BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

MÔN HỌC: CẤU TRÚC DỮ LIỆU VÀ GIẢI THUẬT

SẮP XẾP DỮ LIỆU LỚN

Sinh viên thực hiện:

21120439 – Bùi Minh Duy

21120447 – Nguyễn Nhật Hào

21120453 – Tô Phương Hiếu

Giảng viên hướng dẫn:

Thầy Nguyễn Thái Vũ

Thành phố Hồ Chí Minh, Năm 2022

MỤC LỤC

I. Giới thiệu:

Bài toán: Sắp xếp dữ liệu lớn.

Mô tả: File dữ liệu có kích thước ~ 3GB có định dạng (.csv) bao gồm 3.000.000 dòng dữ liệu.

Mục tiêu: Sắp xếp các dòng dữ liệu theo giá trị "Id" với thứ tự từ bé đến lớn.

II. Thông tin nhóm:

1. Thành viên:

MSSV	Họ và Tên	Đóng góp
21120439	Bùi Minh Duy	100%
21120447	Nguyễn Nhật Hào	100%
21120453	Tô Phương Hiếu	100%

2. Mức độ hoàn thành: 100%

- Thiết kế kiến trúc.

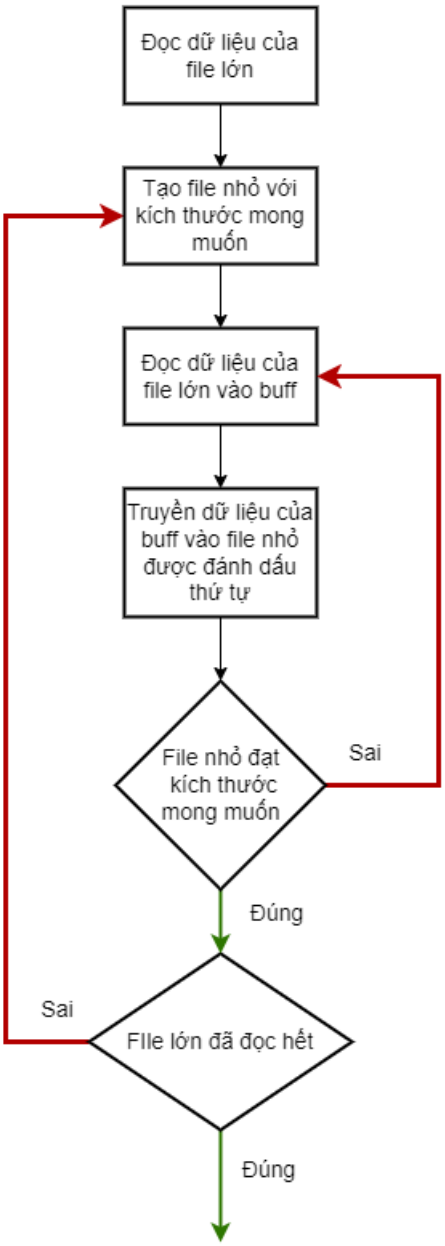
- Áp dụng xử lý bài toán.

- Đánh giá độ phức tạp, tối ưu thuật toán.

III. Kiến trúc và thuật toán:

Bài toán được triển khai qua ba bước chính theo thứ tự sau:

1) Chia File lớn:



- Ý tưởng:

+ Xác định trước kích thước của 1 file nhỏ muốn chia.

+ Dựa vào kích thước đã xác định K ~ 20,000KB, chia thành 140 file nhỏ.

+ Đọc dữ liệu của File lớn vào một bộ nhớ đệm buff (fit trên RAM).

+ Lưu dữ liệu trong bộ nhớ đệm vào file nhỏ với tên X.csv (X là số thứ tự file), lặp lại cho đến khi file nhỏ đạt đến kích thước muốn chia.

+ Khi file nhỏ đạt kích thước mong muốn, kiểm tra ký tự cuối cùng của file nhỏ có phải ký tự xuống dòng không, nếu không thì tiếp tục lưu dữ liệu từ file lớn cho đến khi gặp ký tự xuống dòng để có dữ liệu hoàn chỉnh.

+ Tạo file nhỏ mới lưu với tên X.csv (X là số thứ tự file) và lặp lại quá trình cho đến khi đọc hết dữ liệu từ file lớn.

- Đánh giá độ phức tạp:

Với K là độ lớn của file nhỏ được xác định trước, L là độ lớn của file lớn, buff là độ lớn của bộ nhớ đệm.

Thuật toán cần tạo ra n = L/K file nhỏ, sử dụng vòng lặp for duyệt n file nhỏ, đồng thời phải sử dụng vòng lặp while duyệt m = K/buff lần để lưu dữ liệu từ bộ nhớ đệm vào file nhỏ. Tức là cần O(n.m)

2) Sắp xếp File nhỏ:

- Ý tưởng :

+ Lần lượt xử lí từng file nhỏ. Đọc dữ liệu của từng file nhỏ.

+ Lần lượt chuyển dữ liệu của file nhỏ vào một bộ nhớ đệm buff có kích thước định trước (fit trên RAM).

+ Phân tích dữ liệu từng dòng thành dạng struct với 2 trường id và context để dễ sắp xếp.

+ Sắp xếp mảng struct theo id tăng dần với dữ liệu đã được phân tích trước đó bằng thuật toán Quick Sort.

+ Lưu mảng struct đã sắp xếp vào file với đánh dấu đã sắp xếp (X_sorted.csv, với X là số thứ tự của file nhỏ).

+ Lặp lại toàn bộ quá trình cho đến khi đã sắp xếp tất cả file nhỏ.

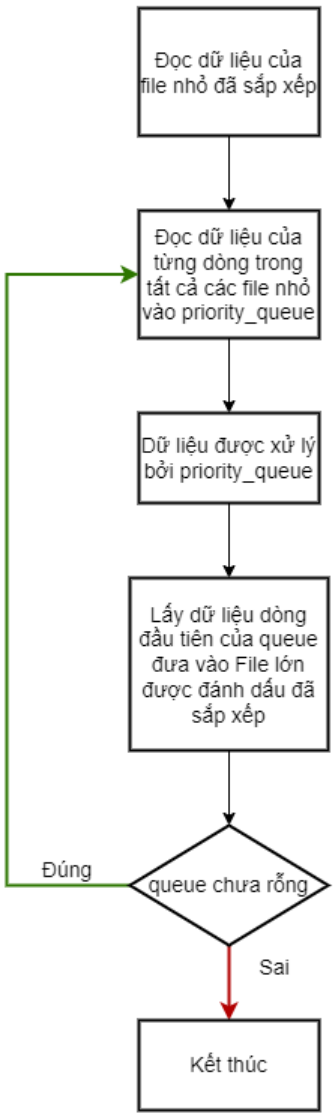
- Đánh giá độ phức tạp:

Với K là độ lớn của file nhỏ, L là độ lớn của file lớn, buff là độ lớn của bộ nhớ đệm, a là số phần tử của mảng struct tương đương với số dòng của file nhỏ).

Thuật toán sử dụng vòng lặp for duyệt n = L/K file nhỏ. Sử dụng vòng lặp while duyệt m = K/buff để phân tích dữ liệu từ file nhỏ vào bộ nhớ đệm rồi vào mảng struct. Sau đó sử dụng thuật toán Quick Sort để sắp xếp mảng dữ liệu với độ phức tạp O(a*log(a)). Cuối cùng là lưu dữ liệu vào file được đánh dấu đã sắp xếp với độ phức tạp O(a).

Vậy độ phức tạp của thuật toán là O(n*(m+a*log(a)+a))

3) Hợp nhất các File nhỏ thành File lớn đã sắp xếp:



- Ý tưởng:

+ Đọc một số lượng dòng nhất định trong tất cả các file nhỏ (giá trị CHUNK_SIZE = 100) để tránh không fit trên Ram.

+ Đưa dữ liệu của từng file nhỏ vào một con trỏ tương ứng, quản lý các con trỏ bằng mảng vector.

+ Đọc dữ liệu từng dòng (số dòng là CHUNK_SIZE) từ con trỏ trong mảng vector, sử dụng kĩ thuật hàng đợi (queue) với dạng priority queue – hàng đợi ưu tiên để sắp xếp các dữ liệu từ các file nhỏ, thuật toán sắp xếp được thư viện hỗ trợ là min-heap, mỗi lần chỉ đưa dữ liệu đầu tiên hiện hành của từng file vào queue.

+ Lấy phần tử đứng đầu trong hàng đợi và đưa vào File lớn (File này sẽ là file kết quả của bài toán) .

+ Kiểm tra nếu trong file nhỏ của phần tử vừa ghi vào File lớn vẫn còn phần kế tiếp thì tiếp tục đưa dữ liệu đó vào queue.

+ Tại một file nhỏ khi đã ghi CHUNK_SIZE dữ liệu vào file lớn, ta tiếp tục đọc CHUNK_SIZE dữ liệu tiếp theo vào con trỏ, việc này dừng lại khi đã lấy hết dữ liệu của file nhỏ.

+ Lặp lại toàn quá trình cho đến khi queue rỗng.

+ Hoàn thành sắp xếp File lớn merge dữ liệu và lưu dưới tên "sorted_big_file.csv"

- Đánh giá độ phức tạp:

Với m = CHUNK_SIZE là số dòng tối đã được đọc trong 1 file nhỏ, n là số file nhỏ n = L/K (k là độ lớn của file nhỏ, L là độ lớn của file lớn), N là số dòng của File lớn.

Thuật toán sử dụng vòng lặp for đọc n file dữ liệu và đưa vào queue, thao tác chèn m dòng dữ liệu vào queue có chi phí O(log(m)) nên vòng lặp có chi phí O(nlog(m)).

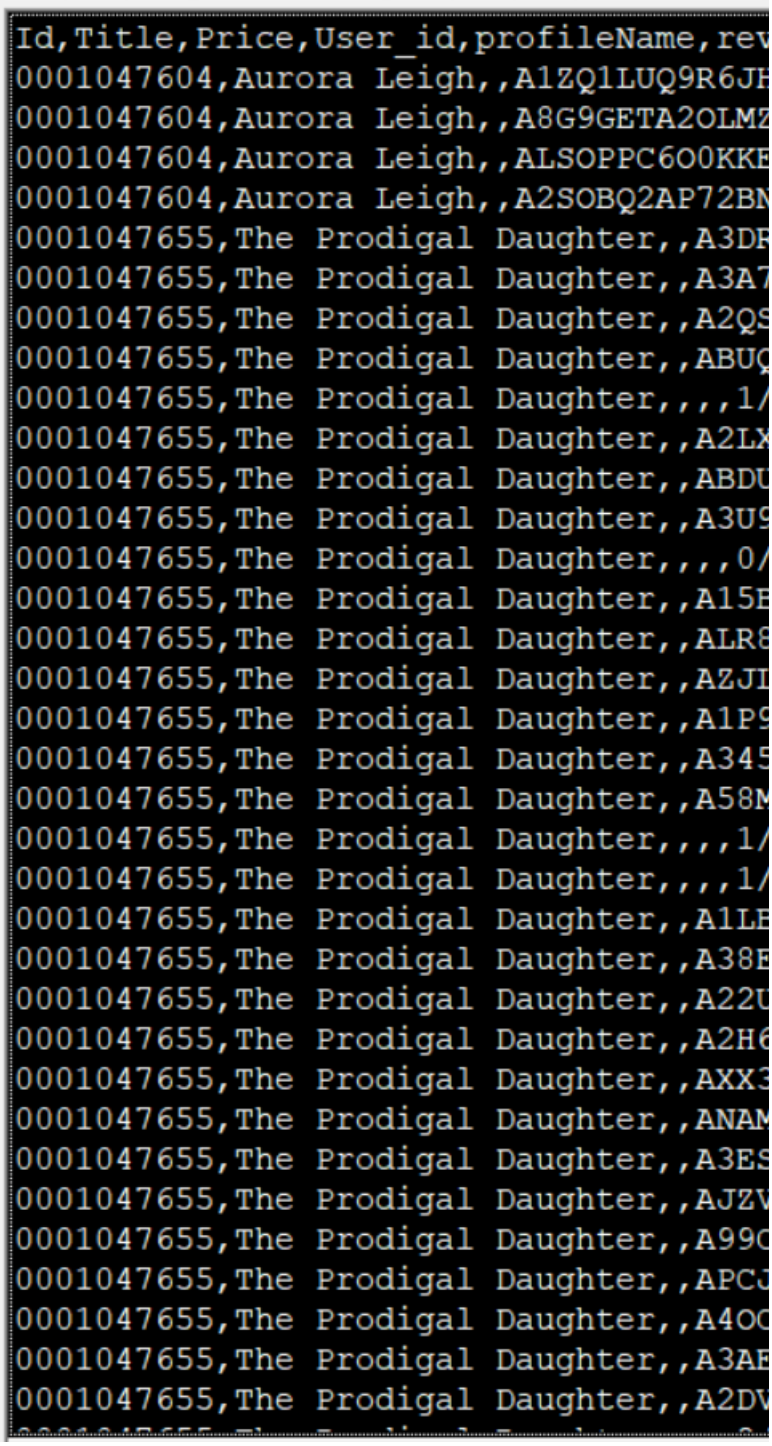
Thuật toán sử dụng vòng lặp while để đưa dữ liệu vào File lớn đã sắp xếp. Về cơ bản, vòng lặp sẽ thực hiện số lần tương đương với số dòng của File lớn và thêm chi phí của việc chèn m dòng dữ liệu vào queue có chi phí O(log(m)). Nên vòng lặp có chi phí O(N + N/m*log(m)).
Vậy độ phức tạp: O(N+ N/m*log(m) + nlog(m))

Kết quả

- Trước khi sắp xếp:

```
Id,Title,Price,User_id,profileName,review
1882931173,Its Only Art If Its Well Hung!
0826414346,Dr. Seuss: American Icon,,A30T
0826414346,Dr. Seuss: American Icon,,A3UH
0826414346,Dr. Seuss: American Icon,,A2MV
0826414346,Dr. Seuss: American Icon,,A22X
0826414346,Dr. Seuss: American Icon,,A2F6
0826414346,Dr. Seuss: American Icon,,A14O
0826414346,Dr. Seuss: American Icon,,A2RS
0826414346,Dr. Seuss: American Icon,,A25M
0826414346,Dr. Seuss: American Icon,,A3VA
0829814000,Wonderful Worship in Smaller C
0829814000,Wonderful Worship in Smaller C
0829814000,Wonderful Worship in Smaller C
0829814000,Wonderful Worship in Smaller C
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
0595344550,Whispers of the Wicked Saints,
```

- Sau khi sắp xếp:



IV. Kết luận:

Đồ án giúp rèn luyện khả năng lập trình. Cung cấp các kiến thức mới và cách vận dụng các kiến thức đã học vào bài toán thực tế. Đồng thời cũng rèn luyện kĩ năng làm việc nhóm thông qua công cụ quản lý source code github và git: https://github.com/nxhawk/Sort-Big-File.git

TÀI LIỆU THAM KHẢO

[1] “priority_queue,” [Trực tuyến]. Available: https://cplusplus.com/reference/queue/priority_queue/.

[2] “ifstream,” [Trực tuyến]. Available: https://cplusplus.com/reference/fstream/ifstream/.

[3] “Tài liệu lý thuyết của môn học,”