

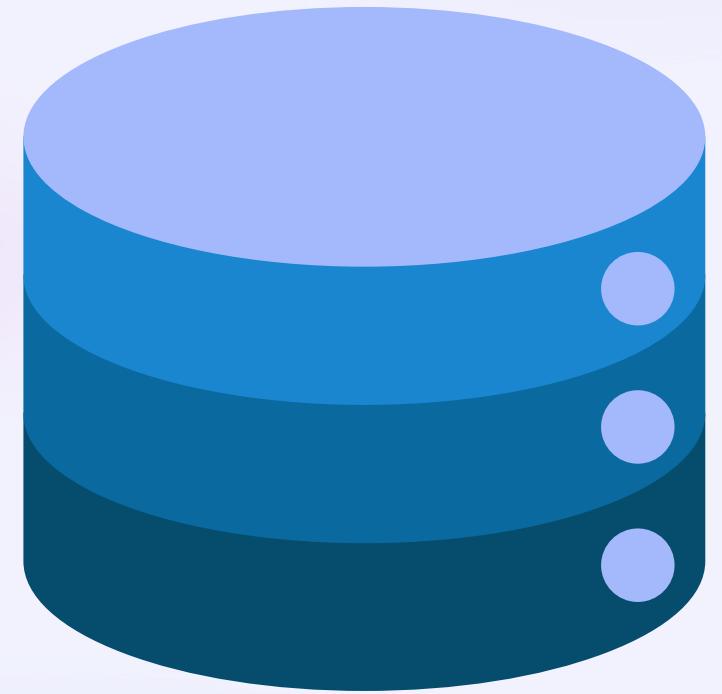
Mise en place d'une infrastructure (Données Météo)



Le Contexte du Projet

Objectif Principal

Migrer l'ensemble des données météorologiques provenant de multiples sources hétérogènes (fichiers CSV et flux Airbyte) vers une base de données MongoDB centralisée et performante.

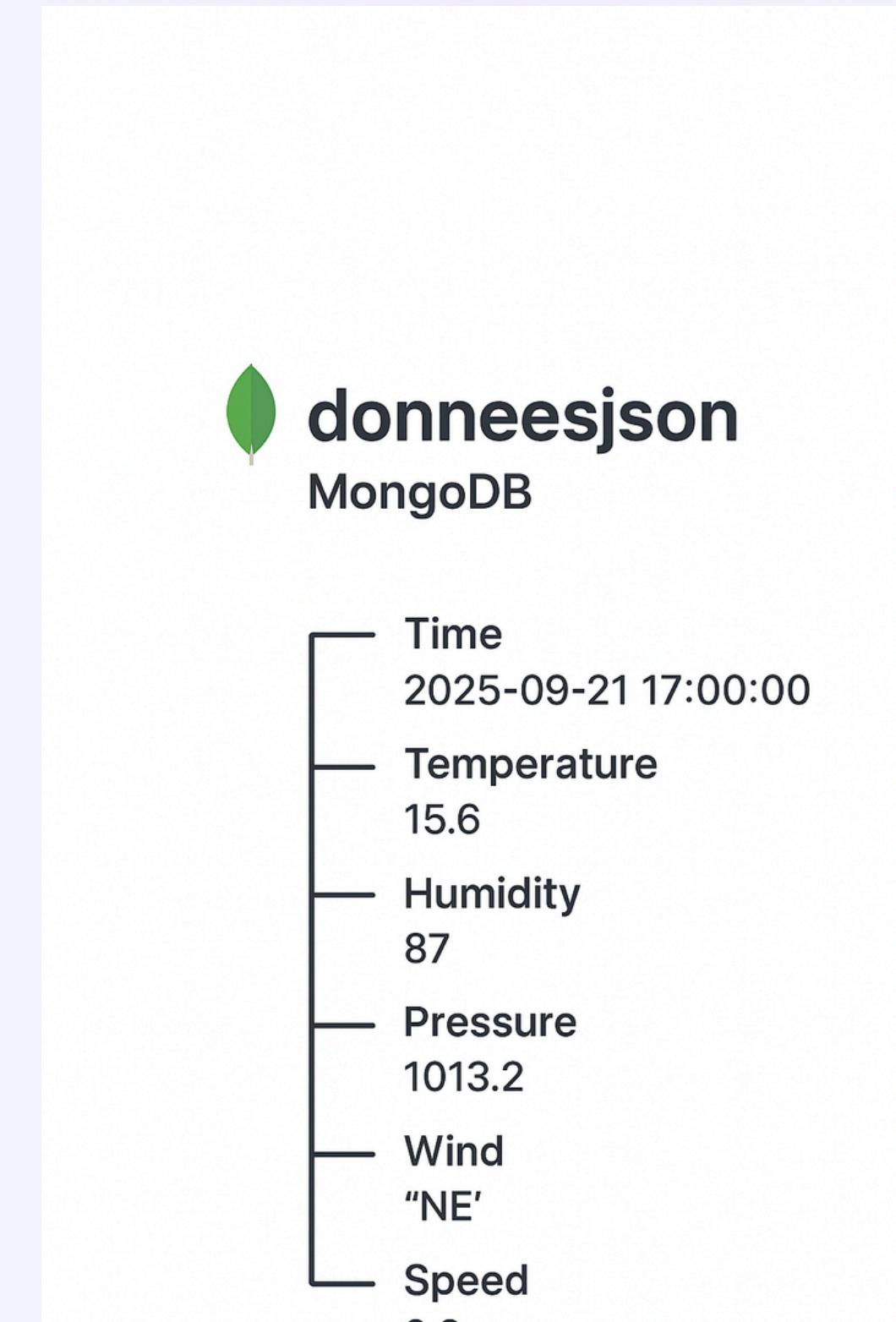
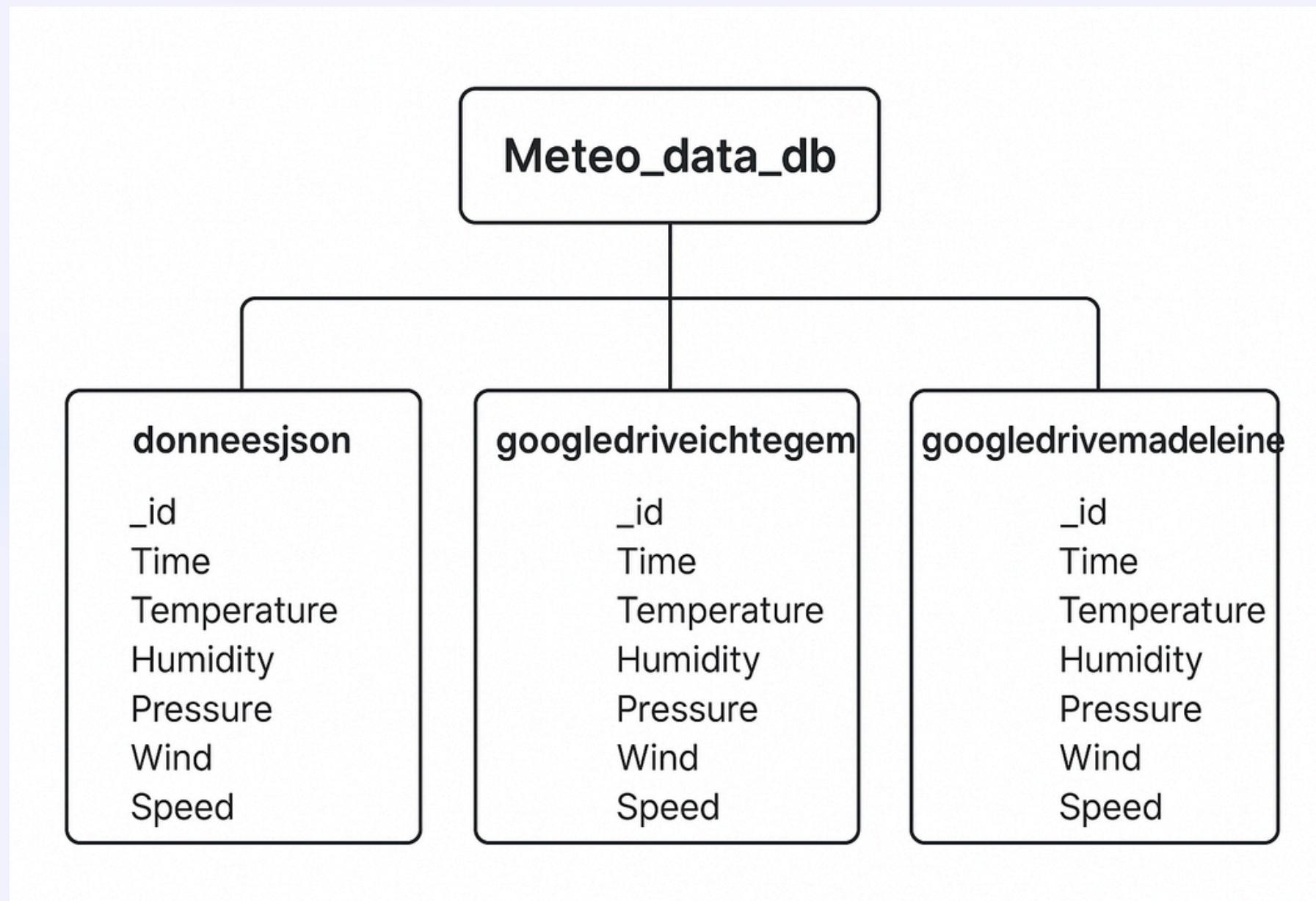


Enjeux Techniques

Mise en place pipeline ETL (Extract, Transform, Load) tout en garantissant la qualité, la cohérence et l'intégrité des données tout au long du processus de migration.



Données & Schémas



Stack Technologique

Catégorie	Technologie
Intégrateur	Airbyte
Base de données	MongoDB
Stockage Cloud	AWS S3
Conteneurisation	Docker + Docker Compose
Cloud Deployment	AWS ECS + ECR
Gestion des secrets	.env / variables d'environnement
Librairies Python	Pandas, pymongo, boto3, python-dotenv

Sources de Données Hétérogènes



Données JSON

Fichiers structurés provenant du répertoire Donnees_JSON avec format standard



Google Drive

Sources Ichtegem et Madeleine stockées sur la plateforme collaborative Google Drive



Format Airbyte

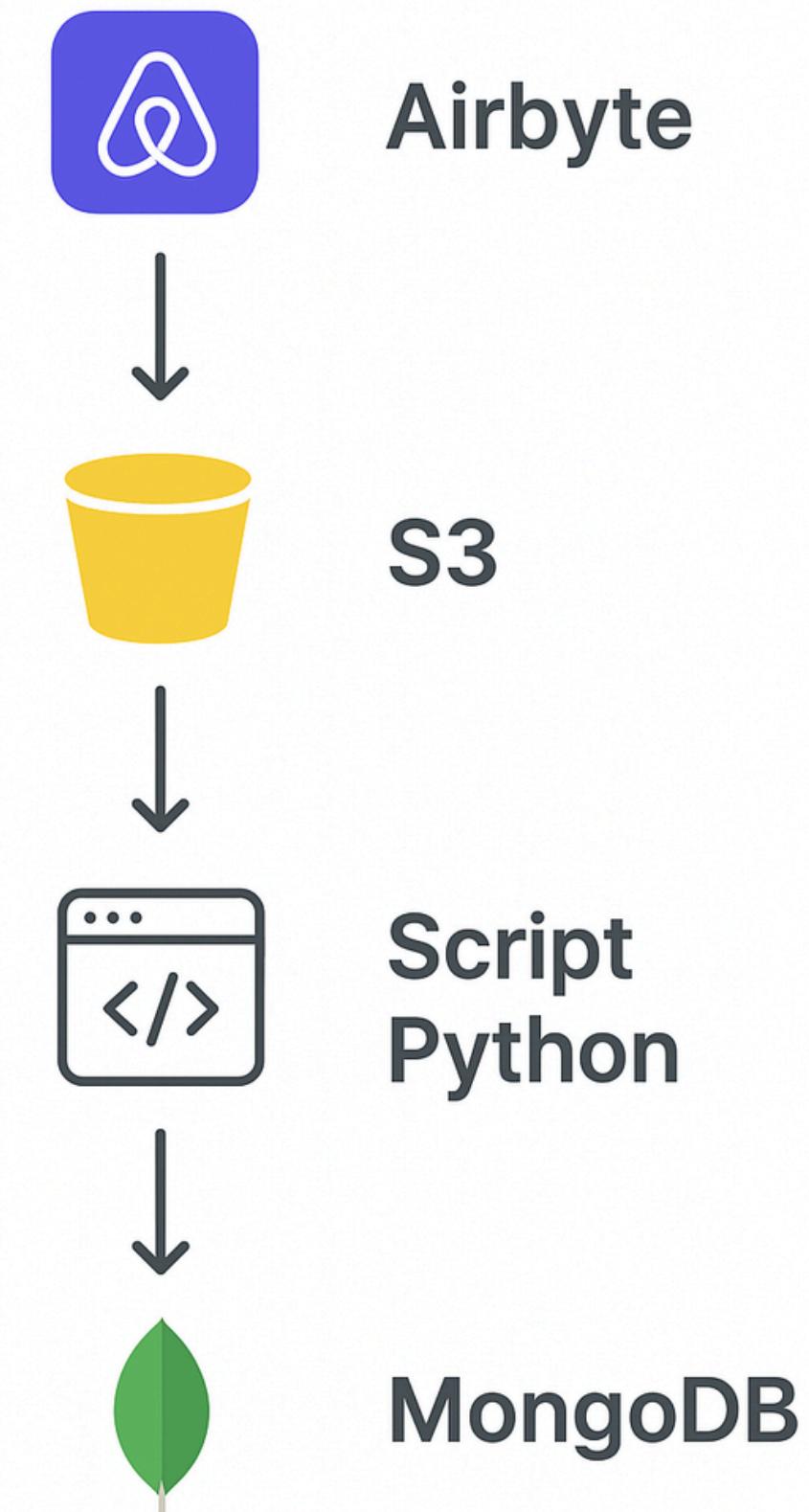
Encapsulé selon la structure sfile CSV du connecteur Airbyte

Le volume total représente plusieurs milliers de lignes par source, nécessitant un traitement automatisé et optimisé pour garantir des performances acceptables lors de la migration.

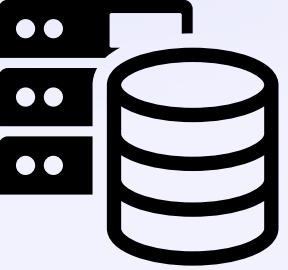
Démonstration intégration Airbyte

NAME	SOURCE NAME	DESTINATION NAME	FREQUENCY	TAGS	LAST SYNC	ENABLED
✓ File (CSV, JSON, Excel, Feather,...	File (CSV, JSON, Excel, Feather,...	S3_destination	Manual		il y a 23 jours	<input checked="" type="checkbox"/>
✓ File (CSV, JSON, Excel, Feather,...	File (CSV, JSON, Excel, Feather,...	S3_destination	Manual		il y a 23 jours	<input checked="" type="checkbox"/>
✓ File (CSV, JSON, Excel, Feather,...	File (CSV, JSON, Excel, Feather,...	S3_destination	Manual		il y a 26 jours	<input checked="" type="checkbox"/>

Schémas de la base de données



Architecture Logicielle



1- Lecture depuis S3

2- Normalisation

3-Nettoyage &
Insertion MongoDB

4- Conteneurisation

5- Déploiement AWS



Workflow ETL Détailé

Configuration

Chargement des variables d'environnement depuis le fichier .env sécurisé

Extraction S3

Connexion au bucket S3 et récupération automatique de tous les fichiers CSV sources

Parsing JSON

Lecture des fichiers CSV et parsing JSON pour les formats Airbyte encapsulés

Nettoyage

Suppression des lignes incomplètes et normalisation des types de données

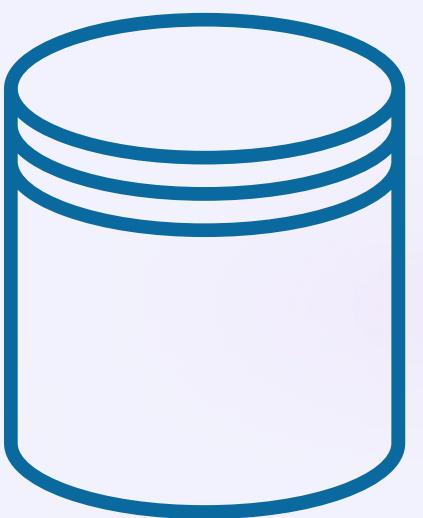
Insertion

Connexion MongoDB et insertion des données dans les collections appropriées

Validation

Contrôle qualité automatique : volume, colonnes, taux de valeurs manquantes

Résultats : validation des données



```
== Traitemet Data_JSON/Donnees_JSON/2025_09_21_1758483200407_0.csv -> donneesjson ==
0 lignes supprimées pour valeurs manquantes.
Connexion à MongoDB réussie.
1 documents insérés dans 'donneesjson'.

== Traitemet Data_JSON/Google_drive_Ichtegem/2025_09_23_1758670114136_0.csv -> googledriveichtegem ==
14 lignes supprimées pour valeurs manquantes.
Connexion à MongoDB réussie.
1892 documents insérés dans 'googledriveichtegem'.

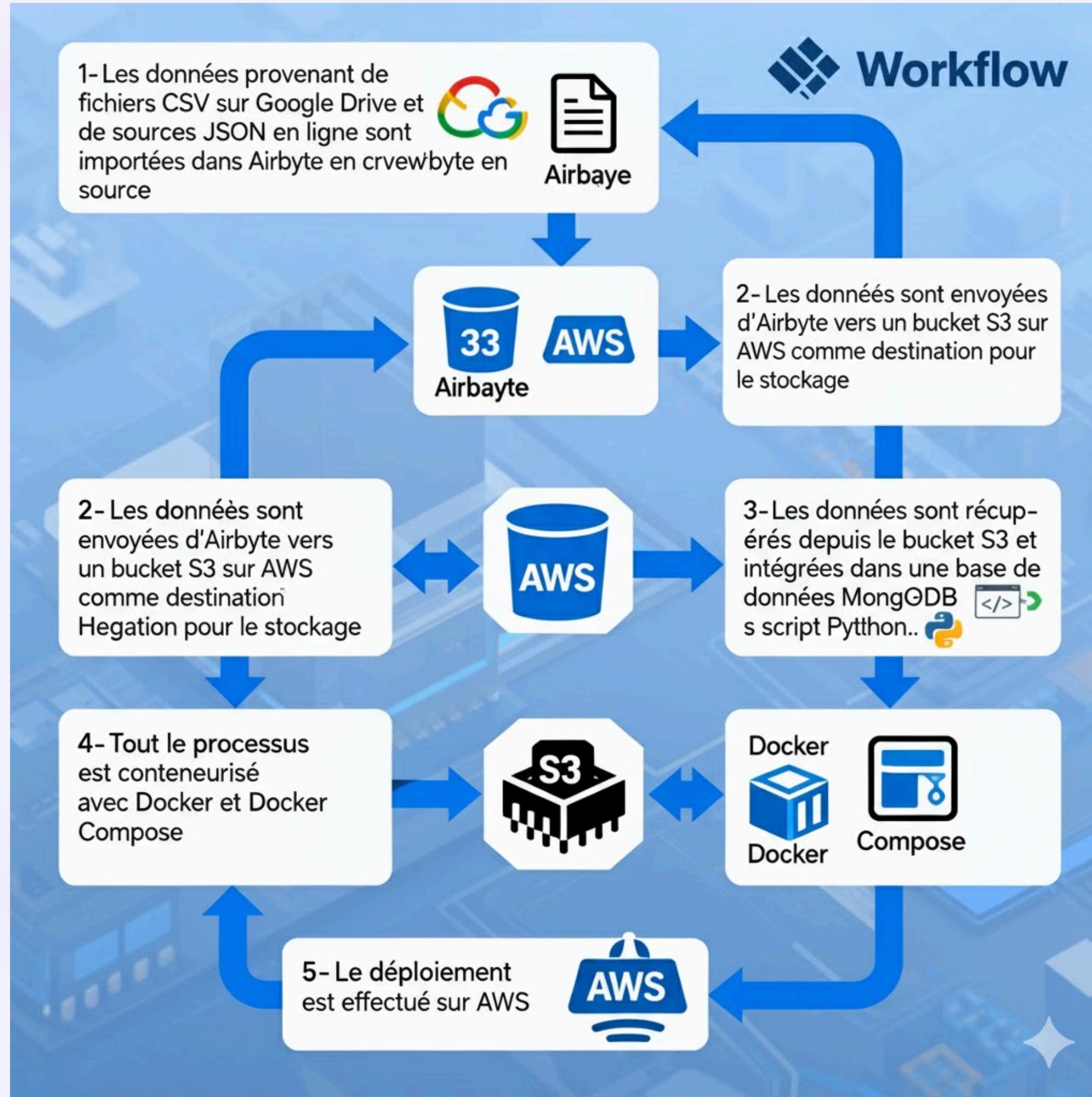
== Traitemet Data_JSON/Google_drive_Madeleine/2025_09_23_1758668827948_0.csv -> googledrivemadeleine
65 lignes supprimées pour valeurs manquantes.
Connexion à MongoDB réussie.
1850 documents insérés dans 'googledrivemadeleine'.
Connexion à MongoDB réussie.
```

```
== Contrôle qualité donneesjson ==
Source : 1 lignes | MongoDB : 1 documents
Colonnes identiques : True
Taux de valeurs manquantes (%) :
status 0.0
errors 0.0
data 0.0
stations 0.0
metadata.temperature 0.0
metadata.pression 0.0
metadata.humidite 0.0
metadata.point_de_rosee 0.0
metadata.visibilite 0.0
metadata.vent_moyen 0.0
metadata.vent_rafales 0.0
metadata.vent_direction 0.0
metadata.pluie_3h 0.0
metadata.pluie_1h 0.0
metadata.neige_au_sol 0.0
metadata.nebulosite 0.0
metadata.temps_omm 0.0
hourly.07015 0.0
hourly.00052 0.0
hourly.000RS 0.0
hourly.STATIC0010 0.0
hourly._params 0.0
dtype: float64
Connexion à MongoDB réussie.
```

```
== Contrôle qualité googledriveichtegem ==
Source : 1906 lignes | MongoDB : 1892 documents
Colonnes identiques : True
Taux de valeurs manquantes (%) :
Time 0.0
Pressure 0.0
Wind 0.0
UV 0.0
Solar 0.0
Dew Point 0.0
Precip. Rate. 0.0
Humidity 0.0
Gust 0.0
Temperature 0.0
Speed 0.0
Precip. Accum. 0.0
dtype: float64
Connexion à MongoDB réussie.
```

```
== Contrôle qualité googledrivemadeleine ==
Source : 1915 lignes | MongoDB : 1850 documents
Colonnes identiques : True
Taux de valeurs manquantes (%) :
Speed 0.0
Humidity 0.0
Gust 0.0
Time 0.0
Wind 0.0
Solar 0.0
Precip. Accum. 0.0
Precip. Rate. 0.0
Pressure 0.0
Dew Point 0.0
Temperature 0.0
UV 0.0
dtype: float64
```

Réprésentation du Workflow



Résultat d'importation

CONNECTIONS (1)

X + ...

Search connections

localhost:27017

- Base_mongo
- Listings
- Meteo_data_db
 - donneesjson
 - googledriveichtegem
 - googledrivemadeleine
- admin
- base_creer
- config
- local
- rainforest
- weather_data_db

+ trash

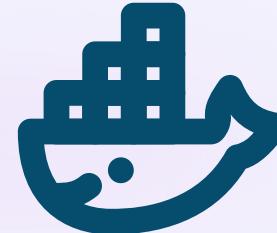
```
_id: ObjectId('68ef93a66238b484222aa884')
Time : "00:04:00"
Pressure : "29.48 in"
Wind : "WSW"
UV : 0
Solar : "0 w/m²"
Dew Point : "53.1 °F"
Precip. Rate. : "0.00 in"
Humidity : "87 %"
Gust : "10.4 mph"
Temperature : "56.8 °F"
Speed : "8.2 mph"
Precip. Accum. : "0.00 in"



---


_id: ObjectId('68ef93a66238b484222aa885')
Time : "00:09:00"
Pressure : "29.47 in"
Wind : "WSW"
UV : 0
Solar : "0 w/m²"
Dew Point : "52.9 °F"
Precip. Rate. : "0.00 in"
Humidity : "87 %"
Gust : "9.8 mph"
Temperature : "56.8 °F"
Speed : "7.9 mph"
Precip. Accum. : "0.00 in"
```

Conteneurisation avec Docker



mongodb_server

Instance MongoDB conteneurisée avec volume persistant pour garantir la durabilité des données

Architecture Docker Compose

Le fichier Docker Compose orchestre deux services essentiels pour l'exécution du pipeline :

migration_container

Conteneur Python exécutant le script Migration_code.py avec toutes ses dépendances

Name	Tag	Image ID	Created	Size	Actions
056743698956.dkr.ecr.eu-west-3.ama	latest	c82b1ed0c54d	2 days ago	924.6 MB	▶ ⋮
projet_donnees-migration_app	latest	c82b1ed0c54d	2 days ago	924.6 MB	▶ ⋮
mongo	7	c258b26dbb77	10 days ago	1.12 GB	▶ ⋮
airbyte/cron	1.8.2	67e6bbd32653	1 month ago	1.33 GB	▶ ⋮
airbyte/server	1.8.2	f8f8da6c2996	1 month ago	1.37 GB	▶ ⋮
airbyte/workload-launcher	1.8.2	e0f239d14440	1 month ago	1.32 GB	▶ ⋮
airbyte/worker	1.8.2	6247c2456603	1 month ago	1.33 GB	▶ ⋮
airbyte/workload-api-server	1.8.2	1402a8e9209d	1 month ago	1.34 GB	▶ ⋮



Isolation

Séparation complète des services pour une meilleure sécurité



Reproductibilité

Environnement identique sur toutes les plateformes

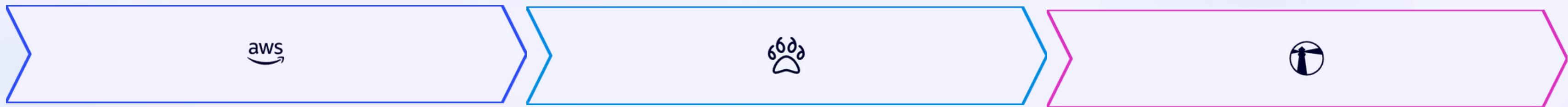


Déploiement

Transition simplifiée vers l'infrastructure cloud



Déploiement



ECR

Création du registre de conteneurs et push de l'image Docker de l'application

ECS

Déploiement d'une tâche ECS utilisant l'image Docker pour exécuter le pipeline ETL

CloudWatch

Gestion centralisée des logs et monitoring en temps réel des tâches ECS

Résultat : Un pipeline ETL entièrement automatisé et exécuté dans le cloud avec MongoDB comme base centrale

Stockage S3

Amazon S3 > Compartiments > bucket-projet-airbyte > Data_JSON/

azon S3

partiments à usage général

partiments de répertoires

partiments de table

partiments de vecteur

ss Grants

ts d'accès (compartiments à
e général, systèmes de
ers FSx)

ts d'accès (compartiments de
rtoires)

ts d'accès de l'objet Lambda

ts d'accès multi-région

rations par lot

Access Analyzer pour S3

Data_JSON/

Objets

Propriétés

Objets (3)



Copier l'URI S3



Copier l'URL

Les objets sont les entités fondamentales stockées dans Ar
Pour que d'autres personnes puissent accéder à vos objets,



Rechercher des objets en fonction du préfixe



Nom



Type



Donnees_JSON/

Dossier



Google_drive_Ichtegem/

Dossier



Google_drive_Madeleine/

Dossier

Amazon ECR

Amazon ECR > Registre privé > Référentiels > projet_donnees-migration_app

Amazon Elastic Container Registry

Private registry

- Repositories
- Summary
- Images
- Permissions
- Lifecycle Policy
- Repository tags
- Features & Settings

Public registry

Images (3)

Rechercher des artefacts

<input type="checkbox"/>	Balise d'image	Type d'artefact	Transmis à	Taille (Mo)
<input type="checkbox"/>	latest	Image Index	15 octobre 2025, 20:08:57 (UTC+02)	224.27
<input type="checkbox"/>	-	Image	15 octobre 2025, 20:08:57 (UTC+02)	0.00
<input type="checkbox"/>	-	Image	15 octobre 2025, 20:08:57 (UTC+02)	224.27

Amazon ECS : le conteneur

Amazon Elastic Container Service > Clusters > migration-cluster > Services > migration-task-service-qs338xdq > Tâches > 71e72f784ba2447a9d92089e8c9052db > Configuration

Amazon Elastic Container Service <

Clusters

Espaces de noms

Définitions de tâches

Paramètres du compte

Amazon ECR []

Référentiels []

AWS Batch []

Documentation []

Découvrir les produits []

Abonnements []

Laissez-nous un commentaire.

Injection de pannes
Désactivé

ECS Exec | Infos
Désactivé

Protection adaptée à l'évolution des tâches

Statut de protection
Désactivé

Groupe de tâches
service:migration-task-service-qs338xdq | Afficher le service

migration-task:3

Conteneurs (1)

Filtrer les conteneurs

Nom du conteneur	ID d'exécution du c...	URI de l'ima...	Résumé des...	Statut	Statut de l'état
mongodb_server	a630e92c7ad9f45...	05674369...	sha256:c8...	Running	Inconnu

Détails du conteneur pour mongodb_server

Détails | Configuration de journaux | Politique de redémarrage | Liaisons réseau | Étiquettes et hôtes Docker | Variables

Détails

URI de l'image <input checked="" type="checkbox"/> 056743698956.dkr.ecr.eu-west-3.amazonaws.com/projet_donnees-migration_appltest	Essentiel Oui
--	------------------

Amazon ECS : les tâches

Amazon Elastic Container Service > Clusters > migration-cluster > Services > migration-task-service-qs338xdq > Tâches

migration-task-service-qs338xdq Infos

Dernière mise à jour à
19 octobre 2025, 15:07 (UTC+2:00)  

Aperçu du service Infos

Statut  Actif

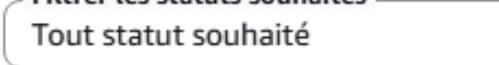
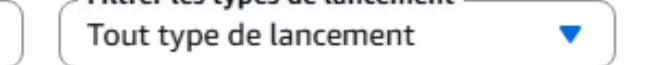
Tâches (1 souhaitées) 

Définition de la tâche : révision [migration-task:3](#)

Statut du déploiement  Réussite

État et métriques       

Tâches (1/2)

Filtrer les tâches par propriété ou par valeur  Filtrer les statuts souhaités  Filtrer les types de lancement 

Tâche	Dernier st...	Statut sou...	Défi...	Statut de l...	Créé à	Démarré par	Démarré à	Ins...
 	 En cours d'exécu	 En cours d'exécu	 migrati...	 Inconnu	il y a 38 minu...	ecs-svc/64154150805...	il y a 37 minu...	
 	 Arrêté Es...	 Arrêté	 migrati...	 Inconnu	il y a 2 heures	ecs-svc/64154150805...	il y a 2 heures	

Conteneurs pour la tâche 71e72f784ba2447a9d92089e8c9052db

Conteneurs (1)

Filtrer les conteneurs 

Contenu d'flux CloudWatch

CloudWatch > Groupes de journaux > /ecs/migration-task

CloudWatch <

Favoris et récents >

Tableaux de bord

Alarmes ⚠️ 0 🗓️ 2 ⏱️ 0

Journaux

Groupes de journaux

Anomalies dans les journaux

Queue en direct

Logs Insights

Contribuer à cette page

< Flux de journaux Balises Détection des anomalies Filtres de métriques Filtres d'abor

Flux de journaux (42)

Filtrer les flux de journaux ou essayer la recherche de préfixe

Correspondance exacte Affiche

Flux de journaux ▾ | Heure du dernier événement

ecs/mongodb_server/5d20141435d64595a93105a1ebf30736	2025-10-17 08:35:32 (UTC)
ecs/mongodb_server/8f49da1617c84edbbe31099534e98786	2025-10-17 07:34:05 (UTC)
ecs/mongodb_server/a3d7312e0fb64c489e68e1416f2ea7ae	2025-10-17 06:32:42 (UTC)

Flux de Jounaux : CloudWatch

CloudWatch > Groupes de journaux > /ecs/migration-task > ecs/mongodb_server/35356fcf7e444788911438f33fda19f2

CloudWatch

Favoris et récents

Tableaux de bord

▶ Alarms ⚠ 0 ✓ 2 ⏰ 0

▼ Journaux

- [Groupes de journaux](#)
- Anomalies dans les journaux
- Queue en direct
- Logs Insights
- Contributor Insights

▶ Métriques [Nouveau](#)

▶ Signaux d'application [Nouveau](#)

Événements de journaux

Vous pouvez utiliser la barre de filtre ci-dessous pour rechercher et faire correspondre des termes, des expressions ou des valeurs dans vos évènements de journal. [En savoir plus](#)

Filtrer les événements : appuyez sur Entrée pour rechercher.

Effacer 1m 30m 1h 1d

Horodatage	Message
Aucun ancien événement pour le moment Réessayer	
▼ 2025-10-19T12:26:16.217Z	Chargement MONGO_URI : mongodb://forecast_writer:Linse3128@mongodb:27017/Meteo_data_db?authSource=Meteo_data_db
	Chargement MONGO_URI : mongodb://forecast_writer:Linse3128@mongodb:27017/Meteo_data_db?authSource=Meteo_data_db
▼ 2025-10-19T12:26:16.217Z	Chargement AWS_S3_BUCKET : bucket-projet-airbyte
	Chargement AWS_S3_BUCKET : bucket-projet-airbyte
▼ 2025-10-19T12:26:16.217Z	==> Traitement Data_JSON/Donnees_JSON/2025_09_21_1758483200407_0.csv -> donneesjson ==>
	==> Traitement Data_JSON/Donnees_JSON/2025_09_21_1758483200407_0.csv -> donneesjson ==>

Résultats et Perspectives

Réalisations

- Base **Meteo_data_db** créée avec 3 collections correspondant aux sources CSV
- Données chargées et validées avec contrôle qualité automatique
- Script et conteneur Docker prêts et déployer sur AWS

Prochaines Étapes

- Automatisation complète sur ECS avec planification cron pour les mises à jour régulières
- Intégration d'un dashboard interactif pour visualiser et explorer les données migrées

