Parquet File

Parquet is an open-source file format that is widely used in the big data ecosystem, particularly in the Apache Hadoop ecosystem. It is designed to store and process large amounts of data more efficiently. Parquet is optimized for complex processing and is particularly well-suited for analytics workloads.

Key features of the Parquet file format include:

1. **Columnar Storage:** Parquet stores data in a columnar fashion rather than the traditional row-based format. This allows for more efficient compression, as data with similar characteristics is stored together.
2. **Compression:** Parquet uses various compression techniques to minimize storage space, making it more efficient for storing and processing large datasets.
3. **Schema Evolution Support:** Parquet files can handle schema evolution, meaning you can easily add, remove, or modify fields in the schema without affecting the data stored in the file.
4. **Splitting and Partitioning:** Parquet files can be split into multiple parts, making it easy to parallelize processing tasks. Additionally, the partitioning feature allows for efficient data pruning and filtering during query execution.
5. **Compatibility:** Parquet files are compatible with a wide range of data processing frameworks and tools, making them a popular choice for data storage in distributed computing environments.

Due to these features, Parquet is commonly used for big data processing tasks, including data warehousing, analytics, and data processing in systems like Apache Hadoop, Apache Spark, and other distributed computing frameworks.