Data Storage Formats in Hive

We will see different file formats for storing data into a Hive table. Using a right file format for Hive table will save a lot of disk space as well as will improve performance of Hive queries.

TEXTFILE

Textfile format stores data as plain text files. Textfile format enables rapid development due to its simplicity but other file formats like ORC are much better when it comes to data size, compression, performance etc. Compressed text files cannot be split for parallel processing

SEQUENCEFILE

Sequencefile is a flat file consisting of binary key/value pairs. Sequencefile provides support for compression at different level(Block/Record). Compressed sequencefiles can be split for parallel processing.

ORC format

ORC (Optimized Row Columnar) file format provides a highly efficient way to store Hive data. Using ORC format  improves performance when reading, writing, and processing data in Hive. We can specify compression to further compress data files. It could result in a small performance loss while writing, but there will be huge performance gain in reading. Compression available are SNAPPY, ZLIB, NONE. The default block size is 256MB.

PARQUET

Parquet is a columnar storage format. Parquet supports very efficient compression and encoding schemes. Using right compression and encoding can have great impact on performance while processing the data stored in parquet files.

AVRO

Avro provides a compact and efficient way to store data in binary format. It uses JSON for storing the data definition.