Large Scale Data Management

Avro+Parquet
Lab Guide 3

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/`.

Steps

1. Define a hierarchical schema for data in `title.basics.tsv.gz`.
2. Convert `title.basics.tsv.gz` to an AvroParquet file.
3. Compute the number of films for each genre from data in the AvroParquet file.
4. Optimize the previous operation with a projection schema.
5. Compare performance with the same operation on text files.

Questions

1. What is the impact of a columnar file and its relation to projection operations?

Learning Outcomes Recognize the impact of columnar formats and projections in MapReduce performance. Apply AvroParquet API and formats in Hadoop programs.