**Project: Sparkify Data Pipelines**

**Introduction**

A music streaming company, Sparkify, has decided that it is time to introduce more monitoring to their data warehouse ETL pipelines.

They expect me to create high grade data pipelines that are dynamic and built from reusable tasks. The source data resides in S3 and needs to be processed in Sparkify's data warehouse in Amazon Redshift. The source datasets consist of JSON logs that tell about user activity in the application and JSON metadata about the songs the users listen to.

This problem entails loading data from S3 into staging tables extracting the data into a Star Schema.

## Table of Contents

## [Installation](#_Installation)

## [Prerequisites](#_Prerequisites)

## [Files](#_Files)

## [Screenshot](#_Screenshot)

## [Instructions](#_Instructions)

## [Running the files](#_Running_the_files)

## [Authors](#_Authors)

## [Acknowledgement](#_Acknowledgement)

**Installation**

There are two ways to run the attached Project scripts.

1. Locally

* You should have Jupyter Notebook installed on your machine
* Python and PySpark API must be installed as well

1. In the Cloud

* You should have an Amazon AWS account

## Prerequisites

You need to create an AWS Redshift cluster and have the Amazon AIM credentials with access to s3 bucket in AWS. The Redshift connection and AWS connection will be setup using DWH configs. Information on where the data resides in AWS S3 is configured using the ‘AWS Redshift Cluster Setup.ipynb’ and save the configuration into the file ‘dwh.cfg’. Once the cluster is set up and the dwh.cfg is saved, you can run the sql scripts from within the create tables.py and etl.py.

## Files

Following files are attached to this repository:

* AWS Redshift Cluster Setup.ipynb – Jupyter Notebook to set up a cluster
* AWS project dwh1.cfg – configuration file to save access key and secret access key run the ‘AWS Redshift Cluster Setup.ipynb’
* dwh.cfg- configuration file to save cluster, IAM Role, S# information
* sql\_queries.py – python to create tables, stage data into table, and insert data tables
* create\_tables.py – python to create tables
* etl.py – python to extract data and insert it into tables
* star diagram.pdf – star schema of all fact and star tables created

## Screenshot

## 

## Instructions

Run the create\_tables.py then the etl.py.

If you are this far, congrats! 🎉 The setup is done. Now you can run the dag 👏

**Running the files**

Once the etl.py is completed you can use the tables to run queries to analyze the dataset.

**Authors**

* **Devora Lewenstein**

**Acknowledgments**

* Project inspired by Livsha Klingman