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.
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

This project deals with analysis of YouTube data using Hadoop MapReduce framework on a cloud platform AWS. Hadoop multi node cluster is setup on private cloud called AWS (Amazon Web Services). Within AWS, I have set up EC2 instances with one name node and 5 data nodes. The video statistics obtained from the API is stored into the HDFS (Hadoop Distributed File System) and the data processing is done by the MapReduce system.

3 Giới thiệu vấn đề

4 Phân tích dữ liệu sử dụng môi trường Hadoop MapReduce

4.1 Research on Hadoop application development

4.2 Extracting video data from YouTupe API

4.3 Storing into hdfs

4.4 Mapper and reducer

4.5 Outcome

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