TÌM HIỂU KHÁI QUÁT VỀ HỆ THỐNG HADOOP

1. HADOOP

1.1 Khái niệm

Hadoop là một Apache framework mã nguồn mở được viết bằng java, cho phép xử lý phân tán (distributed processing) các tập dữ liệu lớn trên các cụm máy tính (clusters of computers) thông qua mô hình lập trình đơn giản. Hadoop được thiết kế để mở rộng quy mô từ một máy chủ đơn sang hàng ngàn máy tính khác có tính toán và lưu trữ cục bộ (local computation and storage).

1.2 Kiến Trúc

- Hadoop gồm 4 module:

- Hadoop Common: Đây là các thư viện và tiện ích cần thiết của Java để các module khác sử dụng. Những thư viện này cung cấp hệ thống file và lớp OS trừu tượng, đồng thời chứa các mã lệnh Java để khởi động Hadoop.
- Hadoop YARN: Đây là framework để quản lý tiến trình và tài nguyên của các cluster.
- Hadoop Distributed File System (HDFS): Đây là hệ thống file phân tán cung cấp truy cập thông lượng cao cho ứng dụng khai thác dữ liệu.
- Hadoop MapReduce: Đây là hệ thống dựa trên YARN dùng để xử lý song song các tập dữ liệu lớn.

1.3 Hoạt động

Giai đoạn 1:

Một user hay một ứng dụng có thể submit một job lên Hadoop (hadoop job client) với yêu cầu xử lý cùng các thông tin cơ bản:

Truyền dữ liệu lên server(input) để bắt đầu phân tán dữ liệu và đưa ra kết quả (output).

Các dữ liệu được chạy thông qua 2 hàm chính là map và reduce.

Map: sẽ quét qua toàn bộ dữ liệu và phân tán chúng ra thành các dữ liệu con.

Reduce: sẽ thu thập các dữ liệu con lại và sắp xếp lại chúng.

Các thiết lập cụ thể liên quan đến job thông qua các thông số truyền vào.

Giai đoạn 2:

Hadoop job client submit job (file jar, file thực thi) và bắt đầu lập lịch làm việc(JobTracker) đưa job vào hàng đợi .

Sau khi tiếp nhận yêu cầu từ JobTracker, server cha(master) sẽ phân chia công việc cho các server con(slave). Các server con sẽ thực hiện các job được giao và trả kết quả cho server cha.

Giai đoạn 3:

TaskTrackers dùng để kiểm tra đảm bảo các MapReduce hoạt động bình thường và kiểm tra kết quả nhận được (quá trình output).

Khi “chạy Hadoop” có nghĩa là chạy một tập các trình nền - daemon, hoặc các chương trình thường trú, trên các máy chủ khác nhau trên mạng của bạn. Những trình nền có vai trò cụ thể, một số chỉ tồn tại trên một máy chủ, một số có thể tồn tại trên nhiều máy chủ.

1.4 Ưu điểm

Hadoop framework cho phép người dùng nhanh chóng viết và kiểm tra các hệ thống phân tán. Đây là cách hiệu quả cho phép phân phối dữ liệu và công việc xuyên suốt các máy trạm nhờ vào cơ chế xử lý song song của các lõi CPU.

Hadoop không dựa vào cơ chế chịu lỗi của phần cứng fault-tolerance and high availability (FTHA), thay vì vậy bản thân Hadoop có các thư viện được thiết kế để phát hiện và xử lý các lỗi ở lớp ứng dụng.

Hadoop có thể phát triển lên nhiều server với cấu trúc master-slave để đảm bảo thực hiện các công việc linh hoạt và không bị ngắt quãng do chia nhỏ công việc cho các server slave được điều khiển bởi server master.

Hadoop có thể tương thích trên mọi nền tảng như Window, Linux, MacOs do được tạo ra từ Java.