Assignment-3: Pig and Hive

- Keshav Chandak(IMT2021003)
- Sunny Kaushik(IMT2021007)
- Muteeb Sheikh(IMT2021008)
- Rishi Nelapati(IMT2021076)

Question-1

Overview

This part focuses on designing and implementing data pipelines using Hive to efficiently analyze and clean educational datasets. The datasets include:

- `Course_Attendance.csv`
- `Enrollment_Data.csv`
- `GradeRosterReport.csv`

The primary tasks include defining schemas, creating Hive tables, loading data, and performing data cleaning operations using HiveQL.

Folder Structure

- **Assignment_3_NoSQL_PiG_Hive.pdf**: The assignment document detailing the tasks and requirements.
- **Course_Attendance.csv**: Contains raw data on course attendance.
- **Enrollment_Data_v7.csv**: Cleaned and processed enrollment data.
- **GradeRosterReport_v4.csv**: Cleaned and processed grade roster data.
- **create_and_load_tables.hql**: HiveQL script to define schemas, create tables, and load raw data.
- **data_cleaning.hql**: HiveQL script to clean and transform data.
- **readme.md**: Part (a) documentation (this file).

Steps and Scripts

1. Define Schemas and Create Tables

The `create_and_load_tables.hql` script defines the schema and creates Hive tables for each dataset:

Course Attendance Table

Schema:

- Course (STRING)
- Instructor (STRING)
- Name (STRING)

- Email_Id (STRING)
- Member_Id (STRING)
- Number_of_classes_attended (INT)
- Number_of_classes_absent (INT)
- Average_Attendance_Percentage (FLOAT)

Enrollment Data Table

Schema:

- Course_Type (STRING)
- Student_ID (STRING)
- Student_Name (STRING)
- Program (STRING)
- Batch (STRING)
- Period (STRING)
- Enrollment_Date (DATE)
- Primary_Faculty (STRING)
- Subject_Code_Name (STRING)
- Section (STRING)

Grade Roster Report Table

Schema:

- Academy_Location (STRING)
- Student_ID (STRING)
- Student_Status (STRING)
- Admission_ID (STRING)
- Admission_Status (STRING)
- Student_Name (STRING)
- Program_Name (STRING)
- Batch (STRING)
- Period (STRING)
- Subject_Code_Name (STRING)
- Section (STRING)
- Faculty_Name (STRING)
- Course_Credit (INT)
- Obtained_Marks_Grade (STRING)
- Out_of_Marks_Grade (STRING)
- Exam_Result (STRING)

---

2. Load Data into Hive Tables

The data from the CSV files is loaded into the corresponding Hive tables using the `LOAD DATA` command in the `create_and_load_tables.hql` script.

---

3. Data Cleaning

The `data_cleaning.hql` script performs the following cleaning operations:

- **Fill Missing Faculty Names**: Uses a self-join to fill in missing faculty names in `GradeRosterReport.csv`.
- **Remove Unnecessary Columns**: Drops unnecessary columns like `Serial No.`, `Status`, and `Academia+LMS` from `Enrollment_Data.csv`.
- **Update Program Name**: Extracts and updates the `Program Name` field from `Program Code/Name` in `GradeRosterReport.csv`.
- **Handle Multiple Faculty Entries**: Extracts a single, primary entry from the `Primary Faculty` column in `Enrollment_Data.csv`.

---

4. Final Output

The cleaned data is saved in:

- `Enrollment_Data_v7.csv`
- `GradeRosterReport_v4.csv`

These are ready for further analysis and reporting.

Usage

1. **Set Up Hive Environment**
   Ensure Apache Hive is properly installed and configured in your environment.

2. **Run Table Creation and Load Script**
   Execute `create_and_load_tables.hql` to define schemas and load the raw data.

   ```
   hive -f create_and_load_tables.hql
   ```

3. **Run Data Cleaning Script** Execute `data_cleaning.hql` to perform all cleaning operations.

   ```
   hive -f data_cleaning.hql
   ```