**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

# 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:

## 1. Kiến trúc:

## 2. Thuật toán

### a) Chia File lớn:

- Ý tưởng thuật toán:

+ Xác định trước kích thước của 1 file nhỏ.

+ Đọc dữ liệu của File lớn vào một bộ nhớ đệm.

+ Lưu dữ liệu trong bộ nhớ đệm vào file nhỏ, lặp lại cho đến khi file nhỏ đầy.

+ Khi file nhỏ đầy, 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 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 để lưu dữ liệu từ bộ nhớ đệm vào file nhỏ. Tức là cần O(n.m)

### b) Sắp xếp File nhỏ:

- Ý tưởng thuật toán:

+ Lần lượt 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.

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

+ Sắp xếp mảng struct với dữ liệu đã được phân tích bằng thuật toán Quick Sort.

+ Lưu mảng struct đã sắp xếp vào file nhỏ với với đánh dấu đã sắp xếp.

- 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))

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

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

- Ý tưởng thuật toán:

+ Đọc một số lượng dòng nhất định trong tất cả các file nhỏ (giá trị CHUNK\_SIZE)

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

+ Đọc dữ liệu các 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.

+ Lấy phần tử đứng đầu trong hàng đợi và đưa vào File lớn được đánh dấu đã sắp xếp.

+ Kiểm tra nếu trong file nhỏ của phần tử vừa ghi vào File lớn vẫn còn phần kết tiếp thì tiếp tục vào queue, nếu đã đọc hết file nhỏ đó thì bỏ qua.

+ Lặp lại cho đến khi queue rỗng.

+ Hoàn thành sắp xếp File lớn hoàn chỉnh.

- 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)

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 chi phí tổng cộng xấp xỉ O(N) với N là số dòng của File lớn.