

Étude comparative (pros & cons) sur les formats de fichiers parquet, orc, avro, apache arrow

Voici une étude comparative des formats de fichiers Parquet, ORC, Avro, et Apache Arrow, présentée sous forme de pros et cons.

1. Parquet

Pros:

- Optimisé pour les requêtes analytiques.
- Compression efficace.
- Large adoption et interopérabilité.
- Support des types complexes.

Cons:

- Moins performant pour l'écriture.
- Plus complexe à mettre en œuvre.

2. ORC

Pros:

- Très performant pour les requêtes analytiques.
- Compression et index intégrés.
- Support des types complexes.
- Optimisé pour Apache Hive.

Cons:

- Moins supporté en dehors de l'écosystème Hadoop.
- Complexe à utiliser.

3. **Avro**

Pros:

- Idéal pour les opérations de lecture/écriture séquentielles.
- Schéma intégré.
- Bonne interopérabilité.
- Sérialisation rapide.

Cons:

- Moins optimisé pour les requêtes analytiques.
- Compression moins efficace.

4. **Apache Arrow**

Pros:

- Conçu pour le traitement des données en mémoire.
- Échange rapide entre applications.
- Performant pour les calculs vectorisés.
- Support des types complexes.

Cons:

- Relativement nouveau et moins mature.
- Écosystème limité.
- Moins adapté pour le stockage persistant.