Project: Adventure Works

Foundation First !!!

Four Key Questions

I. Where do we consolidate our data ? > Storage
II. How will we get it there ? > Ingestion
III. How will we clean it up? > Transformation
IV. How will we analyze it? > Reporting

BRIAN GWAYI

Data Stack

Popular Options

Storage > Snowflake, BigQuery, s3, Redshift
Ingestion > Airbyte, Airflow, Fivetran
Transformation > dbt
Reporting > Tableau, Power BI, Looker, Superset

N/B This is not an exhaustive list.

BRIAN GWAYI

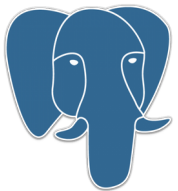# Simple Modern Data Stack Architecture

Source

Ingestion

Storage
BigQuery

Reporting

PostgreSQL

Airflow

+

Python

Transformation
dbt

Raw → Model

Model

Model

Looker

End Goal

Push data

Online Transaction Processing

Online Analytical Processing

BRIAN GWAYI

Content

02 Ingestion

Setting up Apache Airflow

– Airflow Documentation

- Production Deployment Documentation

Writing ELT Python Script

- .py Code - Extract & Load

Python

```python
# importing libraries

from airflow.decorators import dag, task
from datetime import datetime, timedelta
import requests
from google.cloud import bigquery
import pandas as pd
import psycopg2
from io import StringIO
```

02   Ingestion
Setting up Apache Airflow
      – [Airflow Documentation](#)
      [- Production Deployment Documentation](#)

Writing ELT Python Script
      [- .py Code - Extract & Load](#)

Python

```python
# instantiating DAG

args{                                          @dag(
    "owner":"gwayi",                               default_arguments = args
    "retries": 1,                                  schedule=timedelta(minutes=30),
    "retry_delay":timedelta(minutes=5)             start_date=datetime(2024, 7, 29),
    }                                              catchup=False,
                                                   tags=['Team B']
                                                   )
```

Writing ELT Python Script
[- .py Code - Extract & Load](#)

Python

```python
@task()
def gt_tbls(conn):

    sql = """SELECT table_name
    FROM information_schema.tables
    WHERE table_type = 'BASE TABLE'
    AND table_catalog = 'adventure_works'
    AND table_schema NOT IN
    ('pg_catalog','information_schema');"""

    cursor = conn.cursor()
    cursor.execute(sql)
    tbls=cursor.fetchall()

    conn.commit()
    conn.close()

    tbls = [x[0] for x in tbls]
    Return tbls
```

02

Writing ELT Python Script

Python

```python
@task()
def xt_tbls(tbls):

        dataframe = {}
        for tbl in tbls:
                sql = f"SELECT * FROM {tbl} WHERE
                createdAt <= (convert(datetime2, {last_rundate}) OR
                modifiedAt <=(convert(datetime2, {last_rundate})"
                dataframe[tbl] = pd.read_sql(sql, conn)

        return dataframe
```

02
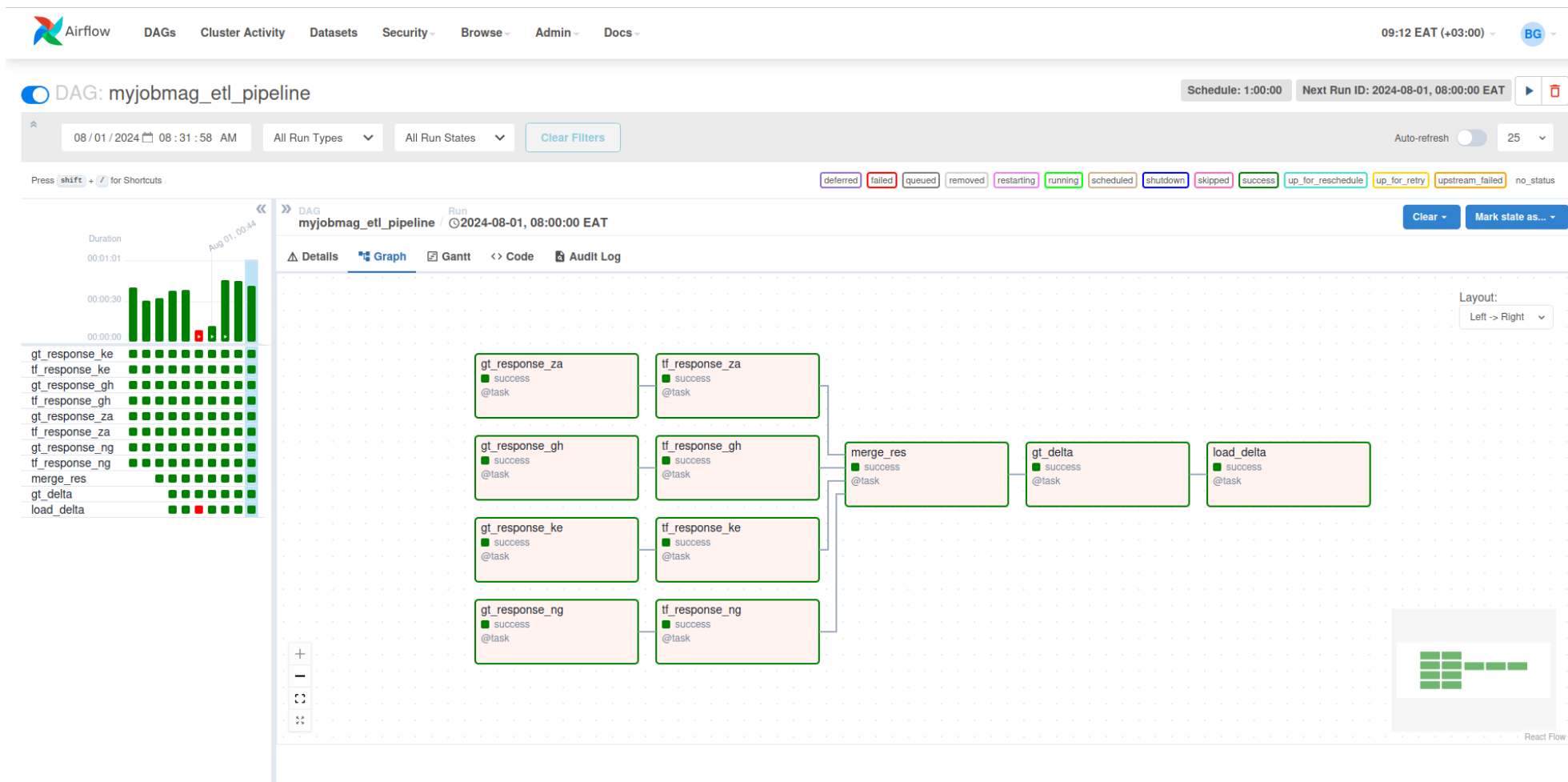
Writing ELT Python Script
- .py Code - Extract & Load

Python

```python
@task()

def upsert_tbls(df):
        client = bigquery.Client()
        table_id = "adventureworks-431609.adw_dwh.customer"
        job = client.load_table_from_dataframe(df, table_id)
        job.result()
        print(f"uploading data to Google BigQuery is {job.state}")
upsert_tbla()
```

02

Ingestion
Orchestrating & Running Workflow – Apache Airflow

02

Ingestion
Data loaded in Google BigQuery