Sparkify

1. Project Prompt:
   1. Music streaming startup Sparkify has grown their user base and song database.
   2. The execs want to move their processes and data onto the cloud.
   3. Their data resides in S3, in a directory of json logs on user activity on the app and a json directory for song metadata on the songs in their app.
   4. As a Data Engineer you need to build an ETL pipeline that extracts their data from the S3 bucket, and stages the data in Redshift which will then be transformed and stored in the correct fact and dimension tables.
   5. This allows the analytics team to build insights for the songs available on their platform.
2. To Complete:
   1. Load the data from S3 to staging tables on Redshift.
   2. Execute SQL commands that build the tables for analysis from staging tables.
3. Datasets:
   1. Song Dataset:
      1. The first dataset is a subset of real data from the [Million Song Dataset](http://millionsongdataset.com/). Each file is in JSON format and contains metadata about a song and the artist of that song. The files are partitioned by the first three letters of each song's track ID. For example, here are file paths to two files in this dataset.
   2. Log Dataset:
      1. The second dataset consists of log files in JSON format generated by this [event simulator(opens in a new tab)](https://github.com/Interana/eventsim) based on the songs in the dataset above. These simulate app activity logs from an imaginary music streaming app based on configuration settings.