**What is Parquet?**

Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk. Apache Parquet is designed to be a common interchange format for both batch and interactive workloads. It is similar to other columnar-storage file formats available in [**Hadoop**](https://databricks.com/glossary/hadoop), namely RCFile and ORC.

**Characteristics of Parquet**

* **Free and open source file format.**
* **Language agnostic.**
* **Column-based format** - files are organized by column, rather than by row, which saves storage space and speeds up analytics queries.
* **Used for analytics (OLAP) use cases**, typically in conjunction with traditional OLTP databases.
* **Highly efficient** data compression and decompression.
* **Supports complex data types** and advanced nested data structures.

**Benefits of Parquet**

* **Good for storing big data of any kind** (structured data tables, images, videos, documents).
* **Saves on cloud storage space** by using highly efficient column-wise compression, and flexible encoding schemes for columns with different data types.
* **Increased data throughput and performance** using techniques like data skipping, whereby queries that fetch specific column values need not read the entire row of data.

Apache Parquet is implemented using the record-shredding and assembly algorithm, which accommodates the complex data structures that can be used to store the data. Parquet is optimized to work with complex data in bulk and features different ways for efficient data compression and encoding types. This approach is best especially for those queries that need to read certain columns from a large table. Parquet can only read the needed columns therefore greatly minimizing the IO.

**Advantages of Storing Data in a Columnar Format:**

* Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases. This way of storage has translated into hardware savings and minimized latency for accessing data.
* Apache Parquet is built from the ground up. Hence it is able to support advanced nested data structures. The layout of Parquet data files is optimized for queries that process large volumes of data, in the gigabyte range for each individual file.
* Parquet is built to support flexible compression options and efficient encoding schemes. As the data type for each column is quite similar, the compression of each column is straightforward (which makes queries even faster). Data can be compressed by using one of the several codecs available; as a result, different data files can be compressed differently.
* Apache Parquet works best with interactive and serverless technologies like AWS Athena, Amazon Redshift Spectrum, Google BigQuery and Google Dataproc.

**Difference Between Parquet and CSV**

CSV is a simple and common format that is used by many tools such as Excel, Google Sheets, and numerous others. Even though the CSV files are the default format for data processing pipelines it has some disadvantages:

* Amazon Athena and Spectrum will charge based on the amount of data scanned per query.
* Google and Amazon will charge you according to the amount of data stored on GS/S3.
* Google Dataproc charges are time-based.

Parquet has helped its users reduce storage requirements by at least one-third on large datasets, in addition, it greatly improved scan and deserialization time, hence the overall costs. The following table compares the savings as well as the speedup obtained by converting data into Parquet from CSV.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Size on Amazon S3 | Query Run Time | Data Scanned | Cost |
| Data stored as CSV files | 1 TB | 236 seconds | 1.15 TB | $5.75 |
| Data stored in Apache Parquet Format | 130 GB | 6.78 seconds | 2.51 GB | $0.01 |
| Savings | 87% less when using Parquet | 34x faster | 99% less data scanned | 99.7% savings |

Reference Link :

[Parquet – Databricks](https://databricks.com/glossary/what-is-parquet)