YouCode / Youssoufia

Développeur Data

2023-2024

Projet : Public Transports (Azure Databricks)

YONLI Fidèle

Table des matières

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.

Date	TransportType	Route	DepartureTime	ArrivalTime	Passengers	DepartureStation	ArrivalStation	Delay
2023-02-01	Tram	Route_4	05:59	07:08	50	Station_10	Station_12	4
2023-02-01	Bus	Route_8	09:14	10:02	73	Station_15	Station_10	3
2023-02-01	Bus	Route_9	20:49	22:22	79	Station_16	Station_18	3
2023-02-01	Tram	Route_5	08:21	09:26	88	Station_9	Station_13	5
2023-02-01	Train	Route_9	14:19	14:56	95	Station_12	Station_13	7
2023-02-01	Metro	Route_8	18:03	18:22	38	Station_16	Station_5	15
2023-02-01	Metro	Route_10	09:07	10:53	7	Station_3	Station_3	0
2023-02-01	Metro	Route_4	17:12	18:30	52	Station_7	Station_10	8
2023-02-01	Metro	Route_9	19:52	20:53	92	Station_19	Station_1	8
2023-02-01	Tram	Route_4	14:03	15:18	47	Station_14	Station_4	8

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 :

Date	TransportType	Route	DepartureTime	ArrivalTime	Passengers	DepartureStation	ArrivalStation	Delay
2023-02-01	Tram	Route_4	05:59	07:08	50	Station_10	Station_12	4
2023-02-01	Bus	Route_8	09:14	10:02	73	Station_15	Station_10	3
2023-02-01	Bus	Route_9	20:49	22:22	79	Station_16	Station_18	3
2023-02-01	Tram	Route_5	08:21	09:26	88	Station_9	Station_13	5
2023-02-01	Train	Route_9	14:19	14:56	95	Station_12	Station_13	7
2023-02-01	Metro	Route_8	18:03	18:22	38	Station_16	Station_5	15
2023-02-01	Metro	Route_10	09:07	10:53	7	Station_3	Station_3	0
2023-02-01	Metro	Route_4	17:12	18:30	52	Station_7	Station_10	8
2023-02-01	Metro	Route_9	19:52	20:53	92	Station_19	Station_1	8
2023-02-01	Tram	Route_4	14:03	15:18	47	Station_14	Station_4	8

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

Route	RetardMoyen	NombreMoyenPassagers	NombreTotalVoyages
Route_9	13.797752808988765	54.2247191011236	89
Route_3	13.020833333333334	50.854166666666664	96
Route_7	14.016260162601625	52.203252032520325	123
Route_8	13.676767676767676	54.24242424242424	99
Route_4	10.054945054945055	52.61538461538461	91
Route_10	14.592592592592593	54.06172839506173	81
Route_1	13.0	50.07608695652174	92
Route_2	13.54320987654321	47.51851851851852	81
Route_6	13.314814814814815	55.129629629629626	108
Route_5	11.81	53.4	100

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.