Hive Different File Formats

Different file formats and compression codecs work better for different data sets in Apache Hive.

Following are the Apache Hive different file formats:

Text File

Sequence File

RC File

AVRO File

ORC File

Parquet File

Hive Text File Format:

Hive Text file format is a default storage format.

We can use the text format to interchange the data with other client application.

The text file format is very common most of the applications.

Data is stored in lines, with each line being a record.

Each lines are terminated by a newline character (\n).

The text format is simple plane file format.

We can use the compression (BZIP2) on the text file to reduce the storage spaces.

Create a TEXT file by add storage option as ‘STORED AS TEXTFILE’ at the end of a Hive CREATE TABLE command.

Hive Text File Format Examples

Below is the Hive CREATE TABLE command with storage format specification:

Create table textfile\_table (column\_specs) stored as textfile;

Hive Sequence File Format:

Sequence files are Hadoop flat files which stores values in binary key-value pairs.

The sequence files are in binary format and these files are able to split.

The main advantages of using sequence file is to merge two or more files into one file.

Create a sequence file by add storage option as ‘STORED AS SEQUENCEFILE’ at the end of a Hive CREATE TABLE command.

Hive Sequence File Format Example

Below is the Hive CREATE TABLE command with storage format specification:

Create table sequencefile\_table (column\_specs) stored as sequencefile;

Hive RC File Format:

RCFile is row columnar file format. This is another form of Hive file format which offers high row level compression rates.

If we have requirement to perform multiple rows at a time then you can use RCFile format.

The RCFile are very much similar to the sequence file format. This file format also stores the data as key-value pairs.

Create RCFile by specifying ‘STORED AS RCFILE’ option at the end of a CREATE TABLE Command:

Hive RC File Format Example

Below is the Hive CREATE TABLE command with storage format specification:

Create table RCfile\_table

(column\_specs)

stored as rcfile;

Hive AVRO File Format

AVRO is open source project that provides data serialization and data exchange services for Hadoop.

We can can exchange data between Hadoop ecosystem and program written in any programming languages.

Avro is one of the popular file format in Big Data Hadoop based applications.

Create AVRO file by specifying ‘STORED AS AVRO’ option at the end of a CREATE TABLE Command.

Hive AVRO File Format Example

Below is the Hive CREATE TABLE command with storage format specification:

Create table avro\_table

(column\_specs)

stored as avro;

Hive ORC File Format

The ORC file stands for Optimized Row Columnar file format. The ORC file format provides a highly efficient way to store data in Hive table.

This file system was actually designed to overcome limitations of the other Hive file formats.

The Use of ORC files improves performance when Hive is reading, writing, and processing data from large tables.

Create ORC file by specifying ‘STORED AS ORC’ option at the end of a CREATE TABLE Command.

Hive ORC File Format Examples

Below is the Hive CREATE TABLE command with storage format specification:

Create table orc\_table

(column\_specs)

stored as orc;

Hive Parquet File Format:

Parquet is a column-oriented binary file format. The parquet is highly efficient for the types of large-scale queries.

Parquet is especially good for queries scanning particular columns within a particular table.

The Parquet table uses compression Snappy, gzip; currently Snappy by default.

Create Parquet file by specifying ‘STORED AS PARQUET’ option at the end of a CREATE TABLE Command.

Hive Parquet File Format Example

Below is the Hive CREATE TABLE command with storage format specification:

Create table parquet\_table

(column\_specs)

stored as parquet;