Data Pipelines with Airflow

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

General

DAG can be browsed without issues in the Airflow UI

The dag follows the data flow provided in the instructions, all the tasks have a dependency and DAG begins with a start_execution task and ends with a end_execution task.

Dag configuration

DAG contains default_args dict, with the following keys:

- Owner
- Depends_on_past
- Start_date
- Retries
- Retry_delay
- Catchup

The DAG object has default args set

The DAG should be scheduled to run once an hour

Staging the data

There is a task that to stages data from S3 to Redshift. (Runs a Redshift copy statement)

Instead of running a static SQL statement to stage the data, the task uses params to generate the copy statement dynamically

The operator contains logging in different steps of the execution

The SQL statements are executed by using a Airflow hook

Loading dimensions and facts

Dimensions are loaded with on the LoadDimension operator

Facts are loaded with on the LoadFact operator

Instead of running a static SQL statement to stage the data, the task uses params to generate the copy statement dynamically

The DAG allows to switch between append-only and delete-load functionality

Data Quality Checks

Data quality check is done with correct operator

The DAG either fails or retries n times

Operator uses params to get the tests and the results, tests are not hard coded to the operator