BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

MÔN HỌC: CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM

BÁO CÁO CHUYÊN ĐỀ

CHUYÊN ĐỀ 4: NGHIÊN CỨU CÁCH XÂY DỰNG CHỈ MỤC PHÂN TÁN SỬ DỤNG MAPREDUCE

GVHD: TS.Nguyễn Thị Tuyết Hải

SVTH	MSSV
1. Nguyễn Viết Tín	N19DCCN171
2. Huỳnh Tuấn Kiệt	N19DCCN079
3. Lê Lâm Tuấn	N19DCCN177

Tp. Hồ Chí Minh, ngày 31 tháng 3 năm 2023

Mục Lục

1. Giới thiệu

a. Chỉ mục phân tán

Các bộ sưu tập thường lớn đến mức chúng ta không thể thực hiện việc xây dựng chỉ mục một cách hiệu quả trên một máy đơn lẻ. Điều này đặc biệt đúng với World Wide Web mà chúng ta cần các cụm máy tính lớn để xây dựng bất kỳ chỉ mục web có kích thước hợp lý nào. Do đó, các công cụ tìm kiếm web sử dụng các thuật toán lập chỉ mục phân tán để xây dựng chỉ mục. Kết quả của quá trình xây dựng là một chỉ mục phân tán được phân vùng trên một số máy - theo thuật ngữ hoặc theo tài liệu. Hầu hết các công cụ tìm kiếm lớn thích chỉ mục được phân vùng tài liệu (có thể dễ dàng tạo từ chỉ mục được phân vùng theo thuật ngữ). [4]

b. MapReduce

MapReduce là một mô hình lập trình và triển khai liên quan để xử lý và tạo các bộ dữ liệu. Người dùng chỉ định một chức năng bản đồ xử lý một cặp khóa/giá trị để tạo một bộ khóa/giá trị trung gian các cặp và một chức năng rút gọn hợp nhất tất cả các trung gian các giá trị được liên kết với cùng một khóa trung gian. Nhiều các nhiệm vụ trong thế giới thực có thể diễn đạt được trong mô hình này. [2]
MapReduce là một mô hình được Google thiết kế độc quyền với khả năng lập trình xử lý một lượng lớn các dữ liệu song song đồng thời phân tán các thuật toán trên cùng một máy tính. Mặc dù MapReduce ban đầu là một công nghệ độc quyền của Google nhưng trong thời gian gần đây, MapReduce đang dần trở thành một trong những thuật ngữ tổng quát hoá. [5]

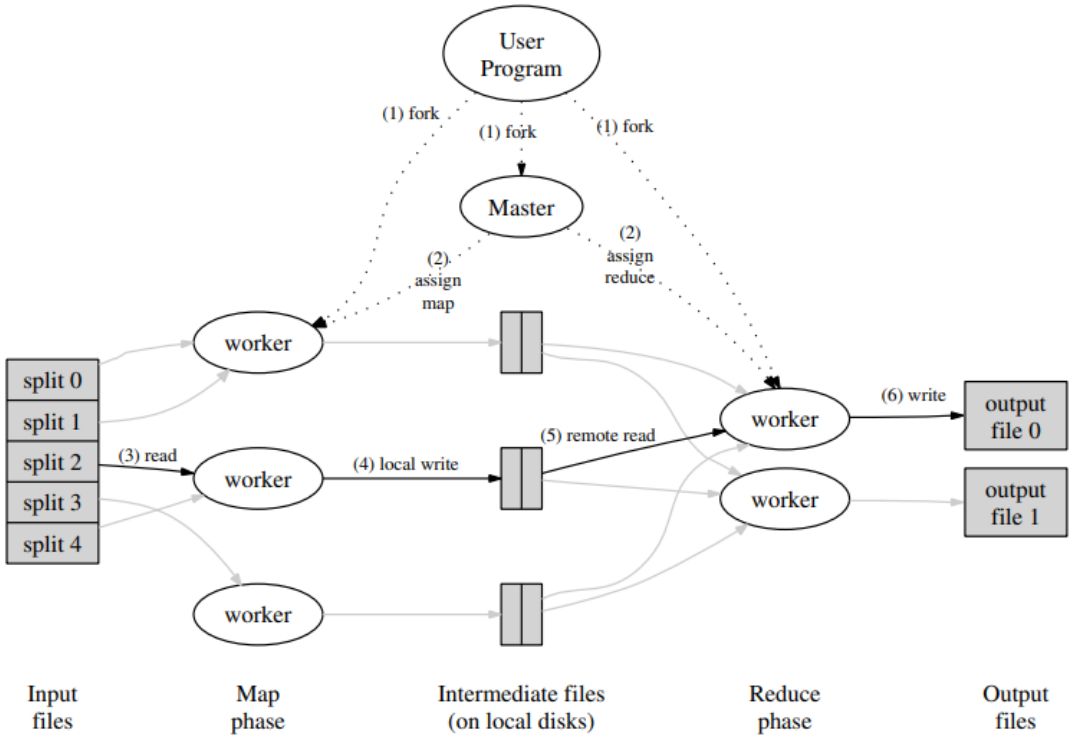
Ưu điểm của MapReduce:

- MapReduce có thể dễ dàng xử lý tất cả mọi bài toán có lượng dữ liệu khổng lồ nhờ khả năng tính toán và tác vụ phân tích phức tạp. Chỉ trong một khoảng thời gian ngắn, nó có thể nhanh chóng xử lý và đưa ra kết quả dễ dàng.
- MapReduce có khả năng chạy song song trên các máy tính có sự phân tán khác nhau với khả năng hoạt động độc lập kết hợp với việc phân tán và xử lý các lỗi kỹ thuật để mang đến hiệu quả cao cho toàn bộ hệ thống.
- MapReduce có khả năng thực hiện được trên đa dạng nhiều loại ngôn ngữ lập trình khác nhau như ngôn ngữ C/C++, Java, Perl, Python, Ruby,... cùng với những thư viện hỗ trợ tương ứng.
- Mã độc trên Internet ngày càng nhiều khiến cho việc xử lý các đoạn mã độc này trở nên phức tạp và tiêu tốn nhiều thời gian hơn. Do đó, MapReduce đang dần hướng quan tâm nhiều hơn cho việc phát hiện các mã độc để có thể nhanh chóng xử lý các đoạn mã độc đó. Nhờ đó, hệ điều hành được đảm bảo vận hành trơn tru với tính bảo mật cao nhất.[5]

2. Tổng quan về MapReduce trong mô hình chỉ mục phân tán

a. Mô hình của MapReduce

Trong bối cảnh lập chỉ mục phân tán, MapReduce là một khung phổ biến để thực hiện quy trình lập chỉ mục. Dưới đây là tổng quan về cách MapReduce hoạt động trong mô hình chỉ mục phân tán:



Hình 1.1: Tổng quan về MapReduce [2]

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. Đây cũng chính là mô hình chia để trị.[3]

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.[3]

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.[3]

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.[3]

b. Giai đoạn Map

Trong giai đoạn Map, mỗi nút (còn được gọi là trình ánh xạ) trong cụm xử lý một tập hợp con các tài liệu hoặc thuật ngữ và tạo một phần chỉ mục cho chúng. Trình ánh xạ đọc dữ liệu từ bộ nhớ cục bộ của nó và thực hiện thao tác lập chỉ mục cục bộ. Các kết quả trung gian do mỗi trình ánh xạ tạo ra sau đó được lưu trữ trong một hệ thống tệp phân tán, chẳng hạn như Hệ thống tệp phân tán Hadoop (HDFS).
Giai đoạn map sẽ chia input data thành các cặp key-value, thuật ngữ gọi là tokenize:

VD: {key = “term”, value = (“docId”, count/position)}

hoặc: {key = “term”, value = “docId”}

Ta có key ở đây chính là từ trong tập tài liệu, value của dictionary này là một tuple chứa ID của tài liệu hoặc vị trí, tùy mục đích sử dụng mà các token được lưu theo nhiều cách khác nhau, sau đó các máy sẽ lưu chúng dưới ổ đĩa và thông báo để thực hiện bước tiếp theo

c. Giai đoạn Shuffle

Thông thường, giai đoạn này sẽ được gộp chung với bước Map, tuy nhiên để bài tìm hiểu mang tính cụ thể, bước này cũng được tách riêng để mang lại cái nhìn tổng quan.

Nhiệm vụ chính của Shuffle là thu thập key/value từ Map và gom chúng thành một nhóm, nhóm đó có thể gom theo nhiều cách thức khác nhau, thông thường, chúng sẽ được gom they key, tức là các từ.[3]

Trong giai đoạn này, các kết quả trung gian do người lập map tạo ra được chuyển đến các node thích hợp. Điều này được thực hiện bằng cách phân vùng các kết quả trung gian dựa trên khóa, thường là thuật ngữ hoặc số nhận dạng tài liệu. Dữ liệu được phân vùng sau đó được gửi đến các node, các node này sẽ xử lý dữ liệu đó trong giai đoạn tiếp theo. Cuối cùng, máy worker sẽ lưu những dữ liệu đã được gom nhóm và lưu chúng vào bộ nhớ cục bộ để chuẩn bị cho giai đoạn Reduce

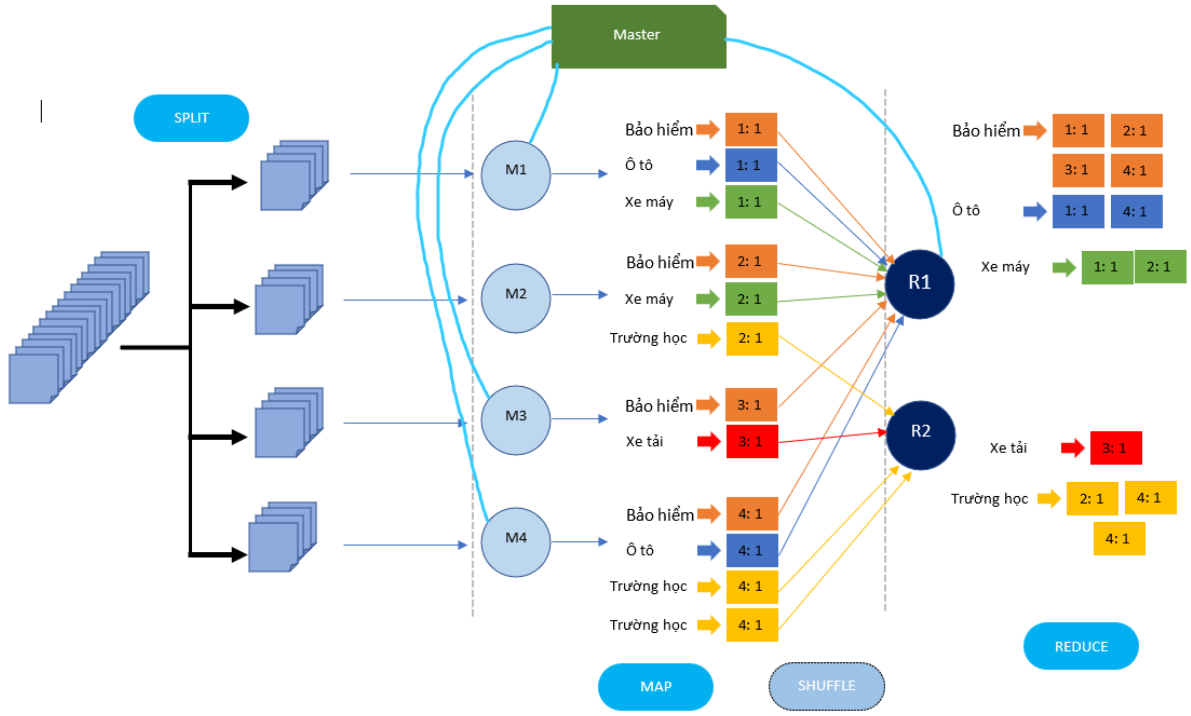
d. Giai đoạn Reduce

Trong giai đoạn Reduce, mỗi nút (còn được gọi là các Reducers) xử lý các kết quả trung gian cho một tập hợp con các thuật ngữ hoặc tài liệu và tạo chỉ mục một phần cho chúng. Để thuận tiện cho bước reduce, ta có thể nhận group các key/value từ bước Shuffle hoặc tiến hành sắp xếp chúng, Reducer đọc các kết quả trung gian từ bộ nhớ cục bộ của nó và thực hiện thao tác lập chỉ mục cục bộ.

Với mỗi cặp key/value, Reduce được áp dụng để tính toán tần suất xuất hiện của từ đó trên toàn bộ tập dữ liệu (không bắt buộc) và các chỉ mục một phần được tạo sau đó được kết hợp để tạo ra một chỉ mục hoàn chỉnh. Kết quả được ghi lại trong chỉ mục phân tán, inverter sẽ là thành phần đảm nhiệm công việc này.[1]

3. Ví dụ điển hình

Các giai đoạn được mô phỏng như sau:



Hình 2: Mô phỏng MapReduce

Tổng thể, ta có máy Master và các worker là: M1,M2,M3,M4,R1,R2 các worker sẽ làm việc độc lập, sau khi xong một bước, chúng sẽ tiến hành lưu cục bộ dữ liệu để chuẩn bị cho bước sau

- Bước Split: bộ từ điển sẽ được chia nhỏ theo từng khối
- Bước Map: Các danh sách thẻ định vị được đọc bởi các worker M1, M2, M3, M4
- Bước Shuffle: bước này có thể gộp chung với bước Map, các danh sách thẻ định vị được gom nhóm lại với nhau và lưu cục bộ.

- Bước Reduce: các R1, R2 tổng hợp lại danh sách thẻ định vị, quy tắc quản lý các thẻ định nào tùy theo cấu hình, ở ví dụ trên, R1: quản lý thẻ cho từ "Bảo hiểm", "Ô tô" và "Xe máy"; R2 quản lý cho "Xe tải" và "Trường học"
  Và cuối cùng tổng hợp lại danh sách thẻ định vị, tổng hợp tf, ta có kết quả như sau:

Hình 3: Kết quả của MapReduce

4. Tài liệu tham khảo

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
[2] Jeffrey Dean and Sanjay Ghemawat: MapReduce: Simplified Data Processing on Large Clusters, Google, Inc
[3] Top dev: MapReduce là gì? Tổng quan về mô hình MapReduce
[4] Distributed indexing - Stanford NLP Group
[5] Mapreduce là gì? Tổng quan thông tin về mô hình lập trình Mapreduce - bizflycloud.vn