Mini-Project: Data Analysis and ETL with Public Data

Project Overview

The goal of this project is to demonstrate your ability to work with data from a public data source, transform it, and perform analysis using various SQL operations. You will create a simple data warehouse schema, perform ETL processes, and generate insights from the data.

Public Data Source

Use the New York City Taxi and Limousine Commission (TLC) Trip Record Data. Download the data for a single month (e.g., January 2023). Link: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Tasks

1. Data Extraction and Initial Exploration

- Download the TLC trip record data for January 2023.
- Load the data into Microsoft SQL Server or SQLite.
- Perform initial data exploration to understand the dataset. Document any observations about the data, such as data types, missing values, and distributions.

2. Schema Design

Design a star schema for a data warehouse with the following components:

- Fact Table: `trips_fact`
  - `trip_id` (Primary Key)
  - `pickup_datetime`
  - `dropoff_datetime`
  - `passenger_count`
  - `trip_distance`
  - `pickup_location_id`
  - `dropoff_location_id`
  - `fare_amount`
  - `tip_amount`
  - `total_amount`
- Dimension Tables:
  - `date_dim` (Date Dimension)

  - `date_id` (Primary Key)
  - `date`
  - `year`
  - `month`
  - `day`
  - `weekday`
- `location_dim` (Location Dimension)
  - `location_id` (Primary Key)
  - `borough`
  - `zone`
  - `service_zone`

3. Data Transformation and Loading (ETL)

- Write scripts to transform the raw data into the designed schema.
- Load the transformed data into the fact and dimension tables.
- Ensure that all foreign key relationships are maintained.

4. Data Analysis

Perform the following analyses using SQL queries:

- Calculate the total number of trips per day.
- Identify the top 5 zones with the highest total fare amount.
- Calculate the average trip distance by borough.
- Determine the most common pickup and dropoff locations.
- Calculate the total tip amount per passenger count.

5. Additional Task (Optional)

Create a simple dashboard (using a tool like Tableau, Power BI, or any open-source alternative) to visualize the results of your analyses.

Submission

Provide a GitHub repository with the following:
- Scripts for data extraction, transformation, and loading.
- SQL queries used for data analysis.
- Documentation of your process, including any assumptions made and challenges encountered.
- (Optional) Dashboard files and screenshots.

This project is designed to assess your ability to work with data, design efficient schemas, perform ETL processes, and extract meaningful insights from data. We look forward to seeing your work and understanding your approach to data engineering tasks.