Cloud Computing

CLOUD COMPUTING PROJECT

PROJECT DETAILS

Student Name:	**Tehsein Firoze Akhtar**		
Student ID:	**22204524**		
Project Title:	**Diagnosis Trends in Medical Reviews**	CC Project:	**2**
Due Date:	**04 / 12 / 2023**	Submitted Date:	**04 / 12 / 2023**

INSTRUCTIONS

1. **Project:** You are asked to choose between two projects. Once you have done that, you need to:
    a. Develop a cloud-based solution for either manage or analyse medical datasets, depending on the project you chose (1 or 2).
    b. Please follow the instructions as described in the project.
2. **Project Report:** produce a project report describing the data set application being studied. This report should contain the following items:
    a. Application overview.
    b. Objectives: Briefly describe each goal/objective and whether they have been completed.
    c. Describe the problem and the collection of the datasets.
    d. Explain your methodology and implementation.
    e. Discuss the suitability of the tool being used to solve the problem.
    f. Describe the features of your software.
    g. Give a worked example.
    h. Conclusion
    i. References
3. **Submission:** The deadline is 04 December 2023. Passed this deadline penalty will apply.
    a. Please submit your project report, the source code, and all the necessary files to execute your implementation.
    b. You may need to prepare a README file to explain how to execute it.
    c. Please submit ONLY one ZIP file, containing all the necessary files and directories, including the project report.

Application Overview

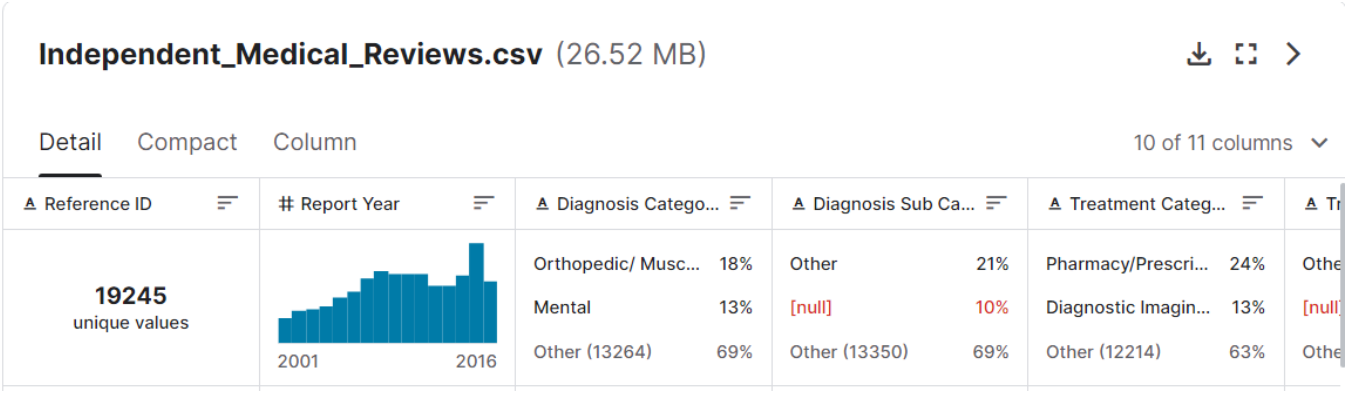
The project focuses on the application of Hadoop/MapReduce for analyzing a medical dataset, specifically the **California Independent Medical Review Dataset**. This dataset comprises various fields such as Reference ID, Report Year, Diagnosis Category, Treatment Category, and more, providing a comprehensive view of medical cases.

Objectives

- **Identify a Suitable Medical Dataset**:
  The "Independent Medical Reviews" dataset was chosen.
- **Define the Analysis**:
  The objective was to analyze treatment and diagnosis categories.
- **Select and Implement an Algorithm**:
  Completed using Apache Pig on a Hadoop cluster.
- **Build a Dashboard**:
  A heatmap was created to visualize the data.

Problem and Data Collection

The primary challenge was to efficiently process and analyze large volumes of medical data to uncover insights into treatment and diagnosis trends. The "Independent Medical Reviews" dataset was utilized, sourced likely from a public health database or a healthcare provider.



Methodology and Implementation

- **Hadoop/MapReduce with Apache Pig**: The dataset was loaded and processed using Apache Pig scripts, which simplified the MapReduce complexities.
- **Hive for Data Storage**: A Hive table was created to store the processed data, allowing for efficient querying and management.

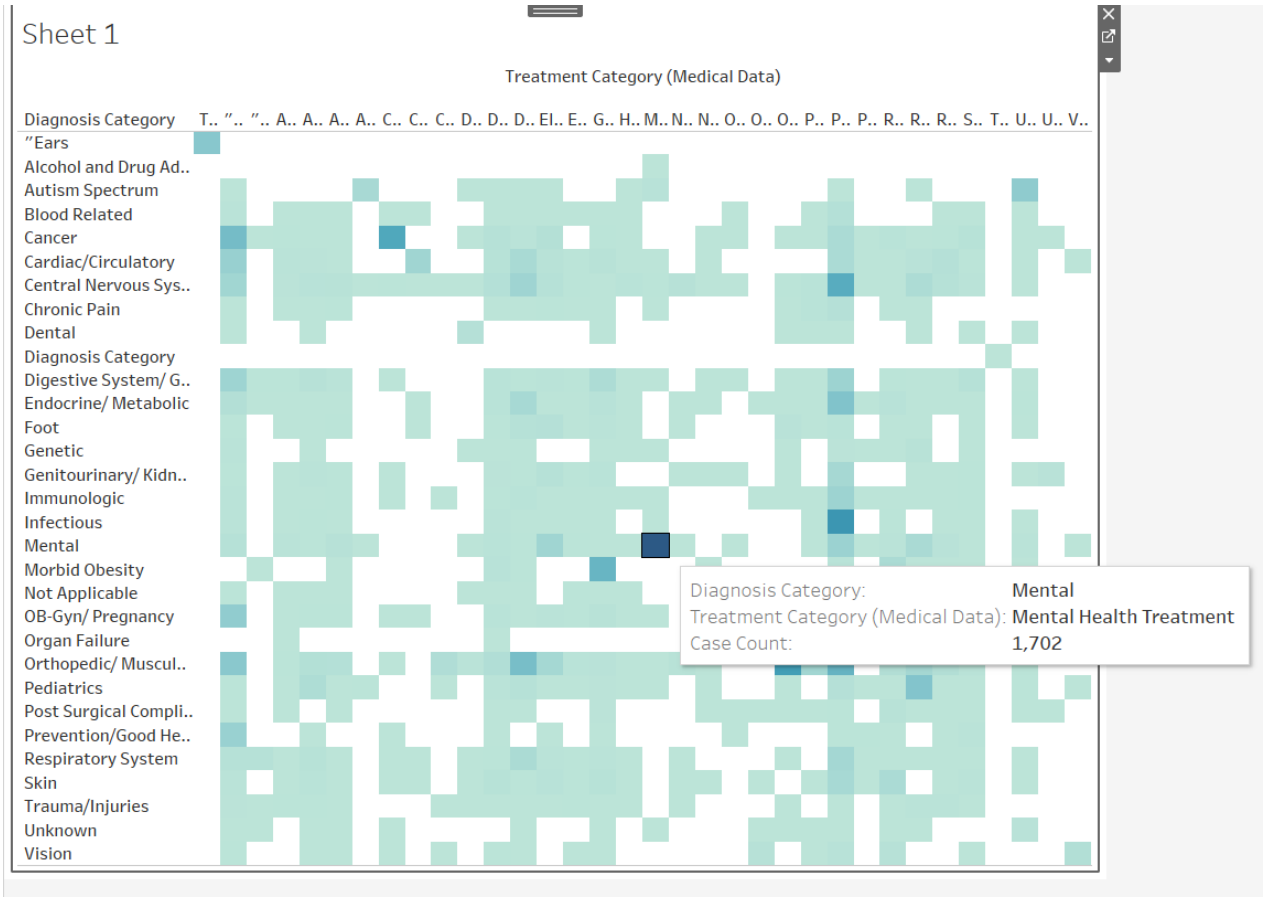
- **ODBC for Connectivity**: ODBC driver was used to connect the Hadoop data with Tableau for visualization.

```
-- Load the dataset
data = LOAD '/medical_reviews/Independent_Medical_Reviews_Filled.csv' USING PigStorage(',')
    AS (reference_id: chararray, report_year: int, diagnosis_category: chararray,
        diagnosis_sub_category: chararray, treatment_category: chararray,
        treatment_sub_category: chararray, determination: chararray,
        type: chararray, age_range: chararray, patient_gender: chararray, findings: chararray);

-- Group the data by treatment and diagnosis categories
grouped_data = GROUP data BY (treatment_category, diagnosis_category);

-- Count the number of records in each group and include field names
counts = FOREACH grouped_data GENERATE group.treatment_category AS treatment_category,
                                       group.diagnosis_category AS diagnosis_category,
                                       COUNT(data) AS record_count;

-- Store the result
STORE counts INTO '/output/treatment_diagnosis_correlation' USING PigStorage(',');
```



Suitability of the Tool

Hadoop/MapReduce, in conjunction with Apache Pig, proved to be highly suitable for handling large-scale data processing. The use of Hive facilitated easy data management, and ODBC provided seamless connectivity to visualization tools.

Software Features

- **Apache Pig for Data Processing**: Simplified scripting for complex data transformations and aggregations.
- **Hive for Data Warehousing**: Structured storage and query capabilities.
- **Tableau for Visualization**: Intuitive dashboard creation and advanced visualization options.

Worked Example

- **Data Loading**: Data was loaded into a Pig script for processing.
- **Pig Script Execution**: The script grouped data by treatment and diagnosis categories, calculating count per group.
- **Visualization in Tableau**: The output was visualized as a heatmap in Tableau, showing the frequency of medical cases across different categories.

Conclusion

The project successfully demonstrates the power of Hadoop/MapReduce for medical data analysis. The combination of Apache Pig for data processing, Hive for data storage, and Tableau for visualization enabled efficient handling and insightful analysis of complex medical data.

References

- Apache Hadoop Documentation.
- Apache Pig User Guide.
- Tableau User Guide.
- Cloudera Quickstart Docker Image Documentation.