HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN 1

BÁO CÁO CƠ SỞ DỮ LIỆU PHÂN TÁN

Nhóm lớp: 10

Nhóm: 19

Thành viên	MSV
Trần Quốc An	B22DCCN007
Đào Đức Duy	B22DCCN145
Nguyễn Quang Phú	B22DCCN621

HÀ NỘI, 06/2025

Mục lục

1: Giới thiệu 3
2. Phân tích yêu cầu và dữ liệu đầu vào 4
2.1. Dữ liệu đầu vào 4
3 Triển khai giải quyết vấn đề 5
3.1 Tạo Metadata 5
3.2 Hàm loadratings(ratingstablename, ratingsfilepath, openconnection) 5
3.3 Hàm rangepartition(ratingstablename, numberofpartitions, openconnection) 6
3.4 Hàm roundrobinpartition(ratingstablename, numberofpartitions, openconnection) 7
3.5 Hàm roundrobininsert(ratingstablename, userid, itemid, rating, openconnection) 8
3.6 Hàm rangeinsert(ratingstablename, userid, itemid, rating, openconnection) 8
4 Phân chia công việc 9

1: Giới thiệu

Bài tập lớn này yêu cầu mô phỏng các phương pháp phân mảnh dữ liệu trên một hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở, cụ thể là PostgreSQL. Nhóm đã phát triển một tập các hàm Python để thực hiện các nhiệm vụ: tải dữ liệu đánh giá phim từ tệp ratings.dat vào một bảng quan hệ, phân mảnh bảng này bằng hai phương pháp phân mảnh ngang là phân mảnh theo khoảng (Range Partitioning) và phân mảnh xoay vòng (Round-Robin Partitioning), và cuối cùng là triển khai các hàm để chèn các bản ghi dữ liệu mới vào đúng phân mảnh tương ứng ().

Mục tiêu của báo cáo này là trình bày chi tiết cách tiếp cận và giải quyết vấn đề của nhóm, bao gồm giải thích logic của từng hàm được triển khai, các quyết định thiết kế quan trọng, và cách quản lý metadata để đảm bảo tính nhất quán và hiệu quả.

2. Phân tích yêu cầu và dữ liệu đầu vào

2.1. Dữ liệu đầu vào

- Dữ liệu được sử dụng là tập dữ liệu MovieLens 10M, chứa trong tệp ratings.dat. Mỗi dòng trong tệp biểu diễn một đánh giá phim và có định dạng UserID::MovieID::Rating::Timestamp ().
    - UserID: Kiểu số nguyên.
    - MovieID: Kiểu số nguyên.
    - Rating: Kiểu số thực, trên thang điểm 5 sao, có thể chia nửa sao (ví dụ: 0.5, 1.0,..., 5.0).
    - Timestamp: Kiểu số nguyên, là số giây kể từ nửa đêm UTC ngày 1 tháng 1 năm 1970.
- Theo yêu cầu của đề bài, bảng Ratings trong cơ sở dữ liệu sẽ có lược đồ UserID (int), MovieID (int), Rating (float) (). Do đó, cột Timestamp từ tệp dữ liệu sẽ không được lưu trữ trong bảng Ratings và các bảng phân mảnh của nó.

2.2 Các hàm cần triển khai

- **Hàm loadratings(ratingstablename, ratingsfilepath, openconnection):** Tải dữ liệu từ ratings.dat vào bảng Ratings.
- **Hàm rangepartition(ratingstablename, numberofpartitions, openconnection)**: Tạo N phân mảnh ngang của bảng ratings_table_name dựa trên các khoảng giá trị đồng đều của thuộc tính Rating.
- **Hàm roundrobinpartition(ratingstablename, numberofpartitions, openconnection):** Tạo N phân mảnh ngang của bảng ratings_table_name sử dụng phương pháp xoay vòng.
- **Hàm roundrobininsert(ratingstablename, userid, itemid, rating, openconnection)**: Chèn một bản ghi mới vào bảng ratings_table_name và vào đúng phân mảnh xoay vòng.
- **Hàm rangeinsert(ratingstablename, userid, itemid, rating, openconnection)**: Chèn một bản ghi mới vào bảng ratings_table_name và vào đúng phân mảnh theo khoảng giá trị Rating.

3 Triển khai giải quyết vấn đề

3.1 Tạo Metadata

Trước khi thực hiện các chức năng chính, định nghĩa hàm create_metadata_tables_if_not_exists(cursor) để tạo các bảng metadata nếu chúng chưa tồn tại. Các bảng này dùng để duy trì trạng thái và cấu hình cho các hoạt động phân mảnh và chèn dữ liệu sau này .

- **round_robin_metadata**: Lưu trữ original_table_name, num_partitions (số lượng phân mảnh N), và current_total_rows (tổng số hàng đã được phân phối cho các phân mảnh round-robin).
- **range_config_metadata**: Lưu trữ original_table_name, num_partitions (số lượng phân mảnh N), min_val (giá trị Rating nhỏ nhất, mặc định 0.0), và max_val (giá trị Rating lớn nhất, mặc định 5.0) cho phân mảnh theo khoảng.

3.2 Hàm loadratings(ratingstablename, ratingsfilepath, openconnection)

Mục tiêu: Tải dữ liệu từ tệp ratings.dat vào bảng Ratings trong PostgreSQL với lược đồ UserID (int), MovieID (int), Rating (float) ().

Triển khai:

1. Hàm tạo bảng Ratings nếu chưa tồn tại với 3 cột yêu cầu.
2. Đọc tệp ratings.dat từng dòng. Với mỗi dòng, tách các trường UserID, MovieID, Rating, Timestamp bằng dấu phân cách ::.
3. Chỉ lấy 3 giá trị UserID, MovieID, Rating để ghi vào một đối tượng io.StringIO. Dữ liệu được định dạng tab-separated để chuẩn bị cho lệnh COPY.
4. Sử dụng phương thức cursor.copy_expert() với lệnh COPY Ratings (UserID, MovieID, Rating) FROM STDIN WITH DELIMITER AS E'\\t' để tải dữ liệu hàng loạt từ io.StringIO vào bảng Ratings. Đây là phương pháp hiệu quả nhất để tải lượng lớn dữ liệu vào PostgreSQL.
5. Thực hiện connection.commit() để lưu các thay đổi.

3.3 Hàm rangepartition(ratingstablename, numberofpartitions, openconnection)

Mục tiêu: Tạo N phân mảnh ngang của bảng ratings_table_name (mặc định là Ratings) dựa trên N khoảng giá trị đồng đều của thuộc tính Rating (từ 0.0 đến 5.0) ().

Triển khai:

1. Kiểm tra N > 0.
2. Tính toán kích thước của mỗi khoảng (step = (5.0 - 0.0) / N).
3. Lặp i từ 0 đến N-1 để tạo từng bảng phân mảnh:
    - Tên bảng phân mảnh: range_part{i}.
    - Tạo bảng range_part{i} với lược đồ UserID (int), MovieID (int), Rating (float) nếu chưa tồn tại.
    - Xóa toàn bộ dữ liệu cũ trong bảng phân mảnh này bằng TRUNCATE TABLE để đảm bảo tính đúng đắn khi chạy lại.
    - Xác định cận dưới (current_lower_bound) và cận trên (current_upper_bound) cho phân mảnh hiện tại. Đảm bảo phân mảnh cuối cùng (i == N-1) bao gồm chính xác giá trị max_rating_val (5.0).
    - Xây dựng điều kiện WHERE cho thuộc tính Rating:
        - Nếu N=1: Rating >= min_rating_val AND Rating <= max_rating_val.
        - Nếu i=0 (phân mảnh đầu tiên và N > 1): Rating >= current_lower_bound AND Rating <= current_upper_bound.
        - Nếu i > 0 (các phân mảnh tiếp theo): Rating > current_lower_bound AND Rating <= current_upper_bound. Logic này tuân thủ chính xác các ví dụ trong đề bài ().
    - Sử dụng câu lệnh INSERT INTO range_part{i} (UserID, MovieID, Rating) SELECT UserID, MovieID, Rating FROM {ratings_table_name} WHERE {condition} để điền dữ liệu vào bảng phân mảnh.
4. Lưu thông tin cấu hình (ratings_table_name, N, min_val=0.0, max_val=5.0) vào bảng range_config_metadata.
5. Thực hiện connection.commit().

3.4 Hàm roundrobinpartition(ratingstablename, numberofpartitions, openconnection)

Mục tiêu: Tạo N phân mảnh ngang của bảng ratingstablename sử dụng phương pháp xoay vòng .

Triển khai:

1. Kiểm tra N > 0.
2. Tạo N danh sách rỗng (partition_tables_data) để lưu trữ tạm thời các hàng cho từng phân mảnh.
3. Lặp i từ 0 đến N-1 để tạo từng bảng phân mảnh:
    - Tên bảng phân mảnh: roundrobin_part{i}.
    - Tạo bảng roundrobin_part{i} với lược đồ UserID (int), MovieID (int), Rating (float) nếu chưa tồn tại.
    - Xóa toàn bộ dữ liệu cũ trong bảng phân mảnh này bằng TRUNCATE TABLE.
4. Truy vấn tất cả các hàng (UserID, MovieID, Rating) từ bảng ratings_table_name.
5. Lặp qua từng hàng đã truy vấn, sử dụng target_partition_index = row_index % N để xác định phân mảnh đích và thêm hàng đó vào danh sách tương ứng trong partition_tables_data.
6. Sau khi phân phối tất cả các hàng, lặp qua N danh sách trong partition_tables_data. Nếu một danh sách có dữ liệu, sử dụng psycopg2.extras.execute_values() để chèn hàng loạt tất cả các hàng trong danh sách đó vào bảng phân mảnh rrobin_part{i} tương ứng. Đây là một cách hiệu quả để thực hiện nhiều lệnh INSERT.
7. Lưu thông tin (ratingstablename, numberofpartitions, total_rows_distributed) vào bảng round_robin_metadata. total_rows_distributed là tổng số hàng đã được đọc từ bảng gốc.
8. Thực hiện connection.commit().

3.5 Hàm roundrobininsert(ratingstablename, userid, itemid, rating, openconnection)

Mục tiêu: Chèn một bản ghi mới (UserID, ItemID, Rating) vào bảng ratings_table_name và vào đúng phân mảnh được tạo theo phương pháp xoay vòng ().

Triển khai:

1. Tham số ItemID được coi là MovieID cho việc lưu trữ.
2. Chèn bản ghi mới vào bảng ratings_table_name (bảng gốc).
3. Truy vấn bảng round_robin_metadata để lấy num_partitions (N) và current_total_rows cho ratings_table_name.
4. Nếu không tìm thấy metadata (ví dụ: RoundRobin_Partition chưa được chạy), báo lỗi và rollback.
5. Tính toán chỉ số phân mảnh đích: target_partition_index = int(current_total_rows % num_partitions).
6. Tên bảng phân mảnh đích: rrobin_part{target_partition_index}.
7. Chèn bản ghi mới vào bảng phân mảnh đích.
8. Cập nhật current_total_rows trong bảng round_robin_metadata bằng cách tăng giá trị lên 1.
9. Thực hiện connection.commit().

3.6 Hàm rangeinsert(ratingstablename, userid, itemid, rating, openconnection)

Mục tiêu: Chèn một bản ghi mới (UserID, ItemID, Rating) vào bảng ratings_table_name và vào đúng phân mảnh được tạo theo phương pháp khoảng giá trị, dựa trên giá trị Rating của bản ghi mới ().

Triển khai:

1. Tham số ItemID được coi là MovieID.
2. Chèn bản ghi mới vào bảng ratings_table_name (bảng gốc).

3. Truy vấn bảng range_config_metadata để lấy num_partitions (N), min_val, max_val cho ratings_table_name.
4. Nếu không tìm thấy metadata, báo lỗi và rollback.
5. Kiểm tra xem giá trị Rating của bản ghi mới có nằm trong khoảng [min_val, max_val] đã cấu hình không. Nếu không, chỉ commit việc chèn vào bảng gốc và không chèn vào phân mảnh nào (theo cảnh báo trong code).
6. Nếu Rating hợp lệ, lặp từ i = 0 đến N-1 để tìm phân mảnh đích:
    - Tính toán current_lower_bound và current_upper_bound cho phân mảnh i tương tự như trong Range_Partition.
    - Áp dụng điều kiện bao gồm/loại trừ cận (giống như trong Range_Partition) để xác định xem Rating có thuộc phân mảnh này không.
    - Nếu tìm thấy, xác định target_partition_name = f"range_part{i}" và thoát khỏi vòng lặp.
7. Nếu target_partition_name được tìm thấy, chèn bản ghi mới vào bảng phân mảnh đó.
8. Thực hiện connection.commit().

4 Phân chia công việc