Large Scale Data Management

Spark streaming
Lab Guide 6

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/`, the IMDb stream generator at `https://github.com/jopereira/streamgen` and a local Spark deployment.

Steps

1. Save incoming data in text files every 1 minute.
2. Compute every minute the top 3 highest ranked movie identifiers in the last 10 minutes.
3. Compute every minute the top 3 highest ranked movie titles in the last 10 minutes.

Questions

1. Would it be possible to do something similar with Hadoop MapReduce, i.e., extending it to stream processing?
2. How do windows behave when the system starts and restarts? Consider checkpointing.

Learning Outcomes Compare batch and streaming data processing paradigms. Formulate Spark streaming and hybrid queries. Employ window transformations to maintain state.