

Big Data file formats

Some of the common big data file formats are noted below:

- **Text/CSV Files:** These are the usual delimited files that you normally see for most raw.
- **Avro:** Apache Avro is a data serialization system that provides a compact, fast binary format. It relies on schemas to make sense of the data in the file.
- **Parquet:** Apache Parquet is a columnar storage format that can be used by different projects in the Hadoop ecosystem. It is built to support very efficient compression and encoding schemes.
- **ORC (optimized Row Columnar):** In this format data is stored in a hybrid fashion, it stores collections of rows and within a collection different columns. It also introduces indexing and statistics like min and max.