Project Summary: Data Pipelines with Airflow

In this project I will be using Apache Airflow, python and SQL to automate the ETL process so that as new data files arrive in S3 they can be processed based on a schedule or a trigger and loaded into AWS Redshift (the data warehouse). There are 4 operators used in the workflow: StageToRedshiftOperator (stage_redshift.py), LoadFactOperator (load_fact.py), LoadDimensionOperator (load_dimension.py), and DataQualityOperator (data_quality.py). Each operator utilizes a python script to execute tasks within each step in the workflow. The entire workflow is contained in a python script called final_project.py where the DAG resides. This python script with the DAG is then loaded into the Airflow server and scheduler where it can then be activated and scheduled.

DAG Workflow in Airflow