

Projet ETL Dagster – S7 ISIBD ENSAB

Réalisé par : Hamza Elmourabit

Encadré par : Pr .Lamia Karim

15 décembre 2025



Table des matières

1	Présentation du projet	2
2	Structure du projet	2
3	Code Python (<code>assets/etl_{assets}.py</code>)	2
4	Configuration (<code>workspace.yaml</code>)	2
5	Instructions d'exécution	3
6	Résultats	3

1 Présentation du projet

Ce projet consiste à construire un pipeline ETL (Extract, Transform, Load) avec **Dagster**. L'objectif est de lire un fichier CSV (input.csv), effectuer des transformations et sauvegarder le résultat dans output.csv. Le job peut être exécuté manuellement ou via un schedule automatique.

2 Structure du projet

```
dagster_etl_scheduled/  
  .venv/                # Environnement virtuel  
  assets/  
    etl_assets.py       # Code du pipeline  
  workspace.yaml        # Config Dagster  
  input.csv             # Données sources  
  output.csv            # Résultat transformé
```

3 Code Python (assets/etl_assets.py)

```
1 import pandas as pd  
2 from dagster import asset, job, materialize  
3  
4 # Extract : charger les données depuis input.csv  
5 @asset  
6 def load_data():  
7     df = pd.read_csv('input.csv')  
8     print('Data loaded:\n', df.head())  
9     return df  
10  
11 # Transform : ajouter colonne 'total'  
12 @asset  
13 def transform_data(load_data):  
14     df = load_data.copy()  
15     df['total'] = df['quantity'] * df['price']  
16     print('Data transformed:\n', df.head())  
17     return df  
18  
19 # Load : sauvegarder les données transformées dans output.csv  
20 @asset  
21 def save_data(transform_data):  
22     df = transform_data  
23     df.to_csv('output.csv', index=False)  
24     print('Data saved to output.csv')  
25     return 'Success'
```

4 Configuration (workspace.yaml)

load_from :

- python_file :
 relative_path : assets/etl_assets.py

5 Instructions d'exécution

1. Activer l'environnement virtuel :

```
1 & .\venv\Scripts\Activate.ps1
```

2. Définir DAGSTER_HOME :

```
1 $env:DAGSTER_HOME = "$HOME\dagster_home"
```

3. Lancer le Webserver Dagster :

```
1 dagster -webserver -w workspace.yaml -p 3001
```

4. Lancer le daemon pour le schedule :

```
1 dagster -daemon run -w workspace.yaml
```


5. Vérifier le fichier output.csv après exécution.

6 Résultats

Après exécution du pipeline, le fichier output.csv contient les données transformées avec la nouvelle colonne total.

The screenshot displays the Dagster web interface for a pipeline run. The top navigation bar includes 'Overview', 'Runs', 'Catalog', 'Jobs', 'Automation', 'Lineage', and 'Deployment'. The 'Runs' section is active, showing a list of runs with columns for 'TIMESTAMP', 'OP', 'EVENT TYPE', and 'INFO'. The 'save_data' operator is highlighted, showing its execution details. The 'INFO' column for 'save_data' shows the output path: 'C:\Users\khadi\dagster_home\storage\c0565df4-cd4f-4c55-963c-0bf518b31010\save_data\result'. The 'EVENT TYPE' column shows 'LOADED_INPUT', 'STEP_INPUT', 'STEP_OUTPUT', 'HANDLED_OUTPUT', 'STEP_SUCCESS', 'ENGINE_EVENT', 'RUN_SUCCESS', and 'ENGINE_EVENT'. The 'INFO' column for 'STEP_SUCCESS' shows 'Finished execution of step "save_data" in 168ms.' The 'ENGINE_EVENT' for 'RUN_SUCCESS' shows 'Finished execution of run for "etl_job".' The 'ENGINE_EVENT' for 'ENGINE_EVENT' shows 'Process for run exited (pid: 29992)'.

TIMESTAMP	OP	EVENT TYPE	INFO
22:28:14,165	save_data	LOADED_INPUT	Loaded input "transform_data" using input manager "io_manager", from output "result" of step "transform_data"
22:28:14,184	save_data	STEP_INPUT	Got input "transform_data" of type "Any". (Type check passed).
22:28:14,210	save_data	STEP_OUTPUT	Yielded output "result" of type "Any". (Type check passed).
22:28:14,264	save_data	HANDLED_OUTPUT	Handled output "result" using IO manager "io_manager" path: C:\Users\khadi\dagster_home\storage\c0565df4-cd4f-4c55-963c-0bf518b31010\save_data\result
22:28:14,283	save_data	STEP_SUCCESS	Finished execution of step "save_data" in 168ms.
22:28:18,456	-	ENGINE_EVENT	Multiprocess executor: parent process exiting after 21.19s (pid: 29992) pid: 29992
22:28:18,473	-	RUN_SUCCESS	Finished execution of run for "etl_job".
22:28:18,583	-	ENGINE_EVENT	Process for run exited (pid: 29992).



Overview

Runs

Catalog

Jobs

Automation

Lineage


Deployment

Search

Hide navigation

Settings

Support



Overview

Runs

Catalog

Jobs

Automation

Lineage

Deployment

Search

Hide navigation

Settings













Support


AllBackfillsQueued (3)In progress (0)FailedScheduled

0:02

FilterShow runs within backfills

←NewerOlder→Actions

	ID	Target	Launched by	Status	Created at	Duration	
<input type="checkbox"/>	<div>c0565df4</div> <div>View all tags (2)</div>	 etl_job	 Manually launched	Success	15 déc., 22:27	0:00:21	<div>View</div>
<input type="checkbox"/>	<div>0adc5fa5</div> <div>View all tags (2)View queue criteria</div>	 etl_job	 Manually launched	Queued	14 déc., 20:07	Queued...	<div>View</div>
<input type="checkbox"/>	<div>f91e7caa</div>	 etl_job	 Manually launched	Success	14 déc., 20:07	0:00:16	<div>View</div>
<input type="checkbox"/>	<div>66c2436c</div>	 etl_job	 Manually launched	Success	14 déc., 20:04	0:00:10	<div>View</div>
<input type="checkbox"/>	<div>eba47ed5</div> <div>View all tags (2)View queue criteria</div>	 etl_job	 Manually launched	Queued	14 déc., 19:57	Queued...	<div>View</div>
<input type="checkbox"/>	<div>57ed3da2</div> <div>View all tags (2)View queue criteria</div>	 etl_job	 Manually launched	Queued	14 déc., 19:48	Queued...	<div>View</div>

Jobs / etl_job Job in __repository__etl_job@etl_assets.pyLatest run: 15 déc., 22:27

OverviewLaunchpadRuns

Search and filter ops

Highlight...

load_data

Any

load_dataAny

transform_data

Any

transform_dataAny

save_data

Any

Search

Zoom

Help

InfoTypes

Job

etl_job

Description

No description provided

Resources

Metadata

Tags

5