Day 3 – 9/15 Storage formats: AVRO vs Parquet vs RC/ORC

This topic in general asks three questions:

* What are file formats?
* What are the common Hadoop file format features?
* Which format should you be using?

Popular Storage Format:

* Text (csv, tsv, JSON)
* Sequence File
* Avro
* Parquet
* ORC

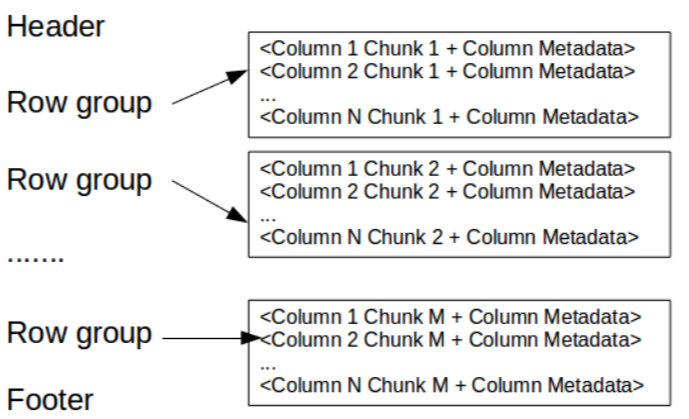
Apache Avro

* A language-neutral data serialization system for Apache Hadoop
* Can be processed by many languages (C, C++, C#, Java, Python, and Ruby)

AVRO File Format

* **Row-based** storage format for Hadoop which is widely used as a serialization platform
* Schema defined with JSON, encoded on the file so the data can be untagged
* Data stored in a binary format
* Support schema evolution
* Ideal for storing data in a data lake landing zone

PARQUET File Format

* Open-source file format for Hadoop which stores nested data structures in a flat **columnar format**
* Ideal for queries that read particular columns from a “wide” table since only needed columns are read and IO is minimized
* Avro files include **markers** that can be used to split large data sets into subsets suitable for Apache MapReduce processing
* Terms:
  + **Row group**: A logical horizontal partitioning of the data into rows. A row group consists of a column chunk for each column in the dataset
  + **Column chunk**: A chunk of the data for a particular column. These column chunks live in a particular row group and are guaranteed to be contiguous in the file
  + **Page**: Column chunks are divided up into pages written back to back. The pages share a common header and readers can skip the page they are not interested in

ORC File Format

* ORC stands for Optimized Row Columnar file format
* Designed to overcome the limitations of other file formats
* Ideal for storing data compact and enables skipping over irrelevant parts without the need for large, complex, or manually maintained indices. The ORC file format addresses all of these issues
* Terms:
  + Stripe footer – contains a directory of stream locations
  + Row data – is used in table scans
  + Index data – include min and max values for each column and the row’s positions within each column. ORC indexes are used only for the selection of stripes and row groups and not for answering queries

Comparisons between different file formats

AVRO vs PARQUET

1. AVRO is row-based storage format whereas PARQUET is a columnar based storage format
2. PARQUET is much better for analytical querying, i.e., reads and querying are much more efficient than writing
3. Write operations in AVRO are better than in PARQUET or ORC
4. AVRO is much matured than PARQUET when it comes to schema evolution. PARQUET only supports schema append whereas AVRO supports a much-featured schema evolution i.e., adding or modifying columns
5. PARQUET is ideal for querying a subset of columns in a multi-column table. AVRO is ideal in case of ETL operations where we need to query all the columns

ORC vs PARQUET

1. PARQUET is more capable of storing nested data
2. ORC is more capable of Predicate Pushdown
3. ORC supports ACID properties
4. ORC is more compression efficient

Resources sited:

Big Data File Formats, https://blog.clairvoyantsoft.com/big-data-file-formats-3fb659903271?gi=3f0ba725ffea

https://www.slideshare.net/mobile/StampedeCon/choosing-an-hdfs-data-storage-format-avro-vs-parquet-and-more-stampedecon-2015