



Wizeline DEB Bootcamp

Capstone Project Presentation



Sidney Ochoki
Isoe

Roles: Software Implementation Consultant |
Business Analyst | Data Enthusiast

Experience: Working with data for over a decade

Hobbies: Music, video games, data

Contact: [in](#) [M](#) [o](#) use [sidneyisoe](#)



Session Goal

The goal is to explain my end-to-end solution used for the capstone project.

1. Introduction
2. Infrastructure
3. Data Pipelines
4. Challenges
5. Lessons Learned

1. Introduction

The objective of the project was to use the GCP technologies and data engineering techniques to build an end-to-end solution for movie analytics data warehouse.



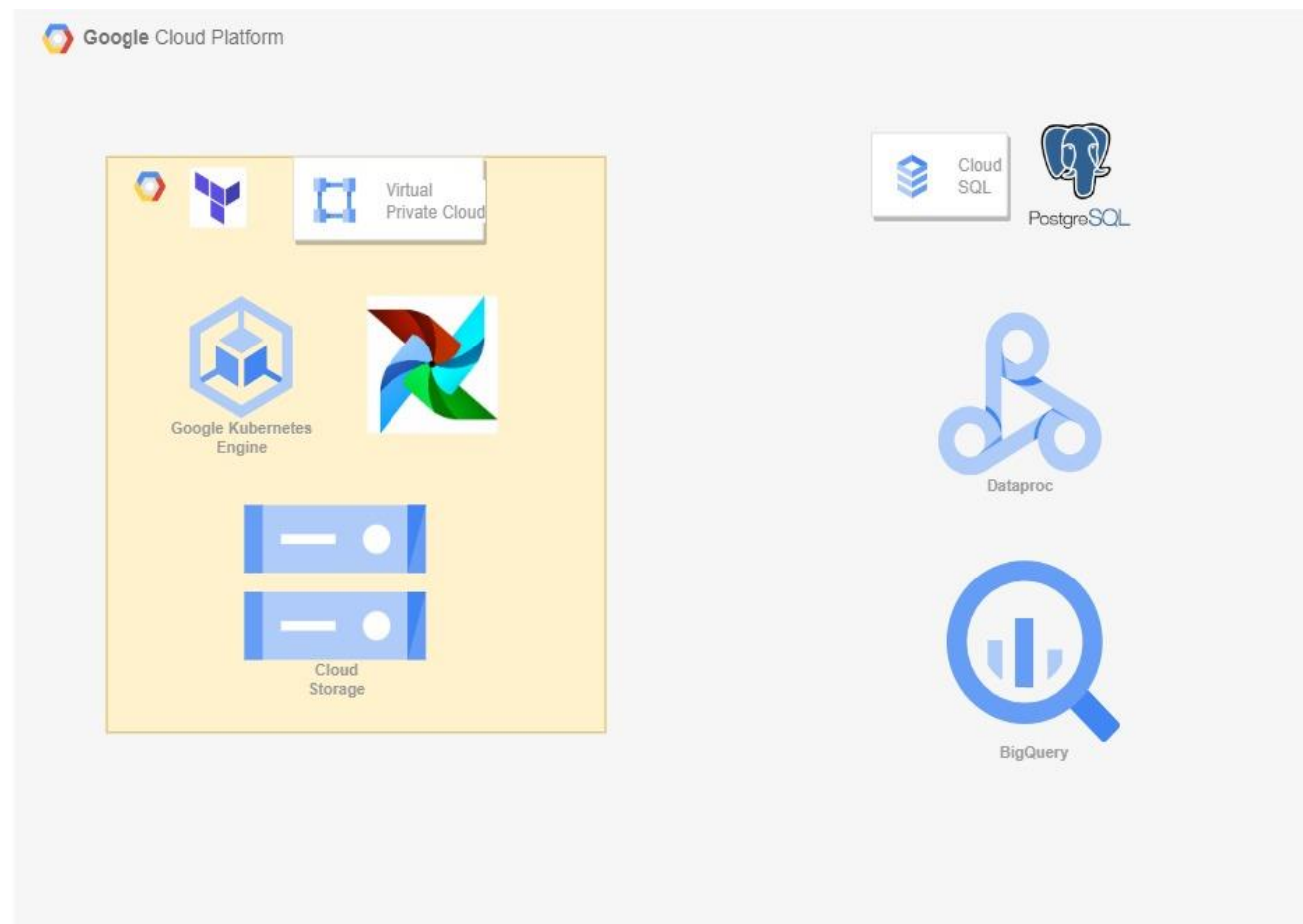
2. Infrastructure

The infrastructure is built using **Terraform** – check the modules and .tf* files in my repo. This infrastructure consists of:

- VPC
- Cloud SQL
- Compute engine
- GKE cluster
- Google Cloud Storage

Other components such as **Dataproc** and **BigQuery** are created using **Airflow**

Prerequisites are listed in the **README.md** file



Airflow Pipeline

DAGs

All 11Active 1Paused 10

Filter DAGs by tag

Show only active DAGsShow only paused DAGs

DAG	Owner	Runs	Schedule	Last Run
<div><div><div>wizeline_capstone_end_to_end_v3</div><div>gcpunstable</div></div></div>	airflow	<div><div></div><div>2</div><div></div><div>3</div></div>	@once	2023-10-23, 15:53:00

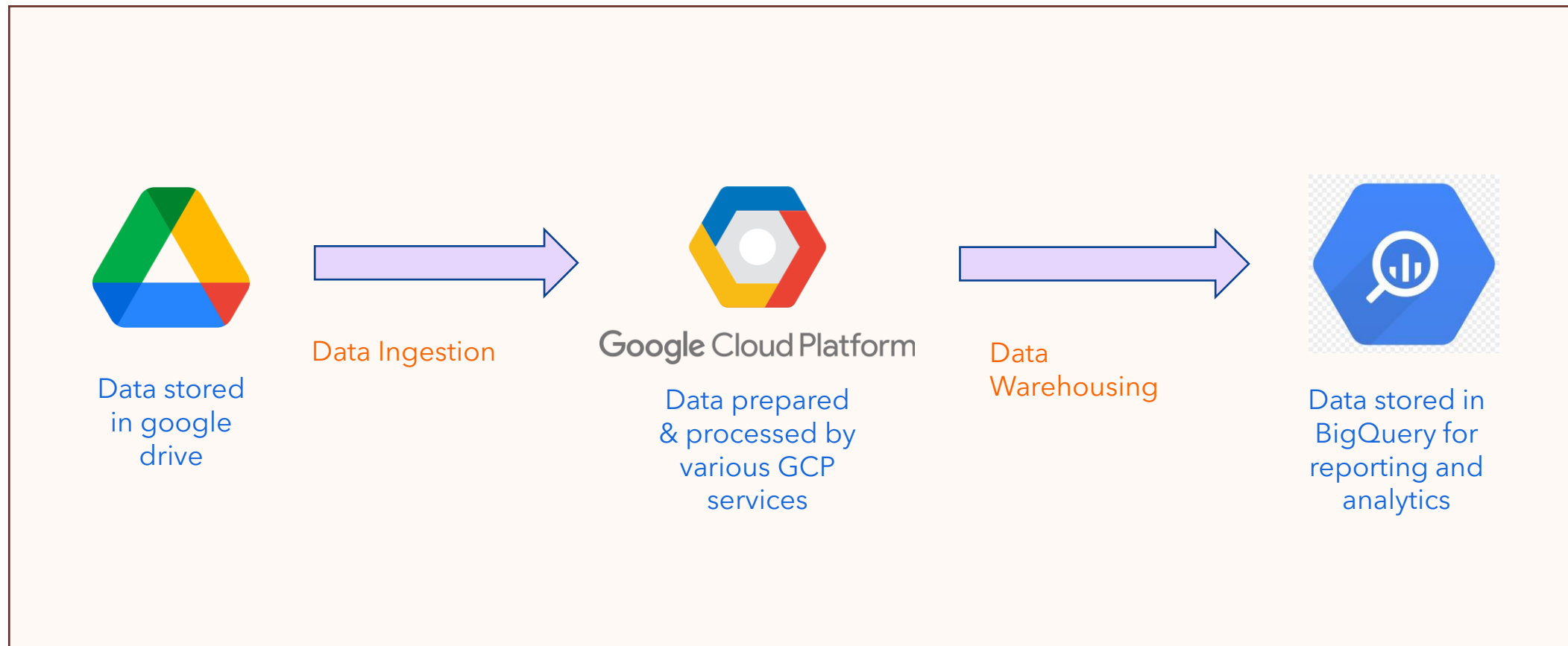
Variables			
	Key	Val	Description
<input type="checkbox"/>	capstone_files_urls_private	{ "log_reviews.csv": "https://drive.g...	
<input type="checkbox"/>	capstone_files_urls_public	{ "log_reviews": "https://drive.googl...	
<input type="checkbox"/>	capstone_pyspark_files_urls_public	{ "log_review_processing": "https://...	

One end-to-end solution

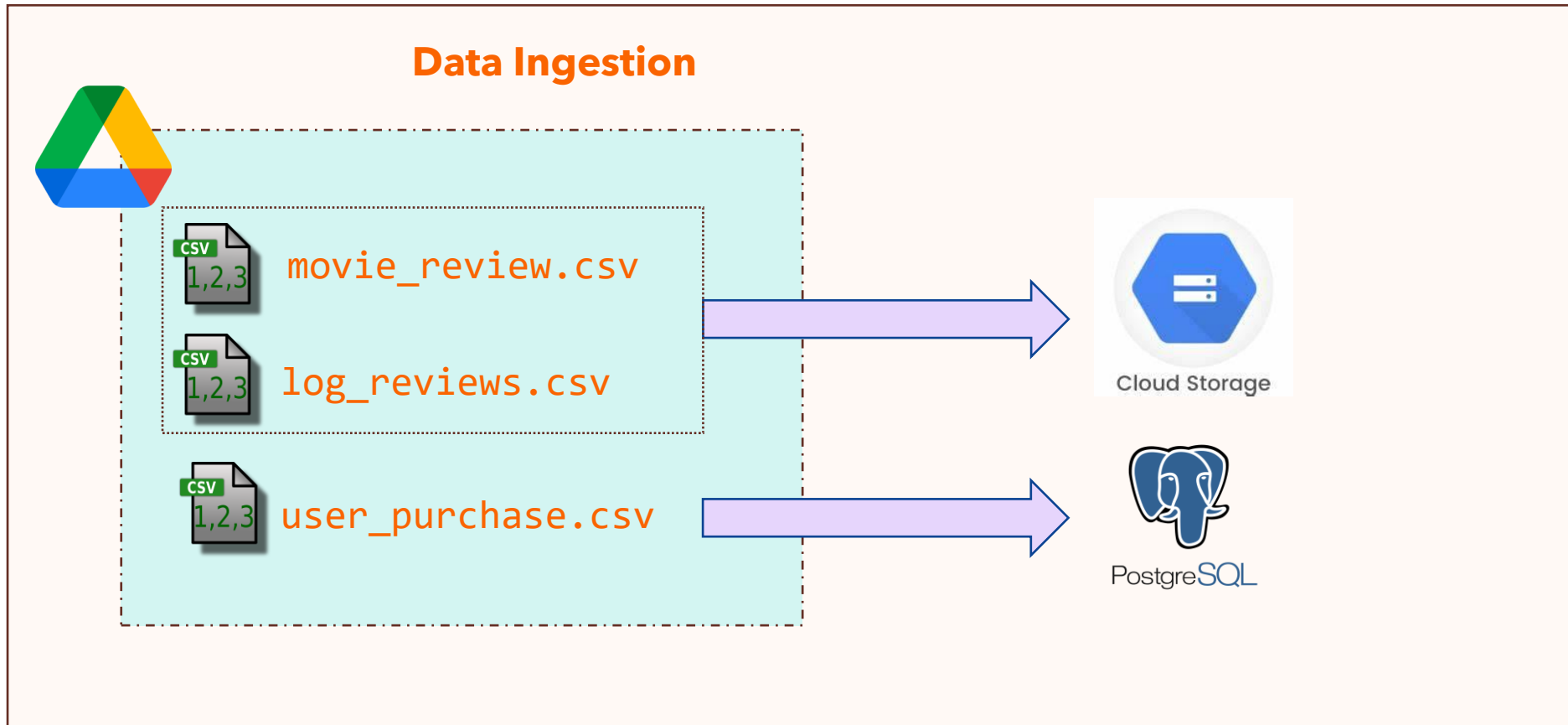
Connections

	Conn Id	Conn Type	Description	Host	Port
<input type="checkbox"/>	capstone_postgres	postgres		34.41.179.52	5432
<input type="checkbox"/>	gcp_default	google_cloud_platform			

Airflow Pipeline high level



Pipeline – load data to RAW

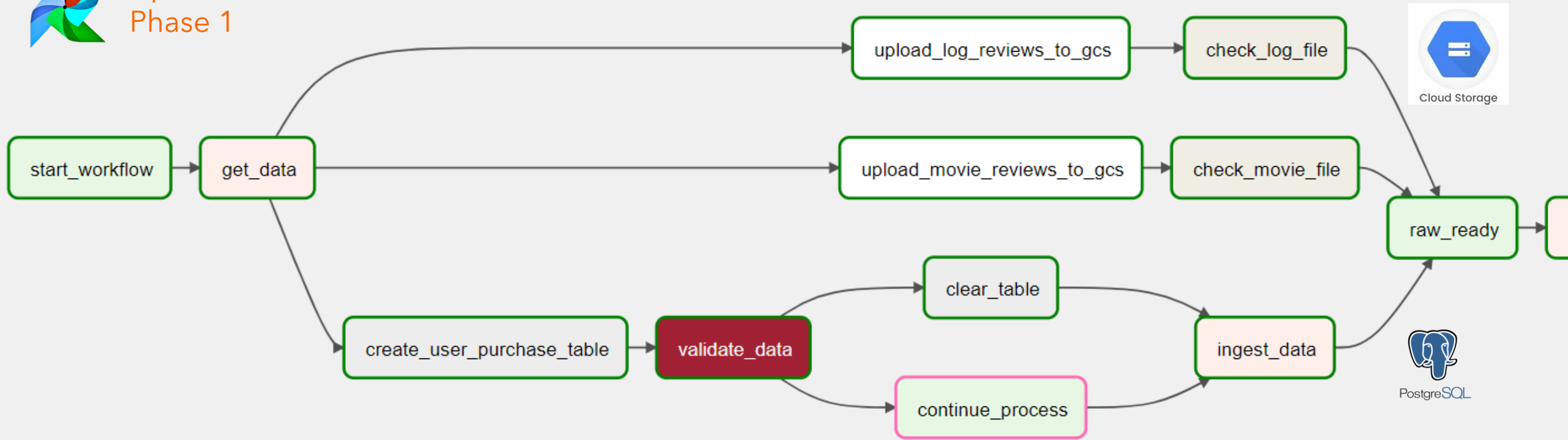


- `user_purchase.csv` file into a PostgreSQL DB
- `movie_review.csv` and `log_reviews.csv` into a storage bucket

Pipeline – load data to RAW



Apache Airflow:
Phase 1





- `user_purchase.csv` file into a PostgreSQL DB
- `movie_review.csv` and `log_reviews.csv` into a storage bucket

Pipelines – ingested data

Buckets > wizeline_bootcamp_bucket > RAW

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER TF

Filter by name prefix only Filter Filter objects and folders

<input type="checkbox"/>	Name	Size
<input type="checkbox"/>	 log_reviews.csv	20.8 MB
<input type="checkbox"/>	 movie_reviews.csv	127.1 MB

Statistics Dependencies Dependents Processes public.user_purchase/capstone/sisoe@capstone

capstone/sisoe@capstone

100 rows

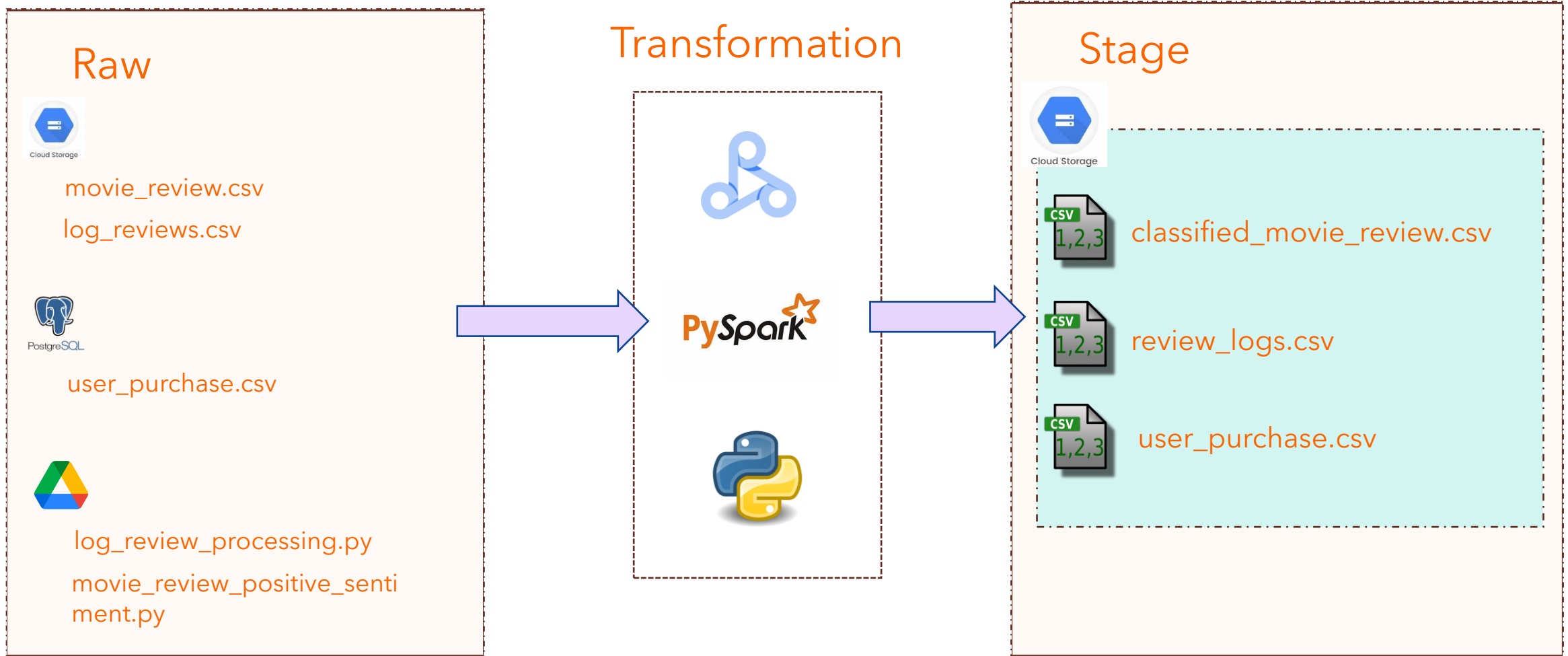
Query Query History

```
1 SELECT * FROM public.user_purchase
2 LIMIT 100
```

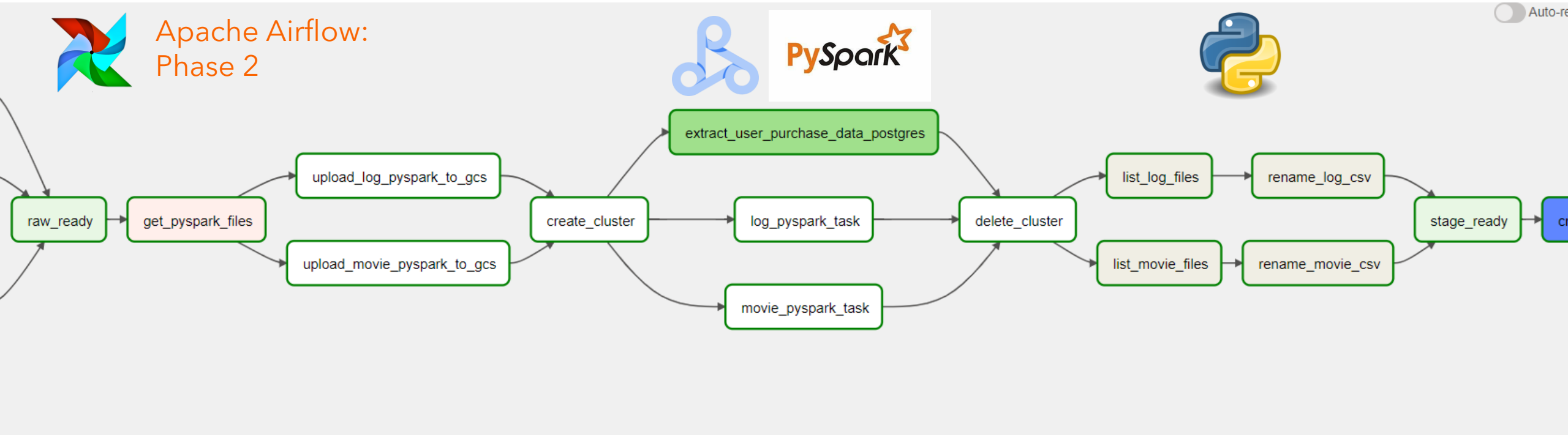
Data Output Messages Notifications

	invoice_number character varying (10)	stock_code character varying (20)	detail character varying (1000)	quantity integer	invoice_date timestamp without time zone	unit_price numeric (8,3)	customer_id integer	country character varying (20)
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.550	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.390	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.750	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.390	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.390	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.650	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.250	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.850	17850	United Kingdom

Pipelines – RAW to STAGING



Pipelines – RAW to STAGING



- **Transforming the data** by submitting a PySpark job into Dataproc
- The **PySpark code** is imported from Google Drive
- Extracting the **user_purchase** data from PostGres
- Combining the part files from Spark into single csv files using a python operator

Pipelines – RAW to STAGE

<input type="checkbox"/>	Name	Size
<input type="checkbox"/>	classified_movie_reviews.csv	1.3 MB
<input type="checkbox"/>	classified_movie_reviews.csv/	—
<input type="checkbox"/>	log_reviews_transformed.csv	6.9 MB
<input type="checkbox"/>	log_reviews_transformed.csv/	—
<input type="checkbox"/>	user_purchase.csv	45.4 MB



STAGE bucket

review_logs.csv



classified_movie_review.csv



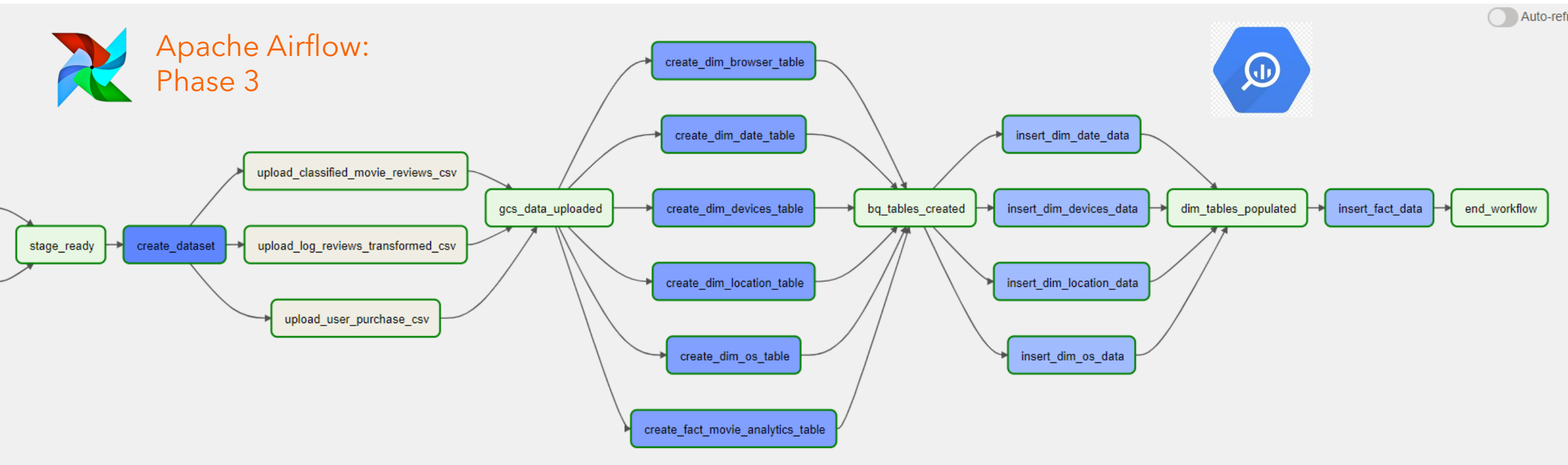
	A	B	C	D	E	F	G
1	log_id	log_date	device	os	location	ip	phone_number
2	1	04-25-2021	Mobile	Apple iOS	Kansas	9.200.232.57	821-540-5777
3	2	03-13-2021	Tablet	Google Android	Oregon	9.200.232.57	819-102-1320
4	3	09-30-2021	Tablet	Apple iOS	Minnesota	9.200.232.57	989-156-0498
5	4	05-24-2021	Tablet	Apple MacOS	Arkansas	9.200.232.57	225-837-9935
6	5	02-01-2021	Tablet	Linux	New Hampshire	9.200.232.57	243-842-4562
7	6	07-23-2021	Tablet	Apple iOS	Pensylvania	9.200.232.57	694-501-4352
8	7	10-13-2021	Computer	Apple MacOS	New York	9.200.232.57	430-449-7136
9	8	09-18-2021	Computer	Microsoft Windows	California	9.200.232.57	633-661-7714
10	9	05-08-2021	Tablet	Apple MacOS	Washington	9.200.232.57	450-036-0504
11	10	03-04-2021	Tablet	Linux	Lousiana	9.200.232.57	738-536-6776
12	11	06-19-2021	Computer	Google Android	Minnesota	9.200.232.57	682-519-9021
13	12	03-23-2021	Mobile	Google Android	Arkansas	9.200.232.57	678-722-6084
14	13	04-08-2021	Mobile	Linux	Pensylvania	9.200.232.57	781-850-8167
15	14	06-29-2021	Mobile	Microsoft Windows	Lousiana	9.200.232.57	208-216-2106
16	15	10-20-2021	Computer	Linux	Idaho	9.200.232.57	693-854-2646
17	16	06-03-2021	Mobile	Linux	Montana	9.200.232.57	805-540-1405
18	17	05-27-2021	Tablet	Apple MacOS	Nebraska	9.200.232.57	896-134-1623
19	18	04-01-2021	Tablet	Apple MacOS	Alabama	9.200.232.57	850-716-4779
20	19	06-12-2021	Tablet	Google Android	Missouri	9.200.232.57	946-565-9757
21	20	05-09-2021	Computer	Google Android	Rhode Island	9.200.232.57	938-623-1577
22	21	05-30-2021	Computer	Microsoft Windows	California	9.200.232.57	544-505-5479
23	22	07-31-2021	Tablet	Apple MacOS	Mississippi	9.200.232.57	461-476-4022
24	23	03-21-2021	Tablet	Apple iOS	New Hampshire	9.200.232.57	756-517-6671
25	24	10-28-2021	Mobile	Microsoft Windows	Texas	9.200.232.57	560-069-9075
26	25	05-11-2021	Mobile	Google Android	West Virginia	9.200.232.57	480-164-5985

	A	B	C
1	user_id	positive_r	review_id
2	13756	0	1
3	15738	0	2
4	15727	0	3
5	17954	0	4
6	16579	1	5
7	14841	0	6
8	18085	0	7
9	16365	1	8
10	17912	0	9
11	15100	1	10
12	16781	0	11
13	16656	1	12
14	14390	1	13
15	17975	0	14
16	17616	0	15
17	14589	0	16
18	17629	0	17
19	13089	0	18

Pipelines – STAGING to Data Warehouse

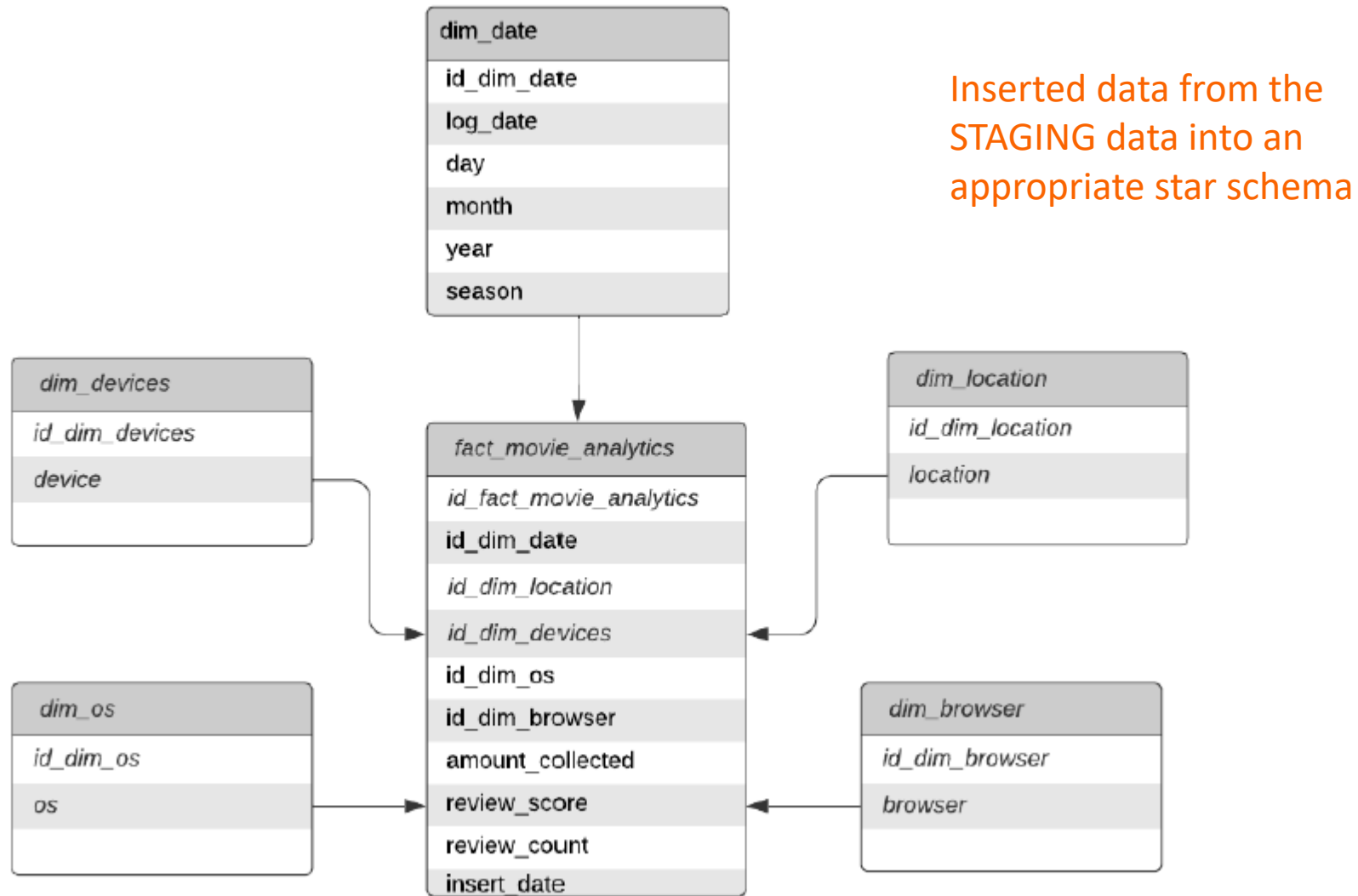


Apache Airflow:
Phase 3



- Uploaded data from GCS to BigQuery
- Created a Data Warehouse **star schema** (dimension and fact tables) in **Google BigQuery**.
- Inserted data from tables linked to staging data into dimension and fact tables

Pipelines –Data Warehouse Schema



BigQuery (DW) - Analytics

▼ wizeline-deb-capstone ☆ ⋮

▶ 🔍 Saved queries (1) ⋮

▶ ➡ External connections ⋮

▼ 📊 capstone_dataset ☆ ⋮

📊 classified_movie_reviews ☆ ⋮

📊 dim_date ☆ ⋮

📊 dim_devices ☆ ⋮

📊 dim_location ☆ ⋮

📊 dim_os ☆ ⋮

📊 fact_movie_analytics ☆ ⋮

📊 log_reviews_transformed ☆ ⋮

📊 user_purchase ☆ ⋮

```
1 SELECT COUNT(*) review_count FROM `wizeline-deb-capstone.capstone_dataset.log_reviews_transformed` AS lr
2 WHERE
3   lr.location IN ('California', 'NY', 'Texas') AND
4   lr.os LIKE '%Apple%' ;
5
```

Query results

JOB INFORMATION		RESULTS	CHART	PREVIEW	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	review_count						
1	1616						

```
1 SELECT COUNT(*) review_count, lr.device FROM `wizeline-deb-capstone.capstone_dataset.log_reviews_transformed` AS lr
2 WHERE
3   lr.location IN ('California', 'NY', 'Texas') AND
4   lr.os LIKE '%Apple%'
5 GROUP BY lr.device;
6
```

Query results

JOB INFORMATION		RESULTS	CHART	PREVIEW	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	review_count	device					
1	537	Mobile					
2	558	Tablet					
3	521	Computer					



Lessons Learned

- Airflow
- IaC / Terraform
- Cloud vs On-premise
- Data storage options
- Big Data processing