ELT – Batch Processing

Airflow

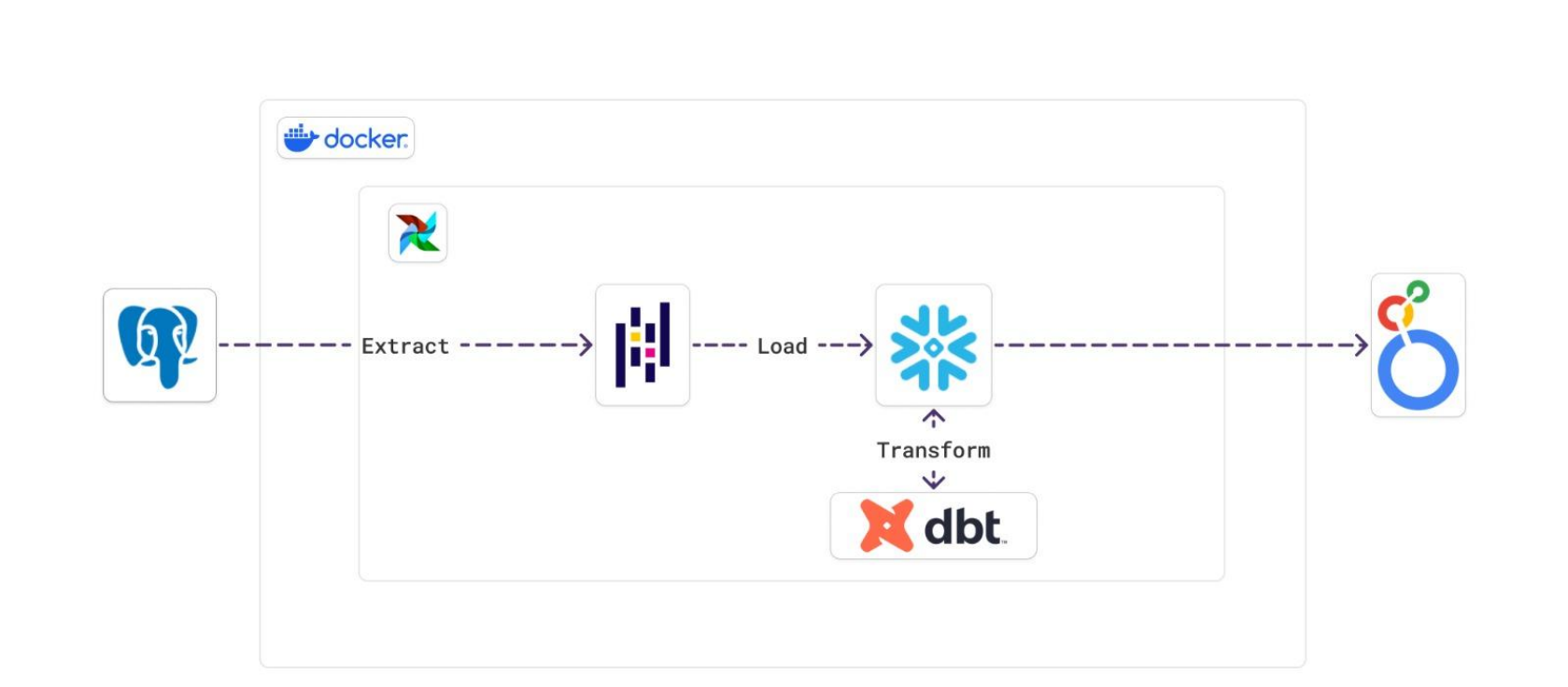
Overview

This project is an improvised result of project 2 where in project 3 we improvise to do scheduling with Airflow.

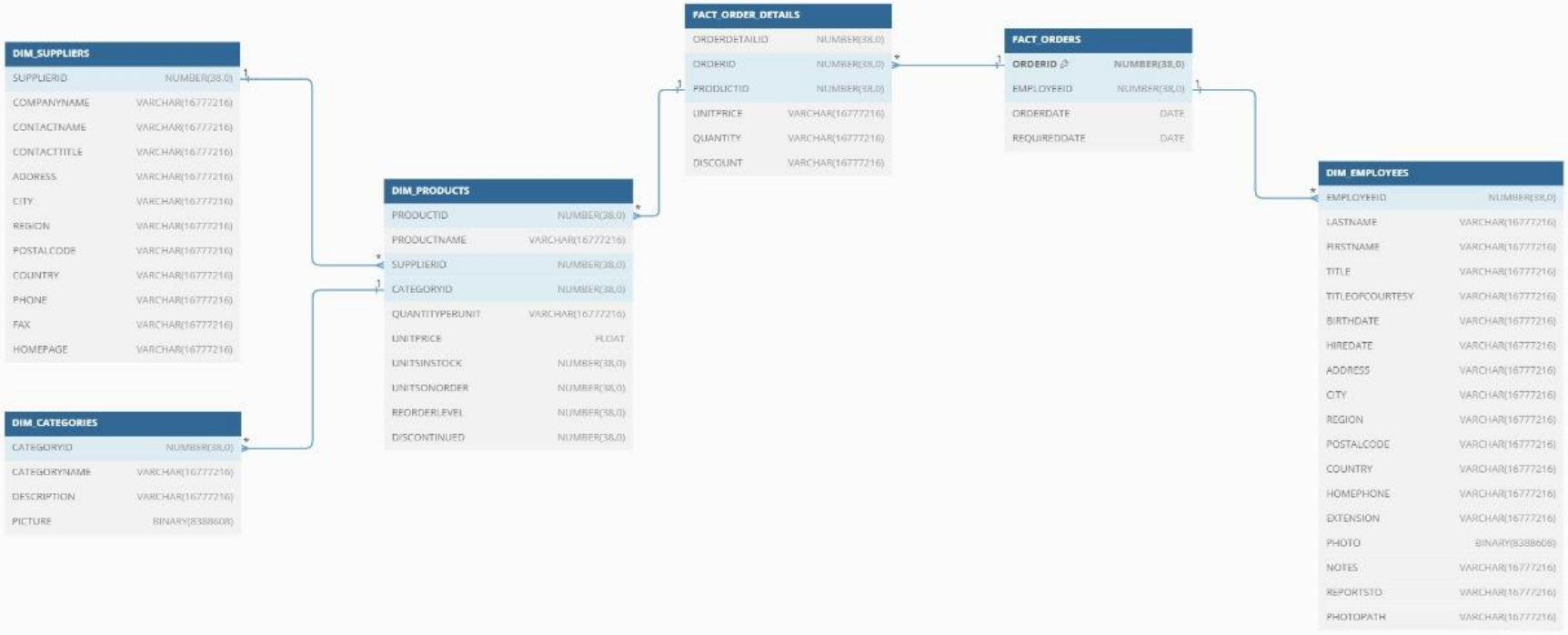
Objective

The objective of this project is to perform scheduling that runs at 12pm every day.

ARCHITECTURE



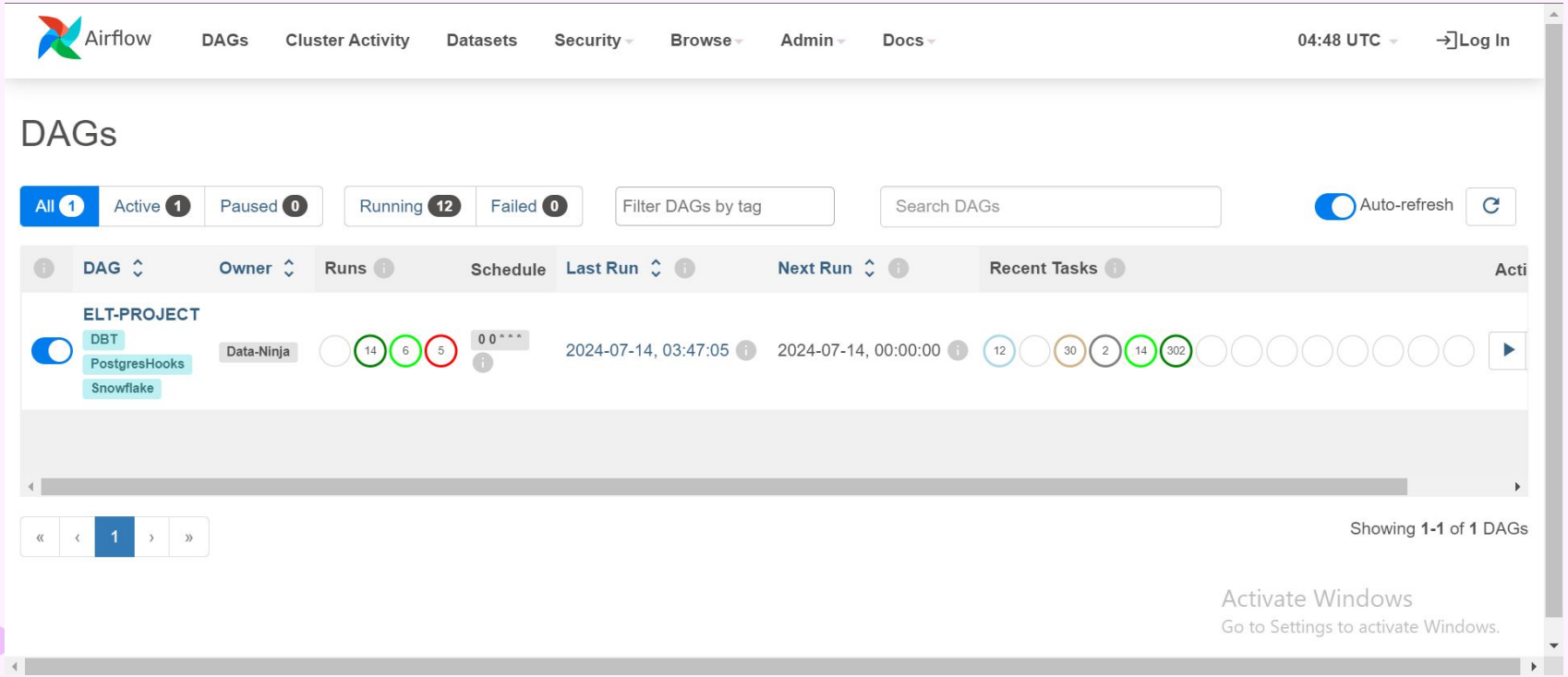
Data Warehouse Modeling



DAG Scripts

```python
from datetime import datetime, timedelta
from airflow.decorators import dag, task, task_group
from airflow.operators.dummy_operator import DummyOperator
from airflow.sensors.time_delta import TimeDeltaSensor
from cosmos import DbtTaskGroup
from dbt.dbt_project.dbt_config import (
    profile_config,
    project_config,
    execution_config,
    render_config
)
from el_module.extractors import extract_func
from el_module.loaders import load_func
from el_module.filenames import names
import os

# Define Date & Run Schedule
startDate = datetime(2024, 6, 27)
schedule = '0 0 * * *'


# DAG Configuration
default_args = {
    "owner": "Data-Ninja",
    "depends_on_past": False,
    "start_date": startDate,
    "retries":1,
    "retry_delay": timedelta(minutes=5),
}

@dag(
    dag_id="ELT-PROJECT",
    default_args=default_args,
    schedule_interval=schedule,
    catchup=True,
    fail_stop=True,
    description="ELT Orchestrator",
    tags=['PostgresHooks', 'Snowflake', 'DBT'],
    )
def elt_dag():
```

Scheduling – Airflow

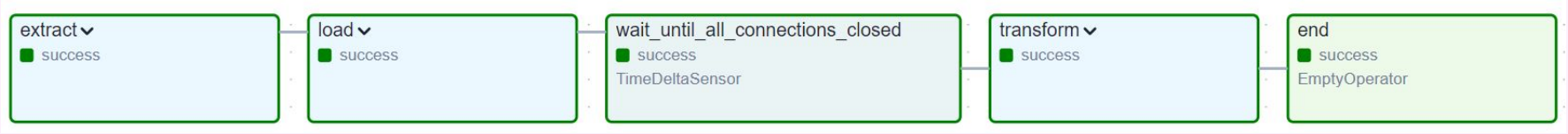


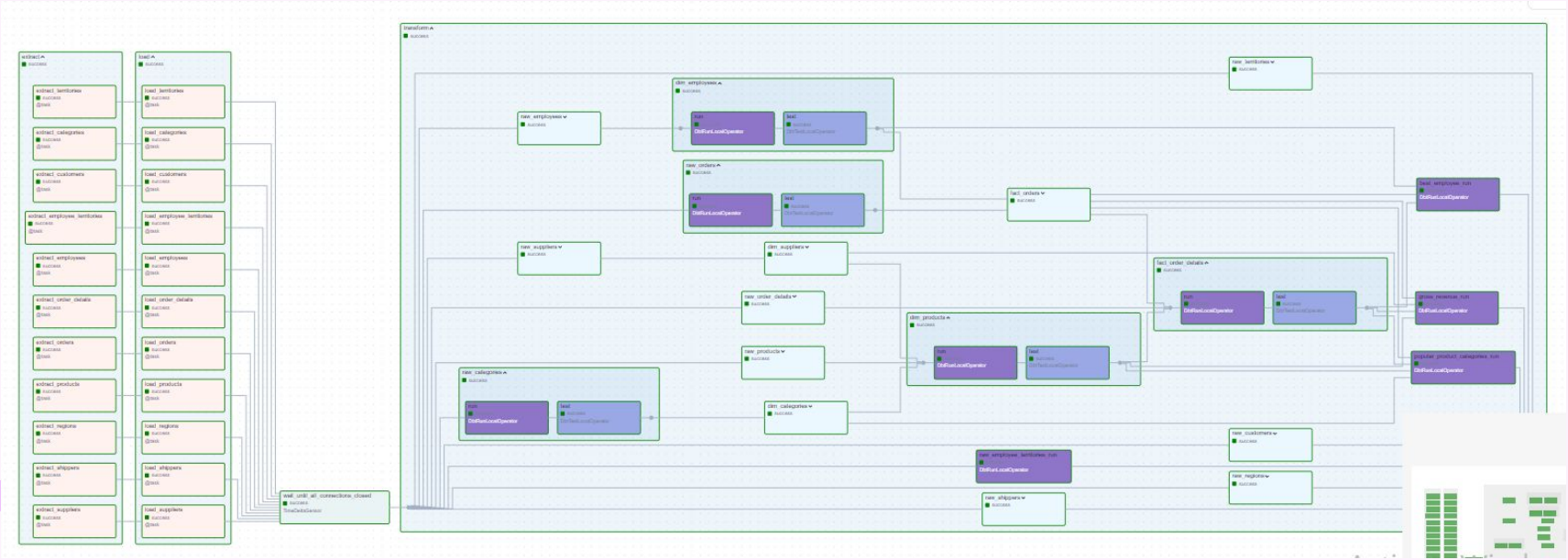
Task Group

```python
@task_group(group_id="extract")
def extract_data():
    tg_extract = []

    # Generate Tasks
    for gen_name in names:
        @task(task_id=f'extract_{gen_name}')
        def extract(name):
            return extract_func(
                table_name=name
            )

        # Populate Task Group
        tg_extract.append(
            extract(name=gen_name)
        )

    return tg_extract

@task_group(group_id="load")
def load_data(dfs):
    tg_load = []

    # Generate Tasks
    for i, gen_name in enumerate(names):
        @task(task_id=f'load_{gen_name}')
        def load(df, table_name):
            load_func(
                df=df, filename=table_name
            )

        # Populate Task Group
        tg_load.append(
            load(df=dfs[i], table_name=gen_name)
        )

    return tg_load
```

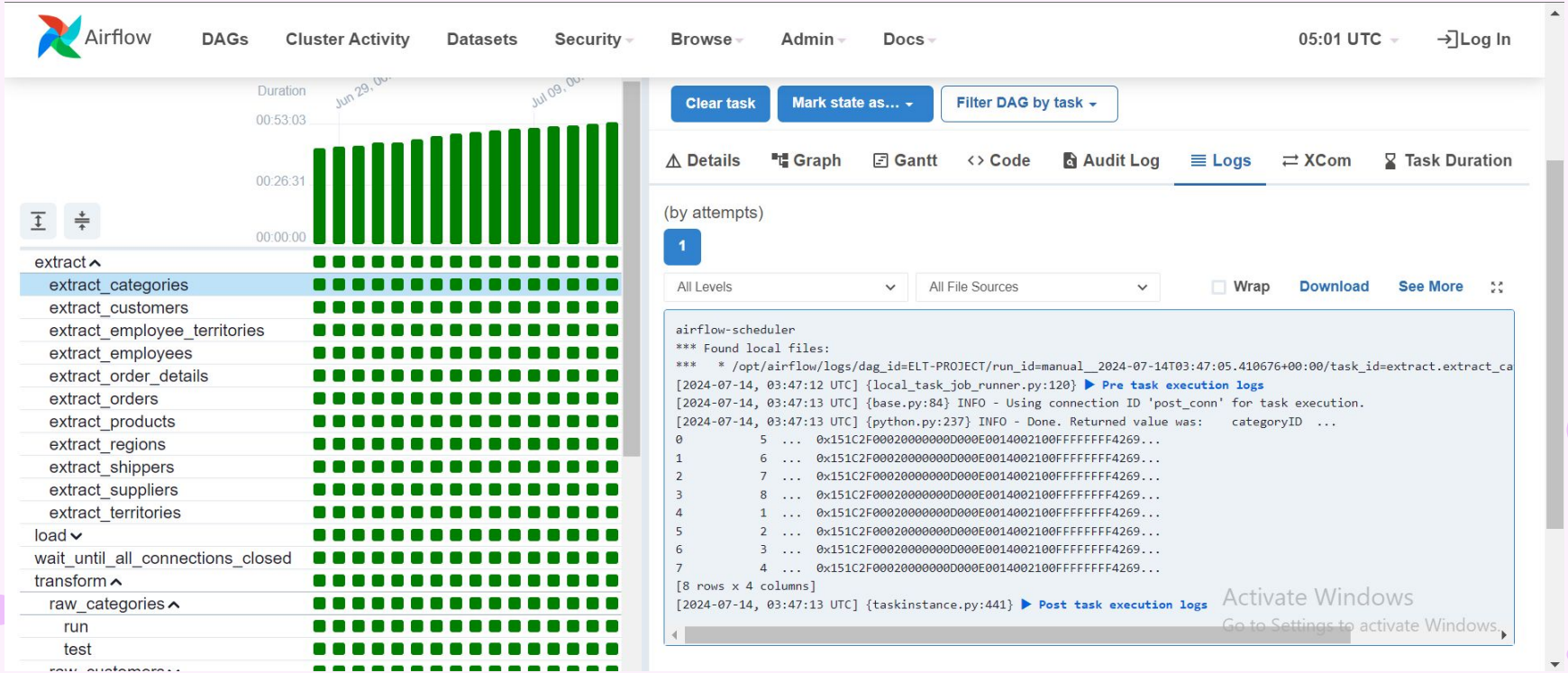
```python
wait = TimeDeltaSensor(
    task_id="wait_until_all_connections_closed",
    delta=timedelta(seconds=10)
)

transform_data = DbtTaskGroup(
    group_id="transform",
    project_config=project_config,
    profile_config=profile_config,
    execution_config=execution_config,
    render_config=render_config,
    operator_args={
        "install_deps": True
    }
)

end = DummyOperator(
    task_id="end"
)

# Task Flow: Mixing Operator
extract = extract_data()
load_data(dfs=extract) >> wait
wait >> transform_data >> end
```

Scheduling – Airflow

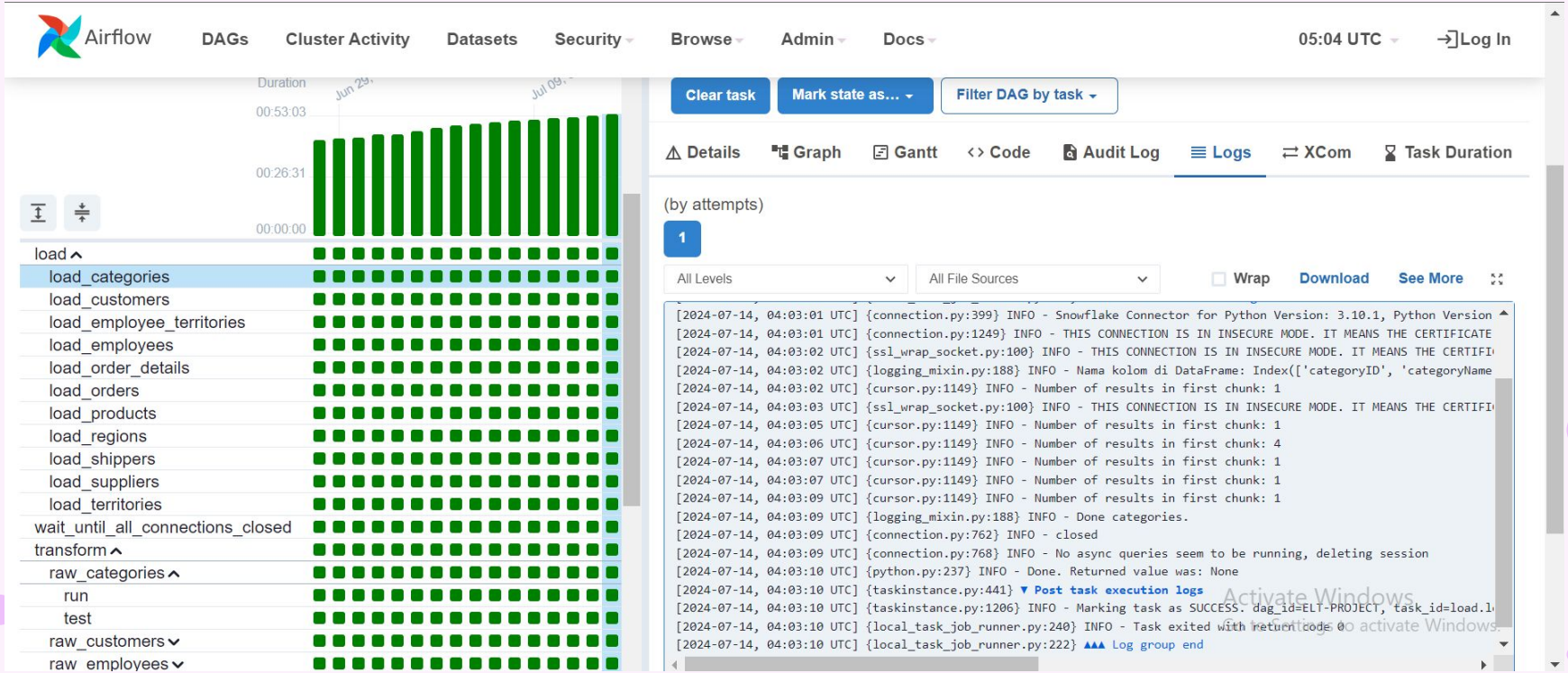




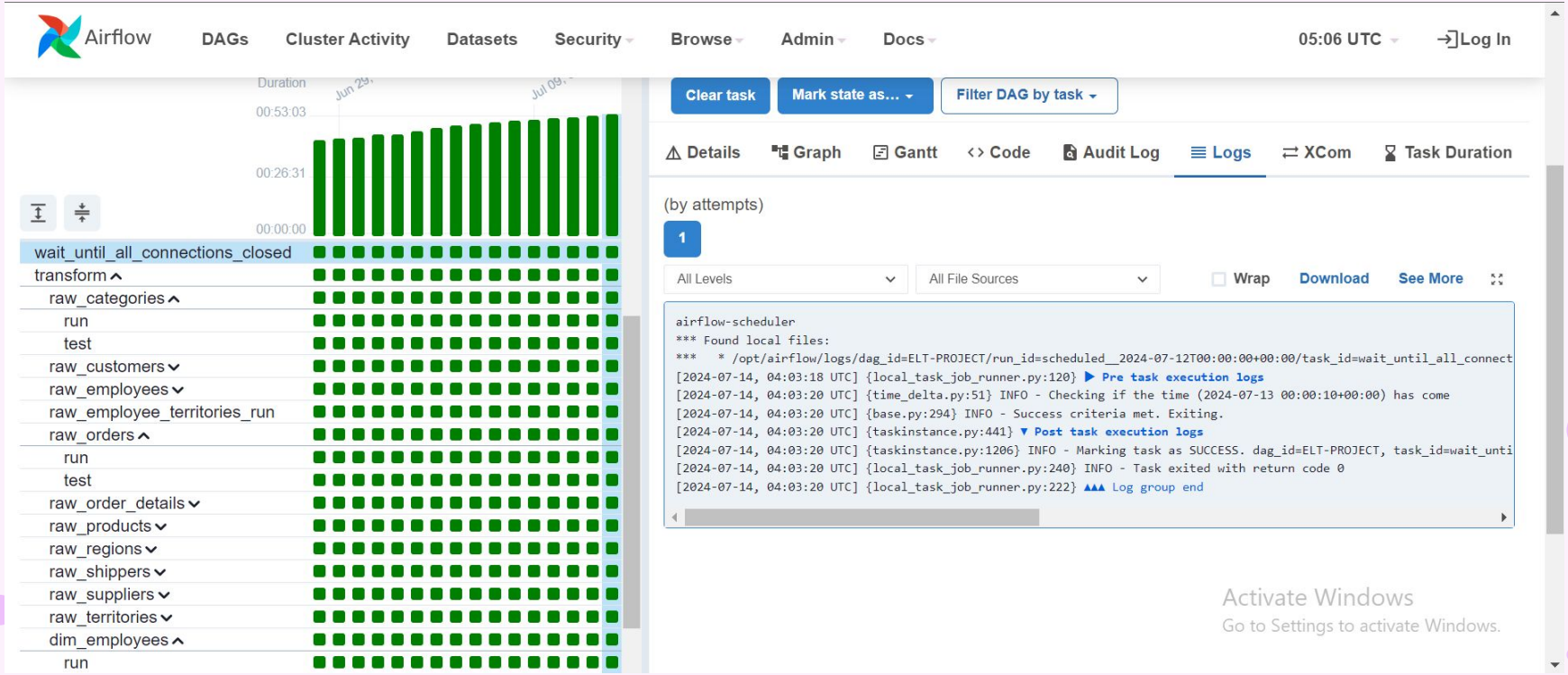
Extract Task



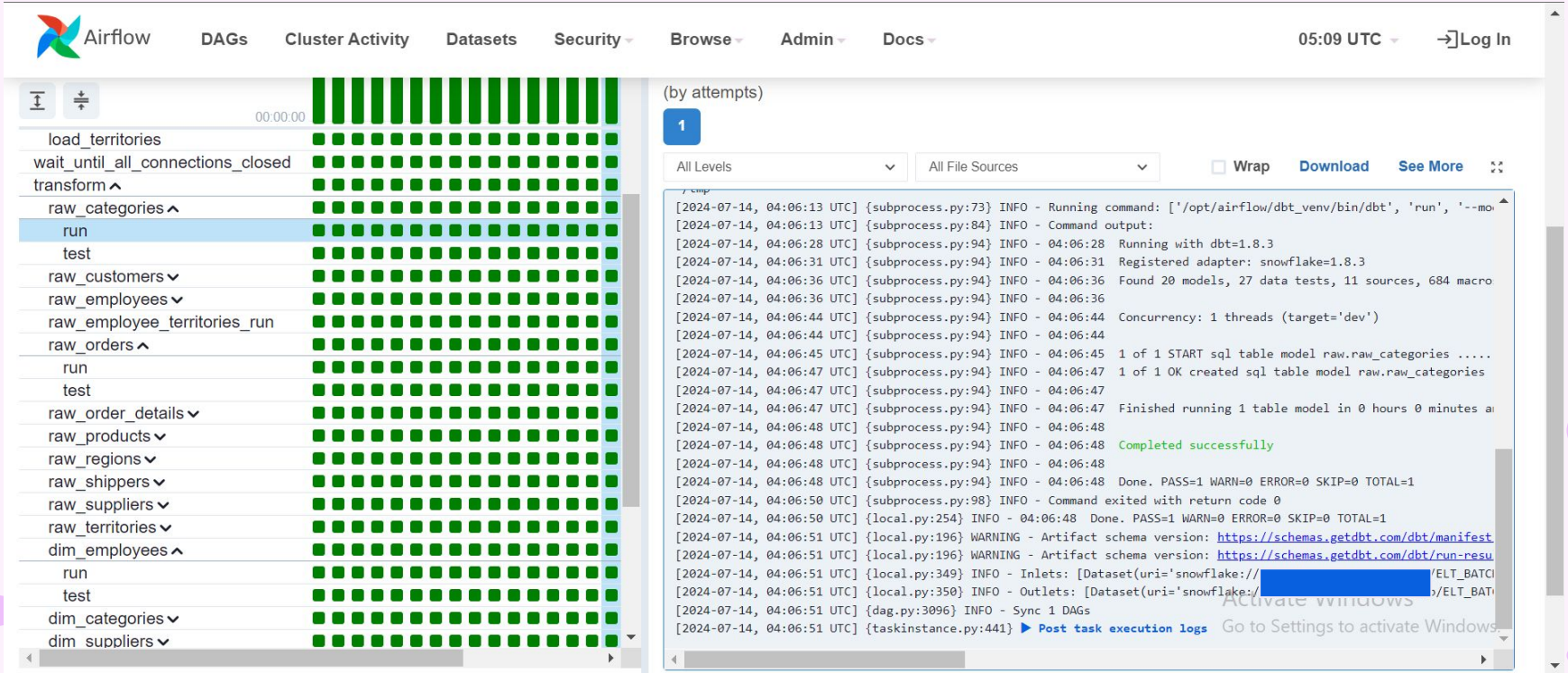
Load Task


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
05:04 UTC
Log In
Duration
00:53:03
00:26:31
00:00:00
load
load_categories
load_customers
load_employee_territories
load_employees
load_order_details
load_orders
load_products
load_regions
load_shippers
load_suppliers
load_territories
wait_until_all_connections_closed
transform
raw_categories
run
test
raw_customers
raw_employees
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
[2024-07-14, 04:03:01 UTC] {connection.py:399} INFO - Snowflake Connector for Python Version: 3.10.1, Python Version
[2024-07-14, 04:03:01 UTC] {connection.py:1249} INFO - THIS CONNECTION IS IN INSECURE MODE. IT MEANS THE CERTIFICATE
[2024-07-14, 04:03:02 UTC] {ssl_wrap_socket.py:100} INFO - THIS CONNECTION IS IN INSECURE MODE. IT MEANS THE CERTIFI
[2024-07-14, 04:03:02 UTC] {logging_mixin.py:188} INFO - Nama kolom di DataFrame: Index(['categoryID', 'categoryName
[2024-07-14, 04:03:02 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 1
[2024-07-14, 04:03:03 UTC] {ssl_wrap_socket.py:100} INFO - THIS CONNECTION IS IN INSECURE MODE. IT MEANS THE CERTIFI
[2024-07-14, 04:03:05 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 1
[2024-07-14, 04:03:06 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 4
[2024-07-14, 04:03:07 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 1
[2024-07-14, 04:03:07 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 1
[2024-07-14, 04:03:09 UTC] {cursor.py:1149} INFO - Number of results in first chunk: 1
[2024-07-14, 04:03:09 UTC] {logging_mixin.py:188} INFO - Done categories.
[2024-07-14, 04:03:09 UTC] {connection.py:762} INFO - closed
[2024-07-14, 04:03:09 UTC] {connection.py:768} INFO - No async queries seem to be running, deleting session
[2024-07-14, 04:03:10 UTC] {python.py:237} INFO - Done. Returned value was: None
[2024-07-14, 04:03:10 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2024-07-14, 04:03:10 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=ELT-PROJECT, task_id=load.l
[2024-07-14, 04:03:10 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2024-07-14, 04:03:10 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end

Wait Task


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
05:06 UTC
Log In
Duration
00:53:03
00:26:31
00:00:00
wait_until_all_connections_closed
transform
raw_categories
run
test
raw_customers
raw_employees
raw_employee_territories_run
raw_orders
run
test
raw_order_details
raw_products
raw_regions
raw_shippers
raw_suppliers
raw_territories
dim_employees
run
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
airflow-scheduler
*** Found local files:
*** * /opt/airflow/logs/dag_id=ELT-PROJECT/run_id=scheduled__2024-07-12T00:00:00+00:00/task_id=wait_until_all_connect
[2024-07-14, 04:03:18 UTC] {local_task_job_runner.py:120} ▶ Pre task execution logs
[2024-07-14, 04:03:20 UTC] {time_delta.py:51} INFO - Checking if the time (2024-07-13 00:00:10+00:00) has come
[2024-07-14, 04:03:20 UTC] {base.py:294} INFO - Success criteria met. Exiting.
[2024-07-14, 04:03:20 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2024-07-14, 04:03:20 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=ELT-PROJECT, task_id=wait_unti
[2024-07-14, 04:03:20 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2024-07-14, 04:03:20 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
Activate Windows
Go to Settings to activate Windows.

Transform Task



DBT Templates

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 05:13 UTC Log In

00:00:00

- load_territories
- wait_until_all_connections_closed
- transform
  - raw_categories
    - run
    - test
  - raw_customers
  - raw_employees
  - raw_employee_territories_run
  - raw_orders
    - run
    - test
  - raw_order_details
  - raw_products
  - raw_regions
  - raw_shippers
  - raw_suppliers
  - raw_territories
  - dim_employees
    - run
    - test
  - dim_categories
  - dim_suppliers

Pool Slots	1
Executor Config	{}
Unix Name	airflow
Max Tries	1
Queue	default
Priority Weight	12

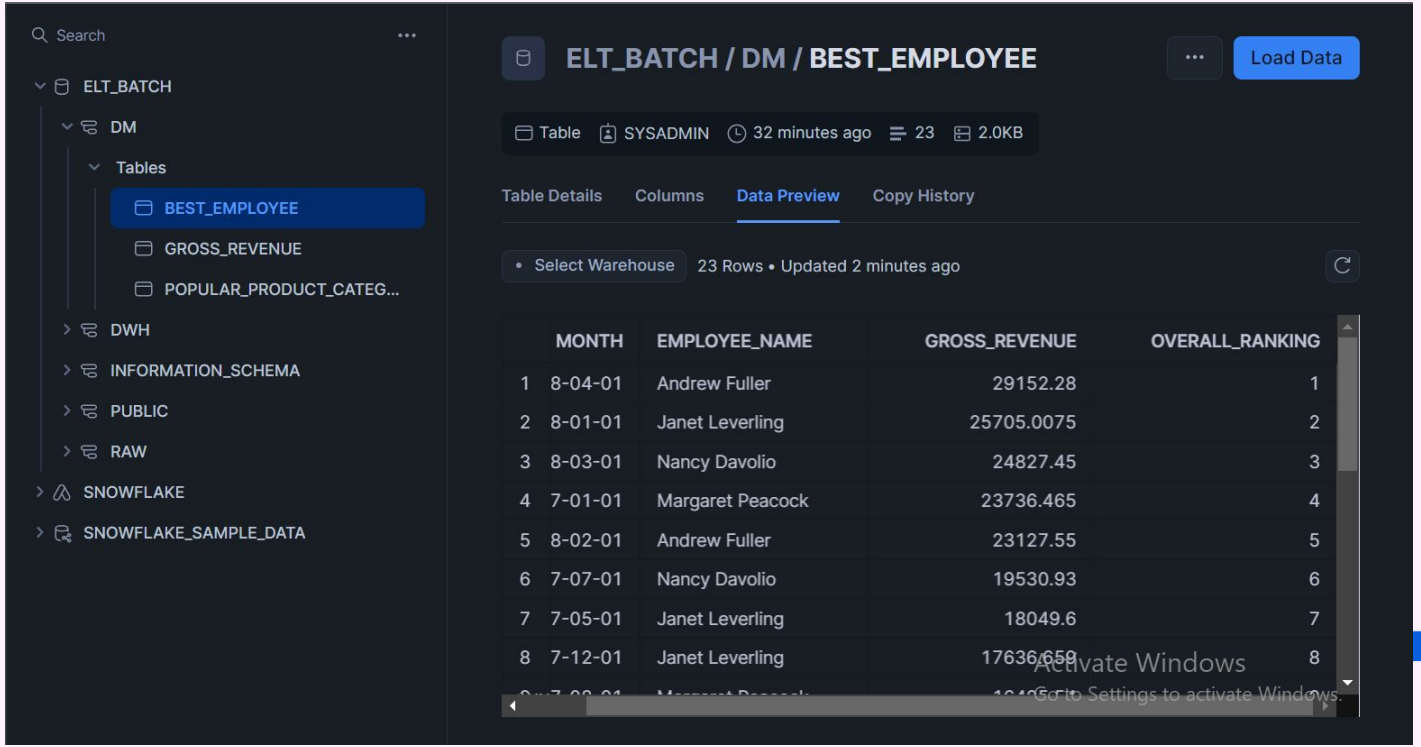
Rendered Templates

compiledSql

```
-- target/compiled/dbt_project/models/staging/raw_categories.sql SELECT DISTINCT CATEGORYID, TRIM(CATEGORYNAME) AS CATEGORYNAME, TRIM(DESCRIPTION) AS DESCRIPTION, TO_BINARY(PICTURE, 'UTF-8') AS PICTURE FROM ELT_BATCH.PUBLIC.categories WHERE CATEGORYID IS NOT NULL AND CATEGORYNAME IS NOT NULL -- target/run/dbt_project/models/staging/raw_categories.sql create or replace transient table ELT_BATCH.raw.raw_categories as (SELECT DISTINCT CATEGORYID, TRIM(CATEGORYNAME) AS CATEGORYNAME, TRIM(DESCRIPTION) AS DESCRIPTION, TO_BINARY(PICTURE, 'UTF-8') AS PICTURE FROM ELT_BATCH.PUBLIC.categories WHERE CATEGORYID IS NOT NULL AND CATEGORYNAME IS NOT NULL );
```

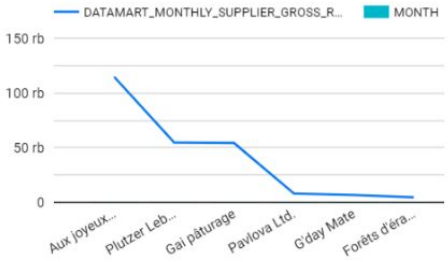
models `raw_categories`

Data Mart



	MONTH	EMPLOYEE_NAME	GROSS_REVENUE	OVERALL_RANKING
1	8-04-01	Andrew Fuller	29152.28	1
2	8-01-01	Janet Leverling	25705.0075	2
3	8-03-01	Nancy Davolio	24827.45	3
4	7-01-01	Margaret Peacock	23736.465	4
5	8-02-01	Andrew Fuller	23127.55	5
6	7-07-01	Nancy Davolio	19530.93	6
7	7-05-01	Janet Leverling	18049.6	7
8	7-12-01	Janet Leverling	17636.659	8

NORTHWIND ANALYTIC



MONTH

Masukkan nilai

Pilih rentang tanggal

EMPLOYEE_NAME

Sama den...

Masukkan nilai

DATAMART_MONTHLY_BEST_E...

6.517,47

30.990,28