|  |
| --- |
| **BIG DATA LABORATORY** |
| **Course Code: ISL75 Credit: 0:0:1 Prerequisite: Nil Contact Hours: 14L**  **Course Coordinator: Dr. Rajeshwari S B** |

**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.

**PART-B**

1. 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.
2. 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.
3. 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.
4. Write **Pig Latin scripts** for the following queries on **Olympic Athletes and Hosts Dataset**:
5. Filter athletes participated in the “Tokyo 2020” games.
6. Filter the games held in “China”.
7. Group games by season and count the number of games in each session.
8. Filter games that occurred after the year 2000.
9. Write **Hive Query** for the following on **Online Retail Dataset**:
10. Total number of unique customers in the "given country".
11. Filter the country from which the maximum revenue was collected from sales in the month of March 2010.
12. Filter the month of 2010 in which maximum number of items were sold.
13. In the month of January 2010, find the country in which maximum number of items were sold.
14. Write **Hive Query** for the following on **Coffee Sales Dataset**:
    1. Number of customers had “Hot Chocolate “coffee on the given date.
    2. Amount of money collected by “Latte” coffee sales.
    3. Number of customers done the payment through card.
    4. Most favourite coffee.

**References:**

1. “Hadoop: The Definitive Guide”, Tom White ,4th Edition, O’Reilly Media, 2015.
2. "Learning Spark: Lightning-Fast Big Data Analysis", Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, O'Reilly Media, 2016.
3. “Programming pig: Dataflow scripting with Hadoop", Gates, Alan, and Daniel Dai. O'Reilly Media, 2016.
4. “Programming Hive: Data warehouse and query language for Hadoop", Capriolo, Edward, Dean Wampler, and Jason Rutherglen. O'Reilly Media, 2012.

**Course Outcomes (COs):**

1. Explore the Hadoop ecosystem and MapReduce programming model to develop and execute a range of applications for Big Data analytics. (PO-1, 2, 3, 5, 9, 10) (PSO- 1, 2)
2. Dissect Apache Spark, a faster alternative to MapReduce, by implementing large-scale data applications. (PO-1, 2, 3, 5, 9, 10) (PSO- 1, 2)
3. Analyse Big Data applications using Pig and Hive for efficient data processing and querying. (PO-1, 2, 3, 5, 9, 10) (PSO- 1, 2)