**Base de données NOSQL et Big Data**

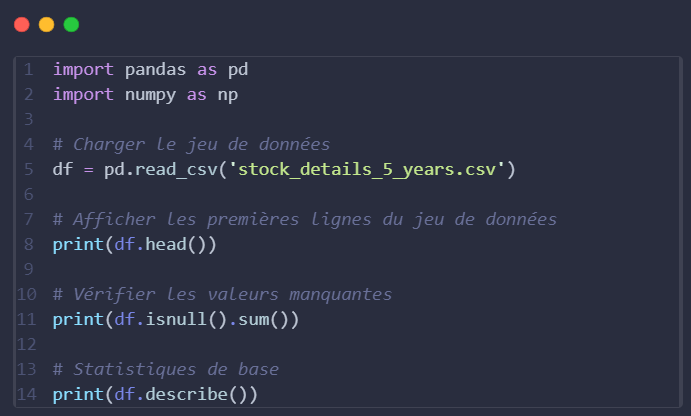
**Projet semestrielle: MapReduce**

**Introduction :**

L'objectif de cette analyse est d'explorer les données historiques sur les prix des actions de 500 grandes entreprises au cours des cinq dernières années. Les données sont extraites de Yahoo Finance et incluent des métriques quotidiennes telles que l'ouverture, la fermeture, le volume, etc. L'analyse vise à fournir des insights pour les investisseurs, les chercheurs et les passionnés d'apprentissage automatique.

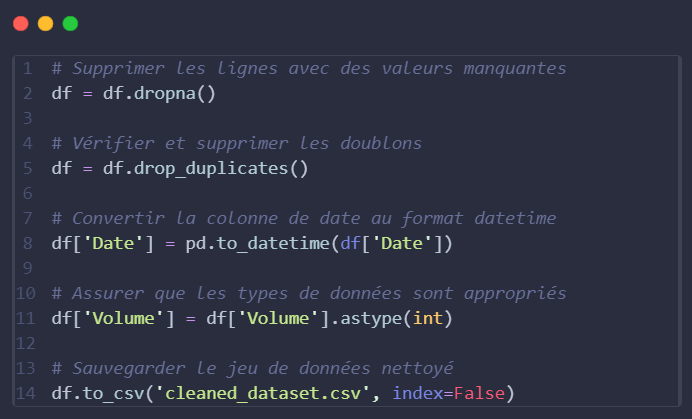
**Étape 1: Exploration des Données**

Dans cette étape, nous avons chargé le jeu de données à l'aide de la bibliothèque Pandas en Python. Nous avons affiché les premières lignes du jeu de données pour comprendre sa structure et vérifié la présence de valeurs manquantes et effectué des statistiques de base.



**Étape 2: Nettoyage des Données**

Dans cette étape, nous avons supprimé les lignes avec des valeurs manquantes, éliminé les doublons et converti la colonne de date au format datetime. Nous avons également assuré que les types de données étaient appropriés.



**Étape 3: Visualisation des Données**

Nous avons utilisé la bibliothèque Matplotlib pour visualiser les prix des actions d'une entreprise spécifique au fil du temps.



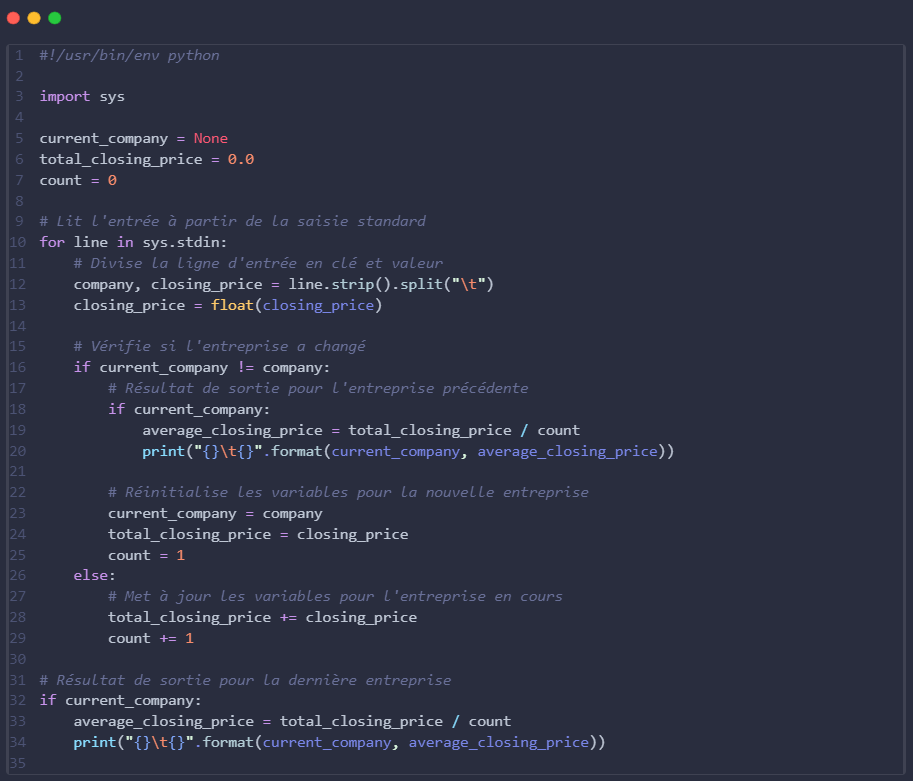
* **MapReduce avec Hadoop**

Nous avons conçu une tâche MapReduce pour calculer le prix de clôture moyen par entreprise sur l'ensemble du jeu de données. Le script MapReduce est écrit en Java et comprend deux parties principales : le mappage et la réduction. Le mappage extrait les paires clé-valeur à partir des données, tandis que la réduction agrège ces paires pour calculer la moyenne.

* **Mapper.py:**



* **Reducer.py:**

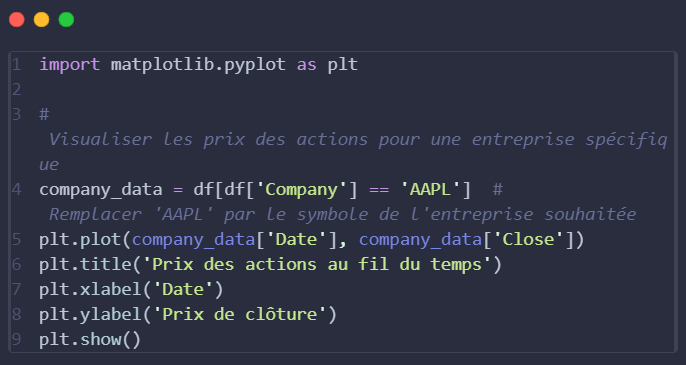


* **MapReduce avec multiprocessing (Python)**

Nous avons également implémenté une version MapReduce en utilisant le module multiprocessing en Python. Cela permet de tirer parti du traitement parallèle sur une machine locale.

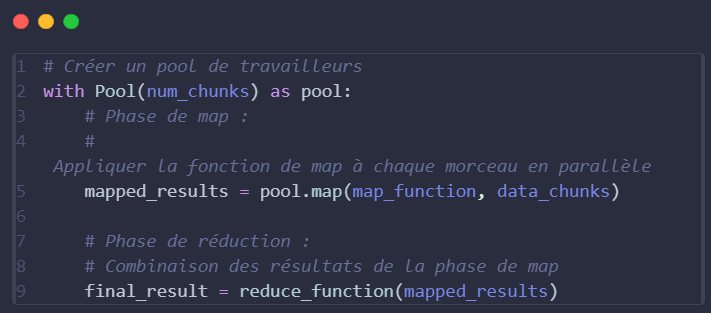
**Étape 4: MapReduce Design**

Dans cette étape, nous concevons le processus MapReduce en définissant les fonctions de mappage et de réduction ainsi que la stratégie de division des données en morceaux pour le traitement parallèle.



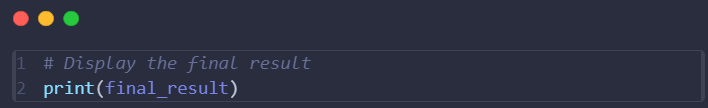
**Étape 5: MapReduce Implementation**

Dans cette étape, nous mettons en œuvre le processus MapReduce en utilisant le module multiprocessing en Python. Cela permet de tirer parti du traitement parallèle sur une machine locale.



**Étape 6: MapReduce Execution**

Dans cette étape, nous exécutons le processus MapReduce en affichant le résultat final, qui peut être utilisé pour des analyses plus approfondies.



**Conclusion**

Ce rapport a couvert toutes les étapes du processus d'analyse des données sur les prix des actions, de l'exploration initiale à l'application de MapReduce avec la version parallèle en utilisant le module multiprocessing en Python. Ces méthodes offrent une approche scalable pour traiter des volumes importants de données. Les résultats peuvent être utilisés pour prendre des décisions éclairées en matière d'investissement, d'analyse de marché et de modélisation prédictive.