

ETL Pipeline: Batch Processing with Airflow

Inspired by: <https://github.com/srbsnkr/ETL-pipeline-Batch-processing-with-Airflow>

This project demonstrates the implementation of an ETL (Extract, Transform, Load) pipeline using Apache Airflow, Google Cloud Composer, Google Cloud Storage (GCS), BigQuery, and Cloud Functions. The pipeline automates the process of extracting trading data, transforming it, and loading it into BigQuery for analysis.

Project Structure:

ETL-pipeline-Batch-processing-with-Airflow/

- ├─ dag.py (Defines the Airflow DAG that orchestrates the ETL workflow)
- ├─ fetch_data.py (Contains the logic to fetch and prepare trading data for processing)
- ├─ metaData/
 - ├─ bq.json (Defines the BigQuery table schema)
 - └─ udf.js (Contains JavaScript UDFs for data transformation)
- ├─ cloud run function/ (Holds logic to pull csv from bucket via event trigger when files are uploaded)
 - ├─ main.py (Picks up every file uploaded by DAG and process it to create a dataflow job)
 - └─ requirments.txt (Contains package dependencies to run the main.py)
- ├─ tradingData.csv (Sample trading data used for testing the pipeline)
- └─ README.md (Provides an overview and instructions for the project)

Workflow Overview

Data Ingestion: The pipeline is triggered by the presence of tradingData.csv in a designated GCS bucket.

Data Transformation: A Cloud Function is invoked to launch a Dataflow job that applies transformations using the provided UDFs.

Data Loading: The transformed data is loaded into a BigQuery table as defined in bq.json.

Dashboard: The data loaded into the BigQuery table is further utilized to prepare data dashboards for different stakeholders.

Landing Storage:

Google Cloud Cloud Storage Bucket details

dev11-source-etl

Location: us-east1 (South Carolina) Storage class: Standard Public access: Not public Protection: Soft Delete

Objects Configuration Permissions Protection Lifecycle Observability **New** Inventory Reports Operations

Folder browser

dev11-source-etl

Create folder Upload Transfer data Other services

Filter by name prefix only Filter objects and folders

Name	Size	Type	Created	Storage class	Last modified	Public access	Version his
tradingData.csv	7.4 KB	text/csv	Jul 31, 2025, 12:07:24 PM	Standard	Jul 31, 2025, 12:07:24 PM	Not public	—

GCP Big data:

Free trial status: RM1,269.83 credit and 70 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud My First Project

Search (/) for resources, docs, products, and more

Explorer + Add data

Search BigQuery resources

Show starred only

avid-atlas-465409-b8

Repositories

Queries

Notebooks

Data canvases

Data preparations

Pipelines

External connections

dev1_gcs_sink

dev1-df-cf-tradin...

Untitled query

Run Save Download Share Schedule Open in More

1 SELECT * FROM `avid-atlas-465409-b8.dev1_gcs_sink.dev1-df-cf-trading-data` --242

Query completed

Query results

Save results Op

Job information	Results	Chart	JSON	Execution details	Execution graph		
Row	timestamp	Open	High	Low	Close	Volume	Open_interes
1	2020-02-29 18:30:00 UTC	21.0	23.4	15.1	19.95	315302519.0	0.0
2	2015-07-31 18:30:00 UTC	18.89	19.45	15.55	16.45	95927560.0	0.0
3	2015-08-31 18:30:00 UTC	16.35	16.89	15.9	16.55	80265192.0	0.0
4	2015-09-30 18:30:00 UTC	16.64	18.5	16.6	18.14	52324181.0	0.0
5	2015-10-31 18:30:00 UTC	18.14	19.35	17.14	19.14	66614311.0	0.0
6	2015-06-30 18:30:00 UTC	19.6	19.89	18.5	18.89	63904266.0	0.0
7	2015-11-30 18:30:00 UTC	19.35	21.35	18.6	21.05	94064974.0	0.0
8	2015-12-31 18:30:00 UTC	21.1	22.4	18.65	20.95	94931538.0	0.0
9	2020-03-31 18:30:00 UTC	19.9	23.55	19.05	20.8	134760873.0	0.0
10	2016-01-31 18:30:00 UTC	21.1	21.95	19.15	20.1	109012986.0	0.0
11	2020-05-31 18:30:00 UTC	19.75	21.25	19.4	20.0	145458636.0	0.0
12	2020-04-30 18:30:00 UTC	20.7	20.7	19.4	19.6	55568471.0	0.0
13	2020-10-31 18:30:00 UTC	20.0	22.15	19.7	21.45	91597173.0	0.0

Composer:

Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Dismiss Activate

Google Cloud My First Project

Search (/) for resources, docs, products, and more

Composer / Environments

Environments Create Refresh Delete

Join Airflow community on October 7th - 9th during the Airflow Summit 2025 conference to learn more about Airflow and share your expertise. Register here

Dismiss

Filter environments

State	Name	Location	Composer version	Airflow version	Creation time	Update time	Airflow webserver	DAG list	Logs	DAGs folder	Labels
On	airflow	us-east1	3	2.10.5-build.10	7/30/25, 12:11 PM	7/30/25, 12:32 PM	Airflow	DAGs	Logs	DAGs	None

Composer → DAG Folder:

Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use. Dismiss Activate

Google Cloud My First Project Search (/) for resources, docs, products, and more Search

Cloud Storage Bucket details Go to path Refresh

Overview Buckets Monitoring Settings Storage Intelligence Insights datasets Configuration

us-east1-airflow-d1137eb1-bucket

Location: us-east1 (South Carolina) Storage class: Standard Public access: Subject to object ACLs Protection: Soft Delete

Objects Configuration Permissions Protection Lifecycle Observability New Inventory Reports Operations

Folder browser

us-east1-airflow-d1137eb1-bucket

dag

scripts

data

logs

plugins

Create folder Upload Transfer data Other services

Filter by name prefix only Filter Filter objects and folders Show Live objects only

Name	Size	Type	Created	Storage class	Last modified	Public access
airflow_monitoring.py	960 B	text/x-python	Jul 30, 2025, 12:31:24 PM	Standard	Jul 30, 2025, 12:31:24 PM	Not public
dag.py	786 B	application/octet-stream	Jul 30, 2025, 12:54:28 PM	Standard	Jul 30, 2025, 12:54:28 PM	Not public
scripts/	—	Folder	—	—	—	—

Place dag.py inside it:

Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use. Dismiss Activate

Google Cloud My First Project Search (/) for resources, docs, products, and more Search

Cloud Storage Bucket details Go to path Refresh Learn

Overview Buckets Monitoring Settings Storage Intelligence Insights datasets Configuration

us-east1-airflow-d1137eb1-bucket

Location: us-east1 (South Carolina) Storage class: Standard Public access: Subject to object ACLs Protection: Soft Delete

Objects Configuration Permissions Protection Lifecycle Observability New Inventory Reports Operations

Folder browser

us-east1-airflow-d1137eb1-bucket

dag

scripts

data

logs

plugins

Create folder Upload Transfer data Other services

Filter by name prefix only Filter Filter objects and folders Show Live objects only

Name	Size	Type	Created	Storage class	Last modified	Public access
airflow_monitoring.py	960 B	text/x-python	Jul 30, 2025, 12:31:24 PM	Standard	Jul 30, 2025, 12:31:24 PM	Not public
dag.py	786 B	application/octet-stream	Jul 30, 2025, 12:54:28 PM	Standard	Jul 30, 2025, 12:54:28 PM	Not public
scripts/	—	Folder	—	—	—	—



dag.py

Create a new folder named scripts and place fetch_data.py inside:

Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use. Dismiss Activate

Google Cloud My First Project Search (/) for resources, docs, products, and more Search

Cloud Storage Bucket details Go to path Refresh Learn

Overview Buckets Monitoring Settings Storage Intelligence Insights datasets Configuration

us-east1-airflow-d1137eb1-bucket

Location: us-east1 (South Carolina) Storage class: Standard Public access: Subject to object ACLs Protection: Soft Delete

Objects Configuration Permissions Protection Lifecycle Observability New Inventory Reports Operations

Folder browser

us-east1-airflow-d1137eb1-bucket

dag

scripts

data

logs

plugins

Create folder Upload Transfer data Other services

Filter by name prefix only Filter Filter objects and folders Show Live objects only

Name	Size	Type	Created	Storage class	Last modified	Public access
fetch_data.py	1.8 KB	application/octet-stream	Jul 30, 2025, 2:14:28 PM	Standard	Jul 30, 2025, 2:14:28 PM	Not public



Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Dismiss

Google Cloud

My First Project

composer

Search

Composer / Environment: airflow / DAGs

Environment details

Open Airflow UI

Open DAGs folder

Save snapshot

Load snapshot

Refresh

Delete

airflow

This environment is running

Monitoring

Logs

DAGs

Environment configuration

Airflow configuration overrides

Environment variables

Labels

Pypi packages

Filter

Filter DAGs

DAG id	State	Description	Schedule interval	Last completed run	Active runs	Successful runs (1h)	Failed runs (1h)
airflow_monitoring	Active	liveness monitoring dag	*/*/*	9 minutes ago	0	6	0
fetch_nse_trading_stats	Active	Runs an external Python script	@daily	2 minutes ago	0	1	0

Cloud run function:

Free trial status: RM1,214.91 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Dismiss

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Run

Services

Jobs

Worker pools

Domain mappings

A service exposes a unique endpoint and automatically scales the underlying infrastructure to handle incoming requests. Deploy a container image, source code or a function to create a service.

Filter

Deployment type: Function

Filter services

Name	Deployment type	Req/sec	Region	Authentication	Ingress	Last deployed	Deployed by	Recommendation
dev11-cf-dataflow	Function	0	us-east1	Public access	All	2 hours ago	arif.warsi.uk@gmail.com	Security

Free trial status: RM1,269.83 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Dismiss

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Run

Services

Jobs

Worker pools

Domain mappings

Service details

Edit & deploy new revision

Connect to repo

Test

Deploying revision

Hide status

Building source (see logs)

Completed

Updating service

Completed

Creating revision

Completed

Routing traffic

Completed

dev11-cf-dataflow

Region: us-east1

URL: https://dev11-cf-dataflow-715856575099.us-east1.run.app

Scaling: Auto (Min: 0)

Metrics

SLOs

Logs

Revisions

Source

Triggers

Networking

Security

YAML

Source

Base image: Python 3.13 (Ubuntu 22)

Function entry point: hello_gcs

Edit source

main.py

requirements.txt

```
1 import functions_framework
2 # added by shiraz start *****
3 import httplib2
4 # added by shiraz end *****
5 from googleapiclient.discovery import build
6 # added by shiraz start *****
7 from oauth2client.client import GoogleCredentials
8 # added by shiraz end *****
9 import json
10 from google.auth import default
11 from google.cloud import storage
12
13 @functions_framework.cloud_event
14 def hello_gcs(cloud_event):
15
16     # Extract metadata from the event
17     data = cloud_event.data
18     event_id = cloud_event["id"]
19     event_type = cloud_event["type"]
20
21     bucket_name = data["bucket"]
22     file_name = data["name"]
23     metageneration = data.get("metageneration")
24     time_created = data.get("timeCreated")
25     updated = data.get("updated")
26
27     region = "us-east1"
28     project = "avid-atlas-465489-b8"
29     template_path = "gs://dataflow-templates-us-east1/latest/GCS_Text_to_BigQuery"
```

Release Notes

Download ZIP



main.py

Free trial status: RM1,269.83 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Run

Services

Jobs

Worker pools

Domain mappings

Service details

Edit & deploy new revision

Connect to repo

Test

Deploying revision

Hide status

Building source (see logs)

Completed

Updating service

Completed

Creating revision

Completed

Routing traffic

Completed

dev11-cf-dataflow

Region: us-east1

URL: <https://dev11-cf-dataflow-715856575099.us-east1.run.app>

Scaling: Auto (Min: 0)

Metrics

SLOs

Logs

Revisions

Source

Triggers

Networking

Security

YAML

Source

Base image: Python 3.13 (Ubuntu 22)

Function entry point: hello_gcs

Edit source

main.py

requirements.txt

1

Functions-framework==3.*

2

google-auth-http1b2

3

oauth2client

4

google-api-python-client==2.105.0

5

google-auth==2.27.0

6

google-cloud-storage==2.14.0



requirements.txt

Setup Trigger:

Free trial status: RM1,269.83 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Run

Services

Jobs

Worker pools

Domain mappings

Service details

Edit & deploy new revision

Connect to repo

Test

Deploying revision

Hide status

Building source (see logs)

Completed

Updating service

Completed

Creating revision

Completed

Routing traffic

Completed

dev11-cf-dataflow

Region: us-east1

URL: <https://dev11-cf-dataflow-715856575099.us-east1.run.app>

Scaling: Auto (Min: 0)

Metrics

SLOs

Logs

Revisions

Source

Triggers

Networking

Security

YAML

Triggers

+ Add trigger

Eventarc trigger

Name

trigger-jwfo8pn

Event provider

Cloud Storage

Event type

google.cloud.storage.object.v1.finalized

Receive events from

[dev11-source-etl \(us-east1\)](#)

Destination

[dev11-cf-dataflow \(us-east1\)](#)

Service URL path

/

Service account

[715856575099-compute@developer.gserviceaccount.com](#)

Invocations

Latency

1 hour

6 hours

1 d

Create alerting

Release Notes

Assigning relevant roles to Service Account:

Free trial status: RM824.27 credit and 63 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google CloudMy First ProjectSearch (/) for resources, docs, products, and moreSearch

IAM & Admin / IAM

IAM

PAM

Principal Access Boun...

Organizations

Identity & Organization

Policy Troubleshooter

Policy Analyzer

Organization Policies

Service Accounts

Workload Identity Fede...

Workforce Identity Fed...

Labels

Tags

Settings

Privacy & Security

Identity-Aware Proxy

IAM

AllowDenyRecommendations history

Permissions for project "My First Project"

These permissions affect this project and all of its resources. [Learn more](#)

View by principalsView by roles

Grant accessRemove access

Filter Enter property name or value

Type	Principal	Name	Role	Security insights
<input type="checkbox"/>	715856575099-compute@developer.gservicesaccount.com	Default compute service account	Cloud Build Service Account	
<input type="checkbox"/>			Cloud Dataflow Service Agent	
<input type="checkbox"/>			Dataflow Developer	
<input type="checkbox"/>			Editor	
<input type="checkbox"/>			Eventarc Event Receiver	
<input type="checkbox"/>			Storage Object Admin	

Free trial status: RM1,269.83 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google CloudMy First ProjectSearch (/) for resources, docs, products, and moreSearch

Cloud Run

Services

Jobs

Worker pools

Domain mappings

Service details

Edit & deploy new revision

Connect to repo

Test

Deploying revision

Hide status

Building source (see logs)Completed

Updating serviceCompleted

Creating revisionCompleted

Routing trafficCompleted

dev11-cf-dataflow

Region: us-east1

URL: <https://dev11-cf-dataflow-715856575099.us-east1.run.app>

Scaling: Auto (Min: 0)

Metrics

SLOs

Logs

Revisions

Source

Triggers

Networking

Security

YAML

Logs

Severity Default

Filter Search all fields and values

Severity	Timestamp	Summary
>	2025-07-31 12:07:27.889 BST	}
>	2025-07-31 12:07:27.889 BST	}
>	2025-07-31 12:07:28.018 BST	Dataflow job launched successfully:
>	2025-07-31 12:07:28.018 BST	{
>	2025-07-31 12:07:28.018 BST	"job": {
>	2025-07-31 12:07:28.018 BST	"id": "2025-07-31_04_07_27-6736734863772400586",
>	2025-07-31 12:07:28.018 BST	"projectId": "avid-atlas-465409-b8",
>	2025-07-31 12:07:28.018 BST	"name": "cf-bq-load-15778222",
>	2025-07-31 12:07:28.018 BST	"type": "JOB_TYPE_BATCH",
>	2025-07-31 12:07:28.018 BST	"currentStateTime": "1978-01-01T00:00:00Z",
>	2025-07-31 12:07:28.018 BST	"createTime": "2025-07-31T11:07:28.0038732Z",
>	2025-07-31 12:07:28.018 BST	"location": "us-east1",
>	2025-07-31 12:07:28.018 BST	"startTime": "2025-07-31T11:07:28.0038732Z"
>	2025-07-31 12:07:28.018 BST	}

Transformation logic:

Create a bucket to hold transformation logic in files bq.json and udf.js

Free trial status: RM1,226.81 credit and 69 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google CloudMy First ProjectSearch (/) for resources, docs, products, and moreSearch

Cloud Storage

Overview

Buckets

Monitoring

Settings

dev11-trading-dataflow-metadata

Location us-east1 (South Carolina)

Storage class Standard

Public access Not public

Protection Soft Delete

Objects

Configuration

Permissions

Protection

Lifecycle

Observability New

Inventory Reports

Operations

Folder browser

dev11-trading-dataflow-metadata

temp/

Buckets > dev11-trading-dataflow-metadata

Create folderUploadTransfer dataOther services

Filter by name prefix onlyFilter objects and folders

Name	Size	Type	Created	Storage class	Last modified	Public acces
bq.json	462 B	application/json	Jul 23, 2025, 2:30:01 PM	Standard	Jul 23, 2025, 2:30:01 PM	Not public
temp/	—	Folder	—	—	—	—
udf.js	355 B	text/javascript	Jul 23, 2025, 2:30:02 PM	Standard	Jul 23, 2025, 2:30:02 PM	Not public

bq.json

udf.js

Analytics:

Display Trading data on time series graph.

