**QUY TRÌNH TRIỂN KHAI KHO DỮ LIỆU**

1. **Mô tả bộ dữ liệu**

Dữ liệu được sử dụng là dữ liệu thô thu thập từ cơ sở dữ liệu bán hàng tại Công Ty TNHH Thương Mại & Dịch Vụ TSG trong 3 năm từ năm 2022 đến năm 2024. Bộ dữ liệu này bao gồm thông tin chi tiết về các giao dịch của khách hàng và các thông tin liên quan đến nhân viên bán hàng và chi nhánh. Tuy nhiên, dữ liệu đã được ẩn đi những thuộc tính chứa thông tin cá nhân của khách hàng, nhân viên nhằm đảm bảo tính bảo mật và quyền riêng tư.

Đầu tiên là file dữ liệu giao dịch chứa thông tin về các giao dịch của khách hàng tại trong 3 năm 2022, 2023, 2024. Dữ liệu được lưu trữ trong các bảng Excel và bao gồm 29129 bản ghi và các trường sau:

* **Thời gian**: Ngày thực hiện giao dịch.
* **Mã đơn hàng**: Mã duy nhất của mỗi giao dịch.
* **Mã khách hàng**: Mã định danh khách hàng thực hiện giao dịch.
* **Mã sản phẩm**: Mã của sản phẩm được khách hàng mua.
* **Số lượng**: Số lượng sản phẩm đã bán trong giao dịch.
* **Doanh thu**: Tổng doanh thu từ mỗi giao dịch (được tính bằng đơn giá nhân với số lượng bán).
* **Lợi nhuận**: Lợi nhuận từ mỗi giao dịch (được tính bằng doanh thu trừ đi tổng giá vốn).
* **Tổng giá vốn**: Giá trị vốn của sản phẩm bán ra (được tính bằng giá vốn nhân với số lượng bán).
* **Mã nhân viên**: Mã của nhân viên bán hàng thực hiện giao dịch.
* **Mã chi nhánh**: Mã chi nhánh nơi giao dịch được thực hiện, bao gồm chi nhánh tại các thành phố Hà Nội (CN HN), thành phố Hồ Chí Minh (CN HCM), và Đà Nẵng (CN DN)

A screenshot of a computer

AI-generated content may be incorrect.

Hình 14: File chứa thông tin giao dịch

Tiếp theo là file chứa thông tin các sản phẩm mà công ty cung cấp, gồm có 784 sản phẩm thuộc 4 nhóm sản phẩm: phần mềm, máy tính, dịch vụ và phụ kiện cùng với giá vốn của sản phẩm.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 15: File chứa thông tin các sản phẩm

Bên cạnh đó, file chứa thông tin khách hàng gồm có 786 khách hàng, bộ dữ liệu mô tả thông tin khách hàng tên khách hàng, số điện thoại, email, nhóm khách hàng tuy nhiên số điện thoại và email có chỉnh sửa để đảm bảo tính bảo mật thông tin khách hàng. File chứa thông tin nhân viên gồm 251 nhân viên gồm các thuộc tính mã nhân viên và tên nhân viên ở ba chi nhánh.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 16: File chứa thông tin khách hàng

A table with text on it

AI-generated content may be incorrect.

Hình 17: File chứa thông tin nhân viên

A close up of a sign

AI-generated content may be incorrect.

Hình 18: File chứa thông tin chi nhánh

Ngoài ra, file chứa thông tin chi nhánh mô tả các chi nhánh và trụ sở của công ty. Bộ dữ liệu xác định các chi nhánh của tổ chức theo mã, tên gọi và địa điểm.

1. **Triển khai xây kho dữ liệu**
   * 1. ***Triển khai kho dữ liệu***

Để sử dụng các dịch vụ như Amazon S3, Amazon Glue hay Amazon Redshift cần cấu hình cho một “IAM” (Identity and Access Management). Vai trò IAM (Quản lý danh tính và quyền truy cập) rất cần thiết khi làm việc với Amazon S3, AWS Glue và Amazon Redshift vì chúng xác định những dịch vụ hoặc người dùng nào được phép thực hiện chẳng hạn như đọc, ghi hoặc thực hiện hành động trên các dịch vụ AWS một cách an toàn và ít đặc quyền nhất.

Các dịch vụ AWS như Glue hoặc Redshift cần truy cập dữ liệu được lưu trữ trong S3. Thay vì nhúng thông tin xác thực, các dịch vụ AWS đảm nhận vai trò IAM với các quyền phù hợp. Khi cấu hình IAM phù hợp sẽ cho phép chúng ta đọc dữ liệu (ví dụ: tệp CSV, JSON, Parquet) từ s3://your-bucket, ghi dữ liệu đã chuyển đổi hoặc đã xử lý trở lại S3, Liệt kê hoặc duyệt nội dung của các bucket.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 19: Tạo các quyền trong IAM

Hình 19 mô tả đã trao quyền cho các chức năng như Redshift, S3 và Glue với đầy đủ chức năng truy cập như đọc, viết. Sau khi đã tạo một IAM để quản lý quyền truy cập tiếp theo đi tới chi tiết các hệ thống S3, Glue và Redshift.

* + 1. ***Cấu hình Amazon S3 (Data Lake Storage)***

Amazon S3 là một dịch vụ lưu trữ đối tượng có khả năng mở rộng, an toàn và tiết kiệm chi phí, đóng vai trò là kho dữ liệu để lưu trữ dữ liệu thô.

Tạo bucket S3 để thu thập dữ liệu thô, dữ liệu đã xử lý và nhật ký.

**A screenshot of a computer

AI-generated content may be incorrect.**

Hình 20: Tạo bucket trong Amazon S3

Thực hiện tạo bucket trong S3 bằng cách chọn “Create bucket” như trong hình 20 và hình ảnh sau khi tạo thành công một bucket có tên là “data-datns3”.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 21: Thực hiện import dữ liệu bằng chức năng upload

Để upload dữ liệu cá nhân, công ty hay tổ chức nên bucket s3 sử dụng chức năng “upload” như trong hình 21.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 22: Upload dữ liệu dạng .csv lên bucket S3

* + 1. ***Thiết lập AWS Glue***

AWS Glue là dịch vụ ETL (Trích xuất, Chuyển đổi, Tải) được quản lý hoàn toàn, tự động hóa các quy trình khám phá dữ liệu, suy luận lược đồ, lập danh mục và chuyển đổi dữ liệu.

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 23: Tạo crawler trong Amazon Glue

AWS Glue Crawler kết nối với bucket S3, đọc các tệp dữ liệu, suy ra lược đồ và tạo siêu dữ liệu (bảng) trong AWS Glue Data Catalog. Điều này giúp tự động hóa việc khám phá dữ liệu và quản lý lược đồ. Để thực hiện tạo crawler chọn như hình 23.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 24: Thực hiện chạy crawl dữ liệu từ S3

Sau khi đã tạo được một Crawler, bước tiếp theo thực hiện crawl dữ liệu từ bucket S3, để thực hiện điều đó chọn 1 như hình 24, lịch sử dữ liệu được crawl sẽ được lưu vào số 2.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 25: Databases sau khi đã được crawl dữ liệu từ S3

Từ hình 25, sau khi crawl dữ liệu từ bucket S3, khi đó Databases trong hệ thống Glue sẽ thể hiện các dữ liệu dạng bảng.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 26: Tạo kết nối giữa Glue và Redshift

Việc tạo kết nối giữa AWS Glue và Amazon Redshift là điều cần thiết khi muốn đọc dữ liệu từ Redshift, ghi dữ liệu vào Redshift hoặc sử dụng Redshift như một phần của quy trình ETL trong Glue. Trong AWS Glue, Kết nối là một tập hợp các thuộc tính kết nối được đặt tên (như tên máy chủ, cổng, tên người dùng và mật khẩu) được AWS Glue sử dụng để kết nối an toàn với các nguồn dữ liệu như Redshift.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 27: Tạo kết nối giữa Glue và Redshift

Sau khi kết nối AWS Glue Studio với Amazon Redshift thông qua “Connection”, AWS Glue Studio ETL Job trở thành công cụ trích xuất, chuyển đổi và tải dữ liệu (ETL) đến hoặc từ Redshift. Đây là lớp chuyển động và chuyển đổi dữ liệu. Dựa vào hình 27 chọn “Create job from a blank graph” để tạo một bộ ETL dữ liệu từ bucket S3 tới Redshift hoặc ngược lại từ Redshift qua Glue Studio để xử lý dữ liệu.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 28: Quy trình tạo một Job để trích xuất dữ liệu từ S3, xử lý dữ liệu và xuất dữ liệu qua Amazon Redshift

Sau khi quá trình ETL đã được thực hiện thành công, việc lập lịch (schedule) cho các công việc ETL giúp tự động hóa quy trình nạp dữ liệu từ Amazon S3 vào Redshift giúp đảm bảo rằng dữ liệu luôn được làm mới theo lịch trình định kỳ đồng thời giảm thiểu các sai sót có thể xảy ra trong quá trình xử lý thủ công cũng như khai thác được tối đa hiệu quả làm việc của hệ thống. Bằng cách này, dữ liệu trong Redshift luôn được cập nhật sẽ giúp Power BI truy xuất và tạo báo cáo chính xác, kịp thời, cung cấp thông tin phục vụ quyết định cho các bộ phận trong doanh nghiệp.

A close-up of a text

AI-generated content may be incorrect.

Hình 29: Cấu hình Schedule cho các job ETL

Để thiết lập schedule cho quá trình ETL trong AWS Glue nhóm tác giả thực hiện các bước cấu hình lịch trình sao cho các Glue Jobs có thể chạy tự động tại các thời điểm được xác định, phục vụ cho mục đích cập nhật dữ liệu định kỳ vào kho dữ liệu của hệ thống. Trong hình trên nhóm đã cấu hình lịch trình công việc ETL cho bảng Fact\_BanHang với tên "Load Fact\_BanHang from S3 to Redshift", chạy hàng tháng vào ngày cuối cùng của mỗi tháng lúc 12:00 AM. Mô tả lịch trình này là "ETL job chạy vào cuối mỗi tháng để nạp dữ liệu từ S3 vào Redshift, cập nhật bảng Fact\_BanHang", giúp dễ dàng nhận diện và quản lý công việc trong hệ thống.

* + 1. ***Thiết lập AWS Redshift***

Amazon Redshift là kho dữ liệu có thể mở rộng, được quản lý hoàn toàn để thực hiện phân tích và báo cáo, được tối ưu hóa cho các truy vấn phân tích quy mô lớn.

Bước đầu làm việc với Amazon Redshift Serverless (theo mục 1 hình 43), hai khái niệm chính là Serverless Dashboard và Workgroups. Chúng giúp quản lý, giám sát và sắp xếp cách kho dữ liệu Redshift Serverless chạy mà không cần tự quản lý cụm. Serverless Dashboard trong Amazon Redshift là giao diện trung tâm để theo dõi, quản lý và hiểu hiệu suất, mức sử dụng và chi phí của môi trường Redshift Serverless.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 30: Tạo Workgroup trong Amazon Redshift Serverless

Nhóm làm việc (Workgroups theo mục 2 hình 30) là ranh giới hoặc môi trường logic trong Redshift Serverless nơi người dùng và khối lượng công việc chạy truy vấn. Giống như một không gian làm việc độc lập cho một nhóm, ứng dụng hoặc trường hợp sử dụng trong môi trường Redshift Serverless.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 31: Tạo Workgroup trong Amazon Redshift Serverless

Để tạo được nhóm làm việc (Workgroups) thực hiện chọn theo ô màu vàng như hình 31.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 32: Query Editor v2 trong Amazon Redshift

Query Editor v2 trong Amazon Redshift là giao diện SQL mạnh mẽ dựa trên trình duyệt cho phép viết, chạy và trực quan hóa các truy vấn trực tiếp từ AWS Management Console mà không cần các công cụ bên ngoài như DBeaver, SQL Workbench hoặc các công cụ BI của bên thứ ba. Chức năng chính của Query Editor v2 trong Amazon Redshift cung cấp môi trường SQL tương tác dựa trên web, nơi người dùng có thể kết nối với cụm Redshift hoặc nhóm làm việc không có máy chủ. Viết và chạy truy vấn SQL, xem kết quả truy vấn và kế hoạch thực thi, khám phá lược đồ và bảng. Lưu và chia sẻ tập lệnh SQL.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 33: Hệ thống lưu trữ database và table trong Query Editor v2

Từ hình 33 chúng ta có cái nhìn tổng quan về cách lưu trữ database, table và cách thực hiện các câu lệnh SQL.

Câu lệnh SQL thực hiện tạo các bảng Dim và Fact trong Query Editor v2 như sau:

|  |
| --- |
| CREATE TABLE Dim\_ChiNhanh (      MaCN VARCHAR (50) PRIMARY KEY,      TenCN VARCHAR (100),      ThanhPho VARCHAR (100)  );  CREATE TABLE Dim\_KhachHang (      MaKH VARCHAR (50) PRIMARY KEY,      TenKH VARCHAR (100),      SDT VARCHAR (20),      Email VARCHAR (100),      NhomKH VARCHAR (50)  );  -- Create Dimension Table: Dim\_NhanVien  CREATE TABLE Dim\_NhanVien (      MaNV VARCHAR (50) PRIMARY KEY,      TenNV VARCHAR (100)  );  CREATE TABLE Dim\_SanPham (      MaSP VARCHAR (50) PRIMARY KEY,      TenSP VARCHAR (100),      NhomSP VARCHAR (50),  );  CREATE TABLE Dim\_ThoiGian (      MaTG INT PRIMARY KEY,      Ngay INT,      Thang INT,      Nam INT,      Quy INT,      ThoiGian DATE  );  CREATE TABLE Fact\_BanHang (      MaTG INT,      TenDH VARCHAR (100),      MaCN VARCHAR (50),      MaNV VARCHAR (50),      MaSP VARCHAR (50),      MaKH VARCHAR (50),      SoLuong INT,      DoanhThu BIGINT,      LoiNhuan BIGINT,      CONSTRAINT FK\_Fact\_BanHang\_Dim\_Time FOREIGN KEY (MaTG) REFERENCES Dim\_ThoiGian (MaTG),      CONSTRAINT FK\_Fact\_BanHang\_Dim\_ChiNhanh FOREIGN KEY (MaCN) REFERENCES Dim\_ChiNhanh (MaCN),      CONSTRAINT FK\_Fact\_BanHang\_Dim\_NhanVien FOREIGN KEY (MaNV) REFERENCES Dim\_NhanVien (MaNV),      CONSTRAINT FK\_Fact\_BanHang\_Dim\_Product FOREIGN KEY (MaSP) REFERENCES Dim\_SanPham (MaSP),      CONSTRAINT FK\_Fact\_BanHang\_Dim\_KhachHang FOREIGN KEY (MaKH) REFERENCES Dim\_KhachHang (MaKH)  ); |

A screenshot of a computer program

AI-generated content may be incorrect.

Hình 34: Lưu trữ các table đã tạo Query Editor v2

Dựa vào hình 34, sau khi thực hiện các dòng lệnh tạo bảng Dim và Fact, khi đó hiển thị danh sách các table nằm trong database dev.

* 1. **Thiết kế kho dữ liệu**
     1. ***Xây dựng các bảng Dim, Fact***

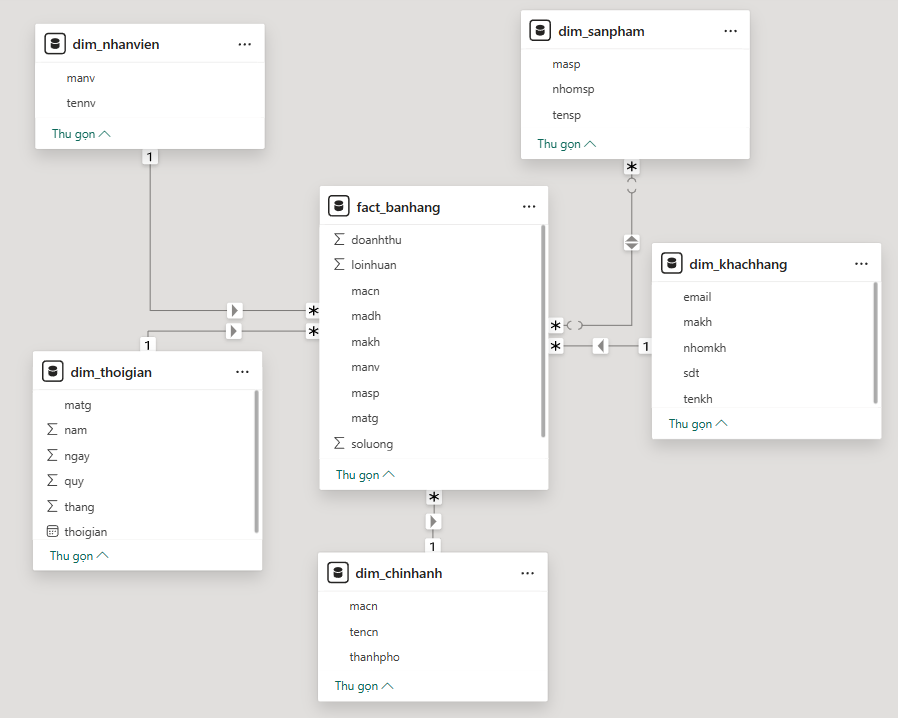
Các bảng chiều (dimension tables) được thiết kế để cung cấp ngữ cảnh cho các dữ liệu sự kiện trong bảng fact, giúp phân tích dữ liệu từ nhiều góc độ khác nhau như thời gian, sản phẩm, khách hàng, địa điểm. Cấu trúc bảng các bảng Dim như sau:

* Dim ChiNhanh: Lưu trữ thông tin địa điểm bán hàng, cụ thể là các tỉnh thành phố lớn tại Việt Nam như Hà Nội, Hồ Chí Minh và Đà Nẵng, bao gồm các trường:
* MaCN: Mã chi nhánh
* TenCN: Tên chi nhánh
* ThanhPho: Thành phố .
* Dim KhachHang: Lưu trữ thông tin của khách hàng mua sản phẩm, với độ phân giải tới mức cụ thể của khách hàng, bao gồm:
* MaKH: Mã khách hàng
* TenKH: Tên khách hàng
* SDT: Số điện thoại khách hàng
* Email: Email khách hàng
* NhomKH: Phân loại nhóm khách hàng
* Dim NhanVien: Lưu trữ thông tin của nhân viên trong công ty, với độ phân giải thông tin của thể của từng nhân viên, bao gồm:
* MaNV: Mã nhân viên
* TenNV: Tên nhân viên
* Dim SanPham: Lưu trữ thông tin của từng sản phẩm có trong công ty, với độ phân giải thông tin từng sản phẩm, bao gồm:
* MaSP: Mã sản phẩm
* TenSP: Tên sản phẩm
* NhomSP: Nhóm sản phẩm
* Dim ThoiGian: Lưu trữ thông tin thời gian bán từng sản phẩm trong công ty, với độ phân giải ngày tháng năm, bao gồm:
* MaTG: Mã thời gian
* Ngay: Ngày
* Thang: Tháng
* Nam: Năm
* Quy: Quý
* Thời Gian: Ngày tháng năm
  + 1. ***Xây dựng các bảng Fact***

Bảng Fact đóng vai trò trung tâm trong lược đồ hình sao, chứa dữ liệu định lượng chính được sử dụng cho các phân tích, cùng các khóa ngoại liên kết đến các bảng chiều. Cấu trúc bảng Fact như sau:

* **Khóa ngoại**:
* MaTG: Mã thời gian
* MaCN: Mã chi nhánh
* MaHD: Mã hóa đơn
* MaNV: Mã nhân viên
* MaSP: Mã sản phẩm
* MaKH: Mã khách hàng
* **Các giá trị đo lường (Measures)**:
* SoLuong: Số lượng sản phẩm trong đơn hàng
* DoanhThu: Doanh thu bán hàng
* LoiNhuan: Lợi nhuận công ty
  + 1. ***Thiết kế lược đồ dữ liệu phục vụ nhu cầu khai thác dữ liệu của người dùng***

Mô hình dữ liệu sử dụng lược đồ hình sao để tăng cường hiệu suất truy vấn bằng cách giảm số lượng liên kết, từ đó tăng tốc quá trình truy xuất dữ liệu. Ngoài ra, lược đồ hình sao có thiết kế đơn giản với bảng Fact trung tâm được kết nối trực tiếp với các bảng Dimension. Sự đơn giản này không chỉ giúp triển khai dễ dàng hơn mà còn cung cấp khả năng mở rộng, giúp dễ dàng thêm các bảng Dimension mới hoặc mở rộng bảng Fact khi các yêu cầu phân tích phát triển.



Hình 19: Lược đồ hình sao