Other File Types

Big data systems sometimes use other file formats in addition to text, Avro, Parquet, and ORC. Two common options are *SequenceFiles* and*RCFiles*. We generally do not recommend using these file formats, but they are briefly described below so that you will be aware of them.

Sequence Files

The SequenceFile format was developed for big data systems as an alternative to text files. SequenceFiles store key-value pairs in a binary container format. They store data more efficiently than text files, and they can store binary data like images.

However, the SequenceFile format is closely associated with the Java programming language, and it is not widely supported outside the Hadoop ecosystem.

Overall, SequenceFiles offer good performance but poor interoperability.

RCFiles

RCFile, which stands for Record Columnar File, is a columnar file format that was developed for use with Hive. RCFile is also supported by some other tools, including Impala, but this support is limited. The RCFile format stores all data as strings, which is inefficient.

Overall, RCFile offers poor performance and limited interoperability.

The ORC file format (described in the previous reading) is an improved version of RCFile with superior performance.