ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

BÁO CÁO BÀI TẬP LỚN MÔN CƠ SỞ DỮ LIỆU NÂNG CAO

Phân tích dữ liệu sử dụng môi trường Hadoop MapReduce

GVHD:
PGS.TS. Đặng Trần Khánh

Sinh viên thực hiện:
Nguyễn Quốc Long - MSSV:1770023

TP.Hồ Chí Minh, 08/11/2018.

Mục lục

Bài luận tìm hiểu về việc sử dụng môi trường Hadoop Map Reduce để phân tích dữ liệu

Nguyễn Quốc Long[1]

MSSV: 1770023

Tóm tắt nội dung. Tài liệu tìm hiểu về thật ngữ phân tích dữ liệu (data analysis) và môi trường hadoop map-reduce, cũng như việc áp dụng việc phân tích dữ liệu của YouTube sử dụng framework Hadoop MapReduce trên nền tảng đám mây AWS (Amazon Web Services).
Topic: proposed by student & approved by the lecturer & tutor (technical requirements: research paper reading & evaluation/empirical test

Từ khóa: algorithm, hadoop map-reduce, phân tích dữ liệu, data analysis

1 Giới thiệu

Phân tích dữ liệu (Tiếng Anh: Data Analytics) là quá trình phát hiện, giải thích và truyền đạt các mô hình có ý nghĩa trong dữ liệu.

Phân tích dữ liệu đóng vai trò quan trọng trong việc xác định kinh doanh và chiến lược tiếp thị. Dự án này có thể đóng một vai trò quan trọng trong giúp doanh nghiệp quảng cáo xác định xu hướng nhất thể loại và đầu tư vào các thể loại video. Dữ liệu YouTube API rất hữu ích để lấy dữ liệu từ trang web và sau đó xử lý nó trong môi trường MapReduce của Hadoop. Để phát triển hơn nữa tầm quan trọng của dự án, công việc trong tương lai có thể được tập trung hơn vào chuyển đổi các dữ liệu này thành các quyết định có tác động tốt đến thế giới thực. Điều này có thể được sử dụng trong các doanh nghiệp chiết xuất hữu ích thông tin từ dữ liệu phi cấu trúc.

Hadoop MapReduce là một framework được Google phát hành năm 2011. Apache Hadoop hay Hadoop là một software framework hỗ trợ các ứng dụng chuyên sâu và được sử dụng miễn phí với giấy phép Apache license 2.0.

Trong bài báo, tập dữ liệu YouTube được phân tích bằng giải thuật MapReduce để tìm ra được những thông số thống kê sau đây:

- Năm danh mục có số lượng video tải lên là lớn nhất.
- Năm tài khoản tải lên số lượng video nhiều nhất.
- Năm video có lượt xem cao nhất.

2 Tóm tắt nội dung

Dự án giải quyết việc phân tích dữ liệu của YouTube sử dụng framework Hadoop MapReduce trên nền tảng AWS (Amazon Web Services). Hệ thống gom cụm Hadoop được thiết lập trên các dữ liệu đám mây cá nhân được gọi là AWS (Amazon Web Services). Với AWS, tác giả thiết lập trên các máy EC2 với một máy name node và 5 máy data nodes.

3 Phân tích dữ liệu sử môi trường Hadoop MapReduce

3.1 Nghiên cứu phát triển ứng dụng Hadoop

3.2 Trích xuất dữ liệu video từ YouTube API

3.3 Lưu trữ dữ liệu trên hdfs (Hadoop Distributed File System)

3.4 Mapper và reducer

3.5 Kết quả đầu ra

4 Đánh giá và kiểm tra thực nghiệm

5 Kết Luận

Bài báo cáo đã trình bày một cách tông quát làm sao để phân tích dữ liệu sử dụng môi trường Hadoop MapReduce. Kiến thức trong bài được dựa trên bài báo cáo khoa học " Data analysis using hadoop MapReduce environment" (thông tin bài báo các bạn có thể xem thêm ở đường dẫn sau đây: https://ieeexplore.ieee.org/document/8258541/metrics).

6 Tài liệu Tham Khảo

http://en.wikipedia.org
https://vi.wikipedia.org/wiki
https://ieeexplore.ieee.org/document/8258541/metrics#metrics