THỰC HÀNH MÔN GIẢI THUẬT XỬ LÝ SONG SONG VÀ PHÂN BỐ

BÀI THỰC HÀNH SỐ 4

ỨNG DỤNG MAPREDUCE TRONG HADOOP: THỰC HÀNH ĐẾM TỪ VỚI DỮ LIỆU KÍCH THƯỚC LỚN

---

1. MỤC TIÊU

- Những công nghệ và mô hình được giới thiệu trong bài thực hành làm sinh viên hiểu được vai trò của Hadoop việc lưu trữ và phân tích dữ liệu lớn trong môi trường phân tán.
- Một số lợi ích khi xử lý bài toán lớn trên môi trường phân tán.

- Xử lý dữ liệu lớn hiệu quả: Mục tiêu chính của việc sử dụng Hadoop là để xử lý dữ liệu lớn một cách hiệu quả, nhanh chóng, và mở rộng theo thời gian. Bài toán Word Count trên file lớn là một ví dụ điển hình để thể hiện khả năng này của Hadoop.
- Phân tích dữ liệu trên hệ thống phân tán: Sử dụng Hadoop giúp phân chia và phân phối dữ liệu trên nhiều máy tính trong một cluster. Mỗi máy tính (hoặc node) sẽ xử lý một phần của dữ liệu. Qua bài toán Word Count, chúng ta có thể thấy rõ sự phân chia này.
- Đảm bảo tính độ tin cậy và khả dụng: Hadoop HDFS lưu trữ nhiều bản sao của cùng một khối dữ liệu trên các node khác nhau, giúp đảm bảo dữ liệu không bị mất khi một số node gặp sự cố.
- Tối ưu hóa tài nguyên: Sử dụng MapReduce giúp tối ưu việc sử dụng tài nguyên trên từng node. Mỗi node chỉ xử lý dữ liệu mà nó lưu trữ, giảm thiểu việc truyền dữ liệu qua lại giữa các node và tăng hiệu suất xử lý.

- Mở rộng linh hoạt: Khi dữ liệu tăng lên, Hadoop Cluster có thể mở rộng bằng cách thêm node mới một cách dễ dàng. Bài toán Word Count giúp thử nghiệm và đánh giá khả năng mở rộng của hệ thống.
- Tiết kiệm chi phí: Hadoop được xây dựng trên phần cứng phổ thông, giúp giảm chi phí đầu tư ban đầu và chi phí bảo dưỡng so với các giải pháp lưu trữ truyền thống.
- Học và nâng cao kỹ năng: Bài toán Word Count cung cấp cơ hội cho sinh viên làm quen và nâng cao kỹ năng làm việc với Hadoop và MapReduce.
- Benchmarking và tối ưu hóa: Word Count có thể được sử dụng như một công cụ để đánh giá hiệu suất của Hadoop Cluster. Dựa vào kết quả, có thể tiến hành tối ưu hóa hệ thống, cấu hình, hoặc mã nguồn.

2. KIẾN THỨC LIÊN QUAN

a. Hadoop:

Hadoop là một dự án mã nguồn mở của Apache Foundation, được thiết kế để lưu trữ và xử lý lượng dữ liệu lớn trên các hệ thống phân tán. Hadoop giúp xử lý hàng petabyte dữ liệu và hỗ trợ các doanh nghiệp và tổ chức trong việc phân tích dữ liệu lớn.

b. Hadoop Cluster:

Hadoop Cluster là một nhóm máy được kết nối với nhau qua mạng, làm việc cùng nhau để lưu trữ và xử lý dữ liệu. Cluster được chia thành hai loại máy chính:

- Master Nodes: Điều phối việc lưu trữ dữ liệu và xử lý tác vụ.
- Slave/Worker Nodes: Lưu trữ dữ liệu thực tế và thực hiện các tác vụ xử lý trên dữ liệu đó.

c. MapReduce:

MapReduce là mô hình lập trình cho xử lý song song và phân tán dữ liệu trên Hadoop Cluster. Gồm hai bước chính:

- Map: Phân loại và xử lý dữ liệu, sau đó tạo ra một tập dữ liệu trung gian dưới dạng cặp (key, value).
- Reduce: Xử lý dữ liệu từ bước Map và tạo ra một tập dữ liệu đầu ra được sắp xếp.

d. HDFS (Hadoop Distributed File System):

HDFS là hệ thống tệp phân tán của Hadoop, thiết kế để lưu trữ lượng dữ liệu rất lớn trên các máy tính phổ thông. Các tính năng chính của HDFS bao gồm:

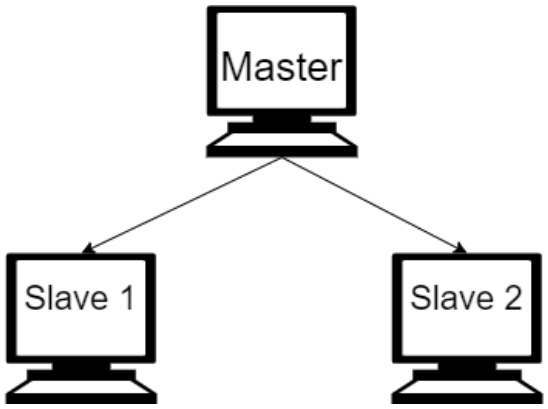
- Độ tin cậy cao: Dữ liệu được sao chép trên nhiều máy (mặc định là 3 bản sao) để đảm bảo dữ liệu không bị mất khi có sự cố về phần cứng.
- Phân tán: Dữ liệu được chia thành các khối có kích thước cố định và được phân tán trên các máy trong cluster.
- Tối ưu cho việc xử lý dữ liệu lớn: HDFS không phù hợp cho việc truy cập dữ liệu ngẫu nhiên hoặc việc đọc/ghi dữ liệu nhỏ, nhưng rất hiệu quả cho việc đọc dữ liệu tuần tự và xử lý dữ liệu lớn.

e. Word Count:

Word Count là bài toán cơ bản và phổ biến khi làm quen với MapReduce. Mục tiêu là đếm số lần xuất hiện của mỗi từ trong một tập dữ liệu lớn.

- Bước Map: Đọc dữ liệu, tách từ và trả về mỗi từ với giá trị là 1 (ví dụ: ("Hello", 1)).
- Bước Reduce: Tổng hợp các giá trị của cùng một từ và trả về tổng số lần xuất hiện của từ đó (ví dụ: ("Hello", 10)).

3. YÊU CẦU BÀI TOÁN



Hình 1. Cấu trúc mạng theo kiến trúc master-slave

- Word Count trên Hadoop với file văn bản ~2GB

- Xây dựng và cấu hình một môi trường Hadoop trên các máy tính cá nhân để thực hiện bài toán đếm số lần xuất hiện của mỗi từ trong một file văn bản có kích thước ~2GB.
- Cài đặt và cấu hình Hadoop trên một cluster gồm 3 máy: 1 máy chủ chính (master) và 2 máy chủ trợ giúp (slave). Mô hình mạng như hình 1. Khuyến khích sử dụng các máy thật để có đủ tài nguyên thực thi bài toán word count trên file lớn.
- Lưu trữ dữ liệu (file text đã cho) trên cụm Với Hadoop Distributed File System (HDFS)
- Thực hiện bài toán Word Count trên một tệp tin có kích thước ~2GB, sử dụng MapReduce. Ghi lại thời gian thực thi bài toán word count cho file text đã cho.

❖ **Hướng dẫn các bước thực hiện:**

- Cài đặt hadoop:

- Cài đặt Java JDK (phiên bản tương thích với Hadoop bạn sẽ cài đặt) và cấu hình biến môi trường JAVA_HOME.
- Tải và cài đặt Hadoop (phiên bản mới nhất hoặc phiên bản tương thích với hệ thống của bạn).
- Cấu hình biến môi trường cho Hadoop: HADOOP_HOME.

- Cấu hình Hadoop:

- Cấu hình core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml và một số file cần thiết khác trên tất cả các máy.
- Cấu hình SSH để máy master có thể truy cập vào tất cả máy slaves.
- Khởi tạo HDFS.

- Chạy Hadoop Cluster:

- Sử dụng các lệnh để khởi động từng dịch vụ của Hadoop Cluster hoặc Khởi động tất cả các dịch vụ bằng script start-all.sh
- Đảm bảo NameNode, DataNode, ResourceManager, và NodeManager đều đang chạy mà không gặp lỗi.

- Tải và chuẩn bị dữ liệu:

- Tải file văn bản ~2GB đã cung cấp trên website môn học.

- Tạo một thư mục trên HDFS để lưu trữ file văn bản
- Sao chép file từ máy local lên HDFS

- Lập trình và biên dịch:

- Viết chương trình Mapper để xử lý dữ liệu đầu vào và xuất ra các cặp (word, 1).
- Viết chương trình Reducer để nhận các cặp từ Mapper và tính tổng số lần xuất hiện của từng từ.
- Biên dịch và tạo package JAR cho chương trình.

- Chạy chương trình MapReduce:

- Thực thi chương trình trên Hadoop Cluster:
- Đảm bảo quá trình thực thi hoàn tất mà không gặp lỗi.

- Kiểm tra kết quả:

- Kiểm tra số lần xuất hiện của một số từ cụ thể và so sánh với kết quả thu được từ các công cụ đếm từ khác. Hiển thị top 5 các từ xuất hiện nhiều nhất trong file.

4. NỘP BÀI VÀ ĐÁNH GIÁ

- Mã nguồn chương trình MapReduce (nếu có tối ưu hóa) nộp trên website môn học.
- Buổi báo cáo ngắn mô tả quá trình cài đặt, cấu hình, vấn đề gặp phải và cách giải quyết, mở dashboard của từng dịch vụ, chạy trực tiếp bài toán Word Count tại buổi báo cáo và hiển thị kết quả thời gian chạy tại yarn web UI.
- Các trường hợp nộp bài trễ, báo cáo trễ, sao chép, gian lận sẽ xử lý tùy mức độ (trừ 10%-100% điểm).

TÀI LIỆU THAM KHẢO:

- Cơ bản về Hadoop và MapReduce: Warrener, P. & Uribe, T. (2019). Mastering Big Data with Hadoop. Packt Publishing.
- Hadoop Tutorial: Hadoop Tutorial (tutorialspoint.com)