Introduction

A music streaming startup, Sparkify, has grown their user base and song database and want to move their processes and data onto the cloud. Their data resides in S3, in a directory of JSON logs on user activity on the app, as well as a directory with JSON metadata on the songs in their app.

In this project, it is to build an ETL pipeline that extracts their data from S3, stages them in Redshift, and transforms data into a set of dimensional tables of star schema for analytics team to continue finding insights in what songs users are listening to.

File Introduction

1. etl.py It is to stage and create tables of schema design;
2. sql_queries.py All SQL Queries that are used to create, drop, insert and staging tables;
3. dwh.cfg All configurayion information of the AWS Account and use this file to connect with Redshift

Major Chanllenges and Solutions

- In original files, timestamp looks like

ts
1541990217796
1541990258796
1541990264796
1541990541796
1541990714796

We need to transform it into the common form. When I load files, I load timeformat AS 'epochmillisecs';

- Fact table: When loading fact tables, not all dimentions could befound in one table, which is different from other dimensional tables.
  Need to find out common columns and join two tables together;
- Primary and Serial Key. In Redshift, Primary key is not required and for serial keys, it should be created as int identity(0,1);
- Null Values. When staging data in Redshift, if we don't specify the path, files could also be loaded but some columns are all null values. If you specify file path, then all values could be loaded. It is little bit strange to me...