Large Scale Data Management

HDFS
Lab Guide 2

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/` and Hadoop deploymnet with `docker-compose` available at `https://github.com/big-data-europe/docker-hadoop`.

Steps

1. Deploy a Hadoop cluster.
2. Load `title.basics.tsv.gz` into HDFS.
3. Compute the number of films for each genre from the file stored in HDFS.
4. Recompress `title.basics.tsv.gz` with *bzip2* and reload it with 16MB blocksize.
5. Repeat step 3.

Questions

1. What is the impact of compression type and block sizes?

Learning Outcomes Recognize the impact of file type and block size in MapReduce performance. Deploy MapReduce jobs in an existing cluster.