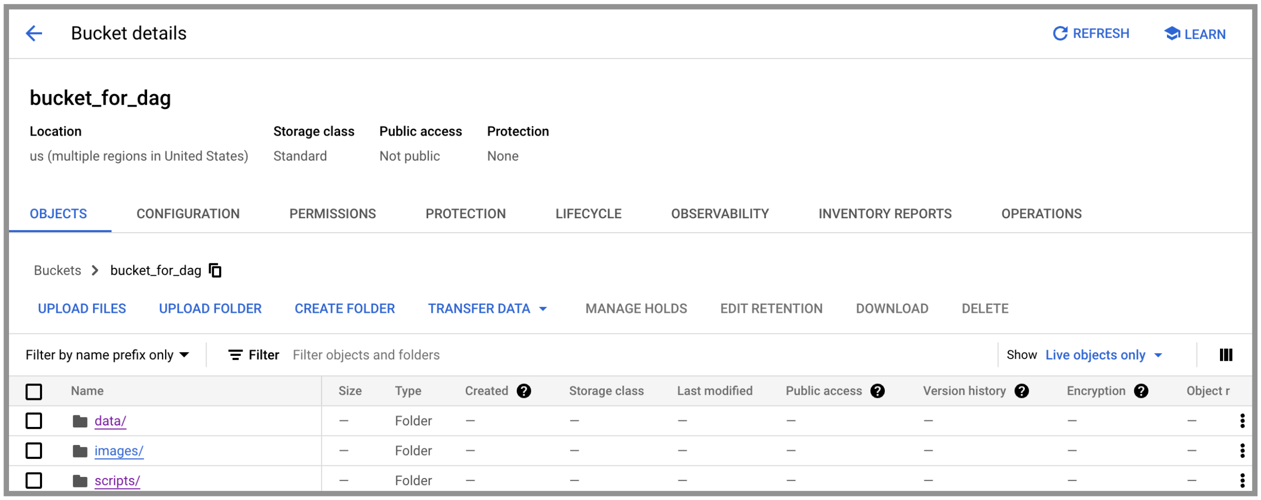
Airflow Initial Configuration Guide:

1. Open a bucket containing the folders :



bucket_for_dag

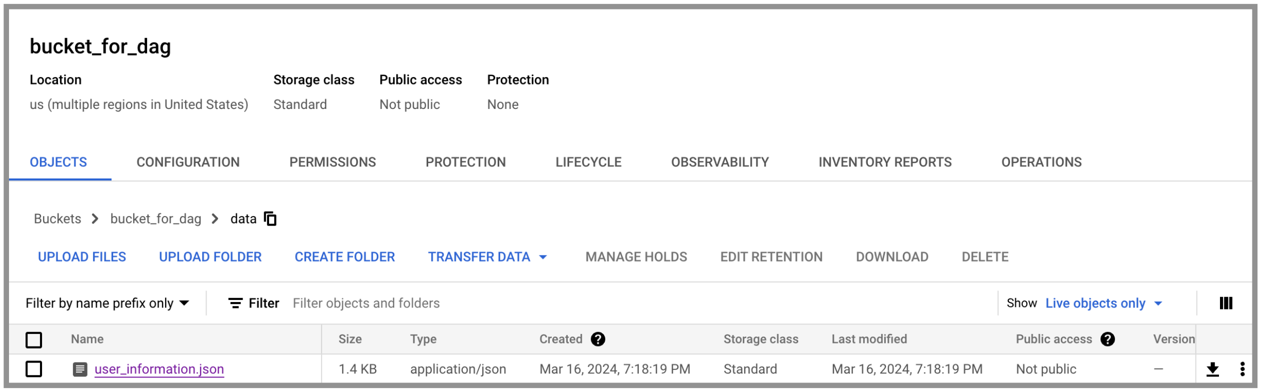
Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > bucket_for_dag

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Object r
□	data/	—	Folder	—	—	—	—	—	—	—
□	images/	—	Folder	—	—	—	—	—	—	—
□	scripts/	—	Folder	—	—	—	—	—	—	—

2. Upload the **'user_information.json'** generated by the **'authorization_code_flow.py'** to the directory:

- **gs://bucket_for_dag/data**



bucket_for_dag

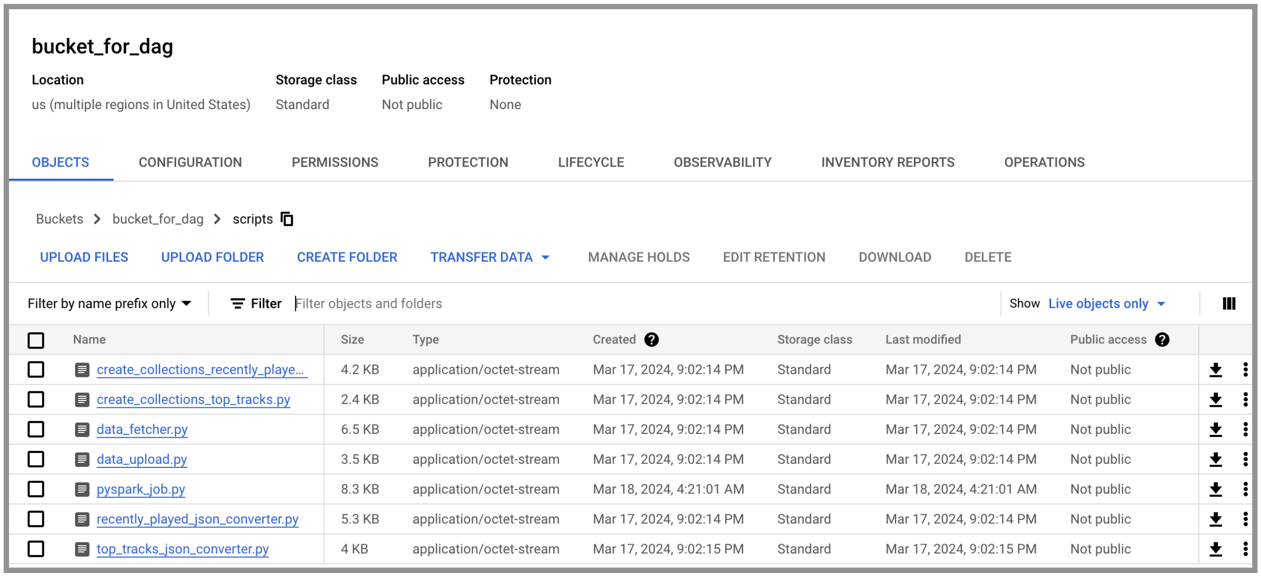
Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > bucket_for_dag > data

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version
□	user_information.json	1.4 KB	application/json	Mar 16, 2024, 7:18:19 PM	Standard	Mar 16, 2024, 7:18:19 PM	Not public	—

3. Upload the scripts to the directory:

- **gs://bucket_for_dag/scripts**



bucket_for_dag

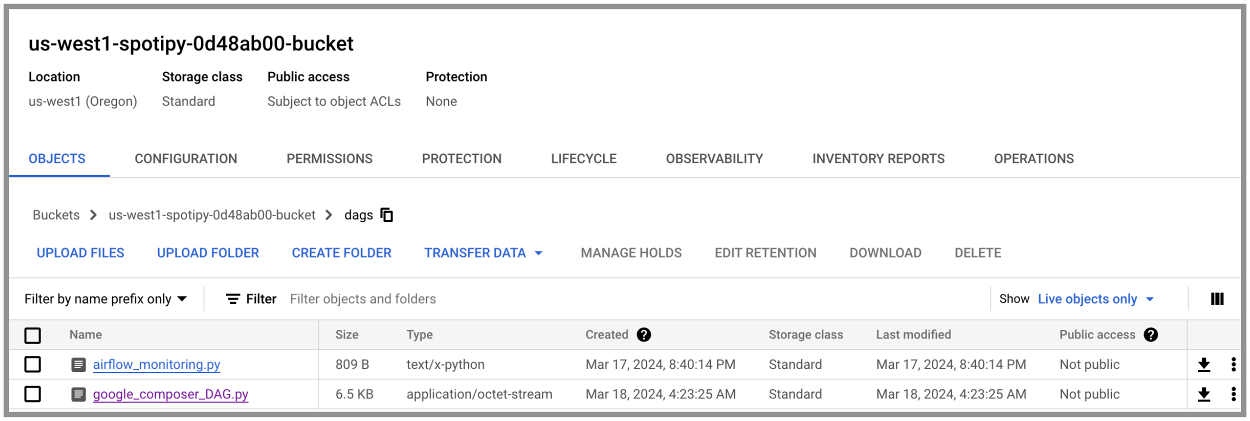
Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > bucket_for_dag > scripts

□	Name	Size	Type	Created	Storage class	Last modified	Public access
□	create_collections_recently_playe…	4.2 KB	application/octet-stream	Mar 17, 2024, 9:02:14 PM	Standard	Mar 17, 2024, 9:02:14 PM	Not public
□	create_collections_top_tracks.py	2.4 KB	application/octet-stream	Mar 17, 2024, 9:02:14 PM	Standard	Mar 17, 2024, 9:02:14 PM	Not public
□	data_fetcher.py	6.5 KB	application/octet-stream	Mar 17, 2024, 9:02:14 PM	Standard	Mar 17, 2024, 9:02:14 PM	Not public
□	data_upload.py	3.5 KB	application/octet-stream	Mar 17, 2024, 9:02:14 PM	Standard	Mar 17, 2024, 9:02:14 PM	Not public
□	pyspark_job.py	8.3 KB	application/octet-stream	Mar 18, 2024, 4:21:01 AM	Standard	Mar 18, 2024, 4:21:01 AM	Not public
□	recently_played_json_converter.py	5.3 KB	application/octet-stream	Mar 17, 2024, 9:02:14 PM	Standard	Mar 17, 2024, 9:02:14 PM	Not public
□	top_tracks_json_converter.py	4 KB	application/octet-stream	Mar 17, 2024, 9:02:15 PM	Standard	Mar 17, 2024, 9:02:15 PM	Not public

4. Create a Cloud Composer and upload **'google_composer_DAG.py'** to the directory:

- **/home/airflow/gcs/dags**



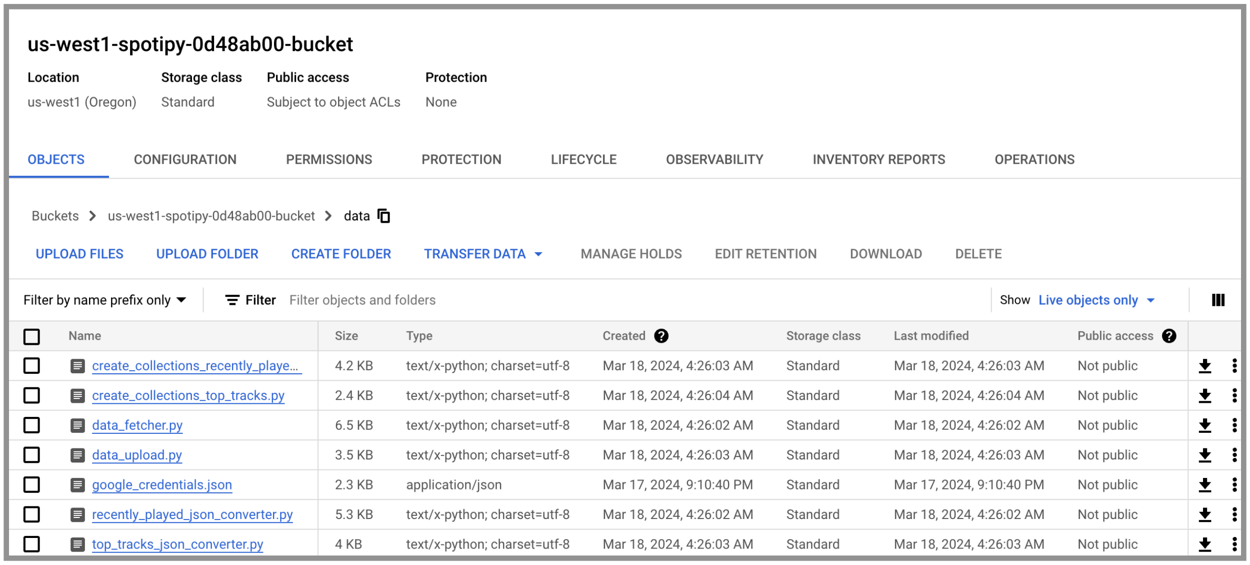
us-west1-spotipy-0d48ab00-bucket

Location	Storage class	Public access	Protection
us-west1 (Oregon)	Standard	Subject to object ACLs	None

Name	Size	Type	Created	Storage class	Last modified	Public access
airflow_monitoring.py	809 B	text/x-python	Mar 17, 2024, 8:40:14 PM	Standard	Mar 17, 2024, 8:40:14 PM	Not public
google_composer_DAG.py	6.5 KB	application/octet-stream	Mar 18, 2024, 4:23:25 AM	Standard	Mar 18, 2024, 4:23:25 AM	Not public

5. Create and download an authentication key (JSON file) from service accounts. Upload it to the directory:

- **/home/airflow/gcs/data**

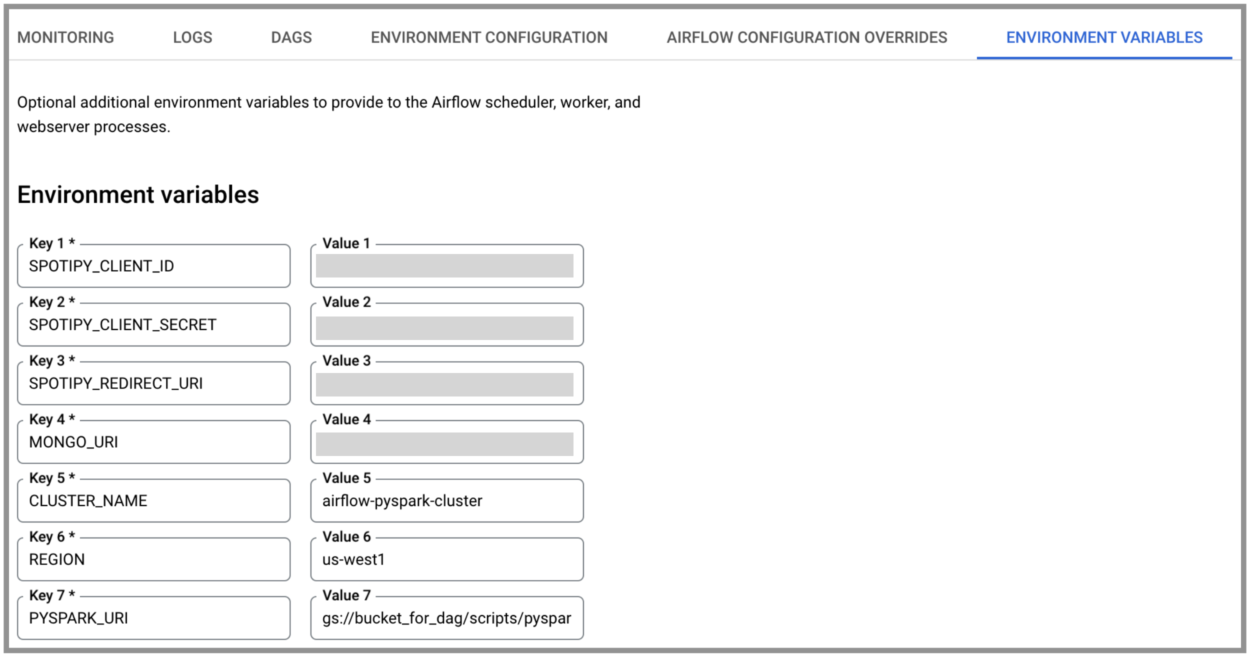


us-west1-spotipy-0d48ab00-bucket

Location	Storage class	Public access	Protection
us-west1 (Oregon)	Standard	Subject to object ACLs	None

Name	Size	Type	Created	Storage class	Last modified	Public access
create_collections_recently_playe...	4.2 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:03 AM	Standard	Mar 18, 2024, 4:26:03 AM	Not public
create_collections_top_tracks.py	2.4 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:04 AM	Standard	Mar 18, 2024, 4:26:04 AM	Not public
data_fetcher.py	6.5 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:02 AM	Standard	Mar 18, 2024, 4:26:02 AM	Not public
data_upload.py	3.5 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:03 AM	Standard	Mar 18, 2024, 4:26:03 AM	Not public
google_credentials.json	2.3 KB	application/json	Mar 17, 2024, 9:10:40 PM	Standard	Mar 17, 2024, 9:10:40 PM	Not public
recently_played_json_converter.py	5.3 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:02 AM	Standard	Mar 18, 2024, 4:26:02 AM	Not public
top_tracks_json_converter.py	4 KB	text/x-python; charset=utf-8	Mar 18, 2024, 4:26:03 AM	Standard	Mar 18, 2024, 4:26:03 AM	Not public

6. Define the following environment variables:

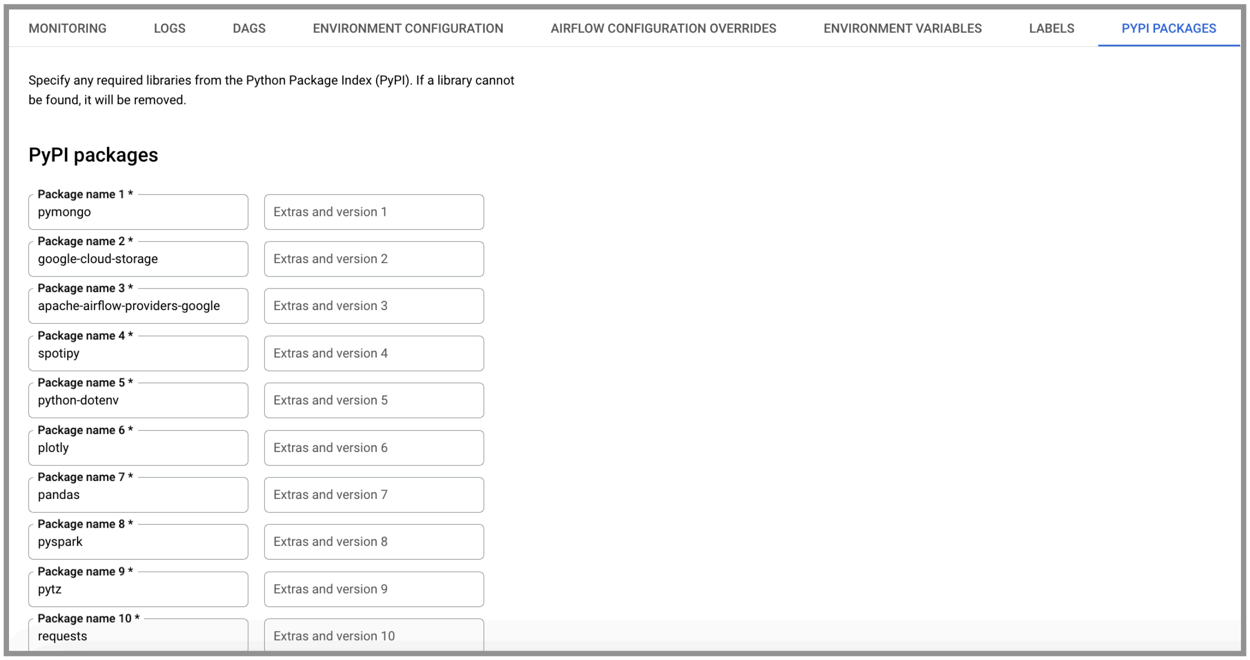


Optional additional environment variables to provide to the Airflow scheduler, worker, and webserver processes.

Environment variables

Key	Value
Key 1 * SPOTIPY_CLIENT_ID	Value 1 [redacted]
Key 2 * SPOTIPY_CLIENT_SECRET	Value 2 [redacted]
Key 3 * SPOTIPY_REDIRECT_URI	Value 3 [redacted]
Key 4 * MONGO_URI	Value 4 [redacted]
Key 5 * CLUSTER_NAME	Value 5 airflow-pyspark-cluster
Key 6 * REGION	Value 6 us-west1
Key 7 * PYSPARK_URI	Value 7 gs://bucket_for_dag/scripts/pyspar

7. Add the following PyPI packages to the environment:



Specify any required libraries from the Python Package Index (PyPI). If a library cannot be found, it will be removed.

PyPI packages

Package name 1 * pymongo	Extras and version 1
Package name 2 * google-cloud-storage	Extras and version 2
Package name 3 * apache-airflow-providers-google	Extras and version 3
Package name 4 * spotipy	Extras and version 4
Package name 5 * python-dotenv	Extras and version 5
Package name 6 * plotly	Extras and version 6
Package name 7 * pandas	Extras and version 7
Package name 8 * pyspark	Extras and version 8
Package name 9 * pytz	Extras and version 9
Package name 10 * requests	Extras and version 10

8. Enable the Dataproc Cluster API.