Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2021

Course: In Memory Processing **Semester: VIII**
Program: B.Tech CSE DevOps **Time 03 hrs.**
Course Code: CSBD 3003 **Max. Marks: 100**

SECTION A

1. Each Question will carry 5 Marks
2. Instruction: Write brief answers

S. No.		**CO**
Q 1	Discuss JobTracker? Explain job scheduling through JobTracker?	**CO1**
Q 2	Explain the concept of Resilient Distributed Dataset (RDD) and how do we create RDDs in Spark?	**CO3**
Q 3	Describe Distributed cache in MapReduce Framework? How does it work?	**CO1**
Q 4	What do you understand by Pair RDD?	**CO3**
Q 5	Explain the components of Spark Ecosystem?	**CO4**
Q 6	Difference between map and flatMap?	**CO4**

SECTION B

1. Each question will carry 10 marks
2. Instruction: Write short notes

Q 7	Define shuffling in Spark. What are the scenerios when it occur ? Does shuffling change the number of partitions?	**CO2**
Q 8	How can differentitate the wide and narrow transformation with the help of suitable examples.	**CO3**
Q 9	Does Spark SQL help in big data analytics through external tools too? Justify	**CO2**
Q 10	Differentiate between the benefits of Apache Spark over Mapreduce framework?	**CO4**
Q 11	Explain what combiners are and when you should use a combiner in a MapReduce Job? **OR**	**CO1**

	Identify what is the Partitioner and its usages? Does Partitioner run in its own JVM or shares with another process?	
SECTION-C **1. Each question carries 20 Marks.** **2. Instruction: Write long answer.**		
Q 12	Write a Python program in Spark to compute the total count of unique words in Spark? Also count the total number of occurence of that word in that paragraph. **OR** Discuss the following: a. Broadcast variables b. Accumulators c. Persistence Storage Levels d. Catalyst Optimizer	**CO5**