders with status COMPLETE and CLOSED. (3marks)</p>	10																						

Section C (40 marks)

3		Bangalore Housing Dataset is provided and loaded as Spark-DataFrame. Using Spark libraries execute the steps, as questioned below.	
	a)	<p>Using PySpark and Spark-SQL libraries process the dataset to find out solutions of queries mentioned below.</p> <ol style="list-style-type: none"> Count the total number of housing-properties listed from 'HSR Layout' location. (4 marks) How many '2 BHK' size housing-properties are listed from 'Whitefield' location? (5 marks) What is the average price of '2 BHK' size housing-properties in 'HSR Layout' location? (6 Marks) 	15
	b)	<p>Using Spark ML execute the steps, as questioned below.</p> <ol style="list-style-type: none"> 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. (5 marks) Convert all string columns into numeric values using StringIndexer transformer and make sure now DataFrame does not have any string columns anymore. (5 marks) 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) Split the vectorized dataframe into training and test sets with one fourth records being held for testing (2 marks) Train default LinearRegression model with features as 'featuresCol' and 'price' as label. (8 marks) 	25