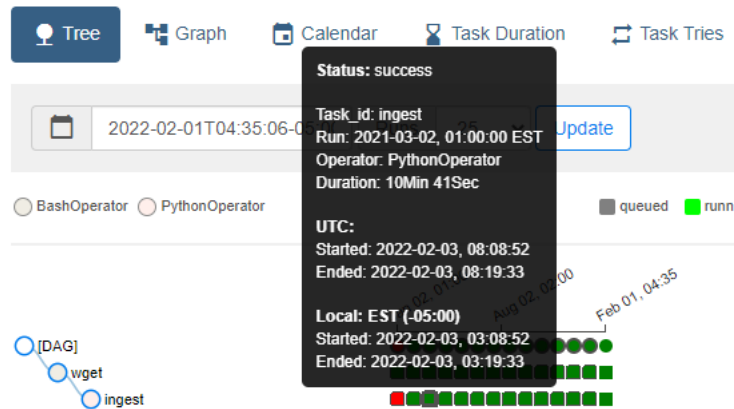


LOADING DATA FASTER USING POSTGRES COPY FROM FEATURE – Takes 30 sec/file instead of 8mins

FEB 3rd – 8mins to load data using reading csv lines

DAG: Wk2_LocalIngestionDag



```
[2022-02-03, 03:08:52 EST] {taskinstance.py:1426} INFO - Exporting the following env vars:
AIRFLOW_CTX_DAG_OWNER=***
AIRFLOW_CTX_DAG_ID=Wk2_LocalIngestionDag
AIRFLOW_CTX_TASK_ID=ingest
AIRFLOW_CTX_EXECUTION_DATE=2021-03-02T06:00:00+00:00
AIRFLOW_CTX_DAG_RUN_ID=scheduled_2021-03-02T06:00:00+00:00
[2022-02-03, 03:08:52 EST] {logging_mixin.py:109} INFO - yellow_taxi_2021_03 /opt/***/output_2021-03.csv
[2022-02-03, 03:08:52 EST] {logging_mixin.py:109} INFO - Connection Successful
[2022-02-03, 03:08:52 EST] {logging_mixin.py:109} INFO - If table exists, drop successful
[2022-02-03, 03:09:27 EST] {logging_mixin.py:109} INFO - Appending
[2022-02-03, 03:10:15 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 48.098 second
[2022-02-03, 03:10:58 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 42.277 second
[2022-02-03, 03:11:29 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 31.413 second
[2022-02-03, 03:12:02 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 32.586 second
[2022-02-03, 03:12:34 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 32.016 second
[2022-02-03, 03:13:04 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 30.846 second
[2022-02-03, 03:13:43 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 39.033 second
[2022-02-03, 03:14:20 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 36.431 second
[2022-02-03, 03:14:54 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 33.974 second
[2022-02-03, 03:15:24 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 30.554 second
[2022-02-03, 03:15:53 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 28.670 second
[2022-02-03, 03:16:24 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 30.759 second
[2022-02-03, 03:16:55 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 31.167 second
[2022-02-03, 03:17:33 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 37.547 second
[2022-02-03, 03:18:04 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 31.576 second
[2022-02-03, 03:18:32 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 28.127 second
[2022-02-03, 03:18:32 EST] {warnings.py:110} WARNING - /home/***/.local/lib/python3.7/site-packages/***/operators/python.py:174: DtypeWarning: Columns (6)
return_value = self.execute_callable()
[2022-02-03, 03:19:01 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 29.061 second
[2022-02-03, 03:19:27 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 25.192 second
[2022-02-03, 03:19:33 EST] {logging_mixin.py:109} INFO - inserted another chunk, took 6.766 second
[2022-02-03, 03:19:33 EST] {logging_mixin.py:109} INFO - completed
[2022-02-03, 03:19:33 EST] {python.py:175} INFO - Done. Returned value was: None
[2022-02-03, 03:19:33 EST] {taskinstance.py:1277} INFO - Marking task as SUCCESS. dag_id=Wk2_LocalIngestionDag, task_id=ingest, execution_date=20210302T0600
[2022-02-03, 03:19:33 EST] {local_task_job.py:154} INFO - Task exited with return code 0
[2022-02-03, 03:19:33 EST] {local_task_job.py:264} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

Vk2_L_IngestionDag

Graph

2021-01-04

PythonOperator

Status: success

Run: 2021-03-02, 01:00:00 EST

Run Id: scheduled__2021-03-02T06:00:00+00:00

Duration: 1Min 59Sec

Data Interval:

Start: 2021-03-02, 01:00:00 EST

End: 2021-04-02, 02:00:00 EDT

UTC:

Started: 2022-02-05, 13:50:30

Ended: 2022-02-05, 13:52:30

Local: EST (-05:00)

Started: 2022-02-05, 08:50:30

Ended: 2022-02-05, 08:52:30

Aug 02, 02:00
Feb 01, 04:35

```
*** Reading local file: /opt/airflow/logs/Wk2_LocalIngestionDag/ingest/2021-03-02T06:00:00+00:00/8.log
[2022-02-05, 08:51:59 EST] {taskinstance.py:1032} INFO - Dependencies all met for <TaskInstance: Wk2_LocalIngestionDag.ingest scheduled__2021-03-02T06:00:00+00:00 [q
[2022-02-05, 08:51:59 EST] {taskinstance.py:1032} INFO - Dependencies all met for <TaskInstance: Wk2_LocalIngestionDag.ingest scheduled__2021-03-02T06:00:00+00:00 [q
[2022-02-05, 08:51:59 EST] {taskinstance.py:1238} INFO -
-----
[2022-02-05, 08:51:59 EST] {taskinstance.py:1239} INFO - Starting attempt 8 of 8
[2022-02-05, 08:51:59 EST] {taskinstance.py:1240} INFO -
-----
[2022-02-05, 08:51:59 EST] {taskinstance.py:1259} INFO - Executing <Task(PythonOperator): ingest> on 2021-03-02 06:00:00+00:00
[2022-02-05, 08:51:59 EST] {standard_task_runner.py:52} INFO - Started process 31288 to run task
[2022-02-05, 08:51:59 EST] {standard_task_runner.py:76} INFO - Running: ['***', 'tasks', 'run', 'Wk2_LocalIngestionDag', 'ingest', 'scheduled__2021-03-02T06:00:00+00
[2022-02-05, 08:51:59 EST] {standard_task_runner.py:77} INFO - Job 78: Subtask ingest
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Running <TaskInstance: Wk2_LocalIngestionDag.ingest scheduled__2021-03-02T06:00:00+00:00 [running]> on host
[2022-02-05, 08:51:59 EST] {warnings.py:110} WARNING - /home/***/.local/lib/python3.7/site-packages/***/utils/context.py:152: AirflowContextDeprecationWarning: Acces
warnings.warn(_create_deprecation_warning(key, self._deprecation_replacements[key]))

[2022-02-05, 08:51:59 EST] {taskinstance.py:1426} INFO - Exporting the following env vars:
AIRFLOW_CTX_DAG_OWNER=***
AIRFLOW_CTX_DAG_ID=Wk2_LocalIngestionDag
AIRFLOW_CTX_TASK_ID=ingest
AIRFLOW_CTX_EXECUTION_DATE=2021-03-02T06:00:00+00:00
AIRFLOW_CTX_DAG_RUN_ID=scheduled__2021-03-02T06:00:00+00:00
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - yellow_taxi_2021_03 /opt/***/output_2021-03.csv
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Connection Successful
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - If table exists, dropping and recreating
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Table creation completed
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Connection Closed
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Connection Successful with psycopg2
[2022-02-05, 08:51:59 EST] {logging_mixin.py:109} INFO - Loading data from file..
[2022-02-05, 08:52:30 EST] {logging_mixin.py:109} INFO - Completed loading data from file
[2022-02-05, 08:52:30 EST] {logging_mixin.py:109} INFO - Total time taken - 30.695 second
[2022-02-05, 08:52:30 EST] {python.py:175} INFO - Done. Returned value was: None
[2022-02-05, 08:52:30 EST] {taskinstance.py:1277} INFO - Marking task as SUCCESS. dag_id=Wk2_LocalIngestionDag, task_id=ingest, execution_date=20210302T060000, start
[2022-02-05, 08:52:30 EST] {local_task_job.py:154} INFO - Task exited with return code 0
[2022-02-05, 08:52:30 EST] {local_task_job.py:264} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

30 sec

