Thành viên:

- Trần Quang Ngọc Huỳnh - N19DCCN075
- Đinh Trường Sơn - N19DCCN159
- Lê Thành Trung - N19DCCN214

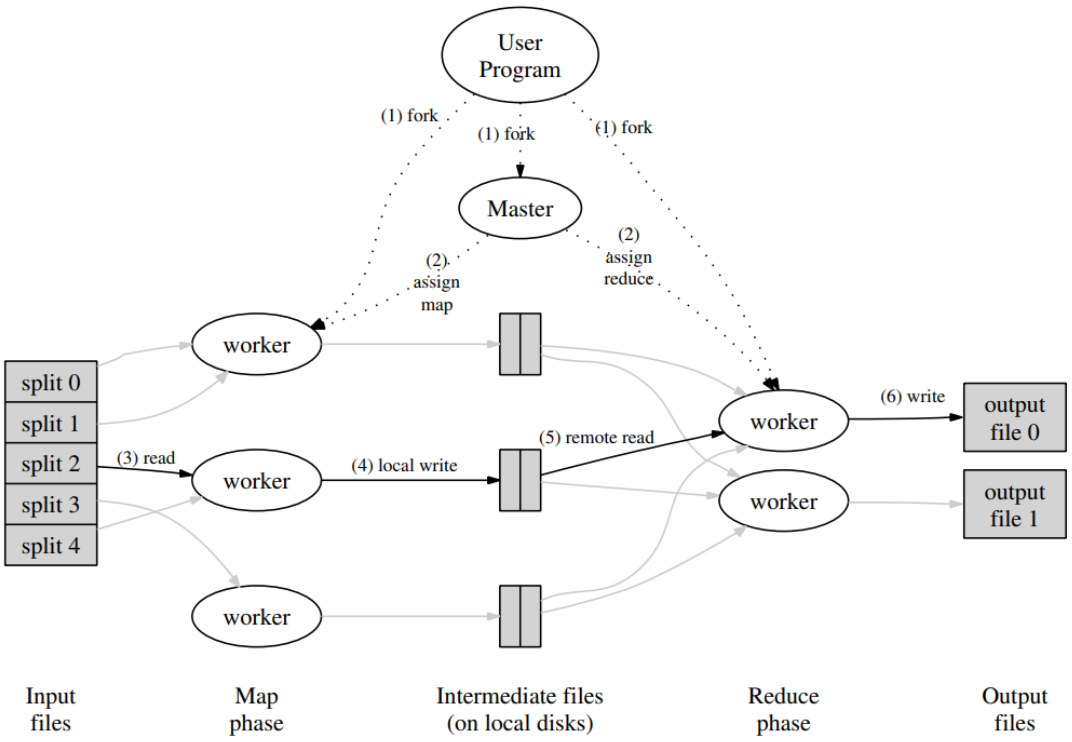
CHUYÊN ĐỀ 4: Nghiên cứu xây dựng chỉ mục phân tán sử dụng MapReduce và ví dụ minh họa

1. **Lý thuyết**

   MapReduce là mô hình được thiết kế độc quyền bởi Google, nó có khả năng lập trình xử lý các tập dữ liệu lớn song song và phân tán thuật toán trên 1 cụm máy tính.

   Quá trình tính toán lấy một tập hợp các cặp key/value đầu vào và tạo ra một tập hợp các cặp key/value trị đầu ra. MapReduce thể hiện tính toán dưới dạng hai hàm: Map và Reduce. Map, được viết bởi user, lấy một cặp đầu vào và tạo ra một tập hợp các cặp key/value trung gian. Thư viện MapReduce nhóm tất cả các giá trị trung gian được liên kết với cùng một khóa trung gian I và chuyển chúng đến hàm Reduce. Hàm Reduce, cũng do người dùng viết, nhận một khóa trung gian I và một tập hợp các value cho key đó. Nó kết hợp các value này lại với nhau để tạo thành một tập hợp các value có thể nhỏ hơn. Thông thường, chỉ có 0 hoặc 1 value đầu ra được tạo cho mỗi lệnh gọi Reduce. Các giá trị trung gian được cung cấp cho hàm Reduce của người dùng thông qua vòng lặp. Điều này cho phép xử lý danh sách các value quá lớn để vừa với bộ nhớ.

2. **Cách thực thi**



- Theo sơ đồ phía trên thì người dùng sẽ thực hiện nhập dữ liệu vào, các dữ liệu đều sẽ được chia nhỏ từ 16MB đến 64MB. Ngay sau đó, thì hệ thống sẽ thực hiện khởi động việc sao chép trên các clusters.
- Hầu hết các máy đều có thể thực hiện xử lý các dữ liệu bao gồm như: master và worker. Trong số đó, máy master có nhiệm vụ điều phối cho những hoạt động bên trong quá trình thực hiện. Các máy worker sau khi đã nhận được dữ liệu thì sẽ tiến hành những nhiệm vụ Map và Reduce. Khi worker đã làm việc xong thì các kết quả đầu ra sẽ xuất hiện các cặp (key và value, các khóa và giá trị) trung gian, những cặp này sẽ được lưu tạm vào bộ nhớ đệm của máy bên trong hệ thống.
- Nếu như Map đã thành công, thì các worker sẽ thực hiện nhiệm vụ tiếp theo là thực hiện phân chia máy trung gian thành những vùng khác nhau. Sau đó, lưu chúng xuống đĩa rồi thông báo kết quả ngược lại cũng như vị trí lưu trữ cho máy master biết.
- Khi đã nhận được thông tin từ worker thì các máy master có thể gán các giá trị trung và vị trí của tệp dữ liệu đó cho máy thực hiện công việc Reduce. Hầu hết,

các máy sẽ được nhận nhiệm vụ xử lý các hàm Reduce rồi xử lý các key, giá trị để có thể đưa ra kết quả cuối cùng.

- Khi quá trình MapReduce đã được hoàn tất thì các máy master đều sẽ được kích hoạt chức năng thông báo cho lập trình viên biết. Khi kết quả đầu ra đã được lưu trữ trên hệ thống thì người dùng có thể dễ dàng sử dụng chúng cũng như quản lý và sao lưu dễ dàng hơn.

★ **Cách thực thi xây dựng chỉ mục phân tán sử dụng MapReduce:**

Hàm map sẽ phân tích mỗi tài liệu và phát đi tập danh sách thẻ định vị và sắp xếp và chia danh sách thẻ định vị thành 3 khoảng theo thứ tự chữ cái lần lượt là a-f, g-p và q-z. Hàm reduce truy xuất các thẻ định vị của từ đang tìm kiếm, sau đó kết hợp các danh sách thẻ định vị của từ này lại thành dạng <từ, danh sách id các tài liệu chứa từ này>. Kết hợp các kết quả trả về từ các worker hoạt động trong bước reduce, ta sẽ thu được bộ chỉ mục ngược cuối cùng.

3. Ví dụ minh họa