Name :

Enrolment No. :

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, Dec 2024

Program Name	: B.Tech CSE (All specializations)	Semester :	VII
Course Name	: Big Data Analysis	Time :	3 hours
Course Code	: CSBD4006P	Max. Marks :	100
No. of Page(s)	: 2		
Instructions	: Attempt all sections.		

SECTION-A

S. No.	Questions	Marks	CO
Q.1	Explain the components of YARN in brief.	4	CO1
Q.2	In the sentence, “Amazon announced a new product in Seattle,” explain how Named Entity Recognition would identify whether ”Amazon” refers to the company or the river.	4	CO2
Q.3	Explain how does the document-based database differs from relational Database.	4	CO3
Q.4	Identify the learning techniques used in the following tasks. a) Face Recognition b) Image Segmentation c) Weather Forecast d) Social Network Analysis	4	CO4
Q.5	List the data reading procedure in Hadoop in step-by-step manner.	4	CO1

SECTION-B

S. No.	Questions	Marks	CO
Q.6 A) B)	Consider a rapidly growing e-commerce platform that handles increasing volumes of user data and transaction requests. Explain why vertical scalability would be a viable approach if improving the performance of the existing system without adding new nodes is crucial for maintaining smooth operations. State any two applications of horizontal scaling.	10 (6+4)	CO1
Q.7	You are working with a large organization that has recently adopted a Data Lake for storing massive volumes of structured and unstructured data. However, you encounter two significant issues during data management: a) The small file problem and b) The shared file problem. Discuss how these challenges impact data storage and retrieval along with potential solutions to address them.	10 (5+5)	CO2

<table>
<tr><td>Q.8</td><td>Consider the following table and a task to find the number of employees per department. Show how MapReduce would perform the operations using a diagram.
<table><tr><th>Employee ID</th><th>Name</th><th>Department</th></tr><tr><td>101</td><td>Alice</td><td>HR</td></tr><tr><td>102</td><td>Bob</td><td>IT</td></tr><tr><td>103</td><td>Charlie</td><td>IT</td></tr><tr><td>104</td><td>Diana</td><td>HR</td></tr><tr><td>105</td><td>Eve</td><td>Finance</td></tr></table></td><td>10</td><td>CO3</td></tr>
<tr><td rowspan="3">Q.9</td><td>Explain the following components of Data Lakes: storage, data format, compute, and metadata.</td><td>10</td><td rowspan="3">CO4</td></tr>
<tr><td>OR</td><td></td></tr>
<tr><td>Compare and Contrast Data Lakes to Data Warehouses on the following points: Modularity, Schema Enforcement, Cost and Data Format.</td><td>10</td></tr>
<tr><td colspan="4">SECTION-C</td></tr>
<tr><td>Q.10</td><td>Differentiate between probabilistic and deterministic classifiers. Write a short note on the following classification algorithms:
a) Logistic Regression
b) Decision Trees
c) Naïve Bayes</td><td>20 (5*4)</td><td>CO1</td></tr>
<tr><td>Q.11</td><td>Explain with examples <b>any four</b> of the following data cleaning and transformation techniques for text data.
a) Sentence Segmentation
b) Tokenization
c) Stop-word Removal
d) Stemming
e) Lemmatization</td><td>20 (5*4)</td><td>CO3</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>A)</td><td>Explain any four factors to consider for replication factor of data in HDFS along with a suitable example of each.</td><td>8</td><td rowspan="2">CO4</td></tr>
<tr><td>B)</td><td>Explain three levels of data locality in HDFS.</td><td>12</td></tr>
</table>