Workshop 6

Part 1

Data at Scale File Formats

General Instructions:

A file format describes the structure of data in a file that applications follow to read or create such files.

File types can be grouped into categories such as text file formats, vector file formats, spreadsheet file formats, video file formats, image file formats, and several others. A file can also be converted into another file format such as Text to Parquet. Finding the right file format for your dataset can be tough. Different applications have also different affinities for file formats.

The purpose of this workshop is to familiarize you with the most popular Data at Scale file formats, **Avro**, **Parquet, and ORC.** It aims to help you to understand each format pros and cons to choose the best one for your use case and optimize storage space and processing time.

Apache spark supports many different data formats like Parquet, JSON, CSV, SQL, NoSQL data sources, and plain text files. Common formats used mainly for big data analysis are Apache Parquet and Apache Avro.

Online resources:

https://avro.apache.org/
https://parquet.apache.org/
https://orc.apache.org/

Tutorials

Learning Activity 1:

In these activities, we will introduce key concepts of File Formats Avro, Parquet and ORC and their usage in Spark.

Choosing The Right Big Data File Format

Avro And Parquet With Spark

Learning Activity 2: Spark and File Formats.

(Refresher) The following case study uses Spark to partition the input data and save it in Parquet/Snappy format. Then the partitions are loaded into Hive / Trino.

H1B Visa Application Analysis