TP — Analyse d'un Data Lake Réseau

Karim Yassine

Introduction

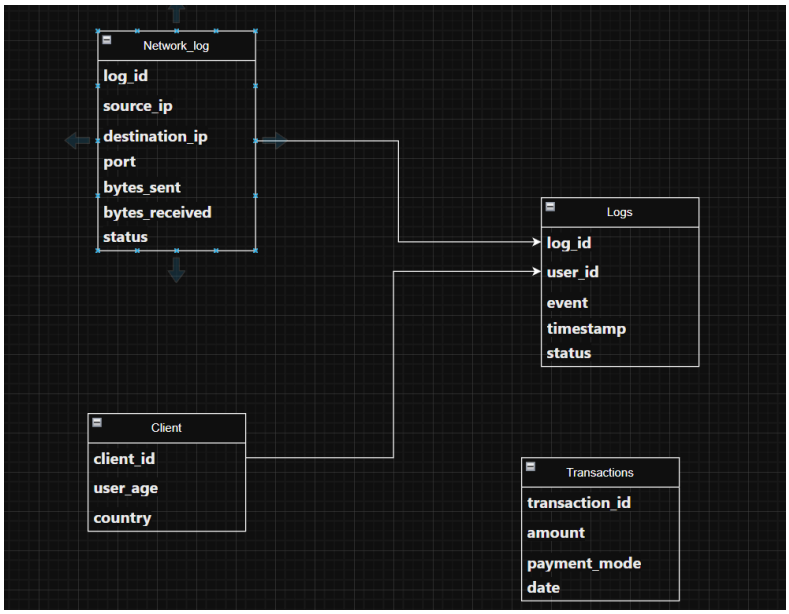
Dans ce TP, nous avons travaillé sur un data lake contenant des données réseau brutes. L'objectif était de transformer ces données désorganisées en informations exploitables, en suivant une méthodologie complète d'analyse de données. Ce type de travail est directement applicable dans un contexte professionnel, notamment en cybersécurité ou en data engineering, pour surveiller un réseau, détecter des anomalies et prendre de meilleures décisions.

Modèle Logique de Données (MLD)

Avant de coder quoi que ce soit, nous avons modélisé la structure de nos données sous forme d'un MLD. Ce schéma représente les 4 tables et leurs relations :

- Network_log est reliée à Logs via le champ log_id, ce qui permet de croiser un événement réseau avec l'activité d'un utilisateur
- Client est reliée à Logs via user_id, ce qui permet d'identifier quel client est derrière chaque connexion
- Transactions reste indépendante car elle ne contient pas de clé permettant de la relier aux autres tables

Ce modèle a directement guidé la création de la base SQL à l'étape 3, notamment pour le choix des index et des vues.



Les données utilisées

Nous avions 4 fichiers à notre disposition : logs.csv qui contient les événements utilisateurs comme les connexions et achats, network_logs.csv qui enregistre le trafic réseau avec les adresses IP et les ports, clients.csv qui regroupe les informations sur les utilisateurs, et transactions.csv qui liste les paiements effectués sur la période.

Étape 1 — Exploration et diagnostic

Avant de toucher aux données, nous avons d'abord analysé leur structure. Nous avons vérifié les dimensions de chaque fichier, les types de colonnes, la présence de valeurs manquantes et de doublons. Nous avons également contrôlé la cohérence métier des valeurs : les ports réseau doivent être entre 0 et 65535, les montants et les volumes de données doivent être positifs. Cette étape est essentielle pour comprendre ce qu'on a entre les mains avant d'agir.

Étape 2 — Nettoyage

Suite au diagnostic, nous avons identifié plusieurs problèmes à corriger. Les colonnes Usr-Ag, tX_ID et pmntMode portaient des noms non standards et ont été renommées. Les colonnes timestamp et date étaient stockées en texte brut et ont été converties au format datetime. Conformément aux bonnes pratiques et aux recommandations du RGPD, les données aberrantes n'ont pas été supprimées mais archivées dans des fichiers séparés pour garder une trace.

Étape 3 — Structuration en base SQL

Les données nettoyées ont été insérées dans une base SQLite structurée. Nous avons créé 4 tables en respectant le modèle logique de données établi au préalable : client, logs, network_log et transactions. Des index ont été ajoutés sur les colonnes les plus utilisées pour optimiser les requêtes. Des vues SQL ont également été créées pour pouvoir réutiliser facilement les requêtes les plus fréquentes.

Étape 4 — Analyse des tendances (KPI)

Nous avons extrait 5 indicateurs clés via des requêtes SQL : le top 10 des utilisateurs les plus actifs, l'évolution du trafic réseau heure par heure, le top 10 des adresses IP sources, les ports les plus sollicités et la répartition des modes de paiement. Ces KPI permettent d'avoir une vue synthétique du comportement du réseau sur la période analysée.

Étape 5 — Visualisation interactive

Les résultats ont été présentés sous forme d'un dashboard interactif développé avec Plotly et Dash. Il contient 6 visualisations accessibles via des onglets : une courbe d'évolution du trafic, des graphiques en barres pour les utilisateurs, les IPs et les ports, un graphique circulaire pour les paiements, et une heatmap croisant les adresses IP sources avec les ports utilisés.

Conclusion

Ce TP nous a permis de parcourir l'intégralité d'un pipeline de données, de la donnée brute jusqu'au dashboard. Chaque étape avait un rôle précis et s'appuyait sur la précédente. Les outils utilisés, Python, Pandas, SQLite et Plotly, sont des standards du marché dans les métiers de la data et de la cybersécurité.