INSTITUT POLYTECHNIQUE DE SAINT-LOUIS

Rapport de projet du cours de :

# Big DATA

Présenté par :
**Oumar FALL**
Élève ingénieur en ing3 génie informatique
2023-2024

Aout 09, 2024

Responsable du cours : Monsieur Djibril MBOUP

# Étude Comparative des Formats de Fichiers en Big Data: Parquet, ORC, Avro, et Apache Arrow

## Introduction

Dans le domaine du Big Data, le choix du format de fichier est crucial pour optimiser le stockage, la performance, et l'intégration des données. Parmi les formats les plus populaires figurent Parquet, ORC, Avro, et Apache Arrow. Cette étude comparative vise à explorer les avantages et inconvénients de chacun de ces formats, afin de mieux comprendre leurs applications et utilisations optimales.

## Apache Parquet

Parquet est un format de stockage en colonnes, optimisé pour les requêtes analytiques à grande échelle. Il est largement utilisé dans les environnements Big Data pour sa capacité à compresser les données et à améliorer les performances de lecture.

### Avantages

- Stockage en colonnes permettant une compression efficace et des performances accrues pour les requêtes analytiques.
- Support étendu dans les écosystèmes Big Data, notamment Hadoop, Spark, et Hive.
- Flexible, prend en charge différents types de données complexes.

### Inconvénients

- Moins adapté aux opérations de lecture-écriture fréquentes.
- Plus complexe à manipuler comparé à d'autres formats comme Avro.

## Optimized Row Columnar (ORC)

Le format ORC est similaire à Parquet mais conçu spécifiquement pour optimiser les performances dans les systèmes d'analyse de données massives comme Hive. Il est également un format en colonnes, offrant des avantages importants en termes de compression et de rapidité d'accès aux données.

### Avantages

- Haute efficacité en termes de compression des données.
- Support pour les index de position, permettant des lectures rapides.
- Conception optimisée pour les environnements Hadoop et Hive.

### Inconvénients

- Moins de support en dehors de l'écosystème Hadoop/Hive.
- Complexité accrue dans la gestion des schémas.

## Apache Avro

Avro est un format de sérialisation des données qui supporte à la fois le stockage en ligne et la communication des messages. Contrairement à Parquet et ORC, Avro est un format orienté ligne, ce qui le rend plus adapté pour les opérations de lecture-écriture fréquentes.

### Avantages

- Très efficace pour la sérialisation et désérialisation des données.
- Supporte les schémas évolutifs, facilitant les mises à jour des données.
- Format compact, bien adapté aux transactions fréquentes.

### Inconvénients

- Moins performant pour les requêtes analytiques comparé à Parquet et ORC.
- Moins efficace en termes de compression de données volumineuses.

## Apache Arrow

Apache Arrow est un format de données en mémoire conçu pour accélérer les processus d'analyse de données. Il est souvent utilisé comme format d'échange entre différents systèmes, permettant une communication efficace entre différents langages et plateformes.

### Avantages

- Hautement performant pour le traitement en mémoire des données.
- Format en colonne, offrant de bonnes performances pour les opérations analytiques.
- Large compatibilité avec d'autres systèmes et formats de données.

### Inconvénients

- Principalement conçu pour l'utilisation en mémoire, moins adapté pour le stockage persistant.
- Complexité accrue pour la gestion de grandes quantités de données persistantes.

# Conclusion

Le choix du format de fichier dans un environnement Big Data dépend largement des besoins spécifiques du projet. Parquet et ORC sont idéaux pour les requêtes analytiques sur de grandes quantités de données, tandis qu'Avro est mieux adapté aux opérations de lecture-écriture fréquentes. Apache Arrow, quant à lui, se distingue par sa performance en mémoire, facilitant l'interopérabilité entre systèmes. En fin de compte, une compréhension approfondie des avantages et inconvénients de chaque format permet de faire un choix éclairé en fonction des objectifs et contraintes du projet.