Large Scale Data Management

Lab Guide 0

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/` (and abridged versions *mini* and *micro* available in Blackboard).

Steps

1. Read and parse the files using Java.
2. Compute the top 10 most popular genres from `title.basics.tsv.gz`.
3. Compute the list of titles identifiers for each person, ordered by person identifier, from `title.principals.tsv.gz`.
4. Measure the time it takes to do the computations for different prefixes of the files.
5. Package and deploy the application as a Docker container.

Questions

1. How does your solution scale with the number of lines in the input file?
2. What is the maximum file size that can be handled with a laptop?

Learning Outcomes Construct simple data processing tools for text-based files. Use Docker containers to deploy data processing applications. Recognize the need for distributed data processing and scale-out.