Big Data – Cas pratique

Introduction

Ce projet est un analyseur de grandes données développées sur VScode en utilisant Python et Jupyter Notebook, il permet d'analyser un catalogue Netflix et de nettoyer et visualiser les données.

1. Choix de Dataset et Technos

J'ai choisi un catalogue Netflix de 8000+ séries et films Netflix présents dans un csv téléchargé de Kaggle. Or sa taille ce Dataset est aussi parfait pour cet exercice car il contient une diversité de données (textuelles, numériques et temporelles) ainsi que ces défis techniques, il contient plusieurs erreurs réalistes comme des données manquantes et formats incohérents.

J'ai choisi d'utilser Python avec **Pandas** et **Numpy** au lieu de **Pyshark** car la taille du Dataset correspondait aux capacités de **Pandas** ce qui va donc permettre à un développement plus rapide et une intégration de visualisions de données plus simples avec **matplotlib** et **seaborn, matplotlib** et pour le stockage de données j'ai utilisé des fichier CSV avec **Pandas DataFrames** ce qui est parfait pour mes tailles de fichier.

2. Traitements Réalisés

a. Nettoyage et transformation de données

J'ai effectué un nettoyage des données qui :

- Supprime les doublons
- Convertit le champ "date_added" en format datetime et extrait les composants mois/année

- Divise les données séparées par virgules (genres, pays, réalisateurs) et les convertit à un meilleu format

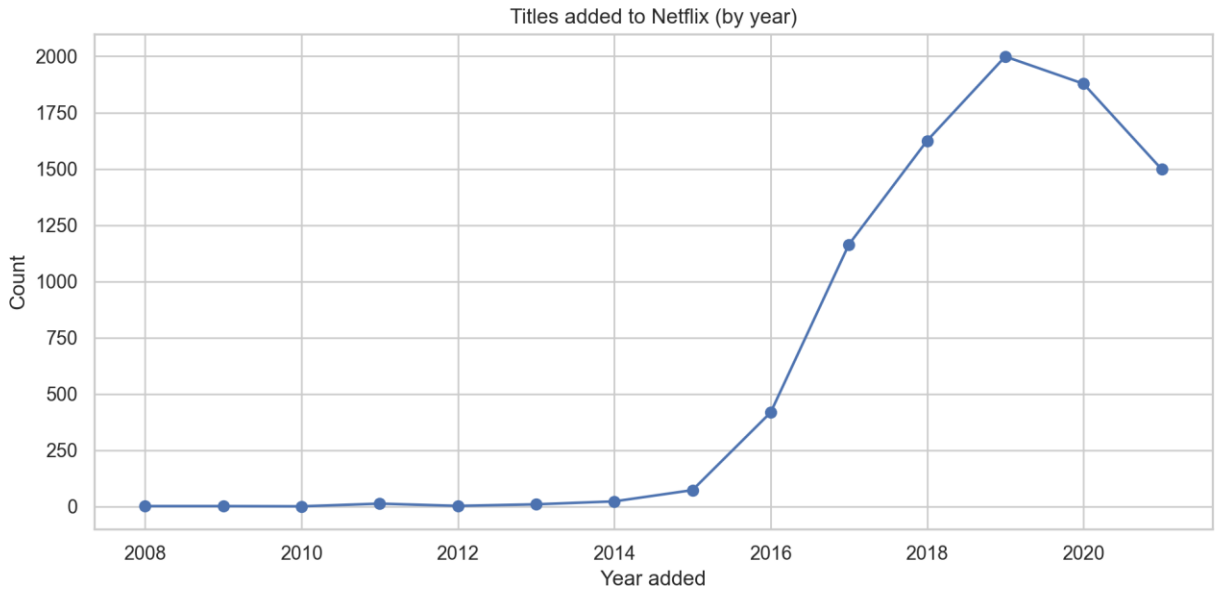
b. Analyse descriptive

J'ai lancé plusieurs analyses pour extraire des insights

- Répartition de Films vs Séries
- Identifications des pays producteurs les plus populaires
- Identification des genres les plus populaires
- Identification du nombre de film par année au fil du temps
- Identification des Réalisateurs les plus prolifiques
- Statistiques de durée

3. Résultats Observés

Il semble y avoir plus de films que de séries sur Netflix et le pays le plus prolifique est les USA. Les 3 genres les plus populaires sont “Films internationaux – Dramas - Comédies” Et le pic des additions par années semble être 2019 et depuis c’est en chute probablement avec la montée des différentes applications de streaming.



4. Les limites et pistes d’amélioration

L’un des aspects limitants est le dataset en lui-même bien que suffisant pour cette exercice 8000+ données ne représentent qu’une fraction des vrais volumes big data, l’utilisation de **Pandas** est donc parfaite pour ce cas spécifique or pour des cas avec des millions de données un calcul en cluster est nécessaire. Aussi bien que çà représente des problèmes réels le Dataset que j’ai utilisé manque de beaucoup trop de données importantes affectant la complétude de l’analyse (mon réalisateur le plus populaire est “unknown”)

Pour ce qui est des potentiels améliorations, une implémentation de **Pyshark** et **AWS EMR** pour des Datasets à des multi millions. **Apache kafka** et **Spark Streaming** permettent aussi le traitement de données en temps réel.
Remplacer les fichiers csv par une base de données **PostgreSQL** et utiliser **Dash** pour une exploration dynamique des visualisations.