**Ramaiah Institute of Technology**

**(Autonomous Institute, Affiliated to VTU)**

**Department of Information Science and Engineering**

|  |  |
| --- | --- |
| Course Name: Big Data Laboratory | Term: October 2024 to January 2025 |
| Course Code: ISL75 | IA Marks: 50 |
| Credits: 0:0:1 | Exam Hours: 03, Max. Marks: 50 |

|  |  |
| --- | --- |
| **Question No.** | **Program Description** |
|  | **PART-A** |
|  | **Write MapReduce programs for the following using Java:** |
| 1. | To count the number of occurrences of each word in a given input text. |
| 2. | To read N natural numbers and display the sum along with Odd and Even count. |
| 3. | To analyse the given Employee Data and generate a statistics report with the total number of Female and Male Employees and their average Salary. |
| 4. | To analyse the Titanic Ship Data and find the average age of the people (both male and female) who died in the tragedy and also how many people are survived in each class. |
| 5. | To analyse the Earthquake Data and generate statistics with region and magnitude/region and depth/region and latitude/region and longitude. |
| 6. | To analyse the given Sales Records over a period of time and generate data about the country’s total sales and the total number of products. Country’s total sales and the frequency of the payment mode. |

|  |  |
| --- | --- |
| **Question No.** | **PART-B** |
| 1. | Write a **Spark program using Python**, to count the number of lines and number of characters of each word in a given paragraph. |
| 2. | Write a **Spark program using Python**, to analyse the given **Weather Report Data** and to generate a report with cities having maximum and minimum temperature for a particular year. |
| 3. | Write a **Spark program using Python**, to analyse the given **Insurance Data** and generate a statistics report with the construction building name and the count of building/ county name and its frequency. |
| 4. | Write **Pig Latin scripts** for the following queries on **Crop Production Dataset**:   1. Calculate total production of each crop. 2. Find the average production per year for each crop. 3. Filter all crops grown in ‘Karnataka’. 4. Calculate the total area used for each crop in the year 2010. |
| 5. | Write **Pig Latin scripts** for the following queries on **Olympic Athletes and Hosts Dataset**:   1. Filter athletes participated in the “Tokyo 2020” games. 2. Filter the games held in “China”. 3. Group games by season and count the number of games in each session. 4. Filter games that occurred after the year 2000. |
| 6. | Write **Pig Latin scripts** for the following on **Online Retail Dataset**:   1. Total number of unique customers in the "United Kingdom". 2. Find total invoices generated in the year 2009. 3. Generate the total revenue collected from Country="France". 4. Find the country in which maximum number of items were sold. |