TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

KHOA HỆ THÔNG THÔNG TIN

BÁO CÁO KIẾN TẬP HƯỚNG NGÀNH BI

NĂM HỌC 2022 - 2023

TÊN ĐỀ TÀI

XÂY DỰNG GIẢI PHÁP BUSINESS INTELLIGENCE TRÊN NỀN TẢNG ĐÁM MÂY MICROSOFT AZURE KẾT HỢP ELT ĐỘNG

Công ty: TNHH Bảo Hiểm Nhân Thọ Prudential Việt Nam

GVHD: Ths. Trường Hoài Phan

Hướng dẫn: Ths. Lê Bá Thiền

DANH SÁCH NHÓM

STT	MSSV	Lớp	Họ và tên	Vai trò
1	K204061440	K20406T	Trần Nhật Nguyên	Nhóm trưởng
2	K204061446	K20406C	Man Đắc Sang	Thành viên

PHIẾU ĐÁNH GIÁ KẾT QUẢ KIẾN TẬP

Tên đơn vị kiến tập: **Công ty TNHH Bảo hiểm nhân thọ Prudential Việt Nam**

Địa chỉ đơn vị: Tầng 25, Tòa nhà Sài Gòn Trade Center, 37 Tôn Đức Thắng, Phường Bến Nghé, Quận 1, Tp. Hồ Chí Minh

Họ và tên người đại diện đơn vị: **Lê Bá Thiền**

Chức vụ: Data Engineer

Thông tin liên hệ: thienlb.ktl@uel.edu.vn

Nhận xét:

- Nhóm sinh viên có thái độ nghiêm túc, tích cực trong việc trao đổi và thực hiện đề tài.
- Thực tế ELT là một quy trình tích hợp dữ liệu được các doanh nghiệp lớn sử dụng, đặc biệt tại công ty Prudential Việt Nam. Tuy là kiến thức mới nhưng các em đã tiếp thu rất nhanh, hiểu khá rõ về sự sai khác của quy trình này so với các quy trình truyền thống như ETL, các em đã có thể mô phỏng lại quy trình dựa trên dữ liệu thực nghiệm. Điều này rất đáng khen ngợi và cần phát huy.

Thành phố Hồ Chi Minh, ngày 20 tháng 06 năm 2023

XÁC NHẬN CỦA ĐƠN VỊ KIẾN TẬP

(Ký và ghi rõ họ tên)

Lê Bá Thiền

PHIẾU ĐÁNH GIÁ KẾT QUẢ CỦA GIẢNG VIÊN HƯỚNG DẪN

GVHD: Ths. Trương Hoài Phan

STT	Tiêu chí	Tiêu chí cụ thể	Điểm	Ghi chú
1	**Hình thức báo cáo**	Trình bày		
		Kết cấu báo cáo		
		Văn phong		
2	**Nội dung báo cáo**	Kỹ năng phân tích		
		Mục tiêu		
		Chuyên môn		
3	**Thái độ của sinh viên**			
4	**Doanh nghiệp đánh giá**			

……., ngày … tháng … năm 2023

GIẢNG VIÊN HƯỚNG DẪN

(Ký và ghi rõ họ tên)

MỤC LỤC

DANH SÁCH NHÓM 1
PHIẾU ĐÁNH GIÁ KẾT QUẢ KIẾN TẬP 2
PHIẾU ĐÁNH GIÁ KẾT QUẢ CỦA GIẢNG VIÊN HƯỚNG DẪN 3
DANH MỤC BẢNG BIỂU 7
DANH MỤC HÌNH ẢNH 8
LỜI NÓI ĐẦU 10
CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 11
1.1 Đặt vấn đề và tính thực tiễn đề tài 11
1.2 Mục tiêu đề tài 11
1.2.1 Đối với doanh nghiệp 11
1.2.2 Đối với nhóm sinh viên kiến tập 11
1.3 Đối tượng và phạm vi đề tài 12
1.3.1 Đối tượng 12
1.3.2 Phạm vi 12
1.4 Tổng quan về công ty kiến tập 12
1.4.1 Thông tin về công ty 12
1.4.2 Quá trình hình thành và phát triển 12
1.4.3 Tầm nhìn và mục tiêu 13
1.4.4 Biểu tượng công ty 13
1.5 Hướng nghề nghiệp của nhóm - Trí tuệ doanh nghiệp (BI) 14
1.5.1 Vai trò 14
1.5.2 Yêu cầu kiến thức, kỹ năng, kinh nghiệm và trình độ 14
1.5.3 Cơ hội 14
1.5.5 Thách thức 15

1.6 Nội dung trình bày 15
CHƯƠNG 2 CƠ SỞ LÝ THUYẾT 17
2.1 Tổng quan về Business Intelligence (BI) 17
2.1.1 BI là gì 17
2.1.2 Kiến trúc BI 17
2.1.3 Lợi ích của BI đối với doanh nghiệp 18
2.2 Quy trình ETL 18
2.3 Quy trình ELT 19
2.4 Sự khác nhau giữa quy trình ETL và ELT 20
2.5 Data warehouse và Data mart 21
2.5.1 Data warehouse (kho dữ liệu) 21
2.5.2 Data mart 22
2.7 Data Lake 22
2.6 KPIs 23
2.7 Azure và các thành phần dùng trong dự án 24
2.7.1 Azure 24
2.7.2 Azure Data Factory 24
2.7.3 Blob storage 25
2.7.4 Azure SQL 25
2.8 Công cụ trực quan hóa dữ liệu - Power BI 25
CHƯƠNG 3: PHÂN TÍCH CHI TIẾT VÀ XÂY DỰNG MÔ HÌNH 26
3.1 Nguồn dữ liệu 26
3.1.1 Cơ sở dữ liệu quan hệ 26
3.1.2 Hệ thống kế toán 27
3.1.3 Trang web thương mại điện tử 28

3.2 Phân tích vấn đề 28
3.4 Đề xuất mô hình 30
CHƯƠNG 4: THỰC NGHIỆM 31
4.1 Xây dựng hồ dữ liệu (Data Lake) 31
4.1.1 Vùng dữ liệu 31
4.1.2 Azure SQL Server bên trong hồ dữ liệu 33
4.1.3 Quy trình ELT động 33
4.2 Xây dựng nhà kho dữ liệu (Data warehouse) 36
4.2.1 Bus matrix 36
4.2.2 Master data 37
4.2.3 Transaction data 37
4.2.4 ETL mapping 38
4.2.5 Bảng Fact và Dimension 42
4.2.6 Mô hình nhà kho dữ liệu (Data warehouse) 44
4.2.7 Quy trình ETL 44
CHƯƠNG 5 PHÂN TÍCH DỮ LIỆU - TRỰC QUAN HÓA 48
5.1 Dashboard là gì? 48
5.2 Dashboard đề xuất 48
CHƯƠNG 6: KẾT LUẬN 52
6.1 Kết luận cho đề tài 52
6.2 Kết luận cho nhóm thực hiện kiến tập 52
6.3 Hạn chế thực hiện đề tài 53
6.4 Phương hướng phát triển 53
TÀI LIỆU THAM KHẢO 54
BẢNG PHÂN CÔNG CÔNG VIỆC 55

DANH MỤC BẢNG BIỂU

Bảng 2-1. So sánh quy trình ETL và ELT 20

Bảng 4-1. Bus Matrix 37

Bảng 4-2. Master data 37

Bảng 4-3. Transaction data 37

Bảng 4-4. ETL mapping 38

Bảng 5-1. Mô tả phân khách hàng dựa vào đặc tính mua hàng 49

DANH MỤC HÌNH ẢNH

Hình 1.1 Biểu tượng tập đoàn PRUNDENTIAL 13
Hình 2-1. Benefits of BI followed by a survey of 2600 users by BI-Survey.com 18
Hình 2-2. Quy trình ETL 19
Hình 2-3. Quy trình ELT 19
Hình 2-4. Sự khác nhau giữa ELT và ETL 20
Hình 2-5. Data warehouse và Data mart là gì 22
Hình 2-6. Ví dụ kiến trúc của Data Lake 23
Hình 2-7. Azure Data Factory là gì? 24
Hình 3-1. Nguồn dữ liệu của công ty 26
Hình 3-2. Mô hình ERD của cơ sở dữ liệu quan hệ 27
Hình 3-3. Giải pháp BI (Nguồn: Tác giả đề xuất) 30
Hình 4-1. Vùng chứa dữ liệu bên trong hồ dữ liệu 31
Hình 4-2. Cấu trúc vùng chứa dữ liệu thô (rawdata) 32
Hình 4-3. Cấu trúc vùng dữ liệu được sắp xếp (curated) 32
Hình 4-4. Đường ống dữ liệu tổng thể quy trình ELT động 34
Hình 4-5. Tổng quan quá trình tải dữ liệu vào vùng chứa curated 34
Hình 4-6. Tổng quan quá trình tải dữ liệu vào Azure SQL Server 35
Hình 4-7. Quá trình kiểm tra và nhập dữ liệu vào Azure SQL Server 36
Hình 4-1. Bảng FactSale 42
Hình 4-2. Bảng DimSeller 42
Hình 4-3. Bảng DimProduct 43
Hình 4-4. Bảng DimTime 43
Hình 4-5. Bảng DimCustomer 43

Hình 4-6. Data Warehouse Star Schema 44
Hình 4-11. Pipeline tổng quát thực hiện trình tự 2 Pipeline “LoadDim” và “LoadFact”. 45
Hình 4-7. Data flow “Src2Dim” thực hiện ETL dữ liệu vào các bảng Dimension 45
Hình 4-8. Pipeline “Pipeline LoadDim” thực thi các Data flow “Src2Dim”..... 46
Hình 4-9. Data flow “Src2FactSales” thực hiện ETL dữ liệu vào bảng FactSales..... 47
Hình 4-10. Pipeline “Pipeline LoadFact” thực thi Data flow “Src2FactSales”..... 47
Hình 5-1. Sales Dashboard..... 48
Hình 5-3. Phân khúc khách hàng sử dụng mô hình RFM..... 49
Hình 5-4. Biểu đồ cột và đường thể hiện doanh số và phần trăm doanh số theo tháng ... 50
Hình 5-5. Biểu đồ thanh thể hiện top 5 danh mục sản phẩm có doanh thu cao nhất..... 50
Hình 5-6. Biểu đồ phân bổ doanh thu theo địa lý 51

LỜI NÓI ĐẦU

Lời đầu tiên, nhóm xin gửi lời cảm ơn sâu sắc đến Ths. Trương Hoài Phan, Ths. Lê Bá Thiền và quý Công ty TNHH Bảo hiểm nhân thọ Prudential Việt Nam đã hỗ trợ nhóm trong suốt quá trình thực hiện kiến tập. Qua quá trình nghiên cứu, học tập môn học này, nhóm đã nhận được sự bài toán rất hay và sự chỉ dẫn tận tâm từ anh hướng dẫn. Các thành viên trong nhóm đã có thêm kiến thức kỹ năng mới bổ ích cho công việc sau này. Đề tài "Xây dựng giải pháp Business Intelligence trên nền tảng đám mây Microsoft Azure kết hợp quy trình ELT động" được nhóm xây dựng nhằm giải quyết vấn đề được đặt ra.

Tuy nhiên, đề tài còn tồn đọng những hạn chế nhất định và nhiều thiếu sót cần bổ sung. Chúng em rất sẵn lòng nhận được ý kiến nhận xét từ thầy, anh hướng dẫn và quý công ty để có thể hoàn thiện nghiên cứu hơn trong tương lai.

Kính chúc thầy và quý công ty nhiều sức khỏe và đạt được nhiều thành công trong công việc, cuộc sống.

Chúng em xin trân trọng cảm ơn.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề và tính thực tiễn đề tài

Trong môi trường kinh doanh hiện nay, các doanh nghiệp đang sử dụng nhiều hệ thống khác nhau và dữ liệu được phân tán ở nhiều nguồn và được định dạng ở các loại tệp khác nhau là vấn đề bất cập cho doanh nghệp. Việc nhập và lưu trữ dữ liệu là một hoạt động quan trọng trong quản lý thông tin của doanh nghiệp, vì thế nếu gặp sai lệch hoặc thất thoát dữ liệu có thể dẫn đến nhiều hậu quả xấu như mất tính nhất quán của dữ liệu, tạo ra chi phí không cần thiết và ảnh hưởng đến quá trình ra quyết định của doanh nghiệp.

Vấn đề đặt ra ngay lúc này đó là làm thế nào để tổng hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau và lưu trữ tại một hệ thống ngay cả khi có hoặc không có yêu cầu phân tích. Dữ liệu được nhập từ nhiều nguồn khác nhau, sản sinh ra nhiều định dạng khác nhau cũng là một trong những khó khăn được đặt ra cho việc dữ liệu cần tính đồng nhất. Và khi có yêu cầu phân tích từ doanh nghiệp thì làm sao để giúp các bộ phân trong doanh nghiệp tiếp cận dữ liệu một cách dễ dàng và thuận tiện hơn, giảm thời gian và chi phí, tăng tính linh hoạt và khả dụng của dữ liệu.

1.2 Mục tiêu đề tài

1.2.1 Đối với doanh nghiệp

Đề tài hỗ trợ đưa ra và triển khai giải pháp liên quan đến vấn đề mà doanh nghiệp đang tốn nhiều thời gian trong quá trình nhập, lưu trữ dữ liệu và đưa ra quyết định kinh doanh. Nhóm đề xuất đưa ra những ý tưởng về hệ thống BI dựa trên nền tảng Cloud Azure chi tiết nhất về vấn đề cũng như cung cấp kết quả giải pháp trong việc triển khai.

1.2.2 Đối với nhóm sinh viên kiến tập

Đối với cá nhân các thành viên nhóm sinh viên kiến tập, thông qua đề tài, nhóm hướng đến mục tiêu và kiến thức nghề trí tuệ kinh doanh trong một tình huống thực tế cũng như hỗ trợ hết mình cho doanh nghiệp kiến tập trong việc đưa ra ý tưởng triển khai để giải quyết vấn đề. Từ những phân tích, thiết kế hệ thống, bài toán phải giải quyết trong đề tài, nhóm mong muốn đạt được những kiến thức sâu rộng hơn về hướng nghề trí tuệ kinh doanh trong thời đại số hiện này, cùng với những kỹ năng cần được trang bị trong nghề

cũng như đạo đức nghề nghiệp và tác phong làm việc trong một doanh nghiệp thực thụ. Đồng thời, nhóm cũng hướng đến các kĩ năng mềm trong quá trình làm việc và trao đổi nhóm trong môi trường chuyên nghiệp.

1.3 Đối tượng và phạm vi đề tài

1.3.1 Đối tượng

Đối tượng hướng đến của đề tài là những đối tượng cần dùng đến dữ liệu từ các hệ thông doanh nghiệp để thực hiện việc phân tích kinh doanh, hỗ trợ quá trình ra quyết định đúng đắn hơn trong doanh nghiệp: Các nhân viên của từng phòng ban. Với tiêu chí đơn giản, tăng tính khả dụng, linh hoạt, khả năng mở rộng cao tiết kiệm thời gian, giảm thiểu công việc lặp đi lặp lại, nâng cao hiệu suất làm việc cho nhân viên của từng bộ phận. Ngoài ra, nhóm cũng mong muốn triển khai hệ thống BI dựa trên nền tảng Cloud Azure sẽ đảm bảo tăng tính bảo mật và ổn định trong việc tối ưu quy trình làm việc cho doanh nghiệp.

1.3.2 Phạm vi

Về nội dung: mô hình triển khai hệ thống BI dựa trên nền tảng Cloud Azure tại công ty Prudential.

Về thời gian: đề tài được thực hiện trong 1 tháng tính từ ngày 01/06/2023.

1.4 Tổng quan về công ty kiến tập

1.4.1 Thông tin về công ty

Thành lập năm 1848 tại Luân Đôn, Tập đoàn Prudential là một trong những tập đoàn tài chính hàng đầu thế giới và đang phục vụ hơn 17 triệu khách hàng tại châu Á và châu Phi, được niêm yết trên các sàn giao dịch chứng khoán ở Luân Đôn, Hồng Kông, Singapore và New York…. Tập đoàn cung cấp các giải pháp bảo hiểm nhân thọ, bảo hiểm sức khỏe, và quản lý tài sản.

Tại Việt Nam công ty TNHH Bảo hiểm nhân thọ Prudential Việt Nam là một trong số các công ty con của tập đoàn, được thành lập vào năm 1999.

1.4.2 Quá trình hình thành và phát triển

Công ty TNHH Bảo hiểm nhân thọ Prudential Việt Nam chính thức hoạt động vào năm 1999 sau 4 năm thành lập văn phòng đại diện tại Hà Nội (1995). Sau nhiều năm hoạt

động công ty đã phát triển vượt bật với những con số ấn tượng như tổng tài sản tính đến cuối năm 2022 là hơn 161 nghìn tỷ VND, luôn nằm trong top những công ty bảo hiểm nhân thọ chiếm thị phần cao nhất Việt Nam.

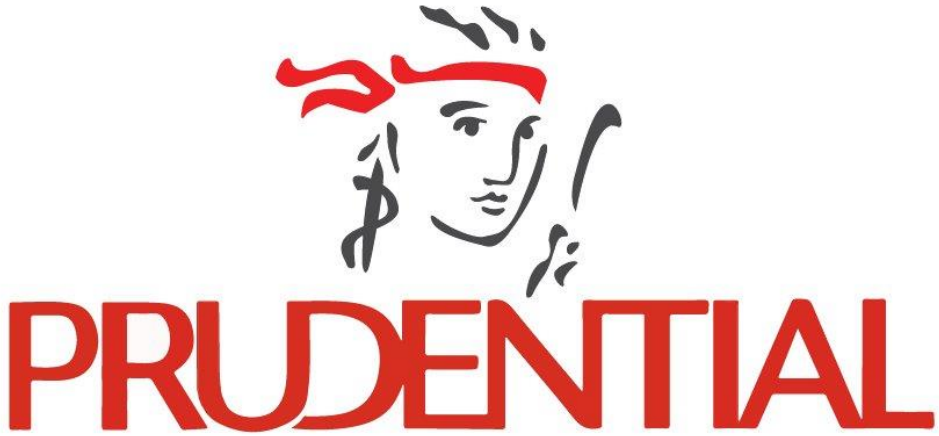
Là công ty bảo hiểm với lượng khách hàng lớn kèm theo số lượng giao dịch quan trọng khổng lồ, việc lưu trữ và sử dụng dữ liệu là vô cùng cần thiết. Do đó, doanh nghiệp đã tiên phong ứng dụng công nghệ hiện đại để đơn giản hóa quy trình, nâng cao trải nghiệm khách hàng và hỗ trợ cộng đồng trên hành trình làm chủ sức khỏe cũng như cuộc sống của chính mình.

1.4.3 Tầm nhìn và mục tiêu

Tầm nhìn: Là lựa chọn số 1 về bảo hiểm nhân thọ cho mọi gia đình Việt.

Mục tiêu: Giúp mọi người đạt được những điều tốt đẹp trong cuộc sống.

1.4.4 Biểu tượng công ty



Hình 1.1 Biểu tượng tập đoàn PRUNDENTIAL

Biểu tượng được lấy cảm hứng từ tranh vẽ nữ thần Prudence của Ngài Joshua Reynolds. Nữ thần được biểu trưng cho 4 yếu tố: sự thận trọng, sự công bằng, liêm chính và tính cách chuẩn mực.

Logo thương hiệu Prudential chính thức sử dụng từ năm 1986 với câu khẩu hiệu "Luôn luôn lắng nghe. Luôn luôn thấu hiểu." ngày nay biểu tượng Prudence đã được cải tiến nhưng vẫn giữ nét cơ bản như trước.

1.5 Hướng nghề nghiệp của nhóm - Trí tuệ doanh nghiệp (BI)

1.5.1 Vai trò

Trong doanh nghiệp, BI là người có nhiệm vụ phân tích và đưa ra các báo cáo, thông tin phân tích và đề xuất giúp tổ chức đưa ra những quyết định kinh doanh chính xác và hiệu quả. Vai trò của BI bao gồm: thu thập và phân tích dữ liệu, xây dựng và phát triển hệ thống BI, xây dựng các báo cáo và thông tin phân tích, đảm bảo tính bảo mật dữ liệu, …

1.5.2 Yêu cầu kiến thức, kỹ năng, kinh nghiệm và trình độ

- Bằng cử nhân thuộc chuyên ngành: Khoa học máy tính, Khoa học dữ liệu, Công nghệ thông tin, Hệ thống thông tin hoặc chuyên ngành khác liên quan;
- Kinh nghiệm làm việc được chứng minh qua các dự án từng tham gia về thiết kế và xây dựng giải pháp BI;
- Kỹ năng toán học và thống kê vững chắc để đo lường, tổ chức và phân tích dữ liệu;
- Kỹ năng sử dụng các loại cơ sở dữ liệu (RDBMS, NoSQL, ...);
- Kỹ năng sử dụng ngôn ngữ SQL;
- Kỹ năng sử dụng ngôn ngữ lập trình (Java, Python, Scala, ...);
- Kiến thức và kỹ năng sử dụng các công cụ trực quan hóa dữ liệu như Power BI, Tableau, Excel, ...;
- Có tư duy tốt, khả năng nghiên cứu, đánh giá và cập nhật công nghệ mới;
- Trình độ tiếng anh tốt: nghe - nói - đọc - viết;
- Kỹ năng giao tiếp, thuyết trình, trình bày vấn đề trực quản, ngắn gọn, hiệu quả.

1.5.3 Cơ hội

Với xu hướng đưa ra quyết định dựa trên số ngày nay, cơ hội mở ra với ngành nghề BI dần trở thành một xu hướng quan trọng trong các tổ chức. Các ngành nghề khác nhau đều có nhu cầu sử dụng BI để phân tích và quản lý dữ liệu. Nghề BI được coi là môi trường tốt cho những người thích trải nghiệm ở nhiều mảng, học hỏi, từ đó nâng cao năng lực bản thân.

Mức thu thập hấp dẫn cũng là một trong những lý do mà nghề BI dần chiếm ưu thế trong thị trường tuyển dụng ngày nay. Với sự phát triển của ngành nghề và các công nghệ

liên quan đến BI, những người làm việc trong ngành này có nhiều cơ hội thăng tiến trong sự nghiệp của mình.

1.5.5 Thách thức

Với sự tăng trưởng nhanh chóng trong thời gian gần đây, việc cạnh trong giữa các ứng viên chuyên viên trí tuệ doanh nghiệp cũng đang tăng lên. Người làm BI phải luôn cập nhật công nghệ mới nhất liên quan đến dữ liệu và phân tích dữ liệu và học cách áp dụng chúng vào công việc của mình.

Độ chính xác của dữ liệu là một vấn đề quan trọng trong ngành nghề BI. Nếu dữ liệu không chính xác, điều này có thể dẫn đến các quyết định kinh doanh sai lầm, gây lãng phí chi phí và nguồn lực. Do đó, BI là người cần phải đảm bảo rằng dữ liệu được trích xuất và phân tích một cách chính xác.

Dữ liệu là tài sản quý giá của các tổ chức và việc bảo mật dữ liệu là cần thiết. Người làm BI phải có đầy đủ kiến thức về bảo mật dữ liệu và biết cách giữ cho dữ liệu được an toàn.

1.6 Nội dung trình bày

Trong báo cáo, nhóm tập trung trình bày các nội dung, bao gồm 5 chương:

- *Chương 1: Giới thiệu đề tài*

Nêu tổng quát vấn đề còn tồn đọng trong với việc lưu trữ và phân tích dữ liệu. Hơn nữa, các thuộc tính tổng quan về nghề nghiệp trí tuệ doanh nghiệp (Business Intelligence).

- *Chương 2: Cơ sở lý thuyết*

Giải thích các lý thuyết, khái niệm được sử dụng trong dự án.

- *Chương 3: Phân tích chi tiết và xây dựng mô hình*

Tại phần này dữ liệu nguồn dùng để thực nghiệm mô hình được phân tích từ tổng quan đến chi tiết. Đồng thời vấn đề được giới thiệu tại chương 1 sẽ được bàn luận, phân tích kỹ hơn từ đó đề xuất mô hình phù hợp.

- *Chương 4: Thực nghiệm*

Thực hiện mô hình đã đề xuất ở chương 3, tiến hành xây dựng Data Lake và Data Warehouse. Ngoài ra quy trình ELT và quy trình ETL được giới thiệu và phân tích chi tiết tại chương này.

- *Chương 5: Phân tích dữ liệu – Trực quan hóa*

Sau khi đã có được mô hình Data warehouse đề xuất, công cụ trực quan hóa Power Bi được kết nối và sử dụng. Sales Dashboard được giới thiệu.

- *Chương 6: Kết luận*

Các kết luận rút ra sau khi hoàn thành dự án, đồng thời nêu lên các điểm hạn chế và phương hướng phát triển trong tương lai.

CHƯƠNG 2 CƠ SỞ LÝ THUYẾT

2.1 Tổng quan về Business Intelligence (BI)

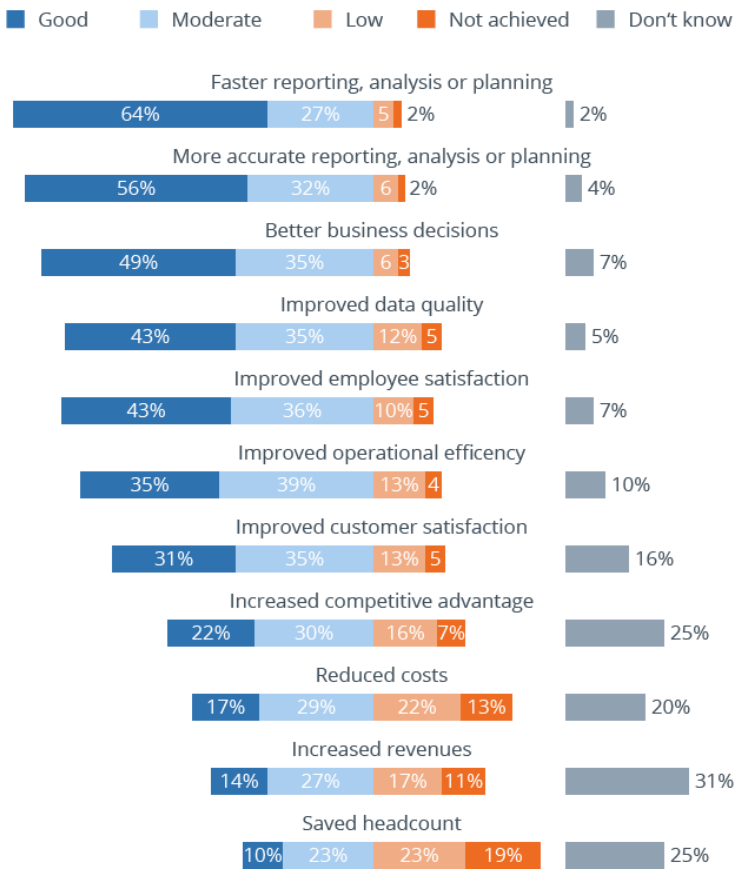
2.1.1 BI là gì

BI là một qui trình có tích hợp công nghệ mà các doanh nghiệp dùng để kiểm soát khối lượng dữ liệu khổng lồ đến từ nhiều nguồn khác nhau và khai thác nguồn dữ liệu đó giúp cho họ có thể đưa các các quyết định hiệu quả hơn trong hoạt động kinh doanh của mình. BI có mặt ở khắp các doanh nghiệp như hệ thống siêu thị, ngân hàng, viễn thông, … đó đều là những nơi cần thu thập, xử lý khối lượng dữ liệu cực lớn.

2.1.2 Kiến trúc BI

- *Nguồn dữ liệu:* Lớp này bao gồm tất cả các nguồn dữ liệu khác nhau, bao gồm cơ sở dữ liệu, bảng tính và các ứng dụng khác, được sử dụng để cung cấp thông tin BI.
- *Tích hợp dữ liệu:* bao gồm việc kết hợp dữ liệu từ nhiều nguồn thành một kho chứa hoặc kho dữ liệu. Các quy trình như trích xuất, chuyển đổi và tải (ETL) dữ liệu thường được sử dụng cho điều này.
- *Lưu trữ dữ liệu:* Lớp này bao gồm việc lưu trữ dữ liệu kết hợp một cách mà các công cụ BI có thể truy cập và đánh giá nhanh chóng. Điều này có thể bao gồm sử dụng một nền tảng kho dữ liệu chuyên dụng hoặc hệ thống quản lý cơ sở dữ liệu quan hệ thông thường (RDBMS).
- *Các công cụ BI:* Chương trình và các công cụ được sử dụng để phân tích và hiển thị dữ liệu, bao gồm các bảng điều khiển, khai thác dữ liệu và các công cụ phân tích dự đoán, được bao gồm trong lớp này.
- *Lớp trình bày:* Lớp này bao gồm việc cung cấp cho người dùng cuối thông tin được sản xuất bởi các công nghệ BI một cách dễ hiểu và có thể thực hiện được.
- *Lớp bảo mật:* Lớp này đảm bảo rằng dữ liệu được an toàn và bảo mật và chỉ những người được ủy quyền mới có quyền truy cập vào dữ liệu nhạy cảm**.**

2.1.3 Lợi ích của BI đối với doanh nghiệp



Hình 2-1. Benefits of BI followed by a survey of 2600 users by BI-Survey.com

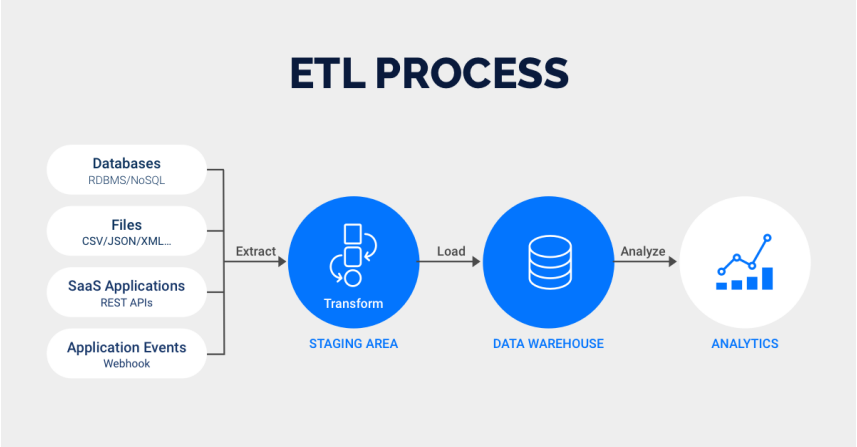
Theo một cuộc khảo sát của 2600 người dùng business intelligence bởi BI-Survey.com, sau đây là bảy lợi ích hàng đầu:

- Báo cáo, phân tích hoặc lập kế hoạch nhanh hơn
- Báo cáo, phân tích hoặc lập kế hoạch chính xác hơn
- Quyết định kinh doanh tốt hơn
- Cải thiện chất lượng dữ liệu
- Nâng cao sự hài lòng của nhân viên
- Tăng cường hiệu quả vận hành
- Nâng cao sự hài lòng của khách hàng

2.2 Quy trình ETL

ETL là viết tắt của Extract, Transform, Load, là quá trình trích xuất dữ liệu từ nguồn, chuyển đổi dữ liệu để đảm bảo tính độc lập và hiệu quả, sau đó tải dữ liệu đã được chuyển đổi vào cơ sở dữ liệu đích.

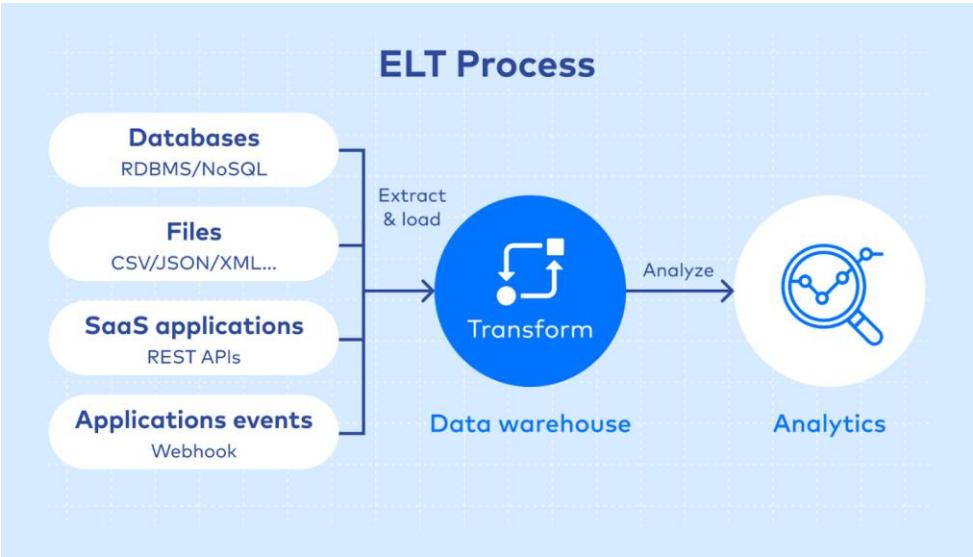
ETL là một quy trình quan trọng trong việc tích hợp dữ liệu từ các nguồn khác nhau để tạo ra một cơ sở dữ liệu toàn diện và được sử dụng rộng rãi trong các dự án phân tích và báo cáo dữ liệu. Công nghệ ETL đơn giản hóa quá trình tích hợp dữ liệu và cải thiện hiệu quả của các chuyên gia phân tích dữ liệu và quản lý dữ liệu.



Hình 2-2. Quy trình ETL

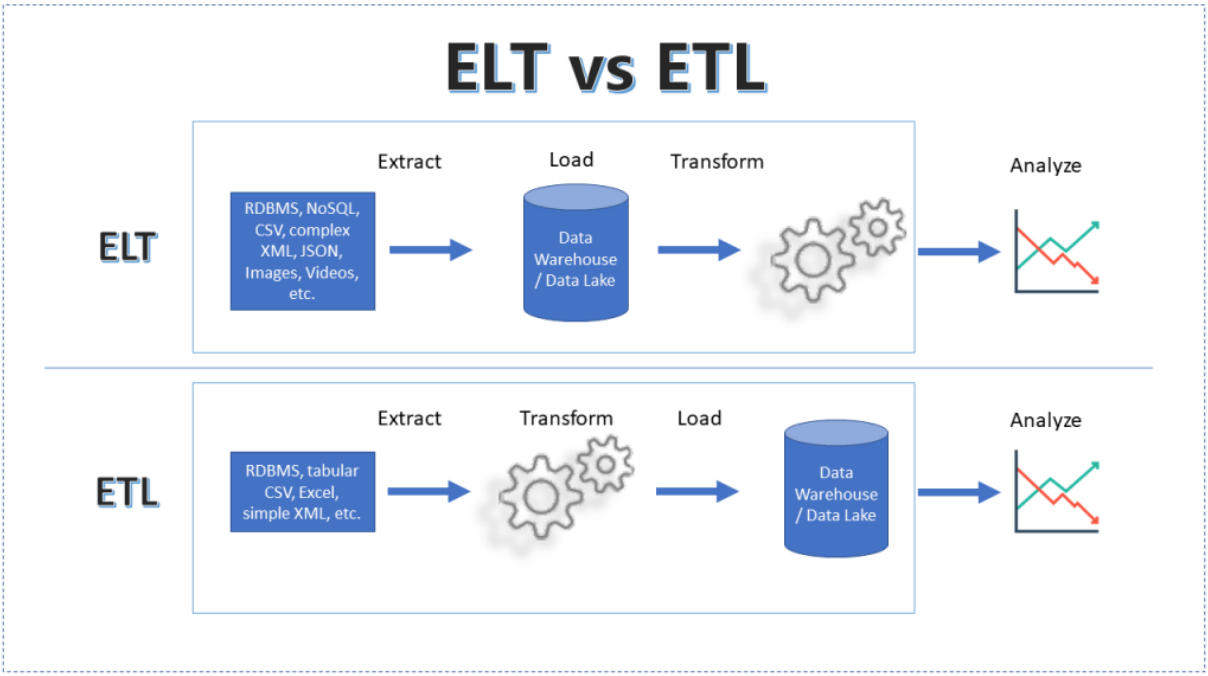
2.3 Quy trình ELT

ELT là viết tắt của "Extract, Load, Transform" tạm dịch là “Trích xuất, Tải, Biến đổi” là một quy trình tích hợp dữ liệu tương tự như ETL (Extract, Transform, Load). Sự khác biệt chính là trong ELT dữ liệu thô được trích xuất từ hệ thống nguồn và sau đó được tải vào nguồn tài nguyên đích, chẳng hạn như kho dữ liệu, trước khi được chuyển đổi thành định dạng có thể sử dụng.



Hình 2-3. Quy trình ELT

2.4 Sự khác nhau giữa quy trình ETL và ELT



Hình 2-4. Sự khác nhau giữa ELT và ETL

Bảng 2-1. So sánh quy trình ETL và ELT

	ETL	ELT
Định nghĩa	Là quá trình Dữ liệu được trích xuất từ hệ thống nguồn, được chuyển đổi trên máy chủ xử lý thứ cấp và được tải vào hệ thống đích.	Dữ liệu được trích xuất từ hệ thống nguồn, được tải vào hệ thống đích và được chuyển đổi bên trong hệ thống đích.
Quy trình	Cần phải xác định rõ nhu cầu phân tích là gì trước, cần sử dụng dữ liệu nào tiếp theo đó mới tiến hành quá trình ETL.	Không cần biết nhu cầu phân tích sẽ bao gồm dữ liệu nào. Tất cả dữ liệu liên quan sẽ được lưu trữ. Do đó, tính sẵn sàng đáp ứng nhu cầu cao hơn.
Tốc độ	Tốn nhiều thời gian, tất cả dữ liệu được chuyển đổi trước khi tải vào một hệ thống đích.	Tiết kiệm thời gian, dữ liệu được tải trực tiếp vào hệ thống đích và chỉ chuyển đổi dữ liệu cần thiết.

Kích thước/loại tập dữ liệu	ETL thích hợp nhất để xử lý các tập dữ liệu quan hệ, nhỏ hơn, yêu cầu các phép biến đổi phức tạp và đã được xác định trước là có liên quan đến các mục tiêu phân tích.	ELT có thể xử lý dữ liệu với kích thước lớn, nhiều kiểu dữ liệu khác nhau và rất phù hợp để xử lý dữ liệu cấu trúc và phi cấu trúc.
Môi trường	ELT thường được sử dụng trong các hệ thống lưu trữ dữ liệu lớn và phân tán. Nó tận dụng sức mạnh của các hệ thống lưu trữ dữ liệu phân tán để xử lý dữ liệu nhanh hơn.	ETL vẫn được sử dụng phổ biến trong các hệ thống dữ liệu truyền thống, trong đó việc biến đổi dữ liệu được thực hiện trước khi dữ liệu được tải vào kho lưu trữ dữ liệu. Điều này có thể làm tăng thời gian xử lý dữ liệu, đặc biệt là khi xử lý các tập dữ liệu lớn.

2.5 Data warehouse và Data mart

2.5.1 Data warehouse (kho dữ liệu)

Là một hệ thống được tạo ra để lưu trữ dữ liệu từ nhiều nguồn và môi trường khác nhau, chẳng hạn như phần mềm bán hàng, kế toán, nhân sự và hệ thống ngân hàng. Nó giúp cải thiện hiệu suất của các truy vấn cho báo cáo và phân tích dữ liệu.

Data warehouse hoạt động như một trung tâm lưu trữ, trong đó dữ liệu được nhập vào từ hệ thống giao dịch và các cơ sở dữ liệu khác. Sau đó, dữ liệu được xử lý và chuyển đổi để người dùng có thể truy cập thông qua các công cụ Business Intelligence, SQL client hoặc bảng tính.

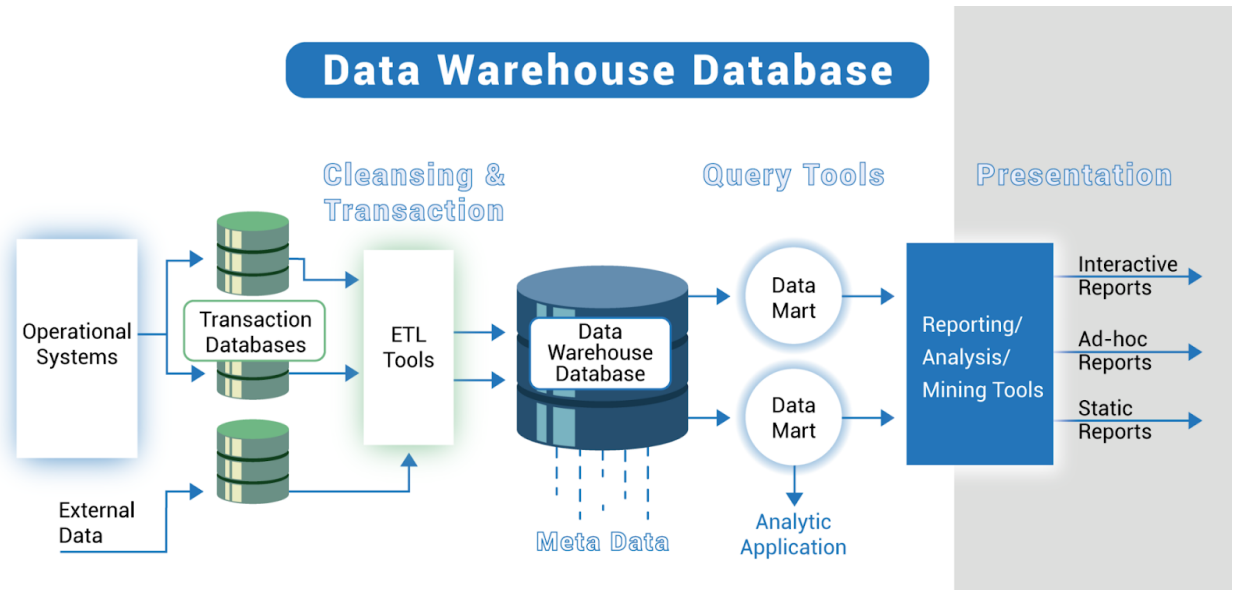
Những đặc tính của Data warehouse:

- *Hướng chủ đề (subject-oriented):* dữ liệu trong Data Warehouse được tổ chức theo các chủ đề hoặc lĩnh vực nghiên cứu khác nhau, chứ không phải theo các ứng dụng hoặc phòng ban cụ thể.

- *Được tích hợp (integrated):* tổ hợp và tích hợp dữ liệu từ nhiều nguồn khác nhau, chẳng hạn như các hệ thống giao dịch, cơ sở dữ liệu liên quan, tệp tin và các nguồn dữ liệu khác, để tạo ra một kho dữ liệu duy nhất, đồng nhất và toàn diện cho toàn bộ doanh nghiệp.
- *Có gán nhãn thời gian (time variant):* Mỗi điểm dữ liệu đều được gắn với một thời điểm cụ thể
- *Bất biến (non-volatile):* Dữ liệu lịch sử không thể sửa và kho dữ liệu chỉ có 2 thao tác chính là tải dữ liệu vào kho và truy cập (đọc) dữ liệu từ kho.

2.5.2 Data mart

Là phiên bản thu gọn của Kho dữ liệu và được thiết kế để sử dụng bởi một bộ phận, đơn vị hoặc nhóm người dùng cụ thể trong một tổ chức. Nó thường được kiểm soát bởi một bộ phận duy nhất trong một tổ chức. Data Mart thường chỉ lấy dữ liệu từ một vài nguồn so với kho dữ liệu. Nó có kích thước nhỏ và linh hoạt hơn so với một Datwarhouse.



Hình 2-5. Data warehouse và Data mart là gì

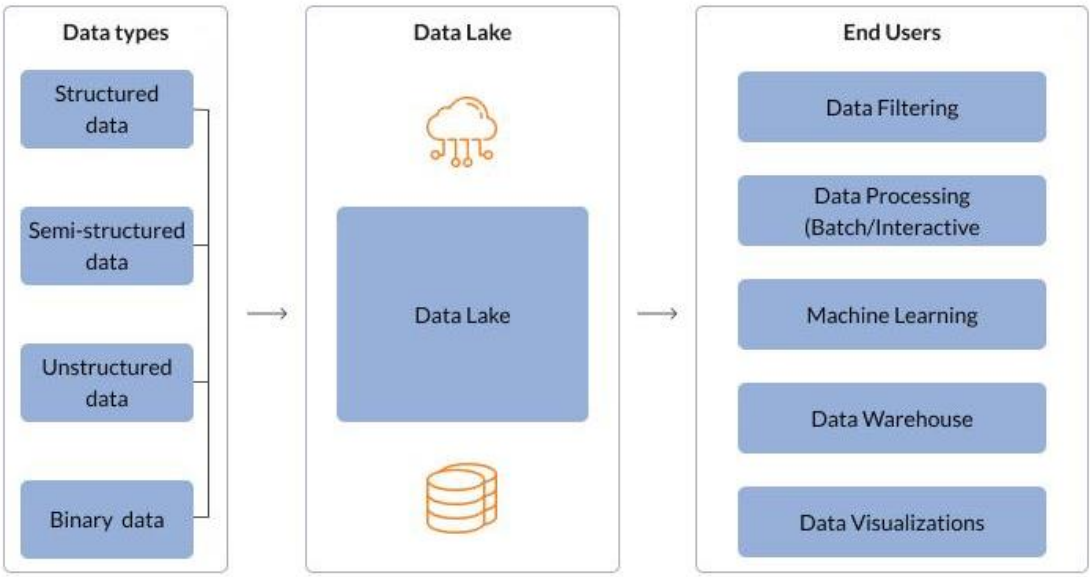
2.7 Data Lake

Data Lake: Là một hệ thống lưu trữ dữ liệu phi cấu trúc và không cố định với khả năng lưu trữ và xử lý dữ liệu ở mọi quy mô và định dạng. Nó cho phép lưu trữ tất cả các loại dữ liệu từ các nguồn khác nhau mà không cần xác định trước cấu trúc hoặc mô hình dữ liệu cụ thể. Data lake cho phép người dùng truy cập và phân tích các tập dữ liệu khác

nhau từ nhiều nguồn khác nhau, giúp họ tạo ra thông tin giá trị và đưa ra quyết định kinh doanh chính xác.

Một Data Lake có thể có nhiều kiểu kiến trúc vật lý khác nhau vì nó có thể được thực hiện bằng nhiều công nghệ khác nhau. Tuy nhiên, có ba nguyên tắc chính giúp phân biệt data lake với các phương pháp lưu trữ dữ liệu lớn khác:

- Tất cả dữ liệu được chấp nhận vào data lake
- Dữ liệu được lưu trữ ở dạng gốc
- Dữ liệu được chuyển đổi theo yêu cầu



Hình 2-6. Ví dụ kiến trúc của Data Lake

2.6 KPIs

Chỉ số hiệu suất chính (KPIs) là một tập hợp các đo lường định lượng được sử dụng để đánh giá hiệu suất của một công ty trong thời gian dài cho một mục tiêu cụ thể. KPIs cung cấp các mục tiêu để các nhóm tiếp cận, và thông tin hữu ích giúp mọi người trong tổ chức đưa ra quyết định tốt hơn. KPIs giúp tất cả các doanh nghiệp tiến lên ở mức chiến lược.

2.7 Azure và các thành phần dùng trong dự án

2.7.1 Azure

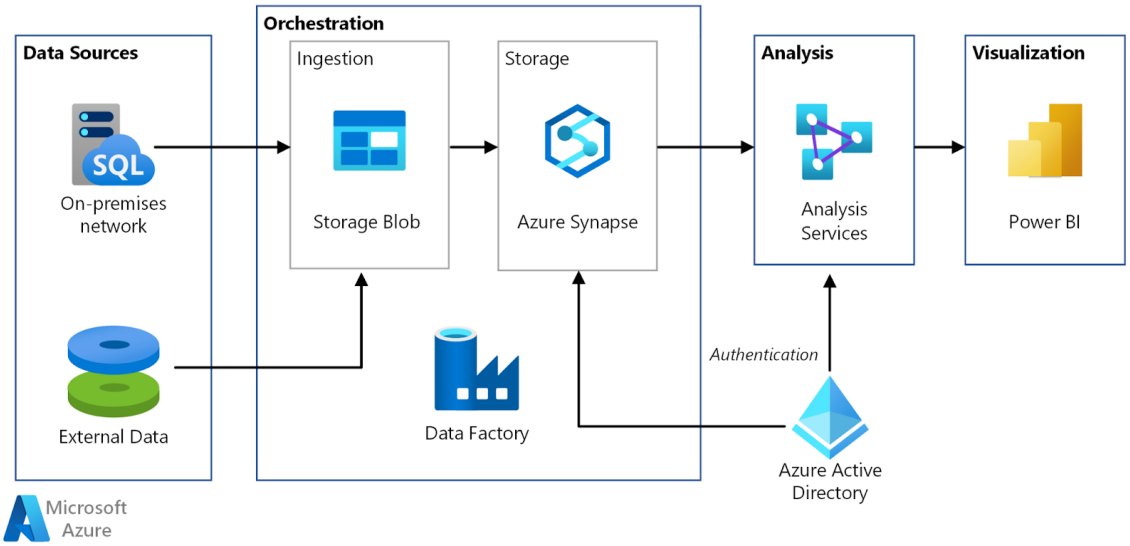
Azure là một nền tảng đám mây rộng lớn của Microsoft, cung cấp nhiều dịch vụ đám mây cho doanh nghiệp và cá nhân. Nó cho phép người dùng lưu trữ và quản lý dữ liệu của họ, chạy các ứng dụng và dịch vụ trên đám mây và tích hợp chúng với các ứng dụng và dịch vụ khác. Azure cung cấp các công cụ và tài nguyên để hỗ trợ việc phát triển, triển khai và quản lý các ứng dụng trên đám mây. Nó cũng cung cấp các giải pháp đám mây cho các ngành công nghiệp khác nhau và được sử dụng rộng rãi trên toàn cầu.

2.7.2 Azure Data Factory

Là một dịch vụ tích hợp dữ liệu dựa trên đám mây của Microsoft Azure. Nó cung cấp khả năng tự động hoá việc di chuyển và chuyển đổi dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu đích hoặc các ứng dụng phân tích dữ liệu.

Azure Data Factory giúp tối ưu hóa quá trình tích hợp dữ liệu bằng cách cho phép người dùng tạo các lưu trữ dữ liệu, lập lịch và quản lý các quy trình tích hợp dữ liệu khác nhau bằng cách sử dụng các giao diện đồ họa hoặc kịch bản mã hóa.

Với tính năng bảo mật, theo dõi, giám sát và báo cáo của nó, Azure Data Factory là một công cụ hữu ích cho các nhà phân tích dữ liệu và quản lý dữ liệu trong việc quản lý và tích hợp dữ liệu một cách hiệu quả.



Hình 2-7. Azure Data Factory là gì?

2.7.3 Blob storage

Blob storage là một dịch vụ lưu trữ đám mây của Microsoft Azure, được thiết kế để lưu trữ các tệp dữ liệu lớn, không cấu trúc và dễ mở rộng. Blob storage là một phần của Azure Storage và cung cấp khả năng lưu trữ dữ liệu độc lập với bất kỳ ứng dụng nào.

Ngoài ra, Blob storage cung cấp nhiều tính năng hữu ích cho việc lưu trữ và quản lý dữ liệu, bao gồm khả năng định cấu trúc theo dạng thư mục, cung cấp các giao thức truy cập đa dạng như HTTP và HTTPS, tính năng bảo mật và quản lý phiên bản dữ liệu

2.7.4 Azure SQL

Azure SQL là một dịch vụ cơ sở dữ liệu quan hệ đám mây được cung cấp bởi Microsoft Azure. Nó cho phép các nhà phát triển và doanh nghiệp tạo, quản lý và mở rộng các cơ sở dữ liệu quan hệ đám mây một cách dễ dàng và hiệu quả. Azure SQL hỗ trợ các phiên bản SQL Server phổ biến, bao gồm SQL Server 2005, 2008, 2012, 2014 và 2016. Với Azure SQL, người dùng có thể tận dụng tính linh hoạt và hiệu quả của đám mây để quản lý và phát triển các ứng dụng cơ sở dữ liệu quan hệ.

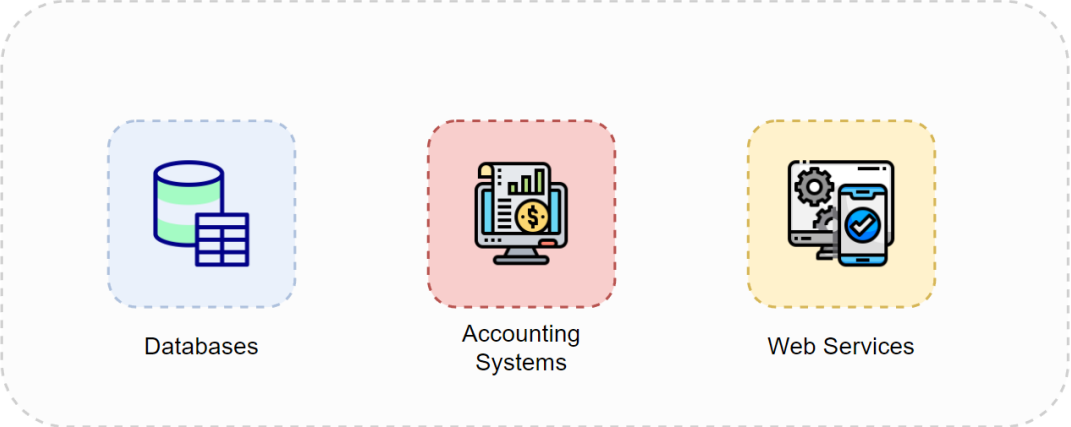
2.8 Công cụ trực quan hóa dữ liệu - Power BI

Power BI là một công cụ trực quan hóa dữ liệu mạnh mẽ của Microsoft, cho phép người dùng kết nối, phân tích và trực quan hóa dữ liệu một cách dễ dàng và nhanh chóng. Với Power BI, người dùng có thể tạo các báo cáo, biểu đồ và bảng điều khiển tùy chỉnh với độ chính xác cao và tính tương tác, giúp họ hiểu rõ hơn về dữ liệu và đưa ra quyết định kinh doanh chính xác hơn. Power BI cũng cung cấp tính năng tự động hoá và liên tục cập nhật dữ liệu, giúp người dùng luôn cập nhật và theo dõi các chỉ số hiệu suất của doanh nghiệp.

CHƯƠNG 3: PHÂN TÍCH CHI TIẾT VÀ XÂY DỰNG MÔ HÌNH

3.1 Nguồn dữ liệu

Prudential là một trong những doanh nghiệp bảo hiểm nhân thọ tiên phong tại Việt Nam với lịch sử hàng trăm năm, bảo hiểm Prudential mang tới những sản phẩm và dịch vụ bảo hiểm theo tiêu chuẩn hàng đầu thế giới. Tuy nhiên do dữ liệu nhạy cảm liên quan tới thông tin khách hàng nên trong phạm vi nghiên cứu này, dữ liệu từ Kaggle sẽ được khảo sát và xem xét là dữ liệu tại doanh nghiệp để thực nghiệm. Tập dữ liệu này được chia làm 3 nguồn chính:



Hình 3-1. Nguồn dữ liệu của công ty

3.1.1 Cơ sở dữ liệu quan hệ

Cơ sở dữ liệu quan hệ ghi nhận lại hoạt động bán hàng của công ty trên sàn thương mại điện tử tại Brazil về các đơn hàng. Nguồn dữ liệu bao gồm thông tin về hơn 100000 đơn đặt hàng từ năm 2016 đến năm 2018 bao gồm tình trạng đơn hàng, ngày đặt hàng, giá cả, chi phí vận chuyển, … Nguồn dữ liệu này bao gồm 6 bảng với nhiều thông tin liên quan tới phân hệ bán hàng. Cơ sở dữ liệu quan hệ bao gồm đầy đủ tất cả các cấu trúc bảng, tên cột, kiểu dữ liệu cột, ràng buộc, khóa chính, khóa ngoại và mối quan hệ giữa các bảng.

Hình 3-2. Mô hình ERD của cơ sở dữ liệu quan hệ

3.1.2 Hệ thống kế toán

Hệ thống kế toán là một phần rất quan trọng trong hoạt động của một công ty. Hệ thống được thiết kế để giúp công ty có thể quản lý và kiểm soát tài chính một cách chính xác và hiệu quả. Hệ thống kế toán của công ty bao gồm nhiều chức năng khác nhau, từ việc ghi nhận các giao dịch kinh tế đến việc lập báo cáo tài chính và phân tích chi phí và doanh thu.

Ngoài ra, hệ thống kế toán còn ghi nhận và quản lý các thông tin thanh toán của khách hàng về các đơn hàng đã được đặt trên sàn thương mại điện tử. Tuy nhiên, nếu cần lấy tập dữ liệu này thì cần xuất ra tệp theo định dạng csv theo định kỳ, chứ nhân viên không thể truy cập toàn quyền vào cơ sở dữ liệu kế toán để lấy được toàn bộ thông tin này.

3.1.3 Trang web thương mại điện tử

Với sự phát triển của công nghệ và internet, việc mua sắm trực tuyến đã trở nên phổ biến hơn bao giờ hết. Trang web thương mại điện tử cung cấp cho khách hàng một nơi để mua sắm và tìm kiếm thông tin về sản phẩm và dịch vụ một cách thuận tiện và nhanh chóng. Tuy nhiên, trang web thương mại điện tử cũng là một nguồn quý giá để thu thập dữ liệu về bình luận của khách hàng về sản phẩm và dịch vụ của công ty. Điều này giúp cho công ty có thể cải thiện sản phẩm và dịch vụ của mình để đáp ứng nhu cầu của khách hàng và tăng trưởng doanh thu. Dữ liệu bình luận khách hàng được thu thập bằng các công cụ tự động để cào dữ liệu và được lưu trữ dưới dạng tệp json.

3.2 Phân tích vấn đề

Phân tích hiệu suất bán hàng là một quá trình quan trọng cho công ty đánh giá được sự thành công của hoạt động kinh doanh của mình. Để giải đáp được vấn đề này, đầu tiên cần phải xác định các chỉ số tiêu chuẩn để đo lường hiệu suất bán hàng bao gồm doanh số bán hàng, lợi nhuận, số lượng sản phẩm bán ra, số lượng khách hàng mới, phân nhóm khách hàng, … Sau khi phân tích chi tiết các chỉ số này sẽ đem lại bức tranh tổng quan về quá trình hoạt động kinh doanh của công ty, từ đó chi phép công ty hướng tới đầu tư các nguồn lực cần thiết và đề xuất các giải pháp cho các khía cạnh hoạt động kém hiệu quả. Phân tích hiệu suất bán hàng là một quá trình liên tục cần được thực hiện thường xuyên để đảm bảo rằng công ty đang hoạt động hiểu quả và đáp ứng được nhu cầu của thị trường. Từ vấn đề trên, việc dữ liệu nằm rải rác nhiều nơi dẫn đến tình trạng tốn thời gian và công sức nhân công trong quá trình nhập và lưu trữ dữ liệu, gây ảnh hưởng đến việc đưa ra quyết định nhanh chóng và kịp thời.

Vì vậy, công ty cần phải xây dựng một giải pháp BI dựa trên nền tảng Cloud Azure bao gồm hồ dữ liệu (data lake) và kho dữ liệu (data warehouse) để tổng hợp dữ liệu từ nhiều nguồn dễ dàng và nhanh chóng giúp tối ưu hóa quy trình ra quyết định. Đồng thời, quá trình thực hiện cần phải đảm bảo tính sẵn sàng cao, dễ dàng quản lý, ổn định và bảo mật cho toàn bộ dữ liệu.

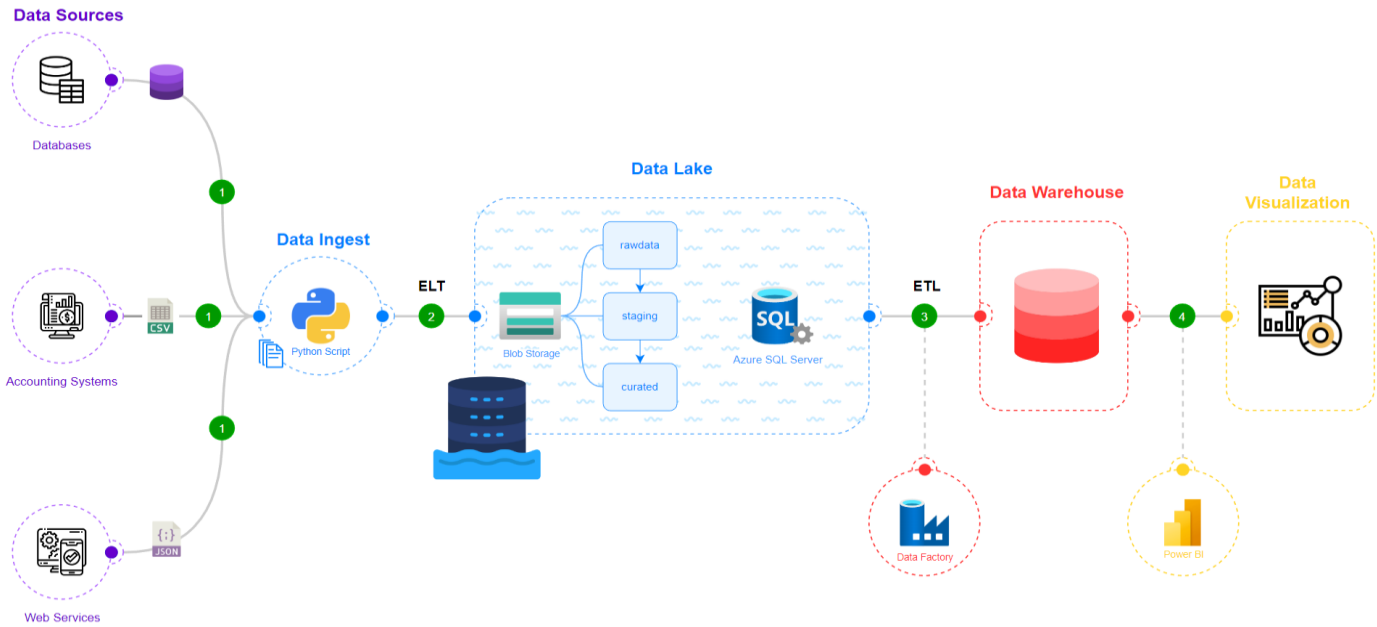
3.3 Yêu cầu kỹ thuật và hệ thống

Để đảm bảo dự án đạt được các mục tiêu và đáp ứng các yêu cầu kinh doanh, các yêu cầu về mặt kỹ thuật và hệ thống sau đây cần được tích hợp vào giải pháp BI:

- *Quy tắc đặt tên tệp*: là một yêu cầu kỹ thuật quan trọng để đảm bảo tính thuận tiện và dễ quản lý cho việc xử lý dữ liệu sau này. Tên tệp nên được mô tả rõ ràng nội dung về thông tin dữ liệu gì cộng với thời gian năm/tháng/ngày tệp được tạo ra để có thể dễ dàng mở rộng và thêm dữ liệu trong tương lai. Đồng thời, việc tuân thủ quy tắc đặt tên tệp đảm bảo tính duy tính, tránh trường hợp trùng lặp khi lưu trữ dữ liệu.
- *Xác định đầy đủ định dạng tệp*: là nhiệm vụ đầu tiên trước khi bắt quá quá trình trích xuất dữ liệu từ nhiều nguồn. Khi xác định đầy đủ định dạng tệp, các quy trình trích xuất, tải và chuyển đổi được thực hiện một cách nhất quán trên toàn bộ nguồn dữ liệu, từ đó đảm bảo tính nhất quán trong cấu trúc và định dạng của dữ liệu được lưu trữ trong hồ dữ liệu. Đồng thời, giảm thiểu thiếu sót tệp trong quá trình xử lý dữ liệu, tránh gây mất thời gian kiểm tra lại đã tải đầy đủ tệp hay chưa.
- *Ghi log*: là quá trình ghi lại các hoạt động và sự kiện diễn ra trong quá trình thực hiện các quy trình xử lý dữ liệu. Tầm quan trọng của việc ghi log trong quá trình xử lý dữ liệu là rất lớn và được coi là một trong những yếu tố quan trọng nhất để đảm bảo tính toàn vẹn và đáng tin cậy của dữ liệu trong Data Lake. Điều này giúp người quản trị có thể dễ dàng phát hiện lỗi phát sinh và sửa chữa nhanh chóng khi có sự cố xảy ra.
- *Tính chính xác*: đối với một doanh nghiệp, dữ liệu không chính xác dẫn đến những hậu quả đáng kể, gây lãng phí nguồn lực không đáng có. Do đó, quá trình ELT và ETL dữ liệu từ nhiều nguồn phải đảm bảo tính chính xác.
- *Trực quan hóa*: là một kỹ thuật để đảm bảo các bảng báo cáo và các biểu đồ được trình bày một cách rõ ràng và dễ hiểu, đáp ứng được các yêu cầu phân tích kinh doanh. Hơn nữa, các biểu đồ báo cáo cần được cập nhật tự động khi có thay đổi dữ liệu trong dữ liệu nguồn. Để đạt những mục tiêu này, các công cụ trực quan hóa như Power BI, Tableau, Qlik, ... cần được thiết kế dành riêng cho việc trực quan hóa và phân tích dữ liệu.

3.4 Đề xuất mô hình

Dưới đây là giải pháp BI dựa trên nền tảng Cloud Azure do nhóm đề xuất để giải quyết các vấn đề đặt ra:



Hình 3-3. Giải pháp BI (Nguồn: Tác giả đề xuất)

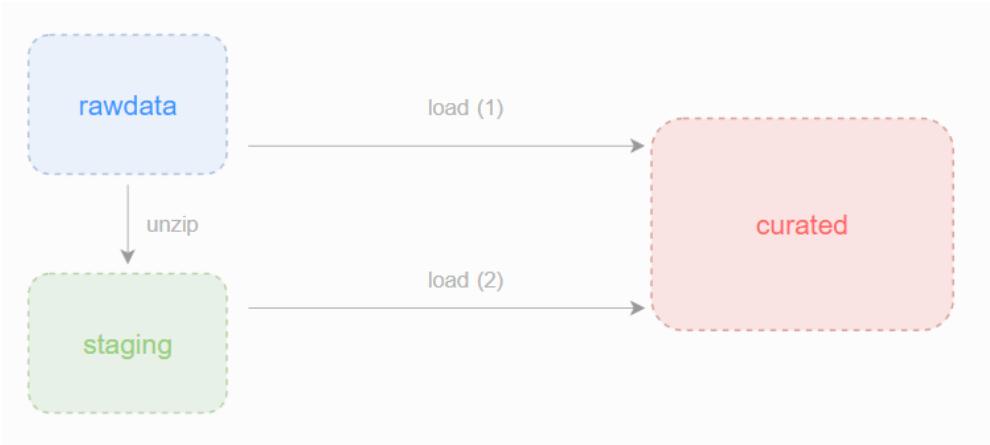
- *Bước 1:* Xác định các nguồn dữ liệu của doanh nghiệp và các loại định dạng tệp của từng nguồn
- *Bước 2:* Toàn bộ dữ liệu được nhập vào vùng chứa "rawdata" bằng đoạn mã Python, sau đó sẽ thực hiện quy trình ELT để tải toàn bộ dữ liệu vào vùng chứa "staging" bằng kỹ thuật ELT động. Những dữ liệu cần thiết cho mục đích kinh doanh được tải vào Azure SQL Server, đồng thời toàn bộ dữ liệu ở vùng chứa staging được đẩy vào vùng chứa "curated" để phục vụ mục đích lưu trữ.
- *Bước 3:* Dữ liệu cần phân tích ở Azure SQL Server được trích xuất, chuyển đổi và tải vào kho dữ liệu đã được thiết kế theo mô hình lược đồ sao từ trước bằng Data Factory.
- *Bước 4:* Dữ liệu từ kho dữ liệu sẽ kết nối với Power BI với mục tích tạo các trang báo cáo phân tích kinh doanh để hỗ trợ quá trình ra quyết định chính xác và nhanh chóng hơn.

CHƯƠNG 4: THỰC NGHIỆM

4.1 Xây dựng hồ dữ liệu (Data Lake)

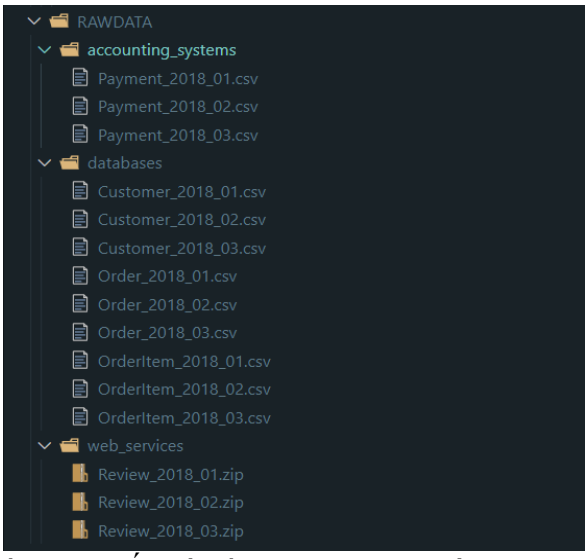
4.1.1 Vùng dữ liệu

Mặc dù hồ dữ liệu cho phép lưu trữ các dữ liệu dưới dạng thô nhưng một hệ thống trong hồ dữ liệu hiệu quả cần được thiết kế thành nhiều vùng khác nhau với thiết kế mức cao để đáp ứng các mục đích sử một của một hồ dữ liệu. Công dụng sử dụng để tạo thiết kế vùng chứa dựa trên nền tảng Cloud Azure là Blog Storage. Hình dưới đây để hình dung được các vùng trong một hồ dữ liệu mà nhóm đề xuất:



Hình 4-1. Vùng chứa dữ liệu bên trong hồ dữ liệu

Vùng dữ liệu thô (rawdata): vùng này được sử dụng để lưu trữ cho các dữ liệu mới được thu thập với các định dạng tự nhiên của chúng và từ nhiều nguồn khác nhau. Đây cũng là một bản sao chính xác của dữ liệu từ các nguồn và được sắp thêm theo thư mục có tổ chức.



Hình 4-2. Cấu trúc vùng chứa dữ liệu thô (rawdata)

Vùng dữ liệu tạm thời (staging): vùng này được sử dụng để giải nén tất cả các tệp được nén thành tệp zip từ các bình luận khách hàng được thu thập trên website thương mại điện tử để chuẩn bị cho quá trình nhập dữ liệu vào vùng dữ liệu được sắp xếp (curated).

Hình 4-3. Cấu trúc vùng dữ liệu được sắp xếp (curated)

Vùng dữ liệu được sắp xếp (curated): vùng này là vùng cho các dữ liệu đã được làm sạch, được chuyển đổi thành cần thiết, được sắp xếp theo cấu trúc thư mục với các quy tắc từ nguồn nào, phòng ban nào, năm nào, tháng nào,… giúp tối ưu trong việc phân phối dữ liệu.

Phân biệt hai luồng nhập dữ liệu vào vùng dữ liệu curated *(hình 4-1):*

- *Luồng (1):* đối với các tệp dữ liệu theo định dạng csv và json được tiến hành phân chia theo cấu trúc và đẩy vào vùng chứa curated.
- *Luồng (2):* đối với các tệp dữ liệu được nén dưới dạng zip được nhập vào vùng dữ liệu staging trước để thực hiện việc giải nén tất cả tập tin nằm bên trong, sau đó mới được đẩy vào vùng dữ liệu curated.

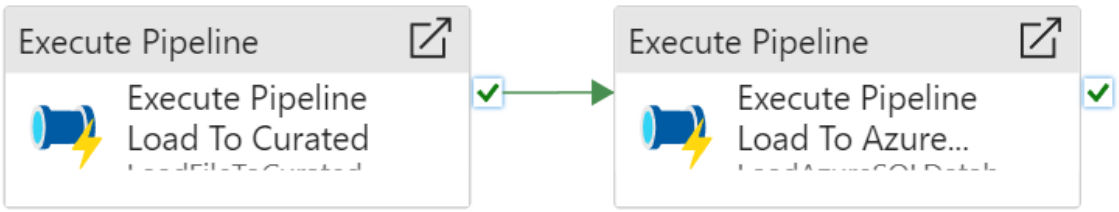
Mỗi luồng dữ liệu đều được ghi log lại để dễ dàng theo dõi, kiểm tra và sửa chửa khi có sự cố trong việc di dời dữ liệu diễn ra.

4.1.2 Azure SQL Server bên trong hồ dữ liệu

Song song với việc lưu trữ dữ liệu trong các vùng chứa được thiết kế trong Blob Storage, việc tạo ra thêm một cơ sở dữ liệu Azure SQL Server kèm theo được nhóm đề xuất giúp giảm thiểu thời gian truy xuất lại toàn bộ vùng chứa curated những dữ liệu cần được phân tích ngay lập tức. Vì các dữ liệu cần dùng để giải quyết yêu cầu kinh doanh đã được di dời sang cơ sở dữ liệu sẵn. Từ đó, kho dữ liệu chỉ cần tải dữ liệu từ nơi này sang để tạo báo cáo, trả lời các câu hỏi kinh doanh nhanh chóng hơn. Đồng thời lưu trữ hai bảng dữ liệu dbo.ELT_Run và dbo.ELT_ImportFiles để ghi nhận lại các sự kiện diễn ra của quá trình ELT động.

4.1.3 Quy trình ELT động

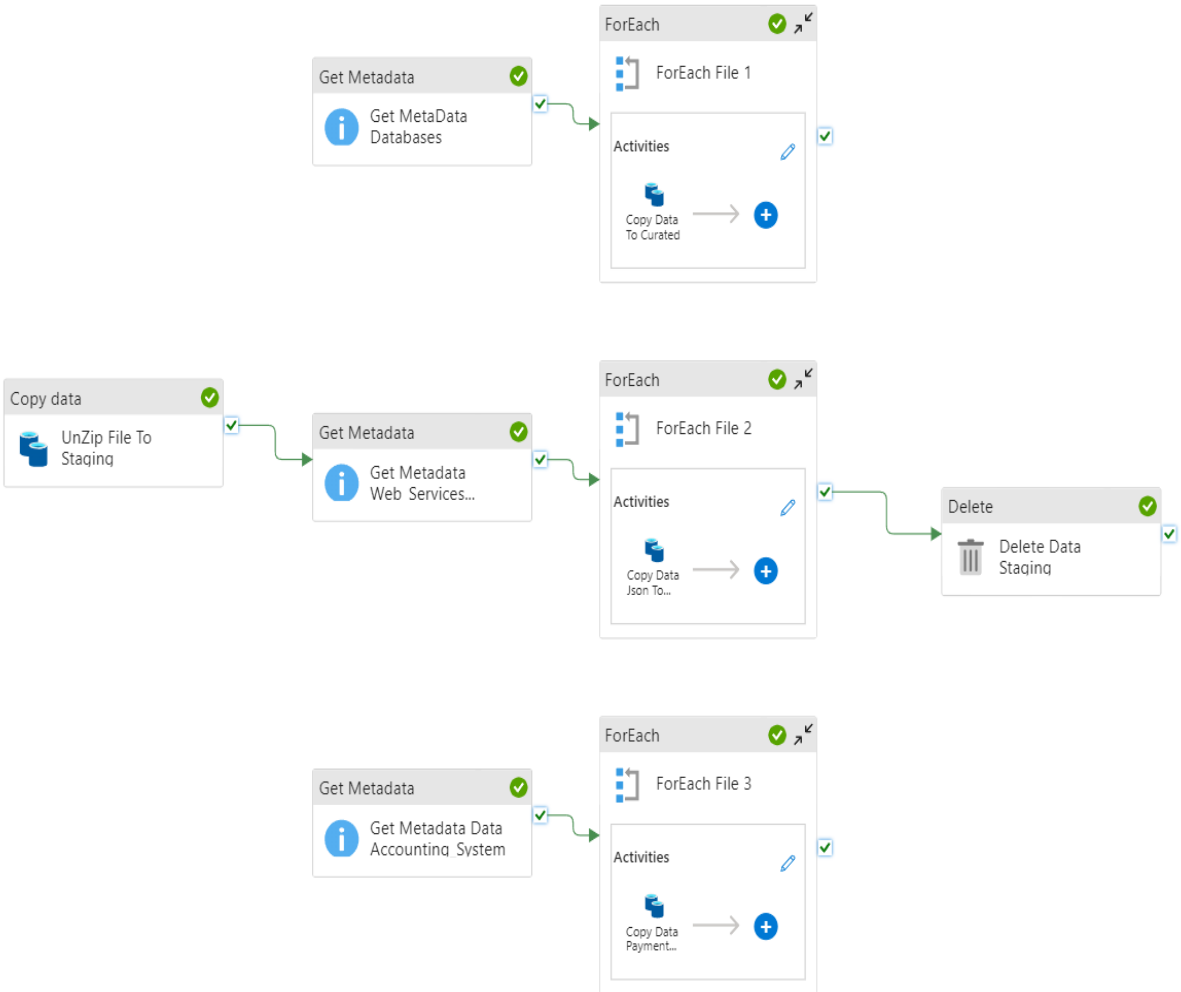
Quy trình ELT động là quá trình trích xuất dữ liệu thô và tải vào các vùng chứa dữ liệu một cách chính xác theo cấu trúc đã định sẵn thông qua việc điều chỉnh các tham số đầu. Quy trình này giải quyết vấn đề các lượng dữ liệu đổ vào các hệ thống ngày càng tăng cộng sự đa dạng định dạng tệp dẫn đến dữ liệu khó quản lý và tốn nhiều thời gian hơn. Hình 4-4. dưới đây mô tả đường ống dữ liệu tổng thể quả quy trình ELT động:



Hình 4-4. Đường ống dữ liệu tổng thể quy trình ELT động

Quá trình thực hiện được chia làm 2 giai đoạn nhỏ là:

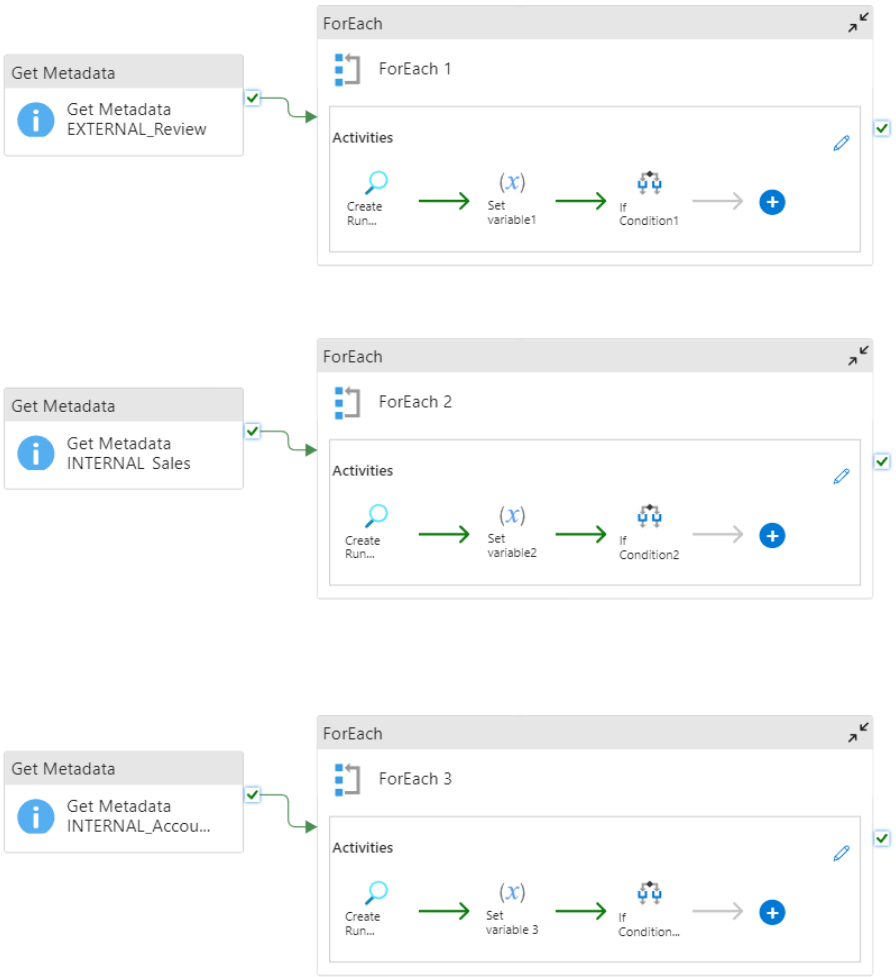
- *Giải đoạn 1:* tải dữ liệu vào vùng chứa curated.
- *Giai đoạn 2:* tải dữ liệu phục vụ phân tích nghiệp vụ kinh doanh vào Azure SQL Server.



Hình 4-5. Tổng quan quá trình tải dữ liệu vào vùng chứa curated

Mỗi đường ống dữ liệu nhỏ trên đại diện cho một nguồn dữ liệu hiện tại của doanh nghiệp đang được tải vào hồ dữ liệu. Ở giai đoạn này, nhóm thực hiện đẩy dữ liệu từ vùng chứa rawdata với đích đến cuối cùng là vùng chứa curated.

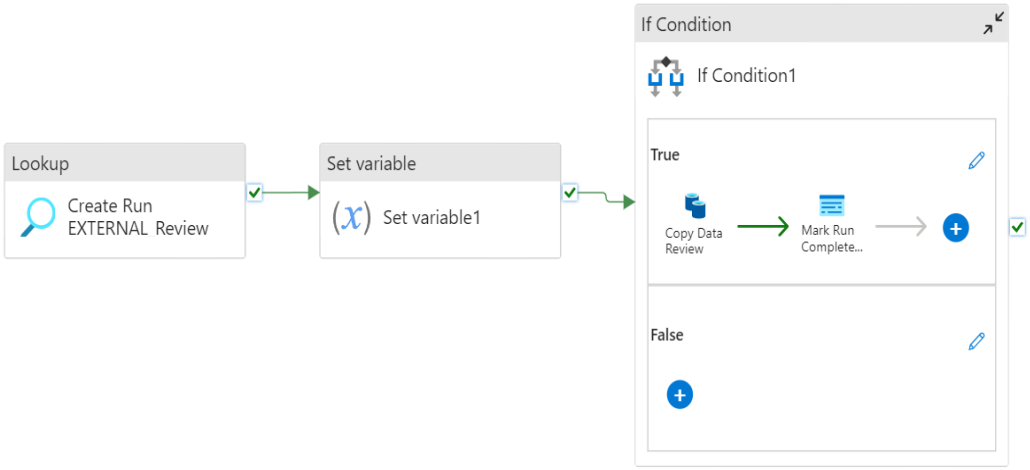
Ở nguồn dữ liệu được thu thập từ website thương mại điện tử thì dữ liệu được định dạng json nhưng được nén vào tệp zip để giảm bớt dung lượng đẩy dữ liệu từ nhiều nguồn lên Blob Storage. Các tệp được di chuyển sang vùng chứa staging để xử lí, đồng thời giải nén toàn bộ các tệp trước khi đưa vào vùng chứa curated. Sau khi quá trình chạy hoàn tất, vùng chứa staging được xóa toàn bộ dữ liệu tại đây và ghi log vào tập tin csv nằm trong vùng chứa logs. Điều này giúp người quản lý dễ dàng kiểm soát các sự kiện diễn ra trong toàn bộ quá trình nhập dữ liệu.



Hình 4-6. Tổng quan quá trình tải dữ liệu vào Azure SQL Server

Tương tự quá trình trên, mỗi đường ống dữ liệu đại diện cho mỗi nguồn dữ liệu đã được phân bổ vào vùng chứa curated với cấu trúc thư mục hợp lý. Sau khi truy xuất siêu

dữ liệu của mỗi đường ống dữ liệu sẽ được đưa vào vòng lặp thực hiện các bước: nhập tên tập tin, kiểm tra tên tập tin có tồn tại trong hệ thống chưa, tải dữ liệu vào Azure SQL Server, ghi nhập quá trình ELT thực hiện thành công.



Hình 4-7. Quá trình kiểm tra và nhập dữ liệu vào Azure SQL Server

Mỗi bước thực hiện trong quá trình trên đều gọi đến thủ tục đã được cấu hình sẵn trong hệ thống Azure SQL Server để ghi nhận lại các tập tin được đẩy vào và quá trình thực hiện ELT hoạt động như thế nào, gặp lỗi ra sao nhằm mục đích giúp người quản trị hệ thống dễ dàng kiểm tra và sửa chữa ghi có sự cố.

4.2 Xây dựng nhà kho dữ liệu (Data warehouse)

4.2.1 Bus matrix

Bus matrix là một công cụ quản lý dữ liệu kinh doanh trong việc xây dựng kiến trúc dữ liệu cho hệ thống data warehouse hoặc các dự án Business Intelligence (BI).

Bus matrix giúp xác định các quy trình kinh doanh và các chiều dữ liệu liên quan đến chúng. Nó bao gồm một bảng hai chiều, trong đó các hàng đại diện cho các quy trình kinh doanh và các cột đại diện cho các chiều dữ liệu. Các ô trong bảng được sử dụng để chỉ ra mối quan hệ giữa các quy trình kinh doanh và các chiều dữ liệu. Bus matrix giúp cho các nhóm phát triển BI có thể đồng bộ hoá kiến trúc dữ liệu và đảm bảo tính nhất quán của các bộ dữ liệu. Nó cũng được sử dụng để ưu tiên các dự án phát triển dữ liệu dựa trên nhu cầu kinh doanh ưu tiên và đảm bảo tính khả thi của các dự án BI.

Bảng 4-1. Bus Matrix

	Customer	Seller	Product	Time
Sales performance tracking	x	x	x	x
Orders	x	x	x	x
Customer segmentations	x			x
Seller distribution		x		

4.2.2 Master data

Master data (dữ liệu chính): là các dữ liệu cơ bản, không thay đổi hoặc thay đổi rất ít trong quá trình hoạt động của doanh nghiệp. Đây là các dữ liệu cần thiết để định nghĩa và quản lý các thực thể như khách hàng, sản phẩm, nhà cung cấp, vật liệu, v.v. Dữ liệu chính thường được quản lý trong các hệ thống quản lý dữ liệu doanh nghiệp (ERP) hoặc các hệ thống quản lý dữ liệu khác. Trong phạm vi dự án này, các Master data bao gồm:

Bảng 4-2. Master data

Customer	Mô tả thông tin về thành phố, tiểu bang, id,... của khách hàng.
Seller	Includes data about the sellers that fulfilled orders. Sử dụng nó để tìm vị trí của người bán và để xác định người bán nào đã hoàn thành bán từng sản phẩm
Product	Thông tin về sản phẩm của công ty như số ký tự trích xuất từ mô tả sản phẩm, chiều dài, trọng lượng, chiều rộng của sản phẩm, ...

4.2.3 Transaction data

Là các dữ liệu thu thập được sau khi xảy ra giao dịch như bán hàng, giao hàng, ... Dữ liệu bao gồm thời gian giao dịch, địa điểm giao dịch, giá của những thứ đã mua, phương thức thanh toán được sử dụng, bất kỳ khoản giảm giá nào cũng như số lượng và chất lượng khác liên quan đến giao dịch.

Bảng 4-3. Transaction data

Dữ liệu giao dịch đặt hàng	Mô tả thông tin về các mặt hàng được mua trong mỗi đơn hàng, giá sản phẩm, số lượng, ngày hạn vận chuyển, giá trị cước, tình trạng...

4.2.4 ETL mapping

Bảng 4-4. ETL mapping

Kho dữ liệu			Nguồn dữ liệu				
Thuộc tính	**Kiểu dữ liệu**	**isNULL**	**Bảng**	**Thuộc tính**	**Kiểu dữ liệu**	**isNULL**	**Quy tắc**
DimCustomer							
customer_key	int				int		Tự động
customer_id	varchar(32)		quantad_customers	customer_id	varchar(50)		Từ nguồn
customer_city	varchar(32)	x	quantad_customers	customer_city	nvarchar(50)	x	Từ nguồn
customer_state	varchar(2)	x	quantad_customers	customer_state	nvarchar(50)	x	Từ nguồn
DimProduct							
product_key	int				int		Tự động
product_id	varchar(50)		quantad_products	product_id	varchar(50)		Từ nguồn
product_category_name	varchar(50)	x	quantad_products	product_category_name	nvarchar(50)	x	Từ nguồn
product_weight_g	float	x	quantad_products	product_weight_g	nvarchar(50)	x	Từ nguồn
product_length_cm	float	x	quantad_products	product_length_cm	int	x	Từ nguồn
product_height_cm	float	x	quantad_products	product_height_cm	int	x	Từ nguồn
product_width_cm	float	x	quantad_products	product_width_cm	int	x	Từ nguồn

DimSeller							
seller_key	Int						Tự động
seller_id	varchar(50)		quantad_sellers	seller_id	nvarchar(50)		Từ nguồn
seller_city	varchar(50)	x	quantad_sellers	seller_city	nvarchar(50)	x	Từ nguồn
seller_state	varchar(50)	x	quantad_sellers	seller_state	nvarchar(50)	x	Từ nguồn
DimDate							
DateKey	int						Tự động
Date	date	x					Tự động
FullDate	char(10)	x					Tự động
DayOfMonth	varchar(2)	x					Tự động
DayName	varchar(9)	x					Tự động
DayOfWeek	char(1)	x					Tự động
Month	varchar(2)	x					Tự động
Quarter	char(1)	x					Tự động
Year	char(4)	x					Tự động

MonthYear	char(10)	x					Tự động
MMYYYY	char(6)	x					Tự động
FirstDayOfMonth	date	x					Tự động
LastDayOfMonth	date	x					Tự động
FirstDayOfQuarter	date	x					Tự động
LastDayOfQuarter	Date	x					Tự động
FirstDayOfYear	date	x					Tự động
LastDayOfYear	date	x					Tự động
FactSales							
seller_key	int		DimSeller	seller_key	int		Truy xuất từ [DimSeller].[seller_key]
product_key	int		DimProduct	product_key	int		Truy xuất từ [DimProduct].[product_key]
customer_key	int		DimCustomer	customer_key	int		Truy xuất từ [DimCustomer].[customer_key]
DateKey	int		DimDate	DateKey	int		Truy xuất từ [DimDate].[DateKey]
order_id	varchar(50)		quantad_orders	order_id	varchar(50)		Từ nguồn
order_item_id	varchar(50)		quantad_order_items	order_item_id	varchar(50)		Từ nguồn

order_status	varchar(32)	x	quantad_orders	order_status	varchar(32)	x	Từ nguồn
order_purchase_timestamp	date	x	quantad_orders	order_purchase_timestamp	Datetime	x	Từ nguồn
order_approved_at	date	x	quantad_orders	order_approved_at	Datetime	x	Từ nguồn
order_delivered_carrier_date	date	x	quantad_orders	order_delivered_carrier_date	Datetime	x	Từ nguồn
order_delivered_customer_date	date	x	quantad_orders	order_delivered_customer_date	Datetime	x	Từ nguồn
order_estimated_delivery_date	float	x	quantad_orders	order_estimated_delivery_date	Datetime	x	Từ nguồn
price	float		quantad_orders_items	price	float		Từ nguồn
freight_value	numeric(8, 2)		quantad_orders_ items	freight_value	numeric(8, 2)		Từ nguồn

4.2.5 Bảng Fact và Dimension

Bảng FactSales: lưu trữ các tính năng được sử dụng để phân tích, bao gồm khóa ngoại được tham chiếu đến từng bảng Dim và đo lường các thuộc tính của từng giao dịch.

Column Name	Data Type	Allow Nulls
seller_key	int	☑
product_key	int	☑
customer_key	int	☑
DateKey	int	☑
order_id	varchar(50)	☐
order_item_id	varchar(50)	☐
order_status	varchar(32)	☑
order_purchase_timestamp	date	☑
order_approved_at	date	☑
order_delivered_carrier_date	date	☑
order_delivered_customer_date	date	☑
order_estimated_delivery_date	date	☑
price	float	☐
freight_value	numeric(8, 2)	☐

Hình 4-1. Bảng FactSale

Bảng DimSeller: lưu trữ các tính năng, bao gồm khóa chính và các thuộc tính về vị trí của người bán

Column Name	Data Type	Allow Nulls
seller_key	int	☐
seller_id	varchar(50)	☐
seller_city	nvarchar(50)	☑
seller_state	nvarchar(50)	☑

Hình 4-2. Bảng DimSeller

Bảng DimProduct: lưu trữ các tính năng bao gồm khóa chính, khóa tự nhiên và các thuộc tính chi tiết của sản phẩm

Column Name	Data Type	Allow Nulls
product_key	int	☐
product_id	varchar(50)	☐
product_category_name	varchar(50)	☑
product_weight_g	float	☑
product_length_cm	float	☑
product_height_cm	float	☑
product_width_cm	float	☑

Hình 4-3. Bảng DimProduct

Bảng DimTime: lưu trữ các đặc trưng, bao gồm khóa chính và các thuộc tính thông tin về thời gian

Column Name	Data Type	Allow Nulls
DateKey	int	☐
Date	date	☑
FullDate	char(10)	☑
DayOfMonth	varchar(2)	☑
DayName	varchar(9)	☑
DayOfWeek	char(1)	☑
Month	varchar(2)	☑
Quarter	char(1)	☑
Year	char(4)	☑
MonthYear	char(10)	☑
MMYYYY	char(6)	☑
FirstDayOfMonth	date	☑
LastDayOfMonth	date	☑
FirstDayOfQuarter	date	☑
LastDayOfQuarter	date	☑
FirstDayOfYear	date	☑
LastDayOfYear	date	☑

Hình 4-4. Bảng DimTime

DimCustomer: lưu trữ các thuộc tính, bao gồm khóa chính, khóa tự nhiên và các thuộc tính về thông tin của khách hàng

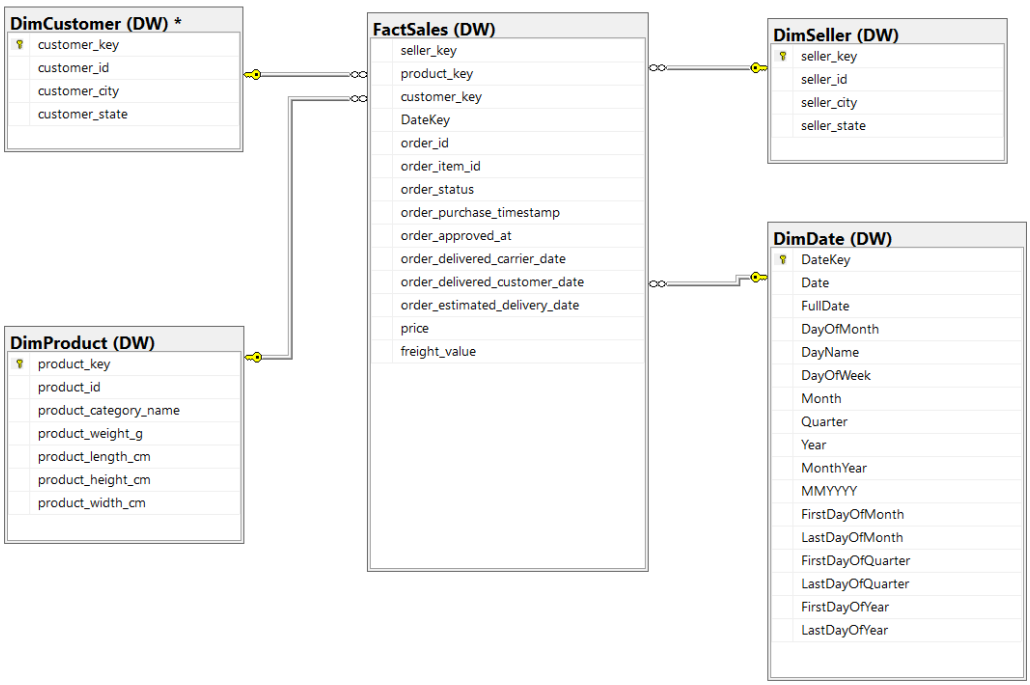
Column Name	Data Type	Allow Nulls
customer_key	int	☐
customer_id	varchar(32)	☐
customer_city	varchar(32)	☑
customer_state	varchar(2)	☑

Hình 4-5. Bảng DimCustomer

4.2.6 Mô hình nhà kho dữ liệu (Data warehouse)

Quá trình xây dựng các lược đồ dựa trên thông tin đầy đủ do khách hàng/chủ sở hữu doanh nghiệp cung cấp trong Mô hình hóa kho dữ liệu, đây là bước đầu tiên trong việc xây dựng hệ thống Kho dữ liệu.

Mô hình Star Schema cũng cho phép các nhà thiết kế dữ liệu có thể dễ dàng thêm hoặc bớt các bảng kích thước mà không làm ảnh hưởng đến cấu trúc tổng thể của hệ thống. Do đó, mô hình Star Schema là một trong những mô hình thiết kế cơ sở dữ liệu phổ biến nhất cho các hệ thống Data Warehouse và mang lại nhiều lợi ích cho các doanh nghiệp trong việc quản lý và phân tích dữ liệu.



Hình 4-6. Data Warehouse Star Schema

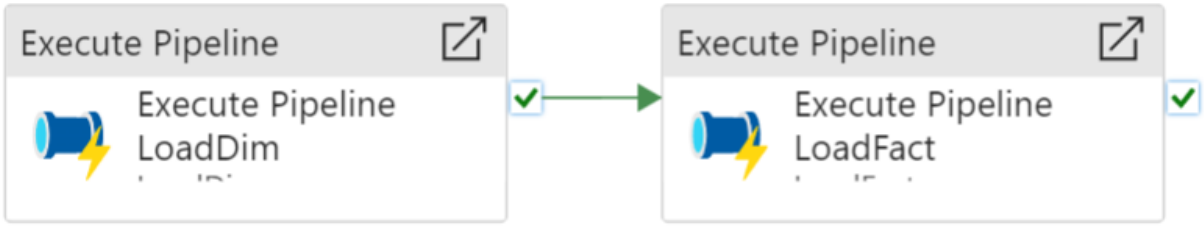
4.2.7 Quy trình ETL

Một trong những giai đoạn quan trọng nhất trong xử lý dữ liệu là ETL. ETL là viết tắt của Extract, Transform, and Load. Đó là một quy trình trích xuất dữ liệu từ một số hệ thống nguồn, biến đổi dữ liệu và sau đó chèn dữ liệu đó vào hệ thống Kho dữ liệu. Hầu hết các thuộc tính trong bảng đều được lấy từ nguồn dữ liệu.

Trong dự án này, sau khi quá trình load dữ liệu vào Data Lake hoàn tất và các tập tin dữ liệu và cơ sở dữ liệu được cập nhật. Quá trình ETL sẽ được tiến hành nhằm

chuyển đổi và lưu trữ dữ liệu vào Data warehouse phục vụ cho quá trình phân tích. Tổng quan quy trình như sau:

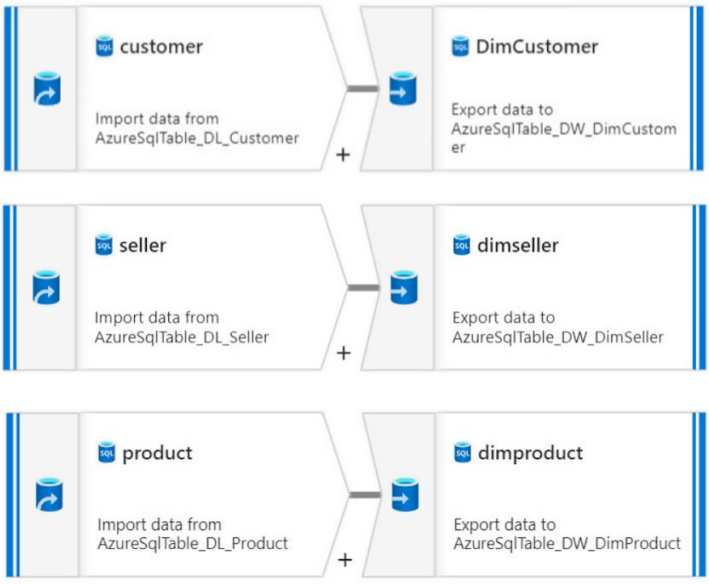
1. *Sử dụng Pipeline tổng quát (tùy chọn)*



Hình 4-11. Pipeline tổng quát thực hiện trình tự 2 Pipeline "LoadDim" và "LoadFact"

Một đường ống ETL tổng quát là một công cụ quan trọng trong việc tổng hợp các đường ống ETL. Nó giúp tối ưu hóa quản lý và bảo trì, giảm thiểu lỗi và xung đột, đảm bảo tính nhất quán và đồng bộ của dữ liệu, tăng tính linh hoạt và mở rộng, và tối ưu hóa hiệu suất và tốc độ xử lý. Do đó, trong dự án này một đường ống tổng quát được sử dụng nhằm thực hiện tuần tự việc tải dữ liệu từ nguồn vào các bảng Dimensions và bảng FactSale

2. *Trích xuất từ Data Lake chuyển đổi và tải dữ liệu vào các bảng Dimension*



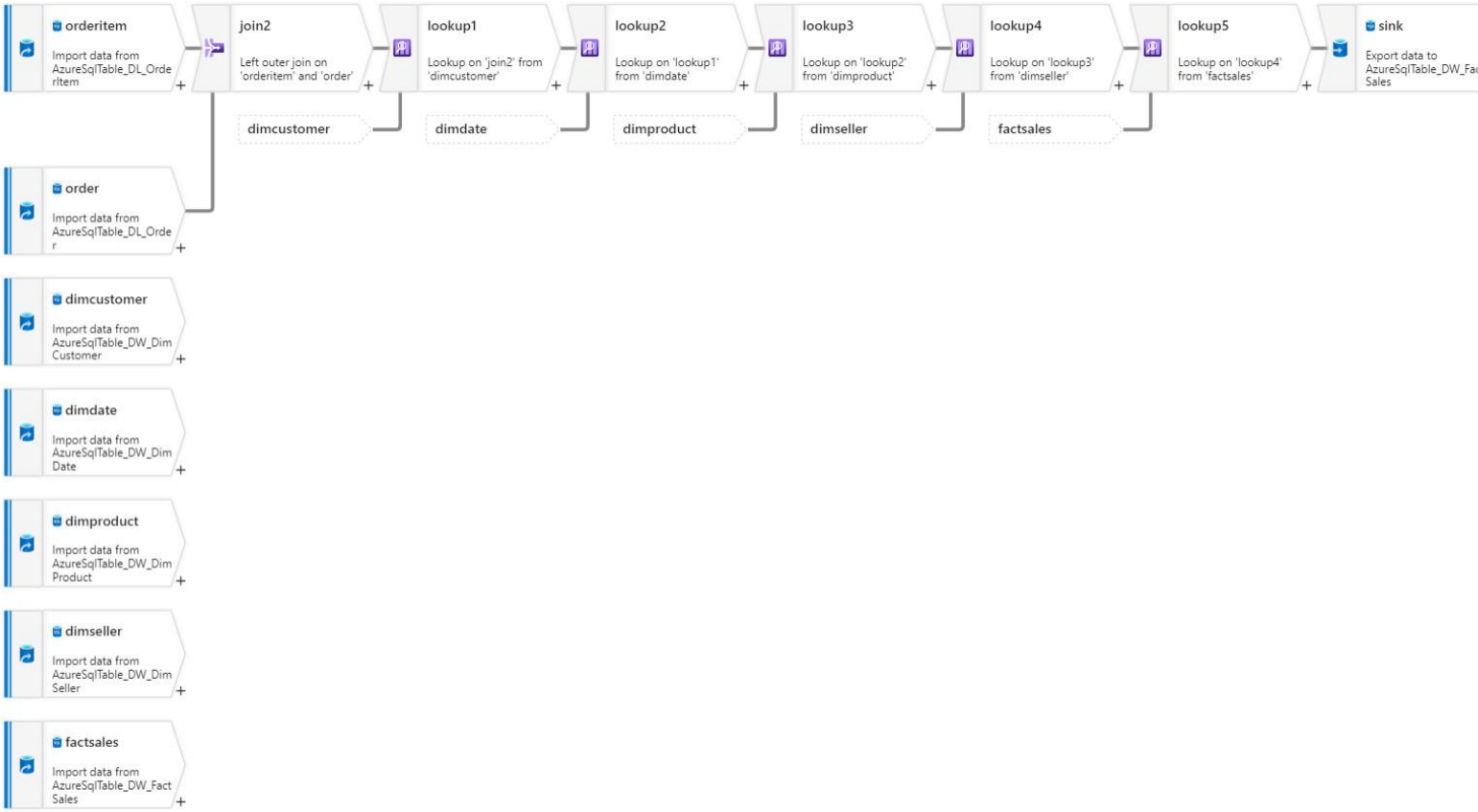
Hình 4-7. Data flow "Src2Dim" thực hiện ETL dữ liệu vào các bảng Dimension

Để thực hiện ETL dữ liệu từ nguồn vào các bảng Dimensions tính năng Data Flow được sử dụng. Dữ liệu được trích xuất từ nguồn sau đó ánh xạ đến các cột đã của từng bảng Dim một cách chính xác.

Hình 4-8. Pipeline "Pipeline LoadDim" thực thi các Data flow "Src2Dim"

Một đường ống thực thi các tính năng Data Flow tải dữ liệu từ nguồn vào các bảng Dimensions đã tạo trước đó.

3. *Tải dữ liệu từ các bảng Dimensions vào bảng FactSale*



Hình 4-9. Data flow "Src2FactSales" thực hiện ETL dữ liệu vào bảng FactSales

Sau khi thực thi hoàn tất quá trình ETL dữ liệu từ nguồn đến các bảng Dimension sẽ đến ETL dữ liệu vào bảng FactSale. Đầu tiên, thực hiện lệnh join hai bảng chứa dữ liệu giao dịch (order, order_item) nhằm lấy được tất cả thuộc tính có liên quan đến một đơn hàng. Tiếp theo đó, lần lượt thực hiện Lookup với các Dimension để được ánh xạ với dữ liệu trong bảng fact trong trường hợp này là các khóa thay thế "Surrogate Key". Cuối cùng là ánh xạ các dữ liệu tương ứng các cột trong bảng fact và thực hiện tải dữ liệu.

Hình 4-10. Pipeline "Pipeline LoadFact" thực thi Data flow "Src2FactSales"

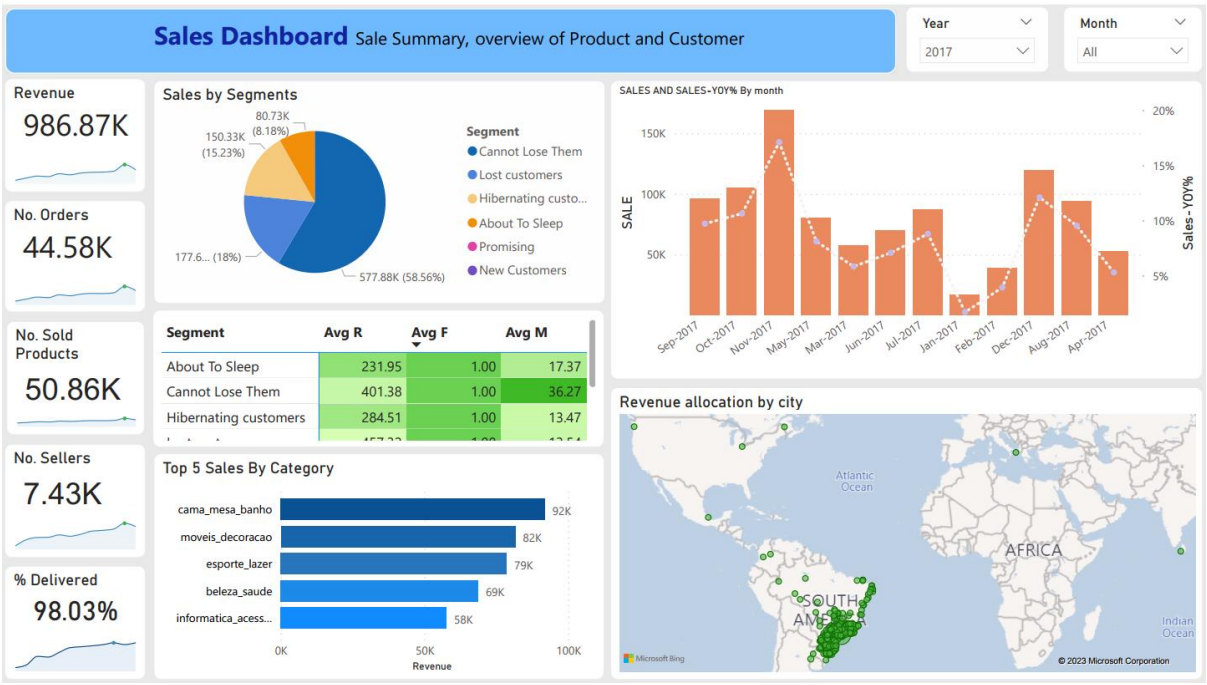
Cần sử dụng một đường ống thực thi Data Flow quá trình ETL dữ liệu vào bảng FactSales.

CHƯƠNG 5 PHÂN TÍCH DỮ LIỆU - TRỰC QUAN HÓA

5.1 Dashboard là gì?

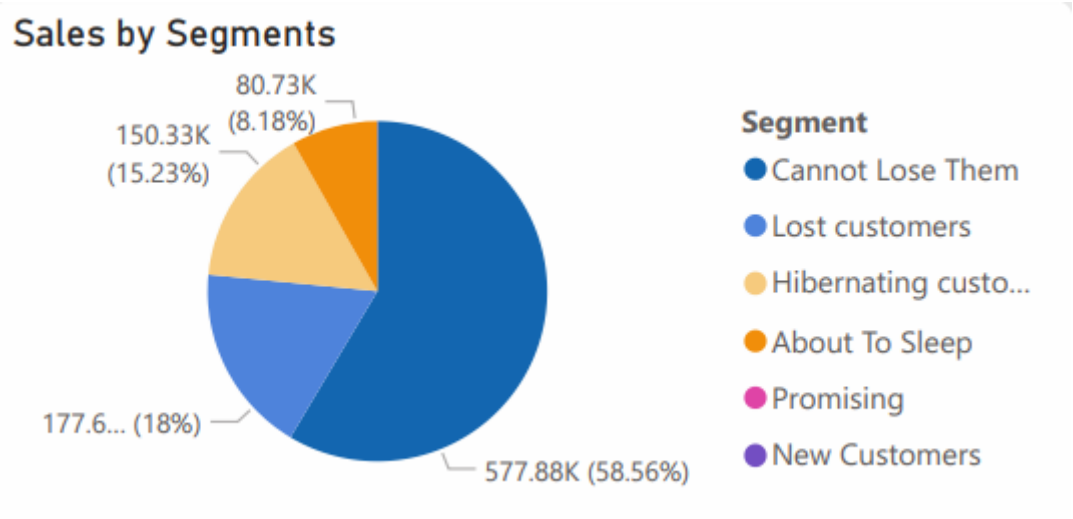
Dashboard tổng quan tình hình kinh doanh là một công cụ trực quan giúp người dùng xem phân tích các chỉ số kinh doanh quan trọng của công ty. Dashboard bao gồm các thông tin về doanh số bán hàng, chi phí, phân khúc khách khách hàng, sản phẩm, v.v và các chỉ số tài chính khác.

5.2 Dashboard đề xuất



Hình 5-1. Sales Dashboard

Với dữ liệu hiện có kết hợp các lý thuyết quan trọng về chỉ số kinh doanh, các KPIs được đề xuất trong Dashborad là doanh thu, số lượng đơn hàng, số lượng khách hàng và phần trăm giao hành thành công. Ngoài ra để có cái nhìn chi tiết hơn về tệp khách hàng, mô hình RFM đã được sử dụng để phân khúc khách hàng.



Hình 5-3. Phân khúc khách hàng sử dụng mô hình RFM

Khách hàng của mỗi nhóm được chia thành dựa trên Recency, Frequency và Monetary Score và các quy tắc kinh doanh sau đây:

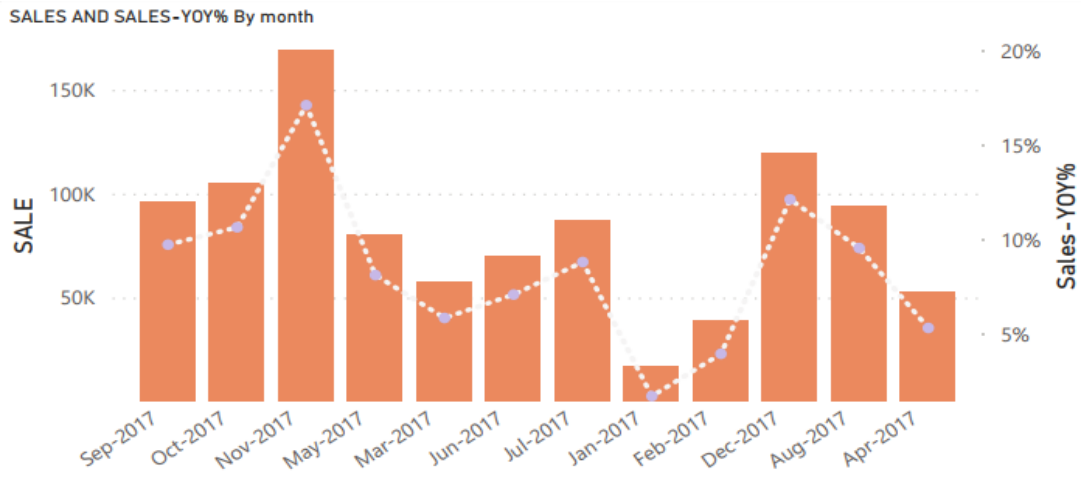
Bảng 5-1. Mô tả phân khách hàng dựa vào đặc tính mua hàng

Phân khúc	**Đặc tính**
Champion	Vừa mới mua, thường xuyên đặt hàng và chi tiêu nhiều nhất.
Loyal	Thường xuyên đặt hàng. Phản hồi tích cực với các chương trình khuyến mãi.
Potential Loyalist	Khách hàng mới, và đã chi tiêu một số tiền lớn.
New Customers	Mua lần đầu và gần đây.
Promising	Có tiềm năng trở thành khách hàng trung thành vài tháng trước. Thường xuyên chi tiêu một số tiền lớn. Nhưng lần mua hàng cuối cùng là vài tuần trước đây.
Core	Khách hàng tiêu chuẩn với lần mua hàng trước không quá lâu
Needs attention	Khách hàng cốt lõi, lần mua hàng cuối cùng đã xảy ra hơn một tháng trước
Can’t lose them but losing	Đã đặt hàng lớn nhất và thường xuyên. Nhưng đã lâu rồi không quay lại mua hàng
At Risk	Không thể bỏ lỡ họ nhưng đang mất dần dần với giá trị đơn hàng và tần suất thấp hơn
Losing but engaged	Đã mua hàng lần cuối cách đây một thời gian dài nhưng trong 4 tuần qua đã truy cập trang web hoặc mở xem email quảng cáo.

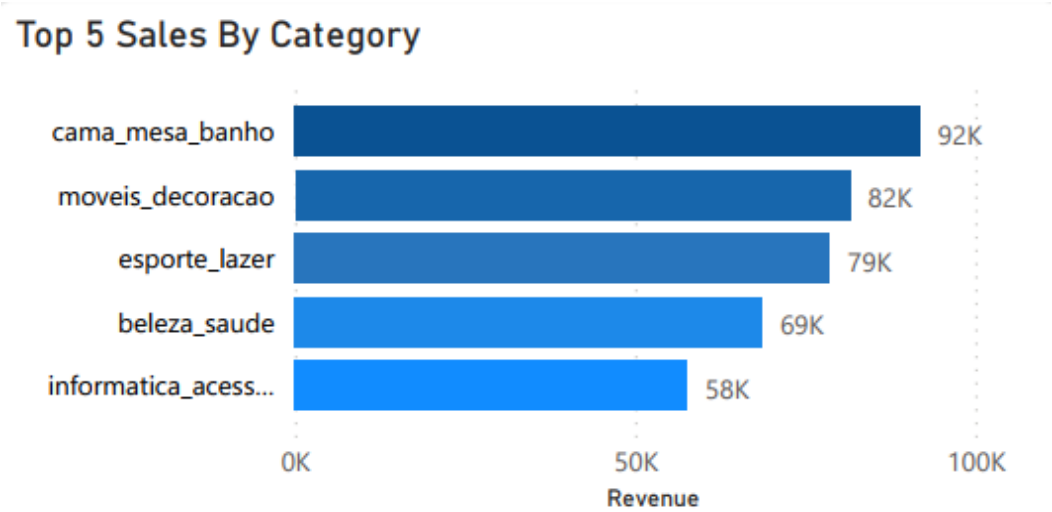
Lost	Đã mua hàng lần cuối cách đây một thời gian dài và không có hoạt động nào trong 4 tuần qua

Nhìn vào biểu đồ ta có thể thấy lượng khách hàng "không thể mất" chiếm hơn 50% vào năm 2017. Điều đó cho thấy tình hình kinh doanh và các chiến lược tiếp thị chưa đạt hiệu quả trong năm này.

Để có thể đánh giá chi tiết doanh thu của các danh mục sản phẩm qua thời gian và địa điểm 3 biểu đồ được đề xuất bên dưới. Dựa vào đó các nhà quản trị có thể tìm hiểu nguyên nhân có sự khác biệt đồng thời ra quyết định giải quyết vấn đề hoặc phát huy điểm mạnh.



Hình 5-4. Biểu đồ cột và đường thể hiện doanh số và phần trăm doanh số theo tháng



Hình 5-5. Biểu đồ thanh thể hiện top 5 danh mục sản phẩm có doanh thu cao nhất

Hình 5-6. Biểu đồ phân bổ doanh thu theo địa lý

5.3 Hàm ý quản trị

Với bộ dữ liệu thực nghiệm đại diện cho doanh nghiệp và dựa vào kết quả từ quá trình áp dụng mô hình nhóm rút ra một số đề xuất chiến lược như sau:

- Tập trung thu hút các khách hàng cũ nhiều hơn, xây dựng lòng trung thành từ khách hàng. Có thể sử dụng các phương pháp như cung cấp các ưu đãi, khuyến mãi đồng thời cập nhật thông tin về các sản phẩm mới đến cho khách hàng.
- Khu vực bán được hàng đang co cụm trong một vài thành phố dẫn đến bão hòa trong tương lai gần, do đó cần thực hiện phân bổ nguồn lực đến đa dạng các thành phố, khu vực hơn nhằm chiếm lĩnh thị phần.
- Cần nắm rõ nguyên nhân dẫn đến tình trạng doanh thu giảm sâu hai tháng đầu năm.

CHƯƠNG 6: KẾT LUẬN

6.1 Kết luận cho đề tài

Với đề tài "Xây dựng giải pháp Business Intelligence trên nền tảng đám mấy Microsoft Azure kết hợp quy trình ELT động" nhóm đã xây dựng giải pháp BI dựa trên nền tảng Cloud Azure cụ thể, tối ưu và mô hình luồng dữ liệu cho hệ thống bao gồm:

- Quy trình tổng hợp dữ liệu từ nhiều nguồn với nhiều định dạng tệp khác nhau.
- Xây dựng hồ dữ liệu (data lake) với mục đích lưu trữ toàn bộ dữ liệu mà doanh nghiệp hiện có.
- Quy trình ELT động thực hiện trích xuất, tải và chuyển đổi dữ liệu vào hồ dữ liệu với đầu vào là toàn bộ loại dữ liệu từ nhiều nguồn với đa dạng định dạng tệp khác nhau.
- Xây dựng kho dữ liệu (data warehouse) hỗ trợ nhân viên có thể truy cập và phân tích yêu cầu kinh doanh một cách nhanh chóng hơn.
- Quy trình ETL thực hiện trích xuất, chuyển đổi và tải dữ vào kho dữ liệu
- Xây dựng báo cáo và các biểu đồ phân tích kinh doanh hỗ trợ quá trình ra quyết định chính xác và kịp thời.

Giải pháp BI dựa trên nền tảng Cloud Azure cung cấp một đường ống dữ liệu hoàn chỉnh từ nguồn đến các báo cáo phân tích kinh doanh trong quá trình thu thập, xử lý, lưu trữ và truy xuất dữ liệu. Kết quả của giải pháp này giúp tối ưu hóa hiệu suất, tiết kiệm thời gian, giảm thiểu chi phí nguồn lực không cần thiết, tăng tính khả dụng và chính xác của dữ liệu.

6.2 Kết luận cho nhóm thực hiện kiến tập

Sau khi trải qua quá trình thực hiện đồ án kiến tập, nhóm đã cơ cơ hội tiếp xúc thêm với nền tảng Cloud kết hợp với giải pháp BI đã được học kỳ học vừa rồi. Từ khâu đầu tiên là nhập dữ liệu đã gây khó khăn trong việc tổng hợp dữ liệu từ nhiều nguồn, từ nhiều loại định dạng tệp khác nhau đến những quy trình xử lý dữ liệu từ dữ liệu thô đến dữ liệu có giá trị cho hoạt động kinh doanh của doanh nghiệp.

Ngoài ra, nhóm còn nhận được sự hướng dẫn tận tình của thầy hướng dẫn, đơn vị kiến tập và các tài liệu liên quan để hoàn thành mặc dù thời gian khá gấp rút, từ đó hình thành cho từng thành viên những kỹ năng nhất định như quản lý thời gian, kỹ năng giải quyết vấn đề, giao tiếp và các kỹ năng mềm khác.

Mở rộng sự hiểu biết của mình ở những phạm vi nghiên cứu lớn hơn về hướng ngành BI giúp mỗi cá nhân có cái hình tổng quan về hệ thống dữ liệu của doanh nghiệp dựa trên nền tảng Cloud Azure trong quá trình xử lý từ dữ liệu nguồn đến các biểu đồ hỗ trợ quá trình ra quyết định chính xác hơn. Điều này bồi dưỡng và rèn luyện thêm kiến thức để chuẩn bị cho một quá trình thực tập, làm việc chính thức với một vị trí chuyên viên phân tích dữ liệu kinh doanh, từ những kiến thức ở trường làm nền tảng để tiếp thu các kiến thức thực tế tại doanh nghiệp.

6.3 Hạn chế thực hiện đề tài

Có một số hạn chế trong quá trình xây dựng giải pháp BI dựa trên nền tảng Cloud Azure, bao gồm việc thiếu nguồn dữ liệu chuẩn chỉnh, giới hạn chi phí trong quá trình triển khai giải pháp trên nền tảng Cloud Azure, quá trình thực hiện đòi hỏi nhiều thời gian thực hiện. Ngoài ra, nhóm chưa có kinh nghiệm triển khai giải pháp BI kết hợp với xây dựng hồ dữ liệu trước đây nên nhóm phải dành nhiều thời gian nghiên cứu kiến thức và kỹ thuật. May mắn thay, đây cũng là cơ hội tốt để mỗi thành viên trong nhóm nâng cao kỹ năng mềm và kiến thức chuyên môn của mình.

6.4 Phương hướng phát triển

Do khả năng và thời gian có hạn, đề tài của nhóm còn nhiều thiếu sót, nhóm còn rất nhiều ý tưởng và dự định để thực hiện. Để đáp ứng nhu cầu ngày càng tăng trong quá trình quản lý và phân tích, nhóm tin rằng giải pháp này sẽ được cải tiến hơn trong tương lai. Nhóm cũng đề xuất một số phương hướng phát triển của đề tài:

- Xây dựng quy trình tự động hóa toàn bộ quy trình.
- Xây dựng hệ thống phân tích dự báo doanh thu và chi phí.
- Xây dựng hệ thống phân tích bình luận khách hàng bằng phương pháp học máy và học sâu.

TÀI LIỆU THAM KHẢO

[1] Davidiseminger (no date) *Bi Solution Architecture in the center of Excellence - Power Bi, Power BI | Microsoft Learn*. Available at: https://learn.microsoft.com/en-us/power-bi/guidance/center-of-excellence-business-intelligence-solution-architecture (Accessed: 10 June 2023).

[2] Wang, C. (2023) *What is ELT?: Blog: Fivetran, RSS*. Available at: https://www.fivetran.com/blog/what-is-elt (Accessed: 10 June 2023).

[3] Sanchez, E. (2022) *ELT vs ETL: Main differences between ETL and ELT (full comparison), Skyvia Blog*. Available at: https://blog.skyvia.com/elt-vs-etl/ (Accessed: 10 June 2023).

[4] Analytics, P.T. (no date) *[11] phân biệt: Database, Data Warehouse, Data Mart, Data Lake, Data Lakehouse, Data Fabric, Data Mesh, LinkedIn*. Available at: https://www.linkedin.com/pulse/11-ph%C3%A2n-bi%E1%BB%87t-data-mesh-vs-lake-warehouse-mart-phuong-thao-analytics/ (Accessed: 10 June 2023).

[5] Nadia Serheichuk, O.S. (2020) *Data Lake vs Data Warehouse: Things you need to know to gain a competitive advantage, N*. Available at: https://www.n-ix.com/data-lake-vs-data-warehouse/ (Accessed: 20 June 2023).

[6] Martinekuan (no date) *Automated enterprise BI - azure architecture center, Azure Architecture Center | Microsoft Learn*. Available at: https://learn.microsoft.com/en-us/azure/architecture/reference-architectures/data/enterprise-bi-adf (Accessed: 20 June 2023).

BẢNG PHÂN CÔNG CÔNG VIỆC

Họ và tên	Nhiệm vụ	Đóng góp (%)	Đánh giá
Trần Nhật Nguyên	- Phân công và giám sát tiến độ thực hiện đề tài - Thực nghiệm xây dựng hồ dữ liệu với Data Factory, Blog Storage, Azure SQL Server - Phụ trách hoàn thành Chương 1: Giới thiệu đề tài - Phụ trác hoàn thành Chương 3: Phân tích chi tiết và xây dựng mô hình - Phụ trách hoàn thành Chương 4: Thực nghiệm – Xây dựng hồ dữ liệu - Phụ trách hoàn thành Chương 6: Kết luận	50%	Tốt
Man Đắc Sang	- Thực nghiệm xây dựng kho dữ liệu với Data Factory, Azure SQL Server - Trực quan hóa dữ liệu với Power BI - Phụ trách hoàn thành Chương 2: Cơ sở lý thuyết - Phụ trách hoàn thành Chương 4: Thực nghiệm – Xây dựng kho dữ liệu - Phụ trách hoàn thành Chương 5: Phân tích dữ liệu – Trực quan hóa - Tổng hợp và định dạng Word	50%	Tốt