YouCode / Youssoufia

Développeur Data

2023-2024

Projet : **Public Transports** **(Azure Databricks)**

**YONLI Fidèle**

**Table des matières**

[**Introduction** 2](#_Toc146733041)

[**Données brutes** 2](#_Toc146733042)

[**Données Transformées** 2](#_Toc146733043)

[**Transformations effectuées** 3](#_Toc146733044)

[**Conclusion** 3](#_Toc146733045)

# **Introduction**

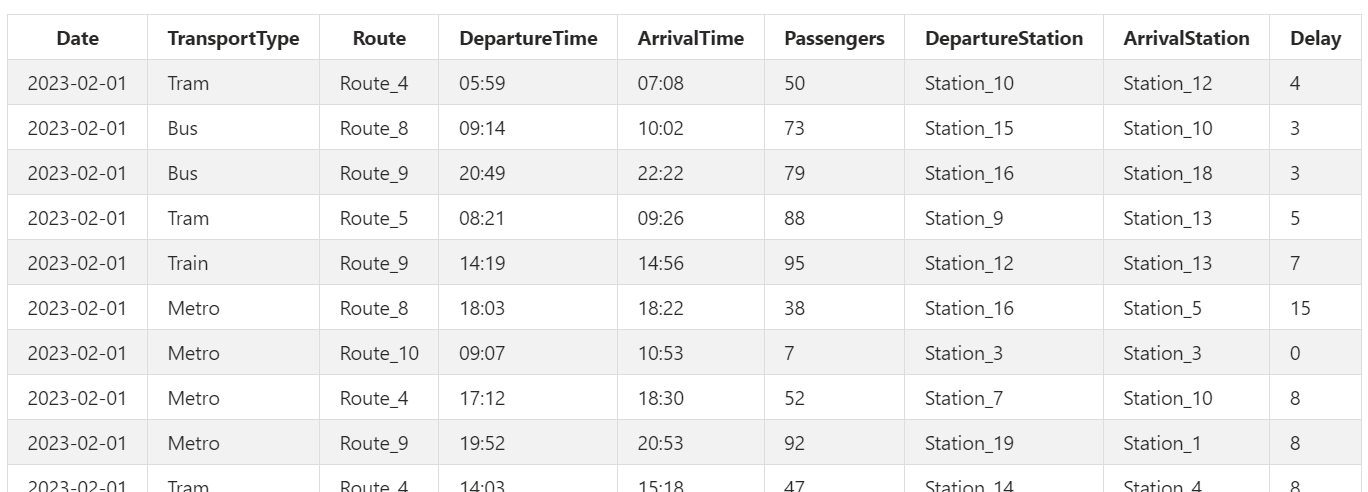
Ce document sert de catalogue de données Azure avec des ensembles de données enregistrées et documentées dans le projet d’intégration et de gestion des données des transports publics. La documentation est essentielle pour la gouvernance des données et pour que les utilisateurs comprennent les données.

Toutes les données sont contenues dans un conteneur (container) public-transport dans **Azure Storage** (compte de stockage). Ce conteneur a trois dossier :

* ***Raw***/
* ***Processed/***
* ***Archives/***

# **Données brutes**

Les données générées sont stockées dans le dossier **raw/)** qui constituent notre datalake. Nous avons 06 fichiers .csv de la même structure comme présenté dans la figure ci-dessous.



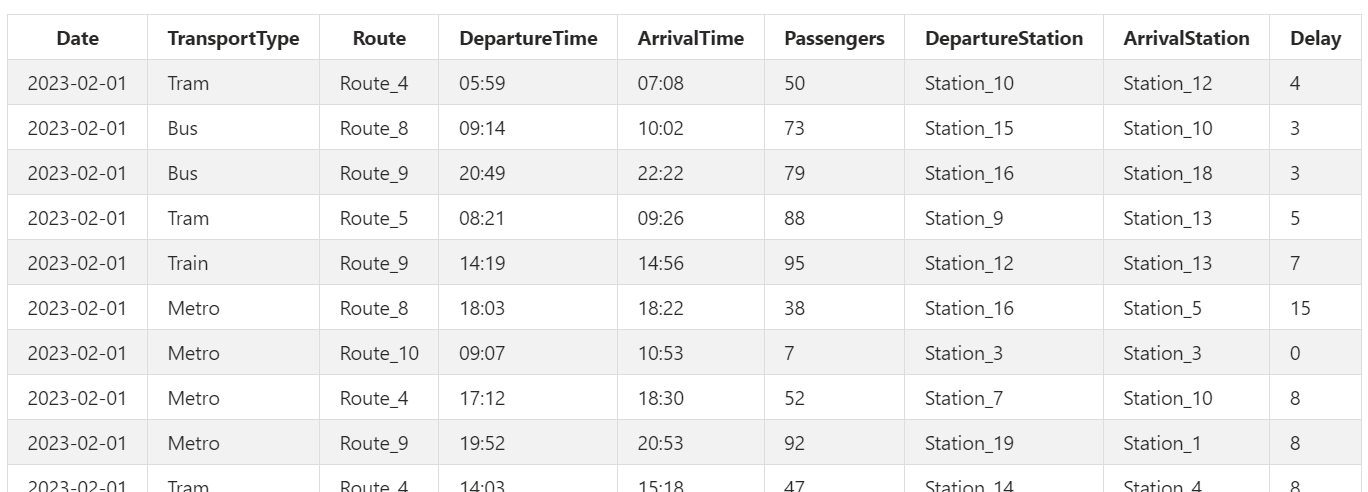
Chaque fichier correspond à un mois (seulement les 06 premiers mois de l’année 2023).

# **Données Transformées**

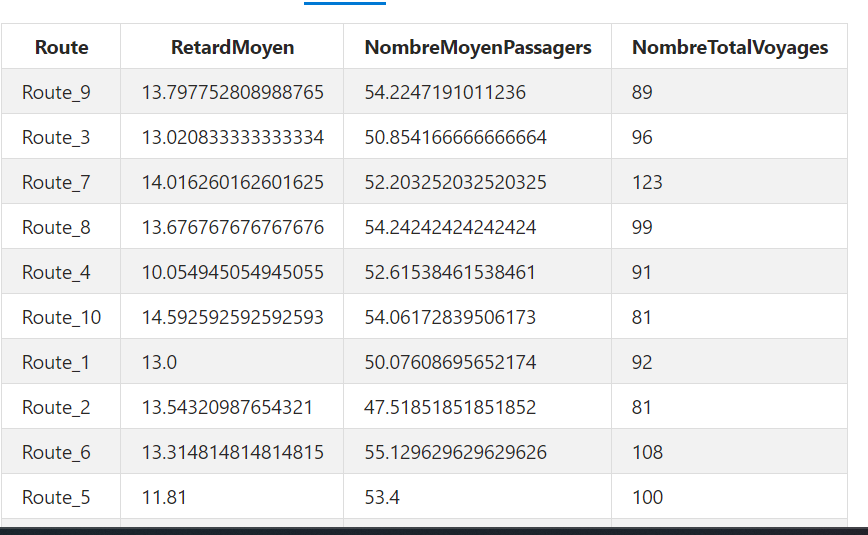
Le dossier ***processed/*** contient deux sous-dossiers :

* ***Data/*** qui contient les fichiers transformés
* ***Analyse/*** qui contient les résultats des analyses effectuées

Les fichiers contenus dans le dossier ***Data/*** ont la structure suivante :



Les fichiers dans ***analyse/*** eux la structure suivante :



Après avoir appliqué les transformations, les fichiers avec les données transformées sont stockés dans ces deux répertoires en fonction de leur nature.

# **Transformations effectuées**

En utilisant **PySpark** dans **Azure Databricks**, nous avons effectué quelques transformations.

* **Transformations de Date** : Extraire l'année, le mois, le jour et le jour de la semaine de la date pour faciliter les requêtes et les rapports.
* **Calculs Temporels** : Calculer la durée de chaque voyage en soustrayant l'heure de départ de l'heure d'arrivée.
* **Analyse des Retards** : Catégoriser les retards en groupes tels que 'Pas de Retard', 'Retard Court' (1-10 minutes), 'Retard Moyen' (11-20 minutes) et 'Long Retard' (>20 minutes).
* **Analyse des Passagers** : Identifier les heures de pointe et hors pointe en fonction du nombre de passagers.
* **Analyse des Itinéraires** : Calculer le retard moyen, le nombre moyen de passagers et le nombre total de voyages pour chaque itinéraire.

# **Conclusion**

Le contenu de ce document est à destination de toute personnes désirant comprendre ou utiliser nos données pour un usage quelconque ultérieur.