PIPELINES

[Jobaoushadancse2021@jerusalemengg.ac.in](mailto:Jobaoushadancse2021@jerusalemengg.ac.in)

Job Aoushadan N – 19.08.2025

To run our pipeline, we need a working Airflow environment. Docker Compose makes this easy and safe – no system-wide installs required. Just open your terminal and run the following:

*# Download the docker-compose.yaml file*

curl -LfO 'https://airflow.apache.org/docs/apache-airflow/stable/docker-compose.yaml'

*# Make expected directories and set an expected environment variable*

mkdir -p ./dags ./logs ./plugins

echo -e "AIRFLOW\_UID=**$(**id -u**)**" > .env

*# Initialize the database*

docker compose up airflow-init

*# Start up all services*

docker compose up

Once Airflow is up and running, visit the UI at http://localhost:8080.

Log in with:

* **Username:** airflow
* **Password:** airflow

You’ll land in the Airflow dashboard, where you can trigger DAGs, explore logs, and manage your environment.

## Create a Postgres Connection

Before our pipeline can write to Postgres, we need to tell Airflow how to connect to it. In the UI, open the **Admin > Connections** page and click the + button to add a new [connection](https://airflow.apache.org/docs/apache-airflow/stable/concepts/connections.html).

Fill in the following details:

* Connection ID: tutorial\_pg\_conn
* Connection Type: postgres
* Host: postgres
* Database: airflow (this is the default database in our container)
* Login: airflow
* Password: airflow
* Port: 5432

Save the connection. This tells Airflow how to reach the Postgres database running in your Docker environment.

Next, we’ll start building the pipeline that uses this connection.

## Create tables for staging and final data

Let’s begin with table creation. We’ll create two tables:

* employees\_temp: a staging table used for raw data
* employees: the cleaned and deduplicated destination

We’ll use the SQLExecuteQueryOperator to run the SQL statements needed to create these tables.

**from** **airflow.providers.common.sql.operators.sql** **import** SQLExecuteQueryOperator

create\_employees\_table = SQLExecuteQueryOperator(

task\_id="create\_employees\_table",

conn\_id="tutorial\_pg\_conn",

sql="""

CREATE TABLE IF NOT EXISTS employees (

"Serial Number" NUMERIC PRIMARY KEY,

"Company Name" TEXT,

"Employee Markme" TEXT,

"Description" TEXT,

"Leave" INTEGER

);""",

)

create\_employees\_temp\_table = SQLExecuteQueryOperator(

task\_id="create\_employees\_temp\_table",

conn\_id="tutorial\_pg\_conn",

sql="""

DROP TABLE IF EXISTS employees\_temp;

CREATE TABLE employees\_temp (

"Serial Number" NUMERIC PRIMARY KEY,

"Company Name" TEXT,

"Employee Markme" TEXT,

"Description" TEXT,

"Leave" INTEGER

);""",

)

You can optionally place these SQL statements in .sql files inside your dags/ folder and pass the file path to the sql= argument. This can be a great way to keep your DAG code clean.

## Load data into the staging table

Next, we’ll download a CSV file, save it locally, and load it into employees\_temp using the PostgresHook.

**import** **os**

**import** **requests**

**from** **airflow.sdk** **import** task

**from** **airflow.providers.postgres.hooks.postgres** **import** PostgresHook

@task

**def** get\_data():

*# NOTE: configure this as appropriate for your airflow environment*

data\_path = "/opt/airflow/dags/files/employees.csv"

os.makedirs(os.path.dirname(data\_path), exist\_ok=**True**)

url = "https://raw.githubusercontent.com/apache/airflow/main/airflow-core/docs/tutorial/pipeline\_example.csv"

response = requests.request("GET", url)

**with** open(data\_path, "w") **as** file:

file.write(response.text)

postgres\_hook = PostgresHook(postgres\_conn\_id="tutorial\_pg\_conn")

conn = postgres\_hook.get\_conn()

cur = conn.cursor()

**with** open(data\_path, "r") **as** file:

cur.copy\_expert(

"COPY employees\_temp FROM STDIN WITH CSV HEADER DELIMITER AS ',' QUOTE '**\"**'",

file,

)

conn.commit()

This task gives you a taste of combining Airflow with native Python and SQL hooks – a common pattern in real-world pipelines.

## Merge and clean the data

Now let’s deduplicate the data and merge it into our final table. We’ll write a task that runs a SQL INSERT … ON CONFLICT DO UPDATE.

**from** **airflow.sdk** **import** task

**from** **airflow.providers.postgres.hooks.postgres** **import** PostgresHook

@task

**def** merge\_data():

query = """

INSERT INTO employees

SELECT \*

FROM (

SELECT DISTINCT \*

FROM employees\_temp

) t

ON CONFLICT ("Serial Number") DO UPDATE

SET

"Employee Markme" = excluded."Employee Markme",

"Description" = excluded."Description",

"Leave" = excluded."Leave";

"""

**try**:

postgres\_hook = PostgresHook(postgres\_conn\_id="tutorial\_pg\_conn")

conn = postgres\_hook.get\_conn()

cur = conn.cursor()

cur.execute(query)

conn.commit()

**return** 0

**except** **Exception** **as** e:

**return** 1

## Defining the DAG

Now that we’ve defined all our tasks, it’s time to put them together into a DAG.

**import** **datetime**

**import** **pendulum**

**import** **os**

**import** **requests**

**from** **airflow.sdk** **import** dag, task

**from** **airflow.providers.postgres.hooks.postgres** **import** PostgresHook

**from** **airflow.providers.common.sql.operators.sql** **import** SQLExecuteQueryOperator

@dag(

dag\_id="process\_employees",

schedule="0 0 \* \* \*",

start\_date=pendulum.datetime(2021, 1, 1, tz="UTC"),

catchup=**False**,

dagrun\_timeout=datetime.timedelta(minutes=60),

)

**def** ProcessEmployees():

create\_employees\_table = SQLExecuteQueryOperator(

task\_id="create\_employees\_table",

conn\_id="tutorial\_pg\_conn",

sql="""

CREATE TABLE IF NOT EXISTS employees (

"Serial Number" NUMERIC PRIMARY KEY,

"Company Name" TEXT,

"Employee Markme" TEXT,

"Description" TEXT,

"Leave" INTEGER

);""",

)

create\_employees\_temp\_table = SQLExecuteQueryOperator(

task\_id="create\_employees\_temp\_table",

conn\_id="tutorial\_pg\_conn",

sql="""

DROP TABLE IF EXISTS employees\_temp;

CREATE TABLE employees\_temp (

"Serial Number" NUMERIC PRIMARY KEY,

"Company Name" TEXT,

"Employee Markme" TEXT,

"Description" TEXT,

"Leave" INTEGER

);""",

)

@task

**def** get\_data():

*# NOTE: configure this as appropriate for your airflow environment*

data\_path = "/opt/airflow/dags/files/employees.csv"

os.makedirs(os.path.dirname(data\_path), exist\_ok=**True**)

url = "https://raw.githubusercontent.com/apache/airflow/main/airflow-core/docs/tutorial/pipeline\_example.csv"

response = requests.request("GET", url)

**with** open(data\_path, "w") **as** file:

file.write(response.text)

postgres\_hook = PostgresHook(postgres\_conn\_id="tutorial\_pg\_conn")

conn = postgres\_hook.get\_conn()

cur = conn.cursor()

**with** open(data\_path, "r") **as** file:

cur.copy\_expert(

"COPY employees\_temp FROM STDIN WITH CSV HEADER DELIMITER AS ',' QUOTE '**\"**'",

file,

)

conn.commit()

@task

**def** merge\_data():

query = """

INSERT INTO employees

SELECT \*

FROM (

SELECT DISTINCT \*

FROM employees\_temp

) t

ON CONFLICT ("Serial Number") DO UPDATE

SET

"Employee Markme" = excluded."Employee Markme",

"Description" = excluded."Description",

"Leave" = excluded."Leave";

"""

**try**:

postgres\_hook = PostgresHook(postgres\_conn\_id="tutorial\_pg\_conn")

conn = postgres\_hook.get\_conn()

cur = conn.cursor()

cur.execute(query)

conn.commit()

**return** 0

**except** **Exception** **as** e:

**return** 1

[create\_employees\_table, create\_employees\_temp\_table] >> get\_data() >> merge\_data()

dag = ProcessEmployees()