Trần Thanh Hùng	N19DCCN072
Nguyễn Văn Danh	N19DCCN028
Vũ Đức Anh	N19DCCN011

Nghiên cứu xây dựng chỉ mục phân tán sử dụng MapReduce

1. Giới thiệu:

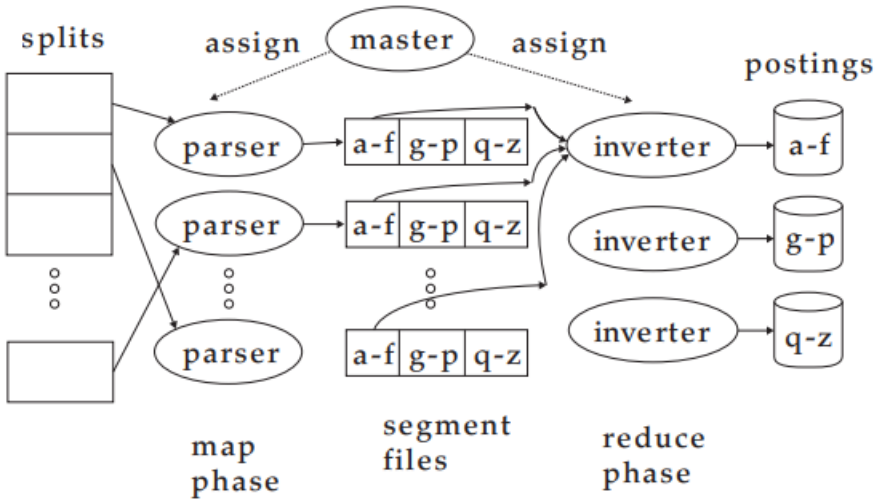
Ngày nay trong các ứng dụng web và phần mềm, việc tìm kiếm thông tin là một nhu cầu thiết yếu. Việc xây dựng một hệ thống chỉ mục phân tán là cần thiết để xử lý tập dữ liệu lớn và cho phép tìm kiếm thông tin một cách hiệu quả. MapReduce là một kỹ thuật mạnh mẽ để xử lý tập dữ liệu lớn bằng cách phân tán các tác vụ trên các nút khác nhau trong một cụm máy tính.

2. Tổng quan:

Lập chỉ mục phân tán với MapReduce là một kỹ thuật được sử dụng trong xử lý dữ liệu quy mô lớn để có thể tạo chỉ mục dữ liệu hiệu quả trên nhiều nút (Node) trong môi trường phân tán. MapReduce là một mô hình lập trình cho phép xử lý song song các tập dữ liệu lớn trên nhiều nút trong một cụm (Parser). Quá trình lập chỉ mục liên quan đến việc chia dữ liệu thành các phần nhỏ hơn, xử lý từng phần trên một nút riêng biệt, sau đó kết hợp các kết quả để tạo chỉ mục cuối cùng.

MapReduce được thiết kế cho các cụm máy tính lớn. Mục đích của một cụm là giải quyết các vấn đề lớn bằng các máy hoặc các node xây dựng từ hàng giá rẻ có các bộ phận tiêu chuẩn (bộ xử lý, bộ nhớ, đĩa) thay vì trên một siêu máy tính có phần cứng chuyên dụng. Mặc dù hàng trăm hoặc hàng ngàn máy có sẵn trong các cụm như vậy, các máy riêng lẻ có thể bị lỗi bất cứ lúc nào. Do đó, để lập chỉ mục phân tán hiệu quả chúng ta cần chia công việc thành các phần có thể dễ dàng chỉ định và trong trường hợp không thành công vẫn có thể chỉ định lại.

3. Cách hoạt động:



Ảnh lấy từ: Figure 4.5 An example of distributed indexing with MapReduce. Adapted from Dean and Ghemawat (2004). (Introduction to IndexInformation Retrieval)

MapReduce bao gồm hai giai đoạn, giai đoạn Map và giai đoạn Reduce.

Trong giai đoạn Map, dữ liệu đầu vào được chia thành các khối nhỏ hơn để có thể xử lí được bởi các máy phổ thông trong thời gian ngắn, rồi được xử lý song song bởi các máy này (các nút trong một cụm). Mỗi nút áp dụng một hàm ánh xạ cho dữ liệu và tạo các cặp khóa giá trị (key-value) làm đầu ra. Trong đó khóa là các từ khóa và giá trị là tên tài liệu chứa các từ khóa. Các khóa được sử dụng để nhóm dữ liệu và các giá trị được sử dụng để thực hiện các thao tác tổng hợp trong giai đoạn Reduce.

Trong giai đoạn Reduce, đầu ra từ giai đoạn Map được tổng hợp để tạo chỉ mục ngược. Các nút trong cụm áp dụng hàm rút gọn cho dữ liệu, hàm này kết hợp các giá trị với cùng một khóa. Đầu ra cuối cùng là danh sách các cặp khóa giá trị, trong đó các khóa là các thuật ngữ được lập chỉ mục và các giá trị là các mục dữ liệu tương ứng. Chỉ mục ngược sẽ được lưu trữ trên nhiều máy chủ khác nhau để đảm bảo khả năng mở rộng và hiệu suất.

Master Node chịu trách nhiệm quản lý toàn bộ quá trình tính toán và truyền dữ liệu trong hệ thống phân tán, cụ thể là quản lí tài nguyên, phân phối các task và quản lý trạng thái.

4. Ví dụ về MapReduce dùng để lập chỉ mục phân tán:

Giả sử chúng ta có một bộ sưu tập lớn các tài liệu văn bản mà chúng ta muốn lập chỉ mục. Chúng ta có thể sử dụng MapReduce để xây dựng một chỉ mục ngược, ánh xạ từng thuật ngữ trong bộ sưu tập vào danh sách tài liệu chứa thuật ngữ đó.

Trong giai đoạn Map, chúng ta chia bộ sưu tập thành các phần nhỏ hơn và mỗi nút trong cụm xử lý một phần tài liệu. Hàm ánh xạ lấy từng tài liệu và trích xuất các thuật ngữ riêng lẻ. Đối với mỗi thuật ngữ, hàm ánh xạ xuất ra một cặp khóa giá trị, trong đó khóa là thuật ngữ và giá trị là tên tài liệu. Ví dụ: nếu một tài liệu chứa các thuật ngữ "big data" và " analytics", thì hàm ánh xạ sẽ xuất ra các cặp khóa giá trị sau:

"big": document1

"data": document1

"analytics": document1

Trong giai đoạn Reduce, các nút trong cụm kết hợp đầu ra từ giai đoạn Map. Hàm rút gọn nhận các cặp khóa giá trị và tổng hợp các giá trị cho từng khóa. Kết quả là một danh sách các cặp khóa giá trị, trong đó mỗi khóa là một thuật ngữ và mỗi giá trị là một danh sách ID tài liệu chứa thuật ngữ đó. Ví dụ:

"big": [document1, document2, document3]

"data": [document1, document3, document4]

"analytics": [document1, document5]

Đầu ra cuối cùng là chỉ mục đảo ngược, có thể được lưu trữ trong hệ thống tệp phân tán, chẳng hạn như Hệ thống tệp phân tán Hadoop (HDFS). Sau đó, chỉ mục ngược có thể được sử dụng để tra cứu nhanh các tài liệu có chứa một thuật ngữ nhất định, điều này rất hữu ích cho các công cụ tìm kiếm và các hệ thống truy xuất thông tin khác.

5. Kết luận:

Lập chỉ mục phân tán với MapReduce có nhiều ưu điểm, bao gồm khả năng mở rộng, khả năng chịu lỗi và sử dụng tài nguyên hiệu quả. Bằng cách xử lý dữ liệu song song trên nhiều nút, quy trình lập chỉ mục có thể được hoàn thành nhanh hơn nhiều so với các kỹ thuật lập chỉ mục truyền thống. Ngoài ra, khung MapReduce cung cấp khả năng chịu lỗi bằng cách tự động phát hiện và xử lý lỗi ở các nút.

Tóm lại, MapReduce là một kỹ thuật mạnh mẽ và đơn giản về mặt khái niệm để triển khai xây dựng chỉ mục trong môi trường phân tán. Bằng cách cung cấp một phương pháp bán tự động để phân chia việc xây dựng chỉ mục thành các nhiệm vụ nhỏ hơn, nó có thể mở rộng quy mô thành các bộ dữ liệu lớn gần như tùy ý, với điều kiện các cụm máy tính có kích thước đủ lớn.