YOUCODE

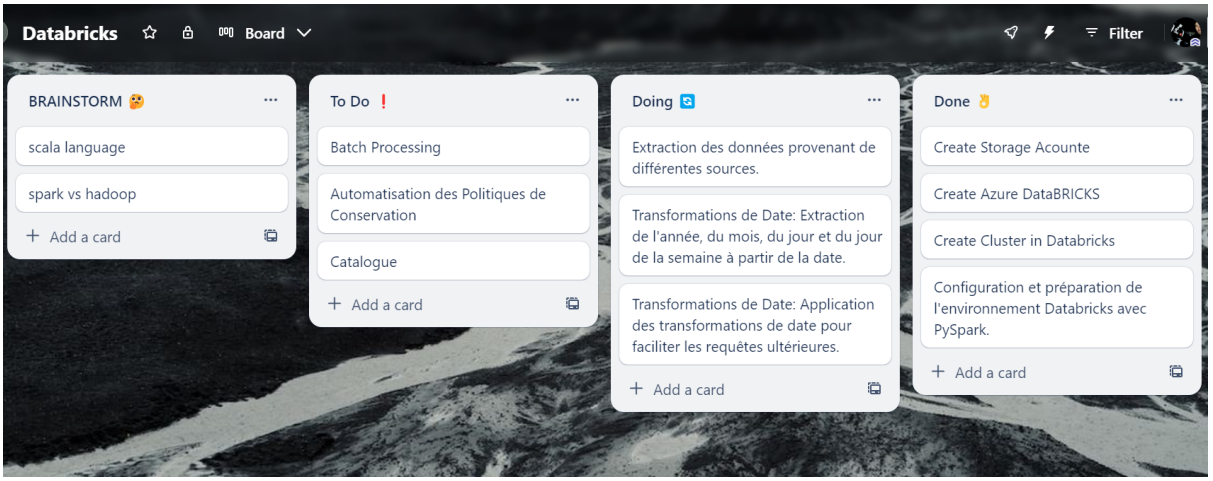
Intégration et de Gestion des Données de Transport Public avec Azure

par Yassine HARRATI
le 27-09-2023

Introduction:

Ce catalogue de données accompagne le projet de gestion des données de transport public, visant à collecter, transformer et gérer les données en utilisant les services Azure tels que Azure Data Lake Storage Gen2 et Azure Databricks. Ce projet a pour objectif d'améliorer les services de transport en tirant parti des données disponibles.

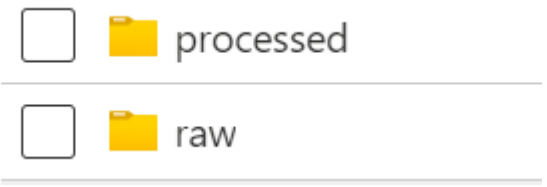
Planification de projet:



Étapes du Projet:

1. Conception de l'Architecture Data Lake:

- Création de l'espace de noms hiérarchique dans Azure Data Lake Storage Gen2.
  - `/public_transport_data/raw/` : pour les données CSV brutes.
  - `/public_transport_data/processed/` : pour les données traitées.



2. Intégration de l'Infrastructure Data Lake:

- Utilisation d'Azure Databricks avec PySpark pour lire et structurer les données.

3. Processus ETL avec Azure Databricks:

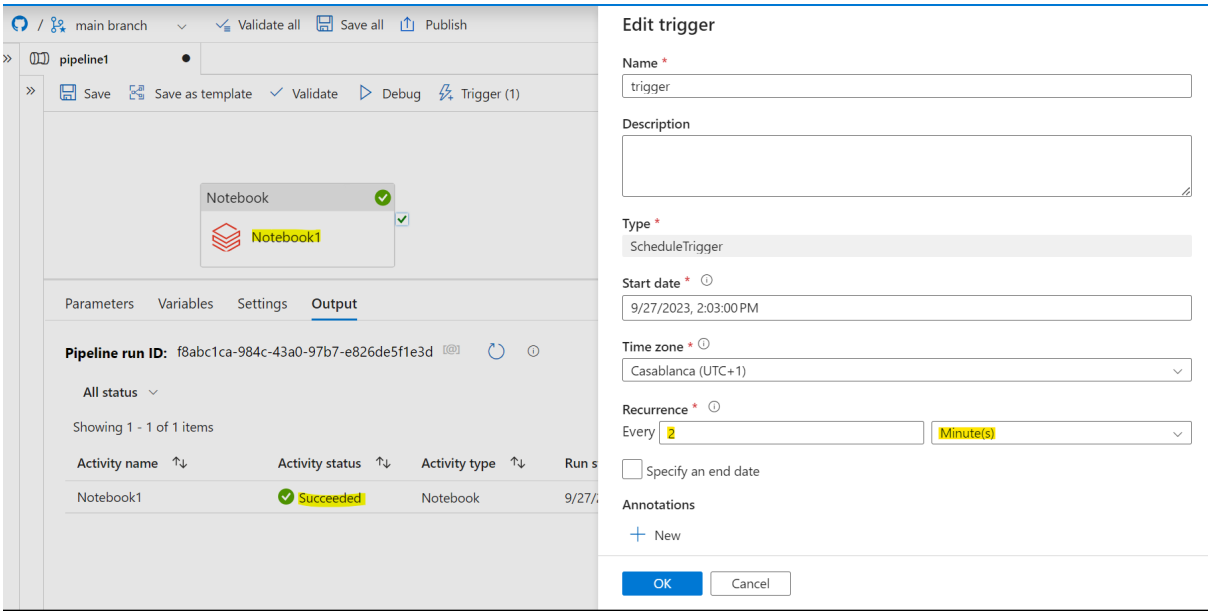
- Utilisation de PySpark pour transformer les données, incluant :
  - Transformations de Date.
  - Calculs Temporels.
  - Analyse des Retards.
  - Analyse des Passagers.
  - Analyse des Itinéraires.

4. Documentation des Transformations et des Données:
- Description des Données : Brève description pour chaque ensemble de données.
- Transformations : Documentation des transformations appliquées aux données brutes.
- Lignage des Données : Indication de la source des données et comment elles ont été traitées.
- Directives d'Utilisation : Informations sur la manière d'utiliser les données et les cas d'utilisation potentiels.

5. Automatisation des Politiques de Conservation:
- Planification d'un cahier pour appliquer les politiques de conservation des données.
  - Archivage des Données.
  - Suppression des Données.

6. Génération de Données à Intervales de Lots (Batch Intervals):
- Création de nouvelles données à intervalles de lots avec des fichiers CSV.

7. Batch Processing
- Traitement automatisé des données collectées dans la journée.



Conclusion

Ce catalogue de données reflète les différentes étapes du projet de gestion des données de transport public. Il décrit l'architecture, les transformations, la documentation, la conservation des données et le traitement automatisé. L'objectif est d'assurer une gestion efficace des données pour améliorer les services de transport.