# Apache Iceberg VS Apache Parquet

## Apache Iceberg:

Apache Iceberg is a table format whose purpose is mainly focuses on data processing.

* Iceberg abstracts the underlying stored data as a table of columns and rows.

## Apache Parquet:

Apache Parquet is a columnar file format whose purpose is mainly focuses on data storage.

* Parquet provides compression of data.

## Why Apache Iceberg is better than Apache Parquet?

### Usecase-1: Schema Evolution

* **Iceberg** allows for schema evolution without requiring the costly re-writes of the entire dataset.
  + It supports adding/dropping/renaming the columns and updates/deletes/inserts on the data efficiently.
* **Parquet** on the other hand needs to re-write the data files to ensure compatibility which is costlier.

### Usecase-2: Data Management and Versioning

#### Transactional Support

* **Apache Iceberg** supports ACID transactions, allowing for more efficient handling of updates/deletes/inserts.
  + It also provides versioned snapshots which makes it easier to track changes and to revert to the previous versions if needed.
* **Parquet**, on the other hand, does not provide built-in transaction support. While you can achieve transactional behavior in Parquet using tools like Apache Hive or Delta Lake, it's not a native feature of the file format.

#### Partitioning

* **Iceberg** offers optimizations like **predicate-pushdown**, **partition pruning** capability which filters the data at the storage level rather than at the processing stage which reduces the I/O and improves performance of the process.
  + This doesn’t require manual intervention.
* **Parquet files** can be partitioned but managing these partitions requires manual effort

### Usecase-3: Storage

#### Columnar Storage vs Flexible Storage

* **Parquet** is a columnar storage file format, which means it stores data by columns rather than rows. This storage format is particularly useful for analytical queries that access a small subset of columns in a table.
* **Iceberg**, on the other hand, is a table format rather than a file format. It can work with **various file formats**, including Parquet, ORC, and Avro. This flexibility allows you to choose the best file format based on your specific use case and storage requirements.

## Why Apache Parquet is better than Apache Iceberg?

## Use case: High-Performance Analytics on Stable, Read-Only data

* **Scenario**:  
  If we need to perform read-only analytical querying which doesn’t have modifications less frequently or not at all mainly for reporting/business purposes.
* **Parquet** is optimized for read performance, providing efficient storage by compressing the data and supports the columnar storage which helps in the faster retrieval of the columnar data.
  + As it is columnar, it facilitates to read-only the relevant columns required for analytics.
  + As the data is stable (which doesn’t undergo modifications), it doesn’t require the features of Iceberg such as ACID transactions etc.