MARIAME DIALLO ING3INFO

Travail à faire : Étude comparative (pros & cons) sur les formats de fichiers parquet, orc, avro, apache arrow.

1. Apache Parquet

Tableau : Avantages et Inconvénients

Avantages	Inconvénients
Colonnes stockées séparément pour une lecture optimisée	Pas optimisé pour les opérations en écriture fréquente
Compression efficace avec plusieurs algorithmes	Complexité de gestion des schémas
Bien supporté par de nombreux outils	
Supporte des types de données complexes	

Résumé et Sources

Parquet est un format de stockage en colonnes très populaire dans les environnements de big data, offrant une compression efficace et une interopérabilité étendue avec des outils tels qu'Apache Spark et Hive. Cependant, il n'est pas idéal pour les mises à jour fréquentes de petites quantités de données et peut nécessiter une gestion complexe des schémas.
[Source](https://parquet.apache.org/documentation/latest/).

2. ORC (Optimized Row Columnar)

Tableau : Avantages et Inconvénients

Avantages	Inconvénients
Compression et stockage très efficaces	Moins supporté en dehors de Hive
Performance accrue avec des fonctionnalités d'indexation	Écriture moins flexible
Optimisé pour Apache Hive	

Résumé et Sources

Le format ORC est optimisé pour les performances et la compression, particulièrement dans les environnements Apache Hive. Il est moins largement adopté que Parquet, mais il offre d'excellentes fonctionnalités d'indexation pour les requêtes analytiques. Sa flexibilité en écriture est limitée, ce qui

peut poser problème pour certaines applications.
[Source](https://orc.apache.org/docs/indexes.html).

3. Apache Avro

Tableau : Avantages et Inconvénients

Avantages	Inconvénients
Sérialisation rapide	Format basé sur les lignes, moins optimisé pour les requêtes analytiques
Support de schéma intégré	Compression moins efficace pour les charges analytiques
Prise en charge de schémas évolutifs	

Résumé et Sources

Avro est souvent utilisé pour la sérialisation des données grâce à sa rapidité et son support de schémas intégrés. C'est un excellent choix pour les cas où les schémas peuvent évoluer au fil du temps. Cependant, en tant que format basé sur les lignes, il n'est pas aussi performant que Parquet ou ORC pour les requêtes analytiques. [Source](https://avro.apache.org/docs/current/).

4. Apache Arrow

Tableau : Avantages et Inconvénients

Avantages	Inconvénients
Mémoire partagée pour des opérations en mémoire rapide	Pas un format de stockage à long terme
Optimisé pour le calcul en colonne	Peu d'outils natifs pour la sérialisation persistante
Bonne interopérabilité avec d'autres frameworks	

Résumé et Sources

Apache Arrow est conçu pour faciliter le calcul rapide en mémoire, avec un fort accent sur l'optimisation des opérations en colonne. Il est particulièrement adapté pour être utilisé en conjonction avec d'autres outils de calcul comme Pandas ou Spark. Arrow n'est pas destiné à être un format de stockage à long terme, ce qui le limite dans certains contextes.
[Source](https://arrow.apache.org/overview/).