1. Project Description

Overview

A fictional company wants to analyze data they collected on songs and user activity on their music streaming app. The data is stored in a directory containing JSON log files on user activity, and another directory with JSON meta data on the songs users play on their app. The analysis team would like to find out the songs that users listen to but finds it difficult to analyze the data in its current form.

Write an ETL pipeline to load the user activity data and song meta data contained in the JSON files to a PostgreSQL database using Python and SQL so that the data is in an analysis-ready format, which can be easily queried.

2. Tools

The tools used are:

- ETL tools – SQL, Python
- Data store – PostgreSQL database

3. Development Process

A star schema was designed to model the data contained in the JSON user activity files and JSON meta data files. In addition, three Python scripts were created implement the star schema on PostgreSQL database and load the data to the database. Refer to table below for filename of schema and scripts developed.

Table-1: Files uploaded

Filename	Description
Create_tables.py	the script was used to create the database on PostgrSQL
Sql_queries.py	used to create the SQL commands to create the tables and insert data into the tables
Etl.py	used to process the JSON files and load the data to the appropriate tables on PostgreSQL utilizing some of the commands in thesql_queriues.py file
Star_schema.pdf	Design of star schema

Procedure
Steps:

1. Create a new database and tables on PostgrSQL
2. Process the JSON files by extracting the relevant data and transform it on source system
3. Load the data to destination, that is, fact and dimension tables on PostgreSQL database

4. Testing

Testing consisted of the following items:

Table-2: Test

Test item	**Activities**	**Status**
Verify database and tables created on PostgreSQL	Open pgAdmin and check that database and tables exist	Verified
Verify desired data exists in tables	Execute SQL query “SELECT *” on all tables created	Verified