

(<https://www.ellicium.com>)

All You Need To Know About ORC File Structure In Depth

Posted by Rohan Karanjawala on June 19, 2017 in [Blog \(https://www.ellicium.com/category/blog/\)](https://www.ellicium.com/category/blog/)

All You Need To Know About ORC File Structure In Depth

Want to store data in Hive tables, just wondering which file format to use, ORC or Parquet?

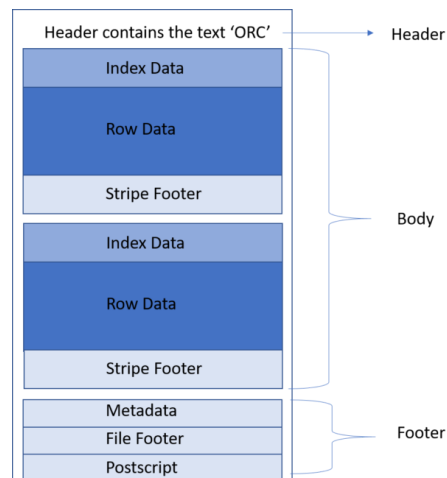
Well this is a question which many have tried to answer in various ways.

Let us understand how is the Optimized Row Columnar (ORC) file format different in comparison to our usual flat file.

ORC is a columnar file format. You can visualize the structure of an ORC file as an area that is divided into Header, body and footer.

Header Section:

The Header contains the text 'ORC' in case some tools require to determine the type of file while processing.



The body contains the actual data as well as the indexes. Actual data is stored in the ORC file in the form of rows of data that are called Stripes. Default stripe size is 250 MB.

Stripes are further divided into three more sections, viz the index section, the actual data and a stripe footer section. One interesting thing to note here is that both index and data section are store as columns so that only the columns where the required data is present, is read. Index data consists of min and max values for each column as well as the row positions within each column. ORC indexes help to locate the stripes based on the data required as well as row groups. The Stripe footer contains the encoding of each column and the directory of the streams as well as their location.

Footer Section:

The footer section consists of three parts viz. file metadata, file footer and postscript.

The file Metadata section contains the various statistical information related to the columns and this information is present at a stripe level. These statistics enable input split elimination based on predicate push down which evaluate for each stripe. The file footer contains information regarding the list of stripes in the file, number of rows per stripe, and the data type for each column. It also contains aggregates counts at column-level like min, max, and sum. The Postscript section contains the file information like the length of the file's Footer and Metadata sections, the version of the file, and the compression parameters like general compression used (eg. none, zlib, or snappy) and the size of the compressed folder.

I'm sure now you would have a much better understanding of the ORC file format structure which would help you make a better decision in selection of the file formats. Of course, now your next step is to compare file formats like ORC and Parquet and get a conclusion which one is better suit to your project.

Confused about where to begin this comparison? Well, in my next blog I would be taking you through on how to compare various file formats, to try and help out those looking forward to doing this activity. Stay Tuned!!!

About Author:

Rohan (<https://www.linkedin.com/in/rohan-karanjawala-2544bb32/>) is senior manager at Ellicium (<https://www.ellicium.com/>) solutions pvt ltd, who looks after projects in Big Data, IOT and Analytics area, helping businesses to stay ahead in the competition.







About Rohan Karanjawala

I work with Ellicium Solutions pvt ltd as an AVP looking after projects in big data analytics area, helping clients to stay ahead in the competition and more importantly to serve their customers well.

Contact us (<https://www.ellicium.com/contact/>) Career (<https://www.ellicium.com/career/>) About us (<https://www.ellicium.com/about/>)

Copyright ©2017 ellicium.com . All Rights Reserved.

 (<https://www.facebook.com/ElliciumSolutionsInc/>)  (https://www.linkedin.com/company/3642801/?trk=tyah&trkInfo=clickedVertical:company,entityType:entityHistoryName,clickedEntityId:company_company_3642801,idx:0)

 (<https://plus.google.com/113844536669345358722>)  (<https://twitter.com/ElliciumSolInc>)