UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2022

Course: Big Data Overview   **Semester:** III
Program: B.Tech CSE+Big Data   **Time** : 03 hrs.
Course Code: CSBD2006   **Max. Marks: 100**

Instructions:

SECTION A

S. No.		Marks	CO
Q 1	Explain how Hadoop is different from the Traditional Processing Systems using Relational Database Management?	4	CO1
Q 2	What is the difference between structured and unstructured data? Explain how each type of data is used in Big Data. Also, give examples of both structured and unstructured data.	4	CO1
Q 3	Define Big Data and discuss the 5V's that are associated with the Big-Data.	4	CO1
Q 4	Explain the Key Value pair NoSql database and its need in big data platforms.	4	CO1
Q 5	Distinguish between Data lake and Data swamp.	4	CO3

SECTION B

Q 6	How do you recover NameNode when it is down? Explain using the block diagram of HDFS. **Or** What are the various steps/stages involved in the deploying of big-data solution? Discuss the Hadoop tools used at each stage.	10	CO2
Q 7	What was the need for introducing YARN in Hadoop? Explain the working YARN daemons with the architecture.	10	CO2
Q 8	Data growth explosion leading to extensive need of storage and processing. Justify the statement with the help of a use case	10	CO3
Q 9	List and explain the essential elements of data lake and analytical platform. Also, explain the data lake challenges.	10	CO1

SECTION-C

Q 10	In the transportation industry, what are the different data sources based on their categories? Discuss the use of data lake in the transportation industry.	20	CO3
Q 11	i. How does Name node determines which data node to write on? Explain the data writing procedure in HDFS with the help of diagram. ii. Imagine that you are uploading a file of 500MB into HDFS. 100MB of data is successfully uploaded into HDFS and another client wants to read the uploaded data while the upload is still in progress. What will happen in such a scenario, will the 100 MB of data that is uploaded will it be displayed? Consider the block size to be 100MB. **Or** i. Describe the scalability and high availability feature of HDFS. How does a name node determines if a data node is not operational? ii. If 8TB is the available disk space per node for data storage and 2TB for operating system. Also, the disk space utilization factor is 70%. Considering there is no storage overhead or compression, how will you estimate the number of data nodes required to store 60TB data?	20	CO2