Rapport de Projet – Architecture Business Intelligence (Scala)

Participant : Lisa AU, Zohra Hussan, Dorine HENRY
Classe : M2TL
Ecole : Digital Campus Paris
Année : 2025
Encadrant : Rakib Sheikh

Contexte

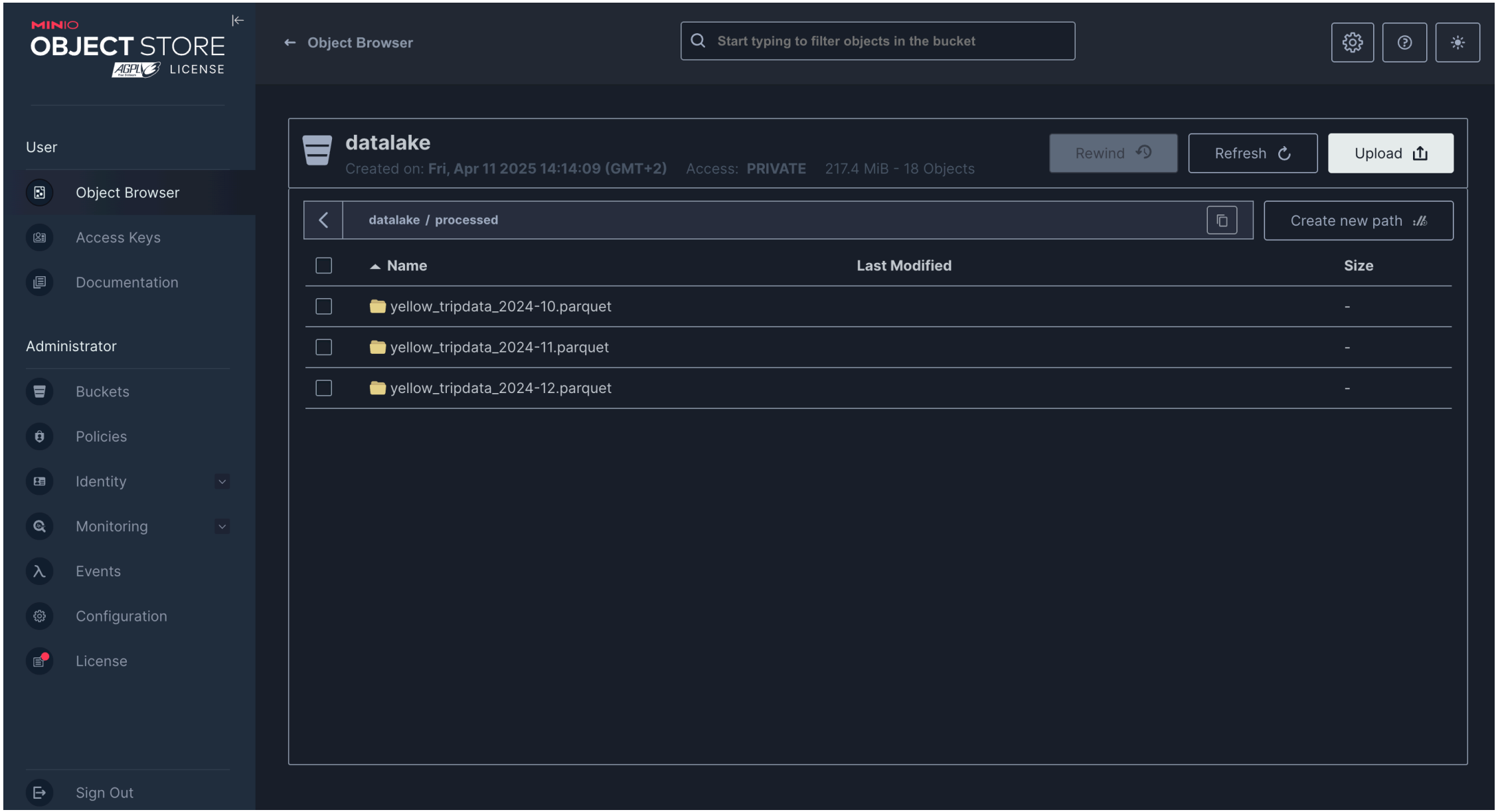
Dans le cadre du cours de Big Data, nous avons travaillé sur la mise en place d'une architecture BI de type CAC40, allant de la collecte de données jusqu'à l'orchestration de pipelines. Nous avons pu réaliser les TPs 1 à 4, mais par manque de temps, le TP5 n'a pas pu être finalisé.

TP1 – Collecte de données via Minio

Le TP1 a été réalisé assez facilement sur un poste sur trois, notamment sur macOS. Cependant, sous Windows, nous avons rencontré des problèmes avec le service Minio, principalement liés à Docker (réseau local, port mapping, compatibilité système de fichiers).

Étapes réalisées :

- Téléchargement des fichiers parquet (taxis jaunes de New York).
- Dépôt dans le dossier `data/raw/`.
- Upload vers le bucket Minio via un script Spark fourni.


MINIO
OBJECT STORE
AGPLv3
LICENSE
User
Object Browser
Access Keys
Documentation
Administrator
Buckets
Policies
Identity
Monitoring
Events
Configuration
License
Sign Out
Object Browser
Start typing to filter objects in the bucket
datalake
Created on: Fri, Apr 11 2025 14:14:09 (GMT+2)
Access: PRIVATE
217.4 MiB - 18 Objects
Rewind
Refresh
Upload
datalake / processed
Create new path
Name
Last Modified
Size
yellow_tripdata_2024-10.parquet
yellow_tripdata_2024-11.parquet
yellow_tripdata_2024-12.parquet

TP2 – Ingestion vers un Data Warehouse

Ce TP a été bien réussi, sur des machines Windows plus performantes.

Étapes :

- Lecture des fichiers parquet depuis Minio.
- Transformation et ingestion dans PostgreSQL en tant que Data Warehouse.
- Utilisation de scripts Scala/Spark pour automatiser l’ingestion.

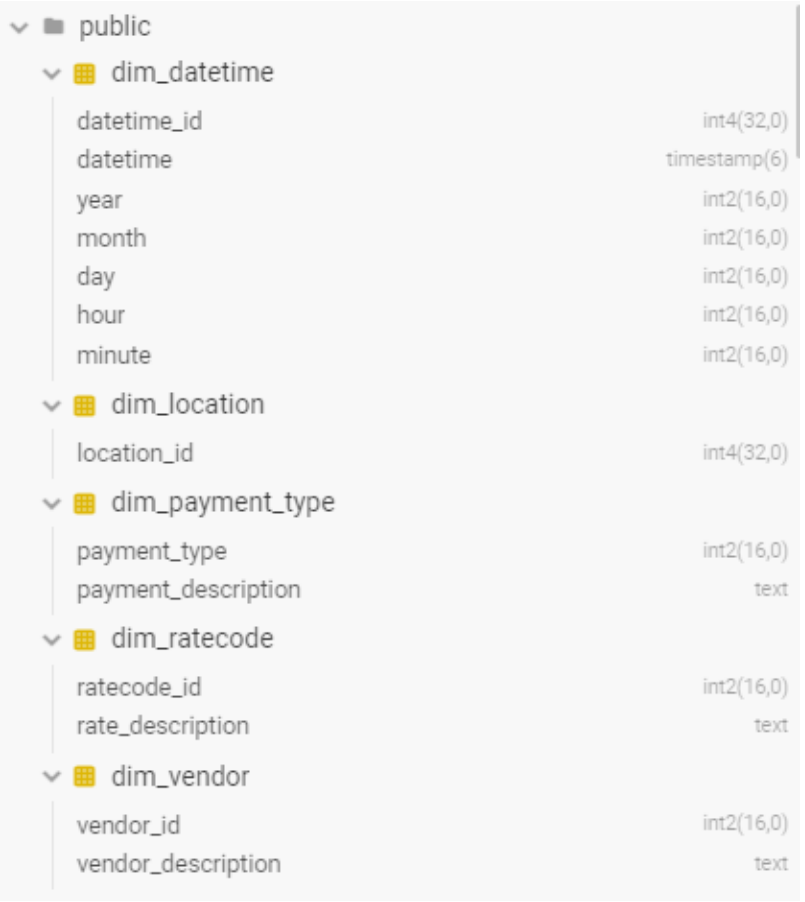
TP3 – Création du Data Mart et modélisation

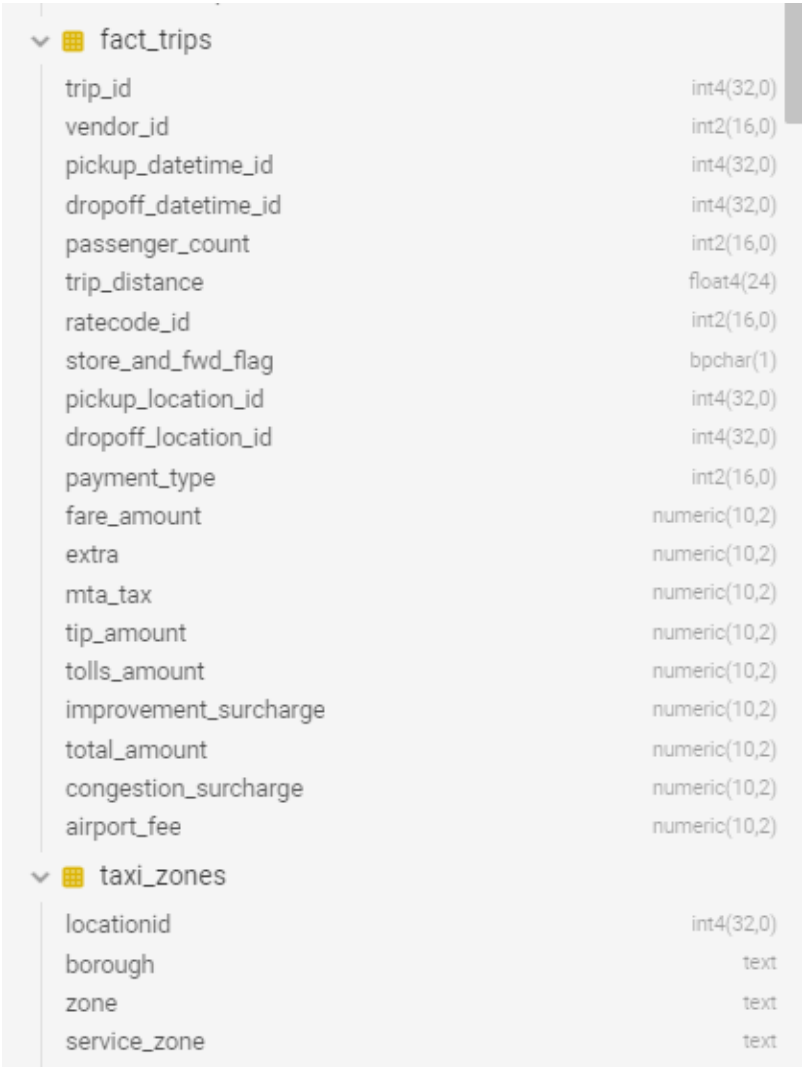
Complications initiales dues à une compréhension floue du besoin.
La partie SQL a été fluide, mais nous avons eu des difficultés de connexion dues à un manque d'informations reçues tardivement.

Points clés :

- Données utilisé : parquet des taxis jaunes de NYC de octobre 2024.
- Création d’un modèle en étoile.
- Scripts `creation.sql` et `insertion.sql`.
- Utilisation de deux serveurs SGBD distincts comme demandé.

Base de donnée





TP4 – Visualisation des données

Réalisé avec JupyterLab au format `.ipynb`.
Réalisation rapide grâce aux bases mathématiques vues le mois précédent.

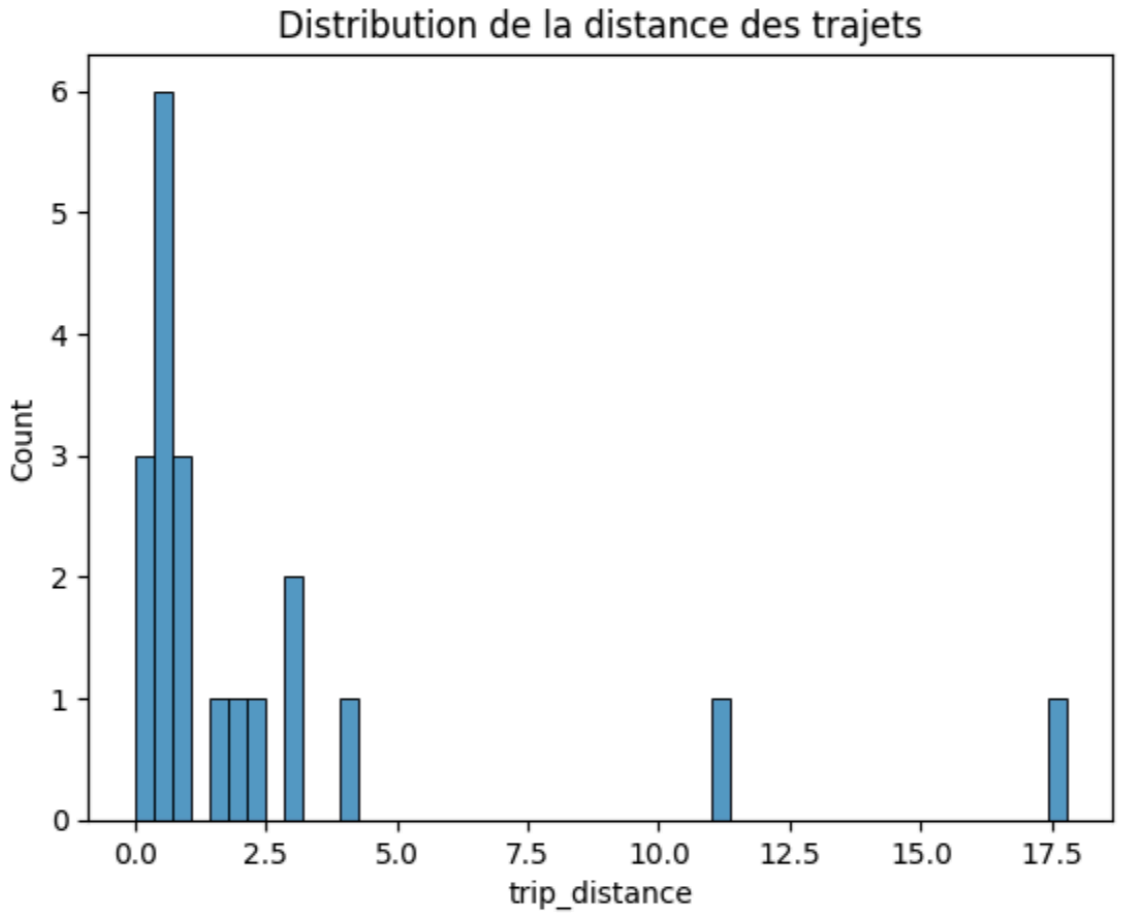
Étapes :

- Connexion au Data Mart.
- Analyse exploratoire des données (EDA).
- Génération de premières visualisations.

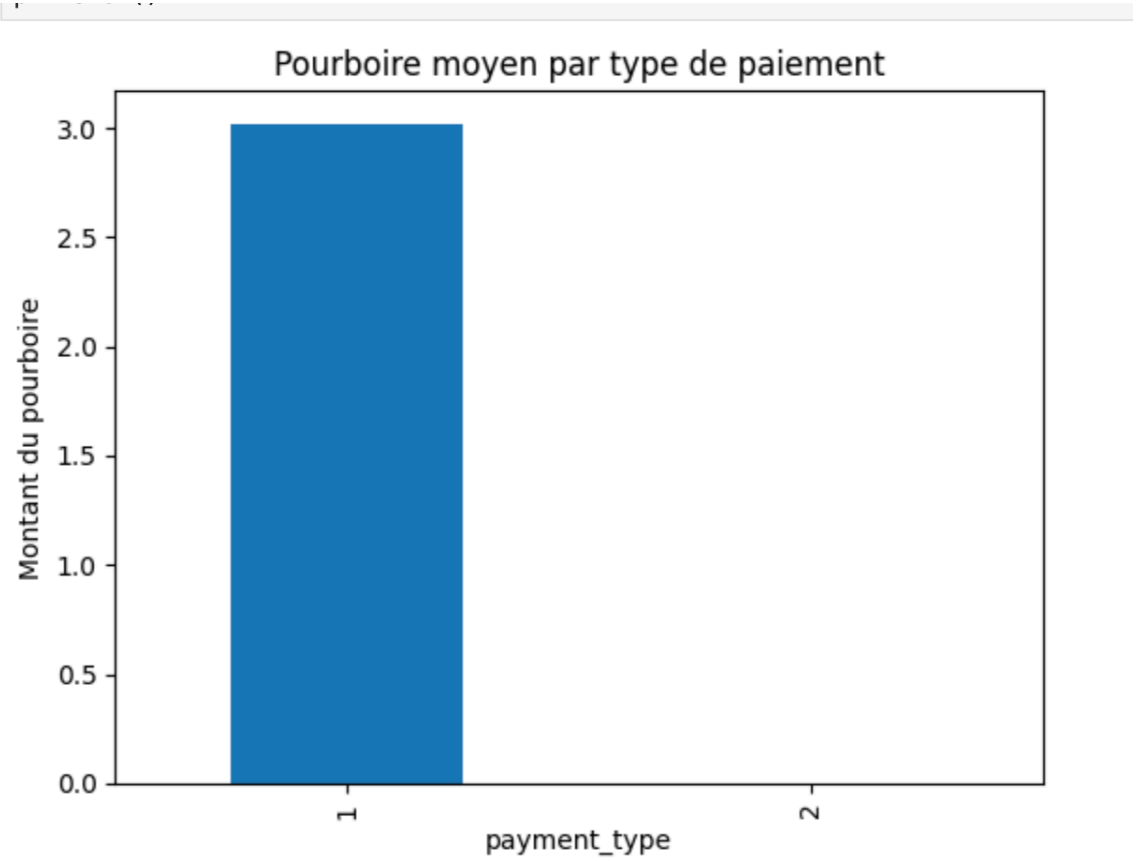
Apperçu des données

	trip_id	vendor_id	pickup_datetime_id	dropoff_datetime_id	passenger_count	trip_distance	ratecode_id	pickup_location_id
count	20.00000	20.000000	2.000000e+01	2.000000e+01	20.000000	20.000000	20.000000	20.000000
mean	10.50000	1.950000	5.250967e+05	9.695724e+05	1.050000	2.549500	1.450000	132.300000
std	5.91608	0.223607	6.197570e+05	7.258437e+05	0.223607	4.356014	1.234376	67.097808
min	1.00000	1.000000	5.497000e+03	2.419300e+04	1.000000	0.000000	1.000000	43.000000
25%	5.75000	2.000000	2.496400e+04	4.356855e+05	1.000000	0.527500	1.000000	79.000000
50%	10.50000	2.000000	1.236685e+05	8.579585e+05	1.000000	0.835000	1.000000	134.500000
75%	15.25000	2.000000	1.049532e+06	1.451914e+06	1.000000	2.540000	1.000000	155.250000
max	20.00000	2.000000	1.645849e+06	2.109584e+06	2.000000	17.770000	5.000000	249.000000

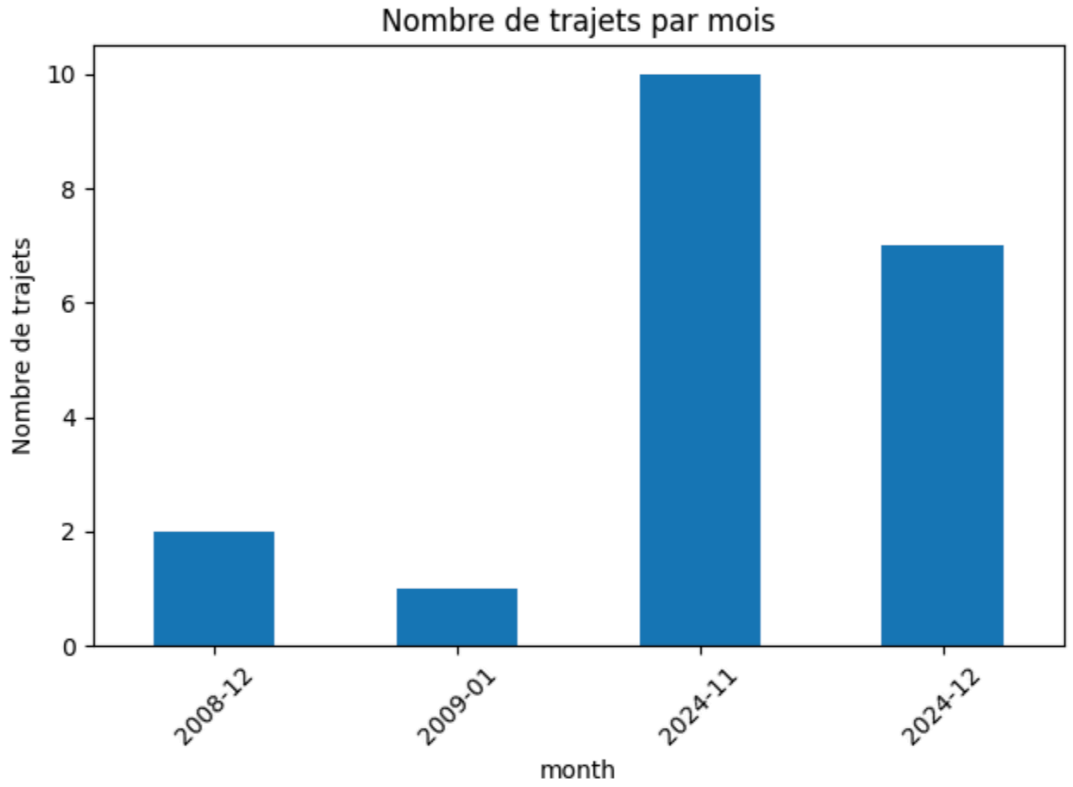
Visualisation des distributions



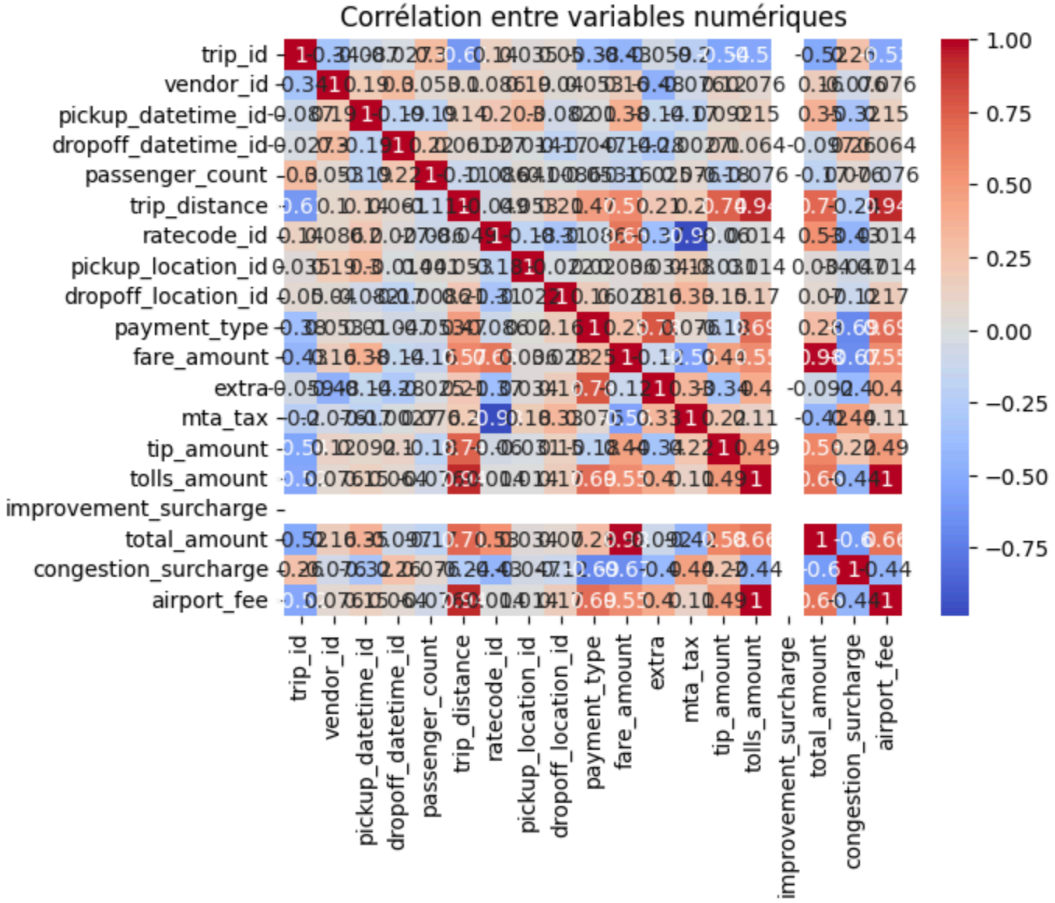
Moyenne des pourboires par type de paiement



Analyse temporelle



Corrélation entre variable



TP5 – Orchestration avec Airflow

Non réalisé.
Le projet étant hébergé sur des machines extérieures, nous n’avons pas eu le temps de finaliser cette partie.

Conclusion

Ce projet nous a permis de comprendre les différentes briques d’une architecture BI : collecte, transformation, modélisation et visualisation. Malgré l’absence du TP5, nous avons mis en œuvre un pipeline BI fonctionnel et structuré.