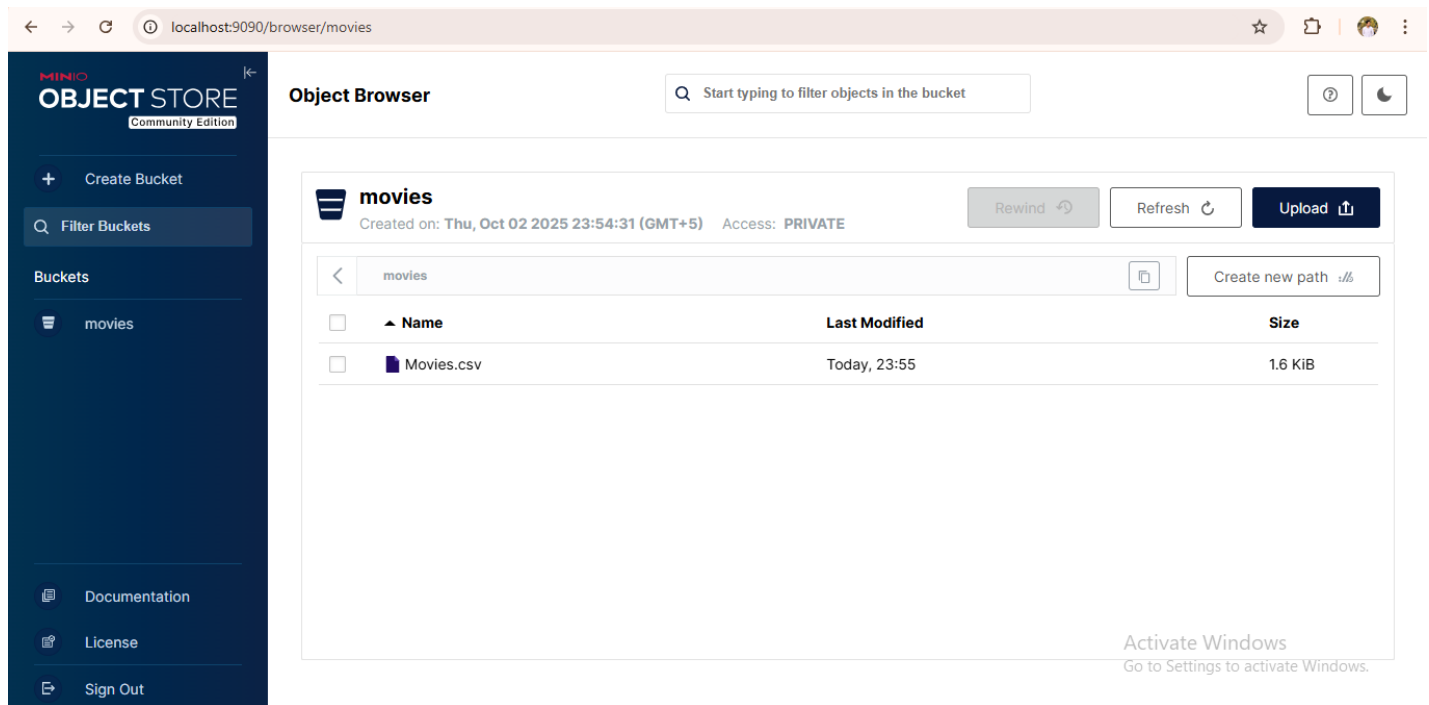


Buildables Task 5

Github Repo: <https://github.com/afnank070/Data-Engineering-Buildables-Fellowship>

MinIO bucket with uploaded CSV



Extracted CSV from MinIO

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS D:\Data-Engineering-Buildables-Fellowship\Task-5> python extract_csv.py
movie_title num_critic_for_reviews duration ... ACTOR_2_facebook_likes imdb_score title_year.1
0 Avatar?y 723 178.0 ... 936.0 7.9 2009.0

[1 rows x 16 columns]
PS D:\Data-Engineering-Buildables-Fellowship\Task-5> |
```

For now, in Transform step we will:

1. Fix missing values → fill empty numeric fields with 0.
2. Ensure numeric columns are numbers.
3. Remove duplicate column title_year.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS D:\Data-Engineering-Buildables-Fellowship\Task-5> python transform_csv.py
Extracted data (first row):
  movie_title  num_critic_for_reviews  duration  ...  ACTOR_2_facebook_likes  imdb_score  title_year.1
0  Avatar?y  723  178.0  ...  936.0  7.9  2009.0

[1 rows x 16 columns]

Transformed data (first row):
  movie_title  num_critic_for_reviews  duration  ...  ACTOR_2_facebook_likes  imdb_score  title_year.1
0  Avatar?y  723  178.0  ...  936.0  7.9  2009.0

[1 rows x 16 columns]

Transformed CSV saved as movies_transformed.csv
PS D:\Data-Engineering-Buildables-Fellowship\Task-5>
```

Loaded in PostgreSQL

```
PS D:\Data-Engineering-Buildables-Fellowship\Task-5> python load_csv.py
Data to load (first row):
  movie_title  num_critic_for_reviews  duration  ...  ACTOR_2_facebook_likes  imdb_score  title_year.1
0  Avatar?y  723  178.0  ...  936.0  7.9  2009.0

[1 rows x 16 columns]

Data loaded into PostgreSQL successfully!
PS D:\Data-Engineering-Buildables-Fellowship\Task-5>
```

pgAdmin 4

File Object Tools Help

Object Explorer

- Catalogs
- Event Triggers
- Extensions
- Foreign Data Wrappers
- Languages
- Publications
- Schemas (1)
 - public
 - Aggregates
 - Collations
 - Domains
 - FTS Configurations
 - FTS Dictionaries
 - FTS Parsers
 - FTS Templates
 - Foreign Tables
 - Functions
 - Materialized Views
 - Operators
 - Procedures
 - Sequences
 - Tables (1)
 - movies
 - Columns (15)
 - movie_title
 - num_critic_for_reviews

public.movies/etl_demo/postgres@PostgreSQL 18

Query

```
1 SELECT * FROM public.movies
2 LIMIT 100
3
```

Data Output

	movie_title character varying (255)	num_critic_for_reviews integer	duration integer	director_facebook_likes integer	actor_3_facebook_likes integer	actor_1_facebook_likes integer	gross bigint
1	Avatar?y	723	178	0	855	0	76050
2	Pirates of the Caribbean: At World's End?y	302	0	0	1000	0	3094
3	Spectre?y	602	148	0	161	0	2000
4	The Dark Knight Rises?y	813	0	0	23000	0	4481
5	John Carter?y	462	132	0	530	0	730
6	Spider-Man 3?y	392	156	0	4000	0	3365
7	Tangled?y	324	0	0	284	0	2008
8	Avengers: Age of Ultron?y	635	141	0	19000	0	4589

Total rows: 14 of 14 Query complete 00:00:00.204 Ln 1, Col 1

Airflow Dashboard:

←→🔄🔍localhost:5000

☆🔖👤⋮

ETL Pipeline Dashboard

Apache Airflow ETL Pipeline Status

Refresh

Pipeline Status: OPERATIONAL

Last Updated: 2025-10-06 01:35:46

DAG: etl_pipeline (Extract → Transform → Load)

Database Status

Total Records: 14

Database: PostgreSQL (etl_demo)

Table: movies

←→🔄🔍localhost:5000

☆🔖👤⋮

Sample Movie Data

Movie Title	IMDB Score	Year	Budget
The Dark Knight Rises	8.5	2012	\$250,000,000
Avatar	7.9	2009	\$237,000,000
Tangled	7.8	2010	\$260,000,000
Avengers: Age of Ultron	7.5	2015	\$250,000,000
Avengers: Age of Ultron	7.5	2015	\$250,000,000

ETL Pipeline Components

- **Extract:** CSV file processing (simulating MinIO)
- **Transform:** Data cleaning and validation with pandas
- **Load:** Insert into PostgreSQL database

Activate Windows

Go to Settings to activate Windows.