Large Scale Data Management

Spark SQL
Lab Guide 7

2020/2021

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

Steps

1. Display on screen the number of films for each year.
2. Compute the best rating in a movie for each year, sorted by year, saving it in a CSV file.
3. Using the previous result and the data, compute the list of movies with the top ratings, saving it as a Parquet file.
4. Measure the time it takes to do the computations for different prefixes of the files.

Questions

1. Can you describe the same operations in terms of RDD?
2. How do solutions with Spark SQL compare with those with Spark RDDs in terms of programming effort?

Learning Outcomes Apply SQL queries to solve data processing tasks on files. Compare SQL and Spark dataflow paradigms for distributed data processing.