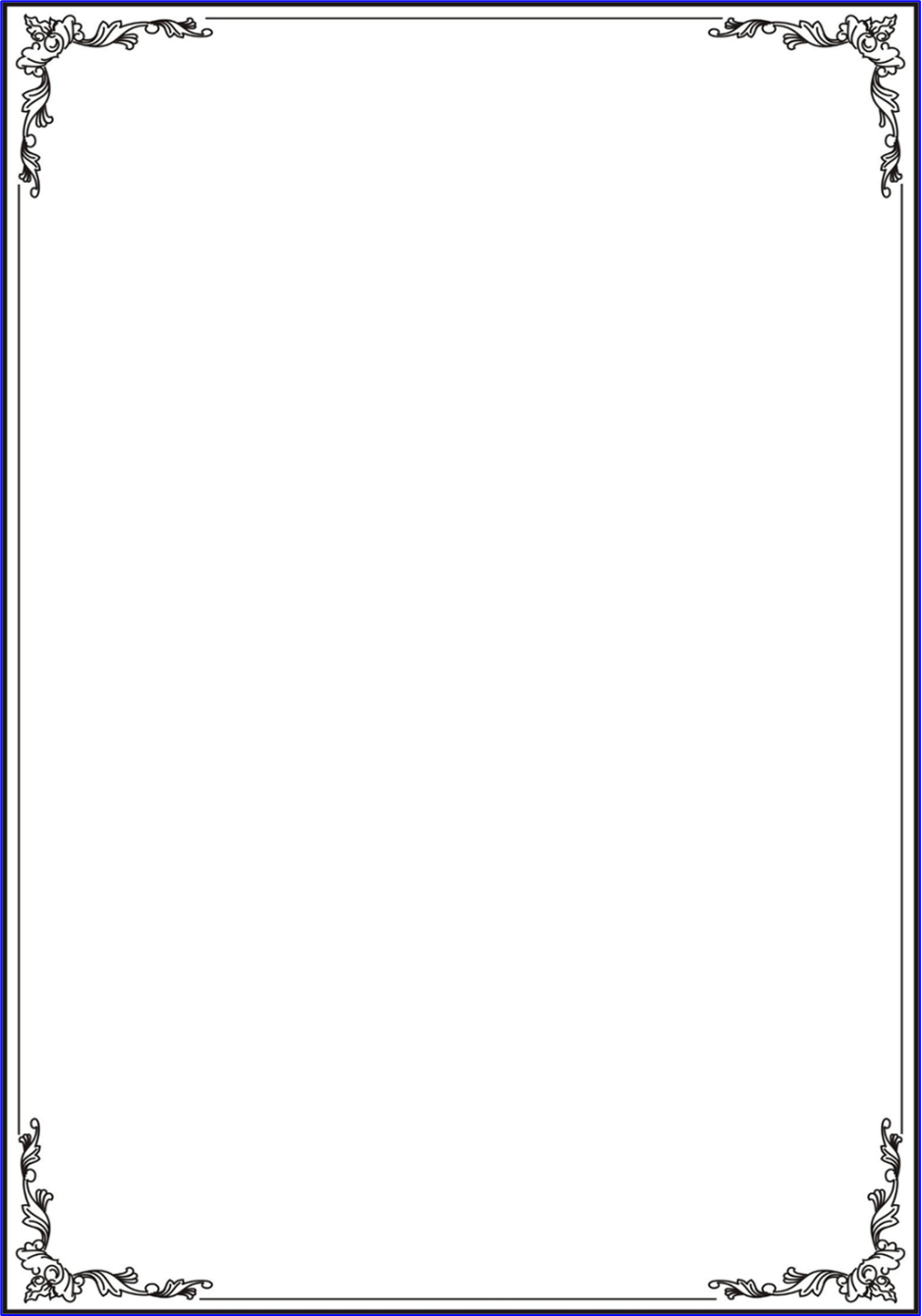
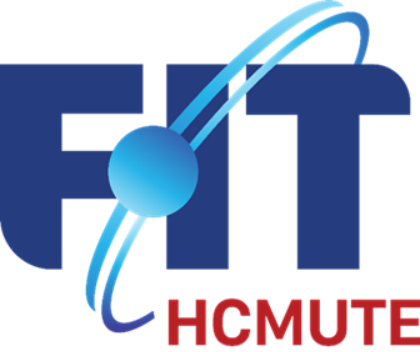
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

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

**KHOA CÔNG NGHỆ THÔNG TIN**

**ỨNG DỤNG BIG DATA TRONG PHÂN TÍCH VÀ   
TRA CỨU THÔNG TIN VIỆC LÀM**

**Giảng viên hướng dẫn: Ts. Trần Quang Khải**

**Thành viên thực hiện**

**1. Đoàn Trường Giang - 22110313**

**2. Bùi Quang Hoàn - 22110328**

**3. Bờ Rông Mỹ Thị Hồng Sáng - 22110005**

**MỤC LỤC**

[PHẦN MỞ ĐẦU 1](#_Toc197910845)

[1.1 Lý do chọn đề tài 1](#_Toc197910846)

[1.2 Mục tiêu nghiên cứu 2](#_Toc197910847)

[1.3 Phạm vi nghiên cứu 2](#_Toc197910848)

[1.4 Kết quả dự kiến đạt được 3](#_Toc197910849)

[PHẦN NỘI DUNG 4](#_Toc197910850)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 4](#_Toc197910851)

[1.1 Tổng quan về thị trường lao động và xu hướng thay đổi nghề nghiệp. 4](#_Toc197910852)

[1.1.1 Khái niệm thị trường lao động 4](#_Toc197910853)

[1.1.2 Xu hướng thay đổi nghề nghiệp 4](#_Toc197910854)

[2.2 Vai trò của Big Data trong phân tích và dự đoán xu hướng nghề nghiệp. 5](#_Toc197910855)

[2.2.1 Khái niệm Big Data 5](#_Toc197910856)

[2.2.2 Ứng dụng của Big Data trong phân tích thị trường lao động 6](#_Toc197910857)

[2.2.3 Thách thức khi ứng dụng Big Data trong dự đoán nghề nghiệp 6](#_Toc197910858)

[CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU 8](#_Toc197910859)

[3.1 Nguồn dữ liệu 8](#_Toc197910860)

[3.2 Quy trình xử lý dữ liệu 9](#_Toc197910861)

[CHƯƠNG 3: TỔNG QUAN DỰ ÁN 10](#_Toc197910862)

[3.1 Thiết lập hạ tầng 10](#_Toc197910863)

[3.2 Thu thập dữ liệu 11](#_Toc197910864)

[3.3 Xử lý và chuẩn bị dữ liệu 11](#_Toc197910865)

[3.4 Lưu trữ dữ liệu 12](#_Toc197910866)

[3.5 Sao lưu và phục hồi dữ liệu 13](#_Toc197910867)

[3.6 Thao tác CRUD trên MySQL Metastore của Hive 14](#_Toc197910868)

[3.7 Xử lý dữ liệu với MapReduce 14](#_Toc197910869)

[3.8 Phát triển ứng dụng web 15](#_Toc197910870)

[3.9 Trực quan hóa dữ liệu với PowerBI 15](#_Toc197910871)

[CHƯƠNG 4: TRIỂN KHAI KỸ THUẬT 17](#_Toc197910872)

[4.1 Dữ liệu 17](#_Toc197910873)

[4.1.1 Thu thập dữ liệu 17](#_Toc197910874)

[4.1.2 Làm sạch và chuẩn bị dữ liệu 18](#_Toc197910875)

[4.2 Lưu trữ 20](#_Toc197910876)

[4.2.1 Lưu trữ dữ liệu trên HDFS 20](#_Toc197910877)

[4.2.2 Tích hợp dữ liệu với Hive 20](#_Toc197910878)

[4.3 Triển khai trên môi trường Windows 21](#_Toc197910879)

[4.4 Chức năng 21](#_Toc197910880)

[4.4.1. Hỗ trợ thực thi truy vấn (query), thao tác với dữ liệu (CRUD) 21](#_Toc197910881)

[4.4.2. Sao lưu, phục hồi dữ liệu 21](#_Toc197910882)

[4.4.3. Trực quan dữ liệu thu thập 22](#_Toc197910883)

[4.5 Giao diện 22](#_Toc197910884)

[4.5.1 Giao diện người dùng (Frontend) 23](#_Toc197910885)

[4.5.2 Xử lý yêu cầu (Backend) 24](#_Toc197910886)

[PHẦN KẾT LUẬN 27](#_Toc197910887)

[DANH MỤC TÀI LIỆU THAM KHẢO 28](#_Toc197910888)

# PHẦN MỞ ĐẦU

### 1.1 Lý do chọn đề tài

Trong bối cảnh toàn cầu hóa và sự bứt phá của khoa học công nghệ, thị trường lao động đang trải qua những biến chuyển sâu sắc. Các ngành nghề truyền thống dần được thay thế hoặc chuyển đổi, trong khi những ngành nghề mới liên quan đến công nghệ số, trí tuệ nhân tạo, và năng lượng xanh liên tục xuất hiện. Những thay đổi này đòi hỏi người lao động phải không ngừng học hỏi và thích nghi để đáp ứng yêu cầu ngày càng cao của thị trường. Tại Việt Nam, sự phát triển mạnh mẽ của cuộc cách mạng công nghiệp 4.0, quá trình chuyển đổi số, và những tác động từ các yếu tố kinh tế - xã hội như già hóa dân số hay đô thị hóa đang khiến việc nắm bắt xu hướng nghề nghiệp trở nên cấp thiết hơn bao giờ hết.

Việc hiểu rõ các xu hướng thay đổi nghề nghiệp mang lại lợi ích thiết thực cho nhiều đối tượng:

Đối với cá nhân: Giúp định hướng con đường sự nghiệp, lựa chọn ngành học phù hợp, và trang bị các kỹ năng cần thiết để tránh nguy cơ thất nghiệp hoặc bị tụt hậu.

Đối với doanh nghiệp: Cung cấp thông tin quan trọng để xây dựng chiến lược phát triển nguồn nhân lực, dự đoán nhu cầu tuyển dụng, và thiết kế các chương trình đào tạo hiệu quả nhằm nâng cao năng lực cạnh tranh.

Đối với nhà hoạch định chính sách: Là cơ sở để xây dựng các chính sách giáo dục, đào tạo nghề, và phát triển thị trường lao động phù hợp với xu hướng kinh tế - xã hội, góp phần đảm bảo an sinh xã hội và giảm thiểu những tác động tiêu cực từ sự thay đổi nghề nghiệp.

Xuất phát từ thực tế đó, tôi lựa chọn đề tài “Ứng dụng Big Data trong Phân tích và Tra cứu Thông tin Việc làm” nhằm làm sáng tỏ các xu hướng biến động của thị trường lao động, đồng thời cung cấp những luận cứ khoa học để hỗ trợ các quyết định liên quan đến nghề nghiệp trong tương lai. Đề tài không chỉ mang tính thời sự mà còn góp phần vào sự phát triển bền vững của nền kinh tế Việt Nam trong bối cảnh hội nhập và chuyển đổi.

### 1.2 Mục tiêu nghiên cứu

Nghiên cứu này đặt ra các mục tiêu cụ thể như sau:

Phân tích các yếu tố chính đang và sẽ tác động đến sự thay đổi cơ cấu nghề nghiệp, bao gồm tiến bộ công nghệ, toàn cầu hóa, biến đổi nhân khẩu học, và các yếu tố kinh tế - xã hội.

Xác định các xu hướng thay đổi nghề nghiệp nổi bật tại Việt Nam trong giai đoạn 2025-2030.

Dự đoán các ngành nghề có tiềm năng phát triển (như công nghệ thông tin, năng lượng tái tạo, y tế), các ngành nghề có nguy cơ suy giảm, và các kỹ năng cần thiết để người lao động thích nghi với thị trường tương lai.

Đề xuất các khuyến nghị thực tiễn cho cá nhân, doanh nghiệp, và nhà hoạch định chính sách nhằm chủ động ứng phó với những biến động của thị trường lao động.

### 1.3 Phạm vi nghiên cứu

Để đảm bảo tính tập trung và khả thi, nghiên cứu này giới hạn phạm vi như sau:

Đối tượng nghiên cứu: Thị trường lao động và cơ cấu nghề nghiệp tại Việt Nam.

Nội dung nghiên cứu: Tập trung vào việc phân tích các yếu tố ảnh hưởng, nhận diện xu hướng thay đổi, và đưa ra dự đoán về các nhóm ngành nghề trọng điểm như công nghệ thông tin, tự động hóa, năng lượng tái tạo, chăm sóc sức khỏe, giáo dục, và các ngành liên quan.

Thời gian nghiên cứu: Dự đoán xu hướng nghề nghiệp trong giai đoạn 2025-2030, dựa trên dữ liệu hiện tại và các yếu tố kinh tế - xã hội dự báo.

Không gian nghiên cứu: Chủ yếu tập trung vào thị trường lao động Việt Nam, nhưng có tham khảo các xu hướng toàn cầu để đảm bảo tính toàn diện.

### 1.4 Kết quả dự kiến đạt được

Nghiên cứu này kỳ vọng mang lại các kết quả sau:

Bức tranh toàn diện về xu hướng nghề nghiệp: Làm rõ các động lực chính thúc đẩy sự thay đổi nghề nghiệp tại Việt Nam trong 5 năm tới, từ tác động của công nghệ đến các yếu tố nhân khẩu học và kinh tế - xã hội.

Danh mục ngành nghề tiềm năng và rủi ro: Xác định các ngành nghề có triển vọng phát triển (như trí tuệ nhân tạo, năng lượng sạch, y tế số) và các ngành nghề có nguy cơ suy giảm (như sản xuất thủ công, công việc lặp lại dễ tự động hóa).

Bộ kỹ năng thiết yếu: Đưa ra danh sách các kỹ năng cần thiết cho người lao động, bao gồm kỹ năng công nghệ, tư duy sáng tạo, khả năng thích nghi, và các kỹ năng mềm như giao tiếp và làm việc nhóm.

Khuyến nghị thực tiễn: Đề xuất giải pháp cụ thể cho:

Cá nhân: Hướng dẫn lựa chọn ngành học, phát triển kỹ năng, và xây dựng lộ trình nghề nghiệp phù hợp với xu hướng thị trường.

Doanh nghiệp: Gợi ý chiến lược tuyển dụng, đào tạo, và tái đào tạo nhân sự để đáp ứng nhu cầu thị trường.

Nhà hoạch định chính sách: Đưa ra các đề xuất cải cách giáo dục, đào tạo nghề, và chính sách lao động nhằm thúc đẩy thị trường lao động bền vững và linh hoạt.

Những kết quả này không chỉ giúp các bên liên quan đưa ra quyết định sáng suốt mà còn đóng góp vào việc xây dựng một thị trường lao động năng động, cạnh tranh, và sẵn sàng cho những thay đổi trong tương lai.

# PHẦN NỘI DUNG

### CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

## 1.1 Tổng quan về thị trường lao động và xu hướng thay đổi nghề nghiệp.

#### 1.1.1 Khái niệm thị trường lao động

Thị trường lao động là một phạm trù kinh tế phức tạp, nơi diễn ra quá trình mua bán sức lao động giữa người có nhu cầu tuyển dụng (người sử dụng lao động) và người có nhu cầu tìm việc (người lao động). Thị trường lao động chịu sự tác động của nhiều yếu tố kinh tế, xã hội, chính trị và công nghệ, thể hiện qua các khía cạnh như cung và cầu lao động, tiền lương, cơ cấu việc làm, tỷ lệ thất nghiệp, và các chính sách liên quan đến lao động.

#### 1.1.2 Xu hướng thay đổi nghề nghiệp

Thị trường lao động luôn chuyển động không ngừng, chịu ảnh hưởng từ nhiều lực đẩy mạnh mẽ. Hiểu và phân tích các xu hướng này không chỉ giúp dự đoán tương lai nghề nghiệp mà còn định hướng chiến lược cho cả người lao động lẫn doanh nghiệp. Dưới đây là những yếu tố chính đang định hình bức tranh nghề nghiệp ngày nay:

Cách mạng công nghệ: Sự phát triển vượt bậc của trí tuệ nhân tạo (AI), tự động hóa, Internet vạn vật (IoT) và các công nghệ tiên tiến đang thay đổi cách chúng ta làm việc. Các ngành nghề mới như chuyên gia dữ liệu lớn (big data) hay kỹ sư trí tuệ nhân tạo đang nổi lên, trong khi những công việc lặp lại, ít đòi hỏi sáng tạo dần bị thay thế. Theo báo cáo của Diễn đàn Kinh tế Thế giới, đến năm 2030, các kỹ năng liên quan đến AI và big data sẽ nằm trong top 3 kỹ năng tăng trưởng nhanh nhất.

Toàn cầu hóa: Thế giới ngày càng kết nối, kéo theo sự cạnh tranh khốc liệt trên thị trường lao động toàn cầu. Các công ty không ngừng tìm kiếm nhân tài với chi phí hợp lý và kỹ năng vượt trội từ khắp nơi, đồng thời mở ra cơ hội lớn cho những ai sẵn sàng thích nghi.

Biến đổi nhân khẩu học: Sự thay đổi trong cơ cấu dân số, từ tỷ lệ sinh, tuổi thọ đến xu hướng di cư, đang tác động mạnh đến nguồn cung lao động và nhu cầu dịch vụ. Chẳng hạn, dân số già hóa làm tăng nhu cầu về chăm sóc y tế, nhưng lại khiến lực lượng lao động trẻ bị thu hẹp.

Môi trường và phát triển bền vững: Khi các vấn đề về biến đổi khí hậu ngày càng cấp bách, các ngành nghề “xanh” như năng lượng tái tạo, quản lý tài nguyên hay công nghệ môi trường đang phát triển mạnh mẽ, trở thành xu hướng tất yếu.

Kỳ vọng mới của người lao động: Thế hệ trẻ ngày nay ưu tiên công việc linh hoạt, có ý nghĩa và cơ hội phát triển cá nhân. Điều này buộc các doanh nghiệp phải đổi mới cách quản lý, xây dựng môi trường làm việc sáng tạo và hấp dẫn để giữ chân nhân tài.

Yếu tố kinh tế - xã hội: Chính sách công, tình hình kinh tế vĩ mô, khủng hoảng tài chính hay các yếu tố như văn hóa, giáo dục đều góp phần định hình thị trường lao động. Một nền kinh tế ổn định hay một chính sách giáo dục tiến bộ có thể mở ra những cơ hội nghề nghiệp mới đầy triển vọng.

Những xu hướng này không chỉ phản ánh sự thay đổi của thị trường lao động mà còn là lời nhắc nhở rằng việc thích nghi, học hỏi và đón đầu là chìa khóa để thành công trong một thế giới nghề nghiệp đầy biến động.

## 2.2 Vai trò của Big Data trong phân tích và dự đoán xu hướng nghề nghiệp.

#### 2.2.1 Khái niệm Big Data

Big data là các tập dữ liệu rất lớn và phức tạp, rất khó để quản lý, lưu trữ và phân tích bằng các công cụ xử lý dữ liệu truyền thống. Điều quan trọng của big data là khả năng phân tích và tìm hiểu thông tin từ những tập dữ liệu này, vì chúng thường chứa nhiều thông tin tiềm ẩn và giá trị quan trọng.

Ví dụ như phương pháp cá nhân hóa nội dung đề xuất cho mỗi người trên các nền tảng Spotify, Netflix, Youtube, các nền tảng Thương mại điện tử như Shopee, Lazada,...

#### 2.2.2 Ứng dụng của Big Data trong phân tích thị trường lao động

Big Data đang mở ra những cơ hội lớn để khám phá và hiểu sâu hơn về thị trường lao động, giúp nhận diện các xu hướng hiện tại và dự đoán những thay đổi trong tương lai một cách rõ nét:

Khám phá nhu cầu tuyển dụng: Bằng cách phân tích dữ liệu từ các trang web việc làm, mạng xã hội nghề nghiệp hay hồ sơ ứng viên, ta có thể xác định những vị trí công việc và kỹ năng đang được săn đón nhiều nhất.

Theo dõi xu hướng kỹ năng: Việc xem xét các tin tuyển dụng, mô tả công việc và thông tin từ người lao động giúp nhận ra những kỹ năng mới nổi, đồng thời phát hiện các kỹ năng dần mất đi tính cạnh tranh.

Dự báo thay đổi nghề nghiệp: Kết hợp dữ liệu lịch sử về thị trường lao động với các yếu tố như kinh tế, công nghệ và xã hội, ta có thể xây dựng mô hình để dự đoán sự xuất hiện hoặc suy giảm của các ngành nghề.

Đánh giá nguồn cung lao động: Thông tin về trình độ học vấn, kinh nghiệm làm việc và khu vực sinh sống của người lao động giúp làm rõ bức tranh về nguồn nhân lực hiện có và tiềm năng.

Hiểu rõ hơn về lương thưởng: Dữ liệu từ tin tuyển dụng hoặc khảo sát lương cung cấp cái nhìn chi tiết về mức lương, cơ cấu đãi ngộ và những yếu tố tác động đến thu nhập.

Nói chung, Big Data không chỉ giúp ta nhìn rõ hơn bức tranh toàn cảnh mà còn mang lại những giải pháp cụ thể, thiết thực cho cả người lao động và nhà tuyển dụng.

#### 2.2.3 Thách thức khi ứng dụng Big Data trong dự đoán nghề nghiệp

Dù Big Data mở ra nhiều cơ hội tuyệt vời để dự đoán xu hướng nghề nghiệp, việc áp dụng nó vào thực tế không hề dễ dàng và đối mặt với không ít rào cản:

Chất lượng dữ liệu chưa đảm bảo: Dữ liệu thu thập được đôi khi thiếu sót, không chính xác hoặc bị nhiễu bởi thông tin không liên quan, làm giảm độ tin cậy của các phân tích và dự báo.

Vấn đề về quyền riêng tư và đạo đức: Sử dụng dữ liệu cá nhân đòi hỏi phải tuân thủ nghiêm ngặt các quy định bảo vệ thông tin, đồng thời đảm bảo cách tiếp cận có trách nhiệm và minh bạch.

Yêu cầu cao về chuyên môn: Phân tích khối lượng dữ liệu khổng lồ cần đến những chuyên gia không chỉ giỏi về kỹ thuật mà còn có khả năng diễn giải kết quả một cách sắc bén và đúng ngữ cảnh.

Tính biến động của thị trường lao động: Thị trường lao động chịu ảnh hưởng từ nhiều yếu tố như kinh tế, công nghệ hay xã hội, vốn thay đổi nhanh và khó đoán, khiến các dự đoán dài hạn dễ bị lệch hướng.

Thiếu dữ liệu ở một số lĩnh vực: Không phải ngành nghề hay khu vực nào cũng có sẵn dữ liệu phong phú và chất lượng, đặc biệt là ở những lĩnh vực mới hoặc vùng sâu, vùng xa.

Tuy nhiên, với sự tiến bộ không ngừng của công nghệ và các phương pháp phân tích dữ liệu, những thách thức này đang dần được khắc phục. Big Data vẫn là một công cụ đầy tiềm năng, ngày càng khẳng định vai trò quan trọng trong việc nắm bắt và dự đoán những chuyển động của thị trường lao động, giúp định hình tương lai nghề nghiệp một cách rõ ràng hơn.

### CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU

## 3.1 Nguồn dữ liệu

Nghiên cứu dựa trên việc khai thác dữ liệu từ nhiều nguồn đa dạng để xây dựng bức tranh toàn diện về thị trường lao động tại Việt Nam và tham khảo các xu hướng toàn cầu. Các nguồn dữ liệu chính bao gồm:

Dữ liệu từ các nền tảng tuyển dụng trực tuyến: Thu thập thông tin từ các trang web việc làm phổ biến tại Việt Nam như VietnamWorks, JobStreet, và TopCV, cũng như các nền tảng quốc tế như LinkedIn. Các dữ liệu này bao gồm tin tuyển dụng, mô tả công việc, yêu cầu kỹ năng, mức lương, và thông tin về ngành nghề.

Báo cáo và thống kê chính thức: Sử dụng dữ liệu từ các cơ quan như Tổng cục Thống kê Việt Nam (GSO), Bộ Lao động - Thương binh và Xã hội, và các tổ chức quốc tế như Tổ chức Lao động Quốc tế (ILO) hay Diễn đàn Kinh tế Thế giới (WEF). Các báo cáo này cung cấp thông tin về tỷ lệ việc làm, cơ cấu ngành nghề, và xu hướng nhân khẩu học.

Dữ liệu mạng xã hội nghề nghiệp: Phân tích nội dung từ các mạng xã hội như LinkedIn để nhận diện xu hướng kỹ năng, các ngành nghề mới nổi, và hành vi tìm kiếm việc làm của người lao động.

Khảo sát và phỏng vấn: Tiến hành khảo sát nhỏ với các nhóm đối tượng như sinh viên, người lao động, và nhà tuyển dụng để thu thập thông tin định tính về kỳ vọng nghề nghiệp, kỹ năng cần thiết, và những thách thức họ đang đối mặt.

Nguồn dữ liệu mở và nghiên cứu học thuật: Khai thác các bài báo, nghiên cứu từ các trường đại học, và cơ sở dữ liệu mở liên quan đến thị trường lao động, công nghệ, và chuyển đổi số tại Việt Nam.

Việc sử dụng đa dạng nguồn dữ liệu giúp đảm bảo tính phong phú và độ tin cậy, đồng thời giảm thiểu rủi ro từ việc phụ thuộc vào một nguồn duy nhất.

## 3.2 Quy trình xử lý dữ liệu

Quy trình xử lý dữ liệu được thiết kế chặt chẽ để chuyển đổi dữ liệu thô thành thông tin có giá trị, phục vụ cho việc phân tích và dự đoán. Các bước chính bao gồm:

Thu thập dữ liệu: Sử dụng các công cụ như API của các trang tuyển dụng, phần mềm thu thập dữ liệu web (web scraping), và khảo sát trực tiếp để thu thập dữ liệu từ các nguồn đã nêu. Dữ liệu được lưu trữ ở định dạng có cấu trúc (như CSV, JSON) để dễ dàng xử lý.

Làm sạch dữ liệu: Loại bỏ dữ liệu trùng lặp, xử lý các giá trị thiếu hoặc không chính xác, và chuẩn hóa định dạng (ví dụ: thống nhất cách viết tên ngành nghề, mức lương). Các kỹ thuật như lọc nhiễu và kiểm tra tính toàn vẹn được áp dụng để đảm bảo chất lượng dữ liệu.

Tích hợp dữ liệu: Kết hợp dữ liệu từ nhiều nguồn khác nhau để tạo ra một tập dữ liệu thống nhất. Ví dụ, thông tin về kỹ năng từ LinkedIn được đối chiếu với yêu cầu công việc từ VietnamWorks để nhận diện sự tương đồng và khác biệt.

Mã hóa và phân loại: Chuyển đổi dữ liệu định tính (như mô tả công việc) thành dạng định lượng bằng cách sử dụng các kỹ thuật như phân tích từ khóa, mã hóa chủ đề, hoặc gán nhãn ngành nghề. Điều này giúp dữ liệu sẵn sàng cho các phương pháp phân tích nâng cao.

Lưu trữ và quản lý: Dữ liệu đã xử lý được lưu trữ trên cơ sở dữ liệu đám mây hoặc hệ thống cục bộ, đảm bảo dễ truy cập và bảo mật. Các công cụ như SQL hoặc MongoDB được sử dụng để quản lý hiệu quả.

Quy trình này không chỉ giúp dữ liệu trở nên “sạch” và có ý nghĩa mà còn tạo nền tảng vững chắc cho các phân tích sâu hơn.

### CHƯƠNG 3: TỔNG QUAN DỰ ÁN

Dự án xây dựng một hệ thống phân tích và tìm kiếm việc làm dựa trên Big Data, tập trung vào thị trường lao động Việt Nam. Hệ thống tích hợp các công nghệ hiện đại để thu thập, xử lý, lưu trữ, trực quan hóa, và cung cấp dữ liệu việc làm từ hai nguồn chính: VietnamWorks (vietnamworks.com) và CareerViet (careerviet.vn). Dự án bao gồm các giai đoạn từ thiết lập hạ tầng, thu thập dữ liệu, xử lý dữ liệu lớn, đến phát triển ứng dụng web và trực quan hóa kết quả, nhằm cung cấp thông tin việc làm chính xác và thân thiện cho người dùng.

## 3.1 Thiết lập hạ tầng

Hạ tầng Big Data và các công cụ liên quan được cài đặt trên hai môi trường: Ubuntu (cho Hadoop, Hive) và Windows (cho Hadoop, Sqoop).

* Trên Ubuntu:
  + Hadoop: Cài đặt Hadoop để sử dụng HDFS làm hệ thống lưu trữ phân tán và MapReduce cho xử lý dữ liệu lớn.
  + Hive: Cài đặt Hive để quản lý và truy vấn dữ liệu trên HDFS, sử dụng định dạng ORC và hỗ trợ ACID cho các thao tác cập nhật.
  + MySQL: Sử dụng làm Hive Metastore để lưu trữ metadata của các bảng Hive, đảm bảo tính nhất quán và dễ quản lý.
  + HiveServer2: Bật dịch vụ HiveServer2 (cổng 10000) để hỗ trợ truy vấn từ xa thông qua JDBC/ODBC.
  + SSH: Cấu hình SSH để các thành viên nhóm truy cập từ xa vào cụm Hadoop/Hive, đảm bảo bảo mật và quản lý dễ dàng.
* Trên Windows:
  + Hadoop: Cài đặt Hadoop để hỗ trợ các tác vụ bổ sung, đồng bộ với cụm trên Ubuntu.
  + Sqoop: Cài đặt Sqoop để chuyển dữ liệu giữa HDFS/Hive và cơ sở dữ liệu quan hệ, sử dụng MySQL làm Sqoop Metastore để lưu trữ metadata của các job Sqoop.

## 3.2 Thu thập dữ liệu

Dữ liệu việc làm được thu thập từ hai nền tảng tuyển dụng trực tuyến:

* VietnamWorks: Sử dụng Python và thư viện requests để gửi yêu cầu POST tới API (https://ms.vietnamworks.com/job-search/v1.0/search). Thu thập 219 trang, mỗi trang 50 bản ghi, bao gồm các trường như jobId, jobTitle, companyName, salaryMin, salary, approvedOn, expiredOn, benefits, workingLocations, companyLogo, salaryCurrency. Dữ liệu được lưu vào file JSON (vietnamworks.json).
* CareerViet: Sử dụng requests.Session để duy trì cookies, truy cập trang HTML (https://careerviet.vn/viec-lam/tat-ca-viec-lam-trang-{page}-vi.html) và gọi API (https://careerviet.vn/search-jobs). Thu thập 567 trang, bao gồm các trường như JOB\_ID, JOB\_TITLE, EMP\_NAME, JOB\_FROMSALARY\_CVR, BENEFIT\_NAME, LOCATION\_NAME\_ARR. Dữ liệu lưu vào file JSON (careerviet.json).

## 3.3 Xử lý và chuẩn bị dữ liệu

Dữ liệu thô được làm sạch, gộp, và chuẩn hóa để đảm bảo chất lượng và sẵn sàng cho lưu trữ:

* Làm sạch dữ liệu:
  + Sử dụng Python với thư viện json, pandas, và datetime để xử lý file JSON.
  + Chuẩn hóa định dạng:
    - Ngày tháng (approvedOn, expiredOn) sang YYYY-MM-DD (ví dụ: từ 2025-05-01T00:00:00Z hoặc 01-05-2025).
    - Trích xuất lợi ích (benefitNames) và thành phố (cities) từ mảng dữ liệu.
    - Chuẩn hóa tiền tệ (salaryCurrency) thành VND.
  + Loại bỏ giá trị thiếu hoặc không hợp lệ.
* Gộp và loại bỏ trùng lặp:
  + Gộp dữ liệu từ vietnamworks.json và careerviet.json thành một danh sách duy nhất.
  + Loại bỏ trùng lặp dựa trên jobId và bộ khóa (jobTitle, companyName, cities, salaryMin, salary, benefitNames).
  + Lưu kết quả vào file CSV (merge.csv).
* Chuẩn hóa dữ liệu:
  + Chia dữ liệu thành các bảng quan hệ:
    - job: Thông tin công việc (jobId, jobTitle, jobUrl, salaryMin, salary, approvedOn, expiredOn, salaryCurrency, company\_id).
    - benefit: Lợi ích (benefit\_id, benefit).
    - city: Thành phố (city\_id, city).
    - company: Công ty (company\_id, companyName, companyLogo).
    - job\_benefit: Liên kết công việc-lợi ích (jobId, benefit\_id).
    - job\_city: Liên kết công việc-thành phố (jobId, city\_id).
  + Lưu các bảng vào file CSV (job.csv, benefit.csv, city.csv, company.csv, job\_benefit.csv, job\_city.csv).

## 3.4 Lưu trữ dữ liệu

Dữ liệu được lưu trữ trên nhiều hệ thống để đảm bảo khả năng truy cập và phân tích:

* HDFS:
  + Các file CSV (job.csv, benefit.csv, city.csv, company.csv, job\_benefit.csv, job\_city.csv) được tải lên HDFS tại thư mục /user/b2hhduser/data/.
* Đầu vào và đầu ra của các chương trình MapReduce:
* Đầu vào: Dữ liệu từ các bảng Hive (work.job, work.benefit, work.city, work.company, work.job\_benefit, work.job\_city, merge\_job) được lưu trữ tại /user/hive/warehouse/work.db/ (thư mục mặc định của database work trong Hive).
* Đầu ra: Kết quả từ 7 chương trình MapReduce được lưu trữ trên HDFS dưới dạng file phân tán tại thư mục /user/b2hhduser/, với các thư mục con cụ thể cho từng chương trình.
* Cơ sở dữ liệu Hive: Các bảng Hive (work.job, work.benefit, work.city, work.company, work.job\_benefit, work.job\_city, merge\_job) được lưu trữ trên HDFS tại /user/hive/warehouse/work.db/ (thư mục mặc định của database work trong Hive).
* Hive:
  + Tạo database work và các bảng Hive (work.job, work.benefit, work.city, work.company, work.job\_benefit, work.job\_city, merge\_job) ánh xạ đến các file CSV trên HDFS.
* MongoDB:
  + Dữ liệu từ các file CSV như job.csv, company.csv, job\_city.csv, job\_benefit.csv, city.csv, và benefit.csv được nhập vào MongoDB để hỗ trợ lưu trữ NoSQL. MongoDB cung cấp khả năng lưu trữ dữ liệu dưới dạng tài liệu JSON, giúp truy vấn linh hoạt và dễ dàng tích hợp với các ứng dụng khác.
  + Sử dụng Python và thư viện pymongo, dữ liệu từ các file CSV này được đọc vào, sau đó xóa dữ liệu cũ trong MongoDB và lưu trữ

## 3.5 Sao lưu và phục hồi dữ liệu

Dữ liệu Hive được sao lưu và phục hồi để đảm bảo an toàn:

* Sao lưu:
  + Sao lưu các bảng Hive bằng cách xuất dữ liệu từ HDFS sang thư mục
* Phục hồi:
  + Phục hồi dữ liệu HDFS từ thư mục sao lưu

## 3.6 Thao tác CRUD trên MySQL Metastore của Hive

Thực hiện các thao tác CRUD (Create, Read, Update, Delete) trên MySQL Hive Metastore để quản lý metadata:

* **Create**: Tạo dữ liệu mới trong bảng.
* **Read**: Truy vấn thông tin và dữ liệu từ bảng.
* **Update**: Cập nhật dữ liệu hiện có trong bảng.
* **Delete**: Xóa dữ liệu khỏi bảng.

## 3.7 Xử lý dữ liệu với MapReduce

Phát triển 7 chương trình MapReduce bằng Python để xử lý dữ liệu từ các bảng Hive:

* Các chương trình xử lý các tác vụ như:
  + job\_salary\_analysis: Mức lương tối thiểu trung bình, mức lương tối đa trung bình, và đơn vị tiền tệ cho mỗi tiêu đề công việc, với định dạng: jobtitle\tavg\_min\_salary\tavg\_max\_salary\tcurrency.
  + job\_count\_by\_company: Tổng số công việc cho mỗi công ty, với định dạng: company\_id\tjob\_count.
  + benefit\_popularity: Số lần xuất hiện của mỗi benefit\_id (mã phúc lợi), với định dạng benefit\_id\tcount.
  + job\_distribution\_by\_city: Số lượng công việc cho mỗi city\_id (mã thành phố), với định dạng: city\_id\tjob\_count.
  + salary\_range\_analysis: Số lượng công việc cho mỗi salary\_range (mức lương), với định dạng: salary\_range\tcount.
  + job\_posting\_duration: Thông tin của mỗi công việc với định dạng: jobid\tjobtitle\tduration.
  + currency\_distribution: Số lượng công việc và phần trăm tương ứng cho mỗi loại tiền tệ, với định dạng: currency\tcount\tpercentage%.
* Dữ liệu đầu vào từ các bảng Hive (merge\_job, job, v.v.).
* Kết quả được lưu lại trên HDF.

## 3.8 Phát triển ứng dụng web

Xây dựng ứng dụng web TopViecLam để hỗ trợ tìm kiếm việc làm:

* Frontend:
  + Sử dụng HTML, CSS, JavaScript, với Bootstrap cho thiết kế đáp ứng và Select2 cho trường chọn đa lựa.
  + Giao diện bao gồm:
    - Biểu mẫu tìm kiếm: Nhập từ khóa, chọn thành phố, lợi ích.
    - Danh sách công việc: Hiển thị thẻ công việc với thông tin như tiêu đề, công ty, lương, địa điểm, lợi ích, thời gian.
    - Phân trang: Hỗ trợ điều hướng qua các trang kết quả.
* Backend:
  + Sử dụng Flask để xây dựng API RESTful, kết nối với HiveServer2 qua PyHive.
  + Các endpoint chính:
    - /api/jobs: Lọc và phân trang công việc theo từ khóa, lương, thành phố, lợi ích.
    - /api/job\_titles, /api/companies: Gợi ý tiêu đề công việc và công ty.
    - /api/benefits, /api/cities: Cung cấp danh sách lợi ích và thành phố.
  + Truy vấn bảng merge\_job với điều kiện lọc động, phân trang thực hiện phía Python.

## 3.9 Trực quan hóa dữ liệu với PowerBI

Sử dụng PowerBI để trực quan hóa dữ liệu từ Hive:

* Kết nối PowerBI với HiveServer2 thông qua ODBC Data Sources, sử dụng driver Hive ODBC.
* Tạo các biểu đồ và bảng biểu:
  + Bảng danh sách công ty, vị trí tuyển dụng và địa chỉ làm việc, như "Công Ty Cổ Phần Sữa Đà Lạt" với vị trí "Chuyên Viên Tiến Lương và Phúc Lợi" tại "Đà Lạt – Lâm Đồng".
  + Biểu đồ top loại việc làm hiển thị các vị trí phổ biến như "Nhân Viên Kinh Doanh" và "Kế Toán Tổng Hợp" được sắp về giảm dần.
  + Bản đồ địa điểm làm việc thể hiện phân bố công việc trên khắp Việt Nam, tập trung chủ yếu tại Hà Nội, Hồ Chí Minh và Đà Nẵng.
  + Biểu đồ top số cơ hội công việc cho thấy Hà Nội và Hồ Chí Minh dẫn đầu về số lượng cơ hội, với hàng chục nghìn vị trí.
* Dữ liệu được truy vấn trực tiếp từ bảng merge\_job và các bảng liên quan, đảm bảo tính tức thời.

Tóm tắt: Dự án xây dựng hệ thống phân tích và tìm kiếm việc làm, sử dụng dữ liệu từ VietnamWorks và CareerViet. Hạ tầng bao gồm Hadoop, Hive trên Ubuntu, Sqoop trên Windows, với MySQL làm Metastore. Dữ liệu được thu thập qua API, làm sạch, chuẩn hóa, lưu trữ trên HDFS, Hive, và MongoDB. Các chương trình MapReduce xử lý dữ liệu lớn, ứng dụng web TopViecLam hỗ trợ tìm kiếm, và PowerBI trực quan hóa kết quả. Quy trình sao lưu/phục hồi và thao tác CRUD trên MySQL đảm bảo an toàn dữ liệu.

### CHƯƠNG 4: TRIỂN KHAI KỸ THUẬT

## 4.1 Dữ liệu

#### 4.1.1 Thu thập dữ liệu

Dữ liệu việc làm được thu thập từ hai trang web tuyển dụng lớn tại Việt Nam: VietnamWorks (vietnamworks.com) và CareerViet (careerviet.vn). Mục tiêu là thu thập thông tin chi tiết về công việc, bao gồm tiêu đề, công ty, mức lương, địa điểm, ngày đăng, ngày hết hạn, và lợi ích.

* Thu thập từ VietnamWorks:
  + Sử dụng Python và thư viện requests để gửi yêu cầu POST tới API của VietnamWorks (https://ms.vietnamworks.com/job-search/v1.0/search).
  + Thiết lập tiêu đề HTTP (user-agent, accept, content-type) để mô phỏng trình duyệt, đảm bảo truy cập thành công.
  + Gửi payload JSON với tham số lấy 50 bản ghi mỗi trang, lặp qua 219 trang để thu thập các trường như jobTitle, companyName, salaryMin, salary, approvedOn, expiredOn, benefits, workingLocations.
  + Lưu dữ liệu thô dưới dạng file JSON (vietnamworks.json).
* Thu thập từ CareerViet:
  + Sử dụng Python với requests.Session để duy trì cookies, truy cập trang HTML (https://careerviet.vn/viec-lam/tat-ca-viec-lam-trang-{page}-vi.html) trước khi gọi API (https://careerviet.vn/search-jobs).
  + Lặp qua 567 trang, gửi yêu cầu POST với payload chứa số trang, thu thập các trường như JOB\_ID, JOB\_TITLE, EMP\_NAME, JOB\_FROMSALARY\_CVR, BENEFIT\_NAME, LOCATION\_NAME\_ARR.
  + Lưu dữ liệu thô vào file JSON (careerviet.json).
* Kết quả:
  + Thu thập hàng chục nghìn bản ghi công việc từ cả hai nguồn, lưu dưới dạng JSON, cung cấp dữ liệu phong phú để phân tích thị trường lao động.

#### 4.1.2 Làm sạch và chuẩn bị dữ liệu

Dữ liệu thô (JSON) từ VietnamWorks và CareerViet có nhiều vấn đề như định dạng không đồng nhất, giá trị thiếu, hoặc trùng lặp. Tôi đã làm sạch, chuẩn hóa, gộp, và loại bỏ trùng lặp để chuẩn bị dữ liệu cho bước lưu trữ.

* Làm sạch dữ liệu VietnamWorks:
  + Đọc file vietnamworks.json bằng Python và thư viện json, lọc các trường cần thiết: jobId, jobTitle, jobUrl, companyName, salaryMin, salary, approvedOn, expiredOn, benefitNames, cities, companyLogo, salaryCurrency.
  + Chuẩn hóa:
    - Chuyển đổi ngày tháng (approvedOn, expiredOn) từ định dạng ISO (2025-05-01T00:00:00Z) sang YYYY-MM-DD (ví dụ: 2025-05-01).
    - Trích xuất danh sách lợi ích (benefitNames) từ benefits và danh sách thành phố (cities) từ workingLocations.
  + Lưu dữ liệu sạch vào file JSON mới (vietnamworks.json).
* Làm sạch dữ liệu CareerViet:
  + Đọc file careerviet.json, ánh xạ các trường (như JOB\_ID thành jobId, JOB\_TITLE thành jobTitle) để đồng nhất với VietnamWorks.
  + Chuẩn hóa:
    - Chuyển đổi ngày tháng (approvedOn, expiredOn) từ DD-MM-YYYY (ví dụ: 01-05-2025) sang YYYY-MM-DD (ví dụ: 2025-05-01).
    - Chuẩn hóa đơn vị tiền tệ (salaryCurrency) thành chữ in hoa (ví dụ: VND).
  + Lưu dữ liệu sạch vào file JSON mới (careerviet.json).
* Gộp và loại bỏ trùng lặp:
  + Gộp hai file JSON (vietnamworks.json, careerviet.json) thành một danh sách duy nhất bằng Python.
  + Loại bỏ trùng lặp dựa trên:
    - Mã công việc (jobId) để bỏ các bản ghi trùng trực tiếp.
    - Bộ khóa (jobTitle, companyName, cities, salaryMin, salary, benefitNames) để phát hiện bản ghi tương tự.
  + Chuẩn hóa văn bản bằng cách loại bỏ khoảng trắng thừa, sắp xếp dữ liệu theo approvedOn giảm dần.
  + Chuyển dữ liệu gộp sang file CSV (merge.csv) với các cột: jobId, jobTitle, jobUrl, companyName, salaryMin, salary, approvedOn, expiredOn, benefitNames, cities, companyLogo, salaryCurrency.
* Chuẩn hóa dữ liệu:
  + Dùng pandas để chia dữ liệu thành các bảng quan hệ:
    - Bảng job: Thông tin công việc (jobId, jobTitle, jobUrl, salaryMin, salary, approvedOn, expiredOn, salaryCurrency, company\_id).
    - Bảng benefit: Lợi ích duy nhất (benefit\_id, benefit).
    - Bảng city: Thành phố duy nhất (city\_id, city).
    - Bảng company: Công ty duy nhất (company\_id, companyName, companyLogo).
    - Bảng job\_benefit: Liên kết công việc-lợi ích (jobId, benefit\_id).
    - Bảng job\_city: Liên kết công việc-thành phố (jobId, city\_id).
  + Loại bỏ khoảng trắng thừa trong benefitNames, cities, gán khóa ngoại (company\_id, benefit\_id, city\_id).
  + Lưu các bảng vào file CSV riêng (job.csv, benefit.csv, city.csv, company.csv, job\_benefit.csv, job\_city.csv) với mã hóa utf-8-sig để hỗ trợ tiếng Việt.
* Kết quả:
  + Từ hàng chục nghìn bản ghi JSON, sau khi làm sạch, gộp, và chuẩn hóa, thu được các bảng CSV quan hệ, sẵn sàng để lưu trữ và phân tích trong bước tiếp theo.

## 4.2 Lưu trữ

Dữ liệu việc làm đã được làm sạch và chuẩn hóa thành các bảng quan hệ (job, benefit, city, company, job\_benefit, job\_city, merge\_job) dưới dạng CSV trong mục 4.1. Trong mục này, dữ liệu được lưu trữ trên HDFS (Hadoop Distributed File System) tại thư mục /user/b2hhduser/data/ và tích hợp vào Hive để hỗ trợ phân tích và truy vấn.

#### 4.2.1 Lưu trữ dữ liệu trên HDFS

HDFS được sử dụng làm hệ thống lưu trữ phân tán để đảm bảo khả năng mở rộng và độ tin cậy cho dữ liệu lớn.

* Quy trình lưu trữ:
  + Các file CSV (job.csv, benefit.csv, city.csv, company.csv, job\_benefit.csv, job\_city.csv, merge\_job.csv) được tạo từ quá trình làm sạch và chuẩn hóa dữ liệu.
  + Thư mục /user/b2hhduser/data/ được tổ chức để lưu trữ tất cả các bảng dữ liệu, đảm bảo dễ dàng quản lý và truy cập.
* Kết quả:
  + Dữ liệu được lưu trữ an toàn trên HDFS, phân tán trên các node của cụm Hadoop (nếu dùng multi-node), sẵn sàng cho xử lý và phân tích.

#### 4.2.2 Tích hợp dữ liệu với Hive

* Tích hợp với Hive: Quá trình tích hợp lên Hive bắt đầu bằng việc cấu hình Hive để hỗ trợ giao dịch với SET hive.support.concurrency=true; và SET hive.txn.manager=org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;, sau đó tạo mới hai cơ sở dữ liệu work\_load, work này. Trong work\_load, các bảng job, benefit, city, company, job\_benefit, job\_city, và merge\_job được tạo với OpenCSVSerde, lưu dưới dạng TEXTFILE, bỏ qua dòng tiêu đề CSV. Tương tự, các bảng trong work được tạo với định dạng ORC, phân cụm (4 hoặc 8 bucket) và hỗ trợ giao dịch. Dữ liệu CSV từ HDFS được tải vào work\_load bằng LOAD DATA INPATH, sau đó chuyển sang work bằng INSERT INTO. Riêng bảng merge\_job tổng hợp dữ liệu từ nhiều bảng bằng LEFT JOIN, sử dụng collect\_set để tạo mảng benefitNames và cities, đồng thời ép kiểu salaryMin, salary sang BIGINT. Kết quả, work\_load lưu dữ liệu thô, còn work chứa dữ liệu tối ưu hóa, với merge\_job cung cấp thông tin tổng hợp về công việc, lợi ích và thành phố.

## 4.3 Triển khai trên môi trường Windows

## 4.4 Chức năng

#### 4.4.1. Hỗ trợ thực thi truy vấn (query), thao tác với dữ liệu (CRUD)

Quá trình hỗ trợ thực thi truy vấn và thao tác dữ liệu (CRUD) trên Hive được thực hiện thông qua các lệnh SQL để chèn (INSERT), truy vấn (SELECT), cập nhật (UPDATE) và xóa (DELETE) dữ liệu. Đầu tiên, lệnh INSERT được sử dụng để thêm bản ghi mới vào các bảng: ví dụ, thêm một công việc với jobId='30000', một lợi ích với benefit\_id=30 và một công ty với company\_id=100 vào các bảng job, benefit và company. Tiếp theo, các truy vấn SELECT cho phép lấy dữ liệu, như truy xuất thông tin công việc với jobId='30000', lợi ích với benefit\_id=30, hoặc thông tin công ty với company\_id=100. Một truy vấn phức tạp hơn tổng hợp dữ liệu từ nhiều bảng (job, company, job\_benefit, benefit, job\_city, city) bằng LEFT JOIN và collect\_set để tạo mảng benefitNames và cities. Lệnh UPDATE được dùng để sửa đổi dữ liệu, như cập nhật lợi ích từ "Nghỉ phép 10 ngày" thành "Nghỉ phép 20 ngày" hoặc đổi tên công ty từ "Nhà báo dân trí" thành "Công ty dân trí". Cuối cùng, lệnh DELETE xóa các bản ghi, như xóa công việc với jobId='30000', lợi ích với benefit\_id=30 hoặc công ty với company\_id=100 và tên "Nhà báo dân trí". Các thao tác này tận dụng cấu hình giao dịch của Hive để đảm bảo tính toàn vẹn dữ liệu trong môi trường phân tán.

#### 4.4.2. Sao lưu, phục hồi dữ liệu

Quá trình sao lưu và phục hồi dữ liệu trên HDFS được thực hiện hiệu quả với các lệnh HDFS. Để sao lưu, lệnh hdfs dfs -cp /user/hive/warehouse/work\_load.db /backup/work\_load.db\_backup được sử dụng để sao chép toàn bộ thư mục dữ liệu của cơ sở dữ liệu work\_load từ thư mục /user/hive/warehouse/ sang thư mục sao lưu /backup/work\_load.db\_backup. Khi cần phục hồi, lệnh hdfs dfs -cp -f /backup/work\_load.db\_backup/\* /user/hive/warehouse/work\_load.db/ được thực thi để sao chép dữ liệu từ thư mục sao lưu trở lại thư mục gốc, với tùy chọn -f để ghi đè nếu cần. Quy trình này đảm bảo dữ liệu trong Hive được bảo vệ và có thể khôi phục nhanh chóng khi xảy ra sự cố.

#### 4.4.3. Trực quan dữ liệu thu thập

Dữ liệu thu thập được trực quan hóa thông qua Power BI, kết nối với HiveServer2 bằng ODBC data source, mang lại giao diện trực quan và dễ hiểu. Giao diện bao gồm bảng danh sách công ty, vị trí tuyển dụng và địa chỉ làm việc. Biểu đồ top loại việc làm hiển thị các tiêu đề việc làm phổ biến như "Nhân Viên Kinh Doanh" và "Kế Toán Tổng Hợp". Biểu đồ top nơi có nhiều việc làm hiện thị các khu vực có việc làm tuyển dụng được sắp xếp giảm dần. Bản đồ địa điểm làm việc thể hiện phân bố công việc trên khắp Việt Nam, tập trung chủ yếu tại Hà Nội, Hồ Chí Minh và Đà Nẵng. Power BI tận dụng kết nối ODBC với HiveServer2 để truy vấn dữ liệu từ các bảng Hive, cho phép tổng hợp và phân tích dữ liệu hiệu quả, hỗ trợ người dùng đưa ra quyết định dựa trên thông tin chi tiết và trực quan.

## 4.5 Giao diện

Đã phát triển một ứng dụng web có tên TopViecLam để hỗ trợ người dùng tìm kiếm việc làm dựa trên dữ liệu lưu trữ trong Hive. Ứng dụng bao gồm giao diện người dùng (frontend) sử dụng HTML, CSS, JavaScript, và hệ thống xử lý yêu cầu (backend) sử dụng Flask, kết nối với HiveServer2 để truy vấn dữ liệu. Giao diện cung cấp các tính năng tìm kiếm, lọc, gợi ý tự động, và phân trang, mang lại trải nghiệm thân thiện và hiệu quả.

#### 4.5.1 Giao diện người dùng (Frontend)

Giao diện web được thiết kế trực quan, đáp ứng (responsive), và tập trung vào việc hỗ trợ người dùng tra cứu công việc phù hợp. Các thành phần chính bao gồm:

* Phần tìm kiếm:
  + Một biểu mẫu tìm kiếm cho phép nhập các tiêu chí:
    - Từ khóa: Nhập tên công việc hoặc công ty (ví dụ: "Kỹ sư phần mềm", "Công ty ABC"). Hệ thống hiển thị gợi ý tự động (autocomplete) dựa trên tiêu đề công việc hoặc tên công ty khi người dùng nhập.
    - Địa điểm: Chọn nhiều thành phố từ danh sách động (ví dụ: Hà Nội, Hồ Chí Minh, Đà Nẵng) thông qua trường chọn đa lựa (select2), được tải từ bảng city.
    - Phúc lợi: Chọn nhiều lợi ích từ danh sách động (ví dụ: Chế độ bảo hiểm, Du lịch, Thưởng) thông qua trường chọn đa lựa, được tải từ bảng benefit.
  + Các nút chức năng:
    - Nút "Tìm kiếm": Gửi yêu cầu tìm kiếm dựa trên các tiêu chí đã chọn.
    - Nút "Đặt lại": Xóa tất cả tiêu chí để bắt đầu lại.
* Phần hiển thị kết quả:
  + Kết quả được hiển thị dưới dạng thẻ công việc (job cards), mỗi thẻ bao gồm:
    - Logo công ty và tên công ty: Từ trường companyLogo và companyName.
    - Tiêu đề công việc: Ví dụ: "Assistant Packaging Manager", "Nhân viên content".
    - Địa điểm: Danh sách thành phố (ví dụ: Hà Nội, Vĩnh Phúc) từ trường cities.
    - Mức lương: Khoảng lương (ví dụ: 10-20 triệu VNĐ/tháng) từ salaryMin và salary.
    - Phúc lợi: Danh sách lợi ích (ví dụ: Chế độ bảo hiểm, Du lịch, Đào tạo) từ benefitNames.
    - Thời gian: Ngày đăng (approvedOn) và ngày hết hạn (expiredOn), hiển thị dạng tương đối (ví dụ: "Đăng 5 ngày trước", "Hết hạn sau 25 ngày").
  + Số lượng công việc tìm thấy được hiển thị ở đầu danh sách (ví dụ: "3 công việc").
* Phân trang:
  + Hỗ trợ phân trang để quản lý số lượng lớn công việc, mỗi trang hiển thị một số bản ghi cố định (ví dụ: 10 công việc).
  + Thanh phân trang cho phép chuyển đổi giữa các trang (ví dụ: trang 1, 289, ..., 293), với nút điều hướng (trước, sau) và hiển thị các trang lân cận.
* Công nghệ sử dụng:
  + HTML và CSS: Tạo cấu trúc và giao diện, sử dụng Bootstrap để đảm bảo hiển thị tốt trên mọi thiết bị.
  + JavaScript: Xử lý tương tác, gọi API, và cập nhật giao diện động. Sử dụng Select2 cho các trường chọn đa lựa (địa điểm, phúc lợi) và gợi ý tìm kiếm.
  + Font Awesome: Cung cấp biểu tượng (icon) cho các nút và trường nhập liệu (ví dụ: kính lúp, bản đồ).

#### 4.5.2 Xử lý yêu cầu (Backend)

Backend được xây dựng bằng Flask, xử lý các yêu cầu từ giao diện và truy vấn dữ liệu từ Hive thông qua HiveServer2.

* API Flask:
  + Tạo các endpoint RESTful:
    - /api/jobs: Lấy danh sách công việc dựa trên các bộ lọc (jobTitle, companyName, salaryMin, salary, salaryCurrency, benefitNames, cities), hỗ trợ phân trang (page, items\_per\_page).
    - /api/job\_titles: Lấy gợi ý tiêu đề công việc dựa trên từ khóa (job\_title), giới hạn số lượng bản ghi (limit).
    - /api/companies: Lấy gợi ý công ty dựa trên tên (company\_name), giới hạn số lượng bản ghi (limit).
    - /api/benefits: Lấy danh sách tất cả lợi ích từ bảng benefit.
    - /api/cities: Lấy danh sách tất cả thành phố từ bảng city.
  + Các hàm xử lý:
    - Hàm get\_jobs:
      * Xây dựng truy vấn Hive SQL trên bảng merge\_job, chứa các trường jobId, jobTitle, jobUrl, companyName, salaryMin, salary, approvedOn, expiredOn, benefitNames, cities, companyLogo, salaryCurrency.
      * Tạo điều kiện lọc động:
        + Tìm kiếm từ khóa: lower(jobTitle) LIKE '%{keyword}%' hoặc lower(companyName) LIKE '%{keyword}%'.
        + Lọc lương: salaryMin >= {value} hoặc salary >= {value}.
        + Lọc tiền tệ: salaryCurrency = '{currency}'.
        + Lọc lợi ích: array\_contains(benefitNames, '{benefit}').
        + Lọc thành phố: array\_contains(cities, '{city}').
      * Thực thi truy vấn để lấy toàn bộ dữ liệu, sau đó phân trang phía Python bằng cách cắt danh sách kết quả (offset, items\_per\_page).
      * Trả về JSON chứa danh sách công việc và số trang tổng (total\_pages).
    - Hàm get\_job\_titles: Truy vấn bảng merge\_job để lấy gợi ý tiêu đề công việc, lọc theo từ khóa và giới hạn số lượng.
    - Hàm get\_companies: Truy vấn bảng company để lấy gợi ý công ty, lọc theo tên và giới hạn số lượng.
    - Hàm get\_benefits và get\_cities: Truy vấn bảng benefit và city để cung cấp danh sách cho các trường chọn đa lựa.
* Kết nối HiveServer2:
  + Sử dụng hàm get\_hive\_connection (dùng PyHive để kết nối với HiveServer2 (cổng 10000).

Thực thi truy vấn SQL trên bảng merge\_job (chứa dữ liệu gộp từ các bảng quan hệ)

* + Xử lý ngoại lệ (lỗi kết nối, truy vấn không hợp lệ) và trả về thông báo lỗi rõ ràng.
* Hiệu suất:
  + Tối ưu truy vấn bằng cách sử dụng bảng merge\_job (gộp sẵn dữ liệu) thay vì nhiều phép nối (JOIN).
  + Phân trang thực hiện phía Python để đơn giản hóa truy vấn Hive, phù hợp với dữ liệu vừa phải.
* Kết quả:
  + Backend xử lý hiệu quả các yêu cầu tìm kiếm, lọc, gợi ý, và phân trang, trả về dữ liệu chính xác từ Hive, tích hợp mượt mà với giao diện frontend.

# PHẦN KẾT LUẬN

# DANH MỤC TÀI LIỆU THAM KHẢO

[1]. Big data là gì? Đặc điểm & ứng dụng Big data vào các ngành, truy cập ngày 11/5/2025, <https://www.pace.edu.vn/tin-kho-tri-thuc/big-data-la-gi#big-data-la-gi>

[2]. The Future of Jobs 2025, truy cập ngày 11/5/2025, <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>

[3].