**Mục Lục**

[**I.** **Xác định yêu cầu bài toán** 2](#_Toc131064719)

[*1.* ***Yêu cầu****:* 2](#_Toc131064720)

[***2.*** ***Phân tích yêu cầu*** 2](#_Toc131064721)

[***3.*** ***Các chỉ số*** 3](#_Toc131064722)

[**II.** **Dữ liệu sử dụng** 6](#_Toc131064723)

[***1.*** ***Nguồn dữ liệu*** 6](#_Toc131064724)

[***2.*** ***Các thuộc tính*** 6](#_Toc131064725)

[**III.** **Data Warehouse** 7](#_Toc131064726)

[***1.*** ***Cấu trúc và các bảng*** 7](#_Toc131064727)

[***a.*** ***Cấu trúc*** 7](#_Toc131064728)

[***b.*** ***Các bảng*** 7](#_Toc131064729)

[**2.** ***Sơ đồ Data Warehouse*** 8](#_Toc131064730)

[**IV.** **Quá trình ETL** 9](#_Toc131064731)

[***1.*** ***Giai đoạn 1:*** etl dữ liệu từ file csv vào bảng staging 9](#_Toc131064732)

[***2.*** ***Giai đoạn 2:*** etl dữ liệu từ bảng staging vào các bảng Dim và Fact 10](#_Toc131064733)

[**V.** **Dashboards** 13](#_Toc131064734)

[***1.*** ***Overview*** 13](#_Toc131064735)

[***2.*** ***Customer*** 15](#_Toc131064736)

[***3.*** ***Product*** 19](#_Toc131064737)

[**VI.** **Những khó khăn khi phân tích** 23](#_Toc131064738)

**Báo cáo phân tích tình hình hoạt động của Superstore**

**giai đoạn 2014 - 2018**

1. **Xác định yêu cầu bài toán**
2. ***Yêu cầu****:*

* xây dựng hệ thống báo cáo để có thể dễ dàng nắm bắt các thông tin và tình hình kinh doanh của doanh nghiệp Superstore

1. ***Phân tích yêu cầu***

* Trong từng thời kỳ hoạt động, kết quả kinh doanh luôn là mục tiêu của mọi doanh nghiệp, là điều kiện để tồn tại và phát triển của doanh nghiệp. Chính vì vậy chúng ta cần phải phân tích, đánh giá tình hình hoạt động kinh doanh để nắm bắt được tình hình kinh doanh và các yếu tố ảnh hưởng đến nó
* Từ yêu cầu trên, các nhà lãnh đạo muốn thấy rõ tình hình kinh doanh của doanh nghiệp Superstore và các thông tin, các yếu tố liên quan đến hoạt động kinh doanh của doanh nghiệp, như là khách hàng và sản phẩm kinh doanh
* Như vậy, chúng ta cần phân tích về hoạt động kinh doanh, khách hàng và sản phẩm kinh doanh của doanh nghiệp

1. ***Các chỉ số***

|  |  |
| --- | --- |
| Các câu hỏi kinh doanh | Các chỉ số |
| Tình hình hoạt động kinh doanh trong 4 năm qua như thế nào | Tổng số đơn hàng, doanh thu, lợi nhuận và số hàng đã bán trong 4 năm qua |
| Tình hình hoạt động kinh doanh của năm hiện tại so với năm trước như thế nào | Tổng số đơn hàng, doanh thu, lợi nhuận và số hàng đã bán của năm hiện tại so với năm trước |
| Hoạt động kinh doanh của doanh nghiệp hàng quý là như thế nào | Tổng số đơn hàng, số hàng đã bán, doanh thu và lợi nhuận mỗi quý |
| Hoạt động kinh doanh của doanh nghiệp hàng thang là như thế nào | Tổng số đơn hàng, số hàng đã bán, doanh thu và lợi nhuận mỗi tháng |
| Hoạt động kinh doanh của doanh nghiệp hàng tuần là như thế nào | Tổng số đơn hàng, số hàng đã bán, doanh thu và lợi nhuận hàng tuần |
| Doanh thu trong 3 tháng tháng tới sẽ như thế nào | Tổng doanh thu trong 3 tháng tiếp theo |
| Hoạt động kinh doanh của doanh nghiệp theo thời gian là như thế nào | Tổng doanh thu và lợi nhuận hàng tháng của năm hiện tại và năm trước |
| Hoạt động kinh doanh của doanh nghiệp tại các bang là như thế nào | Tổng số đơn hàng, số hàng đã bán, doanh thu và lợi nhuận tại từng tiểu bang |
| Những khách hàng nào là khách hàng tiềm năng của doanh nghiệp | Tổng doanh thu của các khách hàng có doanh thu cao nhất |
| Những loại sản phẩm nào được khách hàng ưa thích | Tổng số đơn đặt hàng của các danh mục và danh mục phụ sản phẩm |
| Khách hàng thường thích phương thức giao hàng nào | Tổng số đơn đặt hàng của các phương thức giao hàng |
| Các khách hàng đến từ những khu vực nào | Tổng số khách hàng của từng khu vực trên nước Mỹ |
| Nhóm khách hàng nào chiếm đa số | Tổng số khách hàng của từng phân khúc khách hàng |
| Các nhóm khách hàng có đóng góp như thế nào vào số đơn hàng | Phần trăm số đơn đặt hàng của từng phân khúc khách hàng trên tổng số đơn hàng |
| Các nhóm khách hàng có đóng góp như thế nào vào doanh thu | Phần trăm doanh thu của từng phân khúc khách hàng trên tổng doanh thu |
| Khách hàng thường mua hàng vào các tháng nào trong năm | Tổng số đơn đặt hàng và doanh thu của từng tháng |
| Khách hàng nào là khách hàng mua hàng nhiều nhất | Tổng số đơn đặt hàng của khách hàng có số đơn đặt hàng nhiều nhất |
| Khách hàng nào là khách hàng nhiều hàng nhất | Tổng số sản phẩm đã mua của khách hàng đã mua nhiều sản phẩm nhất |
| Khách hàng nào có doanh thu cao nhất cho doanh nghiệp | Tổng doanh thu của khách hàng có doanh thu cho doanh nghiệp cao nhất |
| Khách hàng nào đem lại nhiều lợi nhuận nhất cho doanh nghiệp | Tổng lợi nhuận của khách hàng có lợi nhuận cho doanh nghiệp cao nhất |
| Các danh mục và danh mục phụ đóng góp như thế nào vào doanh thu | Tổng doanh thu theo các danh mục và danh mục phụ sản phẩm |
| Những sản phẩm nào là sản phẩm có số lượng bán được nhiều nhất | Tổng số sản phẩm đã bán của các sản phẩm được bán với số lượng nhiều nhất |
| Những sản phẩm nào có doanh thu cao nhất | Tổng doanh thu của các sản phẩm có doanh thu cao nhất |
| Những sản phẩm nào mang lại lợi nhuận cao nhất | Tổng lợi nhuận của các sản phẩm có lợi nhuận cao nhất |
| Những sản phẩm nào không đem lại lợi nhuận (bị lỗ) | Tổng lợi nhuận của các sản phẩm có lợi nhuận thấp nhất |
| Các sản phẩm đã đóng góp như thế nào vào hoạt động kinh doanh của doanh nghiệp | Tổng số đơn hàng, số sản phẩm đã bán, doanh thu và lợi nhuận của từng sản phẩm |

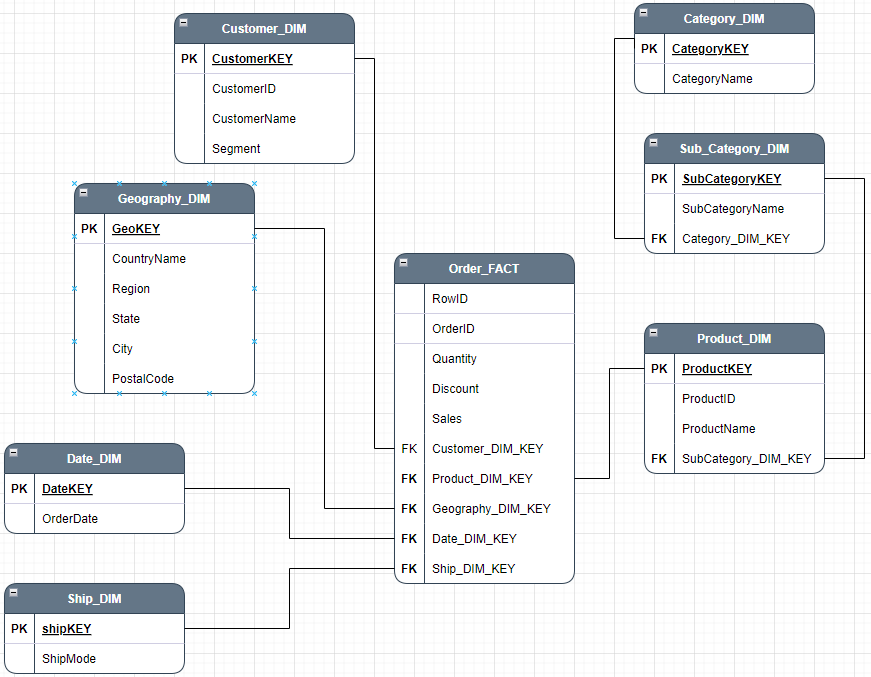
1. **Dữ liệu sử dụng**
2. ***Nguồn dữ liệu***

* Dữ liệu của doanh nghiệp Superstore được thu thập trên trang Kaggle trong giai đoạn 4 năm từ năm 2014 đến 2017

1. ***Các thuộc tính***

|  |  |
| --- | --- |
| Thuộc tính | Mô tả |
| Row ID | ID của mỗi hàng |
| Order ID | Order ID của mỗi khách hàng |
| Order Date | Ngày mua hàng |
| Ship Date | Ngày giao hàng |
| Ship Mode | Phương thức giao hàng |
| Customer ID | ID của khách hàng |
| Customer Name | Tên khách hàng |
| Segment | Loại khách hàng |
| Country | Quốc gia nơi khách hàng đến từ |
| City | Thành phố nơi khách hàng đến từ |
| State | Bang khách hàng đến từ |
| Postal Code | Mã bưu chính của thành phố |
| Region | Vùng khách hàng đến từ |
| Product ID | Mã sản phẩm |
| Product Name | Tên sản phẩm |
| Category | Danh mục của sản phẩm |
| Sub-Category | Danh mục phụ của sản phẩm |
| Sales | Doanh thu của sản phẩm |
| Quantity | Số lượng mua |
| Discount | Mức giảm giá |
| Profit | Lợi nhuận/lỗ phát sinh |

1. **Data Warehouse**
2. ***Cấu trúc và các bảng***
3. ***Cấu trúc***
   * Dữ liệu được thu thập ở dạng file csv nên cần được đưa vào data warehouse để quản lý nhằm hỗ trợ cho việc kinh doanh thông minh (BI), đặc biệt là phân tích
   * Dữ liệu từ file csv ban đầu sẽ được tải vào một staging table và sau đó sẽ được tải vào các bảng Dimension và Fact. Lược đồ sử dụng cho data warehouse này là lược đồ Snowflake giúp dữ liệu được chuẩn hóa hơn
4. ***Các bảng***
   * Customer\_DIM: chứa thông tin của khách hàng
   * Geography\_DIM: chứa thông tin về nơi ở của khách hàng
   * Date\_DIM: chứa các ngày thực hiện mua hàng
   * Ship\_DIM: chứa thông tin phương thức giao hàng
   * Category\_DIM: chứa thông tin danh mục sản phẩm
   * Sub\_Category\_DIM: chứa thông tin danh mục phụ sản phẩm, có khóa ngoại liên kết đến bảng Category\_DIM
   * Product\_DIM: chứa thông tin sản phẩm, có khóa ngoại liên kết đến bảng Sub\_Category\_DIM
   * Order\_FACT: chứa các thông tin giao dịch của khách hàng, có khóa chính đồng thời cũng là khóa ngoại liên kết đến các bảng Customer\_DIM, Geography\_DIM, Date\_DIM, Ship\_DIM, Product\_DIM
5. ***Sơ đồ Data Warehouse***



1. **Quá trình ETL**
2. ***Giai đoạn 1:*** etl dữ liệu từ file csv vào bảng staging

* Quá trình etl từ file csv vào bảng staging được sử dụng công nghệ Incremental Load với mục đích là các dữ liệu nào đã có thì sẽ không tải lại nữa, còn chưa có thì sẽ chỉ thêm dữ liệu đó vào

Diagram

Description automatically generated

* B1: tạo một Flat File Source cho file csv và một OLE DB Source cho bảng staging
* B2: nhóm các dữ liệu trong file csv lại để tránh bị trùng lặp bằng Aggregate và sắp xếp lại với Sort, đồng thời cũng sắp xếp dữ liệu cho bảng staging
* B3: Join dữ liệu từ csv và staging bằng Merge Join, kiểu join là Full Outer Join. Nếu dữ liệu chưa có trong staging thì sẽ Insert vào, nếu đã có thì sẽ không Insert thêm, tránh trùng dữ liệu
* B4: sử dụng Conditional Split để thực hiện logic trên
* B5: tải dữ liệu vào staging

1. ***Giai đoạn 2:*** etl dữ liệu từ bảng staging vào các bảng Dim và Fact

* Sau khi đã giai đoạn 1 hoàn tất, tiếp tục giai đoạn 2 là tải dữ liệu từ bảng staging vào các bảng Dim và Fact
* B1: Tải dữ liệu từ bảng staging vào bảng Customer bằng cách sử dụng các cộng nghệ, logic như ở giai đoạn 1

Diagram

Description automatically generated

* B2: Làm tương tự để có thể tải dữ liệu tương ứng vào các bảng tiếp theo
* B3:. Sau khi đã thiết kế các Data Flow để tải dữ liệu, sử dụng Sequence Container để chạy tất cả các Data Flow 1 lượt, tránh mất thời gian

A picture containing graphical user interface

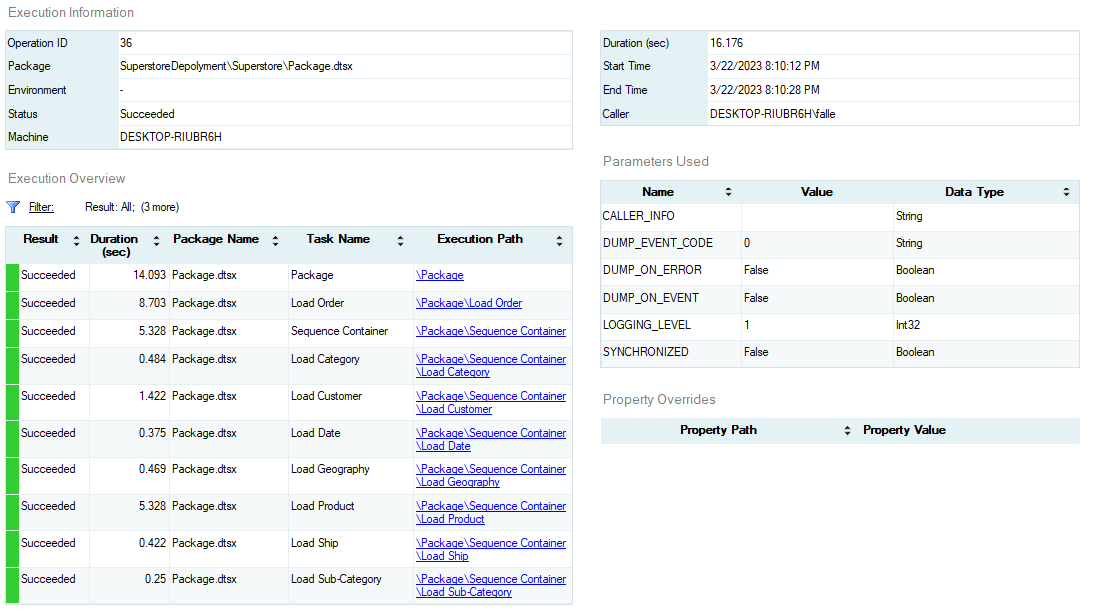
Description automatically generated

* B4: Tải dữ liệu vào bảng Fact bằng cách sử dụng các key của các bảng Dim match với các foreign key của bảng Fact

Graphical user interface

Description automatically generated

* B5: Thực hiện Deployment các SSIS Package để khi thực hiện thêm dữ liệu vào database thì không cần phải chạy SSIS, chỉ cần thực hiện trong SQL Server. Kết quả như hình dưới



1. **Dashboards**
2. ***Overview***

* Doanh thu là tổng thu nhập được tạo ra từ việc bán hàng hoặc dịch vu, lợi nhuận là khoản thu nhập còn lại sau khi đã trừ đi các khoản phỉa trả. Hai chỉ số này thể hiện tình hình hoạt động kinh doanh của doanh nghiệp
* Overview là dashboard không chỉ chứa các phân tích về các chỉ số trên (doanh thu, lợi nhuận) mà còn là đơn hàng và số sản phẩm đã bán. Các chỉ số này nhằm giúp các nhà điều hành thấy được tổng quan tình hình hoạt động kinh doanh của doanh nghiệp trong giai đoạn 4 năm từ 2014 – 2017, năm hiện tại so với năm trước, các tiểu bang và dự đoán trong tương lai

A picture containing graphical user interface

Description automatically generated

* Insights:
* Trong 4 năm qua, có tổng cộng 5008 đơn hàng, bán được 38000 sản phẩm, doanh thu 2.3 triệu đô và đem lại lợi nhuận là 286.36 ngàn đô
* Các chỉ số về số đơn hàng, số sản phẩm bán ra, doanh thu và lợi nhuận đều tăng so với năm trước. Trong đó
  + Số đơn đặt hàng là 1678 so với năm trước là 1314, tăng 28.39%
  + Số sản phẩm bán ra là 12472 so với năm trước là 9828, tăng 26.9%
  + Doanh thu đạt 733.13 ngàn đô so với năm trước là 609.13 ngàn đô, tăng 20.36%
  + Và lợi nhuận thu về là 93.42 ngàn đô so với năm trước là 81.77 ngàn đô, tăng 14.24%
* Qua các năm, doanh thu và lợi nhuận đều bắt đầu tăng ở quý 3 và cao nhất ở quý 4, có thể thấy cuối năm là dịp để người dân mua sắm. Có thể thực hiện các chương trình khuyến mại, tạo các chiến lược marketing để kích thích nhu cầu mua sắm của khách hàng
* Doanh thu và lợi nhuận phần lớn tập trung ở các bang California, Washington, New York, Texas và những bang nằm ở bờ đông nước Mỹ
* Texas là 1 trong các bang mang lại doanh thu lớn cho doanh nghiệp nhưng trong cả 4 năm, lợi nhuận đều bị lỗ vì có tới 53% sản phẩm bị lỗ được giao đến đây

1. ***Customer***

* Khách hàng là các cá nhân, tổ chức có nhu cầu mua sắm, sử dụng dịch vụ mà doanh nghiệp cung cấp và có mong muốn được thỏa mãn nhu cầu đó
* Theo Tom Peters, “Khách hàng là tài sản quan trọng nhất của một doanh nghiệp mặc dù không được ghi chép lại trong sổ sách”. Vì vậy việc phân tích để hiểu rõ hơn về khách hàng như sở thích, thói quen là điều cần thiết để tăng lợi nhuận
* Customer là dashboard chứa các phân tích về khách hàng nhằm tìm ra được các khách hàng tiềm năng, hiểu rõ hơn về khách hàng, phục vụ cho các chương trình quảng cáo, chính sách khuyến mại phù hợp cho các đối tượng khách hàng tương ứng

Chart

Description automatically generated

* Insights tổng quan:
* **B1**: Nhìn vào biểu đồ này ta có thể thấy sự chênh lệch giữa Standard Class và phần còn lại, hầu hết khách hàng đều thích phương thức này
* Standard Class có được gần 3000 đơn giao hàng
* Tiếp theo là Second Class với gần 1000 đơn giao hàng
* Tiệm cận Second Class là First Class với gần 800 đơn giao hàng
* Con số này chỉ khoảng gần 300 đối với Same Day
  + Có thể thấy các khách hàng đến từ các khu vực phía đông và phía tây, có thể thấy là bang California, Texas, New York, Pennsylvania và Illinois
* Đông nhất là khu vực phía Tây với 686 khách, tập trung chủ yếu ở California và Washington
* Tiếp đến là ở phía Tây với 674 khách, tập trung chủ yếu ở New York, Ohio và Pennsylvania
* Khu vực Trung Tâm đứng thứ 3 với 629 khách, phần đông đến từ Texas
* Cuối cùng là khu vực phia Nam với 512 khách, phân phối đều ở các bang, chỉ hơi nhỉnh hơn ở Florida
  + Các khách hàng đa số đều thích các mặt hàng (category) Office Supplies, đặc biệt là các danh mục phụ (sub-category) Binders và Paper
* Các sản phẩm thuộc Office Supplies chiếm hơn 50% số đơn đặt hàng, con số này chỉ hơn 20% với các sản phẩm thuộc Furniture và Technology
* Tuy nhiên doanh thu của các sản phẩm Technology lại nhỉnh hơn các sản phẩm Office Supplies và Funiture là khoảng 100 ngàn đô
  + Hầu hết các khách hàng đều nằm trong phân khúc các khách hàng là Consumer
* Consumer chiếm hơn 50% các đơn đặt hàng và doanh thu
* Corporate với hơn 30% số đơn hàng và doanh thu
* Home Office chiếm chỉ hơn 18% số đơn hàng và doanh thu
  + Khách hàng có thói quen mua hàng vào các dịp từ tháng 9 đến 12 trong các năm, là dịp mua sắm vào các lễ lớn ở Mỹ
  + **Các đề xuất**:
* Các khách hàng rất thích phương thức giao hàng Standard Class nên có thể sử dụng các chương trình khuyến mại, giảm giá,… khi khách hàng đặt giao bằng phương thức này
* Cần hướng sự tập trung sự quan tâm, các chiến lược chăm sóc khách hàng,… vào các khách hàng đến từ các bang như California, Washington, New York, Ohio, Pennsylvania và Texas vì phần đông các khách hàng đến từ đây
* Các sản phẩm Office Supplies được đặt rất nhiều và các khách hàng là Consumer chiếm đa số, có thể suy ra rằng đây là các khách hàng là dân văn phòng, có thể cải thiện hệ thống gợi ý sản phẩm trên website đối với các khách hàng này
* Có thể đưa ra các chiến lược marketing, quảng cáo sản phẩm hấp dẫn hơn nhân các sự kiện cuối năm vì hầu hết các sự kiện lớn ở Mỹ đều tập trung vào cuối năm (Halloween, Lễ Tạ Ơn, Black Friday, Giáng Sinh,…)
* Cần phải phát triển các chiến lược chăm sóc các khách hàng đã tin tưởng đặt mua hàng nhiều nhất, mua nhiều sản phẩm nhất, đem lại doanh thu và lợi nhuận nhiều nhất để cảm ơn và có thể giữ chân họ

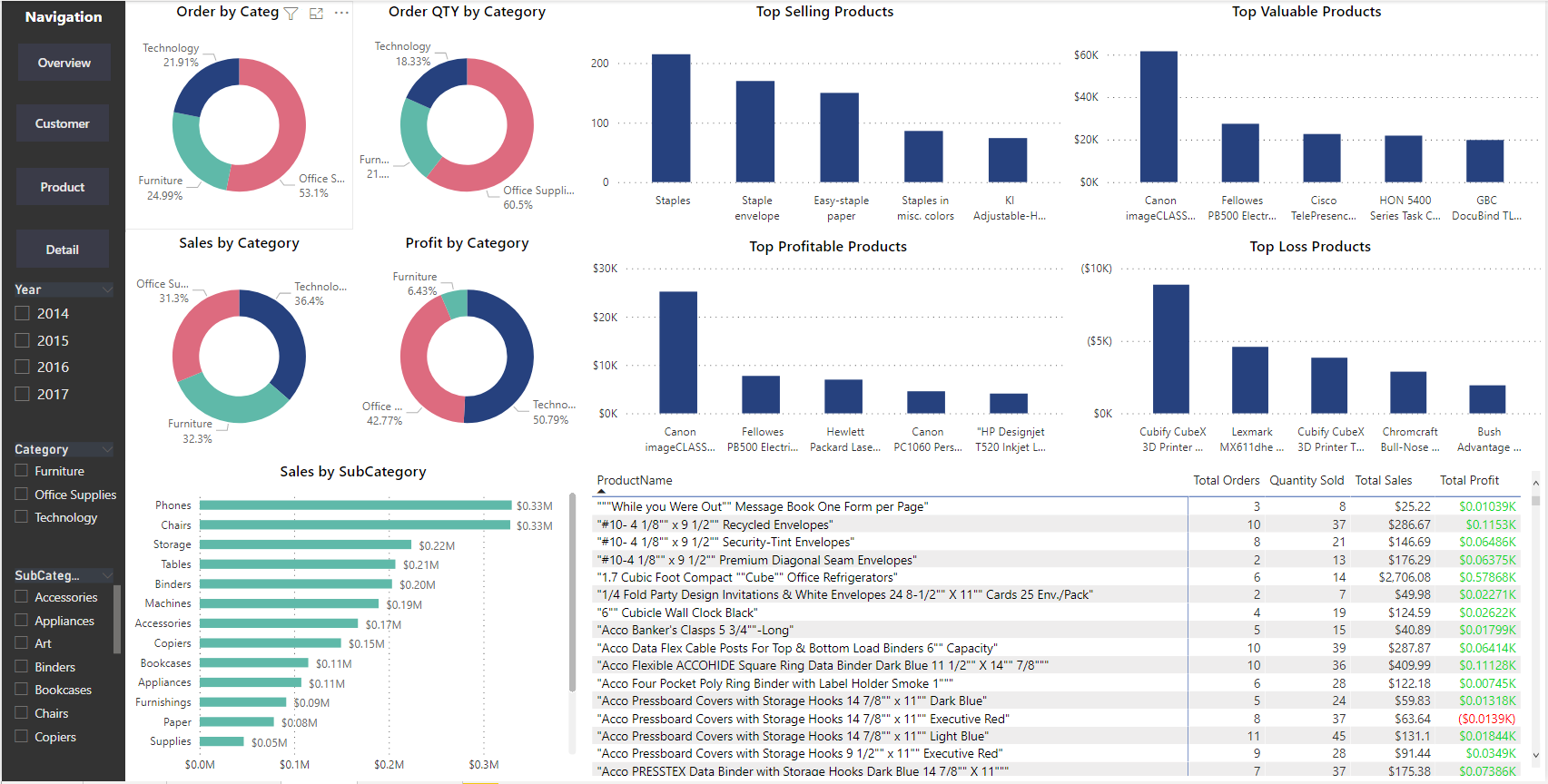
Graphical user interface, application

Description automatically generated

* Insight chi tiết:
* Số đơn hàng và số lượng mua hàng chiếm hơn 50%, hơn 50% doanh thu và gần 47% lợi nhuận
* Nhóm Consumer đều thích các mặt hàng Office Supplies
* Standard Class và Same Day là 2 phương thức được ưa thích và không được ưa thích nhất
* Same Day không được ưa thích ở tất cả các nhóm khách hàng, các tiểu bang và các tháng trong giai đoạn 4 năm kinh doanh
* Các đề xuất
* Consumer là nhóm khách hàng thân thiết, nên áp dụng các chương trình khách hàng thân thiết, các ưu đãi, khuyến mãi khi mua hàng, đặc biệt là Office Supplies
* Cần có những biện pháp như khảo sát ý kiến, đánh giá của khách hàng, cách hoạt động hiện tại của phương thức Same Day,… nhằm có những cải thiện cho tình trạng này

1. ***Product***

* Product là dashboard phân tích sản phẩm nhằm nắm được hiệu suất về kinh doanh của các loại mặt hàng, chi tiết hơn là có thể tìm ra các sản phẩm tiềm năng hoặc các sản phẩm hoạt động chưa được hiệu quả, giúp doanh nghiệp đẩy mạnh sản xuất, quảng cáo các sản phẩm tốt cũng như khắc phục các yếu điểm của các sản phẩm chưa tốt



* Insights tổng quan:
  + Trên tổng doanh thu và lợi nhuận thu được thì các mặt hàng Technology là dẫn đầu. Tuy các sản phẩm Office Supplies được mua nhiều nhưng chỉ là các vật dụng văn phòng phẩm, còn Technology và Funiture là các sản phẩm có giá trị hơn nên phân phối doanh thu ở các ba danh mục là khá đều
* Các sản phẩm Technology chiếm 36.4% doanh thu
* Tiếp theo là các sản phẩm Furniture chiếm 32.3% doanh thu
* Cuối cùng là các sản phẩm Office Supplies chiếm 31.3% doanh thu
  + Phones, Chairs, Storage, Tables là các danh mục phụ sản phẩm hàng đầu
  + Các dòng sản phẩm Staple (Staples, Staple envelope,…) là các sản phẩm được bán với số lượng nhiều nhất
  + “Canon imageCLASS 2200 Advanced Copier” và “Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind” là những sản phẩm có giá trị nhất, đem lại nguồn doanh thu và lợi nhuận nhiều nhất
  + Trong khi đó, “Cubify CubeX 3D Printer Double Head Print” và những sản phẩm Cubify khác là những sản phẩm bị lỗ nặng nhất
  + **Các đề xuất:**
* Cần phải kiểm tra lại chất lượng sản phẩm, mẫu mã, giá cả hoặc có thể là trải nghiệm không tốt của người dùng đối với các sản phẩm đang bị lỗ nhằm cải thiện các sản phẩm này vì bán các sản phẩm kém sẽ ảnh hướng đến uy tín của doanh nghiệp
* Đối với các sản phẩm đem lại doanh thu và lợi nhuận cao thì có thể thực hiện các chương trình khuyến mại đi kèm, giảm giá,… và cũng cần cải tiến thêm về sản phẩm để kích thích sự mua sắm của khách hàng
* Các sản phẩm được mua nhiều lần, mua với số lượng lớn thì có thể tăng giá trong giới hạn nếu có thể nhằm tối đa hóa lợi nhuận

Chart

Description automatically generated

* Insight chi tiết:
  + Office Supplies là các sản phẩm được bán chạy nhất, trong khi doanh thu và lợi nhuận thì Technology là dẫn đầu, đặc biệt là lợi nhuận
  + Phones và Chairs đóng góp rất nhiều vào doanh thu
  + Doanh thu càng cao thì lợi nhuận càng cao và ngược lại
  + Mức giảm giá cang thấp thì số lượng mua hàng, doanh thu, lợi nhuận càng cao và ngược lại
  + Lợi nhuận các sản phẩm Technology đến từ các thành phố như New York, Los Angeles, Seattle
  + Các sản phẩm Technology có mức discount thấp và doanh thu là cao nhất
  + Các sản phẩm có lợi nhuận cao thường có doanh thu cao và mức giảm giá thấp, các sản phẩm lỗ th ngược lại
  + **Các đề xuất**
* Chỉ nên áp dụng mức giảm giá từ 20% trở xuống
* Việc cải tiến, có các chiến dịch marketing tốt cho các sản phẩm Technology sẽ mang lại lợi nhuận lớn, đặc biệt là ở các thành phố trên
* Có thể nâng giá bán, giảm mức giảm giá, thực hiện các quảng cáo để bán chạy hơn, có doanh thu và lợi nhuận hơn đối với các sản phẩm bị lỗ

1. **Những khó khăn khi phân tích**

* Dữ liệu bị trùng lặp, khó phát hiện các điểm không hợp lý
* Dữ liệu không thể đáp ứng toàn bộ các câu hỏi phân tích, chỉ có thể chọn dữ liệu fit nhất có thể