**DS200 – Lab 1**

**Dữ liệu:**

1. **movies.txt**

* **Schema:** MovieID, Title, Genres

1. **ratings\_1.txt, ratings\_2.txt**

* **Schema:** UserID, MovieID, Rating, Timestamp

1. **users.txt**

* **Schema:** UserID, Gender, Age, Occupation, Zip-code

**Bài 1: Tính Điểm Đánh Giá Trung Bình và Tổng Số Lượt Đánh Giá Cho Mỗi Phim**

* **Mục tiêu:**
  + Tính điểm trung bình cho từng phim từ cả 2 file ratings (ratings\_1.txt và ratings\_2.txt).
  + Tính tổng số lượt đánh giá cho mỗi phim.
  + **Output:** MovieTitle AverageRating: xx (TotalRatings: xx)

A screenshot of a report

AI-generated content may be incorrect.

* **Yêu cầu thêm:**
  + Trong quá trình xử lý, tìm ra phim có điểm trung bình cao nhất (**chỉ xét những phim có tối thiểu 5 lượt đánh giá**).
  + Sử dụng biến lớp như maxMovie và maxRating trong reducer và xuất ra kết quả cuối cùng trong phương thức cleanup(), theo định dạng:

*MovieTitle* is the highest rated movie with an average rating of *AverageRating* among movies with at least 5 ratings.

**Bài 2: Phân Tích Đánh Giá Theo Thể Loại**

* **Mục tiêu:**
  + Vì một phim có thể thuộc nhiều thể loại (Genres được phân tách bằng dấu “|”), mapper cần tách riêng từng thể loại của phim đó.
  + Tính điểm trung bình (và tổng số lượt đánh giá nếu cần) cho từng thể loại, dựa trên tất cả các phim thuộc thể loại đó.
* **Output:** Genre: AverageRating (TotalRatings)

A screenshot of a computer

AI-generated content may be incorrect.

**Bài 3: Phân Tích Đánh Giá Theo Giới Tính**

* **Mục tiêu:**
  + Thực hiện join dữ liệu giữa ratings và users (dựa trên UserID) để lấy thông tin giới tính của người đánh giá.
  + Với mỗi phim, tính riêng điểm trung bình từ người dùng nam và nữ.
* **Output:** MovieTitle: Male\_Avg, Female\_Avg

A screen shot of a number

AI-generated content may be incorrect.

**Bài 4 (Tùy Chọn): Phân Tích Đánh Giá Theo Nhóm Tuổi**

* **Mục tiêu:**
  + Phân nhóm người dùng theo độ tuổi (ví dụ: 0-18, 18-35, 35-50, 50+).
  + Với mỗi phim, tính điểm trung bình cho mỗi nhóm tuổi.
* **Output:** MovieTitle: [0-18: AvgRating, 18-35: AvgRating, 35-50: AvgRating, 50+: AvgRating]

A screen shot of a number

AI-generated content may be incorrect.