BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NHA TRANG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC

XỬ LÝ DỮ LIỆU LỚN (INS358)

XÂY DỰNG CHƯƠNG TRÌNH TỔNG HỢP LIÊN KẾT ĐẾN TRANG WEB TRÊN HADOOP/SPARK

Sinh viên thực hiện: **Lý Văn Quỳ**

MSSV: **62133721**

Lớp: **62.CNTT-1**

Giảng viên: **TS. Nguyễn Đình Hưng**

Khánh Hòa - 2024

KẾT QUẢ ĐÁNH GIÁ ĐỒ ÁN MÔN HỌC

Họ và tên sinh viên:

MSSV:

Lớp:

Nội dung	Trọng số	Điểm
1. Giải quyết vấn đề		
1.1. Phân tích bài toán; thu thập, khảo sát và chuẩn bị dữ liệu; thiết kế giải thuật	20%	
1.2. Cài đặt, triển khai ứng dụng trên Hadoop	20%	
1.3. Cài đặt, triển khai ứng dụng trên Spark	20%	
2. Báo cáo bài tập lớn		
2.1. Nội dung báo cáo	20%	
2.2. Vấn đáp	20%	
Điểm trung bình		

Giảng viên

Lời cam đoan

Tôi cam đoan đây là công trình do tôi tự thực hiện. Các nội dung nghiên cứu, số liệu và kết quả thực nghiệm là trung thực. Các số liệu, công trình sử dụng của tác giả khác đều được trích dẫn nguồn gốc rõ ràng.

Tất cả phần mềm sử dụng trong đồ án này đều là mã nguồn mở.

Nếu phát hiện có bất kì sự gian lận nào, tôi xin chịu hoàn toàn trách nhiệm.

Lý Văn Quỳ

Mục lục

Danh sách bảng

Danh sách hình vẽ

Danh sách giải thuật

Chương 1

GIỚI THIỆU

1.1 Tổng quan về dữ liệu lớn

Ngày nay, lĩnh vực dữ liệu lớn (Big Data) đang được quan tâm nghiên cứu và ứng dụng. Việc khai thác hiệu quả dữ liệu lớn giúp khai phá, phát hiện tri thức, giúp doanh nghiệp, tổ chức nâng cao hiệu quả hoạt động.

1.2 Mục tiêu của đề tài

Các mục tiêu chính của đồ án:

- Tìm hiểu tổng quan về dữ liệu lớn và ứng dụng;
- Tìm hiểu các phương pháp, công nghệ, công cụ tiêu biểu trong xử lý dữ liệu lớn;
- Vận dụng kiến thức, công cụ để xây dựng một ứng dụng xử lý dữ liệu lớn đơn giản.

1.3 Cấu trúc của Đồ án

Đồ án gồm các phần như sau:

- Chương 1: Giới thiệu.

- Chương 2: Nội dung và phương pháp thực hiện.
- Chương 3: Kết luận.

Chương 2

NỘI DUNG VÀ PHƯƠNG PHÁP THỰC HIỆN

2.1 Phân tích bài toán

Cho bộ dữ liệu văn bản chứa các cặp (source, target) biểu diễn source -> target. Yêu cầu: Với mỗi trang web, hãy tổng hợp tất cả các trang web có liên kết đến nó.

2.2 Thu thập và chuẩn bị dữ liệu

Trong đồ án này chúng tôi sử dụng dữ liệu từ ĐH Stanford.

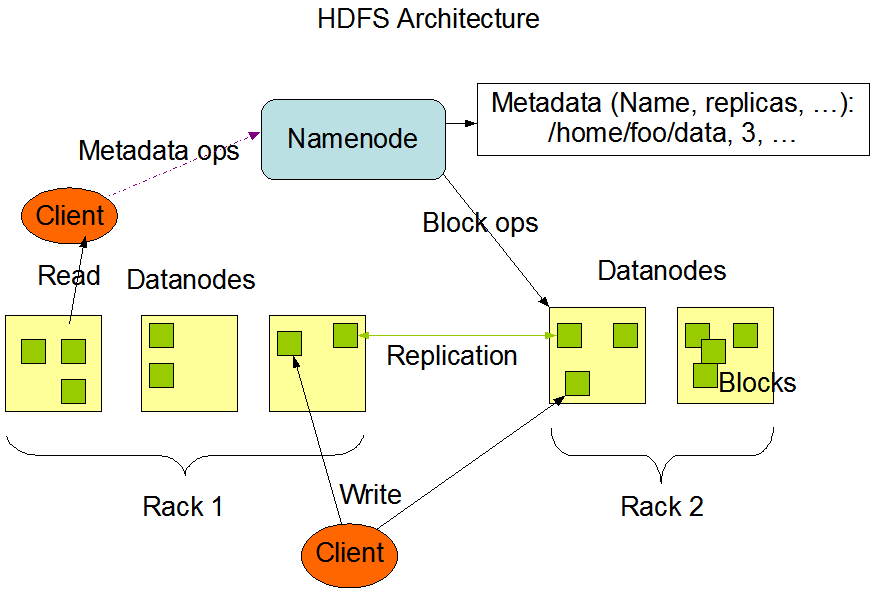
2.3 Cài đặt và triển khai ứng dụng trên Hadoop

Hình 2.1 mô tả kiến trúc của Hadoop [1].

2.3.1 Cài đặt Hadoop

2.3.2 Xây dựng giải thuật

Giải thuật MapReduce tổng hợp liên kết đến trang Web



Hình 2.1: Kiến trúc của Hadoop.

Pha Map

Giải thuật 1: Pha Map xử lý liên kết đến trang web

```
1: for each (source, target) do
2:     print(target, source)
3: end for
```

Pha Reduce

Giải thuật 2: Pha Reduce xử lý liên kết đến trang web

```
1: for each (target) do
2:     print(source)
3: end for
```

2.3.3 Lập trình ứng dụng

Lập trình pha Map

```
#!/usr/bin/python3
"""mapper.py"""

import sys

# Chương trình Python chạy trên Hadoop MapReduce qua tính năng Streaming.
# Dữ liệu vào từ thiết bị nhập chuẩn (STDIN)
# Kết quả xử lý gửi ra thiết bị xuất chuẩn (STDOUT)

for line in sys.stdin.buffer.raw:
    source_site, target_site = line.split()
    print('%s\t%s' % (target_site, source_site))
```

Lập trình pha Reduce

2.3.4 Thực thi ứng dụng

2.4 Cài đặt và triển khai ứng dụng trên Spark

2.4.1 Cài đặt Spark

2.4.2 Lập trình ứng dụng

2.4.3 Thực thi ứng dụng

Chương 3

KẾT LUẬN

3.1 Đánh giá chung

3.1.1 Những kết quả đạt được

3.1.2 Một số hạn chế

3.2 Hướng phát triển

Tài liệu tham khảo

[1] Tom White. *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.