Analyse des Données avec Spark et Polars

Projet Big Data

Ibrahima Ndao & Leopold Dieng , 24 mars 2025

Contexte du Projet

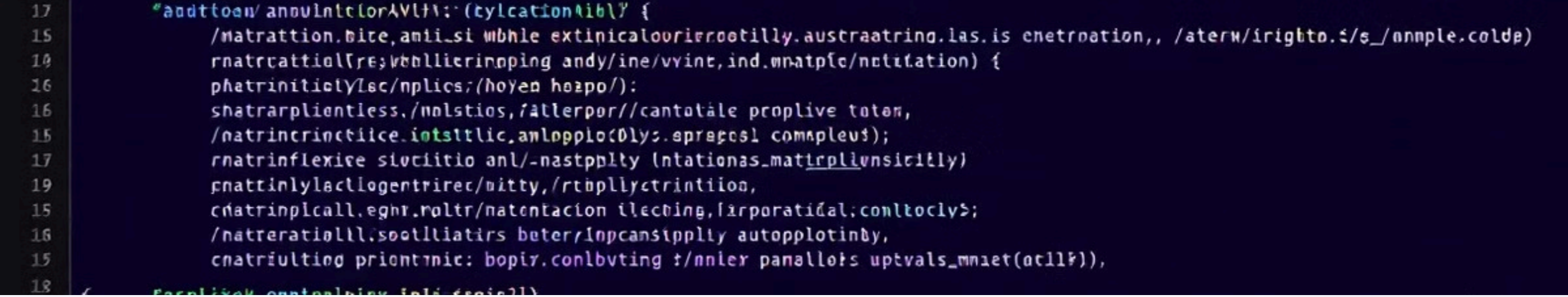
Ce projet vise à comparer Spark et Polars pour l'analyse de données. Nous évaluerons leurs performances et leurs cas d'utilisation. Le jeu de données provient de **Motor Vehicle Crashes**.

Objectifs

- Analyse rigoureuse d’un jeu de données
- Évaluer Spark et Polars

Source des données

- Accidents de voiture sur trois ans



Analyse avec Spark sur Databricks

L'analyse avec Spark sur Databricks a suivi plusieurs étapes clés. Cela comprenait l'importation des données, la création d'un notebook PySpark, et l'analyse des données.

1 Importation des données

Importation des données dans le Databricks File System (DBFS).

2 Création d'un notebook PySpark

Initialisation et configuration du notebook PySpark.

3 Analyse des données

Nettoyage, exploration, statistiques descriptives et visualisation des tendances.

Mise en Place de la Plateforme de Données

Pour la mise en place, nous avons utilisé PostgreSQL, Docker, et Polars. Les données analysées ont été sauvegardées dans MinIO pour une gestion efficace.

1

PostgreSQL et Docker

Déploiement de PostgreSQL avec Docker.

2

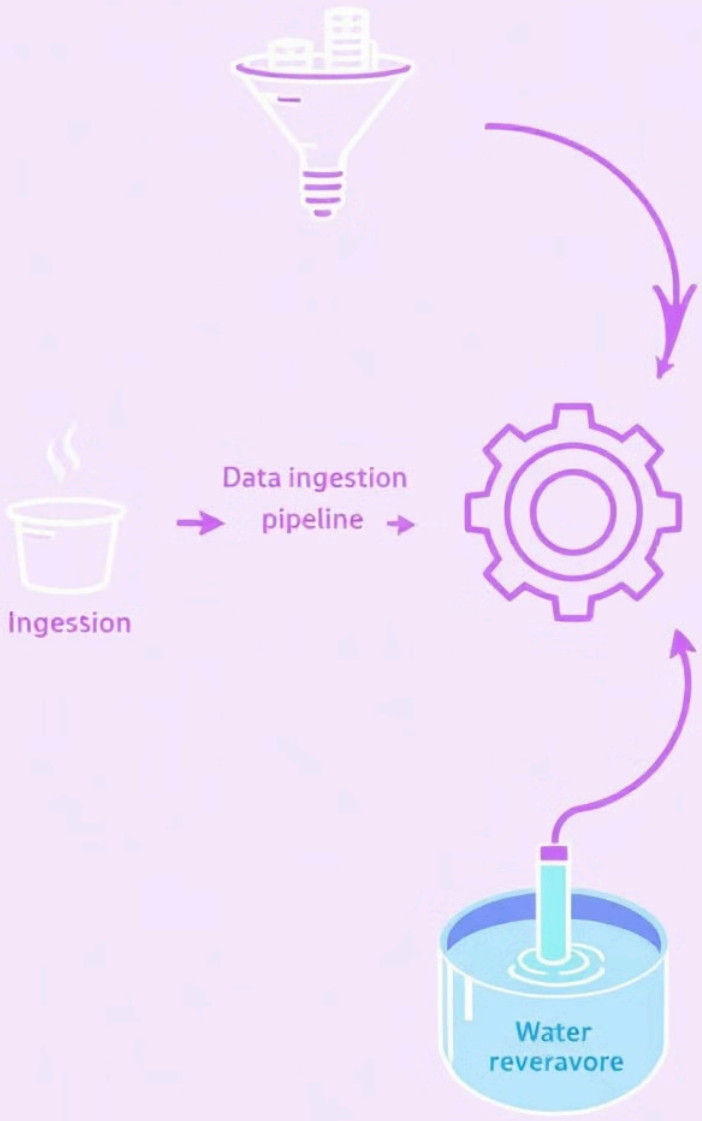
Analyse avec Polars

Connexion à PostgreSQL depuis un notebook en local.

3

Intégration avec MinIO

Sauvegarde des données analysées.



Comparaison Spark vs Polars

Spark excelle pour les workflows distribués, tandis que Polars est idéal pour les analyses rapides et locales. Cette analyse a examiné la performance, les avantages et les inconvénients des deux outils.

Performance

Avantages et inconvénients

Cas d’utilisation

Résultats Finaux

Les deux outils sont complémentaires. Nous avons mis en place une plateforme robuste. Ce processus est reproductible pour des projets similaires.

1

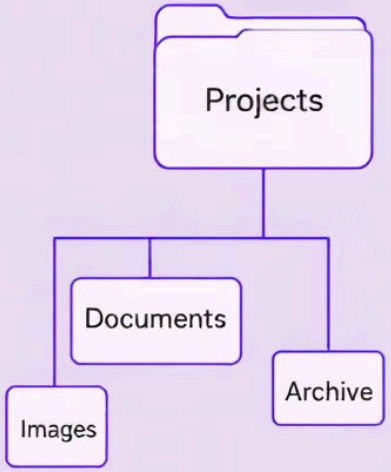
Bilan

Complémentarité des outils.

2

Contributions

Processus reproductible.



Structure des Fichiers

Les notebooks, scripts et documentations sont organisés. Les analyses Spark et Polars sont dans des notebooks distincts. Un document comparatif détaillé est inclus.

Notebooks

- notebook_databricks.ipynb
- notebook_vscode_polars.ipynb

Scripts

- insert_data.py
- docker-compose.yml

Documentation

- Fichier PowerPoint ou PDF
- Document comparatif

Conclusion

Spark et Polars sont complémentaires. La plateforme est scalable. Prochaines étapes: étendre à d'autres jeux de données, améliorer l'automatisation.

1

Outils

Spark et Polars

2

Plateforme

Scalable et conforme

3

Prochaines étapes

Étendre et automatiser