**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue logo with a ball and a swirly circle

Description automatically generated🙠🙟🕮🙝🙢

**ĐỒ ÁN CUỐI KỲ**

**NHẬP MÔN DỮ LIỆU LỚN (BIG DATA ESSENTIALS)**

**MÃ HP: 242BDES333877\_02**

**ĐỀ TÀI**

**HỆ THỐNG PHÂN TÍCH XU HƯỚNG VIỆC LÀM**

**VỀ CÔNG NGHỆ THÔNG TIN**

**GVHD: ThS. Trần Quang Khải**

**Sinh viên thực hiện:**

|  |  |  |
| --- | --- | --- |
| **STT** | **HỌ VÀ TÊN SV** | **MSSV** |
| 1 | Nguyễn Đức Tín | 22110434 |
| 2 | Nguyễn Đức Dũng | 22110301 |
| 3 | Hoàng Mai Công | 22110289 |

Thành phố Hồ Chí Minh, tháng năm 2025

**HỌC KÌ 2 NĂM HỌC 2024-2025**

**NHÓM 6**

*Đề tài*: HỆ THỐNG PHÂN TÍCH XU HƯỚNG VIỆC LÀM VỀ CÔNG NGHỆ THÔNG TIN

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã số sinh viên** | **Họ và tên** | **Nhiệm vụ** | **Tỉ lệ hoàn thành** |
| 22110434 | Nguyễn Đức Tín | * Thu thập dữ liệu từ các nguồn * Lưu trữ dữ liệu vào cơ sở dữ liệu MySQL | 100% |
| 22110301 | Nguyễn Đức Dũng | * Tiền xử lý dữ liệu * Truy vấn, thao tác, rút trích và tìm kiếm dữ liệu | 100% |
| 22110289 | Hoàng Mai Công | * Trực quan hóa dữ liệu * Xây dựng các chương trình MapReduce | 100% |

Nhận xét của giảng viên

TP.HCM, ngày tháng 5 năm 2025

TS. Nguyễn Thành Sơn

**LỜI CẢM ƠN**

Kính thưa thầy Trần Quang Khải,

Nhóm em xin bày tỏ lòng biết ơn sâu sắc đến thầy vì sự tận tình hướng dẫn và hỗ trợ trong suốt quá trình thực hiện đồ án môn Nhập môn dữ liệu lớn (Big Data Essentials). Sự quan tâm và chỉ dạy của thầy đã giúp chúng em không chỉ hoàn thành tốt đồ án mà còn học hỏi được nhiều kinh nghiệm quý báu trong lĩnh vực Big Data.

Nhờ có thầy, nhóm em đã có cơ hội tiếp thu những kiến thức bổ ích và ứng dụng vào thực tiễn một cách hiệu quả. Qua từng bài học, thầy đã giúp chúng em nắm vững những kỹ năng quan trọng, tạo tiền đề vững chắc cho những công việc và thách thức trong tương lai.

Chúng em vô cùng biết ơn thầy vì đã dành nhiều thời gian, tâm huyết để giúp đỡ và đồng hành cùng nhóm trong quá trình học tập. Một lần nữa, nhóm em xin gửi lời cảm ơn chân thành nhất đến thầy Trần Quang Khải.

**MỤC LỤC**

[**1.** **MỞ ĐẦU** 5](#_Toc197927236)

[**1.1.** **Giới thiệu đề tài** 5](#_Toc197927237)

[**1.2.** **Lý do chọn đề tài** 5](#_Toc197927238)

[**1.3.** **Mục tiêu nghiên cứu** 5](#_Toc197927239)

[**1.4.** **Phạm vi và phương pháp thực hiện** 5](#_Toc197927240)

[**2.** **PHẦN NỘI DUNG** 5](#_Toc197927241)

[**2.1.** **CÀI ĐẶT MÔI TRƯỜNG PHÁT TRIỂN** 5](#_Toc197927242)

[**2.1.1.** **Thông tin cấu hình máy tính và hệ thống** 5](#_Toc197927243)

[**2.1.2.** **Cài đặt các công cụ trong hệ sinh thái Hadoop** 5](#_Toc197927244)

[**2.2.** **THU THẬP DỮ LIỆU** 5](#_Toc197927245)

[**2.3.** **LƯU TRỮ DỮ LIỆU: MYSQL** 5](#_Toc197927246)

[**2.4.** **TIỀN XỬ LÝ DỮ LIỆU (LÀM SẠCH DỮ LIỆU)** 5](#_Toc197927247)

[**2.5.** **TRUY VẤN & PHÂN TÍCH DỮ LIỆU** 5](#_Toc197927248)

[**3.** **KẾT LUẬN** 6](#_Toc197927249)

[**4.** **TÀI LIỆU THAM KHẢO** 6](#_Toc197927250)

# **MỞ ĐẦU**

* 1. **Giới thiệu đề tài**
  2. **Lý do chọn đề tài**
  3. **Mục tiêu nghiên cứu**
  4. **Phạm vi và phương pháp thực hiện**

# **PHẦN NỘI DUNG**

* 1. **CÀI ĐẶT MÔI TRƯỜNG PHÁT TRIỂN** 
     1. **Thông tin cấu hình máy tính và hệ thống**
     2. **Cài đặt các công cụ trong hệ sinh thái Hadoop**
  2. **THU THẬP DỮ LIỆU**
* Các trang web để thu thập:
* TopCV: <https://www.topcv.vn/>
* VietnamWorks: <https://www.vietnamworks.com/>
  1. **LƯU TRỮ DỮ LIỆU: MYSQL**

## **TIỀN XỬ LÝ DỮ LIỆU (LÀM SẠCH DỮ LIỆU)**

* Loại bỏ các ký tự đặc biệt
* Downcase các chữ cái
  1. **TRUY VẤN & PHÂN TÍCH DỮ LIỆU**
* **Phân tích**:
  + Thành phố nào có nhiều việc làm CNTT nhất?
  + Mức lương trung bình theo kỹ năng (Java, Python,...).
  + Kỹ năng hot năm 2025

# **KẾT LUẬN**

# **TÀI LIỆU THAM KHẢO**