ORC Files

Apache ORC (Optimized Record Columnar) is a file format originally developed by engineers at Hortonworks and Facebook. (Hortonworks has since merged with Cloudera.)

ORC is very similar to Parquet. ORC and Parquet were designed to meet many of the same needs, and internally they use many of the same techniques to achieve excellent performance.

ORC is often used with Hive. To take advantage of certain features of Hive, you *must* store table data in the ORC file format. (These features are not covered by this course or in the other courses in this specialization.) However, ORC is not widely supported by other tools. Impala cannot query tables whose data is stored in ORC files.

Overall, ORC offers excellent performance but limited interoperability. We recommend choosing ORC when you are using features of Hive that require it. Keep in mind that the Hive tables that use ORC files can not be queried using Impala.