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

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

Course: In Memory Processing **Semester: VI**
Program: B.Tech CSE Big Data **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 the role of Catalyst Optimizer in Spark SQL?	**CO3**
Q 2	Explain the concept of Resilient Distributed Dataset (RDD) and how do we create RDDs in Spark?	**CO2**
Q 3	Differentiate between CreateOrReplaceTempView and createGlobalTempView?	**CO1**
Q 4	What do you understand by Pair RDD?	**CO2**
Q 5	What are the components of Spark Ecosystem?	**CO1**
Q 6	Distinguish between map and flatMap transformation in Spark?	**CO2**

SECTION B

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

Q 7	Define shuffling in Spark. When does it occur ? Does shuffling change the number of partitions?	**CO2**
Q 8	Distinguish between wide transformation and narrow transformation with suitable examples.	**CO1**
Q 9	Does Spark SQL help in big data analytics through external tools too? Justify	**CO2**
Q 10	When do you use apache spark and what are the benefits of Spark over Mapreduce framework?	**CO1**
Q 11	What are the different storage levels of persistence in Spark?	**CO3**

SECTION-C

1. Each question carries 20 Marks.

2. Instruction: Write long answer.

Q 12	IMBD is an online database of movie-related information. IMBD users rate the movies and provide reviews. They rate the movies on a scale of 1 to 5; 1 being the worst and 5 being the best. The dataset also has additional information, such as the release year of the movie. You have to analyse the data collected and answer the following questions. Sample data from the dataset: (Title, release year, rating, number of users) 1,The Nightmare Before Christmas,1993,3.9,4568 2,The Mummy,1932,3.5,4388 3,Orphans of the Storm,1921,3.2,9062 You need to find: 1) The total number of movies (2) 2) The maximum rating of movies (2) 3) The number of movies that have maximum rating (3) 4) The movies with ratings 1 and 2 (3) 5) The list of years and number of movies released each year (5) 6) The number of movies that have a runtime of two hours (5) **OR** a) Discuss Broadcast variables. Why is there a need for broadcast variables when working with Apache Spark? Explain with code snippet. b) How would you compute the total count of unique words in Spark? Write a program in Pyspark?	**CO3**