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

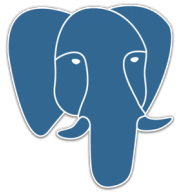
N/B This is not an exhaustive list.

BRIAN GWAYI

Data Pipeline Architecture Design

Source

PostgreSQL



Ingestion

Airflow

+

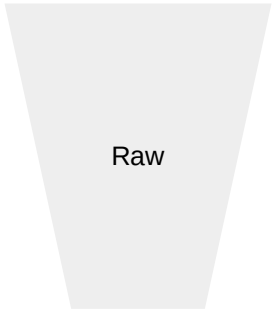
Python

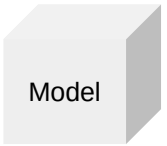
Storage
BigQuery

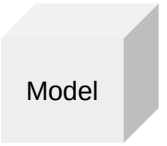
Amazon **Redshift**

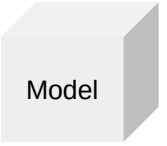
Transformation
dbt

dbt









Reporting

Looker

looker

Content

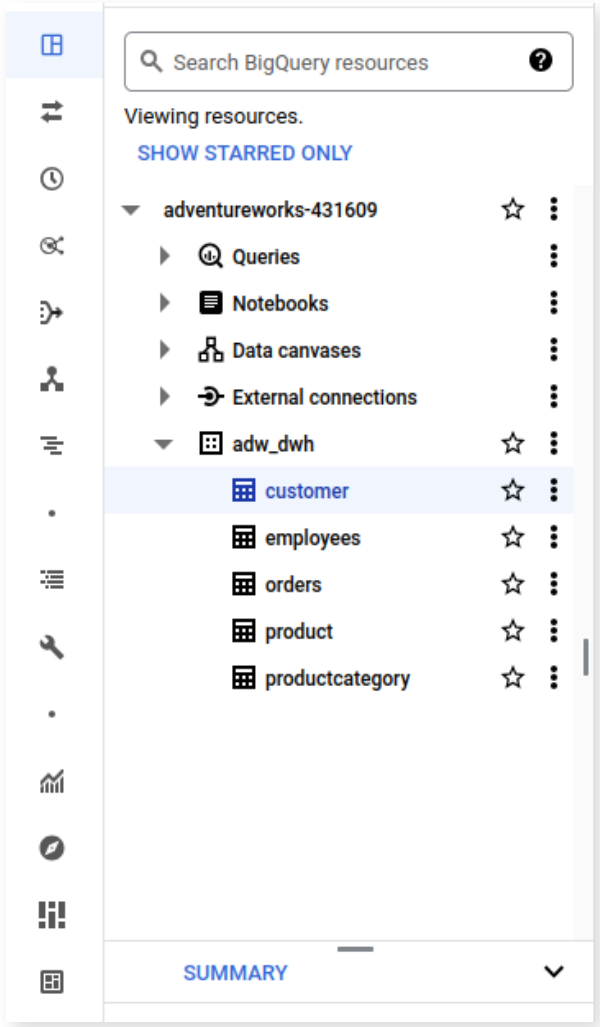
01 **Storage/Database**
Setting Google BigQuery

02 **Ingestion**
Setting up Apache Airflow
Writing elt Python script
Orchestrate data pipeline

03 **Transformation**
Setting up dbt
Transformation

04 **Reporting**
Connecting Looker

01 Storage
Set up BigQuery



customer | Untitled query

customer

QUERY | SHARE | COPY | SNAPSHOT | DELETE | EXPORT

SCHEMA | DETAILS | PREVIEW | TABLE EXPLORER PREVIEW | INSIGHTS PREVIEW | LINEAGE

Row	customerid	firstname	lastname	fullname
1	1305	A.	Leonetti	A. Leonetti
2	829	Ed	Dudenhoefer	Ed Dudenhoefer
3	1953	H.	Valentine	H. Valentine
4	1917	Abe	Tramel	Abe Tramel
5	323	Amy	Alberts	Amy Alberts
6	735	Amy	Consentino	Amy Consentino
7	437	Ann	Beebe	Ann Beebe
8	1033	Ann	Hass	Ann Hass
9	801	Bev	Desalvo	Bev Desalvo
10	919	Bob	Gage	Bob Gage
11	1099	Bob	Hodges	Bob Hodges
12	509	Eli	Bowen	Eli Bowen
13	743	Eva	Corets	Eva Corets
14	949	Guy	Gilbert	Guy Gilbert
15	1467	Jan	Miksovsky	Jan Miksovsky
16	893	Jay	Fluegel	Jay Fluegel
17	1051	Jay	Henningsen	Jay Henningsen

02

Ingestion

Setting up Apache Airflow

– Airflow Documentation

- Production Deployment Documentation

Writing ELT Python Script

- .py Code - Extract & Load

```
# 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

```
# instantiating DAG

@dag(
    schedule=timedelta(minutes=30),
    start_date=datetime(2024, 7, 29),
    catchup=False,
    tags=['Team B']
    )
```

02 Writing ELT Python Script

Task 1
Get Table Lists from PostgreSQL

```
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');"""
```

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

         conn = psycopg2.connect(
            database = "adw_db",
             user = "postgres",
             host= 'localhost',
             password = "password",
             Port = 5432)

        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

Task 2

Extract tables

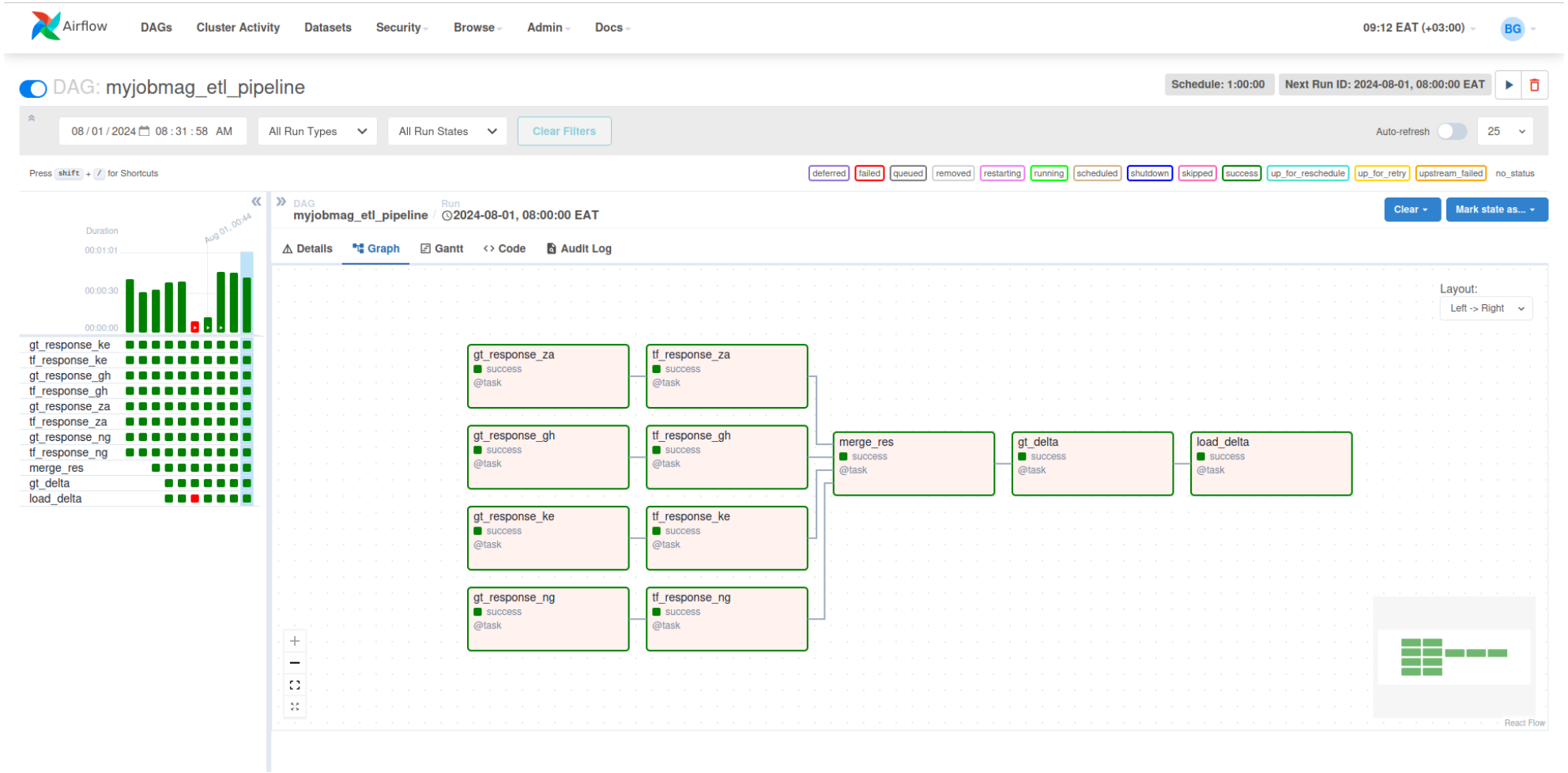
```
@task()
def xt_tbls(tbls):
        dataframe = {}
        for tbl in tbls:
            sql = f"SELECT * FROM {tbl} WHERE
            createdAt <= (convert(datetime2, {last_rundate}) OR
            modifiedAt <=(convert(datetime2, {last_rundate})"
            dataframe[tbl] = pd.read_sql(sql, conn)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

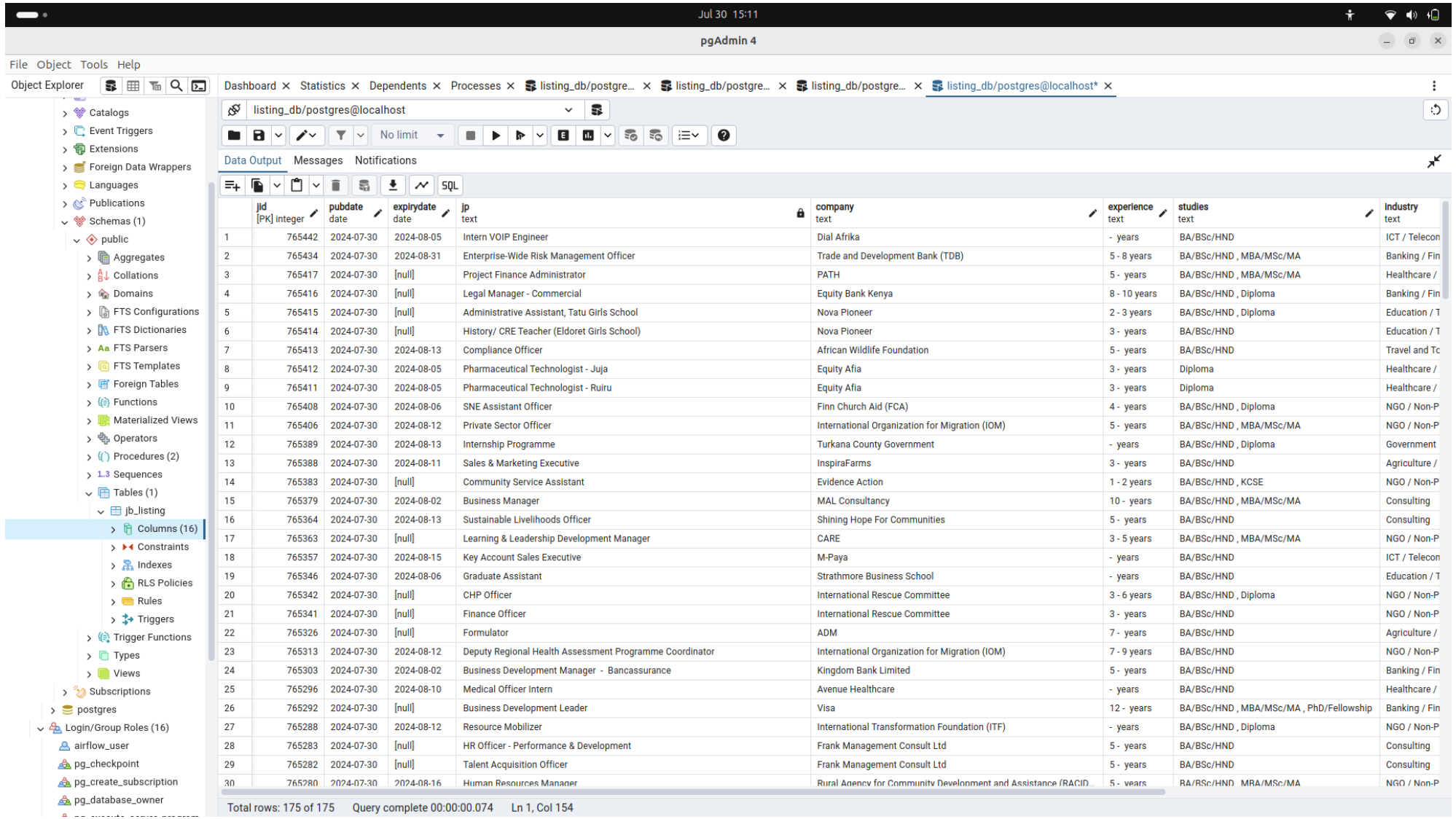
02 Writing ELT Python Script

Task 3

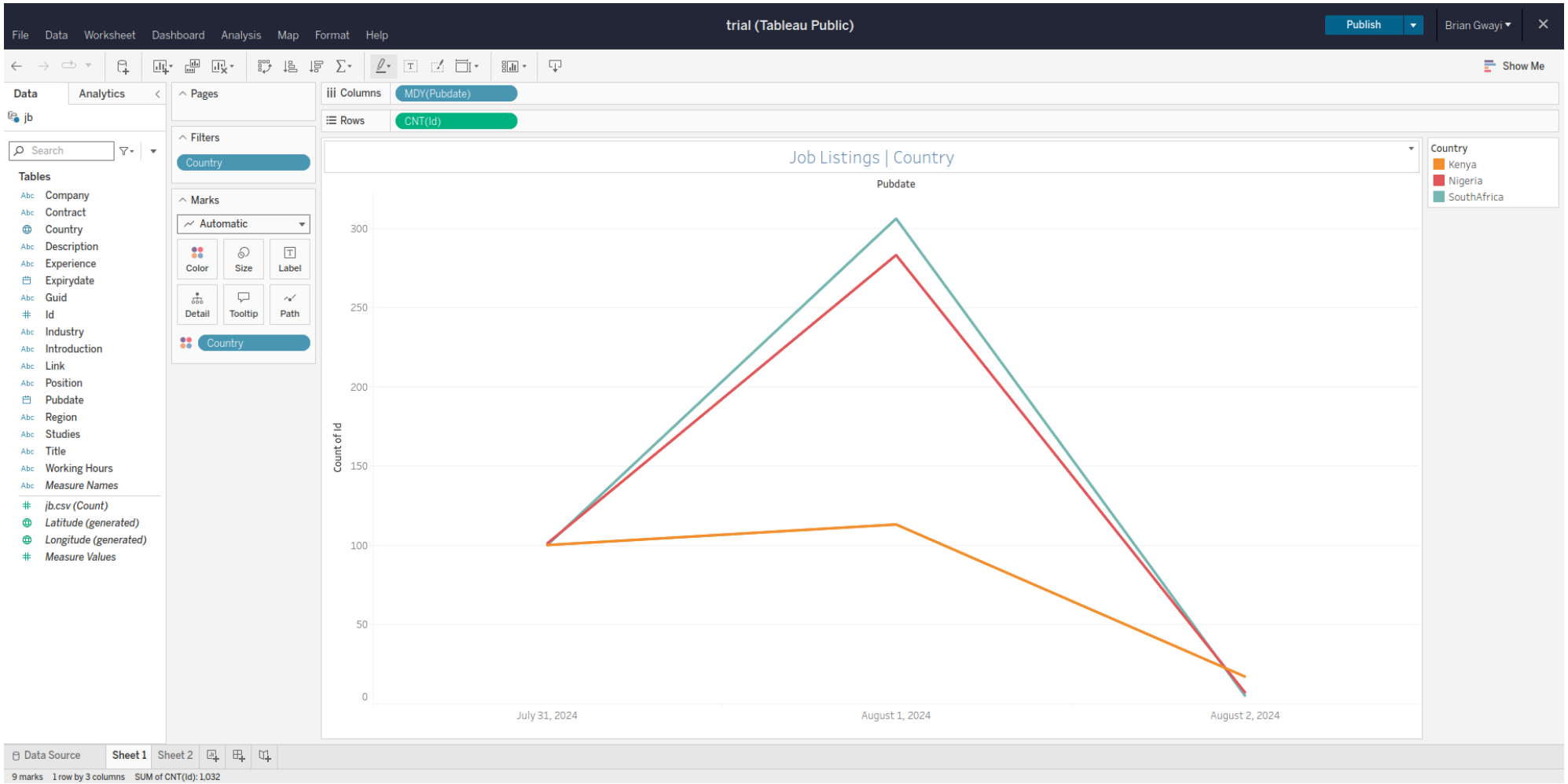
Upsert rowsBigQuery

```
@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()
```


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
09:12 EAT (+03:00)
BG
DAG: myjobmag_etl_pipeline
Schedule: 1:00:00
Next Run ID: 2024-08-01, 08:00:00 EAT
08 / 01 / 2024 08 : 31 : 58 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Aug 01, 00:44
00:01:01
00:00:30
00:00:00
gt_response_ke
tf_response_ke
gt_response_gh
tf_response_gh
gt_response_za
tf_response_za
gt_response_ng
tf_response_ng
merge_res
gt_delta
load_delta
DAG
myjobmag_etl_pipeline
Run
2024-08-01, 08:00:00 EAT
Clear
Mark state as...
Details
Graph
Gantt
Code
Audit Log
Layout:
Left -> Right
gt_response_za
success
@task
tf_response_za
success
@task
gt_response_gh
success
@task
tf_response_gh
success
@task
gt_response_ke
success
@task
tf_response_ke
success
@task
gt_response_ng
success
@task
tf_response_ng
success
@task
merge_res
success
@task
gt_delta
success
@task
load_delta
success
@task
React Flow



	jid [PK] integer	pubdate date	expirydate date	jp text	company text	experience text	studies text	industry text
1	765442	2024-07-30	2024-08-05	Intern VOIP Engineer	Dial Afrika	- years	BA/BSc/HND	ICT / Telecon
2	765434	2024-07-30	2024-08-31	Enterprise-Wide Risk Management Officer	Trade and Development Bank (TDB)	5 - 8 years	BA/BSc/HND , MBA/MSc/MA	Banking / Fin
3	765417	2024-07-30	[null]	Project Finance Administrator	PATH	5 - years	BA/BSc/HND , MBA/MSc/MA	Healthcare /
4	765416	2024-07-30	[null]	Legal Manager - Commercial	Equity Bank Kenya	8 - 10 years	BA/BSc/HND , Diploma	Banking / Fin
5	765415	2024-07-30	[null]	Administrative Assistant, Tatu Girls School	Nova Pioneer	2 - 3 years	BA/BSc/HND , Diploma	Education / T
6	765414	2024-07-30	[null]	History/ CRE Teacher (Eldoret Girls School)	Nova Pioneer	3 - years	BA/BSc/HND	Education / T
7	765413	2024-07-30	2024-08-13	Compliance Officer	African Wildlife Foundation	5 - years	BA/BSc/HND	Travel and Tc
8	765412	2024-07-30	2024-08-05	Pharmaceutical Technologist - Juja	Equity Afia	3 - years	Diploma	Healthcare /
9	765411	2024-07-30	2024-08-05	Pharmaceutical Technologist - Ruiru	Equity Afia	3 - years	Diploma	Healthcare /
10	765408	2024-07-30	2024-08-06	SNE Assistant Officer	Finn Church Aid (FCA)	4 - years	BA/BSc/HND , Diploma	NGO / Non-P
11	765406	2024-07-30	2024-08-12	Private Sector Officer	International Organization for Migration (IOM)	5 - years	BA/BSc/HND , MBA/MSc/MA	NGO / Non-P
12	765389	2024-07-30	2024-08-13	Internship Programme	Turkana County Government	- years	BA/BSc/HND , Diploma	Government
13	765388	2024-07-30	2024-08-11	Sales & Marketing Executive	InspiraFarms	3 - years	BA/BSc/HND	Agriculture /
14	765383	2024-07-30	[null]	Community Service Assistant	Evidence Action	1 - 2 years	BA/BSc/HND , KCSE	NGO / Non-P
15	765379	2024-07-30	2024-08-02	Business Manager	MAL Consultancy	10 - years	BA/BSc/HND , MBA/MSc/MA	Consulting
16	765364	2024-07-30	2024-08-13	Sustainable Livelihoods Officer	Shining Hope For Communities	5 - years	BA/BSc/HND	Consulting
17	765363	2024-07-30	[null]	Learning & Leadership Development Manager	CARE	3 - 5 years	BA/BSc/HND , MBA/MSc/MA	NGO / Non-P
18	765357	2024-07-30	2024-08-15	Key Account Sales Executive	M-Paya	- years	BA/BSc/HND	ICT / Telecon
19	765346	2024-07-30	2024-08-06	Graduate Assistant	Strathmore Business School	- years	BA/BSc/HND	Education / T
20	765342	2024-07-30	[null]	CHP Officer	International Rescue Committee	3 - 6 years	BA/BSc/HND , Diploma	NGO / Non-P
21	765341	2024-07-30	[null]	Finance Officer	International Rescue Committee	3 - years	BA/BSc/HND	NGO / Non-P
22	765326	2024-07-30	[null]	Formulator	ADM	7 - years	BA/BSc/HND	Agriculture /
23	765313	2024-07-30	2024-08-12	Deputy Regional Health Assessment Programme Coordinator	International Organization for Migration (IOM)	7 - 9 years	BA/BSc/HND	NGO / Non-P
24	765303	2024-07-30	2024-08-02	Business Development Manager - Bancassurance	Kingdom Bank Limited	5 - years	BA/BSc/HND	Banking / Fin
25	765296	2024-07-30	2024-08-10	Medical Officer Intern	Avenue Healthcare	- years	BA/BSc/HND	Healthcare /
26	765292	2024-07-30	[null]	Business Development Leader	Visa	12 - years	BA/BSc/HND , MBA/MSc/MA , PhD/Fellowship	Banking / Fin
27	765288	2024-07-30	2024-08-12	Resource Mobilizer	International Transformation Foundation (ITF)	- years	BA/BSc/HND , Diploma	NGO / Non-P
28	765283	2024-07-30	[null]	HR Officer - Performance & Development	Frank Management Consult Ltd	5 - years	BA/BSc/HND	Consulting
29	765282	2024-07-30	[null]	Talent Acquisition Officer	Frank Management Consult Ltd	5 - years	BA/BSc/HND	Consulting
30	765280	2024-07-30	2024-08-16	Human Resources Manager	Rural Agency for Community Development and Assistance (RACID...	5 - years	BA/BSc/HND , MBA/MSc/MA	NGO / Non-P


File Data Worksheet Dashboard Analysis Map Format Help
trial (Tableau Public)
Publish
Brian Gwayi
Show Me
Data
Analytics
jb
Search
Tables
Abc Company
Abc Contract
Country
Abc Description
Abc Experience
Expirydate
Abc Guid
Id
Abc Industry
Abc Introduction
Abc Link
Abc Position
Pubdate
Abc Region
Abc Studies
Abc Title
Abc Working Hours
Abc Measure Names
jb.csv (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Country
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Path
Country
Columns
MDY(Pubdate)
Rows
CNT(Id)
Job Listings | Country
Pubdate
Count of Id
300
250
200
150
100
50
0
July 31, 2024
August 1, 2024
August 2, 2024
Country
Kenya
Nigeria
SouthAfrica
Data Source
Sheet 1
Sheet 2
9 marks 1 row by 3 columns SUM of CNT(Id): 1,032

Modern Data Stack Ecosystem 2024

The right tools for building robust data stack architecture will be bases on Combination of budget, skillset, data sources and preferences.