

ỦY BAN NHÂN DÂN TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN



Họ và tên sinh viên : Võ Quang Đăng Khoa

BÁO CÁO
THỰC TẬP TỐT NGHIỆP

Công ty thực tập : Công ty CP Giải Pháp Dệt May Bền Vững
Chuyên gia hướng dẫn : Ngô Trí Thanh
Giảng viên hướng dẫn : PGS.TS Phạm Thế Bảo

TP. Hồ Chí Minh, tháng 8 năm 2024

**ỦY BAN NHÂN DÂN TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN**



Họ và tên sinh viên : Võ Quang Đăng Khoa

**BÁO CÁO
THỰC TẬP TỐT NGHIỆP**

Công ty thực tập : Công ty CP Giải Pháp Dệt May Bền Vững
Chuyên gia hướng dẫn : Ngô Trí Thanh
Giảng viên hướng dẫn : PGS.TS Phạm Thế Bảo

TP. Hồ Chí Minh, tháng 8 năm 2024

MỤC LỤC

MỤC LỤC.....	i
MỤC LỤC HÌNH.....	iii
MỤC LỤC BẢNG	v
NHẬN XÉT CỦA CHUYÊN GIA DOANH NGHIỆP	vi
NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN.....	vii
LỜI MỞ ĐẦU.....	1
CHƯƠNG 1. GIỚI THIỆU	2
1.1. Giới thiệu công ty thực tập.....	2
1.2. Cơ cấu tổ chức	2
1.3. Cơ sở vật chất của doanh nghiệp.....	3
1.4. Lĩnh vực hoạt động.....	5
1.5. Quy trình của công ty.....	6
1.6. Nhiệm vụ thực tập	6
1.7. Kết luận chương 1	7
CHƯƠNG 2. Quá trình thực tập.....	7
2.1. Quá trình thực tập theo tuần.....	7
2.2. Tìm hiểu về Data Warehouse, Data Mart và các mô hình thiết kế.....	9
2.2.1 Khái niệm về Data Warehouse, Data Mart	9
2.2.2 Các mô hình.....	11
2.3. Tìm hiểu Kiến trúc của hệ thống BI	13
2.4. Tìm hiểu dữ liệu và các thuật ngữ liên quan	15
2.4.1 Các trường dữ liệu	15
2.4.2 HS Code	18
2.4.3 Incoterm	19
2.5. Phân tích thiết kế ERD.....	21

2.5.1 Các bảng và cột.....	21
2.6. Phân tích thiết kế Common Summary Data Value.....	34
2.7. Làm sạch dữ liệu lần 1	37
2.8. Làm sạch dữ liệu lần 2	39
2.9. Kiểm thử dữ liệu.....	40
2.9.1 Mục tiêu	40
2.9.2 Quy trình kiểm thử.....	41
2.9.3 Kết quả kiểm thử năm 2022	42
2.9.4 Kết luận.....	43
2.10. Tải dữ liệu lên hệ thống của công ty.....	43
2.10.1 Chuẩn bị.....	43
2.10.2 Hướng dẫn sử dụng	43
2.11. Cào dữ liệu doanh nghiệp xuất nhập khẩu để làm MasterData	47
2.11.1 Lấy thông tin doanh nghiệp cần thu thập dữ liệu	47
2.11.2 Sử dụng công cụ làm giàu thông tin các doanh nghiệp	50
2.11.3 Sử dụng công cụ để thu thập dữ liệu doanh nghiệp.....	51
2.11.4 Kiểm thử dữ liệu đã thu thập.....	60
CHƯƠNG 3. KẾT QUẢ THỰC TẬP	64
3.1. Kết quả thực tập.....	64
3.2. Các biểu mẫu đánh giá.....	65
CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ	74
4.1. Kết luận.....	74
4.2. Kiến nghị	74
TÀI LIỆU THAM KHẢO	75

MỤC LỤC HÌNH

Hình 1.1: Sơ đồ cơ cấu tổ chức STS	2
Hình 1.2: Không gian tầng trệt tại công ty	3
Hình 1.3: Không gian tầng trệt 200m ² với sức chứa lên đến 100 người	4
Hình 1.4: Không gian tầng 2 tại công ty	4
Hình 1.5: Không gian trình bày các sản phẩm may mặc và trình diễn thời trang	5
Hình 1.6: Cơ sở vật chất tại công ty CP Giải Pháp Dệt May Bền Vững	5
Hình 1.7: Danh sách các nhiệm vụ được giao trên Jira	6
Hình 2.1: Minh họa Data Warehouse	10
Hình 2.2: Mô hình Star Schema	11
Hình 2.3: Mô hình Snowflake Schema	12
Hình 2.4: Mô hình Galaxy Schema	12
Hình 2.5: Kiến trúc của hệ thống BI	13
Hình 2.6: Cấu trúc của dãy số HS Code	18
Hình 2.7: Bảng tra cứu mã HS Code 63	19
Hình 2.8: Các quy tắc incoterms của năm 2020	20
Hình 2.9: Sơ đồ ERD	21
Hình 2.10: Mô hình giải pháp common summary data value	35
Hình 2.11: Sơ đồ Common Summary và Data Value	36
Hình 2.12: Tập dữ liệu có trộn lẫn các tháng	37
Hình 2.13: Tập dữ liệu chỉ xuất hiện một tháng duy nhất	37
Hình 2.14: Code truyền đường dẫn làm sạch dữ liệu	38
Hình 2.15: Code truyền đường dẫn làm sạch dữ liệu	38
Hình 2.16: Kết quả sau khi làm sạch dữ liệu lần 1	39
Hình 2.17: code truyền đường dẫn bộ dữ liệu 63 tỉnh thành VN	39
Hình 2.18: Code truyền đường dẫn bộ dữ liệu	40
Hình 2.19: Kết quả sau khi làm sạch dữ liệu lần 2	40
Hình 2.20: Bảng kiểm tra đối chiếu số liệu để kiểm thử	42
Hình 2.21: Code cài đặt thư viện trước khi tải dữ liệu lên PostgreSQL	43
Hình 2.22: Code thiết lập kết nối tới server PostgreSQL	44

Hình 2.23: Script tạo các schema và bảng khi chưa tồn tại bằng query tool.....	45
Hình 2.24: Code tạo các schema và bảng khi chưa tồn tại	45
Hình 2.25: Code gom tắt các file giao dịch .csv lại thành 1 file duy nhất.....	45
Hình 2.26: Code thực thi tải dữ liệu lên PostgreSQL	46
Hình 2.27: Kết quả sau khi đã tải dữ liệu lên	47
Hình 2.28: Code truyền đường dẫn để thực thi lấy thông tin doanh nghiệp.....	48
Hình 2.29: Kết quả sau khi chạy code lấy thông tin doanh nghiệp.....	48
Hình 2.30 Danh sách mã số thuế của doanh nghiệp cần lấy thông tin.....	49
Hình 2.31: Danh sách nhà cung cấp và địa chỉ cần lấy thông tin.....	50
Hình 2.32: Code truyền đường dẫn để thực thi làm giàu thông tin doanh nghiệp.....	50
Hình 2.33: Kết quả sau khi chạy code làm giàu thông tin doanh nghiệp.....	51
Hình 2.34: Dữ liệu doanh nghiệp của giao dịch nhập khẩu cần thu thập.....	52
Hình 2.35: File chứa kết quả thu thập thông tin doanh nghiệp	52
Hình 2.36: Code dùng để thu thập doanh nghiệp nhập khẩu.....	53
Hình 2.37: Code dùng để thu thập doanh nghiệp nhập khẩu.....	53
Hình 2.38: Danh sách các mã số thuế cần thu thập dữ liệu	54
Hình 2.39: Tiến trình thực hiện thu thập dữ liệu doanh nghiệp.....	55
Hình 2.40: Tiến trình thực hiện tìm kiếm trên trang masothue.com.....	55
Hình 2.41: kết quả thu thập dữ liệu trên console.....	56
Hình 2.42: Các lỗi xảy ra trong quá trình thu thập dữ liệu	56
Hình 2.43: Kiểm tra dataframe trong quá trình chạy tool thu thập dữ liệu.....	57
Hình 2.44: Đoạn code cần thực hiện khi có lỗi xảy ra trong quá trình thu thập	57
Hình 2.45: Kết quả thu thập dữ liệu doanh nghiệp của nguồn dữ liệu nhập khẩu.....	58
Hình 2.46: File dữ liệu đầu vào cho quá trình thu thập doanh nghiệp nhập khẩu	59
Hình 2.47: Code truyền đường dẫn để thu thập dữ liệu doanh nghiệp.....	59
Hình 2.48: Kết quả thu thập dữ liệu doanh nghiệp xuất khẩu.....	59
Hình 2.49: Dữ liệu doanh nghiệp của năm 2019 đã thu thập	60
Hình 2.50: Kết quả kiểm thử dữ liệu doanh nghiệp đã thu thập được.....	61

MỤC LỤC BẢNG

Bảng 2-1: Quá trình thực tập theo tuần	9
Bảng 2-2: Các trường dữ liệu thô.....	18
Bảng 2-3: Bảng FactTransaction	25
Bảng 2-4: Bảng Date.....	25
Bảng 2-5: Bảng Currency.....	26
Bảng 2-6: Bảng PaymentMethod	26
Bảng 2-7: Bảng Incoterms.....	27
Bảng 2-8: Bảng QuantityUnit.....	27
Bảng 2-9: Bảng Transportation.....	27
Bảng 2-10: Bảng TypeOfTransaction	28
Bảng 2-11: Bảng Country.....	28
Bảng 2-12: Bảng Product.....	29
Bảng 2-13: Bảng BondedWarehouse	29
Bảng 2-14: Bảng Port	30
Bảng 2-15: bảng InterpriseInformation	32
Bảng 2-16: Bảng Location	33
Bảng 2-17: Bảng GeographicalRegion.....	33
Bảng 2-18: Bảng Administrative Region	34

NHẬN XÉT CỦA CHUYÊN GIA DOANH NGHIỆP

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

LỜI MỞ ĐẦU

Trong bối cảnh cạnh tranh ngày càng khốc liệt, ngành dệt may đòi hỏi sự cải tiến liên tục và tối ưu hóa quy trình kinh doanh. Data Warehouse (kho dữ liệu) trở thành một công cụ thiết yếu, giúp các doanh nghiệp dệt may quản lý và phân tích dữ liệu từ nhiều nguồn khác nhau, từ sản xuất, tồn kho, vận chuyển đến bán hàng và dịch vụ hậu mãi.

Kho dữ liệu giúp các doanh nghiệp theo dõi và quản lý toàn bộ dữ liệu xuất nhập khẩu, dự đoán nhu cầu khách hàng, và đưa ra các quyết định chiến lược chính xác, từ đó nâng cao năng suất, giảm thiểu rủi ro và tối ưu hóa lợi nhuận. Đồng thời, Data Warehouse cũng hỗ trợ tuân thủ các quy định và tiêu chuẩn quốc tế trong hoạt động xuất nhập khẩu, tăng cường khả năng cạnh tranh trên thị trường toàn cầu.

Hiện tại, em đang thực tập tại Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững, tham gia vào việc xây dựng hệ thống Data Warehouse. Trong kỷ nguyên kinh tế số, việc triển khai Data Warehouse là nền tảng không thể thiếu để các doanh nghiệp xuất nhập khẩu trong ngành dệt may phát triển bền vững và đổi mới sáng tạo.

CHƯƠNG 1. GIỚI THIỆU

1.1. Giới thiệu công ty thực tập



Tên công ty: Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững

Địa chỉ: A07-08 tòa Sarica, Đ. D9, KĐT Sala, P. An Lợi Đông, Q.2, TP. HCM

Email: info@stsgroup.org.vn

Số điện thoại: +84 8 1887 2887

Người đại diện pháp luật: Nguyễn Ngọc Khánh Nhật (Judy Nguyễn)

STS - CTCP Giải Pháp Dệt May Bền Vững, với tầm nhìn đưa ngành dệt may Việt Nam tiệm cận hơn với những tiêu chí phát triển bền vững bằng cách tác động tích cực đến từng dấu chân trong chuỗi giá trị ngành, được tổ chức bởi những người tâm huyết và các doanh nghiệp đang phát triển (SMEs) trong ngành dệt may, cùng với sự đóng góp của đội ngũ nhân viên gen Z và công nghệ AI.

1.2. Cơ cấu tổ chức



Hình 1.1: Sơ đồ cơ cấu tổ chức STS

Công ty cổ phần giải pháp dệt may bền vững (STS) được chia thành 5 phòng ban chính gồm: VTIC (Vietnam Textile Information Center), VTBI (Vietnam Textile Business Intelligence), VFDC (Vietnam Fabric Discovery Center), VTIH (Vietnam Textile Innovation Hub) và VTAF (Vietnam Textile & Apparel Forum). Sự kết hợp của 5 phòng ban đã tạo nên một hoạt động hoàn chỉnh và hiệu quả cho công ty STS, đảm bảo sự phát triển bền vững và đột phá trong ngành dệt may Việt Nam.

Cụ thể hơn về từng phòng ban:

- + Phòng VTIC sẽ chịu trách nhiệm xây dựng nên 1 cổng thông tin thương mại thị trường dệt may Việt Nam được cập nhật đầy đủ, liên tục và có độ tin cậy cao.
- + Phòng VTBI sẽ cung cấp báo cáo xu hướng thị trường & phân tích chuyên sâu về ngành dệt may, cung cấp các thông tin thương mại quốc tế minh bạch.
- + Phòng VFDC sẽ hoạt động như một showroom hiện đại và sáng tạo, nơi trưng bày các sản phẩm vải độc đáo và tiên tiến nhất. Tại đây, khách hàng và đối tác có thể trực tiếp trải nghiệm, kiểm tra chất lượng và khám phá các xu hướng mới nhất trong ngành công nghiệp vải, tạo nên một không gian giao lưu và hợp tác hiệu quả.
- + Phòng VTIH chịu trách nhiệm về Higg Fem 4.0, hay Mô-đun Môi trường Cơ sở Higg, là một công cụ đánh giá tính bền vững nhằm chuẩn hóa cách đo lường và đánh giá hiệu quả giảm thiểu tác động môi trường của các Cơ sở, theo từng năm. Đồng thời, VTIH cũng chịu trách nhiệm tổ chức các lớp đào tạo và huấn luyện định kỳ hàng tháng nhằm thúc đẩy và nâng cao Kinh doanh bền vững và Thực hành công nghiệp có trách nhiệm trong ngành Dệt may, Da - Giày.
- + Phòng VTAF chính là kênh truyền thông và marketing của STS, với mục tiêu “Am hiểu địa phương, làm việc toàn cầu”, VTAF sẽ lập kế hoạch và tổ chức các sự kiện online/offline/hybrid như Triển lãm xúc tiến thương mại nhằm nâng cao giá trị thương hiệu doanh nghiệp Việt Nam.

1.3. Cơ sở vật chất của doanh nghiệp

Không gian tầng trệt với những thiết kế độc đáo đào tạo cảm hứng cho quá trình triển khai những ý tưởng cho công việc. Với diện tích 200 m² cùng sức chứa lên đến 100 người.



Hình 1.2: Không gian tầng trệt tại công ty



Hình 1.3: Không gian tầng trệt 200m² với sức chứa lên đến 100 người

Tiếp theo, không gian tầng 2 với diện tích ~ 200 m² là không gian sáng tạo qua việc sử dụng những thiết kế độc đáo cho nội thất và trên các bức tường được trang trí lấy cảm hứng từ câu chuyện của ngành dệt may Việt Nam. Tầng 2 được trang bị những trang thiết bị hiện đại, tạo nên không gian hoàn hảo cho những sự kiện trưng bày, hội thảo hay thậm chí là một buổi trình diễn thời trang.



Hình 1.4: Không gian tầng 2 tại công ty



Hình 1.5: Không gian trình bày các sản phẩm may mặc và trình diễn thời trang

VENUE	ROOM TYPE (Loại Phòng)	CAPACITY (Sức Chứa)	ACREAGE (Diện Tích)	BENEFITS EQUIPMENT (Trang Thiết bị kèm theo)
STS SPACE	1st FLOOR (Lầu 1)	Only space (~100pax)	200m2 (10 * 12.5m2)	<ul style="list-style-type: none"> • Reception table Bàn tiếp tân • Wellcome-board Bục phát biểu • Standard sound and light system Hệ thống âm thanh, ánh sáng theo tiêu chuẩn • 02 Wireless microphones 02 micro không dây/sàn
	A008 2nd FLOOR (Lầu 2)	Only space (~80pax)	100m2 (9.5 * 10.4m)	
	A007 2nd FLOOR (Lầu 2)	Only space (~80pax)	100m2 (8 * 12.5m)	<ul style="list-style-type: none"> • 1 Flat TV 65inch 1 TV màn hình phẳng 65inch • 6 table (0.8mx1.2m), 20 chair 6 bàn (0.8mx1.2m), 20 ghế
	Full 2nd FLOOR (Toàn Lầu 2)	100 - 150 pax	200m2	

Hình 1.6: Cơ sở vật chất tại công ty CP Giải Pháp Dệt May Bền Vững

1.4. Lĩnh vực hoạt động

Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững hoạt động chủ yếu trong lĩnh vực dệt may, tập trung vào việc phát triển các sản phẩm dệt may bền vững và kinh tế tuần hoàn. Với mục tiêu tạo ra các sản phẩm thân thiện với môi trường và giảm thiểu tác động tiêu cực đến hệ sinh thái, công ty không chỉ dừng lại ở việc sản xuất mà còn tích cực tham gia vào các hoạt động nghiên cứu và phát triển (R&D) nhằm cải tiến và đổi mới quy trình sản xuất. Bên cạnh đó, công ty còn xây dựng một nền tảng thương mại điện tử và cộng đồng thiết kế, tạo điều kiện cho các nhà thiết kế và người tiêu dùng

dễ dàng tiếp cận và tương tác với các sản phẩm bền vững. Việc kết hợp giữa thương mại điện tử và cộng đồng thiết kế không chỉ giúp mở rộng thị trường mà còn thúc đẩy việc áp dụng các giải pháp bền vững trong ngành dệt may, góp phần vào sự phát triển bền vững của ngành công nghiệp này.

1.5. Quy trình của công ty

Quy trình xây dựng data warehouse của công ty được theo dõi và quản lý bằng công cụ làm việc nhóm Jira. Với các nhiệm vụ đã được thảo luận trước đó, leader của nhóm sẽ giao cho từng người trên công cụ kèm theo deadline và tài liệu tìm hiểu (nếu có).

Type	Key	Summary	Status	Assignee	Due date	Labels	Created	Updated	
>	DTV-1	Xây dựng Data Warehouse	IN PROGRESS	Thanh Ngô Tri			Jun 6, 2024	Jun 6, 2024	
>	DTV-2	Hoàn thành tool clean dữ liệu	IN PROGRESS	Thanh Ngô Tri			Jun 6, 2024	Jun 6, 2024	
>	DTV-3	Clean và xử lý dữ liệu 2012-2023	IN PROGRESS	Thanh Ngô Tri			Jun 6, 2024	Jun 6, 2024	
>	DTV-4	Kiểm thử và đối chiếu dữ liệu	IN PROGRESS	Phương Vy			Jun 6, 2024	Jun 6, 2024	
>	DTV-5	Kiểm thử Data Warehouse	TO DO	Huỳnh Phúc Toàn			Jun 6, 2024	Jun 6, 2024	
>	DTV-29	Dựng tool crawl các dữ liệu thông tin doanh nghiệp	IN PROGRESS	Huỳnh Phúc Toàn			Jun 6, 2024	Jun 6, 2024	
>	DTV-30	Crawl các dữ liệu doanh nghiệp để input vào Data Warehouse	IN PROGRESS	Võ Quang Đăng Khoa			Jun 6, 2024	Jun 27, 2024	
	DTV-79	Xây dựng tool Load dữ liệu vào PostgreSQL	DONE	Huỳnh Phúc Toàn	Jun 13, 2024		Jun 12, 2024	Jun 13, 2024	

Hình 1.7: Danh sách các nhiệm vụ được giao trên Jira

1.6. Nhiệm vụ thực tập

Các nhiệm vụ mà em được phân công và thực hiện trong kỳ thực tập này bao gồm:

- Nghiên cứu các tài liệu, thuật ngữ ngành liên quan đến dự án.
- Nghiên cứu hiện trạng dữ liệu thô xuất nhập khẩu dệt may, tìm hiểu các lỗi dữ liệu dự kiến xảy ra.
- Tìm hiểu các khái niệm về Master Data Repos, Infor Sec Policy, Organization Hierarchies, Data Retention Policy... và các khái niệm liên quan.
- Tìm hiểu thêm các khái niệm về Database, Data Warehouse, Data Mart, Big Data, Pipe Data, Galaxy Schema, Snowflake Schema, Star Schema...
- Làm sạch bước 1 dữ liệu thô và xây dựng tool làm sạch bước 2 bằng python để hoàn thiện cho ra dữ liệu tốt nhất.

- Xây dựng tool python kiểm thử lại các số liệu giữa dữ liệu trước và sau khi làm "sạch".
- Xây dựng, thiết kế ERD Master Data.
- Cào dữ liệu doanh nghiệp từ các mã số thẻ trong dữ liệu sau khi làm "sạch".
- Tham gia xây dựng Master Data, tạo khóa chính, khóa ngoại và tham chiếu đến nguồn dữ liệu chính.
- Thực hiện kiểm thử dữ liệu trong cơ sở dữ liệu sau khi load lên VPS của công ty.

1.7. Kết luận chương 1

CTCP Giải pháp dệt may bền vững là một công ty hàng đầu trong lĩnh vực dệt may, chuyên cung cấp các giải pháp và dịch vụ tiên tiến cho ngành công nghiệp này. Với nhiều năm kinh nghiệm, công ty đã trở thành đối tác đáng tin cậy của nhiều doanh nghiệp trong ngành dệt may. Thực tập tại đây, em đã có cơ hội tiếp cận với môi trường làm việc chuyên nghiệp, nắm bắt quy trình công việc và tích lũy kinh nghiệm ứng dụng CNTT trong lĩnh vực dệt may.

CHƯƠNG 2. Quá trình thực tập

2.1. Quá trình thực tập theo tuần

Tuần 1 (24/6 – 28/6)	
24/6	1. Nghiên cứu các tài liệu của dự án và kiến trúc data warehouse. Những gì đã hoàn thành và chưa hoàn thành. Tổng quan về nhiệm vụ của các cá nhân trong team.
25/6	2. Nghiên cứu nguồn dữ liệu đầu vào gồm: các transaction xuất khẩu và nhập khẩu ngành dệt may từ 2012 đến 2023, tìm hiểu về các trường dữ liệu hiện có và cách xử lý dữ liệu ở bước tiếp theo.
26/6	3. Tìm hiểu thêm các khái niệm về database, data warehouse, data mart, big data, pipe data, galaxy schema, snowflake schema, star schema và warehouse cần xây dựng.
27/6	4. Phân tích thiết kế ERD.
Tuần 2 (1/7 – 5/7)	

1/7	1. Tìm hiểu quy trình làm sạch dữ liệu lần 2 (chiết xuất vùng từ địa chỉ chủ thẻ)
2/7	2. Chiết xuất vùng từ địa chỉ chủ thẻ của tệp dữ liệu 2020.
3/7	3. Chiết xuất vùng từ địa chỉ chủ thẻ của tệp dữ liệu 2022.
4/7	4. Kiểm thử dữ liệu sau khi làm sạch lần 2. (Chiết xuất vùng có nghĩa là từ địa chỉ các chủ thẻ như địa chỉ của doanh nghiệp mua hoặc bán có thể xác định được vùng kinh tế của các giao dịch ví dụ Đông Nam Bộ, Trung Du Miền Núi Bắc Bộ... Đã có các tool python dựng sẵn, chỉ cần chạy trên môi trường annacoda của google để lấy kết quả, mất nhiều thời gian vì dữ liệu khá lớn.)
5/7	5. Phân tích thiết kế các Common Summary and Data Value.
Tuần 3 (8/7 – 12/7)	
8/7	1. Thiết kế diagram cho các Master Data sẽ có trong Data Warehouse
9/7	2. Làm sạch dữ liệu lần 1 cho các dữ liệu giao dịch xuất nhập khẩu 2014, 2015.
10/7	3. Làm sạch dữ liệu lần 2 cho các dữ liệu giao dịch xuất nhập khẩu 2014, 2015.
11/7	4. Kiểm thử các dữ liệu đã được giao và báo cáo dữ liệu lỗi phát hiện được.
12/7	5. Load dữ liệu giao dịch của năm 2019, 2020 vào database của công ty và kiểm thử lại dữ liệu sau khi load lên.
Tuần 4 (15/7 – 19/7)	
15/7	1. Xử lý các dữ liệu sai đã tìm được bằng tool python đã xây dựng trước đó.
16/7	2. Tiến hành Crawl dữ liệu chi tiết về các doanh nghiệp dựa vào mã số thuế trong dữ liệu để dựng Master Data của các giao dịch import 2021. (Sử dụng tool selenium python đã xây dựng sẵn)
17/7	3. xử lý dữ liệu bị vấn đề trường total_value bị tính toán sai và trường exchange_rate bằng 0.
18/7	
19/7	
Tuần 5 (22/7 – 26/7)	
22/7	1. Tiếp tục crawl dữ liệu về doanh nghiệp import của năm 2013.

23/7	2. Crawl dữ liệu tỷ giá chuyển đổi giữa USD và VND (thuộc trường exchange_rate trong dữ liệu) từ năm 2012 đến 2023 và load vào Master Data của công ty.
24/7	3. Tạo các Materialized View theo diagram Common Summary và Data Value đã thiết kế trước đó cho dữ liệu từ năm 2012 đến 2023, giúp team Power BI truy vấn dữ liệu nhanh chóng cải thiện tốc độ truy vấn cũng như phân tích dữ liệu.
25/7	
26/7	
Tuần 6 (29/7 – 2/8)	
29/7	1. Crawl dữ liệu doanh nghiệp từ nguồn dữ liệu gồm tên và địa chỉ được lấy từ chi tiết các giao dịch export 2022.
30/7	
31/7	2. Kiểm tra dữ liệu trong database những dòng bị miss data trong quá trình chạy tool.
1/8	
2/8	3. Cài tiến tool crawl dữ liệu doanh nghiệp export python.
Tuần 7 (5/8 – 9/8)	
5/8	1. Load và lọc trùng cho các bảng trong Master Data.
6/8	2. Xây dựng tool python kiểm thử dữ liệu sau khi cào có trùng khớp với địa chỉ và tên công ty trong dữ liệu hay không.
7/8	
8/8	3. Xử lý dữ liệu không trùng khớp và tiếp tục crawl doanh nghiệp export 2021.
9/8	
Tuần 8 (12/8 – 16/8)	
12/8	1. Kiểm thử dữ liệu doanh nghiệp đã lấy được và lọc trùng.
13/8	2. Tạo khóa chính, khóa ngoại của các bảng Master Data và tham chiếu đến dữ liệu chính.
14/8	
15/8	3. Hoàn tất bàn giao công việc (tool, tài liệu,..) cho công ty.
16/8	

Bảng 2-1: Quá trình thực tập theo tuần

2.2. Tìm hiểu về Data Warehouse, Data Mart và các mô hình thiết kế

2.2.1 Khái niệm về Data Warehouse, Data Mart

- Database (cơ sở dữ liệu) là một tập hợp thông tin có tổ chức được lưu trữ theo cách hợp lý và tạo điều kiện cho việc tìm kiếm, truy xuất, thao tác và phân tích dữ liệu dễ dàng hơn. Database lưu trữ thông tin thời gian thực về một bộ phận cụ thể trong doanh

nghiệp của bạn: công việc chính của nó là xử lý các giao dịch hàng ngày mà công ty của bạn thực hiện, ví dụ: ghi lại những mặt hàng đã bán. Database xử lý một lượng lớn các truy vấn đơn giản rất nhanh chóng.

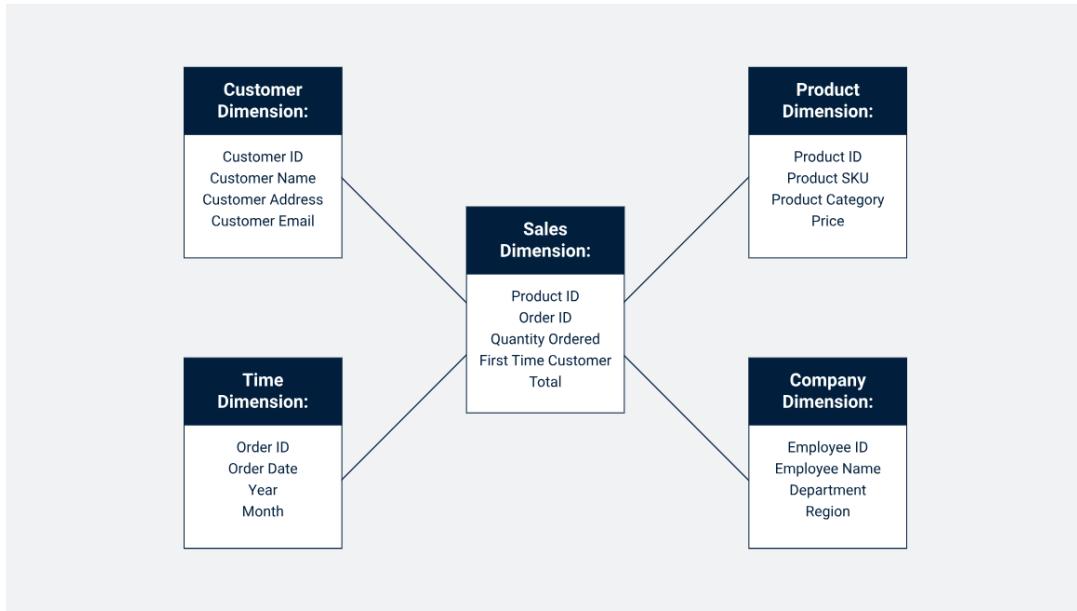


Hình 2.1: Minh họa Data Warehouse

- Data Warehouse (kho dữ liệu) được thiết kế để phân tích, báo cáo, tích hợp dữ liệu giao dịch từ các nguồn khác nhau. Data Warehouse lưu trữ dữ liệu lịch sử về doanh nghiệp của bạn để bạn có thể phân tích và trích xuất thông tin chi tiết từ đó. Nó không lưu trữ thông tin hiện tại, cũng như không được cập nhật theo thời gian thực.
- Sự khác nhau giữa Database và DataWarehouse:
 - Database được thiết kế để thu thập dữ liệu và Data Warehouse được thiết kế để phân tích dữ liệu.
 - Database là một thiết kế hướng đến giao dịch và Data Warehouse là một thiết kế hướng chủ đề.
 - Database thường lưu trữ dữ liệu kinh doanh và Data Warehouse thường lưu trữ dữ liệu lịch sử.
 - Thiết kế Database là để tránh dư thừa càng nhiều càng tốt. Nó thường được thiết kế cho một ứng dụng kinh doanh nhất định. Ví dụ, một bảng User đơn giản có thể ghi dữ liệu đơn giản như tên người dùng và mật khẩu. Nó đáp ứng các ứng dụng kinh doanh nhưng không đáp ứng phân tích. Trong khi đó Data Warehouse lại ngược lại. Các kích thước phân tích và các chỉ tiêu phân tích được thiết kế để đáp ứng yêu cầu phân tích dữ liệu.
- Trong khi kho dữ liệu (Data warehouse) là nơi lưu trữ đa mục đích cho các trường hợp sử dụng khác nhau, thì siêu thị dữ liệu (Mart) là một phần phụ của kho dữ liệu, được thiết kế và xây dựng đặc biệt cho một bộ phận / chức năng kinh doanh cụ thể.

2.2.2 Các mô hình

Star Schema



Hình 2.2: Mô hình Star Schema

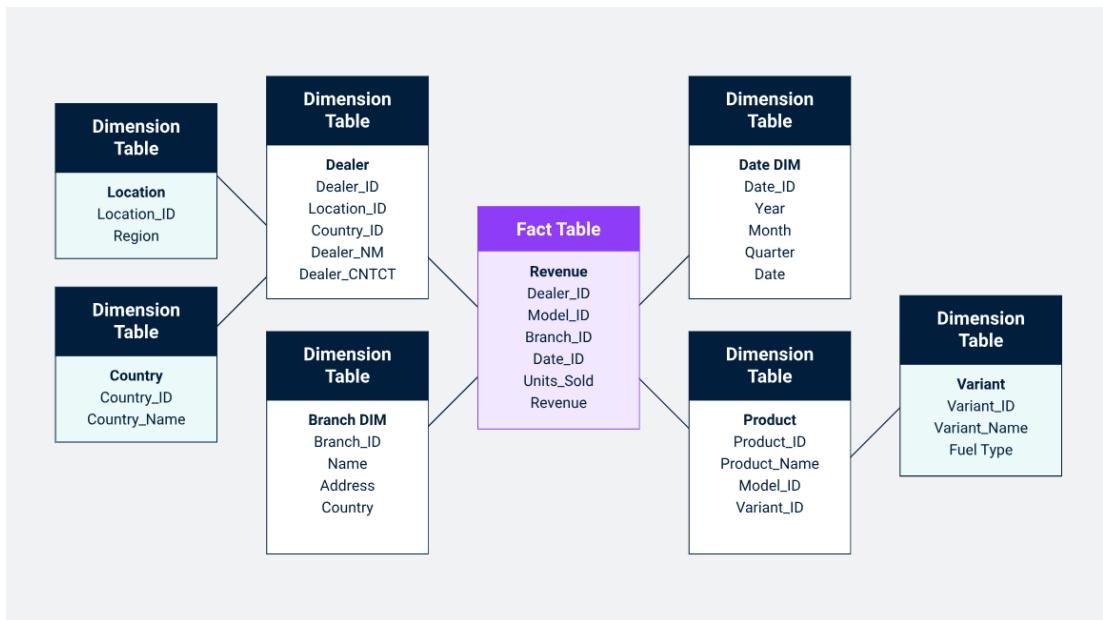
❖ Ưu điểm:

- Thiết kế đơn giản.
- Truy vấn nhanh do không phải join nhiều bảng mỗi lần truy vấn.
- Hỗ trợ khai thác với những truy vấn đòi hỏi sự phức tạp.

❖ Nhược điểm:

- Dư thừa dữ liệu.

Snowflake Schema



Hình 2.3: Mô hình Snowflake Schema

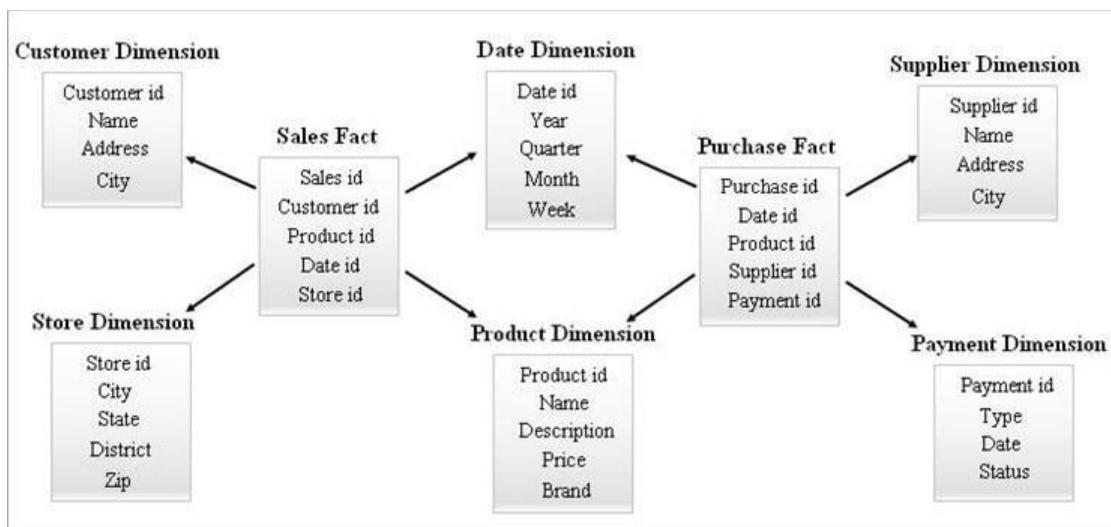
❖ **Ưu điểm:**

- Tránh dư thừa dữ liệu nhưng chưa phải ở mức tối đa.
- Thiết kế đơn giản.

❖ **Nhược điểm:**

- Mỗi lần truy vấn phải join nhiều bảng nên có thể gây chậm.

Galaxy Schema



Hình 2.4: Mô hình Galaxy Schema

❖ **Ưu điểm:**

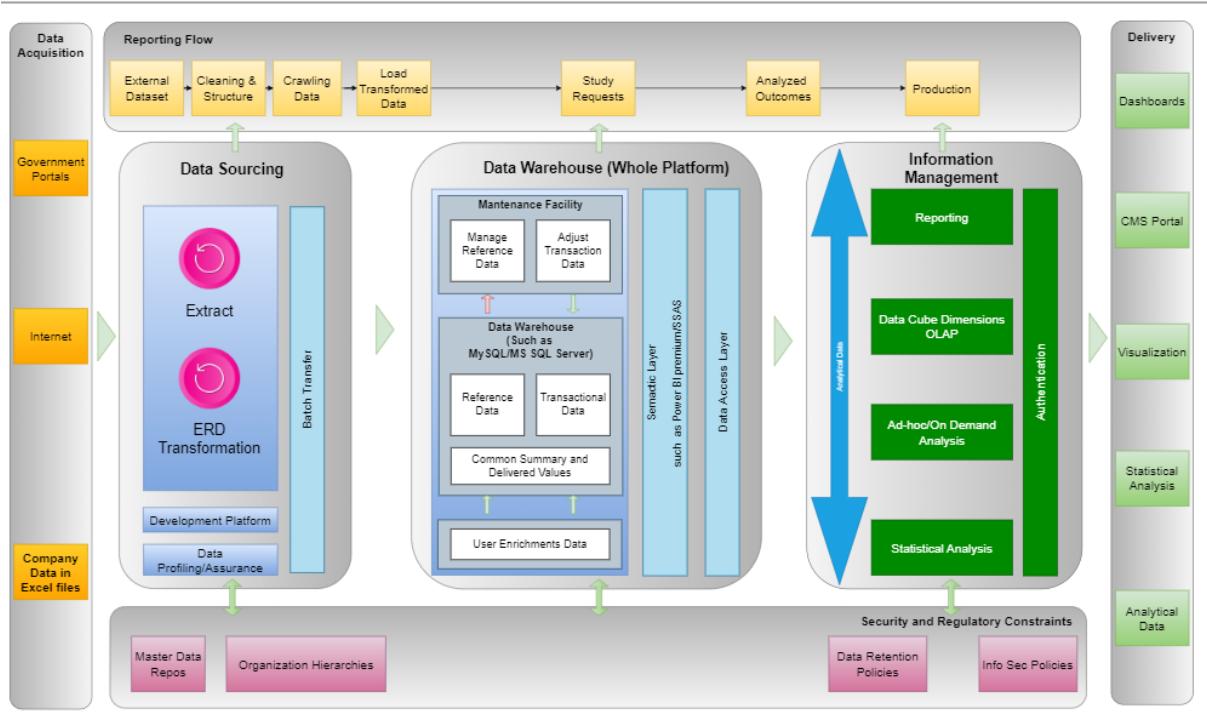
- Tránh dư thừa dữ liệu ở mức tối đa.
- Có thể thực hiện nhiều truy vấn phức tạp.

❖ **Nhược điểm:**

- Mỗi lần truy vấn phải join nhiều bảng nên có thể gây chậm. Vấn đề đặt ra khi các bộ dữ liệu lên đến con số hàng triệu.

2.3. Tìm hiểu Kiến trúc của hệ thống BI

BI System Architecture



Hình 2.5: Kiến trúc của hệ thống BI

Nhiệm vụ chính của Data Intern là tập trung vào xây dựng Data Warehouse.

Giải thích kiến trúc:

- **Data Acquisition (thu thập dữ liệu) từ các nguồn:**

- Government Portals.
- Internet.
- Company Data in Excel files.

- **Reporting Flow:**

- External Dataset (tập dữ liệu thu thập được) đi vào hệ thống Data Sourcing để thực hiện quá trình cấu trúc lại và làm sạch.
- Cleaning & Structure: Dữ liệu đã được làm sạch và cấu trúc lại để làm input cho thu thập dữ liệu.

- Thu thập dữ liệu: tiến hành cào nhũng data còn thiếu ví dụ dữ liệu doanh nghiệp để xây dựng master data repos.
- Load Transformed Data: tải dữ liệu đã được chuyển đổi vào Data WareHouse.
- Study Request: Các yêu cầu data từ bộ phận BI để thực hiện các phân tích, báo cáo sẽ request necessary datas từ Data WareHouse system.
- Analyzed Outcomes: Bộ phận BI phân tích kết quả và kết quả này được xử lý trong Information Management và xuất ra các Production phù hợp với nhu cầu của các bên liên quan.
- Production: sản phẩm đầu ra.

- Data Sourcing:

- Input: dữ liệu chưa được làm sạch.
- Output: Dữ liệu đã được làm sạch và cấu trúc lại.

- Data Warehouse: là một cấu trúc được build trên Postgresql, cấu trúc này bao gồm các Data Reference (Data Master Repos) và các Data Transaction. Từ những dữ liệu này tiến hành xây dựng Common Sumary and Delivered Values (chứa các materialized view lưu trữ các dữ liệu được query dựa trên các tiêu chí phân tích báo cáo của bộ phận BI).

Tại sao phải lưu trữ dữ liệu báo cáo: Vì dữ liệu báo cáo được truy cập với tần số cao nên bắt buộc phải xây dựng những Materialized View để lưu trữ sẵn những dữ liệu báo cáo cần thiết, khi các stakeholder cần thì chỉ cần query vào đúng Materialized View chứa dữ liệu cần thiết để lấy dữ liệu và tổng hợp thông tin. Nếu không có các Materialized View thì mỗi khi cần dữ liệu để lập các report thì hệ thống lại phải join các bảng lại với nhau để trích xuất dữ liệu, trong khi dữ liệu transaction hiện tại lên tới gần 200,000,000 dòng. Việc join bảng và trích xuất như vậy sẽ rất mất nhiều thời gian.

- Information Management: Kết quả sau khi phân tích sẽ phân loại thành các:

- Reporting: Các báo cáo.
- Data Cube Dimension OLAP.
- On Demand Analysis: Thông tin phân tích theo yêu cầu.
- Statistical Analysis: Các phân tích thống kê.

- **Master data repos:** là dữ liệu tĩnh, khó thay đổi trong một tổ chức doanh nghiệp. Muốn thay đổi cần phải tuân theo các Data Retention Policy (Chính sách lưu trữ dữ liệu của doanh nghiệp). Ví dụ: dữ liệu về khách hàng, sản phẩm, nhà cung cấp...

Data Retention Policy: Muốn thay đổi một dữ liệu nào đó trong master data repos phải follow theo các policies của doanh nghiệp.

- **Organization Hierarchies:** Hệ thống phân quyền, không phải ai cũng có quyền truy cập vào data warehouse để lấy dữ liệu dùng, chỉ những ai được ủy quyền mới có thể truy cập và can thiệp vào một số hoặc toàn bộ dữ liệu của doanh nghiệp (đồng thời phải tuân thủ theo Data Retention Policy).

Infor Sec Policy: chính sách thông tin bí mật của doanh nghiệp.

- **Delivery:** Các kênh lấy thông tin để sử dụng.

- Dashboard
- CMS Portal: Hiện tại đang build một cổng Wordpress.
- Visualization
- Statistical Analysis
- Analytical Data

2.4. Tìm hiểu dữ liệu và các thuật ngữ liên quan

2.4.1 Các trường dữ liệu

- *Những trường được tô xám là những trường cần xóa để thay thế thành trường khác hoặc không cần thiết trong thời điểm hiện tại theo sự yêu cầu của User(cụ thể team BI).*

STT	Tên cột	Ý nghĩa
1	Buyer	Tên người mua/ công ty mua hàng
2	Supplier	Tên nhà cung cấp hàng hóa
3	Total Value(USD)	Tổng giá trị được tính theo USD

4	Products	Tên sản phẩm
5	Trade Date	Ngày tiến hành giao dịch
6	Unit Price (Currency)	Giá của mỗi sản phẩm theo tiền tệ (cột Unit Price)
7	Currency	Tiền tệ
8	Supplier address	Địa chỉ của nhà cung cấp
9	Exchange rate	Tỷ giá hối đoái giữa tiền tệ trong giao dịch với tiền Việt Nam
10	B/L Number Or AWB Number	Mã vận đơn
11	Flight/Voyage Number	Số hiệu phương tiện
12	Customs_Warehouse Name In Vietnamese Port	Tên kho hải quan tại cảng Việt Nam
13	Export/Import Declaration Number	Số khai báo xuất khẩu/ nhập khẩu
14	Country of origin	Nước xuất hàng xóa
15	HS Code	HS Code - Harmonized System Codes- Hệ thống hài hòa mô tả và mã hóa hàng hóa - Mã HS Code được sử dụng để phân loại các hàng hóa thành một hệ thống chuẩn IMEX Quốc Tế
16	Export Port	Cảng xuất khẩu
17	Quantity	Số lượng hàng hóa cần xuất/ nhập khẩu
18	Quantity unit	Đơn vị đo số lượng (ví dụ : KGM,Mtr,ROL,set,...)

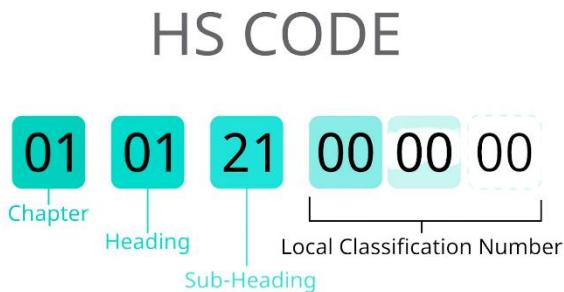
19	Incoterms	Incoterms - international commercial terms- điều khoản thương mại quốc tế (Ví dụ: FOB,CIF,DAF,DAP,..).
20	Importer Code	Mã số thuế Đối với tệp “Export” : đây là mã số thuế của người cung cấp Đối với tệp “Import” : đây là mã số thuế của Buyer
21	Total Value(Currency)	Tổng giá trị được tính theo đồng VND
22	Method Payment	Phương thức thanh toán (Ví dụ : KC,TTR,KHONGTT,...)
23	Buyer address	Đại chỉ người mua hàng
24	Import Tax	Thuế nhập khẩu
25	Landing Port	Cảng nhập khẩu
26	Gross Weight(KG)	Tổng trọng lượng gồm cả bao bì (Kg).
27	Weight Unit	Đơn vị trọng lượng
28	Destination country	Nước đến
29	FOB(USD)	Freight on board (Giá trị hàng hóa khi áp dụng điều khoản này)
30	CIF(USD)	CIF (viết tắt của Cost, Insurance, Freight – tiền hàng, bảo hiểm, cước phí)
31	Transportation	phương thức vận chuyển
32	Buyer (Import)/ Supplier (Export) tel	Số điện thoại Người mua (Nhập khẩu)/ Nhà cung cấp (Xuất khẩu)

33	Customs	Hải quan
34	Port of loading	Cảng chất hàng (hàng giao, nguồn)
35	Type_transaction	Loại giao dịch

Bảng 2-2: Các trường dữ liệu thô

Tổng kết: sử dụng 30 cột dữ liệu để phân tích yêu cầu. Những cột được tô màu sẽ bị drop sau khi làm sạch lần 1.

2.4.2 HS Code



Hình 2.6: Cấu trúc của dãy số HS Code

Hệ thống mã HS được sử dụng ở mọi quốc gia trên thế giới để giúp giao dịch an toàn hơn, nhanh hơn và hiệu quả hơn. Thương nhân, cơ quan Hải quan, bên gửi hàng, bên trung gian chuyển hàng, cảng vụ và nhiều đối tượng khác đều sử dụng hệ thống mã HS thống nhất quốc tế. Điều này đảm bảo mọi người đều hiểu và đồng ý chính xác những gì có trong bất kỳ chuyến hàng nào đi qua biên giới quốc tế.

Hệ thống Hải hòa được tổ chức và duy trì bởi Tổ chức Hải quan Thế giới (WCO). Thành viên WCO là một mạng lưới toàn thế giới các cơ quan Hải quan, có nhiệm vụ đơn giản hóa, hệ thống hóa và thúc đẩy thương mại quốc tế.

Cách hoạt động của hệ thống mã HS - Sáu (6) chữ số đầu tiên của Hệ thống Mã HS

- Hệ thống Mã HS chia tất cả các loại hàng hóa thành: Phần, Chương, Phân chương, Nhóm và Phân nhóm. Với mỗi cấp độ của hệ thống, có các ghi chú giải thích, định nghĩa pháp lý về hàng hóa và mục chi tiết tuân tự của hàng hóa dựa trên cấu trúc thống nhất. Cấu trúc này cùng với các ghi chú và quy tắc đi kèm của nó được gọi là Danh mục Mã HS, hoặc thường chỉ gọi là Danh mục.

- Trong Danh mục, Phần là các nhóm Chương, được tạo ra để nhóm lại với nhau nhiều loại hàng hóa có cùng chủng loại, chức năng, thành phần, ảnh hưởng, mục đích hoặc cách sử dụng.
- Chương bao gồm các Nhóm, tập hợp với nhau các hàng hóa tương tự một cách chặt chẽ hơn và xác định lại các hàng hóa đó bằng hai chữ số. Các chữ số này của Nhóm nằm trong khoảng từ 01 đến 100, với 100 được biểu thị bằng 00. Mục phân loại chi tiết này được lặp lại một lần nữa cho các Phân nhóm (một lần nữa từ 01 đến 00).

Ví dụ minh họa: sản phẩm Chăn điện có mã là **63011000**

- Chương 63** là các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn.
- 6301** là nhóm Chăn và chăn du lịch.
- 63011000** là sản phẩm chăn điện.

Mã HS Việt Nam 63 - Các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn.

Mã HS Việt Nam 63 - Các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn

Tra cứu mã hs số 63 của Việt Nam là Đối với Các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn. tra Mã 2017 HTS hoặc Mã HSN cho Các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn ở Việt Nam.

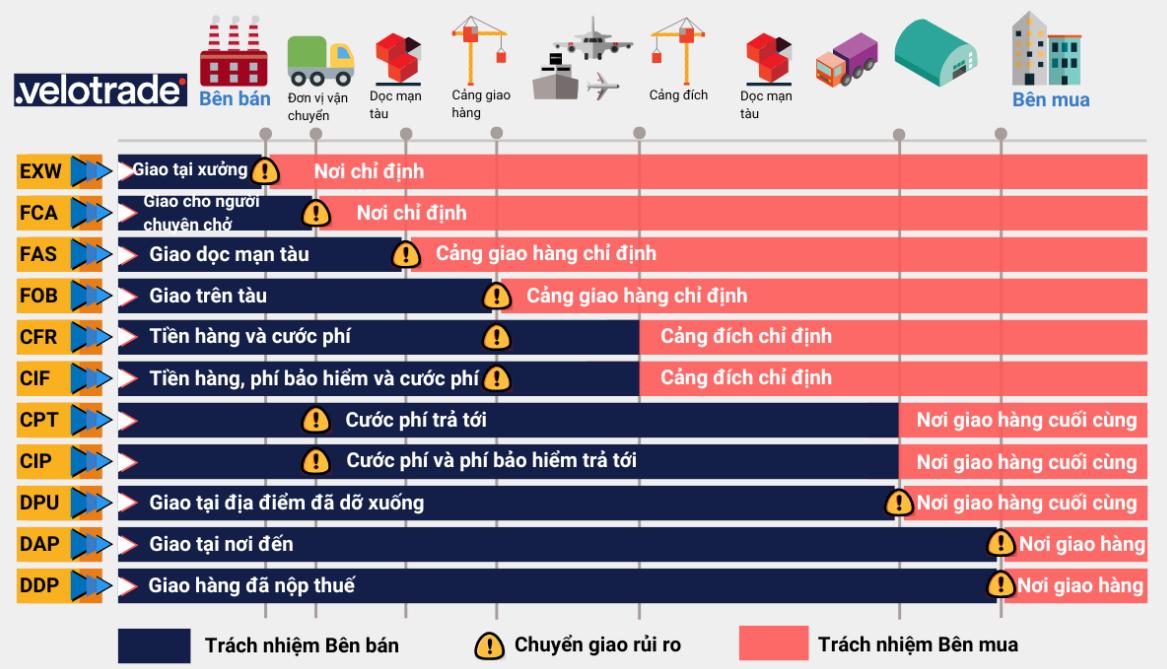
Chương - 63 Các mặt hàng dệt đã hoàn thiện khác; bộ vải; quần áo dệt và các loại hàng dệt đã qua sử dụng khác; vải vụn	
phần mở đầu	mô tả món hàng
6301	Chăn và chăn du lịch.
6302	Khăn trải giường, khăn trải bàn, khăn trong phòng vệ sinh và khăn nhà bếp.
6303	Màn che (kể cả rèm trang trí) và rèm mờ che phía trong; diềm màn che hoặc diềm giường.
6304	Các sản phẩm trang trí nội thất khác, trừ các loại thuộc nhóm 94.04.
6305	Bao và túi, loại dùng để đóng, gói hàng.
6306	Tấm vải chống thấm nước, tấm hiên và tấm che nắng; tảng; buồm cho tàu thuyền, ván lướt hoặc ván lướt cát; các sản phẩm dùng cho cắm trại.
6307	Các mặt hàng đã hoàn thiện khác, kể cả mẫu cắt may.
6308	BỘ VẢI KÈM CHÍ TRANG TRÍ Bó vải bao gồm vải và chi, có hoặc không có phụ kiện dùng để làm chăn, thảm trang trí, khăn trải bàn hoặc khăn ăn đã thêu, hoặc các sản phẩm dệt tương tự, đóng gói sẵn để bán lẻ.
6309	QUẦN ÁO VÀ CÁC SẢN PHẨM DỆT ĐÃ QUA SỬ DỤNG; VẢI VỤNQuần áo và các sản phẩm dệt may đã qua sử dụng khác.
6310	Vải vụn, mẫu dây xe, chéo bện (cordage), thừng và cáp đã qua sử dụng hoặc mới và các phế liệu từ vải vụn, dây xe, chéo bện (cordage), thừng hoặc cáp, từ vật liệu dệt.

Hình 2.7: Bảng tra cứu mã HS Code 63

2.4.3 Incoterm

INCOTERMS 2020

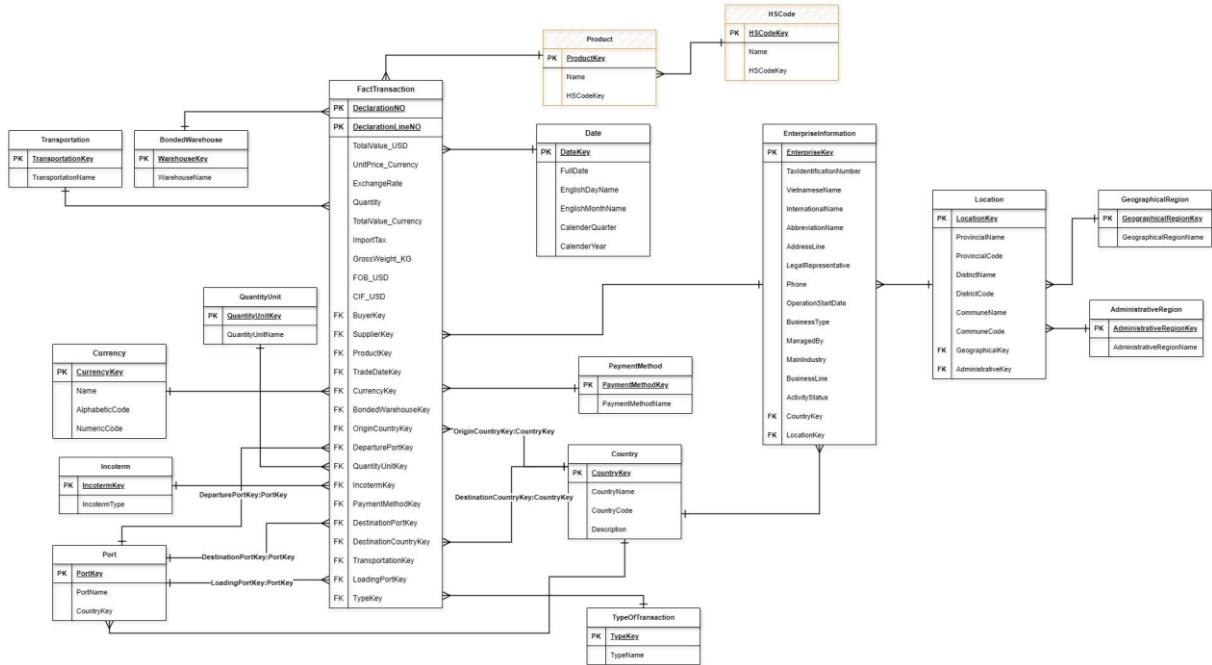
Địa điểm giao hàng và Chuyển giao rủi ro



Hình 2.8: Các quy tắc incoterms của năm 2020

- Incoterms (viết tắt của International Commerce Term) là tập hợp các quy tắc thương mại quốc tế quy định về trách nhiệm của các bên trong hợp đồng ngoại thương.
- Incoterms là các điều khoản thương mại quốc tế được chuẩn hóa và được nhiều quốc gia, vùng lãnh thổ trên thế giới công nhận và sử dụng rộng rãi. Nội dung chính của các điều khoản cần kể đến 02 điểm sau:
 - Trách nhiệm của bên mua, bên bán đến đâu
 - Điểm chuyển giao trách nhiệm, chi phí, rủi ro từ người bán sang người mua
- CIF(USD) Đây là viết tắt của Cost (tiền hàng), Insurance (bảo hiểm), Freight (cước phí). Nội dung của CIF quy định rằng người bán hàng sẽ hoàn thành trách nhiệm của mình khi lô hàng đã được xếp lên boong tàu tại cảng xếp, tuy nhiên lại phải chi trả toàn bộ chi phí vận chuyển trong quá trình vận chuyển hàng đến cảng đích.
- Giá FOB (Free on board - Freight on board nha) chính là giá tại cửa khẩu bến nước của người bán. Giá FOB đã bao gồm toàn bộ chi phí vận chuyển lô hàng ra cảng, thuế xuất khẩu và thuế làm thủ tục xuất khẩu. Lưu ý rằng, giá FOB không bao gồm chi phí bỏ ra để vận chuyển hàng bằng đường biển, cũng không bao gồm chi phí bảo hiểm đường biển.

2.5. Phân tích thiết kế ERD



Hình 2.9: Sơ đồ ERD

Biểu đồ ERD sau khi tham gia phân tích và thiết kế sử dụng công cụ draw.io cùng team Data.

2.5.1 Các bảng và cột

FactTransaction

Mô tả: Là bảng chứa các thông tin liên quan đến giao dịch. Mỗi dòng trong bảng này biểu diễn một giao dịch cụ thể (dựa trên DeclarationNO), mỗi giao dịch sẽ có các chi tiết giao dịch (DeclarationLineNO) và chứa thông tin chi tiết về các khía cạnh của giao dịch đó

Danh sách các cột và kiểu dữ liệu:

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả

1	DeclarationNO	PK	varchar(20)	Khóa chính của bảng, xác định duy nhất mỗi giao dịch. Dùng để nhóm các chi tiết giao dịch thuộc cùng một khai báo.
2	DeclarationLineNO	PK	serial	Số thứ tự của chi tiết trong một khai báo giao dịch. Giúp phân biệt các mục hàng khác nhau trong cùng một khai báo.
3	TotalValue_USD		real	Tổng giá trị của chi tiết giao dịch được tính bằng đô la Mỹ. Đây là giá trị toàn bộ giao dịch sau khi quy đổi sang USD.
4	UnitPrice_Currency		real	Đơn giá của sản phẩm trong đơn vị tiền tệ gốc.
5	ExchangeRate		real	Tỷ giá hối đoái
6	Quantity		real	Số lượng sản phẩm trong 1 chi tiết giao dịch
7	TotalValue_Currency		real	Tổng giá trị của chi tiết giao dịch trong đơn vị tiền tệ gốc.
8	ImportTax		real	Thuế nhập khẩu phải trả
9	GrossWeight_KG		real	Trọng lượng tổng cộng của hàng hóa trong giao dịch

10	FOB_USD		real	Giá trị FOB của chi tiết giao dịch bằng USD
11	CIF_USD		real	Giá trị CIF của chi tiết giao dịch bằng USD
12	BuyerKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Buyer. Dùng để liên kết giao dịch với thông tin chi tiết của người mua.
13	SupplierKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Supplier. Dùng để liên kết giao dịch với thông tin chi tiết của nhà cung cấp.
14	ProductKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Product. Dùng để liên kết giao dịch với thông tin chi tiết của sản phẩm.
15	TradeDateKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Date. Dùng để liên kết giao dịch với thông tin ngày tháng.
16	CurrencyKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Currency.

17	BondedWarehouseKey	FK	serial	Khóa ngoại - tham chiếu đến bảng BondedWarehouse.
18	OriginCountryKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Country
19	DeparturePortKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Port
20	QuantityUnitKey	FK	serial	Khóa ngoại - tham chiếu đến bảng QuantityUnit
21	IncotermKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Incoterm
22	PaymentMethodKey	FK	serial	Khóa ngoại - tham chiếu đến bảng PaymentMethod
23	DestinationPortKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Port
24	DestinationCountryKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Country
25	TransportationKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Transportation
26	LoadingPortKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Port

27	TypeKey	FK	serial	Khóa ngoại - tham chiếu đến bảng TypeOfTransaction
----	---------	----	--------	--

Bảng 2-3: Bảng FactTransaction

Date

Mô tả: Bảng Date là bảng chứa các thông tin về các ngày cụ thể, thường được sử dụng để giúp phân tích dữ liệu theo thời gian. Bảng này có thể lưu trữ thông tin về từng ngày, tuần, tháng, quý, và năm, cũng như các thuộc tính liên quan đến ngày như ngày trong tuần, tuần trong năm,...

Danh sách các cột và kiểu dữ liệu:

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	DateKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi ngày
2	FullDate		date	Ngày đầy đủ dưới định dạng ngày tháng năm
3	EnglishDayName		varchar(10)	Tên ngày trong tuần bằng tiếng Anh (ví dụ: 'Monday', 'Tuesday')
4	EnglishMonthName		varchar(10)	Tên tháng bằng tiếng Anh (ví dụ: 'January', 'February')
5	CalenderQuarter		int	Quý trong năm (từ 1 đến 4).
6	CalenderYear		int	Năm của FullDate

Bảng 2-4: Bảng Date

Currency

Mô tả: Bảng này chứa thông tin về các loại tiền tệ được sử dụng trong hệ thống.

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	CurrencyKey	PK	serial	Mã định danh duy nhất cho mỗi loại tiền tệ trong hệ thống.
2	Name		varchar(100)	Tên của loại tiền tệ. Ví dụ: Đô la Mỹ (USD), Euro (EUR), Yên Nhật (JPY), vv.
3	AlphabeticCode		char(3)	Mã chữ cái 3 ký tự duy nhất đại diện cho loại tiền tệ
4	NumericCode		int	Mã số đại diện cho loại tiền tệ. Đây là một số nguyên duy nhất được gán cho mỗi loại tiền tệ. Thường là các số nguyên nhỏ và không trùng lặp.

Bảng 2-5: Bảng Currency

PaymentMethod

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	PaymentMethodKey	PK	serial	
2	PaymentMethodName		varchar(100)	

Bảng 2-6: Bảng PaymentMethod

Incoterm

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	IncotermKey	PK	serial	
2	IncotermType		varchar(100)	

Bảng 2-7: Bảng Incoterms

QuantityUnit

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	QuantityUnitKey	PK	serial	
2	QuantityUnitName		varchar(100)	

Bảng 2-8: Bảng QuantityUnit

Transportation

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	TransportationKey	PK	serial	
2	TransportationName		varchar(100)	

Bảng 2-9: Bảng Transportation

TypeOfTransaction

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	TypeKey	PK	serial	

2	TypeName		varchar(100)	
---	----------	--	--------------	--

Bảng 2-10: Bảng TypeOfTransaction

Country

Mô tả: Bảng này chứa thông tin về các quốc gia, gồm các thông tin về tên quốc gia

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	CountryKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi quốc gia
2	CountryName		varchar(100)	Tên quốc gia
3	CountryCode		varchar(2)	Mã hiệu quốc gia chứa 2 ký tự (ISO 3166-1 alpha-2)
4	Description		text	Mô tả chi tiết về quốc gia

Bảng 2-11: Bảng Country

Product

Mô tả:

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	ProductKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi sản phẩm
2	Name		text	Tên sản phẩm

3	HSCodeKey	FK	serial	<p>Khóa ngoại - tham chiếu đến bảng HSCode.</p> <p>Mã HS Code phân loại các hàng hóa thành một hệ thống chuẩn IMEX Quốc Tế</p>
---	-----------	----	--------	--

Bảng 2-12: Bảng Product

BondedWarehouse

Mô tả: Bảng BondedWarehouse chứa thông tin về các kho ngoại quan (Bonded Warehouses). Kho ngoại quan là nơi lưu trữ hàng hóa nhập khẩu trước khi chúng được thông quan hoặc xuất khẩu. Bảng này lưu trữ các thông tin cơ bản về các kho ngoại quan như mã kho và tên kho.

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	WarehouseKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi kho ngoại quan
2	WarehouseName		text	Tên kho ngoại quan

Bảng 2-13: Bảng BondedWarehouse

Port

Mô tả: Bảng "Port" là kho lưu trữ các thông tin chi tiết về các cảng vận chuyển quan trọng, cung cấp mã định danh duy nhất và tên của mỗi cảng. Bảng này chủ yếu dùng để quản lý và theo dõi các thông tin về các cảng được sử dụng trong hoạt động xuất nhập khẩu. Điều này giúp tối ưu hóa quá trình vận chuyển và đảm bảo thông tin chính xác và đầy đủ về các điểm giao nhận hàng hóa trong hệ thống.

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả

1	PortKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi cảng. Đây là mã định danh duy nhất cho mỗi cảng trong hệ thống.
2	PortName		text	Tên của cảng. Đây là tên gọi đầy đủ của cảng, giúp nhận diện và phân biệt các cảng khác nhau.
3	CountryKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Country .

Bảng 2-14: Bảng Port

EnterpriseInformation

Mô tả: Bảng này lưu trữ thông tin chi tiết về các doanh nghiệp, bao gồm thông tin về mã số thuế, website, tên trong tiếng Việt và tiếng quốc tế, tên viết tắt, địa chỉ, đại diện pháp luật, số điện thoại, ngày bắt đầu hoạt động, loại hình doanh nghiệp, người quản lý, ngành nghề chính và phụ, và tình trạng hoạt động.

Lưu ý: Đối với các doanh nghiệp nước ngoài, 1 số thông tin sẽ không đầy đủ

Danh sách các cột và kiểu dữ liệu:

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	EnterpriseKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi cảng. Đây là mã định danh duy nhất cho mỗi doanh nghiệp trong hệ thống.
2	TaxIdentificationNumber		varchar(15)	Mã số thuế của doanh nghiệp.

3	VietnameseName		text	Tên của doanh nghiệp trong tiếng Việt.
4	InternationalName		text	Tên của doanh nghiệp trong tiếng quốc tế.
5	AbbreviationName		text	Tên viết tắt của doanh nghiệp.
6	AddressLine		text	Địa chỉ của doanh nghiệp.
7	LegalRepresentative		text	Tên đại diện pháp luật của doanh nghiệp.
8	Phone		varchar(20)	Số điện thoại liên hệ của doanh nghiệp.
9	OperationStartDate		date	Ngày bắt đầu hoạt động kinh doanh của doanh nghiệp.
10	BusinessType		text	Loại hình doanh nghiệp.
11	ManagedBy		text	Đơn vị quản lý doanh nghiệp.
12	MainIndustry		text	Ngành nghề kinh doanh chính của doanh nghiệp.
13	BusinessLine		text	Các ngành nghề kinh doanh khác của doanh nghiệp.

14	ActivityStatus		text	Tình trạng hoạt động của doanh nghiệp (ví dụ: đang hoạt động, ngừng hoạt động).
15	CountryKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Country .
16	LocationKey	FK	serial	Khóa ngoại - tham chiếu đến bảng Location (chỉ những Country nào là VN mới có khóa ngoại này)

Bảng 2-15: bảng *EnterpriseInformation*

Location

Mô tả: Chứa thông tin danh sách tỉnh/thành phố, quận/huyện/thị xã, phường/xã/thị trấn cùng với mã của nó

Danh sách các cột và kiểu dữ liệu:

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	LocationKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi cảng. Đây là mã định danh duy nhất cho mỗi doanh nghiệp trong hệ thống.
2	ProvincialName		text	Tên thành phố trực thuộc trung ương và tỉnh
3	ProvincialCode		char(2)	Mã tỉnh/thành phố

4	DistrictName		text	Tên thành phố thuộc TPTTW, quận, thị xã, huyện, thành phố thuộc tỉnh
5	DistrictCode		char(3)	Mã quận huyện
6	CommuneName		text	Tên phường, xã, thị trấn
7	CommuneCode		char(5)	Mã phường, xã, thị trấn
8	GeographicalKey	FK	serial	Khóa ngoại - tham chiếu đến bảng GeographicalRegion
9	AdministrativeKey	FK	serial	Khóa ngoại - tham chiếu đến bảng AdministrativeRegion

Bảng 2-16: Bảng Location

GeographicalRegion

Mô tả: Bảng "GeographicalRegion" lưu trữ thông tin về ba miền địa lý của Việt Nam, bao gồm Bắc, Trung và Nam.

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	GeographicalRegionKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi miền
2	GeographicalRegionName		text	Tên các miền của đất nước

Bảng 2-17: Bảng GeographicalRegion

AdministrativeRegion

Mô tả: Bảng "GeographicalRegion" lưu trữ thông tin về 6 vùng kinh tế của Việt Nam

STT	Tên cột	Ràng buộc	Kiểu dữ liệu	Mô tả
1	AdministrativeRegionKey	PK	serial	Khóa chính của bảng, xác định duy nhất mỗi vùng kinh tế
2	AdministrativeRegionName		text	Tên các vùng kinh tế của đất nước

Bảng 2-18: Bảng Administrative Region

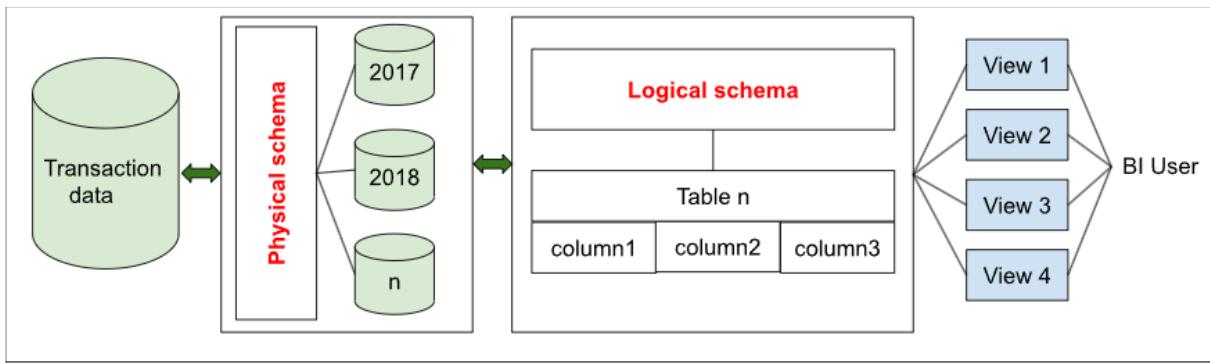
2.6. Phân tích thiết kế Common Summary Data Value

Yêu cầu

- Dữ liệu cần được tổ chức tổng hợp theo dạng bảng, team Data phải chuẩn bị sẵn các View dữ liệu tổng hợp để hỗ trợ team BI trong việc hiển thị dữ liệu.
- Hạn chế số lần query, giải quyết vấn đề về hiệu suất cho mỗi lần query
- Tính toán hết các số liệu có thể tính toán, team BI chỉ cần lấy dữ liệu này hiển thị lên mà không cần phải query lại nhiều lần
- Các bảng hiển thị dữ liệu này cần linh động - có nghĩa là khi dữ liệu năm mới được chèn vào dữ liệu, thì những dữ liệu này sẽ tự động cập nhật giá trị, cập nhật thông tin, thêm và sửa số liệu phù hợp.

Giải pháp dựa trên yêu cầu

Theo sự nghiên cứu và tìm hiểu về hệ quản trị cơ sở dữ liệu PostgreSQL của team Data. Chúng em cho rằng, hệ quản trị này thật sự phù hợp để đáp ứng các yêu cầu mà cấp trên đề xuất. Dưới đây, là sơ đồ minh họa về cách mà PostgreSQL tổ chức dữ liệu một cách linh hoạt và hiệu quả, cho phép quản lý dữ liệu từ nhiều schema vật lý mà không gây ảnh hưởng lẫn nhau, đồng thời cung cấp các công cụ mạnh mẽ cho việc truy xuất dữ liệu thông qua logical schema và views.



Hình 2.10: Mô hình giải pháp common summary data value

Transaction data: Đây là nguồn dữ liệu chính được lưu trữ trong một cơ sở dữ liệu lớn. Nó có thể chứa nhiều schema vật lý khác nhau.

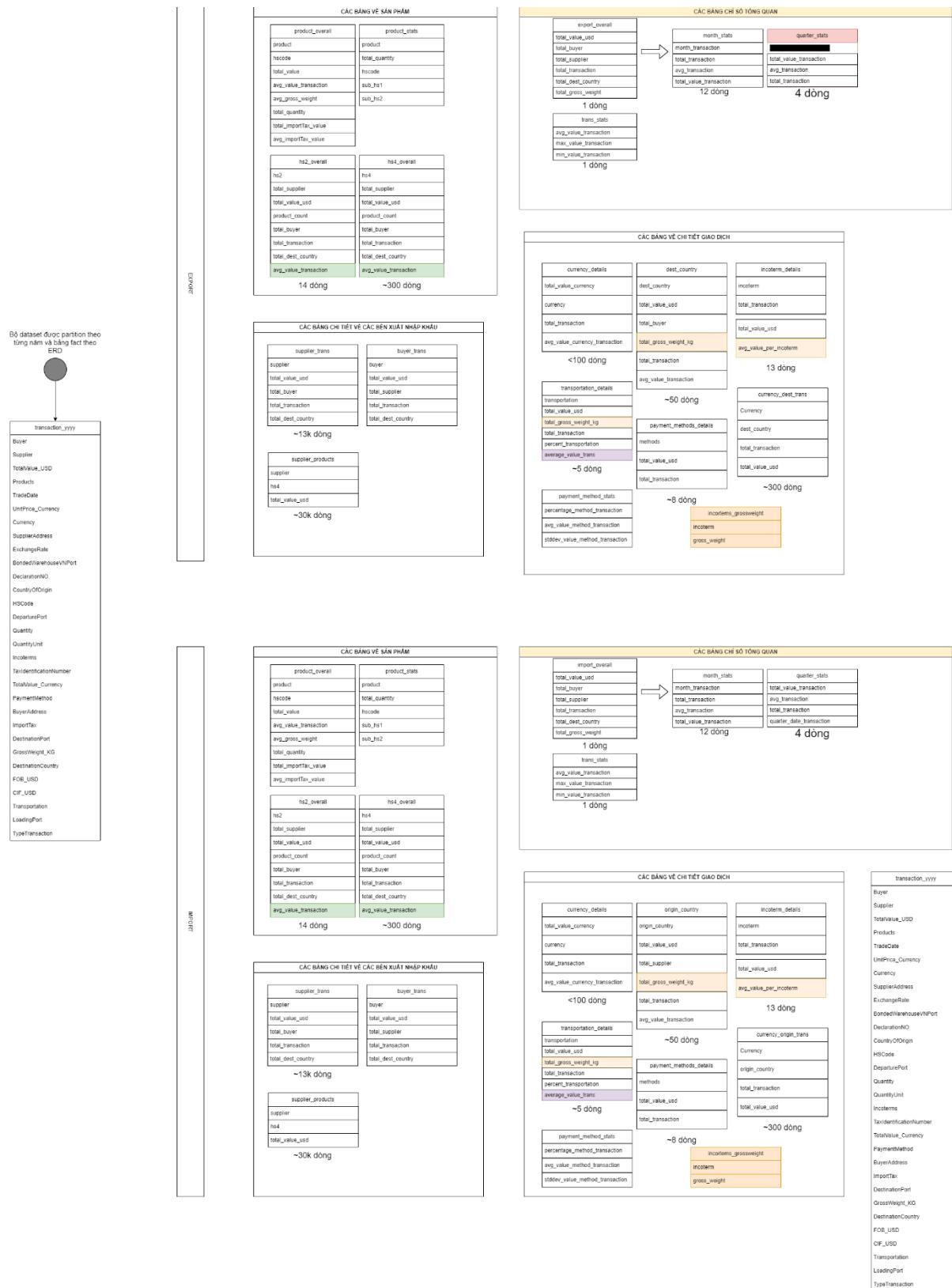
Physical schema: Đây là các schema vật lý tự trị bên trong cơ sở dữ liệu lớn. Mỗi schema này có thể chứa các nhóm đối tượng liên quan với nhau và không ảnh hưởng đến nhau. Ví dụ, schema cho các năm khác nhau như 2017, 2018, n...

Logical schema: Đây là lớp logic được sử dụng để ánh xạ dữ liệu từ các schema vật lý vào các bảng logic, trong trường hợp này các logical schema sẽ là các bảng dữ liệu CSDLV. Logical schema tổ chức dữ liệu từ các schema vật lý vào các bảng chung (Table n) với các cột (column1, column2, column3,...).

Views: Các views được tạo ra từ logical schema để cung cấp dữ liệu cho team BI. Các views này (View 1, View 2, View 3, View 4) cho phép người dùng truy cập và xem dữ liệu từ các bảng logic dưới nhiều góc nhìn khác nhau.

Luồng hoạt động:

- Dữ liệu từ transaction data được phân chia thành các schema vật lý tự trị, ví dụ như schema cho các năm 2017, 2018,..,n.
- Các schema vật lý này tổ chức dữ liệu độc lập và có thể chứa các bảng có cùng tên.
- Logical schema ánh xạ dữ liệu từ các schema vật lý này thành các bảng logic tổng quát.
- Cuối cùng, các views được tạo ra từ các bảng logic để cung cấp dữ liệu cho người dùng dưới nhiều góc nhìn khác nhau.



Hình 2.11: Sơ đồ Common Summary và Data Value

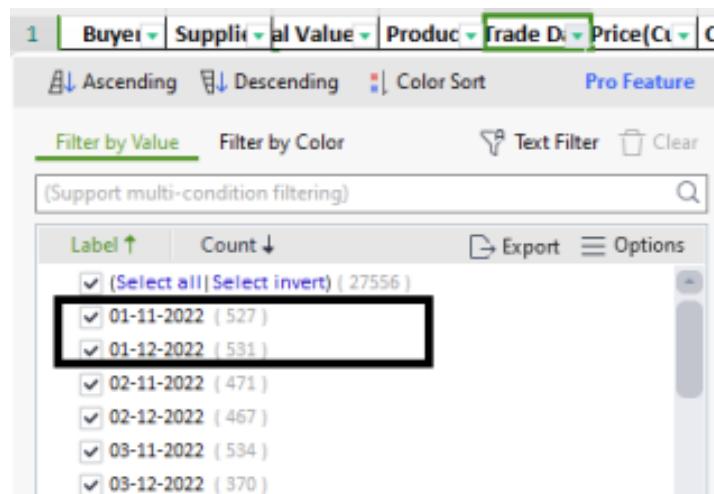
Sơ đồ Common Summary và Data Value cuối cùng sau khi đã được cấp trên xét duyệt.

2.7. Làm sạch dữ liệu lần 1

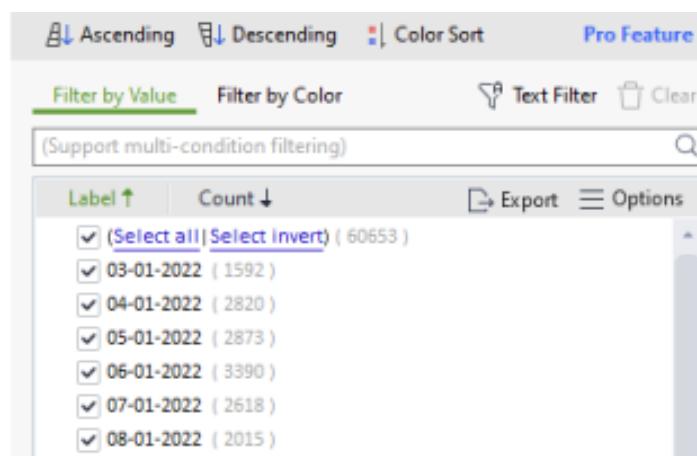
Làm sạch lần 1 theo các tiêu chí đề ra ban đầu của công ty.

Trước khi làm sạch, có hai bộ dữ liệu cần phân biệt : bộ dữ liệu có có một ngày duy nhất, và bộ dữ liệu có trộn lẫn các ngày → Cần tách dữ liệu thành 2 thư mục riêng biệt để tiến hành làm sạch. Cách nhận biết :

Trong bộ dữ liệu có trộn lẫn tháng thì ứng với mỗi tệp sẽ có nhiều tháng xuất hiện trong một tệp như hình sau.



Hình 2.12: Tệp dữ liệu có trộn lẫn các tháng



Hình 2.13: Tệp dữ liệu chỉ xuất hiện một tháng duy nhất

Dựa trên 2 bộ dữ liệu này, ta sẽ tách ra thành các thư mục riêng biệt để sử dụng công cụ làm sạch cho phù hợp.

Bước 1 : Tải mã làm sạch dữ liệu về máy.

- Tệp có tên “**clean_MixedMonths_final.ipynb**” - dành cho bộ dữ liệu có trộn lẫn tháng.

- Tập có tên “***clean_NoMixedMonths_final.ipynb***” - dành cho bộ dữ liệu không trộn lẫn tháng.

Bước 2 : Làm sạch dữ liệu

Bước 2.1 : Làm sạch dữ liệu cho bộ dữ liệu **KHÔNG** trộn lẫn tháng - “***clean_NoMixedMonths_final.ipynb***”

Bước 2.1.1: Tìm kiếm đến vị trí có dòng “ Thủ nghiệm trên folder”

Bước 2.1.2: Sửa lại đường dẫn đến file

- `folder_path` : là nơi đã đến thư mục chứa các file cần làm sạch
- `output_folder` : là nơi sẽ chứa các file sau khi làm sạch
- `processed_files_path` : tệp txt sẽ ghi lại các tệp excel đã làm sạch được, những tệp bị lỗi sẽ không ghi vào.

Hình 2.14: Code truyền đường dẫn làm sạch dữ liệu

Bước 2.1.3 : Nhấn Run All hoặc  để chạy.

Bước 2.2 : Làm sạch dữ liệu cho bộ dữ liệu có trộn lẫn tháng -

“***clean_MixedMonths_final.ipynb***”

Bước 2.2.1: Tìm kiếm đến vị trí có dòng “ Thủ nghiệm trên folder”

Bước 2.2.2: Sửa lại đường dẫn đến file

- `folder_path` : là nơi đã đến thư mục chứa các file cần làm sạch
- `Output_folder` : là nơi sẽ chứa các tệp sau khi làm sạch
- `Processed_files_path` : tệp txt sẽ ghi lại các tệp excel đã làm sạch được, những tệp bị lỗi sẽ không ghi vào.

Hình 2.15: Code truyền đường dẫn làm sạch dữ liệu

Bước 2.1.3 : Nhấn Run All hoặc  để chạy.

Bước 3 : Kết quả làm sạch

	A	B	C	D	E	F	G	H	I	J	K	L
1	Buyer	Supplier	Total Value(USD)	Products	Trade Date	Unit Price(Currency)	Currency	Supplier address	Exchange rate	Customs Warehouse	Export/Import Declar	Country of Origin
2	RESCUE EQUIPMENT Công Ty Cổ Phần Tr		200	Vải bạt che vật r	2014-04-07	200	USD	Undefined	21095	Undefined	Undefined	Viet Nam
3	GOODYEAR DALIAN Công ty TNHH Hyosu		162741.532	TCDVTP-03034101	2014-10-02	3.911	USD	Undefined	21260	Undefined	Undefined	Viet Nam
4	LIANYUNGANG FAR Công ty TNHH Fomic		191441.6	TPVM840GD - Vải n	2014-10-08	5.6	USD	Undefined	21265	Undefined	Undefined	Viet Nam
5	GOODYEAR DALIAN Công ty TNHH Hyosu		163159.088	TCDVTP-03034101	2014-10-02	3.911	USD	Undefined	21260	Undefined	Undefined	Viet Nam
6	LIANYUNGANG FAR Công ty TNHH Fomic		94533.6	TPVM840GD - Vải n	2014-10-02	5.6	USD	Undefined	21260	Undefined	Undefined	Viet Nam
7	HANKOOK TIRE CHI Công ty TNHH Hyosu		179539.32	TCDVTPW-03080772	2014-09-28	5.77	USD	Undefined	21260	Undefined	Undefined	Viet Nam
8	QINGDAO NEXENT Công ty TNHH Hyosu		90711.21	TCDVTPW-03184461	2014-09-25	5.93	USD	Undefined	21217.5	Undefined	Undefined	Viet Nam
9	GOODYEAR DALIAN Công ty TNHH Hyosu		162967.459	TCDVTP-03034101	2014-09-23	3.911	USD	Undefined	21225	Undefined	Undefined	Viet Nam
10	GOODYEAR DALIAN Công ty TNHH Hyosu		162799.286	TCDVTP-03034101	2014-09-21	3.911	USD	Undefined	21225	Undefined	Undefined	Viet Nam
11	TOYO TIRE (ZHANK Công ty TNHH Hyosu		85050	TCDVTPW-03110991	2014-09-17	6.3	USD	Undefined	21200	Undefined	Undefined	Viet Nam
12	TOYO TIRE (ZHANK Công ty TNHH Hyosu		21249.9	TCDVTPW-03110991	2014-09-17	6.3	USD	Undefined	21200	Undefined	Undefined	Viet Nam
13	HANKOOK TIRE CHI Công ty TNHH Hyosu		178494.95	TCDVTPW-03080772	2014-09-09	5.77	USD	Undefined	21200	Undefined	Undefined	Viet Nam
14	GOODYEAR DALIAN Công ty TNHH Hyosu		159271.564	TCDVTP-03034101	2014-09-11	3.911	USD	Undefined	21205	Undefined	Undefined	Viet Nam
15	LIANYUNGANG FAR Công ty TNHH Fomic		95944.8	TPVM840GD - Vải n	2014-09-06	5.6	USD	Undefined	21200	Undefined	Undefined	Viet Nam
16	TOYO TIRE (ZHANK Công ty TNHH Hyosu		29509.05	TCDVTPW-03110961	2014-08-09	6.45	USD	Undefined	21200	Undefined	Undefined	Viet Nam
17	GOODYEAR DALIAN Công ty TNHH Hyosu		162834.485	TCDVTP-03034101	2014-08-30	3.911	USD	Undefined	21200	Undefined	Undefined	Viet Nam
18	GOODYEAR DALIAN Công ty TNHH Hyosu		158571.495	TCDVTP-03034101	2014-08-21	3.911	USD	Undefined	21197.5	Undefined	Undefined	Viet Nam
19	CONG TY VAI HUNG Công Ty TNHH Hạnh		351050	Ruy băng băng vải d	2014-08-09	10.03	USD	Undefined	21200	Undefined	Undefined	Viet Nam
20	CONG TY VAI HUNG Công Ty TNHH Hạnh		180540	Ruy băng băng vải d	2014-08-08	10.03	USD	Undefined	21220	Undefined	Undefined	Viet Nam
21	HANKOOK TIRE CHI Công ty TNHH Hyosu		178316.08	TCDVTPW-03080772	2014-08-29	5.77	USD	Undefined	21195	Undefined	Undefined	Viet Nam
22	CONG TY VAI HUNG Công Ty TNHH Hạnh		180540	Ruy băng băng vải d	2014-08-07	10.03	USD	Undefined	21220	Undefined	Undefined	Viet Nam

Hình 2.16: Kết quả sau khi làm sạch dữ liệu lần 1

2.8. Làm sạch dữ liệu lần 2

Ở bước này thì không cần chia tách bộ dữ liệu. Chỉ cần xác định thư mục các tệp đã làm sạch và thay đổi địa chỉ tương ứng.

- Trong bước này, sẽ xác định lại vị trí các cột và chỉ lấy những cột cần thiết. Những cột không cần thiết sẽ bị loại bỏ ra khỏi dữ liệu.
- Tách địa chỉ và vùng kinh tế

Bước 1 : Tải mã làm sạch dữ liệu về máy và tải tệp excel bộ dữ liệu 63 tỉnh thành được cập nhật - Tệp có tên “**theSecondCleaning.ipynb**”

- Tệp có tên “**QuanHuyenVietNam.xlsx**”

Bước 2 : Làm sạch

Bước 2.1 : Tìm đến vị trí có dòng “Bộ dữ liệu 63 tỉnh thành Việt Nam”

```
path_of = 'C:/Users/Khoa/Desktop/tool/python script/QuanHuyenVietNam.xlsx'
df_country = pd.read_excel(path_of)
print(df_country)
```

Hình 2.17: code truyền đường dẫn bộ dữ liệu 63 tỉnh thành VN

Bước 2.2 : Sửa lại đường dẫn path_of - đây là đường dẫn nơi lưu trữ tệp “**QuanHuyenVietNam.xlsx**” đã được tải về ở Bước 1

Bước 2.3 : Tìm kiếm đến vị trí có dòng “Thử nghiệm trên folder”

Bước 2.4 : Sửa lại đường dẫn đến file

Thử nghiệm trên folder

```
import time
start_time = time.time()
folder_path = 'C:/Users/Khoa/Desktop/tool/file excel' # Đường dẫn đến thư mục chứa file
output_folder = 'C:/Users/Khoa/Desktop/tool/result' # Đường dẫn đến thư mục chứa các file mới excel
# Đọc danh sách đã xử lý từ file nếu có
processed_files_path = 'processed_files_step_2_2022.txt'
processed_file = set() # Use set to avoid duplicates
```

Hình 2.18: Code truyền đường dẫn làm sạch dữ liệu

- folder _path : là nơi đã đến thư mục chứa các file đã làm sạch
- Processed_files_path : tệp txt sẽ ghi lại các tệp excel đã làm sạch được, những tệp bị lỗi sẽ không ghi vào.



Bước 2.5 : Nhấn Run All hoặc để chạy.

Bước 3: Kết quả làm sạch

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	ment	Buyer address	Import Tax	Landing port	Gross Weight(KG)	Destination country	FOB(USD)	CIF(USD)	transportation	Port of loading	Type_transaction	city_province
2	6-2-31 ROPPONGI, #	0 KANSAI INT APT - C	0	Japan	0	0	0 Undefined	HO CHI MINH	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
3	SHINSUNG B/D 3F, #	0 CONG TY TNHH MT	0	Vietnam	0	0	0 Undefined	CONG TY TNHH R.P Export		Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
4	BLK 4010, ANG MO	0 KHO CONG TY ELIT	0	Vietnam	0	0	0 Undefined	KHO CONG TY AVE	Export	Tỉnh Long An	Vùng Đông Bằng Sô	
5	RM 504 HO, 5FL KO	0 INCHEON	0	Korea	0	0	0 Undefined	HA NOI	Export	Thành Phố Hà Nội	Vùng Đông Bằng Sô	
6	RM 504 HO, 5FL KO	0 INCHEON	0	Korea	0	0	0 Undefined	HA NOI	Export	Thành Phố Hà Nội	Vùng Đông Bằng Sô	
7	RM 504 HO, 5FL KO	0 INCHEON	0	Korea	0	0	0 Undefined	HA NOI	Export	Thành Phố Hà Nội	Vùng Đông Bằng Sô	
8	RM 504 HO, 5FL KO	0 INCHEON	0	Korea	0	0	0 Undefined	HA NOI	Export	Thành Phố Hà Nội	Vùng Đông Bằng Sô	
9	RM 504 HO, 5FL KO	0 INCHEON	0	Korea	0	0	0 Undefined	HA NOI	Export	Thành Phố Hà Nội	Vùng Đông Bằng Sô	
10	AM GAMBERG 4, 97	0 OTHER	0	Germany	0	0	0 Undefined	HO CHI MINH	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
11	AM GAMBERG 4, 97	0 OTHER	0	Germany	0	0	0 Undefined	HO CHI MINH	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
12	AM GAMBERG 4, 97	0 OTHER	0	Germany	0	0	0 Undefined	HO CHI MINH	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
13	SO 9 VSIP, DUONG	0 KHO CTY FUJIKURA	0	Vietnam	0	0	0 Undefined	KHO DN TU NHAN T	Export	Tỉnh Đồng Nai	Vùng Đông Nam Bộ	
14	ROOM. 420 BLOCK	0 CONG TY TNHH GIA	0	Vietnam	0	0	0 Undefined	CONGTY TNHH DET	Export	Tỉnh Nam Định	Vùng Đông Bằng Sô	
15	3A WINNER BUILDIN	0 CTY TNHH FREEWE	0	Vietnam	0	0	0 Undefined	CN CTY TNHH XIN C	Export	Tỉnh Bình Dương	Vùng Đông Nam Bộ	
16	16/F,RAILWAY PLAZ	0 KHO CONG TY GIA	0	Vietnam	0	0	0 Undefined	KHO CONG TY YUN	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
17	TRI YEN QUARTER,	0 KHO CTY SHEEN BF	0	Vietnam	0	0	0 Undefined	KHO CTY DASHENG	Export	Tỉnh Đồng Nai	Vùng Đông Nam Bộ	
18	ROOM. 420 BLOCK	0 CONG TY TNHH PHI	0	Vietnam	0	0	0 Undefined	CTY TNHH DET CHÉ	Export	Tỉnh Nam Định	Vùng Đông Bằng Sô	
19	JL. D.I.PANJAITAN I	0 JAKARTA	0	Indonesia	0	0	0 Undefined	HO CHI MINH	Export	Tỉnh Long An	Vùng Đông Bằng Sô	
20	UNIT C & D, 18/F, M	0 CTY TNHH MAY FOI	0	Vietnam	0	0	0 Undefined	CONG TY TNHH R.P	Export	Thành Phố Hồ Chí Minh Vùng Đông Nam Bộ		
21	3/F., SOUTH ASIA B	0 CTY TNHH SAIGON	0	Vietnam	0	0	0 Undefined	CTY TNHH THUONG	Export	Tỉnh Bình Dương	Vùng Đông Nam Bộ	
22	3/F., SOUTH ASIA B	0 KAMPONG CHHNAN	0	Cambodia	0	0	0 Undefined	CUA KHAU MOC BA	Export	Tỉnh Bình Dương	Vùng Đông Nam Bộ	
23	3/F., SOUTH ASIA B	0 KAMPONG CHHNAN	0	Cambodia	0	0	0 Undefined	CUA KHAU MOC BA	Export	Tỉnh Bình Dương	Vùng Đông Nam Bộ	
24	RM # 701, 63, TTUK	0 KHO CONG TY NY H	0	Vietnam	0	0	0 Undefined	KHO CONG TY AVE	Export	Tỉnh Long An	Vùng Đông Bằng Sô	

Hình 2.19: Kết quả sau khi làm sạch dữ liệu lần 2

Lưu ý : Nếu như gặp tệp excel lỗi, tool chạy sẽ dừng lại. Lúc này, cần loại bỏ tệp excel lỗi này sang một thư mục khác để chuyên gia trong lĩnh vực xử lý.

Sau khi đã loại bỏ, chỉ cần nhấn Run và chạy lại, tool sẽ kiểm tra các tệp excel, nếu tệp excel nào nằm trong Processed_files_path thì sẽ bỏ qua.

2.9. Kiểm thử dữ liệu

2.9.1 Mục tiêu

Mục tiêu của quá trình kiểm thử này là để xác minh tính chính xác và hiệu quả của dữ liệu sau khi đã được làm sạch so với số liệu từ phần mềm cũ mà công ty đang sử dụng. Chúng em thực hiện quá trình kiểm thử này để đảm bảo rằng dữ liệu sau khi làm sạch có thể hỗ trợ tốt hơn cho việc ra quyết định và báo cáo của công ty.

2.9.2 Quy trình kiểm thử

1. Làm sạch dữ liệu lần 1

- **Kiểm tra đối chiếu:** So sánh số liệu sau khi làm sạch lần 1 với số liệu tham chiếu từ phần mềm cũ để xác định sự khác biệt và các lỗi có thể tồn tại.

2. Làm sạch dữ liệu lần 2

- **Tối ưu hóa dữ liệu:** Dựa trên kết quả từ lần làm sạch đầu tiên, chúng em tiếp tục tối ưu hóa và làm sạch dữ liệu một lần nữa để đảm bảo tính chính xác và đồng nhất cao nhất có thể.
- **Kiểm tra đối chiếu lần 2:** So sánh số liệu sau khi làm sạch lần 2 với số liệu tham chiếu và số liệu sau khi làm sạch lần 1 để xác định mức độ cải thiện và tính nhất quán của dữ liệu.

Hình 2.20: Bảng kiểm tra đối chiếu số liệu để kiểm thử

2.9.3 Kết quả kiểm thử năm 2022

1. Dữ liệu trước khi làm sạch:

- Tổng số liệu trước khi làm sạch cho phần EXPORT là 48,131,442,464.12 và phần IMPORT là 257,586,582,148.44.

2. Dữ liệu sau khi làm sạch lần 1:

- Tổng số liệu sau khi làm sạch lần 1 cho phần EXPORT là 51,545,535.79 và phần IMPORT là 260,471,845,334.89.
 - Sau khi làm sạch lần 1, chúng em nhận thấy một số cải thiện rõ rệt trong việc loại bỏ các giá trị lỗi và dữ liệu trùng lặp.

3. Dữ liệu sau khi làm sạch lần 2:

- Tổng số liệu sau khi làm sạch lần 2 cho phần EXPORT là 50,882,049,067.49 và phần IMPORT là 260,471,845,334.89.
 - Sau lần làm sạch thứ hai, số liệu cho thấy sự cải thiện về tính nhất quán và đồng nhất của dữ liệu, giúp tăng cường độ tin cậy của dữ liệu trong các báo cáo và phân tích sau này.

4. So sánh với số liệu từ phần mềm cũ:

- o Số liệu từ phần mềm cũ đã được sử dụng làm cơ sở tham chiếu cho quá trình kiểm thử. Việc đổi chiếu cho thấy rằng, sau hai lần làm sạch, dữ liệu hiện tại không chỉ đạt mức độ chính xác cao mà còn được tối ưu hóa tốt hơn so với số liệu cũ.

2.9.4 Kết luận

Quá trình kiểm thử và làm sạch dữ liệu đã chứng minh tính hiệu quả trong việc cải thiện chất lượng dữ liệu, giúp tăng tốc độ truy vấn và hỗ trợ tốt hơn cho quá trình ra quyết định của bộ phận team BI. Số liệu sau khi làm sạch đã trở nên chính xác, nhất quán và đáng tin cậy hơn, giúp công ty cải thiện hiệu quả trong các hoạt động kinh doanh và báo cáo.

2.10. Tải dữ liệu lên hệ thống của công ty

2.10.1 Chuẩn bị

- Nếu sử dụng Google Colab, cần Tài khoản Google Colab kết nối đến folder Drive của mình.
- Nếu sử dụng Jupyter Notebook trên máy tính, cần cài đặt các thư viện liên quan.

2.10.2 Hướng dẫn sử dụng

Cài đặt các thư viện

- Trước khi thực thi chương trình, cần chuẩn bị các thư viện sau: psycopg2, fuzzywuzzy, sqlalchemy
- Nếu sử dụng Google Colab, cần tiến hành cài đặt bằng dòng lệnh, và thực hiện thêm câu lệnh kết nối Drive

```
!pip install fuzzywuzzy  
!pip install psycopg2  
!pip install sqlalchemy
```

```
from google.colab import drive  
drive.mount('/content/drive/')
```

Hình 2.21: Code cài đặt thư viện trước khi tải dữ liệu lên PostgreSQL

Thiết lập Kết nối tới Cơ sở Dữ liệu PostgreSQL

```
db_params = {
    'host': 'your server ip',
    'database': 'database',
    'user': 'user',
    'password': 'password'
}
```

Hình 2.22: Code thiết lập kết nối tới server PostgreSQL

Mục đích của các thông số:

- Host: Xác định địa chỉ của máy chủ cơ sở dữ liệu để client biết nơi để kết nối. Ở đây là **103.110.87.42**
- Database: Xác định cơ sở dữ liệu cụ thể trên máy chủ mà bạn muốn tương tác. Ở đây là database **testdatas**
- User: Xác định người dùng cụ thể đang thực hiện kết nối để xác định quyền và hạn chế truy cập.
- Password: Cung cấp mật khẩu để xác thực danh tính của người dùng.

Sau đó, tiến hành Tạo kết nối tới PostgreSQL, Thiết lập tự động commit và Tạo engine SQLAlchemy

1. Chuẩn bị các schema, table FactTransaction

- Nếu chưa có các schema tương ứng với các năm, tiến hành tạo bằng Query Tool trên pgAdmin4 (nhớ thay giá trị tùy theo năm)

```

# Đặt năm cần sử dụng
year = 2022
# Tên bảng và schema
schema_name = f'{year}'
table_names = {
    "ex": 'FactTransaction_export',
    "im": 'FactTransaction_import1'
}

# Lấy tất cả các khóa dưới dạng danh sách
keys_list = list(table_names.keys())

# Truy xuất trực tiếp giá trị của các khóa
table_name_export = table_names[keys_list[0]]
table_name_import = table_names[keys_list[1]]

# Tạo schema và bảng nếu chúng chưa tồn tại
create_schema_table_query = f"""
BEGIN;

CREATE SCHEMA IF NOT EXISTS "{schema_name}";

CREATE TABLE IF NOT EXISTS "{schema_name}"."{table_name_import}"
(

```

Hình 2.23: Script tạo các schema và bảng khi chưa tồn tại bằng query tool

Hoặc có thể dùng code Python (code nằm trong mục Markdown dưới đây)

Tạo schema & Bảng Facttransaction



```

schema_name = "2017"
table_name = "FactTransaction"
# Tạo schema và bảng nếu chúng chưa tồn tại
create_schema_table_query = f"""
BEGIN;

CREATE SCHEMA IF NOT EXISTS "{schema_name}";

CREATE TABLE IF NOT EXISTS "{schema_name}"."{table_name}"
(
    "ID" bigserial NOT NULL,
    "Buyer" text COLLATE pg_catalog."default",
    "Supplier" text COLLATE pg_catalog."default",
    ...

```

Hình 2.24: Code tạo các schema và bảng khi chưa tồn tại

2. Gom các file .csv lại thành 1 file lớn

Gom các file csv lại thành 1 file lớn:



```

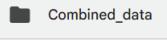
1 # Đường dẫn tới thư mục chứa các file CSV
2 folder_path = f'drive/Shareddrives/Data Team/cleaned_data/{year}/(year)_processed/'
3 # Đường dẫn tới thư mục
4 directory = f'drive/Shareddrives/Data Team/Combined_data/{year}/'
5 os.makedirs(directory, exist_ok=True) # Tạo thư mục nếu chưa tồn tại
6 output_file_im = f'{directory}/combined_data_{year}_im.csv'
7 output_file_ex = f'{directory}/combined_data_{year}_ex.csv'
8
9 start_time = time.time()
10
11 # Kết hợp các file CSV thành hai file CSV lớn
12 with open(output_file_im, 'w') as outfile_im, open(output_file_ex, 'w') as outfile_ex:
13     first_file_im = True
14     first_file_ex = True
15
16     for filename in os.listdir(folder_path):
17         if filename.endswith('.csv'):
18             file_path = os.path.join(folder_path, filename)
19             print(f'Dang xu ly file {file_path}')
20
21             with open(file_path, 'r') as infile:

```

Hình 2.25: Code gom tất các file giao dịch .csv lại thành 1 file duy nhất

Lưu ý: thực hiện thay đổi năm mình muốn import ở biến year

- Đoạn mã này thực hiện việc kết hợp nhiều file CSV từ một thư mục cụ thể thành hai file CSV lớn, dựa trên từ khóa 'im' và 'ex' trong tên file.

- Duyệt qua từng file CSV trong thư mục nguồn, đọc dữ liệu và ghi vào file kết quả tương ứng. Header của file đầu tiên sẽ được giữ lại, còn header của các file tiếp theo sẽ bị bỏ qua. Cuối cùng, tạo hai file kết quả để lưu trữ dữ liệu từ các file CSV có chứa 'im' và 'ex' trong tên file (nằm trong folder )

3. Import dữ liệu .csv vào PostgreSQL

- Nếu lần đầu chạy thì sử dụng code 1
- Nếu đã quen thì sử dụng code 2 cho nhanh

Đầu tiên, duyệt qua tất cả các file .csv nằm trong thư mục đã clean, đọc vào 1 dataframe **df**, sau đó đổi tên các cột cho khớp với cấu trúc đã đưa ra.

Do trong tập dữ liệu có chứa giá trị chuỗi “Undefined”, khi load lên dataframe cần phải được thay thế bằng pd.NA, tức là giá trị null của Pandas.

Sử dụng SQLAlchemy, dữ liệu từ DataFrame được chèn vào bảng FactTransaction trong schema theo từng năm của cơ sở dữ liệu PostgreSQL (với Tùy chọn if_exists='append' chỉ định rằng nếu bảng đã tồn tại, dữ liệu mới sẽ được thêm vào bảng hiện tại và index=False chỉ định rằng các chỉ mục của DataFrame không được chèn vào cơ sở dữ liệu.)

```
# Chèn dữ liệu vào bảng có sẵn
df.to_sql(table_name, engine, schema=schema_name, if_exists='append', index=False)
```

Đoạn code thực thi phần này nằm ở mục Markdown *Dùng thư viện sqlalchemy để import dữ liệu file .csv vào PostgreSQL*

Dùng thư viện sqlalchemy để import dữ liệu file .csv vào PostgreSQL

Nếu bất kỳ giá trị nào trong hàng là "Undefined", nó sẽ được thay thế bằng NULL trong cơ sở dữ liệu.

```
❶ folder_path = 'drive/MyDrive/cleaned_data/2018/'
# Danh sách các file trong folder
files = os.listdir(folder_path)
for csv_file in files:
    if csv_file.endswith('.csv'):
        file_path_csv = os.path.join(folder_path, csv_file)
        print(f'Dang xu ly file {file_path_csv}')
        df = pd.read_csv(file_path_csv)
        # Đổi tên các cột để khớp với bảng PostgreSQL
        df.columns = [
            "Buyer", "Supplier", "TotalValue_USD", "Products", "TradeDate",
            "UnitPrice_Currency", "Currency", "SupplierAddress", "ExchangeRate".
```

Hình 2.26: Code thực thi tải dữ liệu lên PostgreSQL

4. Kết quả

- Sau khi thực thi chương trình, một schema mới đại diện cho từng năm và 1 bảng FactTransaction được thêm thành công vào PostgreSQL.

Buyer	Supplier	TotalValue_USD	Products
JIAXING SHI CHUN HAN TEXTILE CO., LTD	Công Ty TNHH Minh Châu	24000	Phế liệu tơ lõa - FRISON NOT CATED - VN @
JIAXING SHI CHUN HAN TEXTILE CO., LTD	Doanh nghiệp tư nhân tơ lõa thô sợi	214027.847	Tấm tơ trải thảm, mới 100%, xuất xứ Uzbekistan -
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	13924.625	Kén phế tơ tằm vun - VN @
JIANG MEN V APPAREL MANUFACTURING	[null]	254.13	Vải MÀU - VN @
JIANG MEI V	[null]	833.05	vải dệt kim - VN @
JIAXING SHI CHUN HAN TEXTILE CO., LTD	Công Ty TNHH Minh Châu	48976	Phế liệu tơ tằm - FRISON NOT CATED - VN @
SUNG IL INTERNATIONAL CO LTD	[null]	255	Vải MÀU - VN @
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	17015	Kén phế tơ tằm vun - VN @
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	17210.77	Kén phế tơ tằm vun - VN @
SHANGHAI PACIFIC HAT MANUFACTURIN	[null]	249	Vải MÀU - VN @
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	16203.2	Kén phế tơ tằm vun - VN @
JIAXING SHI CHUN HAN TEXTILE CO., LTD	Doanh nghiệp tư nhân tơ lõa thô sợi	258907.878	Tấm tơ trải thảm (tơ tằm thô chưa se), mới 100%
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	17015	Kén phế tơ tằm vun - VN @
ANHUI TINCAI SILK CO., LTD	Công Ty TNHH Một Thành Viên Dịch Vụ Trường Thắng	17015	Kén phế tơ tằm vun - VN @
JIAXING SHI CHUN HAN TEXTILE CO., LTD	Doanh nghiệp tư nhân tơ lõa thô sợi	130956.863	SP2 - Tấm tơ trải thảm, thô, chưa xe, đã qua chín

Hình 2.27: Kết quả sau khi đã tải dữ liệu lên

2.11. Cào dữ liệu doanh nghiệp xuất nhập khẩu để làm MasterData

- Thu thập dữ liệu doanh nghiệp trong các transaction import của năm 2014, 2020 để dựng master data. (Thu thập dữ liệu doanh nghiệp có nghĩa là từ file chứa danh sách các mã số thuế ban đầu, tiến hành chạy một tool python, tool này có nhiệm vụ chạy selenium - Để cào những dữ liệu doanh nghiệp trên trang masothue.com dựa vào danh sách các mã số thuế cung cấp).
- Thu thập dữ liệu doanh nghiệp trong các transaction export của năm 2021, 2022 để dựng master data.

2.11.1 Lấy thông tin doanh nghiệp cần thu thập dữ liệu

Nội dung: Sử dụng tool lấy thông tin doanh nghiệp:

- + Mã số thuế (Importer Code/Tax Identification Number) đối với các giao dịch Nhập khẩu
- + Tên doanh nghiệp (Supplier) và địa chỉ (Supplier address) đối với các giao dịch Xuất khẩu

File dữ liệu cần xử lý: các file data 2014_processed, 2020_processed, 2021_processed, 2022_processed.

File chương trình: Xử lý lấy thông tin doanh nghiệp.ipynb

Hướng dẫn

Thực hiện code ở trong khôi

Thay đổi đường dẫn cho phù hợp với máy mình:

```
year = 2021
def main():
    folder_path = f'.../../Data/{year}/CleanedData/'
    import_folder_path = f'.../../Data/{year}/Import/'
    export_folder_path = f'.../../Data/{year}/Export/'
    os.makedirs(import_folder_path, exist_ok=True)
    os.makedirs(export_folder_path, exist_ok=True)
    processed_import_files = []
    processed_export_files = []
```

Hình 2.28: Code truyền đường dẫn để thực thi lấy thông tin doanh nghiệp

Kết quả sau khi chạy code sẽ xuất hiện 2 thư mục Import và Export trong folder năm

📁 CleanedData	22/06/2024 10:05	File folder
📁 Export	26/06/2024 10:10	File folder
📁 Import	02/07/2024 14:49	File folder

Hình 2.29: Kết quả sau khi chạy code lấy thông tin doanh nghiệp

Trong folder Import sẽ có file *Danh sách ban đầu.xlsx* với nội dung: (file này dùng để **Crawl data cho các doanh nghiệp tham gia vào giao dịch nhập khẩu** để cung cấp dữ liệu đầu vào.

Importer Code
0304655752
0303211289
0304697576
0302279157
0301440244
0305173945
0310477265
4000401369
0301469807
0302483480
3600265469
1001095969
3602482363
3700778993-001
2400515053
0
0315853347
3702816789
1101764652
3901157636
0800304173
5800654554
2600924536

Hình 2.30 Danh sách mã số thuế của doanh nghiệp cần lấy thông tin

Trong folder Export sẽ có file *Danh sách tên công ty và địa chỉ ban đầu.xlsx* với nội dung:

Supplier	Supplier address
công ty cổ	Số 56 Lý Thường Kiệt, Phường 1, Thành Phố Bảo Lộc, Tỉnh Lâm Đồng
công ty cổ	54, Đường 31F, Khu Phố 5, Phường An Phú, Quận 2, Tp.Hcm
công ty trn	71 Nguyen Trong Loi Street, Ward 4, Tan Binh Dist, Hcmc
công ty trn	105Bc Bình Quới, Phường 27, Quận Bình Thạnh, Tp. Hồ Chí Minh
bacninh im	Số 16, Đường Nguyễn Du, Phường Ninh Xá, Thành Phố Bắc Ninh, Tỉnh Bắc Ninh
công ty cổ	Căn Hộ 05, Nhà B13, Khu Đtm Mỹ Đình I, Ngõ 15 Hàm Nghi, Phường Cầu Diễn, Quận Nam Từ Liêm, Tp Hà Nội
doanh ngh	Số 50 Đường 10 Mỹ Tân, Huyện Mỹ Lộc, Tỉnh Nam Định
công ty trn	Số 4 Đường Nguyễn Trãi, Phường Lê Lợi, Thành Phố Hưng Yên, Tỉnh Hưng Yên
công ty trn	Xóm 5 thôn Niêm Ngoại, Xã Kỳ Sơn, Huyện Thuỷ Nguyên, Thành Phố Hải Phòng
hd com b	Số 4 Đường Nguyễn Trãi, Phường Lê Lợi, Thành Phố Hưng Yên, Tỉnh Hưng Yên
công ty trn	56/11-13-15 Đường Ttn17, Kp 4,P. Tân Thới Nhất, Q.12 , Tp.Hcm
cong ty trn	Lô D, Đường Số 1, Kcn Đồng An, Phường Bình Hòa, Tp. Thuận An, Bình Dương
công ty trn	Lô X Đường 11B Kcn Hòa Khánh Mở Rộng Q. Liên Chiểu Tp. Đà Nẵng
undefined	Undefined
thien son p	Số 192 Xóm Đường 10, - Xã Mỹ Tân - Huyện Mỹ Lộc - Nam Định
de nhat im	105Bc Bình Quới, Phường 27, Quận Bình Thạnh, Tp. Hồ Chí Minh
công ty trn	13 Tran Phu, Loc Tien, Bao Loc, Lam Dong
meda ha n	Căn Hộ 05, Nhà B13, Khu Đtm Mỹ Đình I, Ngõ 15 Hàm Nghi, Phường Cầu Diễn, Quận Nam Từ Liêm, Tp Hà Nội
công ty trn	So 21 Ngach 174/27,Lac Long Quan,Tay Ho,Vietnam

Hình 2.31: Danh sách nhà cung cấp và địa chỉ cần lấy thông tin

(file này dùng để **Crawl data cho các doanh nghiệp tham gia vào giao dịch xuất khẩu**) để cung cấp dữ liệu đầu vào.

Trước khi crawl, phải thực hiện các công việc sau.

2.11.2 Sử dụng công cụ làm giàu thông tin các doanh nghiệp

Nội dung: Dựa trên file *Danh sách tên công ty và địa chỉ ban đầu.xlsx* của các giao dịch Export, thực hiện xử lý các thông tin trùng lặp, đã có trong file Doanh nghiệp tổng hợp để làm giảm dữ liệu phải crawl đối với các doanh nghiệp Việt Nam tham gia vào quá trình xuất khẩu

Chuẩn bị

File dữ liệu ban đầu: *Danh sách tên công ty và địa chỉ ban đầu.xlsx*.

File chương trình: *Xử lý MST Export.ipynb*

Hướng dẫn

```
year = '2021'
file_excel_1 = f'.../Data/{year}/Export/Danh sách tên công ty và địa chỉ ban đầu.xlsx'
file_excel_2 = f'.../Data/Dữ liệu doanh nghiệp tổng hợp.xlsx'
file_da_xu_ly1 = f'.../Data/{year}/Export/Dữ liệu doanh nghiệp đã tồn tại của năm {year}.xlsx'
file_da_xu_ly2 = f'.../Data/{year}/Export/Danh sách tên công ty và địa chỉ chưa tồn tại.xlsx'
```

Hình 2.32: Code truyền đường dẫn để thực thi làm giàu thông tin doanh nghiệp

Thực hiện thay đổi năm và đường dẫn dựa trên máy tính của mình. Nhấn Run all để chạy tất cả các khôi code, kết quả cho ra được rất nhiều file

Dữ liệu doanh nghiệp cần crawl của năm...	02/07/2024 15:33	Microsoft Excel W...	449 KB
Danh sách xử lý trùng SupplierAddress kh...	02/07/2024 15:33	Microsoft Excel W...	516 KB
Processed_Suppliers	02/07/2024 15:33	Microsoft Excel W...	121 KB
False_Supplier_Identical	02/07/2024 15:33	Microsoft Excel W...	140 KB
SupplierAddress_ForExportDuplicated	02/07/2024 15:33	Microsoft Excel W...	547 KB
Danh sách xử lý trùng Supplier khác Supp...	02/07/2024 15:33	Microsoft Excel W...	480 KB
False_Addresses_Identical	02/07/2024 15:33	Microsoft Excel W...	98 KB
ImporterCode_ForExportDuplicated	02/07/2024 15:33	Microsoft Excel W...	473 KB
Processed_Addresses	02/07/2024 15:33	Microsoft Excel W...	87 KB
Danh sách tên công ty và địa chỉ chưa tồ...	02/07/2024 15:33	Microsoft Excel W...	469 KB
Dữ liệu doanh nghiệp đã tồn tại của năm...	02/07/2024 15:33	Microsoft Excel W...	2,411 KB
Danh sách tên công ty và địa chỉ ban đầu	02/07/2024 15:26	Microsoft Excel W...	661 KB
ImporterCode_ForExport	06/06/2024 17:54	Microsoft Excel W...	699 KB
~\$Danh sách tên công ty và địa chỉ ban đ...	02/07/2024 16:08	Microsoft Excel W...	1 KB

Hình 2.33: Kết quả sau khi chạy code làm giàu thông tin doanh nghiệp

Tuy nhiên, chỉ cần chú ý đến file *Dữ liệu doanh nghiệp cần crawl của năm 2021.xlsx*, file này là đầu vào của tool **Crawl data cho các doanh nghiệp tham gia vào giao dịch xuất khẩu** ở mục 2.3.2.3.

2.11.3 Sử dụng công cụ để thu thập dữ liệu doanh nghiệp

Nội dung: Sử dụng tool crawl data doanh nghiệp dựa trên thông tin:

- + Mã số thuế (Importer Code/Tax Identification Number) đối với các giao dịch Nhập khẩu
- + Tên doanh nghiệp (Supplier) và địa chỉ (Supplier address) đối với các giao dịch Xuất khẩu

Crawl data cho các doanh nghiệp tham gia vào giao dịch nhập khẩu (tool 1)

Tool này dùng để crawl các doanh nghiệp ở file *Dữ liệu doanh nghiệp cần crawl của năm xxxx.xlsx*

Chuẩn bị

- Jupyter Notebook trên Visual Code
- Google Chrome
- File crx extension giải captcha: *mpbjkejclgfgadiemmegfgebjfooifhl.crx*. Tool này giúp giải captcha tự động.
- Các file đầu vào:

File chương trình: *Crawl data doanh nghiệp nhập khẩu.ipynb*

Dữ liệu doanh nghiệp cần crawl của năm xxxx (ví dụ năm 2019)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Importer Cod																					
2	3603604374																					
3	0306188905																					
4	0313375904																					
5	0107870197																					
6	0200741051																					
7	0302831079																					
8	0302832300																					
9	0311295148																					
10	4900828393																					
11	4900827465																					
12	0108679890																					
13	4601540078																					
14	3503344592																					
15	1200498659																					
16	2300988779																					
17	3901241165																					
18	0900183081																					
19	0302680917-001																					
20	3901224610																					
21	0301452232																					
22	0201558210																					
23	2300945574																					
24	2700676799																					
25	0312312170																					
26	4900836605																					

Hình 2.34: Dữ liệu doanh nghiệp của giao dịch nhập khẩu cần thu thập

File kết quả, có thể tạo trước 1 file excel trống (không cần tạo cũng được vì code sẽ tự tạo nếu file chưa tồn tại)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1																						
2																						
3																						
4																						
5																						
6																						
7																						
8																						
9																						
10																						
11																						
12																						
13																						
14																						
15																						
16																						
17																						
18																						
19																						
20																						
21																						
22																						
23																						
24																						

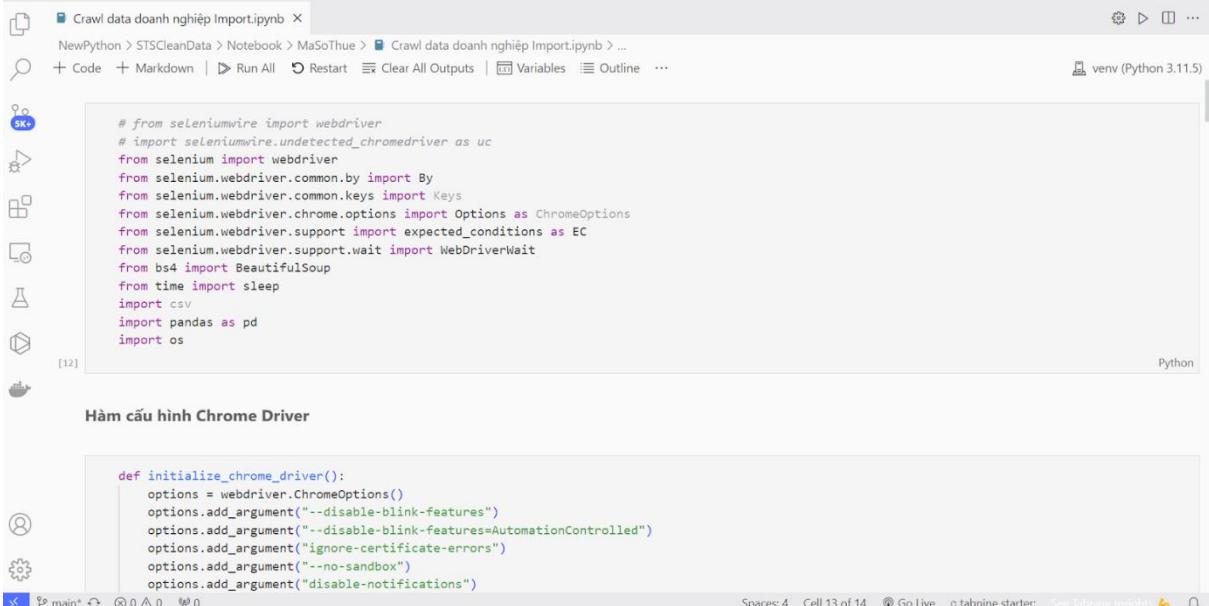
Hình 2.35: File chứa kết quả thu thập thông tin doanh nghiệp

Cài đặt các thư viện trước khi thực thi chương trình: selenium, beautifulsoup4, pandas, webdriver-manager

Hướng dẫn

Bước 1: Chuẩn bị chương trình trên Visual Code, để file

mpbjkejclgfgadiemmfgebjfooflfl.crx cùng đường dẫn với file *Crawl data doanh nghiệp nhập khẩu.ipynb*



The screenshot shows a Jupyter Notebook interface in Visual Studio Code. The title bar says "Crawl data doanh nghiệp Import.ipynb". The code cell contains the following Python code:

```
# from seleniumwire import webdriver
# import seleniumwire.undetected_chromedriver as uc
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.chrome.options import Options as ChromeOptions
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.support.wait import WebDriverWait
from bs4 import BeautifulSoup
from time import sleep
import csv
import pandas as pd
import os
```

Below the code cell, there is a section titled "Hàm cấu hình Chrome Driver" containing:

```
def initialize_chrome_driver():
    options = webdriver.ChromeOptions()
    options.add_argument("--disable-blink-features")
    options.add_argument("--disable-blink-features=AutomationControlled")
    options.add_argument("ignore-certificate-errors")
    options.add_argument("--no-sandbox")
    options.add_argument("disable-notifications")
```

Hình 2.36: Code dùng để thu thập doanh nghiệp nhập khẩu

Bước 2: Di chuyển đến block “**Hàm main**”, chuẩn bị các tham số đầu vào, cần thay đổi cho phù hợp với từng trường hợp:

```
year = "2019"
file_xu_ly = f'../../Data/{year}/Import/Dữ liệu doanh nghiệp cần crawl của năm 2019.xlsx'
file_ket_qua = f'../../Data/{year}/Import/KetQuaCrawl2.xlsx'
start_value = 0
end_value = 122
```

Hình 2.37: Code dùng để thu thập doanh nghiệp nhập khẩu

year: Năm cần thực hiện crawl data doanh nghiệp

file_xu_ly: đường dẫn đến file danh sách ImporterCode là đầu vào để crawl data

file_ket_qua: đường dẫn đến file kết quả sau khi đã crawl data

start_value: khoảng bắt đầu

end_value: khoảng kết thúc

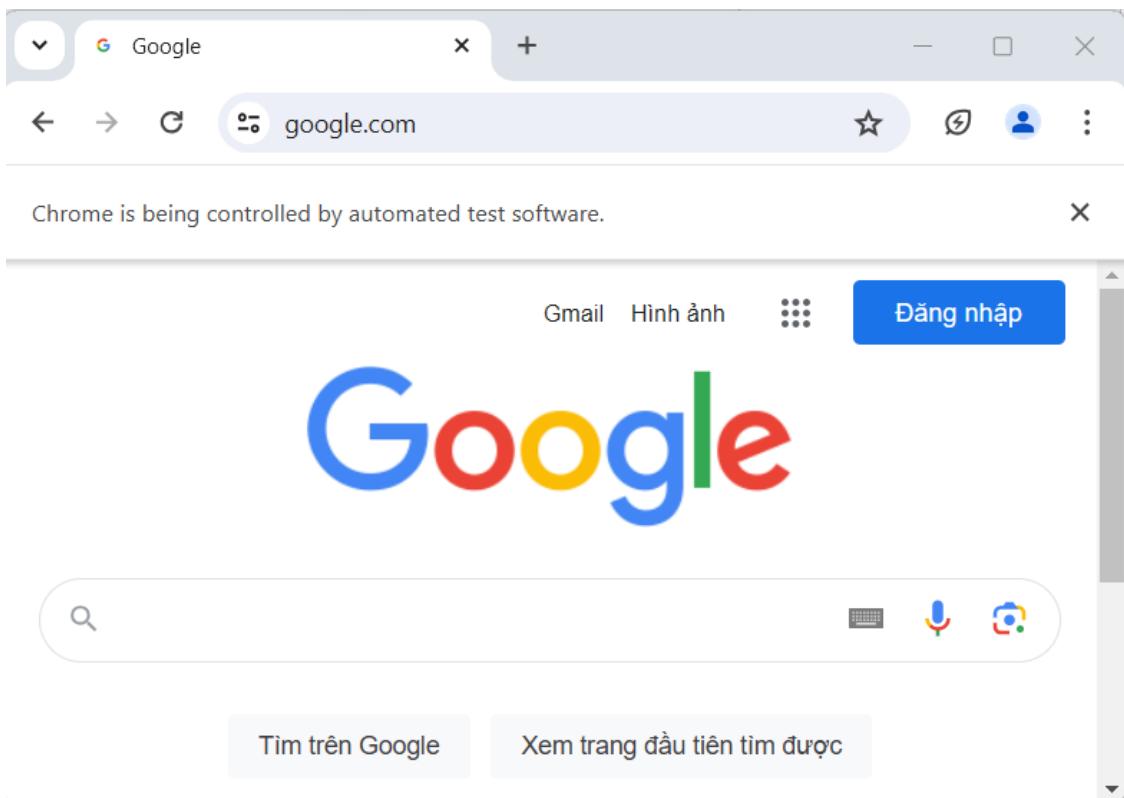
Ví dụ: muốn crawl 14 doanh nghiệp từ cell thứ 2 đến cell thứ 15 trong ảnh dưới đây thì **start_value là 0, end value là 13**

1	Importer Code
2	3603604374
3	0306188905
4	0313375504
5	0107870197
6	0200741051
7	0302831709
8	4900852300
9	0311295148
10	4900828393
11	4900822465
12	0108679690
13	4601540078
14	3502344592
15	1200498659

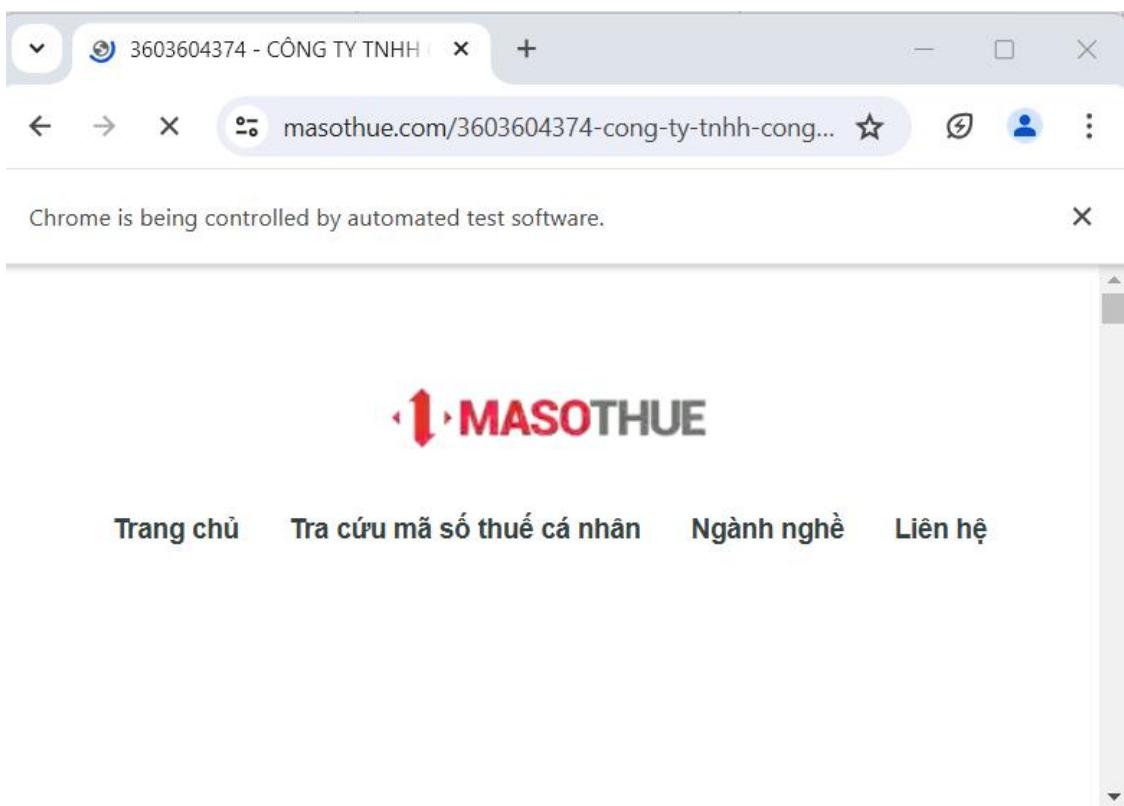
Hình 2.38: Danh sách các mã số thuê cần thu thập dữ liệu

Bước 3: Sau khi đã điền đầy đủ các tham số đầu vào, nhấn nút  Run All

Bước 4: Chương trình sẽ hiển thị 1 cửa sổ Google Chrome thực hiện việc tự truy cập vào trang Google, tìm kiếm Importer Code trên thanh tìm kiếm và truy cập vào trang đầu tiên tìm được, crawl data doanh nghiệp trong trang masothue.com và lưu vào DataFrame **rows_mst**, sau khi chạy hết vòng lặp, xuất kết quả ra DataFrame ra file excel.



Hình 2.39: Tiến trình thực hiện thu thập dữ liệu doanh nghiệp



Hình 2.40: Tiến trình thực hiện tìm kiếm trên trang masothue.com

```

0.....  

3603604374 --masothue.com  

https://masothue.com/3603604374-cong-ty-tnhh-cong-nghe-homewoods-llc  

1.....  

0306188905 --masothue.com  

https://masothue.com/0306188905-cong-ty-tnhh-uyen-dung  

2.....  

0313375504 --masothue.com  

https://masothue.com/0313375504-cong-ty-tnhh-do-an  

3.....  

0107870197 --masothue.com  

https://masothue.com/0107870197-cong-ty-tnhh-thuong-mai-vnexpress  

4.....  

0200741051 --masothue.com  

https://thuvienphapluat.vn/ma-so-thue/cong-ty-trach-nhiem-huu-han-may-yes-vina-mst-0200741051.html  

0200741051 searchsite -masothue.com  

https://www.google.com/search?q=0200741051+searchsite+-masothue.com&btnI=Xem+trang+%C4%91%E1%BA%A7u  

Not found from Google  

5.....  

0302831709 --masothue.com

```

Hình 2.41: kết quả thu thập dữ liệu trên console

Lưu ý: Trong quá trình chạy chương trình, sẽ có những lỗi xảy ra như checkpoint, khi gặp checkpoint thì thực hiện giải captcha, chương trình sẽ thực hiện tiếp tục, không cần phải thoát cửa sổ Chrome đang thực thi sẽ khiến chương trình dừng đột ngột

Nếu chương trình bị lỗi khi đang thực thi, ví dụ như hình dưới:

```

WebDriverException                                     Traceback (most recent call last)
Cell In[8], line 28
    25     combined_data = pd.concat([existing_data, rows_mst], ignore_index=True)
    26     combined_data.to_excel(file_ket_qua, engine='xlsxwriter', index=False)
--> 28 main()

Cell In[8], line 24
    22 # Đọc file Excel hiện tại vào DataFrame, đặt cột "Mã số thuế" thành kiểu dữ liệu object (chuỗi)
    23 existing_data = pd.read_excel(file_ket_qua, dtype={'Mã số thuế': str})
--> 24 crawl_data(start_value,end_value)
    25 combined_data = pd.concat([existing_data, rows_mst], ignore_index=True)
    26 combined_data.to_excel(file_ket_qua, engine='xlsxwriter', index=False)

Cell In[7], line 10
    7 rows_mst.at[row,'MST To Search'] = mst_first
    8 #masothue
    9 # Tìm kiếm masothue
--> 10 if not Google_search(mst_first, '--masothue.com'):
    11     if not Google_search(mst_first, 'searchsite -masothue.com'):
    12         rows_mst.at[row, 'Website masothue'] = 'Not found from Google'

Cell In[5], line 14
    12 driver.find_element('xpath','//*[@id="APjFqb"]').send_keys(url_company)
    13 # sleep(2)
...
    (No symbol) [0x00007FF7C1673592]
    (No symbol) [0x00007FF7C1662F9F]
    Python thread created: [0x00007FF70371E70]::main

```

Hình 2.42: Các lỗi xảy ra trong quá trình thu thập dữ liệu

Lưu ý: Lúc này thực hiện kiểm tra xem dataframe **rows_mst** đã có dữ liệu hay chưa

MST To Search	Tên Việt Nam	Tên quốc tế	Tên viết tắt	Mã số thuế	Địa chỉ	Đại diện pháp luật	Điện thoại	Ngày hoạt động	Quản lý bởi	Loại hình DN	Tình trạng hoạt động	Ngành nghề chính	Ngành nghề kinh doanh
0 3603604374	CÔNG TY TNHH CÔNG NGHỆ HOMWOODS LLC	HOMEWOODS TECHNOLOGY COMPANY LIMITED	HOMEWOODS TECHNOLOGY CO., LTD	3603604374	Tổ 4, ấp 5, Xã An Viễn, Huyện Trảng Bom, Tỉnh ...	LAI NGỌC TRẦM	02513683975	2018-11-30	Chi cục Thuế khu vực hữu hạn ngoài NN	Công ty trách nhiệm hữu hạn thành phố ...	Ngừng hoạt động nhưng chưa hoàn thành thủ tục ...	#1610#Cửa, xe, bảo gỗ và bảo quản gỗ(không hợp pháp, chí hoa...	0220#Khai thác gỗ(từ nguồn gỗ hợp pháp, chí ho...
1 0306188905	CÔNG TY TNHH UYÊN	UYEN DUNG COMPANY	UDT CO., LTD	0306188905	20A Lam Sơn, Phường 2, Quận	BÙI MỸ	028 3848 0000	2008-11-30	Chi cục Thuế	NaN	Ngừng hoạt động nhưng chưa	#1410#May trang phục (trừ trang	1410#May trang phục (trừ trang

Hình 2.43: Kiểm tra dataframe trong quá trình chạy tool thu thập dữ liệu

Nếu đang crawl data giữa chừng mà chương trình dừng đột ngột, để tránh phải crawl lại từ đầu, thực hiện đoạn code ở block ở hình dưới đây:

Chỉ thực hiện đoạn code này khi chương trình gặp lỗi khi đang thực thi

+ Code | + Markdown

```
existing_data = pd.read_excel(file_ket_qua, dtype={'Mã số thuế': str})
# crawl_data(file_excel_export)
combined_data = pd.concat([existing_data, rows_mst], ignore_index=True)
# Loại bỏ các dòng trùng lặp dựa trên cột "NST To Search", chỉ giữ Lại dòng đầu tiên
combined_data = combined_data.drop_duplicates(subset=['NST To Search'], keep='first')
# Lưu Lại DataFrame vào file Excel
combined_data.to_excel(file_ket_qua, engine='xlsxwriter', index=False)
```

Hình 2.44: Đoạn code cần thực hiện khi có lỗi xảy ra trong quá trình thu thập

Lúc này sẽ xuất kết quả hiện có ra file excel, lần crawl tiếp theo lưu ý chỉnh lại **start_value** và **end_value** cho phù hợp, tránh việc crawl lại lần nữa những doanh nghiệp đã có

Kết quả sau khi crawl sẽ là file excel như hình dưới:

Mã số thuế	Website	Tên quốc tế	Tên viết tắt	Địa chỉ	Điện pháp	Điện thoại	Thay hoạt động	Đại hình	Quản lý	bán	nghề	c	nghề kinh	hạng	hoạt	ST To Search			
03061889	https://me	CÔNG TY TUYEN DUN UDT CO.,	I 20A Lam S	BÙI MỸ DL 028 3848	0208-11-14				Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#1410#Ma 1410#May Ngừng học	0306188905		
03133755	https://me	CÔNG TY TNHH ĐỖ AN		51 đường	NGUYỄN N	01663702	2015-07-3	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4641#Báo 1322#Sản	Đang hoạt	0313375504	
01078701	https://me	CÔNG TY VNEXPRO	VNEXPRO	Số 4,	ngách	NGUYỄN T	Bí	theo	2017-06-0	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4669#Báo 4610#Đại	I	Tạm nghỉ	k 0107870197
03028317	https://me	CÔNG TY VIET KHOA VIETKHOA	3 Tân	Thới CAO THỊ	Đ	028371911	2003-01-2	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#1410#Ma 1410#May Ngừng học	0302831709		
49008523	https://me	CÔNG TY THEAVY MIC	CÔNG TY 1Số 140,	tổ	PHẠM THỊ	0986 598	2019-07-2	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4659#Báo 0710#Khai	Tạm nghỉ	k 4900852300	
03112951	https://me	VĂN PHÒNG ĐẠI DIỆN STEINERT	Phòng 506	Dương	Kin	35119800	2011-10-3	Các tổ	chủ	Cục Thuế	Thành phố	Hồ Chí Minh	Ngừng học	Ngừng học	Ngừng học	Ngừng học	Ngừng học	0311295148	
49008283	https://me	CÔNG TY KHANH MINH LANG	S Số 103,	đư	TÔ THỊ	DU	0914 236	2018-05-3	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4649#Báo 4100#Xây	Ngừng học	4900828393	
49008224	https://me	CÔNG TY TTTT TRADI	CÔNG TY 1Số 162B,	T	PHẠM THỊ	088 66555	2018-03-2	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4620#Báo 0710#Khai	Không hoa	4900822465	
01086796	https://me	CÔNG TY DRAGON S	DRAGON S	Số nhà	29,	DƯƠNG	TI 03280019	2019-04-0	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4669#Báo 4610#Đại	I	Đang hoạt	0108679690
46015400	https://me	CÔNG TY TNHH THƯƠNG MẠI	D	Số 696,	tổ	DƯƠNG	TI 0978 413	2019-06-1	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4620#Báo 4620#Bán	Đang hoạt	4601540078	
35023445	https://me	CÔNG TY LAURELIA APPARELS	V	Nhà	xưởng	JOHN	BIJU 02543883:	2017-09-1	Công ty tré	Cục Thuế	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#1410#Ma 1410#May	Không hoa	3502344592	
12004986	https://me	CÔNG TY T NAM OF LONDON	CG	Khu Công	i	Phạm	Minh	02733854:	2010-05-1	Công ty tré	Cục Thuế	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#1410#Ma 1410#May	Đang hoạt	1200498659
23009887	https://me	CÔNG TY CUONG PHUONG IMI	Thửa	đất	NGUYỄN	V	0988 474	12017-07-1	Công ty tré	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	Chi cục	#4669#Báo 0221#Khai	Ngừng học	2300988779	

Hình 2.45: Kết quả thu thập dữ liệu doanh nghiệp của nguồn dữ liệu nhập khẩu

Crawl data cho các doanh nghiệp tham gia vào giao dịch xuất khẩu

Tool này dùng để crawl các doanh nghiệp ở file *Dữ liệu doanh nghiệp cần crawl của năm 2021.xlsx*. Sau khi crawl, sẽ có được thông tin doanh nghiệp cùng với MST chính xác (Các doanh nghiệp Export không có thông tin về MST)

Chuẩn bị

- Jupyter Notebook trên Visual Code
- Google Chrome
- File crx extension giải captcha: *mpbjkejclgfgadiemmegfgebjfooflshl.crx*. Tool này giúp giải captcha tự động.
- Các file đầu vào:

File chương trình: (*Export*) *Crawl data doanh nghiệp.ipynb*

File input: *Dữ liệu doanh nghiệp cần crawl của năm 20xx.xlsx*

Supplier	Supplier address	Importer	Company name	Vietnam Nationality	Address in Vietnam	Address in English	Phone number	Business activity
mounted civinsa.								
huynh de	(Nr Ông Đỗ Trọng Cố) Khu Thượng, Phường Khắc Niệm, Thành Phố Bn, Tỉnh Bn , Việt Nam							
xie cheng	(Thuê:Công Ty Cp Cơ Khí Hùng Cường),Cụm Cn Hợp Lĩnh,P:Hợp Lĩnh,Tp:Bắc Ninh,Tỉnh Bắc Ninh,Việt Na							
chi nhánh 03	Floor, 2A-4A Ton Duc Thang Street, Ben Nghe Ward,District 01, Hochiminh City, Vietnam							
corp corp	03 Nguyễn Oanh Phường 10, Quận Gò Vấp, Tp Hcm							
lim vina p	03 Tân Thới Nhất 17, P. Tân Thới Nhất, Quận 12,, Tp. Hồ Chí Minh							
sung quai	05C, Đường 17A, Khu Phố 6, Thị Trấn Củ Chi, Thành Phố Hồ Chí Minh							
tsuchiya t	05 Độc Lập, Kcn Vsip, P.Bình Hòa, Tp.Thuận An, T. Bình Dương							
hong ngo	05 Quang Trung, P.11, Q.Gò Vấp, Tp.Hcm							
cmi co.ltd	07 Gò Ô Môi, Phường Phú Thuận, Quận 7, Hcm							
vtec corp	07 Lê Minh Xuân Q Tân Bình Tphcm							
me kong	08 Gò Ô Môi, Phường Phú Thuận, Quận 7, Tp.Hcm							
công ty tr	09 Bùi Văn Bình, P.Phú Lợi, Tp Tdm, T.Bình Dương							
mai hoan	09 Nhất Chi Mai, Phường 13, Quận Tân Bình, Tp.Hcm							
nobland v	1-8 Khu A1,4-8 Khu A4, 1-3 & 9-10, Khu Kb1,Khu Cn Tân Thới Hiệp, P Hiệp Thành, Q12, Tphcm							
công ty tr	1/101 Lê Thị Hà, Ấp Đinh, Xã Tân Xuân, H.Hóc Môn, Tphcm							
pao yuan	1/108 Quốc Lộ 13, Phường Hiệp Bình Phước, Quận Thủ Đức, Tp.Hcm							
khang hùi	1/10 Dương Thị Giang, Phường Tân Thới Nhất, Quận 12, Tp. Hcm							
diep binh	1/119 Ấp Đinh , Xã Tân Xuân , Huyện Hóc Môn , Tp Hcm , Việt Nam							
minh thar	1/124C Ấp Đinh, Xã Tân Xuân, Huyện Hóc Môn, Thành Phố Hồ Chí Minh, Việt Nam							
bano k co	1/15-1/17, Đường Trần Bình Trọng, Phường 5, Quận Bình Thạnh, Tp.Hồ Chí Minh							
yu chung	1/1B Tân Thới Nhất 7, P. Tân Thới Nhất. Q12, Tp Hcm							

Hình 2.46: File dữ liệu đầu vào cho quá trình thu thập doanh nghiệp nhập khẩu
Hướng dẫn

Cách sử dụng tương tự như 2 tool trên, thay các tên file và đường dẫn cần thiết để chạy

```
year = 2022

file_xu_ly = f'.../Data/{year}/Export/Dữ liệu doanh nghiệp cần crawl của năm {year}.xlsx'
file_ket_qua = f'.../Data/{year}/Export/KetQuaCrawl.xlsx'
start_value = 10
end_value = 15
```

Hình 2.47: Code truyền đường dẫn để thu thập dữ liệu doanh nghiệp
Kết quả nếu crawl đúng:

```
10.....
cmi co.ltd 07 Gò Ô Môi, Phường Phú Thuận, Quận 7, Hcm masothue.com
https://masothue.com/0303761733-cong-ty-trnhh-corsair-marine-international
11.....
vtec corp 07 Lê Minh Xuân Q Tân Bình Tphcm masothue.com
https://masothue.com/0311263932-tong-cong-ty-co-phan-may-viet-tien-ntnn
```

Hình 2.48: Kết quả thu thập dữ liệu doanh nghiệp xuất khẩu

2.11.4 Kiểm thử dữ liệu đã thu thập

Nội dung: Sau khi đã có được file excel chứa các doanh nghiệp đã crawl từ trang web masothue.com, tiến hành kiểm thử để lọc ra những trường hợp sai sót, sau đó tiến hành crawl thông tin các doanh nghiệp này bằng tay. Các trường hợp sai sót có thể bao gồm:

- + Mã số thuế (Importer Code) đầu vào và mã số thuế crawl được là khác nhau
- + Không tìm thấy mã số thuế đó trên google

Chuẩn bị

- Jupyter Notebook trên Visual Code

- Các file đầu vào:

File chương trình: *Kiểm thử dữ liệu đã cào.ipynb*

Dữ liệu doanh nghiệp đã crawl của năm xxxx (ví dụ năm 2019)

The screenshot shows a Microsoft Excel spreadsheet titled "KetQuaCrawl - Microsoft Excel". The table contains data from column A to column W. The columns represent various fields: A (Mã số thuế), B (Tên doanh nghiệp), C (Địa chỉ), D (Thành phố), E (Tỉnh/TP), F (Số nhà), G (Phường/Huyện), H (Xã/Thị trấn), I (Số điện thoại), J (Email), K (Website), L (Trang web), M (Fanpage), N (Fanpage), O (Fanpage), P (Fanpage), Q (Fanpage), R (Fanpage), S (Fanpage), T (Fanpage), U (Fanpage), V (Fanpage), W (Fanpage). The data consists of approximately 26 rows of crawled business information.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W		
Mã số thuế/mã số doanh nghiệp/nền tảng/tên website																								
2	03061889/ https://ms	CÔNG TY LUẬN ĐƯỜNG CO., LTD	120A Lam Sơn Bùi Mỹ DL 028 3848	2008-11-14	Chi cục: Th	141010M&141010May	Ngưng hoạt	0306188905																
3	03133755/ https://ms	CÔNG TY TNHH ĐỖ AN	51 đường	NGUYỄN A1663702	2015-07-3	Công ty tri	Chi cục: Th	140611B&B1322H&S&	Đang hoạt	0313375504														
4	010787015/ https://ms	CÔNG TY TNHH VIETNEXPRO VINEXPRO	Số 4, ngách	NGUYỄN T&B	2017-06-0	Công ty tri	Chi cục: th	140691B&B14610B&B1	Tạm nghỉ	k0107870197														
5	03028317/ https://ms	CÔNG TY VIỆT KHỔ VIETKHOA	3 Tân Thái	CAO THI	Đ 02837191	2003-01-2	Công ty tri	Chi cục: Th	141010M&141010May	Ngưng hoạt	0302831709													
6	49000852/ https://ms	CÔNG TY HEAVY MACH CÔNG TY	156/140, tổ	PHẠM THI	0986 598 1	2019-07-2	Công ty tri	Chi cục: Th	14059B&B10710B&K&H	Tạm nghỉ	k-4900852300													
7	03112951/ https://ms	VĂN PHÒNG ĐẠI DIỆN STEINERT	Phòng 506	Đường	Đỗ Xanh	2011-10-3	Công ty tri	Chi cục: Th	140659B&B1322H&S&	Đang hoạt	0311295148													
8	490008283/ https://ms	CÔNG TY KHANH MINH LANG	Số 5	103, du	TỐ THI	DU 0914 236	2018-05-3	Công ty tri	Chi cục: Th	14049B&B14100B&X&Y	Ngưng hoạt	49000828393												
9	490008224/ https://ms	CÔNG TY TTT TRADING CO	156/162B,	TP.HCM	THI	08165555	2018-03-2	Công ty tri	Chi cục: Th	14020B&B10710B&K&H	Không hoạt	49000822465												
10	01086796/ https://ms	CÔNG TY DRAGON DRAGONS	Số nhà	29,	ĐƯỜNG	T10328019	2019-04-0	Công ty tri	Chi cục: th	14069B&B14610B&B1	Tạm hoạt	0108679690												
11	46015400/ https://ms	CÔNG TY TNHH THƯƠNG MẠI D&G	696,	đƯỜNG	T10978 413	2019-06-1	Công ty tri	Chi cục: Th	14020B&B14620B&B&	Đang hoạt	4601540078													
12	35023445/ https://ms	CÔNG TY ARTELIA APPARELS	V Nhà	xưởng	JOHN BULL	02543883	2017-09-1	Công ty tri	Chi cục: Th	141010M&141010May	Không hoạt	3502344592												
13	12004986/ https://ms	CÔNG TY TNAM CC	Khu Công	i Phố	Minh	02733854	2010-05-1	Công ty tri	Chi cục: Th	14110B&B141010May	Đang hoạt	1200498659												
14	30009887/ https://ms	CÔNG TY CƯỜNG PHƯƠNG IMI	Thứa	đất	NGUYỄN	V9888 474	2017-07-1	Công ty tri	Chi cục: Th	14069B&B0221#K&H	Ngưng hoạt	3000988779												
15	39012411/ https://ms	CÔNG TY TNHH SUNSTAR	Phòng	506	Đường	Đỗ Xanh	2017-04-2	Công ty tri	Chi cục: Th	14121B&B1313B&H&O	Ngưng hoạt	3901241165												
16	09001830/ https://ms	CÔNG TY VIEBA COM VIEN	CO.	Số	312	đường	TRẦN	TRÒ	0221 3944 2000-12-0	Công ty tri	Chi cục: Th	14020B&B1313B&H&O	Đang hoạt	0900183081										
17	03026897/ https://ms	CHI NHÁNH HAPPY FACE BRANCH	157/1,	Khu	WUJ,	MINC	00503716	2013-12-1	Công ty tri	Chi cục: Th	141010M&1322#S&N	Đang hoạt	0302689717-001											
18	03121522/ https://ms	CÔNG TY TNHH MẶT LANANH C	117-19	TIỀN	HIẾU	008-393255	1993-10-2	Công ty tri	Chi cục: Th	14020B&B141010May	Đang hoạt	0312152232												
19	02015582/ https://ms	CÔNG TY CHANH STAR WOOL	C	Điện	Th	02252992	2014-05-2	Công ty c& Chi	Chi cục: Th	14020B&B1392#S&R	1392#S&M	Đang hoạt	0201558210											
20	23009455/ https://ms	CÔNG TY CHANH STAR WOOL	C	Điện	Th	022526858	2016-08-1	Công ty tri	Chi cục: Th	14020B&B141010May	Ngưng hoạt	2300945574												
21	27006767/ https://ms	CÔNG TY NINH BÌNH NINH BÌNH	Số	nhà	37,	PHẠM	VỊP	01242462	2013-05-1	Công ty tri	Chi cục: Th	141010M&1322#S&N	Đang hoạt	2700676799										
22	03123121/ https://ms	CÔNG TY MEKONG (MEKONG)	Số	6-8	đường	ĐÀNG	VĂN	02462671	2013-05-2	Công ty tri	Chi cục: Th	141010M&141010May	Không hoạt	0312312170										
23	49008366/ https://ms	CÔNG TY TNHH MỲ HOÀNG BÀ	Số	45,	đường	TỐ	NGỌC	0913 277	2018-09-2	Công ty tri	Chi cục: Th	14059B&B1610#C&U	Tạm nghỉ	4900836605										
24	01060476/ https://ms	CÔNG TY XNK LOGIS XNK LOGIS	Tập	thể	CÔ	NGUYỄN	T09132043	2012-11-2	Công ty tri	Chi cục: Th	14020B&B1451#B&N	Đang hoạt	0106047689											
25	39003681/ https://ms	CÔNG TY KORNBEST KORNBEST	Cụm	công	Chen	Weij	0276 3821	2008-04-2	Công ty tri	Chi cục: Th	14130B&S&R	1410#M	Đang hoạt	3900368105										
26	03149148/ https://ms	CÔNG TY GIUM HO V-GUM HO	V/69/5	Ấp	T	NGUYỄN	09079330	2018-03-1	Công ty tri	Chi cục: Th	14069B&B0220#K&H	Không hoạt	0314914893											

Hình 2.49: *Dữ liệu doanh nghiệp của năm 2019 đã thu thập*

- Cài đặt các thư viện trước khi thực thi chương trình: pandas

Hướng dẫn

Bước 1: Di chuyển đến block “Kiểm thử dữ liệu mã số thuế ban đầu khác với mã số thuế đã crawl (File Import)”, chuẩn bị các tham số đầu vào, cần thay đổi cho phù hợp với từng trường hợp:

```

year = 2019
file_path_excel = f'../../Data/{year}/Import/KetQuaCrawl.xlsx'

```

year: Năm cần kiểm thử dữ liệu doanh nghiệp

file_path_excel: đường dẫn đến file kết quả đã crawl của năm đó, chỉnh lại cho phù hợp với máy tính cá nhân

Bước 2: Chạy lần lượt 2 đoạn code sau, kết quả ra được 2 file **Danh sách sai sót năm xxxx.xlsx** và **Danh sách chính xác năm xxxx.xlsx**

```

year = 2019
file_path_excel = f'../../Data/{year}/Import/KetQuaCrawl.xlsx'
df = pd.read_excel(file_path_excel)

4]   ✓  0.9s           Python

df = df.drop_duplicates(subset='MST To Search', keep='first')
# Kiểm tra các dòng có giá trị khác nhau giữa hai cột 'Mã số thuế' và 'MST To Search'
different_rows = df[df['Mã số thuế'] != df['MST To Search']]
output_file_path = f'../../Data/{year}/Import/Danh sách sai sót năm {year}.xlsx'
output_file_path1 = f'../../Data/{year}/Import/Danh sách chính xác năm {year}.xlsx'

different_rows.to_excel(output_file_path, index=False)

trungnhau = df[df['Mã số thuế'] == df['MST To Search']]
trungnhau.to_excel(output_file_path1, index=False)

5]   ✓  1.9s           Python

```

Bước 3: Kiểm tra file kết quả *Danh sách sai sót năm xxxx.xlsx*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Mã số thuế	website maso	Việt Nam	địa chỉ	tên viết tắt	Địa chỉ	diện pháp	Điện thoại	lý do	đóng	Đại diện	Đoàn lý	bình	nghề	kinh
2															hạng hoạt
3															ST To Search
4															0200741051
5	030652782	https://ma	VPDD CÔNG TY TNHH DỊCH VỤ F53 Bùi Tá HÙNG MẠNH HÙN	2009-11-0	Công ty trá	Chi cục Thuế thành phố Thủ Đức	Tam	Đức	Đại	Đại	Đại	Đại	Đại	Đại	4900798195

Hình 2.50: Kết quả kiểm thử dữ liệu doanh nghiệp đã thu thập được

Thực hiện tìm kiếm các trường hợp mã số thuế bị sai sót ở cột MST To Search, trường hợp đầu tiên là Not found from Google, tìm kiếm trên trang masothue.com, ví dụ như 0200741051

CÔNG TY TRÁCH NHIỆM HỮU HẠN MAY YES VINA	
Tên quốc tế	YES VINA GARMENT COMPANY LIMITED
Tên viết tắt	YES VINA CO., LTD
Mã số thuế	0200741051
Địa chỉ	Thôn 3, Xã Kiền Bái, Huyện Thủ Nglookup, Thành phố Hải Phòng, Việt Nam
Người đại diện	OH JUNG SUK
Điện thoại	Bị ẩn theo yêu cầu người dùng
Ngày hoạt động	2007-05-21
Quản lý bởi	Cục Thuế TP Hải Phòng
Loại hình DN	Công ty trách nhiệm hữu hạn ngoài NN
Tình trạng	Ngừng hoạt động và đã đóng MST

Như đã thấy, doanh nghiệp này có thông tin trên trang masothue.com, sai sót xảy ra khi crawl có thể do chương trình chạy quá nhanh hoặc chuỗi tìm kiếm không chính xác, tiến hành cập nhập thông tin doanh nghiệp này bằng tay vào file excel.

Trường hợp thứ 2 là MST tìm kiếm ví dụ là 0306527837, nhưng kết quả crawl lại là doanh nghiệp 0306527837-002 (là chi nhánh của doanh nghiệp trên), trường hợp này cũng tiến hành cập nhập thông tin doanh nghiệp này bằng tay vào file excel.

Cuối cùng, copy dữ liệu của file *Danh sách sai sót năm xxxx.xlsx* vào *Danh sách chính xác năm xxxx.xlsx*, file *Danh sách chính xác năm xxxx.xlsx* là file kết quả chính xác để up lên Drive

Sau khi hoàn tất, nếu vẫn không tra ra được thông tin doanh nghiệp, cần ghi lại thông tin mã số thuế của chúng để có cách giải quyết riêng.

Up kết quả lên Google Drive

Up file *Danh sách chính xác năm xxxx.xlsx* và file danh sách mã số thuế lỗi (nếu có) vào trong thư mục Drive tương ứng theo năm.

Nhận định và đánh giá của bản thân về nhiệm vụ này:

Nhiệm vụ "Cào dữ liệu doanh nghiệp để dựng Master data" là một công việc đòi hỏi sự cẩn trọng và chi tiết trong từng bước thực hiện. Đây là một phần quan trọng trong quy trình xây dựng hệ thống dữ liệu tổng thể của doanh nghiệp, giúp thu thập thông tin cần thiết về các đối tác kinh doanh, bao gồm mã số thuế, tên doanh nghiệp, và địa chỉ.

Trong quá trình thực hiện nhiệm vụ em đúc kết ra được những điều như sau:

- Việc thu thập dữ liệu từ các nguồn nhập khẩu và xuất khẩu giúp xây dựng cơ sở dữ liệu chính xác và đầy đủ về các doanh nghiệp đối tác. Dữ liệu này không chỉ hỗ trợ trong

việc quản lý quan hệ đối tác mà còn đóng vai trò quan trọng trong việc phân tích và ra quyết định kinh doanh. Thông qua việc thu thập thông tin như mã số thuế và địa chỉ doanh nghiệp, em có thể xác định chính xác các bên liên quan và từ đó tối ưu hóa các chiến lược kinh doanh.

- Công việc này yêu cầu sử dụng nhiều công cụ hỗ trợ như Jupyter Notebook, Selenium, và các thư viện như BeautifulSoup và Pandas để tự động hóa quá trình cào dữ liệu. Sự kết hợp giữa công nghệ và quy trình tự động giúp giảm bớt khối lượng công việc thủ công, tiết kiệm thời gian và giảm thiểu sai sót. Tuy nhiên, việc cấu hình và sử dụng các công cụ này đòi hỏi kiến thức chuyên môn và sự hiểu biết sâu rộng về công nghệ thông tin.

- Một trong những thách thức lớn trong nhiệm vụ này là xử lý các sai sót có thể phát sinh, chẳng hạn như thông tin mã số thuế không khớp hoặc không tìm thấy trên trang web. Việc kiểm thử dữ liệu và xử lý thủ công những trường hợp ngoại lệ là bước quan trọng để đảm bảo tính chính xác và tin cậy của dữ liệu. Đây là một quá trình cần sự tỉ mỉ và kiên nhẫn, đặc biệt khi làm việc với dữ liệu lớn và đa dạng.

Kết quả cuối cùng là bộ dữ liệu doanh nghiệp chính xác, được lưu trữ và cập nhật lên hệ thống. Dữ liệu này sẽ được sử dụng trong các phân tích sâu hơn, hỗ trợ cho các chiến lược kinh doanh, và giúp tối ưu hóa mối quan hệ với các đối tác. Điều này không chỉ nâng cao hiệu quả hoạt động mà còn giúp doanh nghiệp hiểu rõ hơn về thị trường và các bên liên quan.

=> Nhiệm vụ này đòi hỏi một quy trình làm việc khoa học, sử dụng các công cụ hiện đại, và khả năng xử lý dữ liệu chi tiết. Qua đó, em đã học được rất nhiều về quản lý dữ liệu, từ khâu thu thập, xử lý, cho đến kiểm thử và phân tích. Những kỹ năng và kiến thức này sẽ rất hữu ích trong các dự án tiếp theo và trong sự phát triển sự nghiệp của em trong lĩnh vực công nghệ thông tin.

CHƯƠNG 3. KẾT QUẢ THỰC TẬP

3.1. Kết quả thực tập

Trong 8 tuần thực tập tại Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững, em đã hoàn thành một số nhiệm vụ quan trọng và đạt được nhiều kết quả tích cực. Những nhiệm vụ này bao gồm:

- **Nghiên cứu và xây dựng hệ thống kho dữ liệu (Data Warehouse):** Em đã tham gia vào việc thiết kế và triển khai hệ thống kho dữ liệu, tập trung vào việc làm sạch và tổ chức dữ liệu từ nhiều nguồn khác nhau. Điều này giúp tối ưu hóa quy trình xử lý và phân tích dữ liệu, từ đó hỗ trợ ban lãnh đạo đưa ra các quyết định chiến lược chính xác và kịp thời.
- **Triển khai materialized view:** Em đã thực hiện các script để tạo materialized view, cải thiện đáng kể hiệu suất hệ thống. Các view này không chỉ đáp ứng nhu cầu hiện tại của bộ phận BI mà còn mở ra khả năng mở rộng và nâng cấp hệ thống trong tương lai.
- **Kiểm thử và làm sạch dữ liệu:** Em đã tham gia vào quá trình làm sạch dữ liệu thông ban đầu, đảm bảo chất lượng và tính chính xác của dữ liệu. Công việc này bao gồm việc sử dụng các công cụ và công nghệ hiện đại để tối ưu hóa quá trình quản lý và phân tích dữ liệu.

Qua đó em đã học hỏi được thêm nhiều kỹ năng chuyên môn như:

- **Quản lý dữ liệu và làm sạch dữ liệu:** Qua quá trình thực tập, em đã nắm vững các phương pháp làm sạch dữ liệu, thiết kế và triển khai các bảng dữ liệu, cũng như sử dụng các công cụ và công nghệ liên quan như Python và SQL.
- **Thiết kế hệ thống:** Em đã học được cách thiết kế hệ thống kho dữ liệu và các thành phần liên quan, bao gồm cả việc thiết lập các khóa chính, khóa ngoại, và các materialized view trong cơ sở dữ liệu.
- **Phân tích và xử lý dữ liệu:** Kinh nghiệm thực tế đã giúp em cải thiện kỹ năng phân tích và xử lý dữ liệu, hiểu rõ hơn về tầm quan trọng của dữ liệu trong việc hỗ trợ ra quyết định chiến lược và quản lý doanh nghiệp.

Dù đã đạt được nhiều kết quả đáng khích lệ, em nhận thấy còn một số kỹ năng cần cải thiện và học hỏi thêm:

- **Nâng cao kỹ năng về phân tích dữ liệu nâng cao:** Việc tìm hiểu sâu hơn về các kỹ thuật phân tích dữ liệu tiên tiến và các công cụ BI (Business Intelligence) sẽ giúp em cải thiện hiệu suất công việc.
- **Học thêm về quản lý dự án:** Nắm vững các phương pháp và kỹ năng quản lý dự án sẽ giúp em đóng góp hiệu quả hơn vào các dự án tương lai.
- **Cập nhật công nghệ mới:** Tiếp tục cập nhật và học hỏi các công nghệ mới trong lĩnh vực quản lý dữ liệu và phân tích dữ liệu là điều cần thiết để duy trì và nâng cao năng lực chuyên môn.

Dù với vị trí là thực tập sinh, nhưng em đã rất may mắn khi là một trong những người được hướng dẫn bởi anh Ngô Trí Thanh. Anh Thanh đã rất tận tình hướng dẫn và tạo cho em cơ hội được trải nghiệm trong một môi trường làm việc đầy năng động. Nhờ sự chỉ dẫn của anh, em đã nhận ra và cải thiện nhiều điểm yếu của bản thân, từ đó không ngừng hoàn thiện các kỹ năng mềm cần thiết trong công việc.

Thêm vào đó, là em cũng xin gửi lời cảm ơn chân thành đến bạn Vy, bạn Toàn và bạn Hùng chung team. Các bạn đã hỗ trợ, training và phổ cập kiến thức cho em, giúp em học hỏi và tích lũy được nhiều kiến thức để phục vụ cho quá trình hoàn thành các nhiệm vụ được giao.

Bên cạnh đó, em cũng may mắn nhận được sự dấn dát nhiệt tình của giảng viên hướng dẫn thực tập – thầy Phạm Thế Bảo. Thầy đã đồng hành và hướng dẫn cho chúng em trong suốt quá trình thực tập, giúp chúng em có thể hoàn thành kỳ thực tập một cách thành công.

Em xin cảm ơn thầy Bảo, anh Thanh, bạn Toàn, bạn Vy và bạn Hùng đã đồng hành và hỗ trợ em trong suốt quá trình thực tập.

3.2. Các biểu mẫu đánh giá

TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Tp. Hồ Chí Minh, ngày 17 tháng 08 năm 2024

BẢNG GHI NHẬN KẾT QUẢ THỰC TẬP HÀNG TUẦN

Một số thông tin liên hệ

Họ và tên: Võ Quang Đăng khoa

Ngày sinh: 21/03/2002

Mã số sinh viên: 3120560047

Lớp: DKP1201

Ngành học: Kỹ thuật phần mềm

Email: dangkhoa014@gmail.com

Điện thoại: 0702788634

Chuyên gia doanh nghiệp: Ngô Trí Thanh

Email: thanh.ngo@stsgroup.org.vn

Điện thoại: 0334004946

Giảng viên hướng dẫn: PGS.TS Phạm Thế Bảo

Email:

Điện thoại

Tuần	Nội dung thực tập (do chuyên gia của doanh nghiệp giao)	Kết quả thực tập (do chuyên gia của doanh nghiệp đánh giá)
1 Từ ngày 24/06/2024 đến ngày 29/06/2024	<p>1. Nghiên cứu các tài liệu của dự án và kiến trúc data warehouse. Những gì đã hoàn thành và chưa hoàn thành. Tổng quan về nhiệm vụ của các cá nhân trong team.</p> <p>2. Nghiên cứu nguồn dữ liệu đầu gồm: các transaction xuất khẩu và nhập khẩu ngành dệt may từ 2012 đến 2023, tìm hiểu về các trường dữ liệu hiện có và cách xử lý dữ liệu ở bước tiếp theo.</p> <p>3. Tìm hiểu thêm các khái niệm về database, data warehouse, data mart, big data, pipe data, galaxy scheme, snowflake schema, star schema và warehouse cần xây dựng.</p> <p>4. Phân tích thiết kế ERD.</p>	
2 Từ ngày 01/07/2024 đến ngày 06/07/2024	<p>1. Tìm hiểu quy trình làm sạch dữ liệu lần 2 (chiết xuất vùng từ địa chỉ chủ thẻ)</p> <p>2. Chiết xuất vùng từ địa chỉ chủ thẻ của tệp dữ liệu 2020.</p> <p>3. Chiết xuất vùng từ địa chỉ chủ thẻ của tệp dữ liệu 2022.</p> <p>4. Kiểm thử dữ liệu sau khi làm sạch lần 2. (Chiết xuất vùng có nghĩa là từ địa chỉ các chủ thẻ như địa chỉ của doanh nghiệp mua hoặc bán có thể xác định được vùng kinh tế của các giao dịch ví dụ Đông Nam Bộ, Trung Du Miền Núi Bắc Bộ... Đã có các tool python dựng sẵn, chỉ cần chạy trên môi</p>	

	<p>trường annacoda của google để lấy kết quả, mất nhiều thời gian vì dữ liệu khá lớn.)</p> <p>5. Phân tích thiết kế các Common Summary and Data Value.</p>	
3 Từ ngày 08/07/2024 đến ngày 13/07/2024	<p>1. Thiết kế diagram cho các Master Data sẽ có trong Data Warehouse</p> <p>2. Làm sạch dữ liệu lần 1 cho các dữ liệu giao dịch xuất nhập khẩu 2014, 2015.</p> <p>3. Làm sạch dữ liệu lần 2 cho các dữ liệu giao dịch xuất nhập khẩu 2014, 2015.</p> <p>4. Kiểm thử các dữ liệu đã được giao và báo cáo dữ liệu lỗi phát hiện được.</p> <p>5. Load dữ liệu giao dịch của năm 2019, 2020 vào database của công ty và kiểm thử lại dữ liệu sau khi load lên.</p>	
4 Từ ngày 15/07/2024 đến ngày 20/07/2024	<p>1. Xử lý các dữ liệu sai đã tìm được bằng tool python đã xây dựng trước đó.</p> <p>2. Tiến hành Crawl dữ liệu chi tiết về các doanh nghiệp dựa vào mã số thuế trong dữ liệu để dựng Master Data của các giao dịch import 2021. (Sử dụng tool selenium python đã xây dựng sẵn)</p> <p>3. xử lý dữ liệu bị vần đề trường total_value bị tính toán sai và trường exchange_rate bằng 0.</p>	
5 Từ ngày 22/07/2024 đến ngày 27/07/2024	<p>1. Tiếp tục crawl dữ liệu về doanh nghiệp import của năm 2013.</p> <p>2. Crawl dữ liệu tỷ giá chuyển đổi giữa USD và VND (thuộc trường exchange_rate trong dữ liệu) từ năm 2012 đến 2023 và load vào Master Data của công ty.</p> <p>3. Tạo các Materialized View theo diagram Common Summary và Data Value đã thiết kế trước đó cho dữ</p>	

	liệu từ năm 2012 đến 2023, giúp team Power BI truy vấn dữ liệu nhanh chóng cải thiện tốc độ truy vấn cũng như phân tích dữ liệu.	
6 Từ ngày 29/07/2024 đến ngày 03/08/2024	<p>1. Crawl dữ liệu doanh nghiệp từ nguồn dữ liệu gồm tên và địa chỉ được lấy từ chi tiết các giao dịch export 2022.</p> <p>2. Kiểm tra dữ liệu trong database những dòng bị miss data trong quá trình chạy tool.</p> <p>3. Cải tiến tool crawl dữ liệu doanh nghiệp export python.</p>	
7 Từ ngày 05/08/2024 đến ngày 10/08/2024	<p>1. Load và lọc trùng cho các bảng trong Master Data.</p> <p>2. Xây dựng tool python kiểm thử dữ liệu sau khi cào có trùng khớp với địa chỉ và tên công ty trong dữ liệu hay không.</p> <p>3. Xử lý dữ liệu không trùng khớp và tiếp tục crawl doanh nghiệp export 2021.</p>	
8 Từ ngày 12/08/2024 đến ngày 17/08/2024	<p>1. Kiểm thử dữ liệu doanh nghiệp đã lấy được và lọc trùng.</p> <p>2. Tạo khóa chính, khóa ngoại của các bảng Master Data và tham chiếu đến dữ liệu chính.</p> <p>3. Hoàn tất bàn giao công việc (tool, tài liệu,...) cho công ty.</p>	

Chuyên gia doanh nghiệp hướng dẫn thực tập

(Ký tên và ghi họ tên)

**ỦY BAN NHÂN DÂN TP. HỒ CHÍ
MINH TRƯỜNG ĐẠI HỌC SÀI GÒN**

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc**

KHOA CNTT

**BẢNG ĐÁNH GIÁ QUÁ TRÌNH THỰC TẬP TỐT NGHIỆP
(do chuyên gia doanh nghiệp đánh giá).**

Họ và tên sinh viên: Võ Quang Đăng Khoa .

Ngày sinh: 21/03/2002

Mã số sinh viên: 3120560047

Lớp : DKP1201

Thời gian thực tập: 24/06/2024 – 17/08/2024

Doanh nghiệp thực tập: Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững

Địa chỉ doanh nghiệp: A07-08 tòa Sarica, Đ. D9, KĐT Sala, P. An Lợi Đông, Q.2, TP. HCM

Chuyên gia doanh nghiệp hướng dẫn: Anh Ngô Trí Thanh

I. ĐÁNH GIÁ VỀ QUÁ TRÌNH THỰC TẬP

STT	Nội dung đánh giá	ĐIỂM		
		0	0.5	1
1	Khả năng thực hành			
2	Khả năng làm việc nhóm			
3	Tính thân thiện			
4	Tính năng động			
5	Tính thần sáng tạo			
6	Chấp hành nội quy cơ quan			
7	Giờ giấc làm việc			
8	Phương pháp làm việc			
9	Khối lượng công việc			
10	Báo cáo thực tập tốt nghiệp			

(theo thang điểm 10).

II. CÁC ĐÁNH GIÁ KHÁC:

.....
.....
.....
.....

III. KẾT QUẢ TỔNG HỢP:

Điểm tổng cộng :

XÁC NHẬN CỦA DOANH NGHIỆP
(đóng mộc tròn của doanh nghiệp, họ tên, ký tên)

Chuyên gia hướng dẫn
(Ký và ghi họ tên)

KHOA CNTT

PHIẾU ĐÁNH GIÁ KẾT QUẢ THỰC TẬP TỐT NGHIỆP
(do giảng viên hướng dẫn đánh giá)

Họ và tên sinh viên: Võ Quang Đăng Khoa .

Ngày sinh: 21/03/2002

Mã số sinh viên: 3120560047

Lớp : DKP1201

Thời gian thực tập: 24/06/2024 – 17/08/2024

Doanh nghiệp thực tập: Công Ty Cổ Phần Giải Pháp Dệt May Bền Vững

Địa chỉ doanh nghiệp: A07-08 tòa Sarica, Đ. D9, KĐT Sala, P. An Lợi Đông, Q.2, TP. HCM

Chuyên gia doanh nghiệp hướng dẫn: Anh Ngô Trí Thanh.

I. ĐIỂM CỦA CHUYÊN GIA DOANH NGHIỆP :

(thang điểm 10)

II. ĐIỂM CỦA GIẢNG VIÊN HƯỚNG DẪN:

(thang điểm 10)

II. ĐIỂM TỔNG KẾT:

(trung bình cộng 2 cột điểm trên, thang điểm 10)

Xếp loại :

TP Hồ Chí Minh ngày tháng năm
Giảng viên hướng dẫn

Kết luận chương 3:

Thực tập tại Công ty Cổ Phần Giải Pháp Dệt May Bên Vũng là một trải nghiệm đáng nhớ. Em đã được tham gia vào dự án xây dựng kho dữ liệu, một công việc đòi hỏi sự tỉ mỉ và kiến thức chuyên môn. Qua quá trình làm việc, em đã không chỉ nâng cao kỹ năng xử lý dữ liệu mà còn hiểu rõ hơn về tầm quan trọng của thông tin trong việc ra quyết định kinh doanh. Những kiến thức và kinh nghiệm tích lũy được sẽ là hành trang quý giá giúp em tự tin bước vào môi trường làm việc chuyên nghiệp.

CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ

4.1. Kết luận

Qua kỳ thực tập tại Công ty Cổ Phần Giải Pháp Dệt May Bền Vững, em đã có những trải nghiệm thực tế vô cùng quý báu. Môi trường làm việc chuyên nghiệp tại công ty đã giúp em rèn luyện khả năng ứng dụng kiến thức lý thuyết vào thực tiễn, đồng thời em cũng nhận thức rõ hơn về tầm quan trọng của việc quản lý dữ liệu hiệu quả. Bên cạnh đó, các kỹ năng mềm như làm việc nhóm, giao tiếp và giải quyết vấn đề cũng được em nâng cao đáng kể. Đây là hành trang quý giá giúp em tự tin hơn trong công việc và cuộc sống.

4.2. Kiến nghị

Để nâng cao hiệu quả hoạt động và phát triển bền vững, công ty nên:

- Số hóa sản xuất:** Đầu tư hệ thống ERP, xây dựng kho dữ liệu lớn, ứng dụng AI vào kiểm soát chất lượng. Tập trung vào việc ứng dụng công nghệ để nâng cao hiệu quả sản xuất, giảm thiểu lãng phí và tăng năng suất.
- Nâng cao năng lực nhân viên:** Đào tạo chuyên sâu, xây dựng chương trình mentoring, phát triển kỹ năng mềm. Việc đầu tư vào con người rất quan trọng, giúp nhân viên nâng cao kiến thức và kỹ năng để đáp ứng yêu cầu công việc.
- Cải thiện môi trường làm việc:** Xây dựng văn hóa doanh nghiệp, cải thiện cơ sở vật chất, áp dụng chính sách linh hoạt. Tạo ra một môi trường làm việc tích cực, khuyến khích sự sáng tạo và gắn kết giữa các thành viên trong công ty.
- Tối ưu hóa quy trình:** Áp dụng Lean Manufacturing, tìm kiếm giải pháp tự động hóa.

Cuối cùng, em xin gửi lời cảm ơn chân thành đến thầy Phạm Thế Bảo và toàn thể đồng nghiệp tại Công ty Cổ phần Giải Pháp Dệt May Bền Vững đã tạo điều kiện và tận tình hướng dẫn em trong suốt thời gian qua.

TÀI LIỆU THAM KHẢO

[1]. Khái niệm về Database, Data Warehouse, Data Mart:

<https://mastering-da.com/11-phan-biet-database-data-warehouse-data-mart-data-lakehouse-data-mesh-p1/>

[2]. Khái niệm về Big Data:

<https://topdev.vn/blog/big-data/>

[3]. Khái niệm về Galaxy Schema, Snowflake Schema, Star Schema:

https://www.softwareag.com/en_corporate/blog/streamsets/schemas-data-warehouses-star-galaxy-snowflake.html