|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Roll No. |  |  |  |  |  |  |  |  |  |  |  |  |



 **PRESIDENCY UNIVERSITY**

  **Bengaluru**

|  |
| --- |
| **End - Term Examinations – JANUARY 2025** |
| **Date:** 10 – 01- 2025 **Time:** 09:30 am – 12:30 pm |

|  |  |
| --- | --- |
| **School:** SOIS | **Program:** BCA/BSD |
| **Course Code :** CSA3004 | **Course Name :** Big Data Analytics |
| **Semester**: V | **Max Marks**: 100 | **Weightage**: 50% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CO - Levels** | **CO1** | **CO2** | **CO3** | **CO4** |
| **Marks** | **24** | **24** | **28** | **24** |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Do not write anything on the question paper other than roll number.*

**Part A**

|  |
| --- |
| **Answer ALL the Questions. Each question carries 2marks. 10Q x 2M=20M** |
| **1** | Outline the main challenges faced in traditional data processing compared to Big Data processing.  | **2 Marks**  | **L1**  | **CO1**  |
| **2** | Identify the purpose of replication management in HDFS?  | **2 Marks**  | **L1**  | **CO1**  |
| **3** | Define Shuffle and Sort in the context of MapReduce  | **2 Marks**  | **L1**  | **CO2**  |
| **4** | Define Distributed File System (DFS)?  | **2 Marks**  | **L1**  | **CO2**  |
| **5** |  State the features of Hadoop 2.0.  | **2 Marks**  | **L1**  | **CO2**  |
| **6** | Define the concept of Hive Table partitioning  | **2 Marks**  | **L1**  | **CO3**  |
| **7** | Identify the differences between Hbase and RDBMS.  | **2 Marks**  | **L1**  | **CO3**  |
| **8** | Label the Limitations of Hive.  | **2 Marks**  | **L1**  | **CO3**  |
| **9** | Identify the main role of Apache Spark in Big Data analytics?  | **2 Marks**  | **L1**  | **CO4**  |
| **10** | Tabulate the differences between Apache Spark and Hadoop.  | **2 Marks**  | **L1**  | **CO4**  |

**Part B**

|  |
| --- |
| **Answer the Questions Total 80 Marks.** |
| **11.** | **a.** | Explain in detail the various types of data (structured, unstructured, semi-structured, and quasi-structured) in the context of Big Data. | **20 Marks** | **L2** | **CO1** |
| **Or** |
| **12.** | **a.** | Describe the architecture of YARN in detail, including its components and their functions. | **20 Marks** | **L2** | **CO1** |
|  |  |  |  |  |  |
| **13.** | **a.** | Employ the scenario below and suggest an efficient approach to overcome the same. The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. The analyst wants to know the age of senior most lady passenger who had travelled in titanic ship. Help analyst by developing a MapReduce program. Use your own data set with the following fields such as Passenger ID, Survived, Name, Gender, Age, Ticket.  | **20 Marks** | **L3** | **CO2** |
| **Or** |
| **14.** | **a.** | Employ the scenario below and suggest an efficient approach to overcome the same.  A data scientist is someone who makes value out of data. Such a person proactively fetches information from various sources and analyses it for a better understanding of how the business performs and to build AI tools that automate certain processes within the company. Data scientist duties typically include creating various machine learning-based tools or processes within the company, such as recommendation engines or automated lead scoring systems. People in this role should also be able to perform statistical analysis. Create a Hadoop MapReduce for the application to process the data(fields given below) to find the maximum salary of a newly recruited data scientist. Use your own data set with the fields such as Company\_Name, Job\_Title, Min\_Experience, Salary, Number.  | **20 Marks** | **L3** | **CO2** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **15.** | **a.** | Describe in detail the various types of operations (DDL and DML) in Hive with examples | **20 Marks** | **L2** | **CO3** |
| **Or** |
| **16.** | **a.** | Explain the architecture and working of HBase. Discuss its main components and how it is used for managing large-scale data. | **20 Marks** | **L2** | **CO3** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **17.** | **a.** | Explain the architecture of Apache Spark and its different components like Spark Core, Spark SQL, and Spark MLlib. | **20 Marks** | **L2** | **CO4** |
| **Or** |
| **18.** | **a.** | Describe the process of creating RDDs in Spark and explain the different operations (transformations and actions) that can be performed on them. | **20 Marks** | **L2** | **CO4** |

**\*\*\*\*\* BEST WISHES \*\*\*\*\***