Sébastien Branly Guillaume Maugin Louis Réville

M1 DEV MANAGER FULLSTACK

Big Data

"Intégration et Analyse de Données Environnementales :

Une Approche Multi-Dataset pour Comprendre les Défis Globaux"

Présentation de l'étude sur les données environnementales

Contexte

Aujourd'hui, les enjeux environnementaux occupent une place centrale dans les débats sociétaux et politiques. L'impact des activités humaines sur l'environnement, notamment en matière d'émissions de CO_2, de gestion des déchets plastiques et de déforestation, soulève des préoccupations majeures pour la durabilité de notre planète.

Dans ce contexte mondialement préoccupant, analyser et comprendre ces phénomènes à l'aide de données fiables devient indispensable. Ces analyses permettent de mieux cerner les tendances actuelles, de mesurer les impacts environnementaux, et de soutenir la prise de décisions éclairées pour des politiques plus durables.

À travers cette étude, nous avons mobilisé plusieurs ensembles de données riches et variés:

- Les émissions de CO_2 par pays, permettant d'évaluer l'évolution des gaz à effet de serre à l'échelle globale et nationale.
- Les projections et analyses des émissions de CO_2 pour identifier les scénarios futurs possibles et anticiper les impacts environnementaux.
- Les informations sur la surface forestière mondiale, un indicateur clé pour évaluer les pertes en biodiversité et les changements climatiques.

Cette combinaison d'ensembles de données offre une vue d'ensemble des défis environnementaux majeurs et constitue un outil précieux pour explorer des solutions potentielles à ces problématiques complexes.

Objectif

L'objectif principal de cette étude est de mieux comprendre les impacts environnementaux en analysant des données sur les émissions de CO_2, la gestion des déchets plastiques et la déforestation. Cela vise à soutenir les initiatives durables et à orienter les décisions politiques et économiques vers des solutions plus respectueuses de l'environnement.

Pour atteindre cet objectif, plusieurs axes d'analyse sont envisagés. Il s'agit tout d'abord de suivre et comparer les émissions de CO_2 par pays afin d'identifier les zones géographiques où des efforts de réduction sont nécessaires. Ensuite, l'évaluation de la gestion des déchets plastiques permettra de mieux comprendre leur ampleur et de proposer des stratégies adaptées aux spécificités locales. L'analyse de l'évolution des surfaces forestières mondiales constitue également un volet clé pour mettre en lumière les régions les plus touchées par la déforestation et ses impacts sur la biodiversité.

Enfin, cette étude proposera des indicateurs de performance environnementale pour mesurer l'efficacité des politiques actuelles. Elle inclura aussi des prédictions basées sur les

tendances observées afin d'anticiper les futurs scénarios et de permettre une planification proactive. L'ensemble de ces objectifs repose sur des données fiables collectées à partir de diverses sources, garantissant une vision globale et détaillée des enjeux environnementaux actuels.

Datasets

- World Forest Area (km and %)
- Carbon (CO2) Emissions
- Country-Level CO_2 Emissions Analysis and Prediction | Kaggle

Scripts utilisés

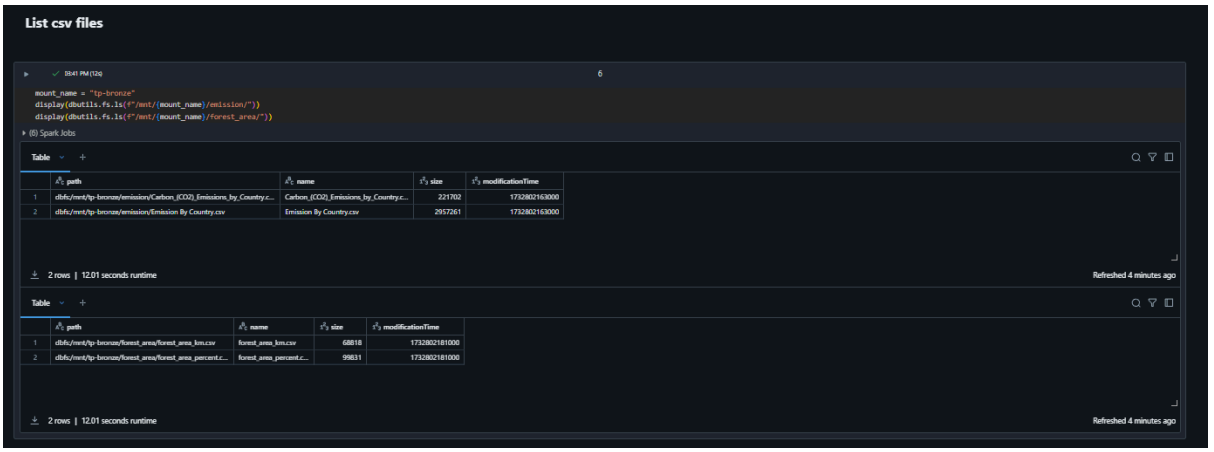
L'ensemble des scripts utilisés sont présents sur ce Repository Github.

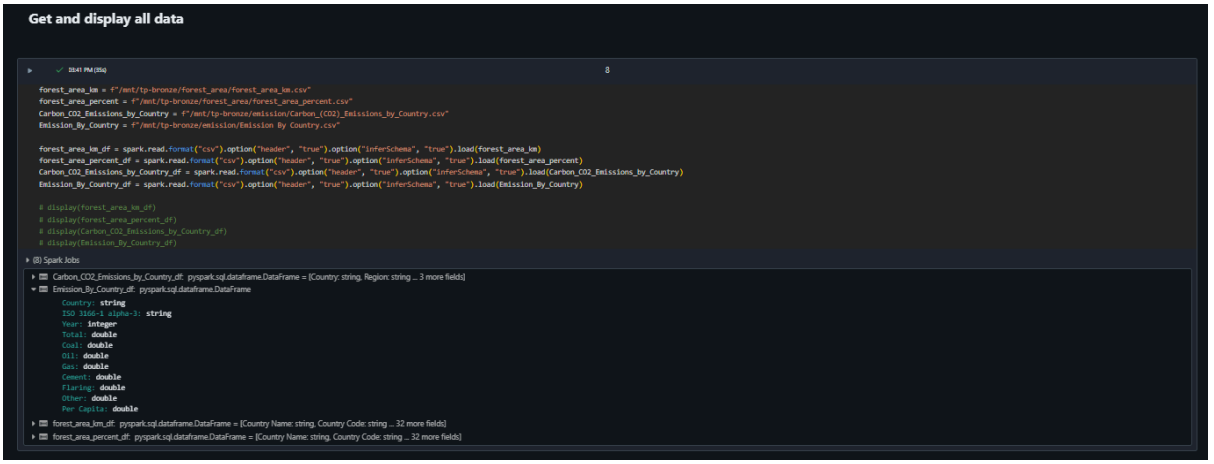
Traitement des données

Ce code établit un workflow avec Databricks et Azure Blob Storage pour traiter des données sur les forêts et les émissions de CO2 dans le monde de 1970 à 2021.

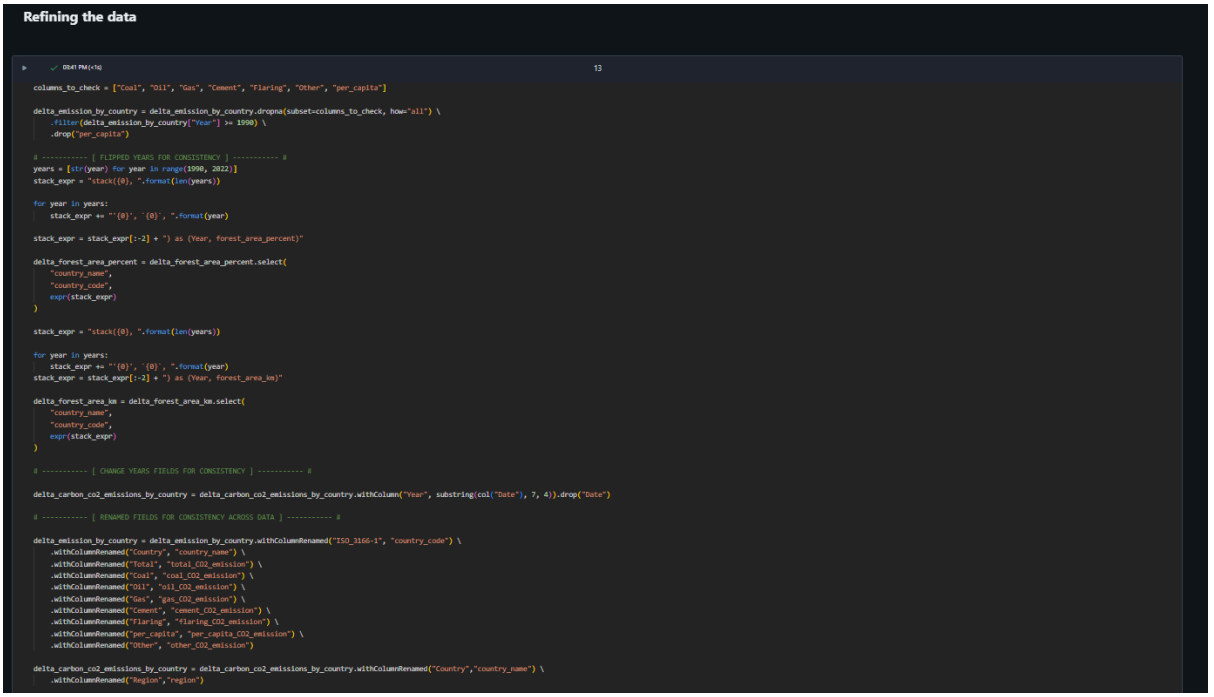
1. Montage des conteneurs Azure : Les conteneurs **tp-bronze**, **tp-silver**, et **tp-gold** sont montés dans Databricks, permettant d'accéder aux fichiers CSV.

2. Chargement des données brutes : Les fichiers CSV du conteneur **tp-bronze** (contenant des données sur les forêts et les émissions) sont lus en Dataframes.

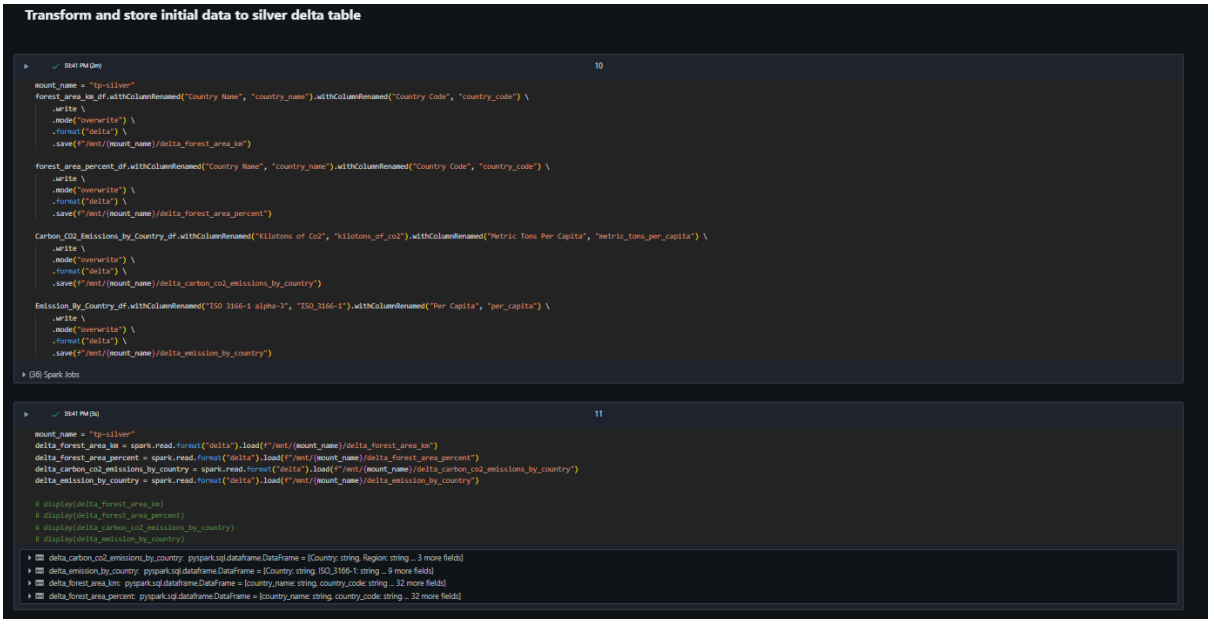




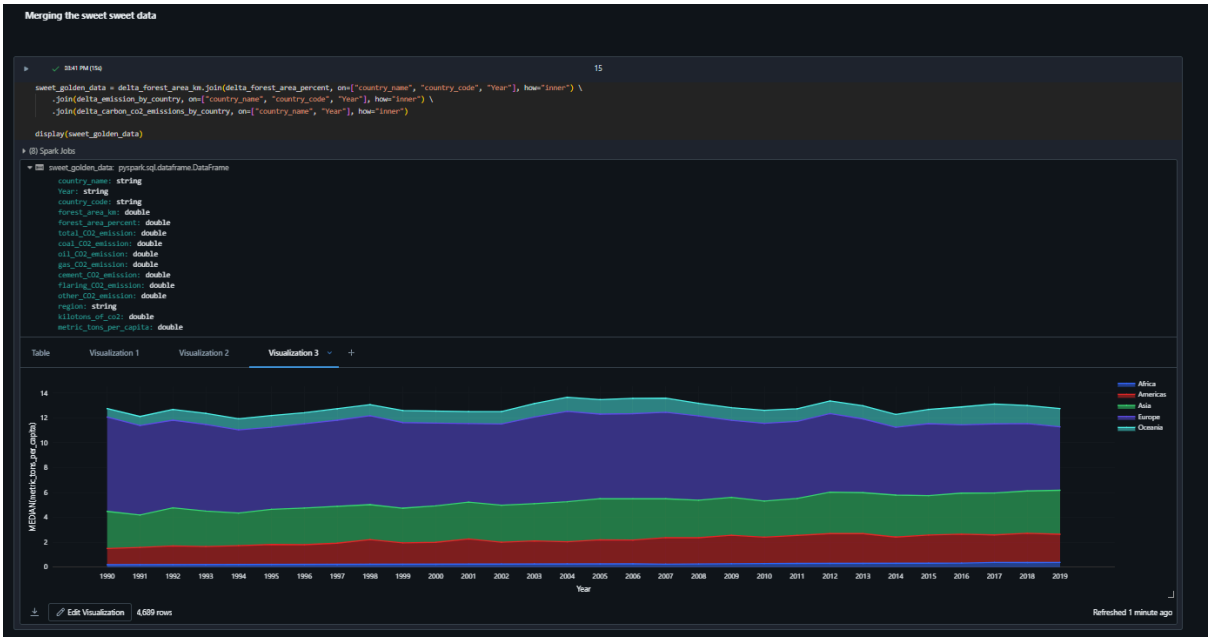
3. Transformation des données :
    - Les DataFrames sont nettoyés, normalisés (renommage des colonnes, suppression des colonnes inutiles, formatage des années), et les données manquantes sont filtrées.
    - Les colonnes de type "année" sont pivotées (stack) pour structurer les données de manière cohérente.
    - Les jeux de données sont harmonisés avec des noms de colonnes standardisés.



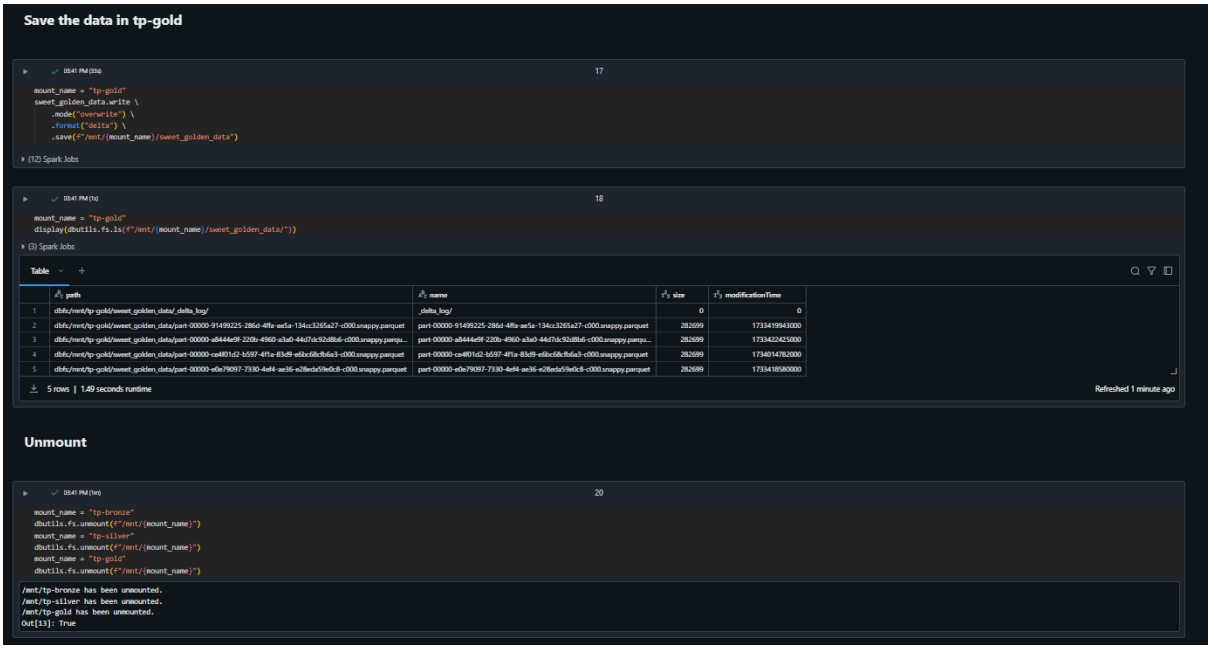
4. Stockage des données transformées : Les données sont écrites en format Delta dans le conteneur **tp-silver**, assurant un format optimisé pour des requêtes analytiques.



5. Fusion des données nettoyées : Un Dataframe unifié, **sweet_golden_data**, est créé en joignant les différents jeux de données harmonisés.

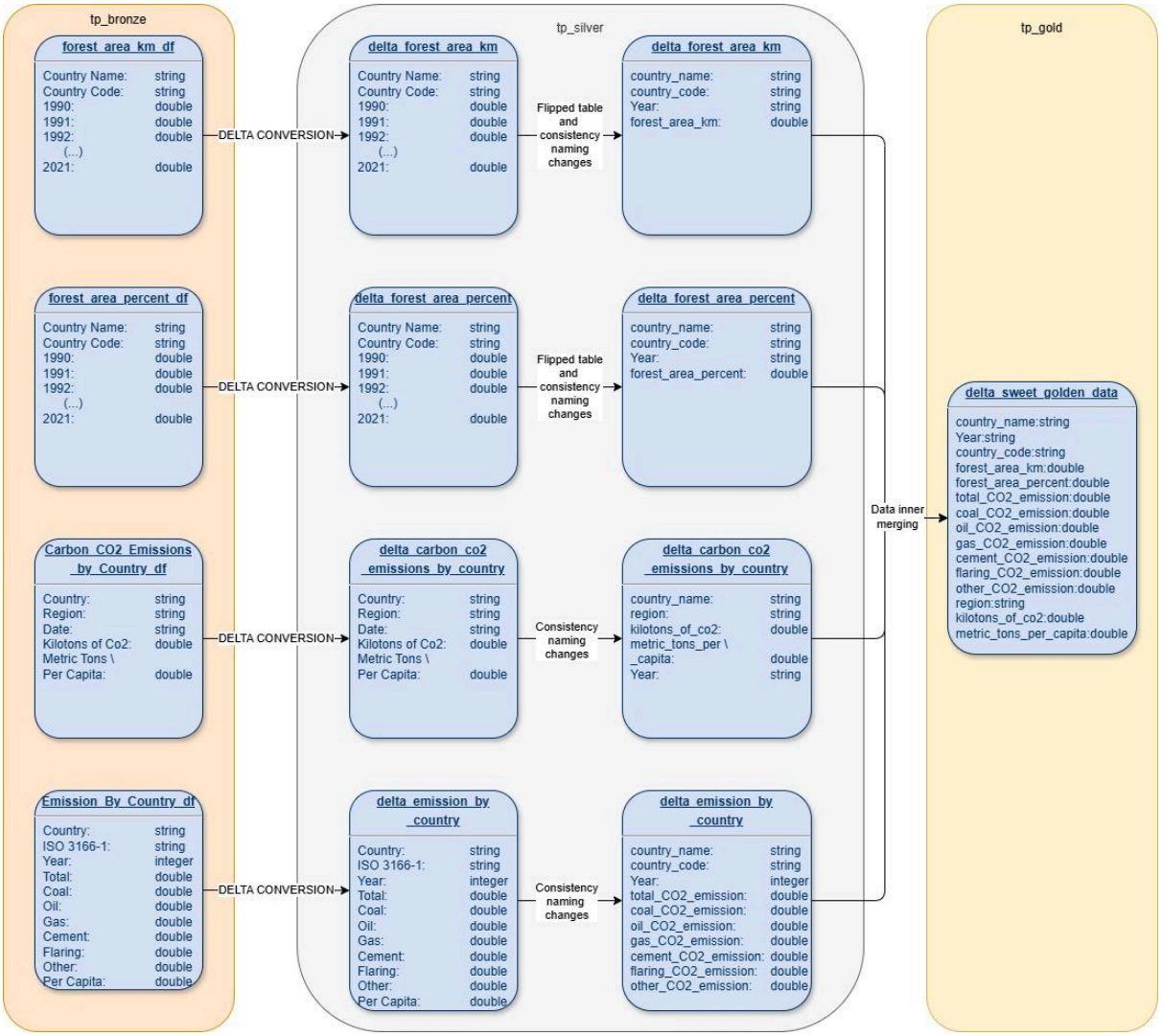


6. Stockage final : Le Dataframe fusionné est sauvegardé dans le conteneur **tp-gold** au format Delta.

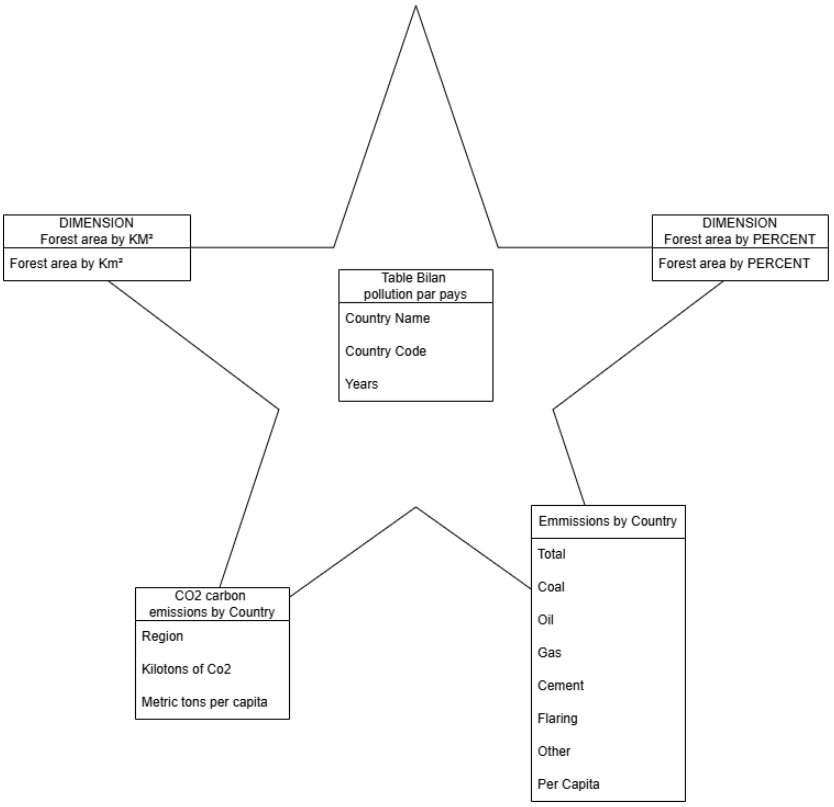


7. Démontage des conteneurs : Une fois le traitement terminé, les montages de **tp-bronze**, **tp-silver**, et **tp-gold** sont démontés pour libérer les ressources.

Modèle Conceptuels des Données



Modèle en étoile



Rapport Databricks

screens tableau de bord