TRƯỜNG ĐẠI HỌC THĂNG LONG

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÀI TẬP LỚN**

**XÂY DỰNG DATAWAREHOUSE CHO CÔNG TY CPI CARD GROUP**

|  |  |
| --- | --- |
| **Giáo viên hướng dẫn:** | **Thành viên nhóm:** |
|  | A41674\_ Nguyễn Thiện Đức |
| Vũ Minh Hoàng | A41791\_ Ninh Thị Thanh |
|  | A41798\_ Đỗ Quỳnh Thi |
|  | A41944\_ Nguyễn Huy Cường |
|  | A45602\_ Trương Thị Yến Nhi |

**HÀ NỘI – 2024**

**MỤC LỤC**

[Chương I. GIỚI THIỆU ĐỀ TÀI 1](#_Toc177074570)

[1. Cơ sở lý thuyết 1](#_Toc177074571)

[2. Công ty CPI Card Group 2](#_Toc177074572)

[Chương II. Xác định quy trình 3](#_Toc177074573)

[1. Quy trình marketing 3](#_Toc177074574)

[2. Quy trình sản xuất 3](#_Toc177074575)

[3. Quy trình bán hàng: 4](#_Toc177074576)

[Chương III. Mô hình hóa chiều 6](#_Toc177074577)

[1. Tổng quan dữ liệu 6](#_Toc177074578)

[1.1. ERP 6](#_Toc177074579)

[1.1.1. Customer 7](#_Toc177074580)

[1.1.2. CustLocation 8](#_Toc177074581)

[1.1.3. Invoice 9](#_Toc177074582)

[1.1.4. Job 10](#_Toc177074583)

[1.1.5. Subjob 11](#_Toc177074584)

[1.1.6. Shipment 11](#_Toc177074585)

[1.1.7. Location 12](#_Toc177074586)

[1.1.8. MachineType 12](#_Toc177074587)

[1.1.9. SalesAgent 12](#_Toc177074588)

[1.1.10. SalesClass 13](#_Toc177074589)

[1.2. Financial Summary và LeadFile 13](#_Toc177074590)

[1.2.1. Lead File 13](#_Toc177074591)

[1.2.2. Sales Summary 14](#_Toc177074592)

[1.2.3. Cost Summary 14](#_Toc177074593)

[2. Bus matrix 15](#_Toc177074594)

[2.1. Common Dimensions 16](#_Toc177074595)

[3. Quy trình marketing 18](#_Toc177074596)

[3.1. Mục tiêu 18](#_Toc177074597)

[3.2. Xây dựng star schema 19](#_Toc177074598)

[4. Quy trình sản xuất 20](#_Toc177074599)

[4.1. Sản xuất 20](#_Toc177074600)

[4.2. Vận chuyển 21](#_Toc177074601)

[5. Quy trình bán hàng 23](#_Toc177074602)

[5.1. Tiếp nhận đơn hàng 23](#_Toc177074603)

[5.2. Phát hành và quản lý hóa đơn 24](#_Toc177074604)

[5.3. Tính toán chi phí 26](#_Toc177074605)

[5.4. Báo cáo doanh thu 27](#_Toc177074606)

[Chương IV. Thiết kế data pipeline 29](#_Toc177074607)

[1. Quy trình ETL 29](#_Toc177074608)

[2. Triển khai ETL 30](#_Toc177074609)

[2.1. Tạo CSDL lưu trữ 30](#_Toc177074610)

[2.2. Extract 30](#_Toc177074611)

[2.2.1. ERP 30](#_Toc177074612)

[2.2.2. Lead và Financial Summary 33](#_Toc177074613)

[2.3. Transform 34](#_Toc177074614)

[2.4. Load 34](#_Toc177074615)

[3. Kết quả quá trình ETL và truy vấn 35](#_Toc177074616)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Lead schema ( Marketing) 19](#_Toc177077957)

[Hình 2. Subjob schema (Sản xuất) 20](#_Toc177077958)

[Hình 3. Job\_Shipment schema (Vận chuyển) 22](#_Toc177077959)

[Hình 4. Job Schema (Lập đơn hàng) 23](#_Toc177077960)

[Hình 5. Invoice schema (Phát hành và quản lý hóa đơn) 24](#_Toc177077961)

[Hình 6. Cost Financial Summary Schema ( Tính toán chi phí) 26](#_Toc177077962)

[Hình 7. Sales Financial Summary Schema (Báo cáo doanh thu) 27](#_Toc177077963)

[Hình 8. Lược đồ schema tổng hợp 30](#_Toc177077964)

**BẢNG PHÂN CÔNG CÔNG VIỆC**

|  |  |  |
| --- | --- | --- |
| MSV | Thành viên | Công việc |
| A41944 | Nguyễn Huy Cường | -Giới thiệu đề tài  -Tạo CSDL lưu trữ  -Powerpoint |
| A41798 | Đỗ Quỳnh Thi | -Mô tả dữ liệu  -Lý thuyết ETL chương 4  -Powerpoint |
| A41791 | Ninh Thị Thanh | -Quy trình marketing  -Thiết kế bus matrix + schema marketing  -Code |
| A41674 | Nguyễn Thiện Đức | -Quy trình sản xuất  -Thiết kế schema quy trình sản xuất + schema tiếp nhận đơn hàng |
| A45602 | Trương Thị Yến Nhi | -Quy trình bán hàng  -Thiết kế schema hóa đơn + chi phí +doanh thu  -Powerpoint |

# GIỚI THIỆU ĐỀ TÀI

## Cơ sở lý thuyết

* **Data Warehouse (Kho Dữ Liệu)** là một hệ thống quản lý và lưu trữ dữ liệu, được thiết kế đặc biệt để hỗ trợ quá trình phân tích dữ liệu và ra quyết định. Nó thu thập và lưu trữ dữ liệu từ nhiều nguồn khác nhau, sau đó tổng hợp và sắp xếp dữ liệu để phục vụ cho các mục đích phân tích, báo cáo và ra quyết định chiến lược
* **Kiến trúc data warehouse:**
* **Data Source**: Đây quá trình thu thập dữ liệu từ nhiều nguồn khác nhau, bao gồm cơ sở dữ liệu, tệp CSV,.. Nguồn dữ liệu có thể bao gồm cả dữ liệu cấu trúc và phi cấu trúc, đóng vai trò quan trọng trong việc đảm bảo tính toàn vẹn và phong phú của thông tin.
* **Data Staging:** Sau khi được thu thập, dữ liệu sẽ được chuyển đến giai đoạn tạm thời để thực hiện các quy trình ETL (Extract, Transform, Load). Trong giai đoạn này, dữ liệu được trích xuất từ các nguồn, chuyển đổi để phù hợp với định dạng mong muốn, và cuối cùng được tải vào kho dữ liệu.
* **Data Storage:** Dữ liệu đã qua xử lý được lưu trữ trong Data Warehouse, nơi mà nó được tổ chức một cách có hệ thống. Từ Data Warehouse, có thể phát triển các Data Marts nhằm phục vụ các nhu cầu cụ thể của các phòng ban hoặc tổ chức, giúp cho việc truy xuất và phân tích dữ liệu hiệu quả.
* **Data Presentation:** Dữ liệu từ Data Warehouse được sử dụng cho các hoạt động phân tích và báo cáo. Người dùng có thể thực hiện các phân tích dữ liệu, tạo báo cáo chi tiết, và tiến hành khai thác dữ liệu để phát hiện các mẫu và xu hướng tiềm ẩn, từ đó hỗ trợ quá trình ra quyết định dựa trên dữ liệu
* **Mô hình ngôi sao (Star Schema):** Một bảng trung tâm (fact table) chứa dữ liệu giao dịch và được liên kết với các bảng dimension (bảng chiều) để mô tả các thuộc tính liên quan.
* **Fact Table** lưu trữ các dữ liệu giao dịch, sự kiện hoặc các số liệu cần được phân tích. Nó bao gồm các giá trị số liệu (metrics) như doanh thu, số lượng bán, chi phí, và các khóa ngoại (foreign key) liên kết đến các bảng dimension.
* **Dimension Tables** lưu trữ thông tin mô tả (thuộc tính) về các thực thể kinh doanh như sản phẩm, khách hàng, thời gian, địa điểm, v.v. Các bảng này cung cấp ngữ cảnh cho các số liệu trong Fact Table.

## Công ty CPI Card Group

CPI Card Group là một trong những công ty hàng đầu trong lĩnh vực sản xuất, chuyên cung cấp các giải pháp thẻ tín dụng, thẻ ghi nợ và thẻ trả trước dưới cả hai dạng vật lý và kỹ thuật số.

* **Mục tiêu xây dựng kho dữ liệu:**
* Theo dõi và so sánh doanh thu và chi phí: Công ty cần theo dõi doanh thu và chi phí theo các chiều chính như địa điểm, loại hình doanh thu, đại lý bán hàng và khách hàng qua các khoảng thời gian khác nhau. Điều này giúp họ hiểu rõ hơn về hiệu suất kinh doanh và nhận diện các khu vực cần cải thiện.
* Quản lý chất lượng sản phẩm và hợp đồng: Giám sát chất lượng sản phẩm và hiệu suất hợp đồng là một phần quan trọng trong việc duy trì sự hài lòng của khách hàng và tối ưu hóa quy trình sản xuất. Công ty cần theo dõi các ngày giao hàng, số lượng sản phẩm bị trả lại và các vấn đề chất lượng khác để đảm bảo sản phẩm đáp ứng các tiêu chuẩn cao nhất.
* Phân tích tài chính và đối chiếu: Công ty cần theo dõi và đánh giá hiệu suất tài chính thông qua các truy vấn phân tích tài chính, bao gồm dự báo doanh thu và lập ngân sách, cũng như đối chiếu các báo cáo tài chính để loại bỏ các giao dịch nội bộ và chi phí trùng lặp
* **Kết quả mong muốn:**
* Tích hợp dữ liệu: Tạo ra một kho dữ liệu tập trung từ các hệ thống khác nhau, bao gồm dữ liệu bán hàng, sản xuất, tài chính và dịch vụ khách hàng. Sử dụng công cụ ETL (Extract, Transform, Load) để tích hợp dữ liệu từ các nguồn khác nhau vào một nền tảng duy nhất
* Cải thiện khả năng ra quyết định: Một kho dữ liệu cung cấp dữ liệu sạch, thống nhất và được chuẩn hóa. Ngoài ra còn lưu trữ dữ liệu theo thời gian, cho phép các nhà phân tích so sánh, theo dõi và dự đoán xu hướng. Điều này rất quan trọng đối với các tổ chức cần hiểu được sự thay đổi qua thời gian và đưa ra dự đoán cho tương lai.
* Cải thiện hiệu suất truy vấn và phân tích: Một kho dữ liệu được tối ưu hóa cho các truy vấn phức tạp và các báo cáo dữ liệu lớn. Điều này cho phép truy xuất dữ liệu nhanh chóng, hiệu quả và thực hiện các phân tích sâu sắc mà không ảnh hưởng đến hiệu suất của hệ thống giao dịch.

# Xác định quy trình

## Quy trình marketing

**A close-up of a black background

Description automatically generated**

* **Tìm kiếm khách hàng tiềm năng (lead)**: Khi khách hàng tiềm năng được xác định hoặc quan tâm đến sản phẩm/dịch vụ, một lead được tạo và lưu vào CRM.
* **Báo giá và theo dõi**: Mỗi lead được gửi một hoặc nhiều báo giá (gồm vật liệu, nhân công, máy móc và các chi phí khác để tạo ra sản phẩm cuối cùng.). Đại lý bán hàng chịu trách nhiệm giao tiếp với lead để tư vấn và đàm phán, nhằm biến lead thành khách hàng thực sự.
* **Phân tích hiệu quả**: Kết quả của báo giá cho thấy tỷ lệ thành công (Y) hoặc thất bại (N) của từng lead. Dựa trên dữ liệu này, phòng marketing có thể phân tích tỷ lệ thành công của các chiến dịch tiếp thị và điều chỉnh chiến lược.
* **Báo cáo và đánh giá**:
* Phân loại và tối ưu hóa: Mỗi khách hàng và lead có thể được phân loại theo sản phẩm, giúp marketing xác định các phân khúc khách hàng và điều chỉnh chiến dịch theo từng nhóm.
* Dữ liệu được lưu trữ và báo cáo thường xuyên để đánh giá hiệu quả tiếp thị, hiệu quả của các đại diện bán hàng, và khả năng chuyển đổi lead thành khách hàng.

## Quy trình sản xuất

* **Nhận công việc:**
* **Tiếp nhận công việc:** Khi có một công việc được ghi vào hệ thống, bộ phận lập kế hoạch sản xuất sẽ xem xét báo giá bán hàng cần thiết để thực hiện công việc
* **Thiết kế và lập kế hoạch sản xuất**
* **Duyệt Thiết Kế:** Thiết kế thẻ sẽ được tạo hoặc sửa đổi dựa trên yêu cầu của khách hàng. Thiết kế được gửi cho khách hàng để phê duyệt trước khi tiến hành sản xuất.
* **Lập kế hoạch sản xuất:** chuyển đổi báo giá bán hàng và dữ liệu công việc thành kế hoạch sản xuất thực tế (Subjob) để tạo ra sản phẩm cuối cùng.
* **Sản Xuất:**
* **In Ấn:** Thẻ được in ấn với thiết kế đã được phê duyệt. CPI Card Group sử dụng các công nghệ in tiên tiến để đảm bảo chất lượng cao.
* **Chế Tạo:** Thẻ được cắt, cán màng bảo vệ, và thực hiện các bước gia công cần thiết như in số seri, mã vạch, hoặc chip từ.
* **Kiểm Tra Chất Lượng:** Các thẻ được kiểm tra chất lượng để đảm bảo chúng đáp ứng các tiêu chuẩn kỹ thuật và yêu cầu của khách hàng.
* **Đóng Gói:**
* **Đóng gói thẻ:** Các thẻ hoàn thành được đóng gói theo yêu cầu của đơn hàng.
* **Vận Chuyển:**
* **Gửi đơn hàng:** Đơn hàng được gửi đến khách hàng qua các dịch vụ vận chuyển. CPI Card Group có thể cung cấp các tùy chọn vận chuyển khác nhau tùy thuộc vào yêu cầu của khách hàng.
* **Theo dõi đơn hàng:** Quá trình vận chuyển được theo dõi để đảm bảo rằng đơn hàng đến nơi đúng thời gian và trong tình trạng tốt.

## Quy trình bán hàng:

* **Tiếp nhận và xử lý đơn hàng**
* **Tiếp nhận đơn hàng:** tạo một công việc sản xuất mới (Job) (đơn hàng), hợp nhất báo giá từ CRM vào hệ thống.
* **Xử lý đơn hàng:** xem xét báo giá bán hàng,thu thập số liệu, chi phí sản xuất
* **Phát hành và quản lý hóa đơn**
* **Phát hành hóa đơn:** Sau khi xử lý đơn hàng (giao hàng), hóa đơn được lập để gửi cho khách hàng, ghi rõ chi tiết các sản phẩm, số lượng, giá cả, thuế và tổng số tiền cần thanh toán.
* **Quản lý hóa đơn:** sau khi đơn hàng được vận chuyển và thanh toán, tiến hành xuất hóa đơn
* **Ghi nhận và tóm tắt chi phí**
* **Ghi nhận chi phí liên quan**: Bao gồm chi phí mua hàng, vận chuyển, nhân công, khuyến mãi, và các chi phí phụ khác.
* **Phân tích chi phí:** Xác định các loại chi phí chính yếu ảnh hưởng đến lợi nhuận, từ đó đưa ra các chiến lược cải thiện.
* **Báo cáo doanh thu**
* **Phân tích tài chính:** Tổng hợp và báo cáo dựa trên chi phí, giá trị và lợi nhuận đơn hàng
* **Đánh giá hiệu suất**: So sánh tổng chi phí với doanh thu để đánh giá lợi nhuận và hiệu quả kinh doanh, từ đó tối ưu hóa quy trình bán hàng.

# Mô hình hóa chiều

## Tổng quan dữ liệu

Dữ liệu được lấy từ 3 nguồn khác nhau (ERP, Lead File (Hoạt động marketing), Financial Summaries (Hoạt động tài chính)).

A diagram of a data processing process

Description automatically generated

### ERP

Cơ sở dữ liệu ERP là nguồn dữ liệu chính được sử dụng bởi sản xuất để quản lý toàn diện, giúp theo dõi và tối ưu hóa công việc, công việc phụ, hoạt động kinh doanh từ quản lý khách hàng, hóa đơn, sản xuất đến vận chuyển lô hàng.

Cơ sở dữ liệu ERP bao gồm 10 bảng dữ liệu:

* **Customer**: Lưu trữ thông tin về khách hàng, gồm 500 dòng và 17 cột.
* **CustLocation**: Lưu trữ thông tin về địa điểm của khách hàng, gồm 1481 dòng và 18 cột.
* **Invoice**: Chứa thông tin về hóa đơn, bao gồm ngày xuất, số tiền và số lượng hàng hóa, gồm 20632 dòng và 9 cột.
* **Job**: Ghi lại thông tin về các công việc, từ mô tả đến ngày ký hợp đồng, gồm 2569 dòng và 19 cột.
* **Subjob**: Ghi lại thông tin về các công việc phụ, bao gồm chi phí và số lượng sản phẩm, gồm 8057 dòng và 12 cột.
* **Shipment**: Cung cấp thông tin về các lô hàng, bao gồm ngày gửi và số lượng hàng hóa, gồm 24235 dòng và 13 cột.
* **Location**: Lưu trữ thông tin về vị trí, như tên và mô tả, gồm 12 dòng và 2 cột.
* **MachineType**: Thông tin về các loại máy móc, bao gồm nhà sản xuất và chi phí sử dụng, gồm 7 dòng và 5 cột.
* **SalesAgent**: Thông tin về đại lý bán hàng, bao gồm tên và khu vực hoạt động, gồm 500 dòng và 5 cột.
* **SalesClass**: Lưu trữ thông tin về các sản phẩm, bao gồm mô tả và giá cơ bản, gồm 6 dòng và 2 cột.

A diagram of a computer

Description automatically generated

#### Customer

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Definition** |
| Cust\_Key | LONG | Khóa chính, định danh khách hàng |
| Cust\_Name | VARCHAR | Tên khách hàng |
| Active | BOOLEAN | Đánh dấu xem khách hàng có đang hoạt động hay không |
| Address\_1 | VARCHAR | Dòng 1 của địa chỉ khách hàng |
| Address\_2 | VARCHAR | Dòng 2 của địa chỉ khách hàng |
| Address\_3 | VARCHAR | Dòng 3 của địa chỉ khách hàng |
| City | VARCHAR | Thành phố của khách hàng |
| Country | VARCHAR | Quốc gia của khách hàng |
| Credit\_Limit | CURRENCY | Hạn mức tín dụng của khách hàng |
| Date\_First\_Order | DATE | Ngày đặt hàng đầu tiên của khách hàng |
| E\_Mail\_Address | VARCHAR | Địa chỉ email của khách hàng |
| Fax | CHAR(10) | Số fax của khách hàng |
| Phone | CHAR(10) | Số điện thoại của khách hàng (Có thể được phân tích) |
| Sales\_Tax\_Code | CHAR(10) | Mã thuế bán hàng của khách hàng |
| State | CHAR(2) | Tiểu bang của khách hàng |
| Terms\_Code | CHAR(10) | Chỉ ra điều khoản thanh toán |
| ZIP | CHAR(10) | Mã bưu điện của khách hàng |

#### CustLocation

|  |  |  |
| --- | --- | --- |
| **Column name** | **Data type** | **Defition** |
| Cust\_Loc\_Key | LONG | Mã định danh duy nhất cho địa điểm của khách hàng |
| Cust\_Name | VARCHAR | Tên của địa điểm khách hàng |
| Active | BOOLEAN | Đánh dấu xem địa điểm của khách hàng có đang hoạt động hay không |
| Address\_1 | VARCHAR | Dòng 1 của địa chỉ địa điểm khách hàng |
| Address\_2 | VARCHAR | Dòng 2 của địa chỉ địa điểm khách hàng |
| Address\_3 | VARCHAR | Dòng 3 của địa chỉ địa điểm khách hàng |
| City | VARCHAR | Thành phố của địa điểm khách hàng |
| Country | VARCHAR | Quốc gia của địa điểm khách hàng |
| Credit\_Limit | CURRENCY | Hạn mức tín dụng của địa điểm khách hàng |
| Date\_First\_Order | DATE | Ngày đặt hàng đầu tiên của địa điểm khách hàng |
| E\_Mail\_Address | VARCHAR | Địa chỉ email của địa điểm khách hàng |
| Fax | CHAR(10) | Số fax của địa điểm khách hàng |
| Phone | CHAR(10) | Số điện thoại của địa điểm khách hàng (có thể được phân tích) |
| Sales\_Tax\_Code | CHAR(10) | Mã số thuế bán hàng của địa điểm khách hàng (chỉ dùng cho khách hàng không thương mại) |
| State | CHAR(2) | Tiểu bang/tỉnh của địa điểm khách hàng |
| Terms\_Code | CHAR(10) | Điều khoản thanh toán của khách hàng |
| ZIP | CHAR(10) | Mã bưu điện của địa điểm khách hàng |
| Cust\_Key | LONG | Mã định danh của khách hàng (liên kết với bảng khách hàng) |

#### Invoice

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Invoice\_Id | LONG | Mã định danh duy nhất của hóa đơn |
| Date\_Invoiced | DATE | Ngày lập hóa đơn |
| Date\_Due | DATE | Ngày đến hạn thanh toán (phụ thuộc vào điều khoản thanh toán) |
| Posting\_Date | DATE | Ngày ghi nhận thanh toán |
| Invoice\_Desc | VARCHAR | Mô tả nội dung hóa đơn |
| Invoice\_Amount | CURRENCY | Tổng số tiền của hóa đơn |
| Invoice\_Quantity | INTEGER | Số lượng hàng hóa trên hóa đơn |
| Invoice\_Shipped | INTEGER | Số lượng hàng hóa đã giao; Số lượng trả lại bằng hiệu số giữa số lượng trên hóa đơn và số lượng đã giao |
| Cust\_Key | LONG | Mã định danh của khách hàng được lập hóa đơn |

#### Job

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Job\_Id | LONG | Mã định danh duy nhất của công việc |
| Job\_Desc | CHAR(50) | Mô tả công việc |
| Contract\_Date | DATE | Ngày lập hợp đồng công việc |
| Date\_Promised | DATE | Ngày dự kiến giao hàng cuối cùng |
| Number\_Of\_Subjobs | INTEGER | Số lượng công việc con |
| PO\_Number | CHAR(10) | Mã số đơn đặt hàng |
| Record\_Active | BOOLEAN | Trạng thái hoạt động của công việc |
| Date\_Ship\_By | DATE | Ngày dự kiến giao hàng đầu tiên |
| Job\_Complete | BOOLEAN | Trạng thái hoàn thành công việc |
| Print\_Flow\_Status | CHAR(10) | Trạng thái sản xuất |
| Contact\_Name | CHAR(50) | Tên người liên hệ |
| Unit\_Price | CURRENCY | Giá đơn vị sản phẩm |
| Quantity\_Ordered | SHORT | Số lượng đặt hàng |
| Quotation\_Amount | CURRENCY | Tổng giá trị báo giá |
| Quotation\_Ordered | SHORT | Số lượng đặt hàng ban đầu |
| Location\_Id | LONG | Mã định danh vị trí |
| Sales\_Class\_id | LONG | Mã định danh loại bán hàng |
| Sales\_Agent\_Id | LONG | Mã định danh nhân viên bán hàng |
| Cust\_Key | LONG | Mã định danh khách hàng |

#### Subjob

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Job\_Id | LONG | Mã định danh của công việc chính và là một phần của khóa chính |
| Sub\_Job\_Id | SHORT | Mã định danh của công việc con trong công việc chính |
| Sub\_Job\_Desc | CHAR(50) | Mô tả của công việc con |
| Cost\_Labor | CURRENCY | Chi phí nhân công cho công việc con |
| Cost\_Material | CURRENCY | Chi phí vật liệu cho công việc con |
| Cost\_Overhead | CURRENCY | Chi phí chung cho công việc con |
| Machine\_Hours | DECIMAL | Số giờ máy sử dụng cho công việc con |
| Date\_Prod\_Begin | DATE | Ngày bắt đầu sản xuất công việc con |
| Date\_Prod\_End | DATE | Ngày kết thúc sản xuất công việc con |
| Quantity\_Produced | INTEGER | Số lượng sản phẩm sản xuất được cho công việc con |
| Sub\_Job\_Amount | CURRENCY | Giá trị tiền tệ của sản phẩm sản xuất được cho công việc con |
| Machine\_Type\_Id | LONG | Mã định danh loại máy sử dụng cho công việc con |

#### Shipment

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Shipment\_id | LONG | Mã định danh duy nhất của lô hàng |
| Actual\_Ship\_Date | DATE | Ngày thực tế lô hàng diễn ra |
| Requested\_Ship\_Date | DATE | Ngày khách hàng yêu cầu giao hàng |
| Actual\_Quantity | INTEGER | Số lượng thực tế hàng hóa được giao |
| Requested\_Quantity | INTEGER | Số lượng hàng hóa khách hàng yêu cầu |
| Boxes | INTEGER | Số lượng hộp đầy trong lô hàng |
| Quantity\_Per\_Box | INTEGER | Số lượng hàng hóa trong mỗi hộp đầy |
| Quantity\_Per\_Partial\_Box | INTEGER | Số lượng hàng hóa trong hộp không đầy |
| Job\_Id | LONG | Mã định danh của công việc liên quan đến lô hàng |
| Shipment\_Amount | CURRENCY | Số tiền cần thanh toán cho lô hàng |
| Sub\_Job\_Id | LONG | Mã định danh của công việc con liên quan đến lô hàng |
| Invoice\_Id | LONG | Mã định danh của hóa đơn liên quan đến lô hàng; null nếu chưa lập hóa đơn |
| Cust\_Loc\_Key | LONG | Mã định danh của địa điểm khách hàng liên quan |

#### Location

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Location\_Id | LONG | Mã định danh duy nhất của địa điểm |
| Location\_Name | CHAR(50) | Tên của địa điểm |

#### MachineType

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Machine\_Type\_Id | LONG | Mã định danh duy nhất của loại máy |
| Manufacturer | LONG | Nhà sản xuất của loại máy |
| Model | VARCHAR | Mô hình cụ thể của loại máy |
| Rate Per Hour | CURRENCY | Giá thuê mỗi giờ cho loại máy |
| Number\_Of\_Machines | INTEGER | Số lượng máy có sẵn |

#### SalesAgent

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Sales\_Agent\_id | LONG | Mã định danh duy nhất của nhân viên bán hàng |
| Sales\_Agent\_Name | VARCHAR | Tên của nhân viên bán hàng |
| State | CHAR(2) | Tiểu bang nơi nhân viên bán hàng làm việc |
| Country | CHAR(25) | Quốc gia nơi nhân viên bán hàng làm việc |
| Record\_Active | BOOLEAN | Trạng thái hoạt động: True nếu nhân viên đang làm việc cho công ty |

#### SalesClass

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Difinition** |
| Sales\_Class\_Id | LONG | Mã định danh duy nhất cho loại sản phẩm bán hàng |
| Sales\_Class\_Desc | VARCHAR | Loại sản phẩm được sản xuất trong công việc |

### Financial Summary và LeadFile

Lead File và Financial Summaries là các nguồn dữ ở định dạng bảng tính và được chuẩn bị từ các nguồn dữ liệu khác được sử dụng bởi các phòng ban marketing và kế toán.

Dữ liệu thứ cấp bao gồm 3 tập là:

* **lead.csv**: có 3000 dòng và 10 cột, cung cấp thông tin chi tiết về khách hàng tiềm năng, giúp theo dõi và đánh giá khả năng thành công.
* **sales summary.csv**: có 2160 dòng và 9 cột, chứa dữ liệu tài chính liên quan đến doanh thu, giúp theo dõi kết quả kinh doanh và so sánh giữa số liệu thực tế và dự báo.
* **cost\_summary:** có 13.440 dòng và 16 cột, bao gồm các chi phí tài chính, giúp phân tích mối liên hệ giữa chi phí và hiệu suất sản xuất.

#### Lead File

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Definition** |
| Lead\_Id | LONG | Mã định danh duy nhất của khách hàng tiềm năng |
| Quote\_Qty | INTEGER | Số lượng sản phẩm được báo giá cho khách hàng tiềm năng |
| Quote\_Price | CURRENCY | Giá mỗi sản phẩm được báo giá cho khách hàng tiềm năng |
| Quote\_Value | CURRENCY | Tổng giá trị báo giá cho khách hàng tiềm năng |
| Success | BOOLEAN | Đánh dấu xem khách hàng tiềm năng có trở thành đơn hàng hay không |
| PO\_Number | LONG | Số đơn đặt hàng nếu khách hàng tiềm năng trở thành đơn hàng |
| Created\_Date | DATE | Ngày tạo khách hàng tiềm năng |
| Cust\_Id | LONG | Mã định danh của khách hàng liên quan đến khách hàng tiềm năng |
| Location\_Id | LONG | Mã định danh của vị trí liên quan đến khách hàng tiềm năng |
| Sales Agent Id | LONG | Mã định danh của nhân viên bán hàng liên quan đến khách hàng tiềm năng |
| Sales Class Id | LONG | Mã định danh của loại sản phẩm liên quan đến khách hàng tiềm năng |

#### Sales Summary

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Definition** |
| Summary\_Sales\_Id | LONG | Mã định danh duy nhất cho bản tóm tắt doanh số |
| Actual\_Units | INTEGER | Tổng số lượng sản phẩm bán được thực tế đến cuối kỳ |
| Actual\_Amount | CURRENCY | Tổng doanh thu thực tế đến cuối kỳ |
| Forecast\_Units | INTEGER | Số lượng sản phẩm dự kiến bán được trong năm |
| Forecast\_Amount | CURRENCY | Doanh thu dự kiến trong năm |
| Location\_Id | LONG | Mã định danh của địa điểm |
| Sales\_Class\_id | LONG | Mã định danh của loại sản phẩm |
| Begin\_Date | DATE | Ngày bắt đầu tính toán số lượng và doanh thu thực tế, thường là ngày đầu tiên của tháng |
| End\_Date | DATE | Ngày kết thúc tính toán số lượng và doanh thu thực tế, thường là ngày cuối cùng của tháng |

#### Cost Summary

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Definition** |
| Summary\_Cost\_Id | LONG | Mã định danh duy nhất cho bản tóm tắt chi phí |
| Actual\_Units | INTEGER | Tổng số lượng thực tế đến cuối kỳ |
| Actual\_Labor\_Cost | CURRENCY | Tổng chi phí lao động thực tế đến cuối kỳ |
| Actual\_Material\_Cost | CURRENCY | Tổng chi phí vật liệu thực tế đến cuối kỳ |
| Actual\_Machine\_Cost | CURRENCY | Tổng chi phí máy móc thực tế đến cuối kỳ |
| Actual\_Overhead\_Cost | CURRENCY | Tổng chi phí chung thực tế đến cuối kỳ |
| Budget\_Units | INTEGER | Tổng số lượng dự toán cả năm |
| Budget\_Labor\_Cost | CURRENCY | Tổng chi phí lao động dự toán cả năm |
| Budget\_Material\_Cost | CURRENCY | Tổng chi phí vật liệu dự toán cả năm |
| Budget\_Machine\_Cost | CURRENCY | Tổng chi phí máy móc dự toán cả năm |
| Budget\_Overhead\_Cost | CURRENCY | Tổng chi phí chung dự toán cả năm |
| Location\_Id | LONG | Mã định danh của vị trí |
| Machine\_Type\_id | LONG | Mã định danh của loại máy |
| Sales\_Class\_Id | LONG | Mã định danh của loại sản phẩm |
| Begin\_Date | DATE | Ngày bắt đầu tính toán chi phí thực tế, thường là ngày đầu tiên của tháng |
| End\_Date | DATE | Ngày kết thúc tính toán chi phí thực tế, thường là ngày cuối cùng của tháng |

## Bus matrix

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Common Dim**  **Process** | **Time** | **Customer** | **Customer Location** | **SalesClass** | **SalesAgent** | **Location** | **Machine Type** |
| Marketing | x | x |  | x | x | x |  |
| Tiếp nhận đơn hàng | x | x |  | x | x | x |  |
| Sản xuất | x | x |  | x |  | x | x |
| Vận chuyển | x |  | x | x |  | x |  |
| Phát hành và quản lý hóa đơn | x | x |  | x | x | x |  |
| Tính toán chi phí | x |  |  | x |  | x | x |
| Báo cáo doanh thu | x |  |  | x |  | x |  |

### Common Dimensions

* W\_Time\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Time\_Id | INT | Mã định danh duy nhất cho mỗi đơn vị thời gian. |
| Time\_Year | INT | Năm. |
| Time\_Quarter | INT | Quý của năm (Q1, Q2, ...). |
| Time\_Month | INT | Tháng trong năm. |
| Time\_Day | INT | Ngày trong tháng. |
| Time\_Week | INT | Tuần trong năm. |

* W\_Customer\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Cust\_Key | INT | Mã định danh duy nhất cho mỗi khách hàng. |
| Cust\_Name | VARCHAR | Tên của khách hàng. |
| City | VARCHAR | Thành phố của khách hàng. |
| County | VARCHAR | Quận/huyện nơi khách hàng cư trú. |
| Credit\_Limit | DECIMAL | Giới hạn tín dụng của khách hàng. |
| E\_Mail\_Address | VARCHAR | Địa chỉ email của khách hàng. |
| Terms\_Code | VARCHAR | Mã điều khoản thanh toán. |
| Cust\_State | VARCHAR | Bang/Tỉnh nơi khách hàng cư trú. |
| ZIP | VARCHAR | Mã ZIP của khách hàng. |

* Bảng W\_Cust\_Location\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Cust\_Loc\_Key | INT | Mã định danh duy nhất cho mỗi địa điểm. |
| Cust\_Name | VARCHAR | Tên của khách hàng. |
| City | VARCHAR | Thành phố của khách hàng. |
| Country | VARCHAR | Quận/huyện nơi khách hàng cư trú. |
| Credit\_Limit | DECIMAL | Giới hạn tín dụng của khách hàng. |
| E\_Mail\_Address | VARCHAR | Địa chỉ email của khách hàng. |
| Cust\_Location\_State | VARCHAR | Bang/Tỉnh nơi khách hàng cư trú. |
| Zip | VARCHAR | Mã ZIP của khách hàng. |
| Cust\_Key | INT | Mã định danh duy nhất cho mỗi khách hàng. |

* Bảng W\_Sales\_Agent\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Sales\_Agent\_Id | INT | Mã định danh duy nhất cho mỗi đại lý bán hàng. |
| Sales\_Agent\_Name | VARCHAR | Tên của đại lý bán hàng. |
| State | VARCHAR | Bang/Tỉnh mà đại lý bán hàng hoạt động. |
| Country | VARCHAR | Quốc gia nơi đại lý bán hàng hoạt động. |

* W\_Sales\_Class\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Sales\_Class\_Id | INT | Mã định danh duy nhất cho mỗi loại bán hàng. |
| Sales\_Class\_Desc | VARCHAR | Mô tả về loại bán hàng. |
| Base\_Price | DECIMAL | Giá cơ bản liên quan đến loại bán hàng. |

* Bảng W\_Location\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Location\_Id | INT | Mã định danh duy nhất cho mỗi địa điểm. |
| Location\_Name | VARCHAR | Tên của địa điểm. |

* Bảng W\_Machine\_Type\_D

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Machine\_Type\_Id | INT | Mã định danh cho loại máy móc. |
| Manufacturer | VARCHAR | Tên nhà sản xuất máy móc. |
| Machine\_Model | VARCHAR | Mô hình máy. |
| Rate\_Per\_Hour | DECIMAL | Chi phí thuê hoặc sử dụng máy móc theo giờ. |
| Number\_Of\_Machines | INT | Số lượng máy thuộc loại máy móc này. |

## Quy trình marketing

### Mục tiêu

Xây dựng báo cáo kết hợp các chiều dữ liệu để có cái nhìn tổng quan về hiệu suất marketing:

* Đánh giá tỷ lệ chuyển đổi lead thành khách hàng của đại lý bán hàng, theo các tiêu chí như địa lý, thời gian ( ngày, tuần,tháng,quý, năm)
* Xác định những địa điểm có tỷ lệ thành công cao nhất.
* Theo dõi xu hướng báo giá và mặt hàng ưa chuộng

### Xây dựng star schema

Hình 1. Lead schema ( Marketing)

Mô tả:

* Bảng W\_Lead\_Fact chứa các dữ liệu liên quan đến lead (khách hàng tiềm năng) bao gồm :

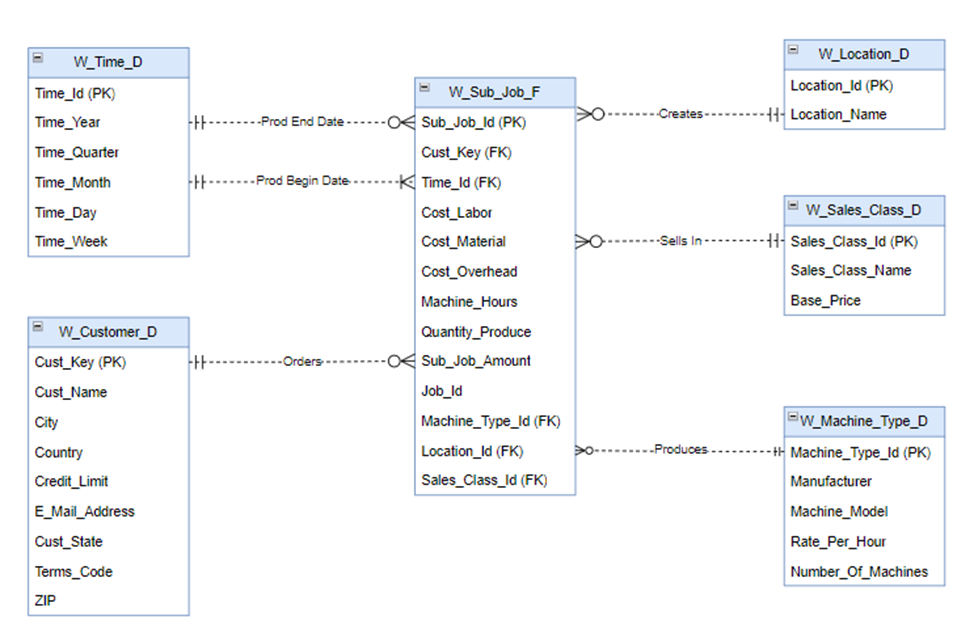
|  |  |  |
| --- | --- | --- |
| **Tên trường** | **Mô tả** | **Kiểu dữ liệu** |
| Lead\_Id | Mã của KH tiềm năng | int |
| Quote\_Qty | Số lượng báo giá mà khách hàng tiềm năng nhận được. | int |
| Quote\_Price | Giá báo cho khách hàng | float |
| Lead\_Success | Kết quả báo giá | char |
| Location\_ID (FK) | Mã địa điểm | int |
| Sales\_Agent\_Id (FK) | Mã nhân viên | int |
| Time\_Id (FK) | Mã thời gian (Thời gian tạo Lead) | int |
| Cust\_Key (FK) | Mã khách hàng | int |
| Sales\_Class\_Id (FK) | Mã loại sản phẩm | int |

* Các bảng Dimension: W\_Time\_D, W\_Customer\_D, W\_Location\_D, W\_Sales\_Agent\_D, W\_Sales\_Class\_D

## Quy trình sản xuất

### Sản xuất

* Mục tiêu:
* Phân tích chi phí sản xuất cho từng loại mặt hàng trên mỗi khách hàng bao gồm chi phí về vật liệu, nhãn dán,…
* Phân tích hiệu suất sản xuất từng từng loại máy theo từng loại sản phẩm theo thời gian (ngày, tuần,..)



Hình 2. Subjob schema (Sản xuất)

Ảnh **Error! Use the Home tab to apply 0 to the text that you want to appear here.**.2.Subjob schema (Sản xuất)

Mô tả:

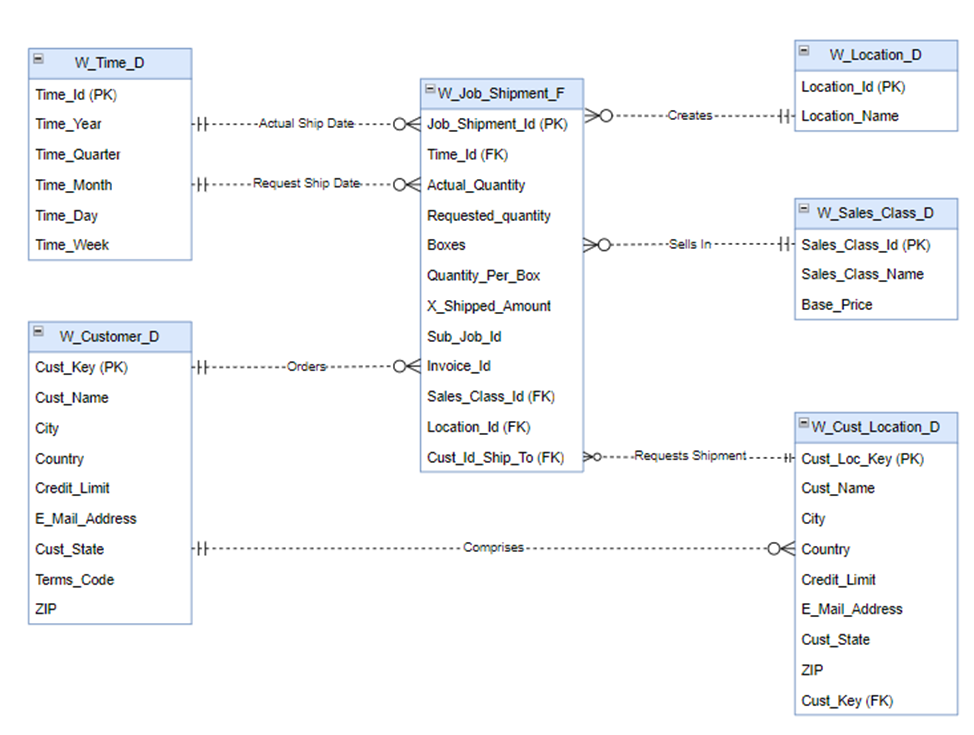
* Bảng W\_Sub\_Job\_Fact chứa các dữ liệu liên quan đến việc theo dõi các chi tiết về công việc phụ trong một công ty sản xuất nhằm mục đích phân tích hiệu quả công việc và chi phí. bao gồm:

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| Sub\_Job\_Id (PK) | INT | Khóa chính của bảng, định danh duy nhất cho từng công việc phụ (sub-job). |
| Cust\_Key (FK) | INT | Khách hàng liên quan đến công việc phụ này. |
| Time\_Id (FK) | INT | Thời gian liên quan đến công việc phụ này. |
| Cost\_Labor | CURRENCY | Chi phí nhân công liên quan đến công việc phụ |
| Cost\_Material | CURRENCY | Chi phí vật liệu sử dụng cho công việc phụ. |
| Cost\_Overhead | CURRENCY | Chi phí quản lý hoặc chi phí chung cho công việc phụ. |
| Machine\_Hours | DECIMAL | Số giờ máy móc được sử dụng để hoàn thành công việc phụ. |
| Quantity\_Produce | INT | Số lượng sản phẩm được sản xuất trong công việc phụ. |
| Sub\_Job\_Amount | CURRENCY | Tổng giá trị (số tiền) của công việc phụ. |
| Job\_Id (FK) | INT | Mã công việc chính |
| Machine\_Type\_Id (FK) | INT | Mã loại máy móc sử dụng trong công việc phụ |
| Location\_Id (FK) | INT | Mã vị trí |
| Sales\_Class\_Id (FK) | INT | Mã phân loại sản phẩm |

* Các bảng Dimension: W\_Time\_D, W\_Customer\_D, W\_Location\_D, W\_Sales\_Class\_D, W\_Machine\_Type\_D

### Vận chuyển

* Mục tiêu:
* Phân tích hiệu suất thực tế dựa trên yêu cầu và số sản phẩm thực tế được giao
* Phân tích độ tin cậy dựa trên so sánh thời gian vận chuyển thực tế và yêu cầu
* Phân tích số lượng hàng hóa được vận chuyển đến mỗi khách hàng



Hình 3. Job\_Shipment schema (Vận chuyển)

Mô tả:

* Bảng W\_Job\_Shipment\_Fact chứa các dữ liệu liên quan đến việc vận chuyển đơn hàng nhằm mục đích phân tích và theo dõi hiệu suất vận chuyển, đồng thời hỗ trợ việc tối ưu hóa quy trình vận chuyển và phân phối bao gồm:

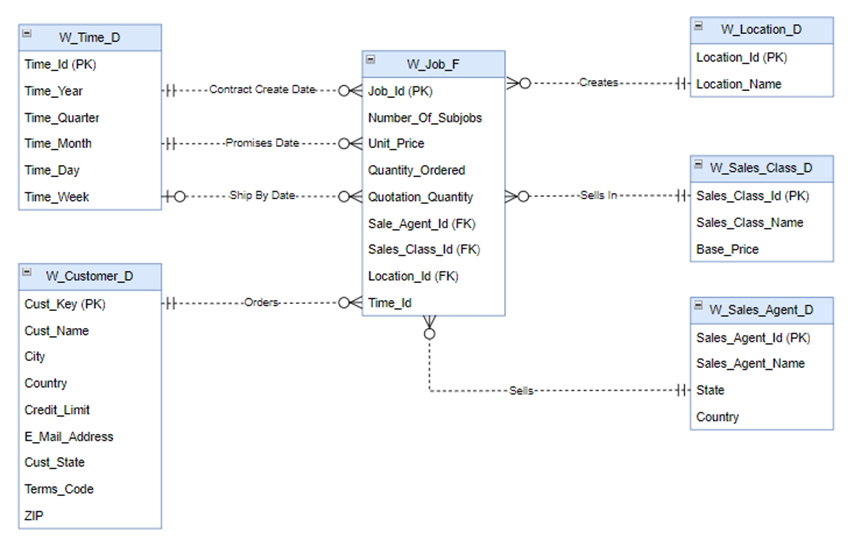
|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| Job\_Shipment\_Id (PK) | INT | Mã đơn hàng vận chuyển |
| Time\_Id (FK) | INT | Mã thời gian vận chuyển (thời gian yêu cầu, thời gian giao thực tế) |
| Actual\_Quantity | INT | Số lượng thực tế đã vận chuyển cho đơn hàng này. |
| Requested\_Quantity | INT | Số lượng mà khách hàng hoặc đơn hàng yêu cầu vận chuyển. |
| Boxes | INT | Số lượng hộp hoặc thùng sử dụng để vận chuyển  hàng |
| Quantity\_Per\_Box | INT | Số lượng sản phẩm trong mỗi hộp hoặc thùng. |
| X\_Shipped\_Amount | CURRENCY | Giá trị tiền của hàng hóa đã được vận chuyển. |
| Sub\_Job\_Id | INT | Tham chiếu đến công việc phụ (sub-job) liên quan đến quá trình vận chuyển này. |
| Invoice\_Id | INT | Mã hóa đơn liên quan đến việc vận chuyển. |
| Sales\_Class\_Id (FK) | INT | Mã lớp bán hàng của sản phẩm được vận chuyển. |
| Location\_Id (FK) | INT | Mã vị trí xuất hàng hoá. |
| Cust\_Id\_Ship\_To (FK) | INT | Mã khách hàng mà hàng hóa được vận chuyển tới. |

* Các bảng Dimension: W\_Time\_D, W\_Customer\_D, W\_Location\_D, W\_Sales\_Class\_D, W\_Cust\_Location\_D

## Quy trình bán hàng

### Tiếp nhận đơn hàng

* Mục tiêu
* Phân tích giá trị đơn hàng trên số lượng đặt hàng của từng loại hàng trên từng khách hàng ở các địa điểm khác nhau
* Phân tích hiệu suất bán hàng của đại lý, số lượng công việc trên đơn đặt hàng



Hình 4. Job Schema (Lập đơn hàng)

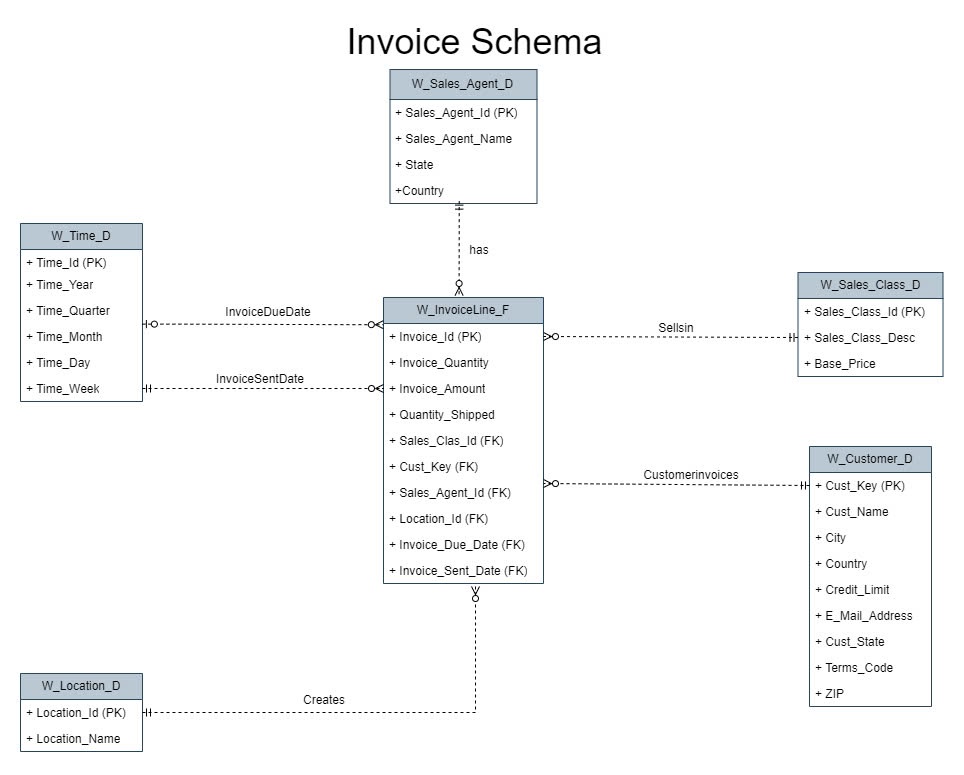
* Bảng W\_Job\_Fact chứa các dữ liệu liên quan đến các đơn hàng bao gồm:

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| Job\_Id | INT | Mã đơn hàng |
| Number\_Of\_Subjobs | INT | Số lượng các công việc phụ (subjobs) được chia nhỏ từ một công việc chính (job) |
| Unit\_Price | CURRENCY | Giá của một đơn vị sản phẩm |
| Quantity\_Ordered | INT | Số lượng sản phẩm đã đặt hàng |
| Quotation\_Quantity | INT | Số lượng sản phẩm trong báo giá |
| Sales\_Agent\_Id (FK) | INT | Mã nhân viên bán hàng |
| Sales\_Class\_Id (FK) | INT | Mã phân loại hàng hoá, sản phẩm |
| Location\_Id (FK) | INT | Mã vị trí |
| Time\_Id (FK) | INT | Mã thời gian (thời gian tạo đơn hàng,thời gian hẹn giao, thời gian giao thực tế) |

* Các bảng Dimension: W\_Time\_D, W\_Customer\_D, W\_Location\_D, W\_Sales\_Agent\_D, W\_Sales\_Class\_D

### Phát hành và quản lý hóa đơn

Mục tiêu: Phân tích doanh thu bán hàng theo từng mặt hàng, đại lý , khách hàng trong các khoảng thời gian tại các địa điểm



Hình 5. Invoice schema (Phát hành và quản lý hóa đơn)

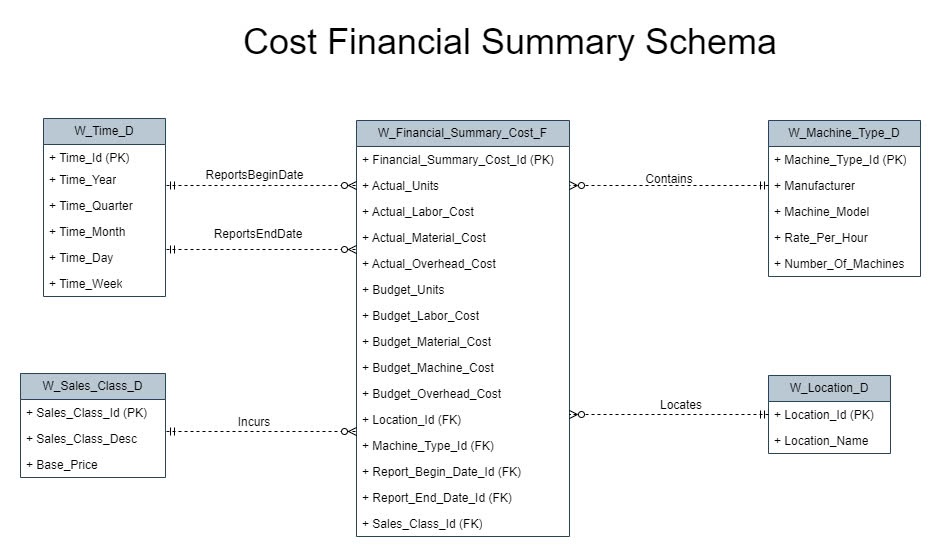
* Bảng W\_InvoiceLine\_F

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Invoice\_Id | INT | Mã định danh duy nhất cho mỗi hóa đơn. |
| Invoice\_Quantity | INT | Số lượng sản phẩm trong hóa đơn. |
| Invoice\_Amount | DECIMAL | Tổng số tiền của hóa đơn. |
| Quantity\_Shipped | INT | Số lượng sản phẩm đã được giao. |
| Sales\_Class\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Sales\_Class \_D. |
| Cust\_Key | INT | Khóa ngoại liên kết đến bảng W\_ Customer \_D. |
| Sales\_Agent\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Sales\_Agent \_D. |
| Location\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Location \_D. |
| Invoice\_Due\_Date | INT | Khóa ngoại liên kết đến bảng W\_Time\_D, xác định ngày đến hạn thanh toán của hóa đơn |
| Invoice\_Sent\_Date | INT | Khóa ngoại liên kết đến bảng W\_Time\_D, xác định ngày hóa đơn được gửi đi cho khách hàng. |

* Các bảng Dim: W\_Sales\_Agent\_D, W\_Time\_D, W\_Location\_D, W\_Sales\_Class\_D, W\_Customer\_D

### Tính toán chi phí

* Mục tiêu
* Phân tích chi phí sản của từng loại hàng
* Phân tích chênh lệch chi phí giữa ngân sách và thực tế



Hình 6. Cost Financial Summary Schema ( Tính toán chi phí)

* Bảng W\_Financial\_Summary\_Cost\_F

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Financial\_Summary\_Cost\_Id | INT | Mã định danh duy nhất cho mỗi bản tóm tắt chi phí tài chính. |
| Actual\_Units | INT | Số lượng thực tế của sản phẩm hoặc dịch vụ. |
| Actual\_Labor\_Cost | DECIMAL | Chi phí lao động thực tế phát sinh. |
| Actual\_Material\_Cost | DECIMAL | Chi phí nguyên liệu thực tế phát sinh. |
| Actual\_Machine\_Cost | DECIMAL | Chi phí máy móc thực tế phát sinh. |
| Actual\_Overhead\_Cost | DECIMAL | Chi phí chung thực tế phát sinh. |
| Budget\_Units | INT | Số lượng dự toán của sản phẩm hoặc dịch vụ. |
| Budget\_Labor\_Cost | DECIMAL | Chi phí lao động dự toán. |
| Budget\_Material\_Cost | DECIMAL | Chi phí nguyên liệu dự toán. |
| Budget\_Machine\_Cost | DECIMAL | Chi phí máy móc dự toán. |
| Budget\_Overhead\_Cost | DECIMAL | Chi phí chung dự toán. |
| Location\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Location\_D. |
| Machine\_Type\_Id | INT | Khóa ngoại liên kết đến bảng W\_Machine\_Type\_D. |
| Report\_Begin\_Date\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Time \_D. |
| Report\_End\_Date\_Id | INT | Khóa ngoại liên kết đến bảng W\_ Time \_D. |
| Sales\_Class\_Id | INT | Khóa ngoại liên kết đến bảng W\_Sales\_Class\_D. |

* Các bảng W\_Time\_D, W\_Location\_D, W\_Sales\_Class\_D,W\_Machine\_Type\_D

### Báo cáo doanh thu

* Mục tiêu:

Phân tích dữ kiệu bán hàng thực tế và ước tính của từng loại hàng tại các địa điểm

A diagram of a financial summary scheme

Description automatically generated

Hình 7. Sales Financial Summary Schema (Báo cáo doanh thu)

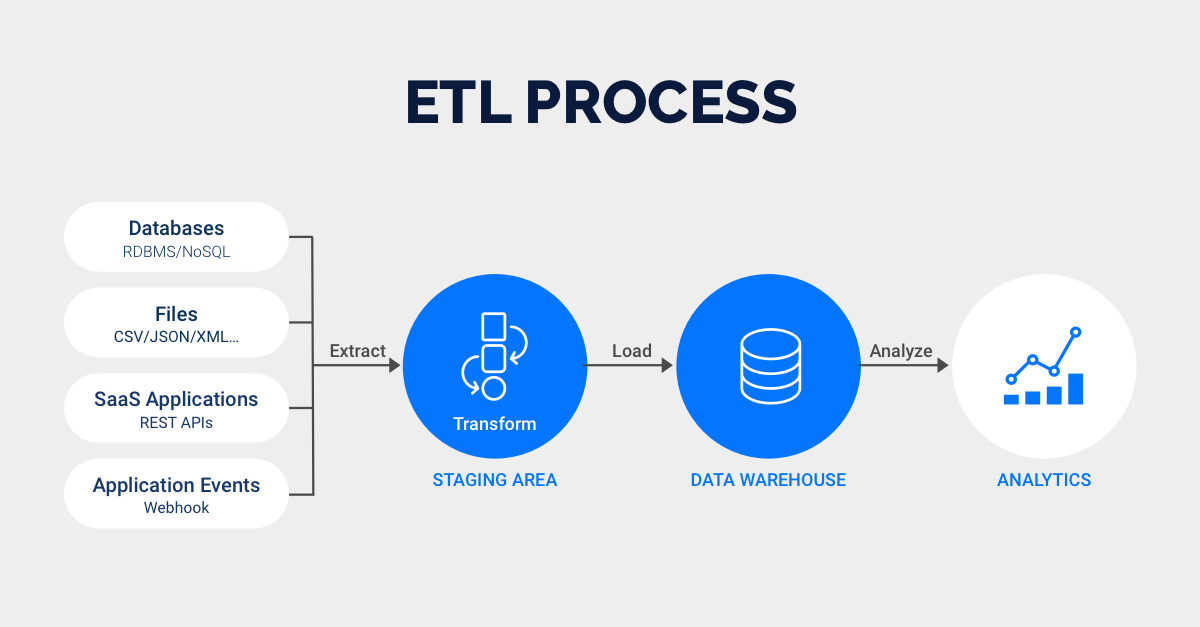
* Bảng W\_Financial\_Summary\_Sales\_F

|  |  |  |
| --- | --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** | **Mô Tả** |
| Financial\_Summary\_Sales\_Id | INT | Mã định danh duy nhất cho mỗi bản ghi tóm tắt tài chính bán hàng (Primary Key). |
| Actual\_Units | INT | Số lượng sản phẩm thực tế đã bán. |
| Actual\_Amount | DECIMAL | Số tiền bán hàng thực tế. |
| Forecast\_Units | INT | Số lượng sản phẩm dự báo sẽ bán. |
| Forecast\_Amount | DECIMAL | Số tiền bán hàng dự báo. |
| Location\_Id | INT | Khóa ngoại tham chiếu đến bảng W\_Location\_D |
| Report\_Begin\_Date\_Id | INT | Khóa ngoại tham chiếu đến bảng W\_Time\_D |
| Report\_End\_Date\_Id | INT | Khóa ngoại tham chiếu đến bảng W\_Time\_D |
| Sales\_Class\_Id | INT | Khóa ngoại tham chiếu đến bảng W\_Sales\_Class\_D |

* Các bảng Dim: W\_Time\_D, W\_Location\_D, W\_Sales\_Class\_D

# Thiết kế data pipeline

## Quy trình ETL



* **Extract (Trích Xuất)**: Giai đoạn đầu tiên trong quy trình ETL là trích xuất dữ liệu. Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau, bao gồm cơ sở dữ liệu, API, thu thập dữ liệu từ web, dữ liệu cảm biến, và các nền tảng mạng xã hội, cùng nhiều nguồn khác. Giai đoạn trích xuất yêu cầu xác định các nguồn dữ liệu liên quan và thu hồi dữ liệu cần thiết, có thể bao gồm việc truy vấn cơ sở dữ liệu, thực hiện các yêu cầu API, hoặc áp dụng các kỹ thuật thu thập dữ liệu từ web. Dữ liệu được trích xuất có thể là cấu trúc, bán cấu trúc hoặc phi cấu trúc, bao gồm văn bản, giá trị số, hình ảnh, hoặc các dạng dữ liệu khác.
* **Transform (Chuyển Đổi)**: Giai đoạn chuyển đổi trong quy trình ETL là rất quan trọng đối với khoa học dữ liệu. Giai đoạn này bao gồm các công việc làm sạch, tiền xử lý và thao tác dữ liệu đã được trích xuất để làm cho nó phù hợp cho việc phân tích và mô hình hóa. Các nhiệm vụ trong giai đoạn này có thể bao gồm làm sạch dữ liệu, bù đắp giá trị thiếu, chuẩn hóa dữ liệu, kỹ thuật đặc trưng, giảm chiều dữ liệu, và tổng hợp dữ liệu. Các nhà khoa học dữ liệu có thể sử dụng nhiều kỹ thuật và thuật toán khác nhau trong giai đoạn này, tùy thuộc vào tính chất của dữ liệu và mục tiêu của phân tích.
* **Load (Tải Dữ Liệu)**: Giai đoạn cuối cùng trong quy trình ETL cho khoa học dữ liệu là tải dữ liệu. Sau khi dữ liệu đã được chuyển đổi, nó cần được tải vào định dạng hoặc cấu trúc phù hợp cho việc phân tích tiếp theo. Quá trình này có thể liên quan đến việc lưu trữ dữ liệu trong cơ sở dữ liệu, hồ dữ liệu, hoặc một định dạng tệp cụ thể. Việc đảm bảo tính toàn vẹn và an ninh của dữ liệu trong suốt quá trình tải là rất quan trọng, cũng như thiết lập các thực tiễn quản trị dữ liệu thích hợp để tuân thủ các quy định và chính sách nội bộ.

## Triển khai ETL

Sử dụng python để triển khai ETL

### Tạo CSDL lưu trữ

* Lưu trữ tại PostgreSQL

A screenshot of a computer

Description automatically generated

Hình 8. Lược đồ schema tổng hợp

### Extract

#### ERP

Kết nối với CSDL để đọc và lấy dữ liệu từ các bảng

* Location

A screenshot of a computer program

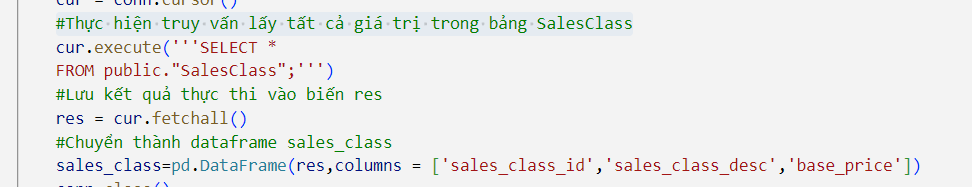
Description automatically generated

Sale Agent

A screenshot of a computer

Description automatically generated

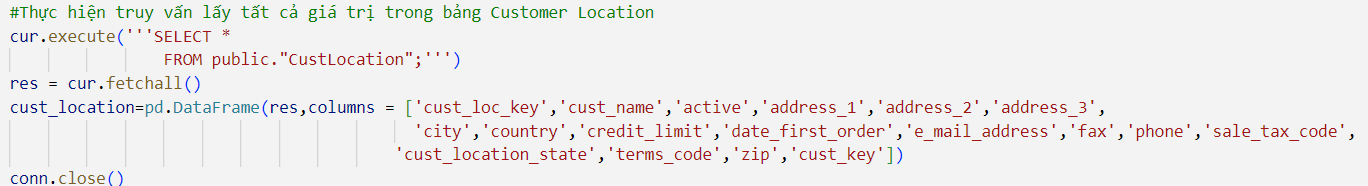
Sale Class



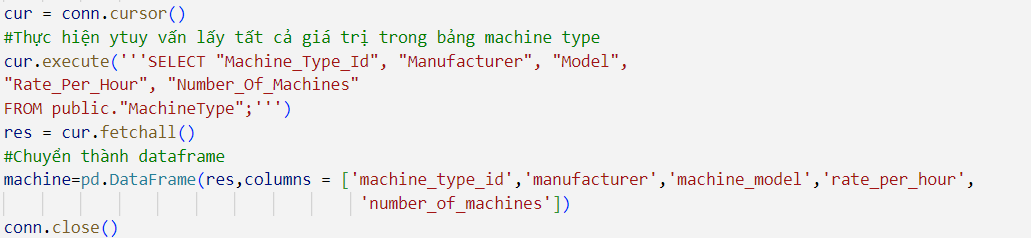
Customer



Customer Location



Machine Type

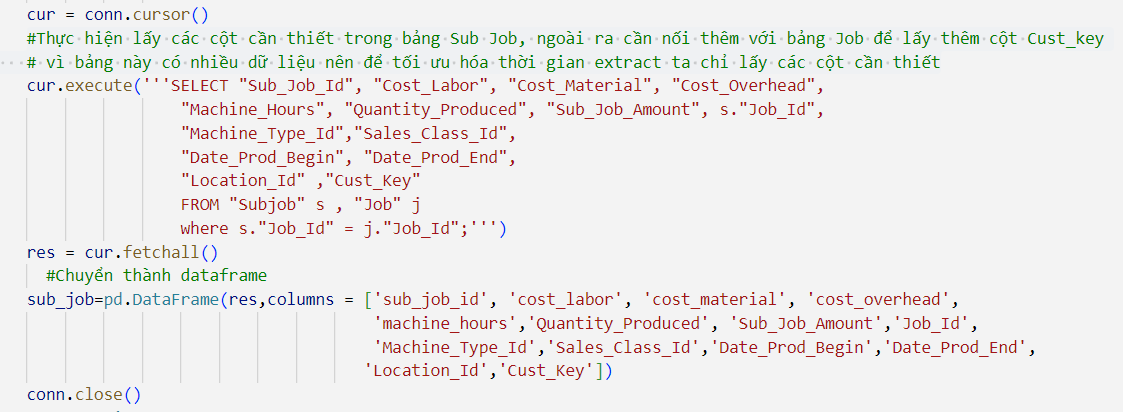


Job

A screenshot of a computer code

Description automatically generated

Sub job



Invoice

**A screen shot of a computer

Description automatically generated**

Shipment

A screenshot of a computer screen

Description automatically generated

#### Lead và Financial Summary

* Lấy dữ liệu từ file csv

A white rectangular object with a white background

Description automatically generated

A white rectangular object with red and blue text

Description automatically generated

Sale

A close-up of a person

Description automatically generated

### Transform

* Xóa các cột không cần thiết

A screen shot of a computer code

Description automatically generated

* Chuyển đổi kiểu dữ liệu các cột thời gian

A blurry image of a computer code

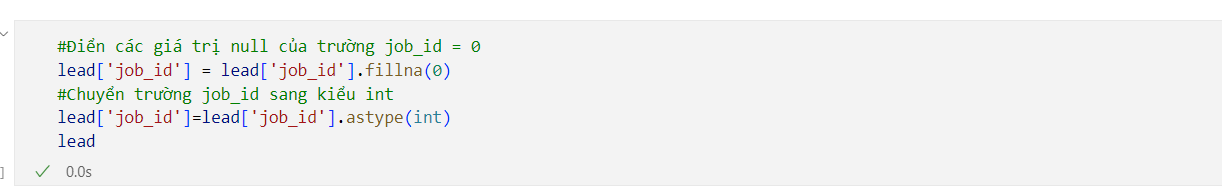
Description automatically generated with medium confidence

* Xóa các giá trị ngoại lai

A computer screen shot of a number

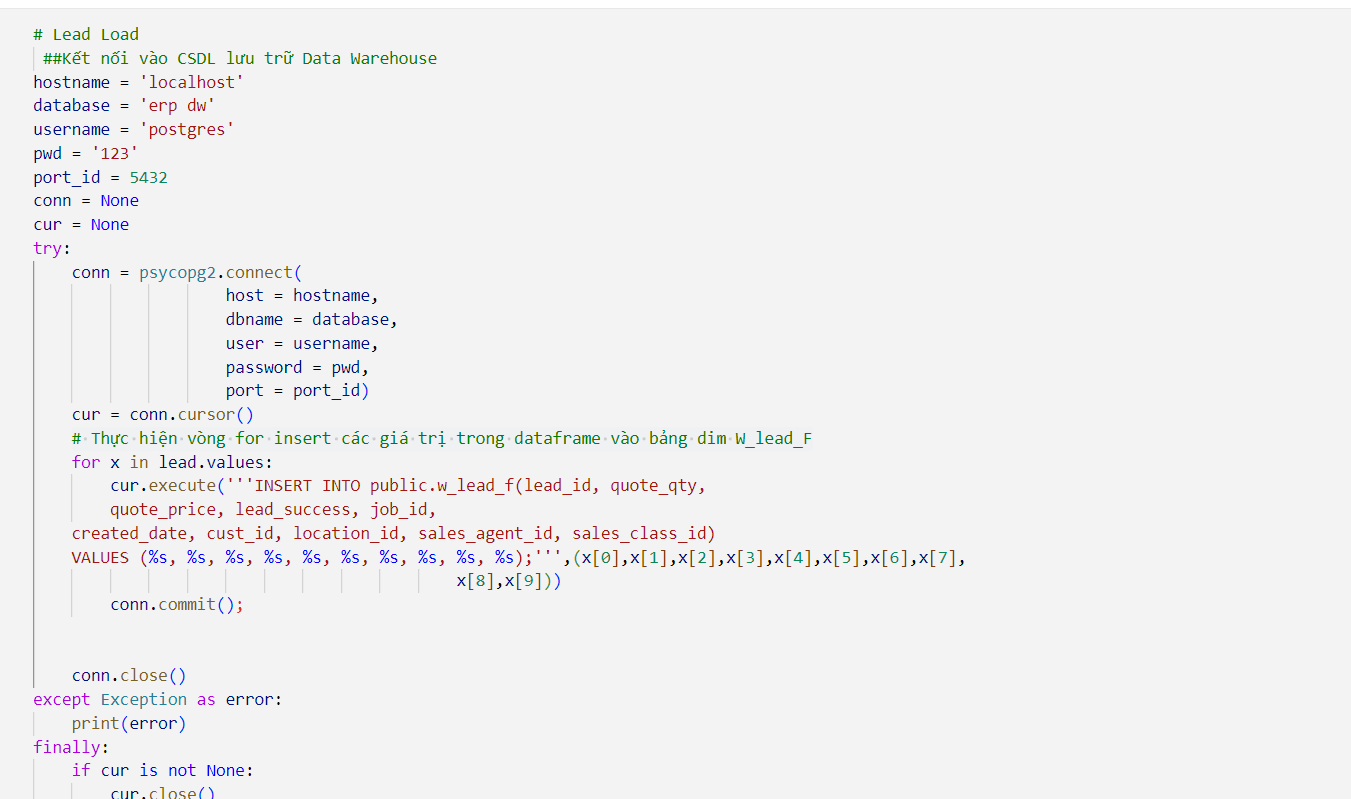
Description automatically generated with medium confidence

* Điền vào các giá trị null



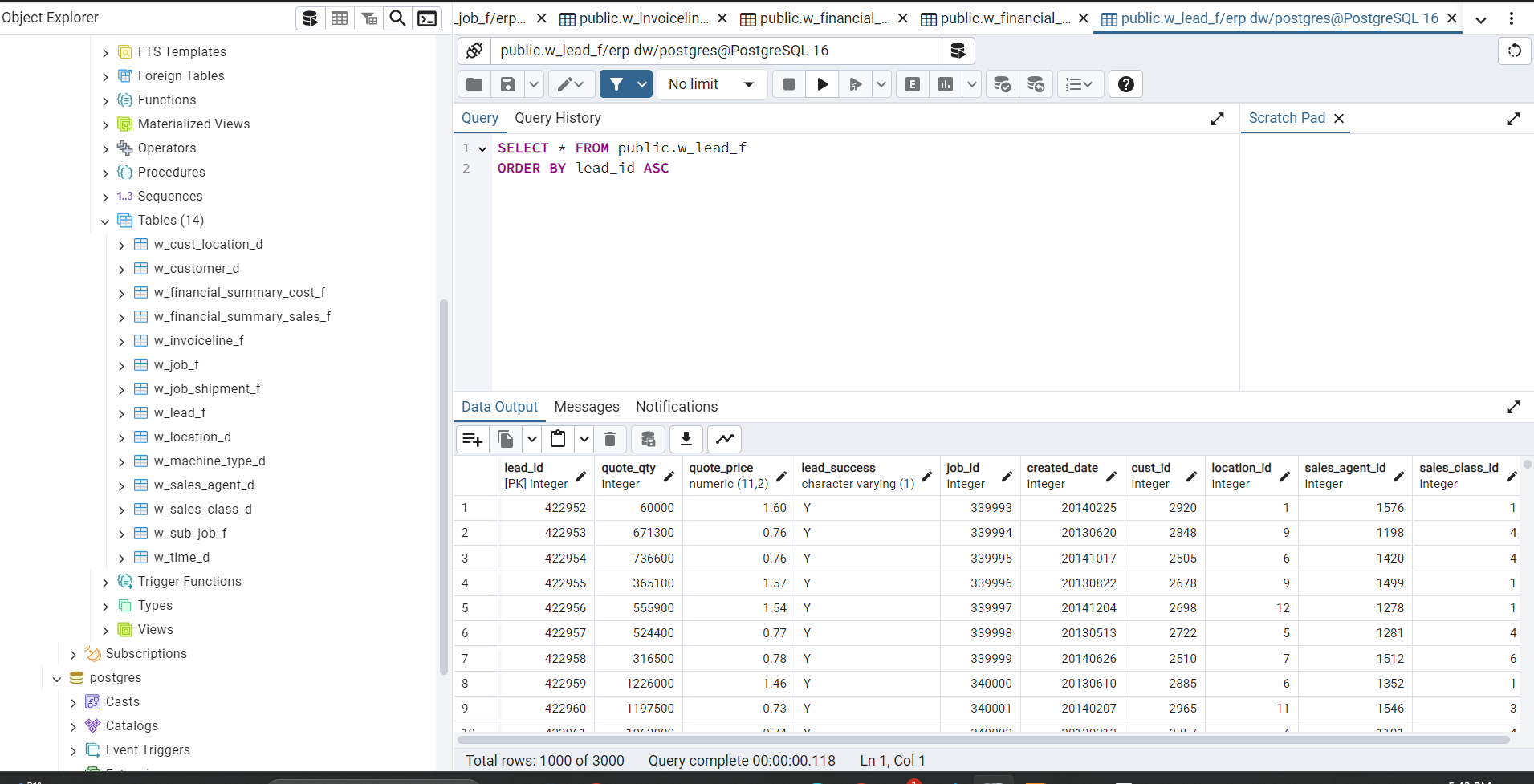
### Load

* Kết nối đến kho dữ liệu và tải dữ liệu lên
* Hàm tải lên của các bảng tương tự nhau với các cột tương ứng



## Kết quả quá trình ETL và truy vấn

* Dữ liệu đã được chèn vào các bảng phù hợp và có thể truy vấn



* Ứng dụng để truy vấn một số mục tiêu đề ra khi xây dựng kho dữ liệu:
* Quy trình marketing: Đánh giá hiệu suất chuyển đổi từ Lead thành khách hàng của từng đại lý bán hàng

A screenshot of a computer

Description automatically generated

* Quy trình sản xuất:
* Tính tỷ lệ giao hàng trễ theo từng khu vực theo các năm

A screenshot of a computer

Description automatically generated

* Quy trình bán hàng
  + Khách hàng mua hàng nhiều nhất

A screenshot of a computer

Description automatically generated

* + Loại hàng bán được nhiều nhất

A screenshot of a computer

Description automatically generated