Large Scale Data Management

Hive
Lab Guide 8

2020/2021

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

Steps

1. Load *name.basics* as a Parquet file, and then as a Parquet file partitioned by birth year.
2. Using RDD operations compute the total number of actors/actresses and then count those born in the 1960s decade.
3. Using SQL compute the total number of actors/actresses and then count those born in the 1960s decade.
4. Measure the times needed by each of these operations with the two Parquet files.

Questions

1. How do solutions with Spark SQL compare with those using Spark RDDs in terms of efficiency?
2. Can you explain the different results?

Learning Outcomes Compare SQL and Spark dataflow paradigms for distributed data processing. Recognize the impact of metadata and optimization in query efficiency.