# Sparkify S3 to Redshift ETL Project

1. Purpose of the database

The purpose of this database is to help Sparkify, a music streaming startup, be able to do analytical processing on the data that they have collected and are continuing to collect. This database has data on their users, songs and what songs the users are listening to. The database is organized into tables that form a schema that will help the data analysts to easily perform SQL statements on the database.

2. Database schema design and ETL pipeline

The database schema design is in the form of a star schema with a fact table and 4 dimension tables. The ETL pipeline is taking files located in AWS S3 in the log_data and song_data folders and transfering the data in these files to staging tables located in the AWS redshift cluster. Once the staging tables have been loaded with all the data in S3 they are then loaded into the fact and dimension tables of the Star Schema. Once the Star Schema Database is loaded with data it is then ready to be queried with SQL statements. The use of the star schema design is to enable fast query performance and it is easy to understand from the perspective of business users.

3. How to run the python scripts.

Using jupyter open a new workbook and type '%run create_tables.py' to create the staging tables as well as the tables in the database. Once that has completed type '%run etl.py' to extract data from the data files on S3 and transfer into the staging tables, which will then transform and transfer the data from the staging tables to the tables in the database.

4. Description of the files in S3

The files located in S3 are the log data and song data. The log data contains data related to the consumption of songs that have been heard by the users. The song data contains data related to the songs that are available for listening. All of the files are in JSON format.