**ĐẠI HỌC ĐÀ NẴNG**

**TRƯỜNG ĐẠI HỌC KINH TẾ**



**BÁO CÁO THỰC TẬP TỐT NGHIỆP**

**Chuyên ngành: Thương Mại Điện Tử**

**Tên đề tài**

**Xây dựng Hệ thống Số hóa và Phân tích Hóa đơn Đa**

**định dạng tại HTX Kim Thanh.**

**Mã sinh viên** : 201124029136

**Lớp** : 46K29.1

**Khoá** : 46K

**Giảng viên hướng dẫn** : Th.S Trần Văn Lộc

*Đà Nẵng, ngày … tháng … năm 2025*

**ĐẠI HỌC ĐÀ NẴNG**

**TRƯỜNG ĐẠI HỌC KINH TẾ**



**BÁO CÁO THỰC TẬP TỐT NGHIỆP**

**Chuyên ngành: Thương Mại Điện Tử**

**Tên đề tài**

**Xây dựng Hệ thống Số hóa và Phân tích Hóa đơn Đa**

**định dạng tại HTX Kim Thanh.**

**Giảng viên hướng dẫn Sinh viên thực hiện**

*(Ký và ghi rõ họ tên) (Ký và ghi rõ họ tên)*

*Th.s Trần Văn Lộc Đàm Hồng Nhật*

**MỤC LỤC**

[**PHẦN MỞ ĐẦU 5**](#_4cbrikmd4fbz)

[**CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 7**](#_7trhq1wmeuba)

[1.1. Nhận dạng ký tự quang học (OCR - Optical Character Recognition) 7](#_9vis2lsg5h1g)

[1.1.1. Khái niệm và tầm quan trọng 7](#_9jm8cwxbepg)

[1.1.2. Các công nghệ OCR phổ biến 8](#_irob2frdzifm)

[1.1.3. Các thách thức trong nhận dạng hóa đơn 8](#_im0wujl2cd0a)

[1.2. Kho dữ liệu (Data Warehouse) 8](#_jofv011qhru1)

[1.2.1. Khái niệm 8](#_er4h568wtlq3)

[1.2.2. Đặc điểm chính của kho dữ liệu 9](#_fesvyb19r7wx)

[1.2.3. Mô hình kho dữ liệu 9](#_mt85due6wz91)

[1.2.4. Các thành phần của kho dữ liệu 13](#_ksk8hu7xt6y8)

[1.2.5. Lợi ích của việc xây dựng kho dữ liệu 14](#_65pa4y41plpm)

[1.3. Trích xuất, biến đổi, nạp dữ liệu (ETL - Extract, Transform, Load) 15](#_49cfrdm5ovo4)

[1.3.1. Định nghĩa và Tầm quan trọng của ETL 15](#_9fbu0m66evui)

[1.3.2. Giai đoạn Trích xuất (Extract) chi tiết 15](#_k32zcajduhv3)

[1.3.3. Giai đoạn Biến đổi (Transform) chuyên sâu 16](#_ydw4u525zkbd)

[1.3.4. Giai đoạn Nạp (Load) chi tiết 16](#_ej96dby2it5r)

[1.3.5. Tự động hóa quy trình ETL 16](#_18vtpffmxgf)

[1.3.6. Các kiến trúc ETL 17](#_avsdf8k70mtg)

[1.3.7. Metadata trong ETL 17](#_ocbtkms246a1)

[1.4. Phân tích dữ liệu kinh doanh 17](#_q8ovaut3zkl8)

[1.4.1. Định nghĩa 17](#_6cq04ia2kevx)

[1.4.2. Các Nguyên tắc và Khái niệm Thống kê Nền tảng 18](#_6v1s8b53csca)

[1.4.3. Cơ sở Lý thuyết của Khai thác Dữ liệu và Học Máy 18](#_cwkjx3lc67xp)

[1.4.4. Lý thuyết về Trực quan hóa Dữ liệu 19](#_yduj8pygxt19)

[1.4.5. Cơ sở Lý thuyết về Báo cáo Dữ liệu Kinh doanh 19](#_a5rkc596luvj)

[**CHƯƠNG 2: YÊU CẦU VÀ THIẾT KẾ HỆ THỐNG 19**](#_sz2d3a7qvnva)

[2.1. Yêu cầu hệ thống 19](#_t6ijon6grig)

[2.1.1. Yêu cầu trích xuất dữ liệu từ hóa đơn 19](#_t4s77dr7u3f3)

[2.1.2. Lưu trữ dữ liệu trích xuất vào file Excel 20](#_y3soaqabf091)

[2.1.3. Đưa dữ liệu từ file Excel vào kho dữ liệu SQL Server 20](#_xh1u70cve44b)

[2.1.4. Hỗ trợ truy vấn, báo cáo và phân tích tình hình kinh doanh 21](#_plu63cbjah7h)

[2.2. Thiết kế hệ thống 21](#_vcse9267rmsv)

[2.2.1. Trích xuất dữ liệu 21](#_xte4x4o0jx83)

[2.2.2. Lưu trữ tạm thời và xử lý dữ liệu 21](#_jifzcpvxgl53)

[2.2.3. Thiết kế kho dữ liệu (Data Warehouse) 22](#_p66nldp9gb8p)

[2.2.4. Báo cáo và phân tích 22](#_9cdo1ebierhz)

[**CHƯƠNG 3: XÂY DỰNG VÀ TRIỂN KHAI 22**](#_rqpausykjngy)

[3.1. Vintern – Mô hình kết hợp OCR và xử lý văn bản 22](#_n6h4ips6cvr)

[3.1.1 Cấu trúc mô hình 22](#_dtglqph41qfw)

[3.2. Tổng quan sơ đồ hệ thống 24](#_sy09g2rtal3o)

[3.3. Trích xuất và xử lý dữ liệu từ hóa đơn 24](#_drg0282lmnvm)

[3.3.1. Xử lý hóa đơn giấy 24](#_121zcqbg90q8)

[3.3.2. Xử lý hóa đơn điện tử (file PDF) 24](#_1krwjvvxwv2f)

[3.4. Lưu trữ tạm thời và làm sạch dữ liệu 25](#_16azwjp3xk89)

[3.4.1. Làm sạch dữ liệu 25](#_19npvgvdo313)

[3.4.2. Lưu trữ vào file Excel 25](#_4jp31v4mehqv)

[3.5. Quá trình ETL – Đưa dữ liệu vào kho dữ liệu SQL Server 25](#_uf9qqo6tjnlw)

[3.5.1. Extract – Trích xuất dữ liệu 25](#_85g8zteub6fb)

[3.5.2. Transform – Chuyển đổi dữ liệu 25](#_o3kau0njscsi)

[3.6. Thiết lập hệ thống báo cáo và phân tích dữ liệu 26](#_4t0gaeu6mybu)

[3.6.1. Cấu hình công cụ báo cáo 26](#_ifvbcl7bmlyo)

[4.1.1. Quy trình triển khai báo cáo 28](#_9h4ibx59jypw)

[**CHƯƠNG 4: KHUYẾN NGHỊ VÀ KẾT LUẬN 28**](#_xp0wlqgqchee)

[4.1 Kết luận 28](#_4s0hl986wjs4)

[4.1.1 Tích hợp công nghệ và quy trình tự động 28](#_kwvhnv3k4dx3)

[5.1.1. Hiệu quả trong lưu trữ và quản lý dữ liệu 28](#_jq5upz8mrgqr)

[5.1.2. Khả năng mở rộng và tích hợp báo cáo 28](#_iji2s3yricvn)

[5.1.3. Đánh giá tổng thể và hạn chế gặp phải 28](#_sikiqce28phb)

[5.2. Khuyến nghị 29](#_tqppkns37rq2)

[5.2.1. Cải thiện quá trình số hóa hóa đơn giấy 29](#_apbxrfbe77pr)

[5.2.2. Nâng cao chất lượng xử lý file PDF 29](#_tq6isdvhvjb1)

[5.2.3. Tối ưu hóa quy trình ETL 29](#_jqlgtp7mh9c2)

[5.2.4. Mở rộng khả năng báo cáo và phân tích dữ liệu 29](#_7ibsern8xdze)

[5.2.5. Đảm bảo bảo mật và sao lưu dữ liệu 29](#_qdtwyp9hbdln)

[**TÀI LIỆU THAM KHẢO 30**](#_x930i51x3gxu)

# **PHẦN MỞ ĐẦU**

1. **Tính cấp thiết của đề tài**
   1. **Bối cảnh chung**

Cuộc cách mạng công nghiệp 4.0 đang diễn ra với tốc độ chóng mặt, kéo theo sự phát triển vượt bậc của các công nghệ như trí tuệ nhân tạo (AI), dữ liệu lớn (Big Data), Internet vạn vật (IoT), điện toán đám mây, và blockchain. Các công nghệ này đang thay đổi cách thức vận hành của mọi lĩnh vực, từ sản xuất, kinh doanh, dịch vụ, đến quản lí doanh nghiệp. Các tổ chức, doanh nghiệp, và cả các hợp tác xã (HTX) cần phải nhanh chóng thích ứng với xu hướng này để tồn tại và phát triển.

Chuyển đổi số đang trở thành xu hướng tất yếu, tác động sâu rộng đến mọi mặt đời sống, kinh tế, văn hóa và xã hội. Đây là quá trình tích hợp công nghệ số vào tất cả các lĩnh vực của một tổ chức hoặc doanh nghiệp, nhằm thay đổi cách thức vận hành, mô hình kinh doanh và cung cấp giá trị mới cho khách hàng.

Quá trình chuyển đổi số mang lại nhiều lợi ích cho doanh nghiệp, như cắt giảm chi phí vận hành, tiếp cận được nhiều khách hàng hơn, ra quyết định nhanh chóng và chính xác hơn nhờ hệ thống báo cáo thông suốt kịp thời, tối ưu hóa năng suất làm việc của nhân viên, từ đó nâng cao hiệu quả hoạt động và tăng tính cạnh tranh. Tuy nhiên, chuyển đổi số cũng đặt ra những thách thức về thay đổi văn hóa tổ chức, yêu cầu doanh nghiệp liên tục cập nhật, học hỏi và chấp nhận cả thất bại bên cạnh những thành công của sự đổi mới.

Nhìn chung, chuyển đổi số không chỉ là việc áp dụng công nghệ, mà còn là quá trình thay đổi toàn diện về phương thức làm việc, tư duy và văn hóa trong tổ chức, nhằm tạo ra giá trị mới và nâng cao hiệu quả hoạt động trong thời đại số hóa.

* 1. **Thực trạng tại HTX Kim Thanh**

HTX Kim Thanh hiện đang xử lý một lượng lớn hóa đơn đầu vào, bao gồm cả hóa đơn viết tay truyền thống và hóa đơn điện tử định dạng PDF.

Quy trình xử lý hóa đơn hiện tại chủ yếu là thủ công, gây ra nhiều bất cập:

* Tốn kém thời gian và nguồn lực: Nhân viên phải nhập liệu thủ công thông tin từ hóa đơn giấy vào hệ thống, và trích xuất dữ liệu từ các file PDF, điều này mất nhiều thời gian và công sức.
* Sai sót dữ liệu: Việc nhập liệu thủ công dễ dẫn đến sai sót, ảnh hưởng đến tính chính xác của dữ liệu và báo cáo tài chính.
* Khó khăn trong phân tích và báo cáo: Việc trích xuất và tổng hợp dữ liệu từ nhiều nguồn khác nhau là khó khăn, làm chậm quá trình phân tích và báo cáo, ảnh hưởng đến việc đưa ra các quyết định kinh doanh kịp thời.
* Khả năng mở rộng hạn chế: Khi quy mô hoạt động của HTX tăng lên, hệ thống thủ công hiện tại không thể đáp ứng được nhu cầu xử lý lượng lớn dữ liệu, gây cản trở cho sự phát triển.

Vì những khó khăn trên nên việc đề xuất giải pháp chuyển đổi số để đáp ứng được những nhu cầu của Hợp tác xã Kim Thanh là tối quan trọng. Và giải pháp đưa ra là: Hệ thống Số hóa và Phân tích Hóa đơn đa định dạng tại HTX Kim Thanh. Việc triển khai một giải pháp tự động hóa quy trình ETL (Extract, Transform, Load) và xây dựng kho dữ liệu để số hóa hóa đơn đa định dạng (PDF và thủ công) là vô cùng cần thiết.

Giải pháp này sẽ giúp HTX Kim Thanh:

* Tiết kiệm thời gian và chi phí xử lý hóa đơn.
* Giảm thiểu sai sót dữ liệu.
* Tăng cường khả năng phân tích và báo cáo dữ liệu.
* Hỗ trợ đưa ra các quyết định kinh doanh chính xác và kịp thời.
* Tăng khả năng mở rộng hệ thống.

1. **Mục tiêu nghiên cứu của đề tài**

* Mục tiêu tổng quát: Xây dựng một hệ thống tự động hóa ETL và kho dữ liệu hiệu quả để số hóa hóa đơn đa định dạng (PDF và thủ công) tại HTX Kim Thanh, từ đó nâng cao hiệu quả quản lý và phân tích dữ liệu kinh doanh.
* Mục tiêu cụ thể:
  + Phát triển hệ thống nhận dạng ký tự quang học (OCR) để trích xuất dữ liệu từ hóa đơn viết tay với độ chính xác cao.
  + Thiết kế và triển khai kho dữ liệu (Data Warehouse) để lưu trữ và quản lý dữ liệu hóa đơn một cách hiệu quả và dễ dàng truy xuất.
  + Xây dựng quy trình ETL tự động để trích xuất dữ liệu từ hóa đơn PDF và hóa đơn viết tay đã số hóa một cách chính xác.
  + Phát triển các công cụ báo cáo và phân tích dữ liệu linh hoạt, để hỗ trợ người dùng trích xuất thông tin hữu ích từ kho dữ liệu, phục vụ cho việc ra quyết định.
  + Đánh giá hiệu quả của hệ thống đã xây dựng, bao gồm tính chính xác, hiệu suất, và khả năng đáp ứng yêu cầu của người dùng.

1. **Đối tượng và phạm vi nghiên cứu**
   1. **Đối tượng nghiên cứu**

* Quy trình xử lý hóa đơn hiện tại tại HTX Kim Thanh.
* Công nghệ OCR, ETL, và kho dữ liệu (Data Warehouse).
* Các phương pháp phân tích và báo cáo dữ liệu kinh doanh.
* Hóa đơn PDF, và hóa đơn thủ công.
  1. **Phạm vi nghiên cứu**
* Đề tài tập trung vào việc số hóa hóa đơn đầu vào (hóa đơn mua hàng và bán hàng) của HTX Kim Thanh.
* Dữ liệu được xử lý trong phạm vi nghiên cứu là dữ liệu hóa đơn trong vòng 2 năm gần nhất, để bảo đảm tính thống kê và đưa ra cái nhìn tổng quan.
* Hệ thống được xây dựng và triển khai trong phạm vi HTX Kim Thanh.
* Hạn chế về các loại hoá đơn đặc biệt, có những định dạng quá khác biệt.

1. **Phương pháp nghiên cứu**

* Nghiên cứu tài liệu: Nghiên cứu các tài liệu liên quan đến công nghệ OCR, ETL, kho dữ liệu, phân tích dữ liệu kinh doanh, và các giải pháp số hóa hóa đơn, bao gồm các bài báo khoa học, sách, tài liệu hướng dẫn kỹ thuật, và các báo cáo nghiên cứu.
* Khảo sát và phỏng vấn: Khảo sát và phỏng vấn cán bộ, nhân viên của HTX Kim Thanh để hiểu rõ về quy trình xử lý hóa đơn hiện tại, các vấn đề gặp phải, và nhu cầu của người dùng..
* Phân tích và thiết kế hệ thống: Phân tích yêu cầu và thiết kế hệ thống số hóa hóa đơn, bao gồm kiến trúc hệ thống, quy trình ETL, và cấu trúc kho dữ liệu , sau đó sử dụng các công cụ sơ đồ để thể hiện rõ hơn về luồng dữ liệu của hệ thống, cũng như sơ đồ về Data Warehouse.
* Xây dựng và triển khai hệ thống: Xây dựng và triển khai hệ thống số hóa hóa đơn bằng cách sử dụng các công nghệ phù hợp, bao gồm lựa chọn ngôn ngữ lập trình, cơ sở dữ liệu, và công cụ ETL. Việc chia hệ thống ra thành các module là cần thiết để có thể dễ dàng quản lí, và nâng cấp về sau.
* Thử nghiệm và đánh giá: Thử nghiệm hệ thống với dữ liệu thực tế để đánh giá hiệu quả của hệ thống dựa trên các tiêu chí đã đề ra, bao gồm tính chính xác, hiệu suất, và khả năng đáp ứng yêu cầu của người dùng, áp dụng các số liệu cụ thể để cho thấy sự hiệu quả của hệ thống sau khi vận hành.

1. **Kết cấu đề tài**

Kết cấu đề tài gồm các chương cụ thể như sau:

Chương 1: Cơ sở lý thuyết

Chương 2: Yêu cầu và thiết kế hệ thống

Chương 3: Xây dựng và triển khai

Chương 4: Kết luận và khuyến nghị

# **CHƯƠNG 1: CƠ SỞ LÝ THUYẾT**

## Nhận dạng ký tự quang học (OCR - Optical Character Recognition)

### Khái niệm và tầm quan trọng

Nhận dạng ký tự quang học (OCR) là quá trình chuyển đổi một hình ảnh văn bản thành định dạng văn bản mà máy có thể đọc được.(link)

Ví dụ: Nếu bạn quét một biểu mẫu hoặc biên lai, máy tính của bạn sẽ lưu bản quét đó dưới dạng tệp hình ảnh. Bạn không thể sử dụng trình soạn thảo văn bản để chỉnh sửa, tìm kiếm hoặc đếm số từ trong tệp hình ảnh. Tuy nhiên, bạn có thể sử dụng OCR để chuyển đổi hình ảnh thành tài liệu văn bản, trong đó phần nội dung sẽ được lưu trữ dưới dạng dữ liệu văn bản.

Hầu hết các luồng công việc kinh doanh sẽ bao gồm việc nhận thông tin từ các phương tiện truyền thông dạng bản in. Các biểu mẫu, hóa đơn dạng giấy, bản quét tài liệu pháp lý và bản in hợp đồng đều là một phần trong quy trình kinh doanh. Khối lượng giấy tờ lớn như vậy làm mất rất nhiều thời gian và không gian để lưu trữ và quản lý. Mặc dù quản lý tài liệu không cần giấy tờ là cách thức hiệu quả nhất, nhưng việc quét tài liệu thành hình ảnh sẽ tạo ra nhiều thách thức. Quá trình này đòi hỏi phải có thao tác can thiệp thủ công và cũng có thể khá buồn tẻ và lề mề.

Hơn nữa, việc số hóa nội dung tài liệu này sẽ tạo ra các tệp hình ảnh với văn bản ẩn bên trong. Không thể xử lý văn bản trong hình ảnh bằng phần mềm xử lý văn bản theo cách giống như đối với tài liệu văn bản. Công nghệ OCR giải quyết vấn đề này bằng cách chuyển đổi hình ảnh văn bản thành dữ liệu văn bản mà phần mềm kinh doanh khác có thể phân tích. Sau đó, bạn có thể sử dụng dữ liệu để tiến hành phân tích, hợp lý hóa hoạt động, tự động hóa các quy trình và cải thiện năng suất.

### Các công nghệ OCR phổ biến

Các phương pháp truyền thống:

* Dựa trên so sánh mẫu: So sánh các ký tự đã được phân đoạn với các mẫu ký tự đã biết.
* Phân tích đặc trưng: Trích xuất các đặc điểm của ký tự (đường cong, góc cạnh, v.v.) và so sánh chúng với các đặc điểm của ký tự đã biết.

Các phương pháp dựa trên học máy:

* Sử dụng mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi quy (RNN) để cải thiện độ chính xác.
* Các công cụ OCR thông dụng như Tesseract OCR và các API OCR của Google Cloud Vision.

### Các thách thức trong nhận dạng hóa đơn

* Đa dạng định dạng và bố cục hóa đơn.
* Chữ viết tay không đồng nhất, gây khó khăn cho việc trích xuất.
* Chất lượng hình ảnh hóa đơn thấp (nhòe, mờ, nếp gấp, v.v.).
* Sự đa dạng về phông chữ và kích thước chữ

## Kho dữ liệu (Data Warehouse)

### Khái niệm

Theo định nghĩa của Ralph Kimball và Margy Ross trích trong tác phẩm The Data Warehouse Toolkit: "Kho dữ liệu là một bản sao của dữ liệu giao dịch được tối ưu hóa cho truy vấn và phân tích. Kho dữ liệu thường liên quan đến việc tích hợp dữ liệu từ nhiều nguồn khác nhau vào một kho lưu trữ thống nhất.".

Theo W.H. Inmon trình bày trong tác phẩm Building the Data Warehouse: "Kho dữ liệu là một tập hợp dữ liệu hướng chủ đề, tích hợp, thay đổi theo thời gian và không biến động, được tổ chức để hỗ trợ việc ra quyết định của nhà quản lý.".

Từ định nghĩa của hai tác giả trên có thể rút ra được định nghĩa chung: “Kho dữ liệu là hệ thống lưu trữ dữ liệu tập trung, được thiết kế để hỗ trợ báo cáo và phân tích dữ liệu, nhằm phục vụ cho việc ra quyết định trong doanh nghiệp.”

Kho dữ liệu là công cụ quan trọng để biến dữ liệu thô thành thông tin hữu ích, hỗ trợ việc ra quyết định chiến lược nhằm giúp doanh nghiệp hiểu rõ hơn về khách hàng, sản phẩm, thị trường và các yếu tố khác ảnh hưởng đến hoạt động kinh doanh, tạo nền tảng cho các hoạt động phân tích dữ liệu nâng cao như khai phá dữ liệu (data mining) và học máy (machine learning).

### Đặc điểm chính của kho dữ liệu

Các đặc điểm chính của Kho Dữ Liệu

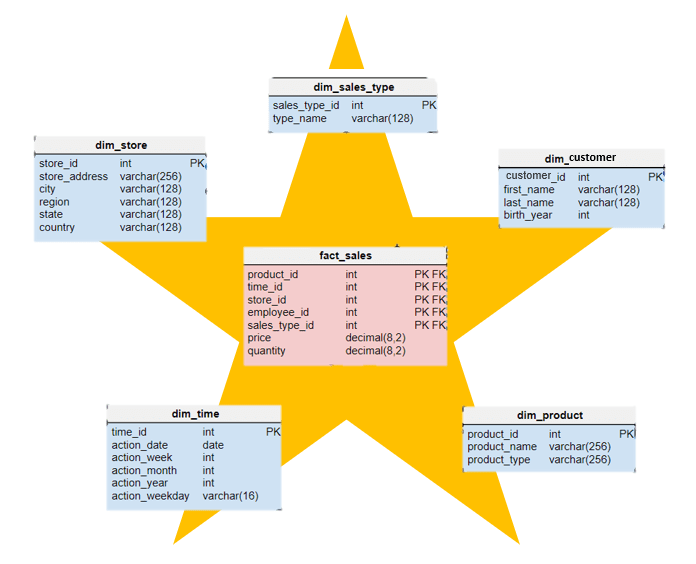
* Hướng Chủ Đề (Subject-Oriented): Kho dữ liệu được tổ chức xung quanh các chủ đề chính của doanh nghiệp (ví dụ: khách hàng, sản phẩm, bán hàng), thay vì các ứng dụng cụ thể. Điều này cho phép người dùng dễ dàng truy xuất và phân tích dữ liệu theo các lĩnh vực kinh doanh quan trọng.
* Tích Hợp (Integrated): Dữ liệu từ nhiều nguồn khác nhau (ví dụ: CRM, ERP, hệ thống bán hàng) được tích hợp vào một kho lưu trữ thống nhất. Quá trình tích hợp bao gồm làm sạch, chuyển đổi và chuẩn hóa dữ liệu để đảm bảo tính nhất quán.
* Ổn Định (Non-Volatile): Dữ liệu trong kho dữ liệu không bị thay đổi hoặc xóa bỏ sau khi đã được tải vào. Dữ liệu chỉ được thêm mới hoặc cập nhật định kỳ, đảm bảo tính lịch sử và khả năng so sánh theo thời gian.
* Theo Thời Gian (Time-Variant): Kho dữ liệu lưu trữ dữ liệu theo thời gian, cho phép phân tích xu hướng và thay đổi dữ liệu theo thời gian. Mỗi bản ghi dữ liệu thường đi kèm với thông tin về thời gian, cho phép truy vấn dữ liệu theo các khoảng thời gian khác nhau. (Ralph Kimball 2019.)

### Mô hình kho dữ liệu

* + - 1. **Lược đồ hình sao (Star schema):**

Cấu trúc kho dữ liệu với bảng chiều (dimension table) và bảng sự kiện (fact table): Mô hình hình sao bao gồm một bảng sự kiện (fact table) nằm ở trung tâm, được bao quanh bởi các bảng chiều (dimension tables). Bảng sự kiện (Fact Table): Chứa các số liệu định lượng về các hoạt động kinh doanh. Mỗi bản ghi trong bảng sự kiện đại diện cho một sự kiện hoặc giao dịch cụ thể và bao gồm các khóa ngoại tham chiếu đến các bảng chiều liên quan. Bảng chiều (Dimension Table): Lưu trữ thông tin mô tả về các khía cạnh của dữ liệu, chẳng hạn như thông tin về khách hàng, sản phẩm, thời gian, địa điểm. Các bảng chiều giúp cung cấp ngữ cảnh cho các số liệu trong bảng sự kiện, cho phép phân tích dữ liệu theo nhiều chiều khác nhau. Mối quan hệ giữa bảng sự kiện và bảng chiều được thiết lập thông qua các khóa ngoại trong bảng sự kiện, liên kết đến các khóa chính trong các bảng chiều. Điều này tạo nên cấu trúc giống như hình ngôi sao khi biểu diễn sơ đồ quan hệ.

Hình ảnh lược đồ hình sao:



Hình 1.1 Lược đồ hình sao

**Ví dụ về các bảng sự kiện và bảng chiều trong ngữ cảnh hóa đơn:**

Giả sử chúng ta có một hệ thống quản lý bán hàng với các bảng sau:​

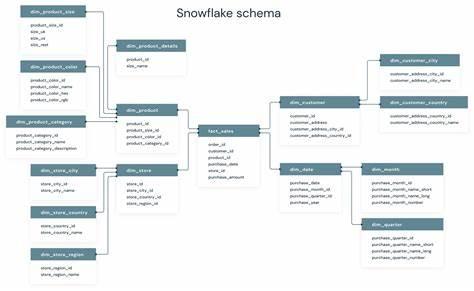
* **Bảng sự kiện "Hóa đơn" (Fact Table):** Lưu trữ thông tin về các giao dịch bán hàng, bao gồm các cột như:​
  + MaHoaDon (Khóa chính)
  + MaKhachHang (Khóa ngoại đến bảng Khách hàng)
  + MaSanPham (Khóa ngoại đến bảng Sản phẩm)
  + MaThoiGian (Khóa ngoại đến bảng Thời gian)
  + SoLuong (Số lượng sản phẩm)
  + DoanhThu (Doanh thu từ giao dịch)
* **Bảng chiều "Khách hàng" (Dimension Table):** Chứa thông tin về khách hàng, bao gồm:​
  + MaKhachHang (Khóa chính)
  + TenKhachHang
  + DiaChi
  + SoDienThoai
* **Bảng chiều "Sản phẩm" (Dimension Table):** Chứa thông tin về sản phẩm:​
  + MaSanPham (Khóa chính)
  + TenSanPham
  + LoaiSanPham
  + Gia
* **Bảng chiều "Thời gian" (Dimension Table):** Chứa thông tin về thời gian
  + MaThoiGian (Khóa chính)
  + Ngay
  + Thang
  + Nam

Cấu trúc này cho phép phân tích doanh thu theo từng khách hàng, sản phẩm và thời gian cụ thể.

* + - 1. **Lược đồ bông tuyết (Snowflake schema):**

Mô hình bông tuyết là một biến thể của mô hình hình sao, trong đó các bảng chiều được chuẩn hóa và phân tách thành nhiều bảng nhỏ hơn để loại bỏ sự dư thừa dữ liệu. Cấu trúc này bao gồm một bảng sự kiện (fact table) trung tâm chứa các số liệu kinh doanh và nhiều bảng chiều (dimension tables) liên kết, nhưng khác với mô hình hình sao, các bảng chiều trong mô hình bông tuyết được phân tách thành các bảng con, tạo thành cấu trúc giống như một bông tuyết khi biểu diễn sơ đồ quan hệ. ​**Bảng sự kiện (Fact Table):** Chứa các số liệu định lượng về các hoạt động kinh doanh. Mỗi bản ghi trong bảng sự kiện đại diện cho một sự kiện hoặc giao dịch cụ thể và bao gồm các khóa ngoại tham chiếu đến các bảng chiều liên quan.​ **Bảng chiều (Dimension Table):** Trong mô hình bông tuyết, các bảng chiều được chuẩn hóa và phân tách thành nhiều bảng con để loại bỏ sự dư thừa dữ liệu. Ví dụ, một bảng chiều "Sản phẩm" có thể được chia thành các bảng nhỏ hơn như "Danh mục sản phẩm", "Thương hiệu", và "Nhà cung cấp".​ Mối quan hệ giữa bảng sự kiện và các bảng chiều được thiết lập thông qua các khóa ngoại trong bảng sự kiện, liên kết đến các khóa chính trong các bảng chiều. Trong mô hình bông tuyết, do các bảng chiều được phân tách thành nhiều bảng con, các phép nối (joins) giữa các bảng sẽ phức tạp hơn so với mô hình hình sao. ​[Guru99](https://www.guru99.com/vi/snowflake-schema-in-data-warehouse-model.html?utm_source=chatgpt.com)

Ví dụ về lược đồ bông tuyết:



Hình 1.2 Lược đồ bông tuyết

**Ví dụ về các bảng sự kiện và bảng chiều cụ thể trong ngữ cảnh hóa đơn:**

Giả sử chúng ta có một hệ thống quản lý bán hàng với các bảng sau trong mô hình bông tuyết:​

* **Bảng sự kiện "Hóa đơn" (Fact Table):** Lưu trữ thông tin về các giao dịch bán hàng, bao gồm các cột như:​
  + MaHoaDon (Khóa chính)
  + MaKhachHang (Khóa ngoại đến bảng KhachHang)
  + MaSanPham (Khóa ngoại đến bảng SanPham)
  + MaThoiGian (Khóa ngoại đến bảng ThoiGian)
  + SoLuong (Số lượng sản phẩm)
  + DoanhThu (Doanh thu từ giao dịch)
* **Bảng chiều "KhachHang" (Dimension Table):** Chứa thông tin về khách hàng:​
  + MaKhachHang (Khóa chính)
  + TenKhachHang
  + DiaChi
  + SoDienThoai
* **Bảng chiều "SanPham" (Dimension Table):** Trong mô hình bông tuyết, bảng "SanPham" được chuẩn hóa và phân tách thành các bảng nhỏ hơn:​
  + **Bảng "SanPham":**
    - MaSanPham (Khóa chính)
    - TenSanPham
    - MaLoaiSanPham (Khóa ngoại đến bảng LoaiSanPham)
    - MaNhaCungCap (Khóa ngoại đến bảng NhaCungCap)
  + **Bảng "LoaiSanPham":**
    - MaLoaiSanPham (Khóa chính)
    - TenLoaiSanPham
  + **Bảng "NhaCungCap":**
    - MaNhaCungCap (Khóa chính)
    - TenNhaCungCap
    - DiaChi
* **Bảng chiều "ThoiGian" (Dimension Table):** Chứa thông tin về thời gian:​
  + MaThoiGian (Khóa chính)
  + Ngay
  + Thang
  + Nam

Cấu trúc này cho phép phân tích doanh thu theo từng khách hàng, sản phẩm và thời gian cụ thể, đồng thời giảm thiểu sự dư thừa dữ liệu thông qua việc chuẩn hóa các bảng chiều.

### Các thành phần của kho dữ liệu

Kho dữ liệu là một hệ thống phức tạp bao gồm nhiều thành phần khác nhau, phối hợp với nhau để cung cấp một nền tảng vững chắc cho việc phân tích và báo cáo dữ liệu. Dưới đây là mô tả chi tiết về các thành phần chính:

* + - 1. **Hệ thống nguồn (Source Systems)**

Đây là các hệ thống giao dịch ban đầu, nơi dữ liệu được tạo ra và lưu trữ dưới dạng dữ liệu chi tiết (transactional data). Các hệ thống này cung cấp dữ liệu gốc, đa dạng về nguồn, cần được tập hợp và xử lý để phục vụ cho mục đích phân tích

* + - 1. **Quá trình ETL(Extract – Transform – Load)**

Giai đoạn này có nhiệm vụ:

* **Extract (Trích xuất):** Lấy dữ liệu từ các hệ thống nguồn vận hành.
* **Transform (Chuyển đổi):** Làm sạch, tích hợp và chuyển đổi dữ liệu theo các quy tắc nghiệp vụ đã được xác định. Qua đó, dữ liệu được chuẩn hóa và định dạng phù hợp với mô hình dữ liệu chiều.
* **Load (Tải dữ liệu):** Đưa dữ liệu đã được chuyển đổi vào kho dữ liệu.  
  Quá trình ETL không chỉ đảm bảo chất lượng dữ liệu mà còn tối ưu hiệu suất xử lý và cập nhật dữ liệu vào Data Warehouse.
  + - 1. **Data Warehouse (Kho Dữ Liệu Chính)**

Kho dữ liệu trung tâm (Data Warehouse) – Mô hình chiều, dữ liệu trong kho được tổ chức theo mô hình chiều, gồm:

* **Bảng sự kiện (Fact Tables):** Lưu trữ các số liệu đo lường, thông tin định lượng (ví dụ: doanh số, số lượng giao dịch) và chứa các khóa liên kết đến các bảng thuộc tính.
* **Bảng thuộc tính (Dimension Tables):** Cung cấp ngữ cảnh mô tả cho các sự kiện (ví dụ: thời gian, sản phẩm, khách hàng, địa lý).
* **Lược đồ sao (Star Schema):** Là cách tổ chức dữ liệu phổ biến trong kiến trúc Kimball, giúp tối ưu hóa các truy vấn phân tích và báo cáo.

Ngoài ra, quy hoạch theo “bus architecture” với các “conformed dimensions” (các chiều thống nhất) giúp đảm bảo tính nhất quán và khả năng tích hợp dữ liệu từ nhiều nguồn khác nhau.

* + - 1. **Khu vực trình bày ( Presentation Area)**

Đây là nơi dữ liệu từ kho dữ liệu được cung cấp cho người dùng cuối qua các công cụ báo cáo, phân tích và khai thác dữ liệu (BI – Business Intelligence). Các công cụ này có thể bao gồm OLAP, dashboards, báo cáo tương tác… Nhờ đó, người dùng có thể truy vấn và phân tích dữ liệu một cách trực quan, nhanh chóng để đưa ra quyết định.

* + - 1. **Ứng dụng Business Intelligence (BI Application)**

Các ứng dụng này giúp khai thác, phân tích dữ liệu từ kho dữ liệu nhằm hỗ trợ các quyết định kinh doanh. Chúng bao gồm các công cụ phân tích, báo cáo, thậm chí là các hệ thống cảnh báo sớm giúp quản lý hiệu quả hoạt động kinh doanh.

### Lợi ích của việc xây dựng kho dữ liệu

Việc xây dựng kho dữ liệu mang lại nhiều lợi ích thiết thực cho tổ chức, trước hết là cải thiện đáng kể khả năng truy cập và phân tích dữ liệu bằng cách tập trung, làm sạch và cấu trúc dữ liệu từ nhiều nguồn khác nhau, tạo điều kiện thuận lợi cho việc phân tích sâu rộng và theo thời gian. Nhờ đó, tổ chức có thể đưa ra các quyết định kinh doanh chính xác và kịp thời hơn, dựa trên những bằng chứng và phân tích dữ liệu cụ thể, giúp nhận diện sớm cơ hội, thách thức và đánh giá hiệu quả chiến lược. Cuối cùng, việc phân tích dữ liệu từ kho dữ liệu còn giúp nâng cao hiệu quả hoạt động của tổ chức thông qua việc tối ưu hóa quy trình, cá nhân hóa trải nghiệm khách hàng, nâng cao hiệu suất nhân viên và quản lý rủi ro hiệu quả hơn, góp phần tạo dựng lợi thế cạnh tranh bền vững.

## Trích xuất, biến đổi, nạp dữ liệu (ETL - Extract, Transform, Load)

### Định nghĩa và Tầm quan trọng của ETL

ETL là quy trình tích hợp dữ liệu gồm ba bước chính:

* Extract (Trích xuất): Lấy dữ liệu từ các nguồn khác nhau (cơ sở dữ liệu quan hệ, file, dịch vụ web, …).
* Transform (Biến đổi): Xử lý, làm sạch, chuẩn hóa và chuyển đổi dữ liệu theo các quy tắc kinh doanh.
* Load (Nạp): Chuyển dữ liệu đã được xử lý vào kho dữ liệu hoặc hệ thống lưu trữ mục tiêu để phục vụ cho phân tích và báo cáo.

Quá trình này không chỉ giúp hợp nhất dữ liệu mà còn đảm bảo chất lượng và tính nhất quán của thông tin từ các nguồn khác nhau .

Tầm quan trọng

* Tích hợp dữ liệu: ETL cho phép kết nối dữ liệu từ nhiều nguồn khác nhau, giúp xây dựng một tập hợp dữ liệu thống nhất phục vụ phân tích và ra quyết định .
* Đảm bảo chất lượng: Thông qua các bước làm sạch và chuẩn hóa, ETL giảm thiểu lỗi và sai lệch dữ liệu, góp phần nâng cao độ tin cậy của các báo cáo kinh doanh.
* Hỗ trợ quyết định: Dữ liệu được xử lý một cách chính xác và kịp thời tạo điều kiện thuận lợi cho các hệ thống BI (Business Intelligence) và hỗ trợ quyết định chiến lược trong doanh nghiệp .

### Giai đoạn Trích xuất (Extract) chi tiết

Quá trình trích xuất dữ liệu bao gồm các bước sau:

* Xác định nguồn dữ liệu: Xác định các hệ thống, cơ sở dữ liệu, file log, dịch vụ API,… là nguồn cung cấp dữ liệu.
* Kết nối và truy xuất: Sử dụng giao thức kết nối (SQL, REST API, …) để truy vấn và lấy dữ liệu từ nguồn.
* Xử lý dữ liệu phi cấu trúc: Với các nguồn dữ liệu phi cấu trúc hoặc bán cấu trúc, cần áp dụng các công cụ chuyên dụng để chuyển đổi dữ liệu về dạng có cấu trúc.
* Đảm bảo tính toàn vẹn: Các biện pháp kiểm tra đảm bảo dữ liệu được trích xuất đầy đủ và không bị thay đổi so với bản gốc.

### Giai đoạn Biến đổi (Transform) chuyên sâu

Quá trình biến đổi dữ liệu là bước then chốt nhằm đảm bảo dữ liệu có thể sử dụng một cách hiệu quả:

* Làm sạch dữ liệu: Xác định và loại bỏ các bản ghi lỗi, giá trị thiếu hụt, dữ liệu trùng lặp và không hợp lệ.
* Chuẩn hóa: Đưa dữ liệu về định dạng thống nhất (định dạng ngày tháng, đơn vị đo lường, ký tự mã hóa, …) nhằm tạo điều kiện thuận lợi cho quá trình phân tích sau này.
* Tích hợp và kết nối: Kết hợp dữ liệu từ các nguồn khác nhau, xử lý các mâu thuẫn và đảm bảo sự đồng bộ của thông tin.
* Áp dụng quy tắc kinh doanh: Tính toán, chuyển đổi dữ liệu theo các công thức và quy tắc đã định nghĩa, nhằm phản ánh đúng yêu cầu phân tích của tổ chức.
* Tối ưu hóa: Sắp xếp và định dạng dữ liệu sao cho phù hợp với mô hình của hệ thống đích, tối ưu hiệu năng truy xuất.

### Giai đoạn Nạp (Load) chi tiết

Giai đoạn nạp dữ liệu bao gồm:

* Xác định hệ thống đích: Chọn kho dữ liệu (data warehouse), cơ sở dữ liệu quan hệ hoặc hệ thống lưu trữ NoSQL tùy thuộc vào mục tiêu sử dụng.
* Chiến lược nạp dữ liệu:
* Nạp theo lô (Batch Loading): Phù hợp với các hệ thống không yêu cầu cập nhật theo thời gian thực, thường được thực hiện trong giờ thấp điểm nhằm giảm tải hệ thống.
* Nạp thời gian thực (Real-time Loading): Dùng cho các ứng dụng cần cập nhật liên tục, hỗ trợ các dịch vụ thời gian thực và giao dịch trực tuyến.
* Kiểm tra và xác nhận: Sau khi nạp, tiến hành kiểm tra tính toàn vẹn của dữ liệu, đảm bảo không có lỗi phát sinh trong quá trình truyền tải.

### Tự động hóa quy trình ETL

Lợi ích của tự động hóa

* Tiết kiệm thời gian và chi phí: Giảm thiểu sự can thiệp thủ công, qua đó hạn chế sai sót và tăng tốc độ xử lý dữ liệu.
* Độ chính xác và nhất quán: Các quy trình tự động đảm bảo rằng các bước ETL được thực hiện theo một chuẩn mực nhất định, nâng cao độ tin cậy của dữ liệu.
* Khả năng mở rộng: Hệ thống tự động dễ dàng mở rộng khi khối lượng dữ liệu tăng đột biến, giúp đáp ứng yêu cầu của doanh nghiệp mà không cần thay đổi quy trình thủ công nhiều.
* Giám sát và cảnh báo: Các công cụ ETL tự động cung cấp các tính năng giám sát thời gian thực, phát hiện sớm các lỗi và cảnh báo cho quản trị viên, đảm bảo hệ thống luôn vận hành trơn tru.

### Các kiến trúc ETL

Các mô hình kiến trúc ETL phổ biến:

* **ETL truyền thống:**
  + Dữ liệu được trích xuất, biến đổi và nạp theo lịch trình định kỳ (batch processing).
  + Thích hợp với các hệ thống báo cáo không yêu cầu cập nhật theo thời gian thực.
* **ELT (Extract – Load – Transform):**
  + Dữ liệu được trích xuất và nạp trực tiếp vào hệ thống đích, sau đó thực hiện các bước biến đổi trên kho dữ liệu hiện đại có khả năng xử lý song song cao.
  + Tối ưu hóa hiệu năng cho các hệ thống phân tích lớn .
* **ETL thời gian thực:**
  + Sử với các hệ thống giao dịch hoặc báo cáo thời gian thực.
* **Kiến trúc đa tầng (Multi-tier ETL):**
  + Chia quy trình ETL thành các tầng riêng biệt như staging, cleansing, và integration.
  + Mỗi tầng được kiểm soát chặt chẽ nhằm đảm bảo tính ổn định và khả năng mở rộng của toàn bộ hệ thống .

### Metadata trong ETL

Metadata được hiểu là “dữ liệu về dữ liệu”, cung cấp các thông tin mô tả về nguồn gốc, cấu trúc, chất lượng và quá trình xử lý của dữ liệu.

Vai trò:

* Quản lý và kiểm soát: Metadata ghi lại các bước và quy trình ETL, giúp các nhà quản trị theo dõi quá trình xử lý dữ liệu và dễ dàng phát hiện lỗi khi có sự cố xảy ra.
* Minh bạch: Cung cấp thông tin chi tiết về nguồn gốc, thời điểm và cách thức biến đổi dữ liệu, qua đó giúp người dùng hiểu rõ hơn về hành trình của dữ liệu từ nguồn đến đích.
* Tối ưu hóa quy trình: Phân tích metadata cho phép đánh giá hiệu năng của các bước ETL, từ đó đưa ra các biện pháp cải tiến và tối ưu hóa hệ thống.
* Hỗ trợ truy vấn và phân tích: Khi có metadata chi tiết, người dùng có thể dễ dàng truy xuất các thông tin liên quan, từ đó tạo điều kiện cho việc phân tích và báo cáo dữ liệu chính xác.

## Phân tích dữ liệu kinh doanh

### Định nghĩa

Phân tích dữ liệu kinh doanh là quá trình thu thập, xử lý, phân tích và diễn giải dữ liệu liên quan đến các hoạt động kinh doanh để đưa ra những kết luận và khuyến nghị hỗ trợ ra quyết định chiến lược. Quá trình này bao gồm việc áp dụng các phương pháp thống kê, khai thác dữ liệu và học máy nhằm phát hiện các mẫu, xu hướng và mối liên hệ tiềm ẩn trong dữ liệu kinh doanh. (Wikipedia, Business Analytics)

### Các Nguyên tắc và Khái niệm Thống kê Nền tảng

Theo IBM(IBM Analytics, “Statistical Foundations for Business Analytics”. Truy cập tại: https://www.ibm.com/analytics (xem các hướng dẫn về thống kê và phân tích dữ liệu)), các nguyên tắc thống kê là nền tảng của phân tích dữ liệu, bao gồm:

* **Mô tả thống kê:** Các chỉ số như trung bình, trung vị, phương sai, độ lệch chuẩn giúp tóm tắt đặc trưng dữ liệu.
* **Xác suất:** Đo lường mức độ không chắc chắn, hỗ trợ việc dự báo và ra quyết định trong môi trường rủi ro.
* **Phân phối xác suất:** Hiểu biết về các phân phối (ví dụ: phân phối chuẩn, Poisson) giúp lựa chọn các mô hình thống kê phù hợp.
* **Kiểm định giả thuyết:** Đánh giá độ tin cậy của các kết quả phân tích thông qua các bài kiểm tra thống kê và mức ý nghĩa.
* **Hồi quy và tương quan:** Xác định mối quan hệ giữa các biến số và dự báo xu hướng tương lai.

### Cơ sở Lý thuyết của Khai thác Dữ liệu và Học Máy

Theo IBM ( IBM Analytics, “Data Mining and Machine Learning”. Truy cập tại: https://www.ibm.com/analytics (xem tài liệu về khai thác dữ liệu và học máy trong kinh doanh)

**Khai thác dữ liệu (Data Mining)**

* Mục tiêu: Tìm kiếm mẫu, phân nhóm (clustering), phân loại (classification) và luật kết hợp (association rules) từ kho dữ liệu lớn.
* Ứng dụng: Giúp doanh nghiệp phát hiện xu hướng, nhận diện dị thường và hỗ trợ xây dựng chiến lược kinh doanh dựa trên dữ liệu thực tế.

**Học máy (Machine Learning)**

* Các phương pháp học máy:
  + Học giám sát (Supervised Learning): Dự báo và phân loại dựa trên dữ liệu đã gán nhãn.
  + Học không giám sát (Unsupervised Learning): Nhóm và phân tích các mẫu ẩn trong dữ liệu.
  + Học tăng cường (Reinforcement Learning): Tối ưu hoá quá trình ra quyết định qua trải nghiệm.
* Lợi ích của học máy: Xây dựng các mô hình dự báo, tối ưu hóa hiệu năng và giảm thiểu sai số trong các ứng dụng kinh doanh.

### Lý thuyết về Trực quan hóa Dữ liệu

Theo Statista, trực quan hóa dữ liệu giúp chuyển đổi các kết quả phân tích thành hình ảnh dễ hiểu như biểu đồ, đồ thị và dashboard:

* **Nguyên tắc thiết kế:**
  + Tính đơn giản và rõ ràng là yếu tố then chốt để truyền đạt thông tin.
  + Áp dụng các nguyên tắc của Edward Tufte như “data-ink ratio” nhằm tối ưu hóa hiển thị thông tin.
* **Công cụ trực quan hóa:**
  + Sử dụng các công cụ như Tableau, Power BI, D3.js để tạo biểu đồ tương tác và trực quan.
  + Lựa chọn biểu đồ phù hợp (biểu đồ cột, đường, tròn, phân tán) tùy theo mục đích phân tích.

### Cơ sở Lý thuyết về Báo cáo Dữ liệu Kinh doanh

Dựa trên IBM Analytics, báo cáo dữ liệu kinh doanh là công cụ truyền tải thông tin phân tích đến các bên liên quan, bao gồm:

* **Định dạng và cấu trúc báo cáo:**
  + Các báo cáo có thể bao gồm văn bản, bảng biểu và trực quan hóa dữ liệu.
  + Dashboard tương tác cung cấp cập nhật theo thời gian thực cho quản lý.
* **Quy trình và chuẩn mực báo cáo:**
  + Đảm bảo chất lượng, độ tin cậy và tính nhất quán của dữ liệu thông qua việc tuân thủ các chuẩn mực báo cáo.
  + Tự động hoá quy trình báo cáo nhằm giảm thiểu sai sót và tăng cường hiệu quả truyền đạt thông tin.
* **Ứng dụng thực tiễn:** Các báo cáo không chỉ giúp giám sát hoạt động kinh doanh mà còn hỗ trợ việc ra quyết định chiến lược và quản lý rủi ro.

# **CHƯƠNG 2: YÊU CẦU VÀ THIẾT KẾ HỆ THỐNG**

## Yêu cầu hệ thống

### Yêu cầu trích xuất dữ liệu từ hóa đơn

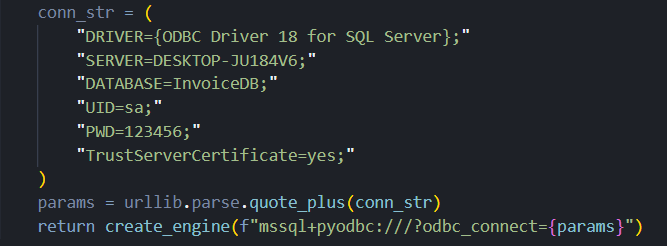
* + - 1. **Xử lý hóa đơn giấy**
* Quét và nhận diện ký tự (OCR): Sử dụng thư viện như hỗ trợ OCR để chuyển ảnh quét hóa đơn giấy thành dạng dữ liệu json.
* Phân đoạn trường thông tin: Dùng thuật toán định vị khung (bounding box) hoặc template matching để xác định vị trí các mục quan trọng (ngày tháng, tên khách hàng, tên mặt hàng, tổng tiền…).
* Kiểm tra và hiệu chỉnh lỗi: So sánh giá trị trích xuất với danh sách tham chiếu (định dạng ngày tháng) và áp dụng quy tắc làm sạch (regex, từ điển) để sửa lỗi chính tả hoặc cú pháp.
  + - 1. **Xử lý hóa đơn điện tử (file PDF)**
* Chuyển định dạng file PDF thành IMG để có thể sử dụng model Vintern 1b\_V3.5 thực hiện việc OCR trích xuất thông tin
* Xử lý PDF có cấu trúc lẫn không cấu trúc: Áp dụng OCR tương tự hóa đơn giấy cho các PDF dạng hình ảnh hoặc scan.
* Đối soát dữ liệu: Kiểm tra tính toàn vẹn (checksum).

### Lưu trữ dữ liệu trích xuất vào file Excel

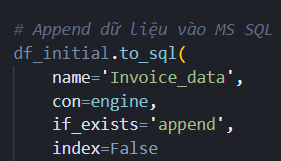
* + - 1. **Xây dựng cấu trúc file Excel với các trường dữ liệu**
* **Xác định các trường chính:** Ngày hóa đơn, mã khách hàng, tên khách hàng, tổng tiền, …
* **Định dạng cột và kiểu dữ liệu:** Thiết lập dữ liệu ngày tháng, số thập phân, văn bản, …
  + - 1. **Sử dụng thư viện Pandas cho quá trình ghi dữ liệu và xử lý làm sạch**
* **Đọc/ghi file Excel:** Dùng pandas.read\_excel() và DataFrame.to\_excel() để thao tác.
* **Làm sạch dữ liệu:** Xử dụng hàm dropna(), fillna(), astype() và str.replace() để chuẩn hóa giá trị thiếu, định dạng, xóa khoảng trắng.
* **Kiểm tra tính hợp lệ:** Áp dụng các điều kiện và báo lỗi hoặc ghi log nếu phát hiện ngoại lệ.

### Đưa dữ liệu từ file Excel vào kho dữ liệu SQL Server

* + - 1. **Quy trình ETL (Extract, Transform, Load)**
* **Extract:** Kết nối và đọc dữ liệu từ file Excel.
* **Transform:** Áp dụng các bước làm sạch, tính toán bổ trợ
* **Load:** Đưa dữ liệu đã xử lý vào bảng staging, sau đó đẩy vào bảng đích trong kho dữ liệu.
  + - 1. **Sử dụng pyodbc hoặc các công cụ hỗ trợ khác cho kết nối SQL Server**
* **Thiết lập kết nối:** Sử dụng chuỗi kết nối ODBC:



* **Thao tác với cơ sở dữ liệu:** Dùng pandas.to\_sql() kết hợp SQLAlchemy hoặc thủ công với cursor.executemany() để batch insert.



### Hỗ trợ truy vấn, báo cáo và phân tích tình hình kinh doanh

* + - 1. **Yêu cầu về báo cáo**
* **Báo cáo tổng hợp:** Doanh thu theo ngày/tuần/tháng/quý/năm; số lượng hóa đơn; doanh thu theo từng khách hang, phân khúc khách hang, ….
* **Báo cáo chi tiết:** chi tiết mặt hàng bán ra.
* **Báo cáo cảnh báo:** Hóa đơn sai định dạng, trùng lặp.
  + - 1. **Công cụ truy vấn**
* **SQL Management Studio :** Cho người dùng kỹ thuật.
* **Power BI / Tableau / Metabase:** Cho người dùng nghiệp vụ, hỗ trợ dashboard trực quan, drill-down.

## Thiết kế hệ thống

### Trích xuất dữ liệu

* + - 1. **Lựa chọn thư viện và công nghệ:**
* **OCR:** Sử dụng mô hình Vintern 1B\_V3.5 là mô hình tích OCR kết hợp với xử lý văn bản nhỏ nhẹ giành cho Tiếng Việt
* **Xử lý PDF:** Chuyển định dạng PDF thành IMG để tận dụng sức mạnh của mô hình Vintern 1B\_V3.5
* **Ngôn ngữ lập trình:** Python (thư viện phong phú)
  + - 1. **Xây dựng mô-đun xử lý ngoại lệ và kiểm tra chất lượng dữ liệu**

### Lưu trữ tạm thời và xử lý dữ liệu

* + - 1. **Quá trình làm sạch dữ liệu sau khi trích xuất**
* **Chuẩn hóa định dạng:** Ngày giờ chuyển về UTC hoặc múi giờ chung, số hợp lệ không chứa ký tự phụ…
* **Kiểm tra trùng lặp:** So sánh số hóa đơn + ngày, nếu trùng báo cảnh báo.
  + - 1. **Sử dụng pandas để xử lý và lưu dữ liệu vào file Excel**

Thực hiện như mô tả tại mục 2.1.2.2.

* + - 1. **Xác minh và kiểm tra tính đầy đủ của dữ liệu trước khi chuyển đổi**
* **Record Count Check:** Số bản ghi trong Excel và số hóa đơn đã xử lý khớp.
* **Integrity Check:** Kiểm tra khóa chính, khóa ngoại, không có giá trị null trong trường bắt buộc.

### Thiết kế kho dữ liệu (Data Warehouse)

* + - 1. **Quy trình ETL**
* **Extract:** trích xuất dữ liệu thô
* **Transform:** Tạo thêm mã tạo đơn cho dữ liệu mới, kiểm tra mã tạo đơn lớn nhất trong kho dữ liệu và tạo mã tạo đơn theo thứ tự tang dần
* **Load:** Đưa dữ liệu vào kho dữ liệu

### Báo cáo và phân tích

* + - 1. **Lựa chọn công cụ báo cáo**

 **Power BI:** Tích hợp dễ dàng với SQL Server, hỗ trợ DAX mạnh mẽ.

* + - 1. **Thiết kế giao diện báo cáo và dashboard**
      2. **Tích hợp hệ thống truy vấn và báo cáo trực tiếp với kho dữ liệu**

Sử dụng API của Gemini 2.0 flash để thực hiện ứng dụng kết nối với kho dữ liệu doanh nghiệp cho phép thực hiện truy vấn và báo cáo trực tiếp bằng cách nhập promt bằng tiếng việt

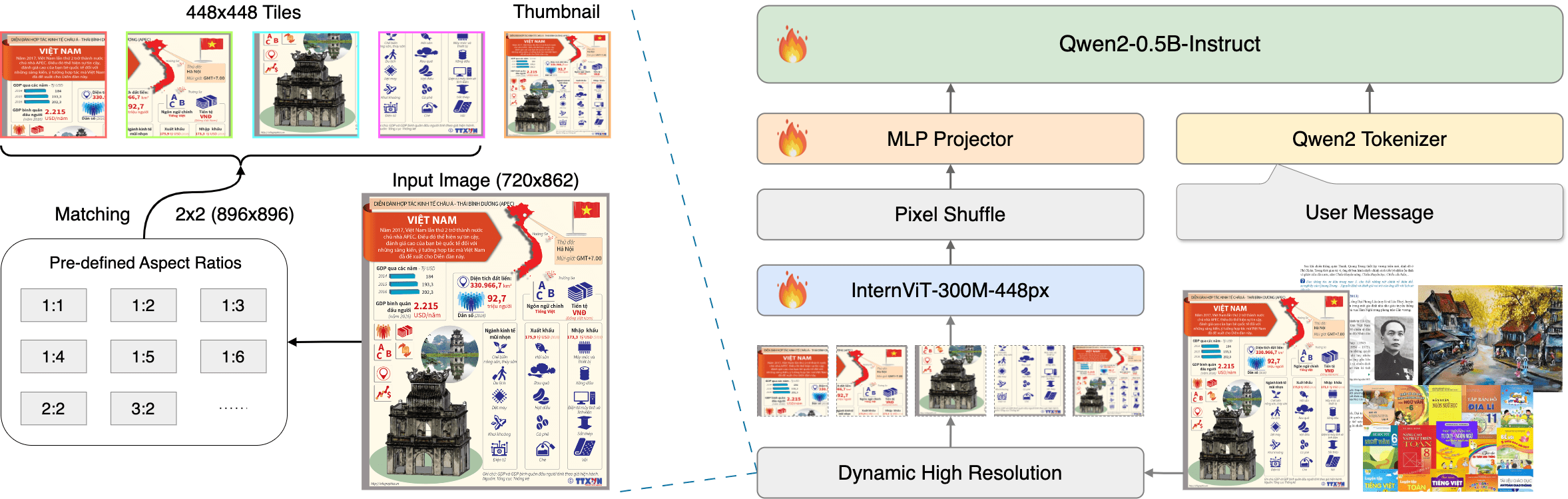
# **CHƯƠNG 3: XÂY DỰNG VÀ TRIỂN KHAI**

## Vintern – Mô hình kết hợp OCR và xử lý văn bản

### 3.1.1 Cấu trúc mô hình

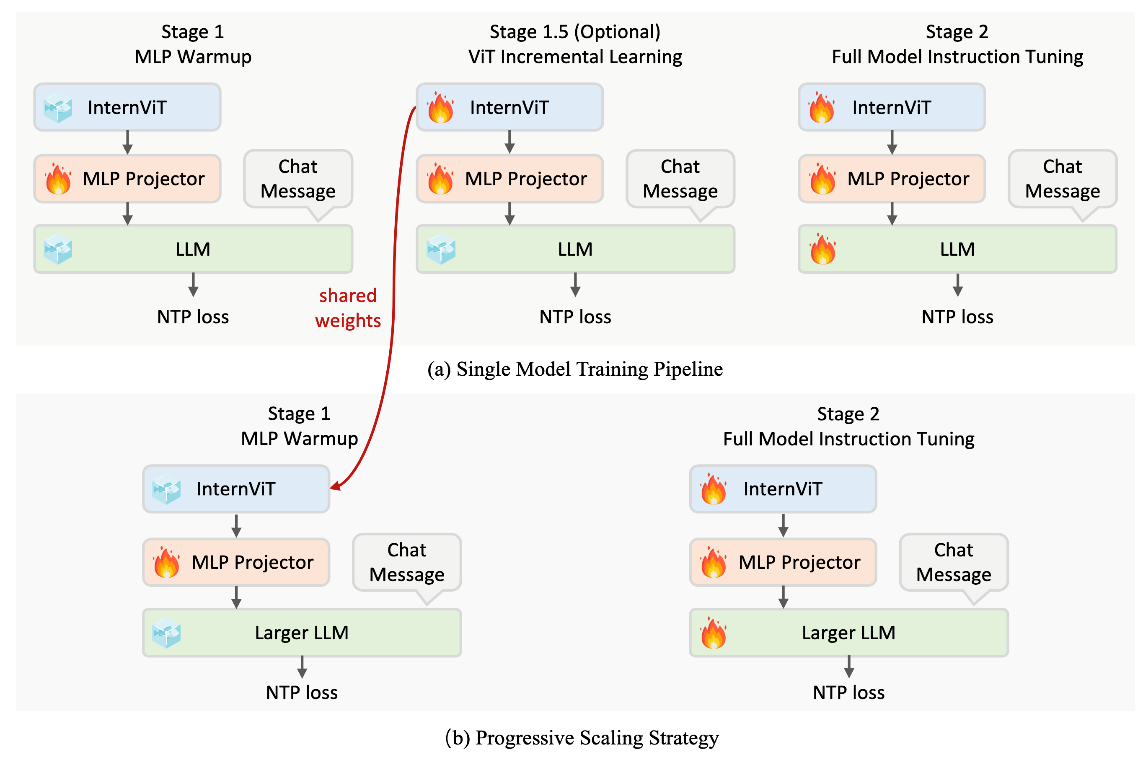
Vintern-1B-v3.5 sử dụng kiến trúc "ViT-MLP-LLM" phổ biến cho các mô hình đa phương thức (MLLM), gồm 3 thành phần chính:

* **InternViT-300M-448px**: Mô hình thị giác, huấn luyện từ InternViT-6B, xử lý ảnh độ phân giải 448x448.
* **Qwen2-0.5B-Instruct**: Mô hình ngôn ngữ đa ngôn ngữ mạnh, hỗ trợ tốt tiếng Việt.
* **MLP projector**: Cầu nối giữa mô hình thị giác và ngôn ngữ, giúp liên kết thông tin hình ảnh và văn bản hiệu quả.



Hình 1: Kiến trúc tổng thể

Hình ảnh đầu vào được chia thành các phần 448x448 pixel cùng với một ảnh thu nhỏ, rồi đưa vào InternViT-300M để trích xuất đặc trưng. Sau đó, qua bước Pixel Shuffle và MLP projector, các đặc trưng được căn chỉnh với embedding của Qwen2-0.5B-Instruct. Mô hình ngôn ngữ này tiếp nhận các token thị giác đã căn chỉnh cùng câu hỏi để tạo ra câu trả lời.



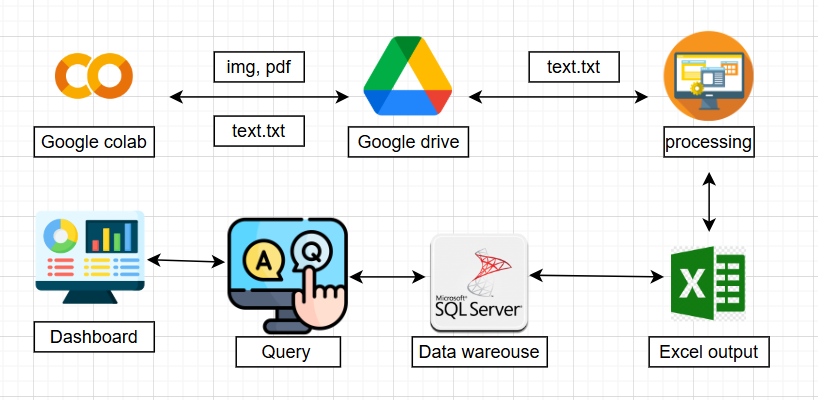
Hình 2: Single Model Training Pipeline

Pipeline huấn luyện được thiết kế thành ba giai đoạn nhằm nâng cao khả năng nhận thức hình ảnh và xử lý đa phương thức:

* **Giai đoạn 1 – Khởi động MLP:** Chỉ huấn luyện MLP, trong khi vision encoder và mô hình ngôn ngữ bị đóng băng. Áp dụng chiến lược độ phân giải cao động để tăng hiệu quả liên kết đa mô thức.
* **Giai đoạn 1.5 – Học tăng cường ViT (tùy chọn):** Huấn luyện thêm vision encoder cùng MLP nhằm cải thiện khả năng trích xuất đặc trưng, đặc biệt hữu ích với dữ liệu ít phổ biến như OCR hoặc biểu đồ. Encoder sau đó có thể tái sử dụng mà không cần huấn luyện lại.
* **Giai đoạn 2 – Huấn luyện đa phương thức:** Toàn bộ mô hình được huấn luyện với dữ liệu hướng dẫn đa phương thức chất lượng cao. Kiểm soát dữ liệu nghiêm ngặt để tránh làm giảm chất lượng đầu ra của mô hình ngôn ngữ.

Đây là một lựa chọn đáng cân nhắc cho bất kỳ ai đang tìm kiếm một giải pháp OCR và xử lý văn bản hiệu quả, tối ưu cho ngôn ngữ và bối cảnh Việt Nam.

## Tổng quan sơ đồ hệ thống

****

Hình 3: Tổng quan hệ thống

## Trích xuất và xử lý dữ liệu từ hóa đơn

### Xử lý hóa đơn giấy

* + - 1. **Số hóa hóa đơn**

**Thiết lập quy trình scan**

* Sử dụng máy scan/phần mềm scanner tự động quét hàng loạt hoặc đơn giản hơn là chụp bằng điện thoại
* Đặt chuẩn DPI (300–600 dpi) để cân bằng chất lượng và kích thước file

**Lưu trữ hình ảnh**

Định dạng JPG hoặc PNG giữ độ nét cao

* + - 1. **Áp dụng OCR – Model Vintern 1B\_ V3.5**

Đây có thể nói là một **lựa chọn sáng giá** có thể quyết nhanh các bài toán OCR tiếng Việt nhưng vẫn khả dụng trên phần cứng phổ thông.

### Xử lý hóa đơn điện tử (file PDF)

Với sức mạnh của Vintern 1B\_V3.5 , việc sử lý hoá đơn điện tử trở nên đơn giản hơn rất nhiều, chỉ cần chuyển định dạng của PDF sang định dạng của ảnh như JPG hay PNG và Vintern sẽ thực hiện việc còn lại

## Lưu trữ tạm thời và làm sạch dữ liệu

### Làm sạch dữ liệu

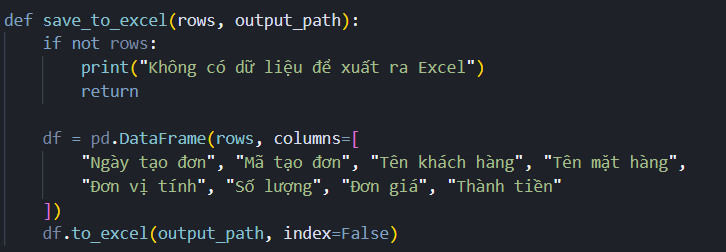
**Chuẩn hóa định dạng trường:**

* + Ngày tháng → MM-DD-YYYY với pd.to\_datetime().
  + Số tiền → xóa dấu phân cách, chuyển kiểu float, bổ sung thêm 000 cho chuẩn hoá đơn vị tiền tệ VND

### Lưu trữ vào file Excel

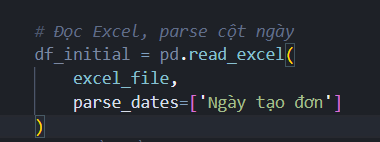
 **Xác định template:** file mẫu có header, danh sách Lookup.

 **Ghi dữ liệu:**

****

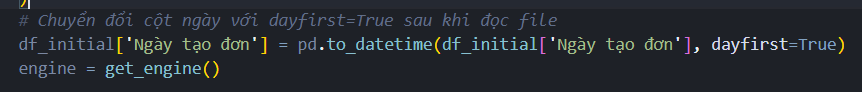
## Quá trình ETL – Đưa dữ liệu vào kho dữ liệu SQL Server

### Extract – Trích xuất dữ liệu

Đọc từ file Excel 

### Transform – Chuyển đổi dữ liệu

* + - 1. **Làm sạch và định dạng lại dữ liệu**

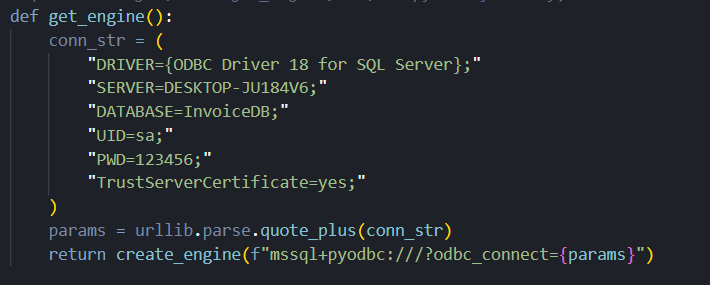


* + - 1. **Chuẩn hóa dữ liệu**

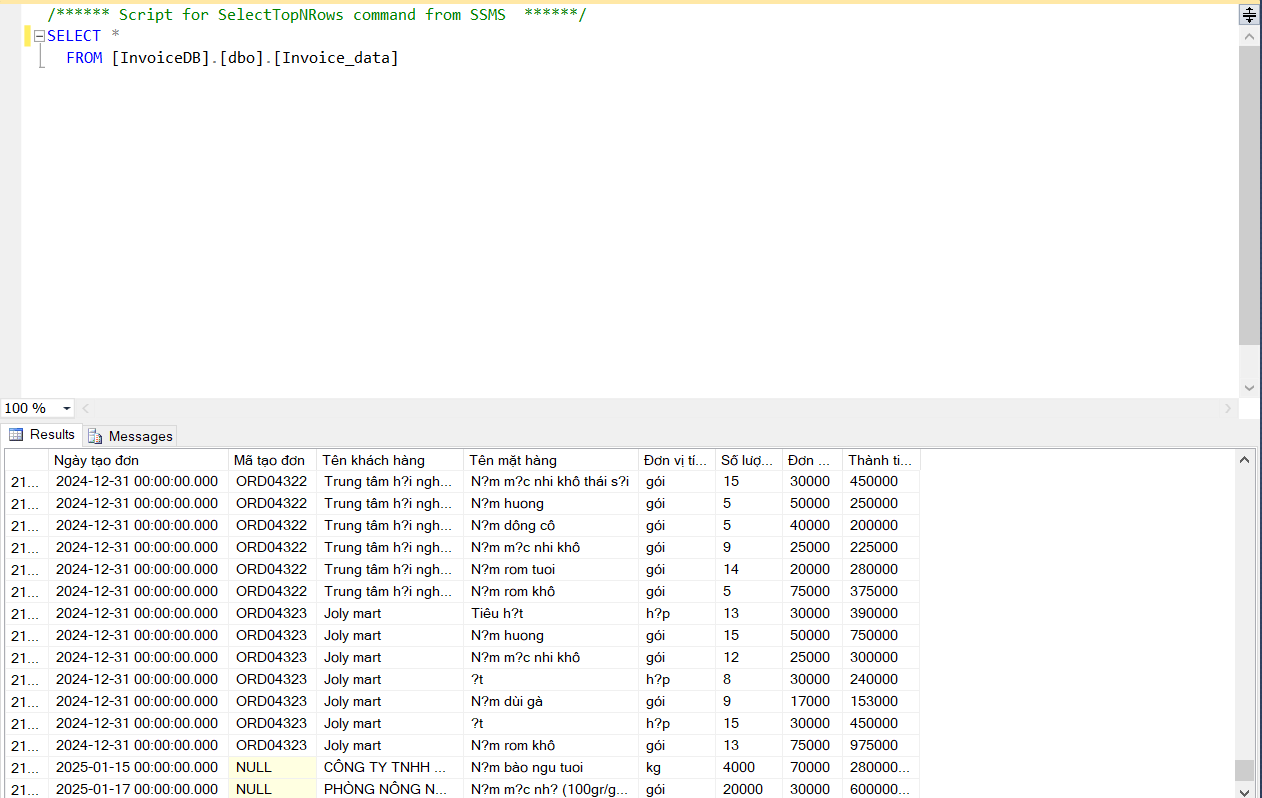
Các bước chuẩn hoá : Gán Mã tạo đơn tăng dần cho từng nhóm đơn => Merge về dataframe chi tiết=> Sắp xếp lại cột giống bảng MSSQL **Load – Nạp dữ liệu vào SQL Server**

* + - 1. **Kết nối và truyền dữ liệu (sử dụng pyodbc hoặc công cụ ETL)**

Chuỗi kết nối ODBC:

****

* + - 1. **Kiểm tra sau khi nạp dữ liệu**



## Thiết lập hệ thống báo cáo và phân tích dữ liệu

### Cấu hình công cụ báo cáo

Xây dựng ứng dụng để thực hiện truy vấn và báo cáo kết quả kinh doanh bằng ngôn ngữ tự nhiên tiếng Việt bằng cách gõ promt, ví dụ: Cơ cấu doanh thu của từng mặt hàng

**Tổng quan:** ứng dụng cho phép người dùng tạo và phân tích biểu đồ cơ cấu doanh thu thông qua việc nhập prompt tự nhiên.

Workflow Xử Lý Prompt Vẽ Biểu Đồ Cơ Cấu Doanh Thu

**1. Nhận Prompt từ Người Dùng**

* Người dùng nhập prompt mô tả yêu cầu vẽ biểu đồ cơ cấu doanh thu
* Ví dụ: "Vẽ biểu đồ cơ cấu doanh thu theo ngành hàng năm 2023"

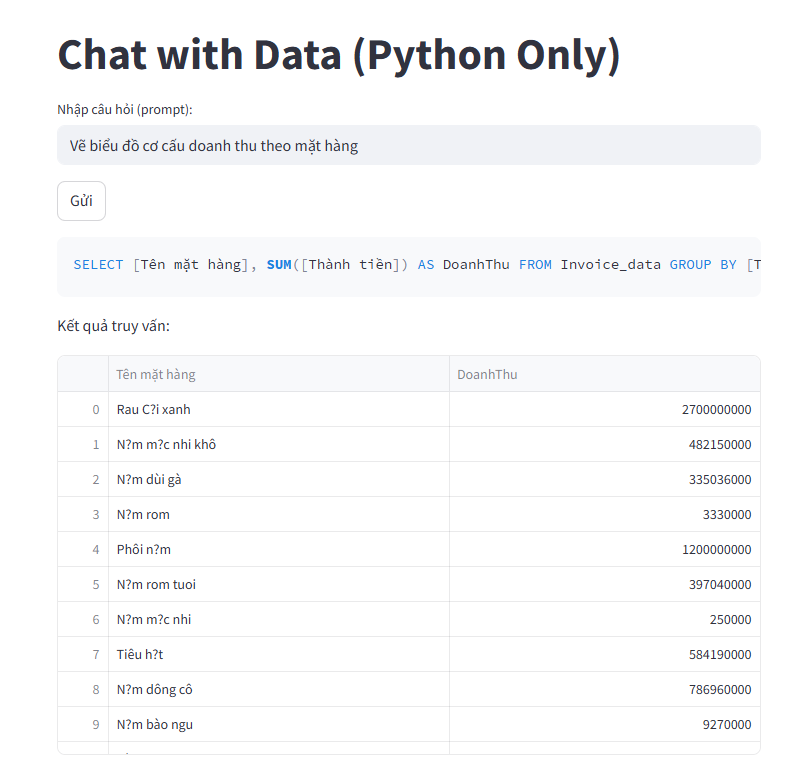
**2. Xử Lý và Phân Tích Prompt**

* Hệ thống phân tích prompt để xác định:
* Loại biểu đồ cần vẽ (pie chart, bar chart, etc.)
* Thời gian phân tích (năm, quý, tháng)
* Các tiêu chí phân loại doanh thu
* Các yêu cầu đặc biệt về định dạng hoặc hiển thị

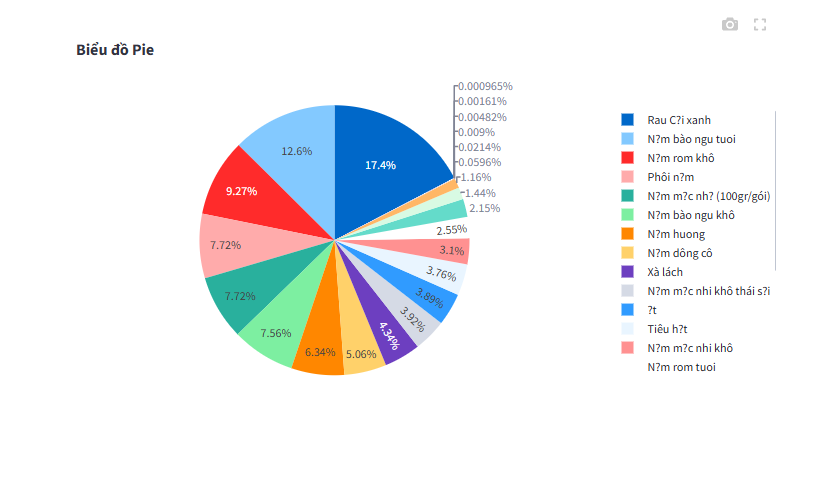
**3. Truy Vấn Dữ Liệu**

- Hệ thống truy vấn cơ sở dữ liệu để lấy thông tin doanh thu

* Lọc dữ liệu theo các tiêu chí đã xác định từ prompt
* Tính toán các chỉ số cần thiết (tỷ lệ phần trăm, tổng doanh thu, etc.)



1. **Tạo Biểu Đồ và Hiển thị kết quả**



### Quy trình triển khai báo cáo

Tích hợp và tự động hóa cập nhật báo cáo: Sử dụng Apache Airflow để thực hiện việc lập lịch hằng ngày

# **CHƯƠNG 4: KHUYẾN NGHỊ VÀ KẾT LUẬN**

## 4.1 Kết luận

### 4.1.1 Tích hợp công nghệ và quy trình tự động

Hệ thống đã thành công trong việc kết hợp các công nghệ OCR, xử lý PDF và thư viện Python (Pandas, pyodbc) để tự động hóa quy trình trích xuất dữ liệu từ cả hóa đơn giấy và hóa đơn điện tử. Việc tự động hóa giúp giảm thiểu thao tác thủ công, tiết kiệm thời gian và nâng cao độ chính xác.

### 4.1.2 Hiệu quả trong lưu trữ và quản lý dữ liệu

Việc lưu trữ tạm thời qua file Excel với các bảng tra cứu (lookup tables) và chuyển tiếp vào kho dữ liệu SQL Server đã đảm bảo tính toàn vẹn, dễ truy vấn và dễ bảo trì. Các bước làm sạch và kiểm tra dữ liệu liên tục loại bỏ được bản ghi lỗi và trùng lặp, đảm bảo chất lượng dữ liệu đầu ra.

### 4.1.3 Khả năng mở rộng và tích hợp báo cáo

Kiến trúc Data Warehouse cho phép mở rộng dễ dàng . Hệ thống báo cáo (Power BI, Tableau) kết nối trực tiếp hoặc qua Data Mart, cho phép tạo dashboard đa dạng, drill-down và tự động refresh, phục vụ nhu cầu phân tích kinh doanh linh hoạt.

### Đánh giá tổng thể và hạn chế gặp phải

* **Ưu điểm:** Tự động hóa cao, giảm thiểu lỗi nhập liệu, báo cáo nhanh chóng, khả năng mở rộng và tích hợp tốt.
* **Hạn chế:** Chất lượng OCR phụ thuộc vào độ nét của ảnh scan; với PDF scan kém chất lượng, tỷ lệ trích xuất chính xác giảm. Quy trình ETL ban đầu cần tinh chỉnh nhiều để xử lý các trường hợp ngoại lệ.

## 4.2 Khuyến nghị

### 4.2.1 Cải thiện quá trình số hóa hóa đơn giấy

* Đầu tư scanner có tính năng tự động cân bằng sáng và tự động cắt khung ảnh.
* Thiết lập tiêu chuẩn scan (độ phân giải, định dạng) nghiêm ngặt và hướng dẫn vận hành rõ ràng cho nhân viên vận hành.

### 4.2.2 Nâng cao chất lượng xử lý file PDF

* Sử dụng kết hợp nhiều engine OCR (ví dụ: Tesseract + Google Vision) để so sánh và chọn kết quả tốt nhất.
* Với các hóa đơn điện tử có cấu trúc XML nhúng, ưu tiên parse trực tiếp thay vì OCR để đảm bảo độ chính xác 100%.

### 4.2.3 Tối ưu hóa quy trình ETL

* Xây dựng pipeline ETL sử dụng Apache Airflow hoặc Azure Data Factory để tự động hóa, lên lịch và quản lý dependency giữa các bước.

### 4.2.4 Mở rộng khả năng báo cáo và phân tích dữ liệu

* Triển khai thêm các mô hình Advanced Analytics (như dự báo doanh thu, phân tích giỏ hàng) sử dụng Python/R kết hợp với Power BI.
* Xây dựng API RESTful để cấp dữ liệu báo cáo cho các ứng dụng web/mobile.

### 4.2.5 Đảm bảo bảo mật và sao lưu dữ liệu

* Thiết lập chính sách backup & restore tự động cho SQL Server (full backup, differential, transaction log).
* Mã hóa dữ liệu nhạy cảm (PII)
* Quản lý phân quyền truy cập nghiêm ngặt, audit log các hoạt động thay đổi dữ liệu.

# **TÀI LIỆU THAM KHẢO**

[1] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (Third Edition). Indianapolis, IN: Wiley.

[2] IBM Analytics, “Statistical Foundations for Business Analytics” <https://www.ibm.com/think/topics/business-analytics>

[3] IBM Analytics, “Data Mining and Machine Learning”.

<https://www.ibm.com/think/topics/data-mining>

<https://www.ibm.com/think/topics/machine-learning>

[4] Statista, “Data Visualization in Business Analytics”

[5] IBM Analytics, “Business Reporting and Dashboarding”.

<https://www.ibm.com/docs/en/pm-and-q/2.6.3?topic=product-reports-dashboards>

[6] [Vintern-1B-v3.5] Mô hình kết hợp OCR và xử lý văn bản nhỏ nhẹ dành cho tiếng Việt

[https://viblo.asia/p/vintern-1b-v35-mo-hinh-ket-hop-ocr-va-xu-ly-van-ban-nho-nhe-danh-cho-tieng-viet-AZoJj8WOVY7#\_22-cau-truc-mo-hinh-3](https://viblo.asia/p/vintern-1b-v35-mo-hinh-ket-hop-ocr-va-xu-ly-van-ban-nho-nhe-danh-cho-tieng-viet-AZoJj8WOVY7%23_22-cau-truc-mo-hinh-3)

[7] <https://huggingface.co/5CD-AI/Vintern-1B-v3_5>

[8] <https://github.com/OpenGVLab/InternVL>