Large Scale Data Management

Spark
Lab Guide 4

2020/2021

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

Steps

1. Compute the number of films for each genre from `title.basics.tsv`.
2. Measure the time it takes to do the computations for different prefixes of the files.
3. Compute a list showing the rating for each title using also `title.ratings.tsv`.
4. Compute a list of films with rating of at least 9.0, sorted by rating.

Questions

1. How do solutions with Spark compare with those with MapReduce in terms of programming effort?
2. How do solutions with Spark compare with those with MapReduce in terms of performance?

Learning Outcomes Formulate Spark jobs for simple pipeline queries. Compare Hadoop MapReduce and Spark dataflow paradigms for distributed data processing.