Lab 03

NoSQL

Task

Your task is to analyze the Netflix Prize dataset, which you already transformed into a CSV format, by importing the data into a database technology.

1. Choose one database technology among following list (use docker if possible):
    a. Redis
    b. MongoDB
    c. Neo4j
    d. Cassandra
2. Import your processed Netflix Prize data set and movie titles data set (see movie_titles.csv at Kaggle) via Python script.
3. Execute/Compute following queries on the chosen database technology:
    a. Compute the average rating, lowest rating, highest rating and number of ratings per movie
    b. Get the movie with the lowest overall rating. If the movies are tied for rating, take the movie with the lowest movie_id
    c. Get the movie with the highest overall rating. If the movies are tied for rating, take the movie with the lowest movie_id
    d. Get the number of ratings in February 2002
    e. Get the user that creates the lowest average rating and has rated at least 5 times. If two users are tied, use the one with the lower user_id
    f. For the movie “The Spy Who Loved Me”: Get the average rating for each year and month sorted by year and month in ascending order

Output

- A PDF document containing the queries and results for each step
- Furthermore, the document shall contain all setup steps and the (Python) script which was used for importing the data
- Include also a conclusion why certain queries might not be ideal for your chosen database technology

Outcome

- Basic understanding of ingesting and analyzing data in the chosen database technology

Note

- You are allowed to shrink the dataset accordingly if you hit certain performance bottlenecks on your laptop. If so, for subtask **3f** another movie might be freely chosen.