Large Scale Data Management

Spark cluster
Lab Guide 5

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/` and Spark deployment with `docker-spark` available at `https://github.com/big-data-europe/docker-spark`.

Steps

1. Compute the number of movies for each actor and the top 10 actors by number of movies.
2. Compute the top 3 highest ranking movies for each actor.
3. Compute the set of collaborators for each actor (i.e., other actors that have participated in a common movie).

Questions

1. Can you predict execution plans for each of the data processing tasks?
2. What is the impact of caching and number of partitions in performance?

Learning Outcomes Formulate Spark jobs for complex and inter-related queries. Explain how queries map to execution plans. Assess the impact of the number of partitions used in each stage of the plan.