**AdventureWorks Bike Company**

**Data Modeling**

Team Members: Ang LI

Bo Yang

Li Zhuang

Xiaomeng Xun

Ziang Cui

Submission Date: April 8, 2024

Course Name: Data Modeling

# Executive Summary

This project conducts an in-depth data modeling analysis for "Adventure Works" bicycle company, aiming to optimize data analysis and decision support processes through the construction of a star schema. We focused on the company's five main business processes: membership management, inventory management, manufacturing processes, purchasing, and sales activities. By transforming the original 3NF normalized tables into a star schema, we successfully simplified the data structure, enhanced the efficiency of data queries, and provided clear, intuitive data support for decision-makers.

# Assumptions

* The dataset is complete and up-to-date, reflecting the current business conditions of the company.
* All software and tools used in the analysis process are accurate and reliable.
* Market and economic conditions remain stable during the analysis period.

# Introduction

"Adventure Works" is a leading bicycle manufacturer with a wide range of products and a global market. The company leverages its rich data resources to optimize operational efficiency and market strategy. We chose this dataset because it provides a complex business environment suitable for demonstrating the depth and breadth of data modeling and analysis.

# Dataset Overview

The "Adventure Works" dataset is an intricate compilation initially comprising 71distinct tables that encapsulate a wide array of the company’s operational aspects. From this extensive collection, we have crafted a refined data model that consolidates the original tables into a more analytically conducive format. The transformation from the raw data into our final model was a comprehensive process that streamlined the numerous tables into a structured star schema suitable for sophisticated data analysis.

In the process of data preparation and transformation, we initially dealt with a complex set of 27 raw tables, each holding valuable information pertaining to different business processes such as customer management, sales, inventory management, purchasing, and manufacturing. These tables have been meticulously merged and expanded, resulting in a final model that includes a total of 5 fact tables at the heart of our star schema. These fact tables—centered around sales, purchasing, manufacturing, and inventory—are directly tied to the business's transactional data streams and are key to performance measurement and analysis.

Surrounding these fact tables are shared dimension tables, which have been reduced and optimized from the original set to a precise number that provides comprehensive descriptive contexts without redundancy. The shared dimension tables are fewer in number than the original set due to the consolidation of related attributes, eliminating duplicates, and ensuring a high level of data integrity and consistency. In total, the model leverages 17dimension tables that intersect multiple fact tables, providing a shared analytical framework that enhances our ability to perform multi-faceted queries.

This optimized arrangement of tables facilitates an efficient and robust analytical environment. The shared dimension tables not only conserve space and processing power but also enforce a single point of truth for dimensions like date, product details, and customer information across the entire data model. With this configuration, we are now poised to uncover insights through rich, multidimensional analyses, empowering "Adventure Works" with the intelligence to make informed, data-driven decisions.

A screenshot of a computer screen

Description automatically generated

# Business Process Analysis

## Member

* **Responsible:** Ziang Cui
* **Process** **Overview**

The business process represented by the Person, PhoneNumberType, StateProvince, and PersonAddressType and PersonType tables in the AdventureWorks database is membership management. This process encompasses person’s basic information, registration time, address information, etc.

* **Data Modeling**

A screenshot of a computer

Description automatically generated

* + **Fact Table Consolidation**

Initially, the model featured several separate tables: Person, PersonEmail, PerspnPhone and PersonAddress. I have merged the data into the fact table (Person) with the exception of the column which store the type of it. Such operation improves the performance of analyzing person-centric data and facilitates the segmentation of people given specific classifications.

* **Insights**
  + Most U.S. and Canadian users are concentrated on the West Coast and are found substantially in metropolitan areas.
  + There is no significant difference in the number of cell phones and home phones left behind by users.
  + User registrations were spread between 2017 and 2014, but there was a significant boost in the number of users between 2013 and 2014.
  + Since person did not provide much information about themselves, but according to the analysis of their surnames, the majority of users are still of European descent, with a small number of Asians. It is important to recognize that this indirect analysis is not necessarily accurate and needs to be corroborated by other data.

## Inventory

* **Responsible:** Ang Li
* **Process Overview:**

The business process represented by the Product, Location, ProductInventory, and ProductSubcategory tables in the AdventureWorks database is inventory management. This process encompasses tracking and managing the stock of products, the locations where these products are stored, and the categorization of products into various groups or subgroups. Inventory management involves ensuring that the right amount of product is available in the right location at the right time, managing the flow of goods from suppliers to warehouses and then to customers, and maintaining optimal inventory levels to meet customer demand while minimizing cost and maximizing efficiency.

* **Granularity:**

Product: Each record in the Product table represents a single product offered by the company. The granularity is at the product level.

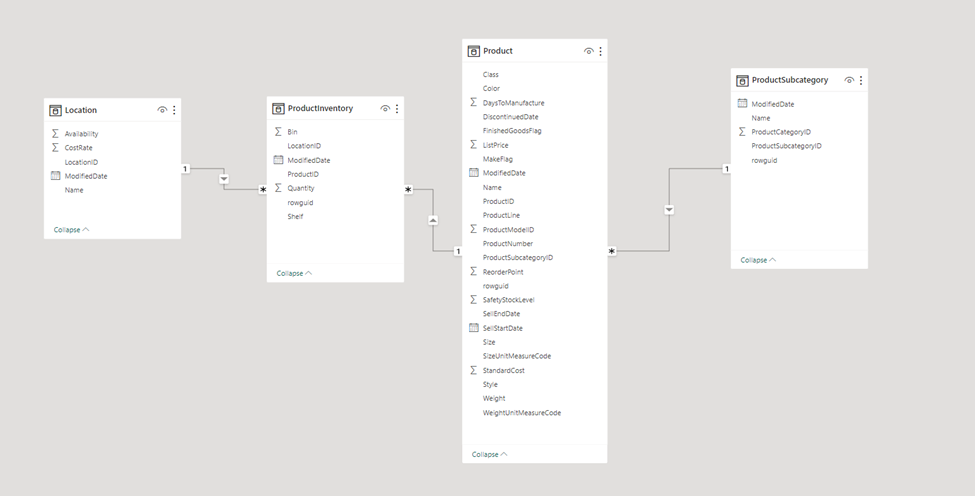
Location: Each record represents a specific location or warehouse where products are stored. The granularity is at the location or warehouse level.

ProductInventory: Each record in the ProductInventory table represents the quantity of a specific product at a specific location, making the granularity at the level of the product-location combination.

ProductSubcategory: Each record represents a single category or subcategory under which products are classified. The granularity is at the category or subcategory level.

* **Data Modeling:**

After loading these tables into Power BI, the original relationship can be connected as follows:



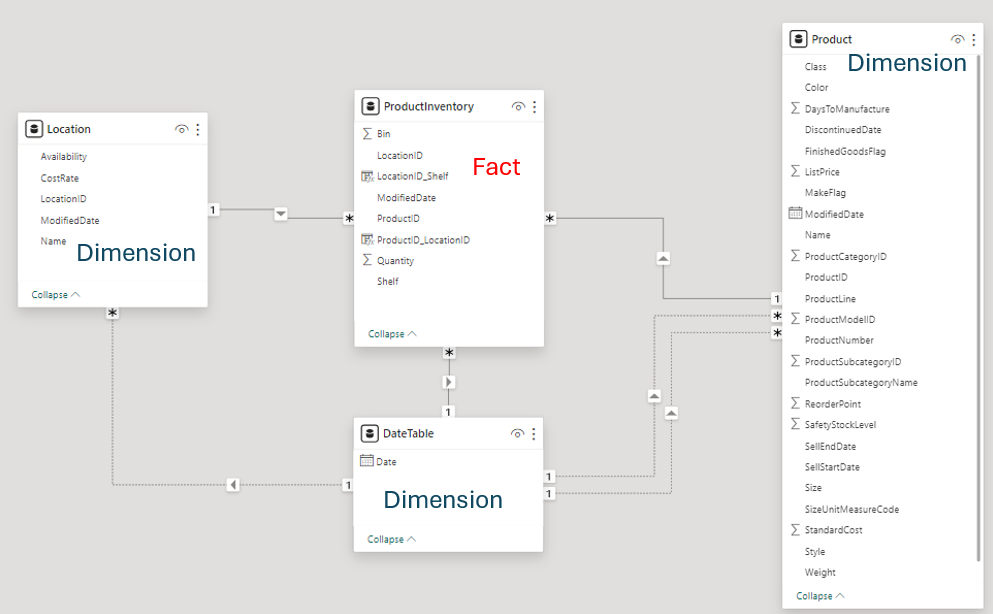
Location and ProductInventory connected by LocationID, ProductInventory and Product connected by ProductID, Product and ProductSubcategory connected by ProductSubcategoryID.

Then begin the process to transform these models into a star schema model.

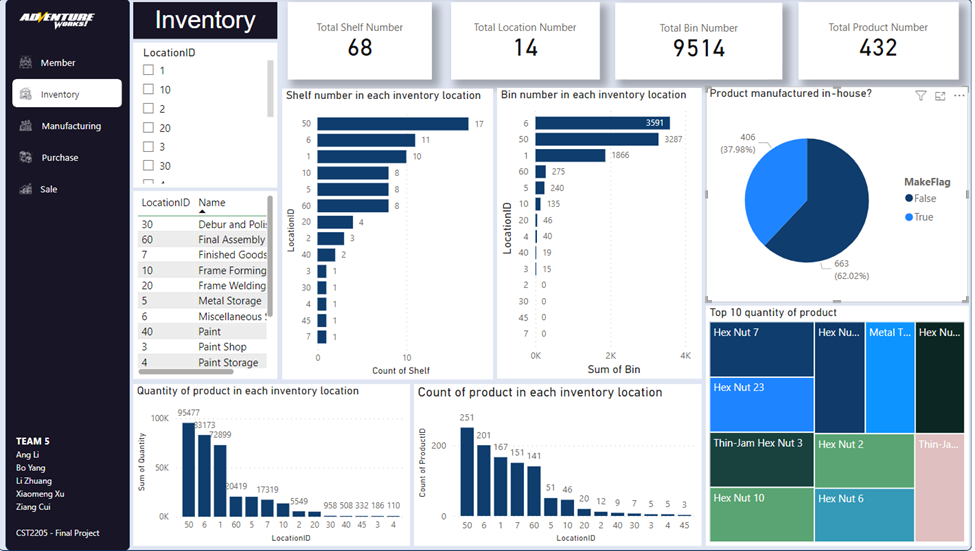
First, rowguid was deleted in all the tables. The rowguid is not useable in the models.

Second, the ProductSubcategory table was merged into Product table. And ProductCategoryID, ProductSubcategoryName column was added to Product table.

Then, a date table was created to connect all the date attributes.



* **Dashboard:**



From this Dashboard, we can easily get some important information:

1. In total, there are 14 Inventory locations, 68 shelves, 9514 bins, and 432 products.
2. In total, 23.66% products are manufactured in-house.
3. Location 50 has the most Shelves (17). Location 7, 45, 4, 30, 3 has the least Shelves (1).
4. Location 6 has the most Bins (3591). Location 2, 30, 45, 7 has the least Bins (0).
5. Location 50 has the most quantity of product (95477). Location 4 has the least quantity of product (110).
6. Location 50 has the most count of product (251). Location 45 has the least count of product (3).
7. Top 10 quantity of product are these: Hex Nut 7, Hex Nut 23, Thin-Jam Hex Nut 3, Nex Nut 10, Hex Nut 4, Metal Tread Plate, Hex Nut 18, Hex Nut 2, Hex Nut 6, Thin-Jam Hex Nut 7.

* **Insights:**

After interactive with the dashboard, we can find some insights:

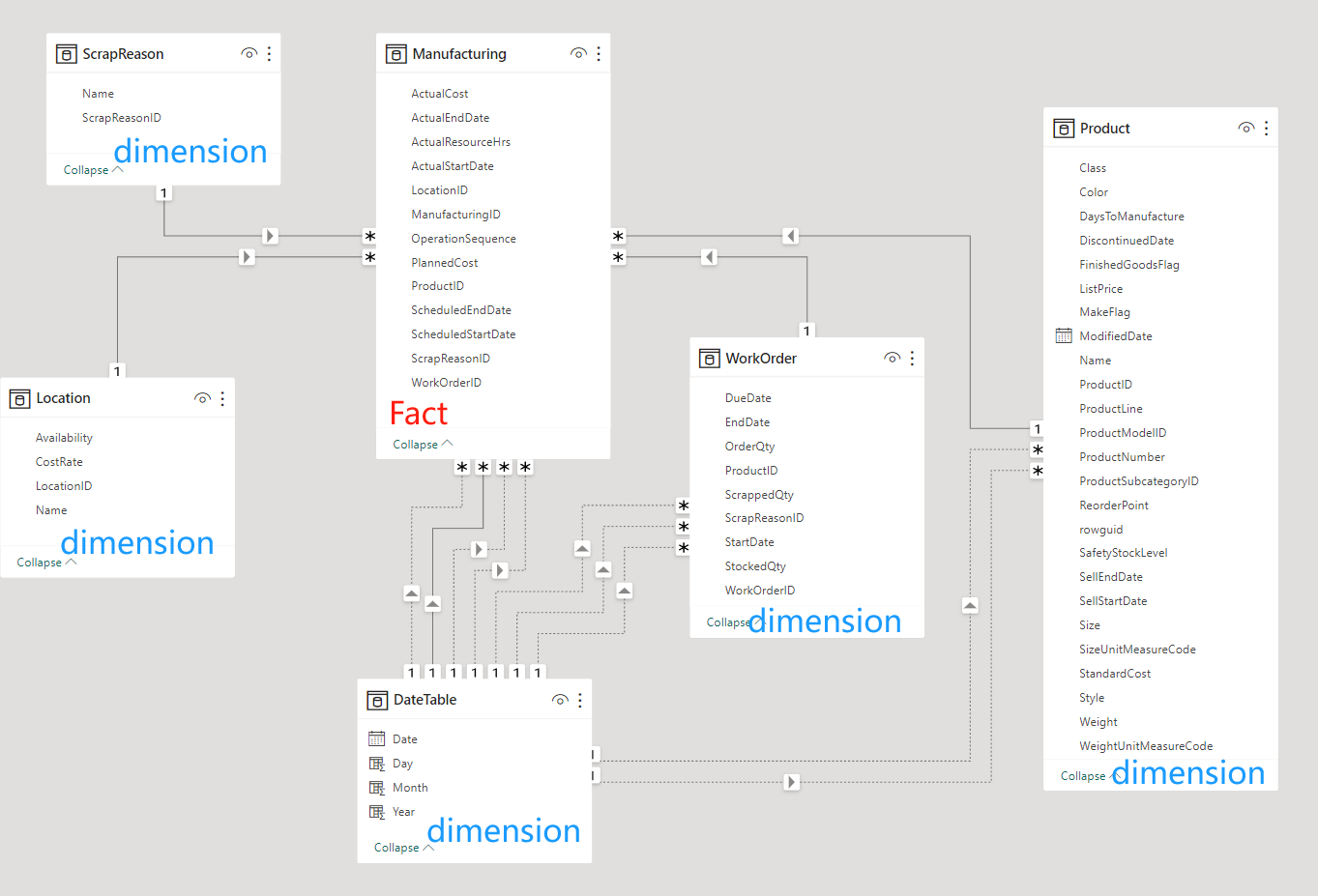
1. All top 10 quantity of product are stored in the Location 1, 6, 50, which is Tool Crib, Miscellaneous Storage, and Subassembly. At the same time, Location 1, 6, 50 are the top 3 locations with the most shelves, bins, quantity of product, and count of product. It indicates the Location 1, 6, 50 are the core locations for the inventory.
2. For the Location 50, it has the most shelves, quantity of product and count of product. But the bins number is the second. For the Location 6, it has the 2nd highest number of shelves, quantity of product and count f product. But the bins number is the most. If we assume that the shelves and bins are the same size, then it may because of many products in Location 50 may be not need a bin or cannot hold in a bin. So the number of Bin is smaller than Location 6.
3. For Location 5, 10, 60, they have the same shelves, but Location 60 has more quantity of product and count of product. It may because of the product size in Location 60 is smaller than in Location 5 and 10.
4. For Location 7, it has the least shelve (1) and no bin. But the quantity of product ranks the 6th and count of product ranks the 4th. It may because all the products in Location 7 don’t need bin, and the size of the products are small.
5. The products in Location 2, 3, 4 are all not manufactured in-house (purchased). The products in Location 30, 45 are all manufactured in-house.

## Manufacturing

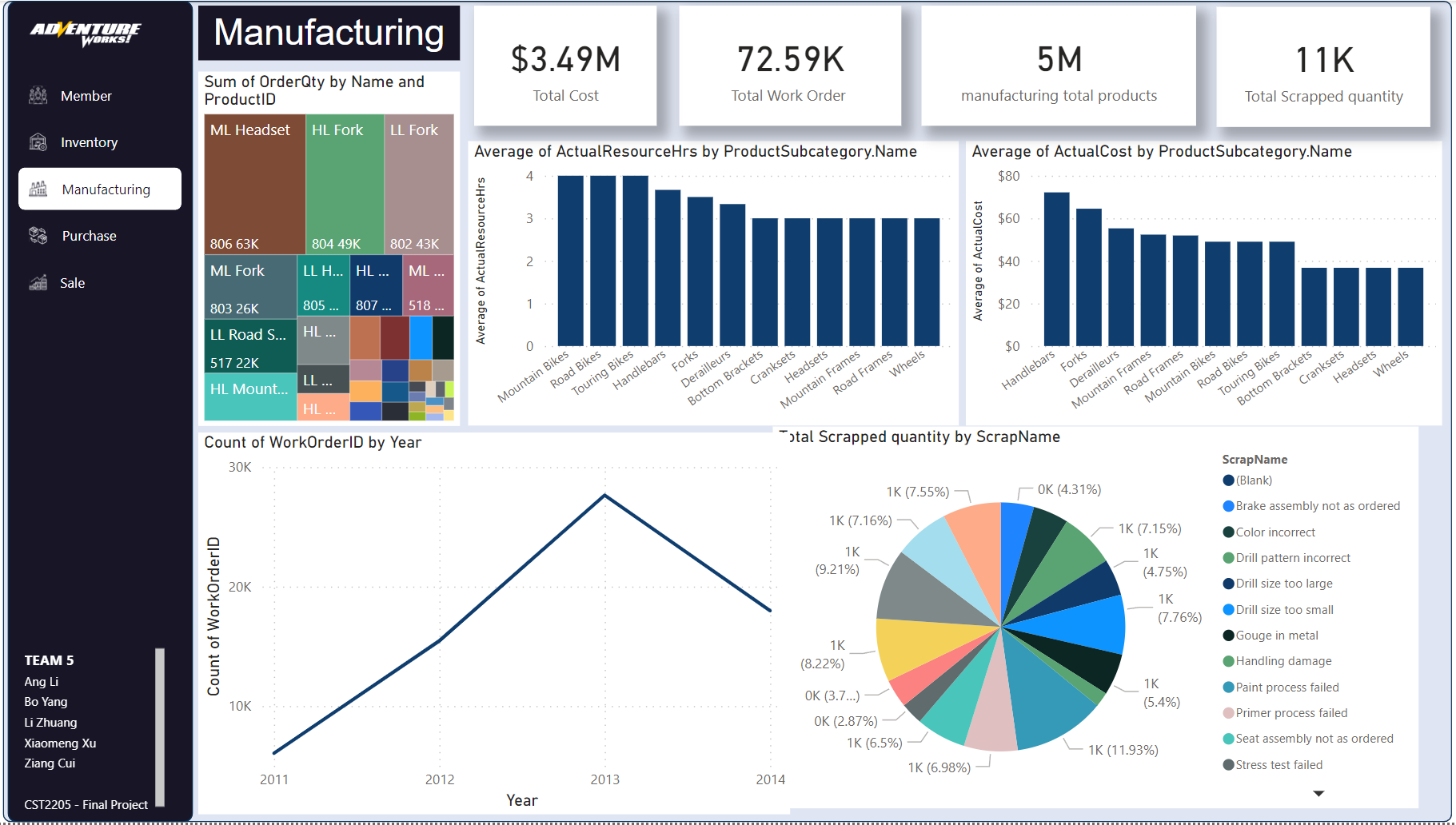
* **Responsible**: Xiaomeng Xu
* **Process Overview**

The business process is about manufacturing bicycles. The company places the work orders and factory receive the work orders and do their job to manufacture bicycles and track scrapped materials.

* **Data Model**



* + A start schema
  + Fact table:
    - Manufacturing
  + Dimension table:
    - WorkOrder
    - Product
    - Location
    - ScrapReason
    - Date
  + Relationship
    - Manufacturing table to WorkOrder table is many to one
    - Manufacturing table to Product table is many to one
    - Manufacturing table to Location table is many to one
    - Manufacturing table to ScrapReason table is many to one
    - Manufacturing table to Date table is many to one
* **Key Findings and Charts**



* + Time period: 2011 - 2014
  + Key performance indicators:
    - Total Cost: 3.49 M
    - Total work orders: 72.59 K
    - Manufacturing total products: 5 M
    - Total Scrapped quantity: 11 K
    - The year with most work orders: 2013 (27, 652)
  + As we can see in the dashboard, the top four products we manufactured are ML headset, HL Fork, LL Fork and ML Fork. A headset is the set of components on a bicycle that provides a rotatable interface between the bicycle fork and the head tube of a bicycle frame, and a fork is used for holding the front wheel. Basically, these are two very important parts of bicycle. Therefore, we should consider the quality of these two kinds of products first to ensure our bicycle having solid foundation.
  + Products with the top three average costs are Handlebars, Forks and Derailleurs. So, if we aim to reduce the overall cost of bicycle, we need to consider reduce the cost of these three products. However, we cannot compromise on quality to cut costs, especially when it comes to Forks
  + The accurate scrap rate = 10,651 / 4,507,721 = 0.24%, which is a positive indicator. It means we have efficient production processes and effective quality control. But we still have something to notice, as the pie chart shows the top three scrapped reasons are Paint process failed, Trim length too long, and Wheel misaligned. We can improve these manufacturing process in the future.

## Purchase

* **Responsible**: Li Zhuang
* **Process Overview**  
  The Fact table of Purchase Process captures the purchasing order process at "Adventure Works," encompassing the complete journey from order initiation to product delivery. It meticulously records each order's creation date, due fulfillment time, actual quantities received, and financial details including pricing, taxes, and shipping costs. Additionally, it tracks the order's status, responsible employees, and supplier information, providing comprehensive data support for purchasing efficiency, cost control, and supply chain management.
* **Data Modeling**

The data model for the purchasing process at "Adventure Works" has been meticulously architected to facilitate in-depth analysis of supplier interactions and purchasing efficiencies. By consolidating the original nine tables related to the purchasing process into a single fact table and a streamlined set of dimension tables, we have established a robust framework for our analytical methods.

A screenshot of a computer

Description automatically generated

**1. Fact Table Consolidation**:

Initially, the model featured two separate fact tables: **PurchaseOrderHeader** and **PurchaseOrderDetail**. The **PurchaseOrderHeader** table provided a snapshot of the orders, encapsulating critical order metadata such as identification numbers, timestamps, financials, vendor information, and employee details. Meanwhile, the **PurchaseOrderDetail** table broke down each order into line items, detailing product-specific transactions, quantities, and line totals. To enhance the model’s analytical potential, these fact tables have been merged into one encompassing fact table. This allows for an integrated view of each purchase, bridging high-level order summaries with detailed line item data.

**2. Dimension Table Refinement**:

The dimension tables within the purchasing model were refined from the original seven to a smaller, more cohesive set.

* Product-related data were methodically integrated by merging the **Product**, **ProductCategory**, and **ProductSubCategory** tables using their respective identifiers: **ProductID**, **ProductCategoryID**, and **ProductSubCategoryID**. This merge resulted in a unified Product Dimension table that encapsulates the hierarchical relationship between the products and their classifications.
* Vendor information was streamlined by combining the **Vendor** and **PurchaseVendor** tables into one cohesive table. This consolidation was achieved by merging records based on the **VendorID**, ensuring a single, comprehensive view of the vendor information within the procurement system.

**3.** **Transform Many: Many to 1: Many**

To facilitate a star schema and resolve the many-to-many relationship between **PurchaseVendor** and **PurchaseOrderDetail**, the **ProductID** and **VendorID** from both tables are merged to form a new column. This reconstruction creates a unique identifier for each product-vendor pair and transforms the relationship into a one-to-many model, aligning with the star schema design principles.

**4. Add Date table as Date dimension**

In the absence of a pre-existing **Date** dimension within the original tables, we capitalized on Power BI's built-in ‘Calendar’ function to generate a comprehensive Date dimension. This newly created Date dimension table has been integrated into our model, providing a connection point for all date-related fields present in the fact table. This ensures that any temporal analysis can be consistently applied across various date fields such as order dates, ship dates, and due dates, enabling a holistic view of time-based trends and patterns within the purchasing process.

**The final Star Schema model for Purchase Process is as follows:**

A screenshot of a computer

Description automatically generated

* **Key Findings and Charts**

"Adventure Works" has a strategic procurement division that categorizes purchases based on their end-use within the company’s operations. The purchasing categories service two main areas: the production line and direct consumer sales. Components such as pedals, brakes, and rims are procured primarily for the manufacturing process, whereas accessories and clothing items like helmets, water bottles, and jerseys are intended for retail to end customers.

A screenshot of a computer

Description automatically generated

**The "Blank" Category**

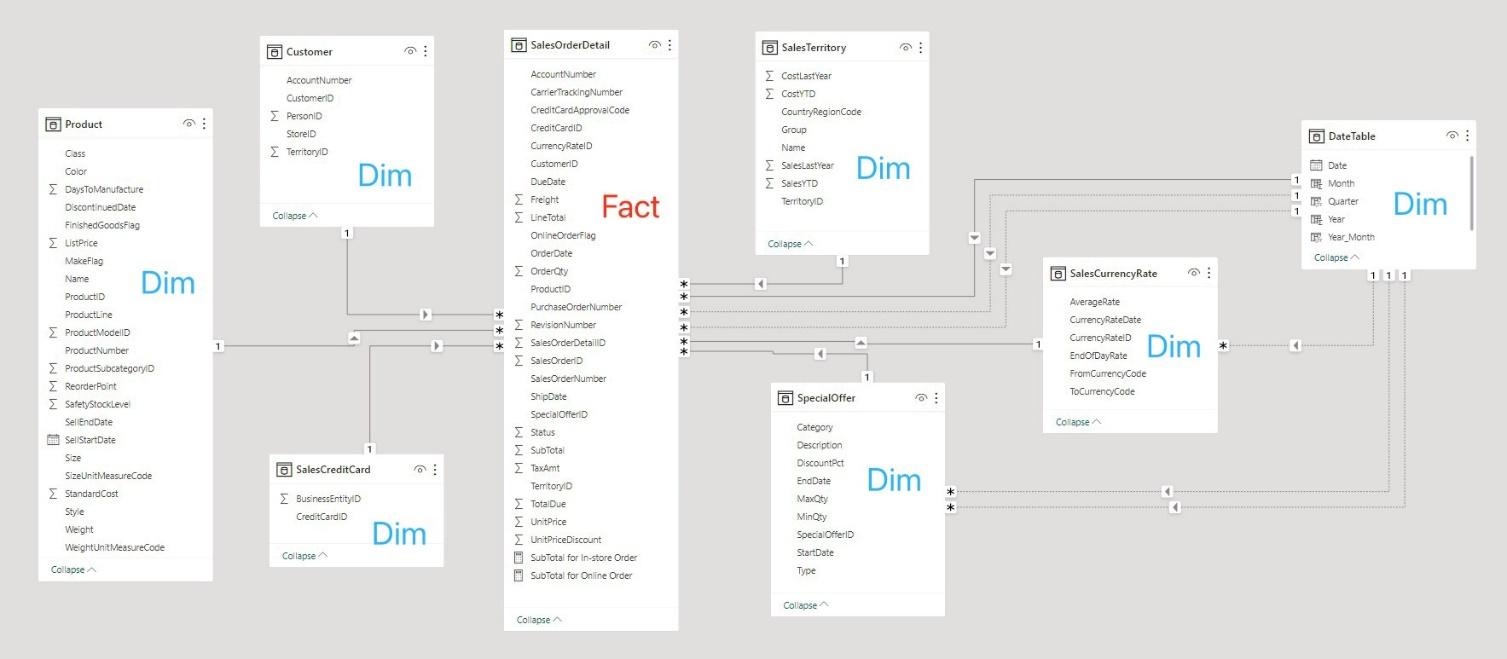
The "Blank" category within the procurement data likely represents items that have not been assigned to a specific subcategory. These could include miscellaneous parts or new products that have not yet been integrated into the established classification system. The presence of such a category indicates a potential area for process improvement, as unclassified items can affect the efficiency of inventory management and procurement processes.

**Insights from the Dashboard**

The dashboard highlights several insights, including the concentration of spending in certain categories and the roles individuals play within the procurement process. For example, the fact that buyers are the primary initiators of purchase orders might suggest a centralized purchasing strategy. The data also reveal spending patterns that could indicate which items are in high demand and may benefit from bulk purchasing.

## Sales

* Responsible: Bo Yang
* Process Overview: Explore the impact of different factors on sales trends, aiming to identify the main influencing elements and underlying issues. By employing analysis methods that align with industry requirements, propose viable adjustments to Adventure Works' sales process. These recommendations are intended to lay the foundation for the company's stable growth in sales revenue.
* Data Modeling



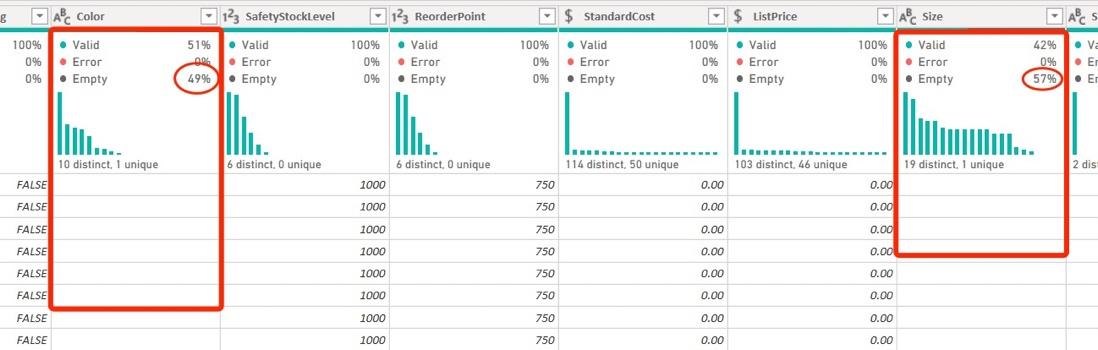
The analytical foundation is predicated on a star-schema data model, with the SalesOrderDetail fact table at its core, interfacing with multiple dimension tables including Product, Customer, SalesTerritory, SpecialOffer, SalesCurrencyRate, and DateTable. The fact table catalogues intricate details of sales transactions such as quantity, revenue, and discounts, while the dimension tables contextualize this data for analysis.

* Product Dimension offers a deep dive into product-specific attributes like classification, color, manufacturing lead time, and pricing, aiding in dissecting how product features impact sales.
* Customer Dimension logs essential identifiers like account numbers and personal IDs, essential for decoding buying patterns across customer segments.
* SalesTerritory Dimension delineates regional segmentation, enabling a geographical breakdown of sales performance.
* SpecialOffer Dimension maintains promotional details such as discount types and validity, instrumental in gauging promotional campaign effectiveness.
* SalesCurrencyRate Dimension encompasses exchange rate metrics, a critical component for transnational sales scrutiny.
* DateTable Dimension provides a hierarchical timeline framework, facilitating trend and seasonal pattern analysis in sales data.

Data Modeling steps:

1. Data cleaning and preprocessing:

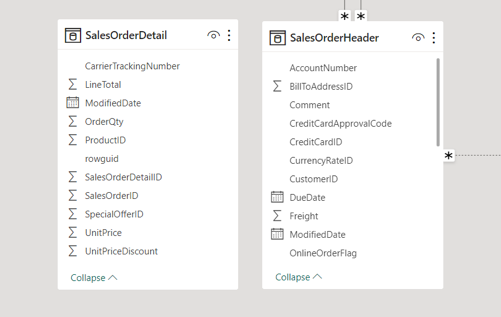
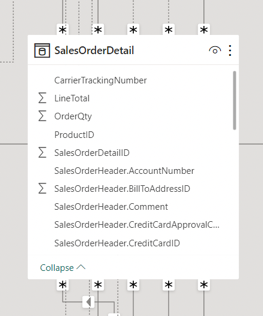
* Checking the data quality by using Data Transform



* Deleting duplicated
* Fix or delete empty items
* Changing format (price amount, text, and whole number values)

2. Data preprocessing:

* Energy tables (outer-left join SalesOrderHeader to SalesOrderDetail by using the common SalesOrderDetailID) to get the final fact table.

Before -> After

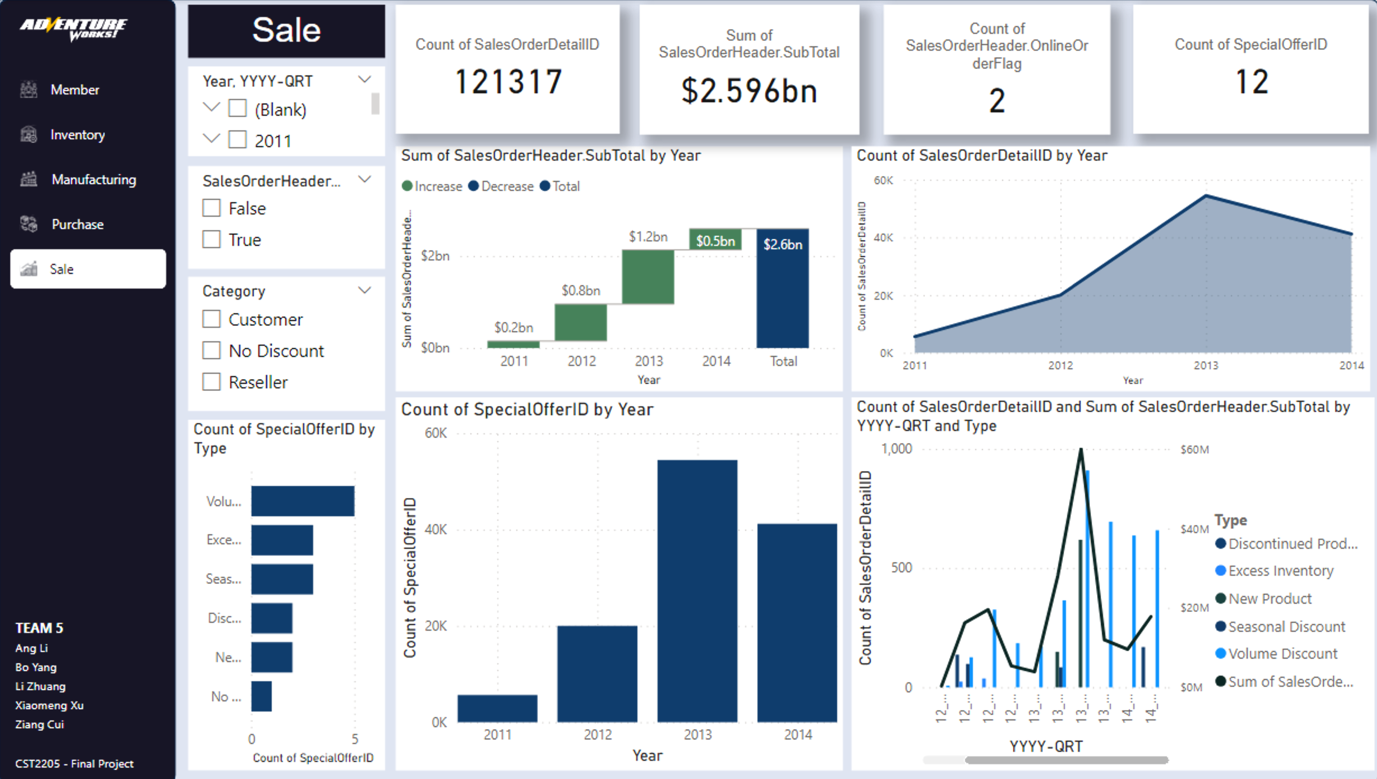
3. Design logical model: Design the structure and relationships of the data model based on an Entity-Relationship (ER) model or other logical models.

* Create the relationship between fact table and dimension tables. Make sure it is a **Star Schema** NOT Snowflakes Schema.
* Check if all cardinalities are set correctly as “Many to One”.
* Set the cross-filter direction to “Single” for most relationship for enhance the performance of data query.

4. Normalization: Normalize the data model based on normalization principles to eliminate data redundancy and improve the optimization of data structure.

* Adjust the format type of Date through all the tables (Like YY-MM and YY-QQ)

Data Visualization:



1. Choose the appropriate visualization type: Choose the most suitable visualization type based on the nature and purpose of the data, such as line charts, bar charts, pie charts, scatter charts, etc.

2. Design visual layout: Design the visual layout, including determining the size, position, coordinate axis, legend, etc. of the chart to ensure that the information is clear and easy to understand.

3. Add interactive functions: Add interactive functions, such as mouse hover to display data information, draggable adjustment of view range, filter function, etc., to improve user experience and data exploration.

Key Findings:

1. In 2012-2013, sales accelerated significantly. But in 2014, the sales trend reversed and began to decline.
2. The rising and falling trends in sales are affected by the order type. In other words, the proportion of online or offline sales has always been stable. It is a key factor affecting the sales growth trend, especially for the instore purchase.
3. From 2012 to 2013, the number of new products launching discounts increased significantly, and from 2013 to 2014, the number of discontinue product discounts increased significantly.

# Conclusion

By conducting detailed analysis on the five main business processes of "Adventure Works", we successfully transformed the complex 3NF data into a more analyzable star schema. This transformation significantly improved the efficiency of data queries and analysis depth, providing valuable insights for decision-makers.

* Member
* The member dashboard gives the general information about where the users come from how we can contact to them and the registration date in the whole time. Generally, most of the member is based on west coast and most of them registered between the 2012 to 2014. Just 5% of them lest their phone number and half of phone number are cell phone.
* Inventory

The Inventory dashboard gives a clear indication about how many Inventory Locations, Total Shelf Number, Total Bin Number, Total Product Number, which is a high level of understanding of the inventory part. Meanwhile, the shelf number, bin number, quantity of products, and count of product is also shown based on different inventory locations. By choosing different location, the top 10 quantity of product can also be easily to figure out. In general, the top 10 quantity of product in all the locations are all stored in Location 1, 6, 50. Meanwhile, Location 1, 6, 50 are the top 3 locations with the most shelves, bins, quantity of product, and count of product. It indicates the Location 1, 6, 50 are the core locations for the inventory.

* Manufacturing

The analysis of the manufacturing dashboard reveals critical insights into the production efficiency and cost management of our bicycle components. The top four manufactured products, namely the ML headset, HL Fork, LL Fork, and ML Fork, are foundational to bicycle integrity, highlighting the importance of prioritizing quality in these components. Additionally, Handlebars, Forks, and Derailleurs emerge as the products with the highest average costs, indicating potential areas for cost optimization. Despite an impressively low scrap rate of 0.24%, indicating efficient production processes and effective quality control, the primary reasons for scrapping—Paint process failure, Trim length too long, and Wheel misalignment—point to specific areas for improvement.

* Purchase

Through the meticulous redesign of the data model for the purchasing process at "Adventure Works," we have significantly enhanced our analytical capabilities, allowing for a deeper and more accurate understanding of procurement activities. The consolidation of fact and dimension tables has simplified the data structure, making analysis more direct and efficient. Notably, by addressing many-to-many relationships and introducing a date dimension, we have laid a solid foundation for analyzing temporal trends. Moreover, our deep dive into the procurement data has not only identified key patterns in the purchasing process but also highlighted the "Blank" category, pointing out potential areas for improvement in data classification and management.

While this analysis has provided valuable insights and revealed opportunities for optimizing purchasing efficiency and cost control, it also underscores the necessity of further refining the procurement strategy. The existence of the "Blank" category suggests a need for improvements in the classification system to eliminate potential impacts on the efficiency of the procurement process. Additionally, the analysis of purchasing categories not only shows a trend of spending concentration but also suggests that procurement strategies could be more personalized and targeted.

* Sales

After researching the reasons for this key finding: "From 2012 to 2013, sales accelerated significantly. But by 2014, the sales trend reversed and began to decline." I concluded that the type of order is what affects the sales growth trend One of the key factors (when the proportion of offline sales increases, the sales trend of total sales will also be faster). Another key factor is the special order aspect. We found that when the number of new product discounts increases significantly, it will promote an increase in sales, while when the number of new product discounts decreases significantly or a large number of products are removed from the shelves, sales will decrease significantly.

# Recommendations and Suggestions

**Members**

We have several suggestion bases on the analytics.

* + - 1. Increase advertising investment in the central and western regions.
      2. Short message will not reach most of the members, we should focus on the email campaign.
      3. Most of the member are European descent and a small amount of them are Asian. It’s not in high priority to develop a multilingual purchase system.
      4. Promotion through email will reach to the most amount member in records.

**Inventory**

In the data of this process, it is hard to find out any inventory change based on time, which is a challenge for data analyst to figure out more insights based on product inventory change. But in the inventory process, it is important for the executive or operation team to get better vision about inventory change. So if the company record more information related with inventory change and time, there will be more possibility to get better insights.

**Manufacturing**

We have three aspects of suggestions:

* Quality Focus: Continue to prioritize and enhance the quality of headsets and forks to ensure the structural foundation of our bicycles remains uncompromised.
* Cost Reduction Strategies: While focusing on Handlebars, Forks, and Derailleurs for cost optimization, explore innovative approaches to reduce manufacturing expenses without sacrificing quality, especially for Forks which are critical to bike safety and performance.
* Process Improvements: Focusing on improving the process including paint process failed, trim length too long, and wheel misaligned.

**Purchase Process**

For unclassified "Blank" items, it is recommended that "Adventure Works" enhances its classification management. By doing so, the company can improve the accuracy of inventory tracking, streamline the procurement process, and better align purchasing with production needs and market demand. Proper classification will also enable more precise data analysis, leading to informed decision-making and potential cost savings. Ensuring every item is categorized correctly can reduce procurement delays, avoid stockouts or overstock situations, and overall, optimize the supply chain’s responsiveness.

**Sales**

In order to achieve the goal of stable sales growth, I suggest that Adventure Works make the following two improvements in its business strategy:

1. Increase the supervision of offline orders and deeply explore the key driving forces for offline order transactions through questionnaires. In addition, you can also stimulate the number and rate of offline order transactions by issuing physical coupons or in-store only.

2. Adjust product release and related discount strategies. The planning of product release and removal can be more flexibly combined with seasons or appropriate frequency cycles to avoid centralized release and several types of removal. For discount strategies, the intensity and frequency of discounts can be adjusted based on data analysis results, flexibly respond to market demand and competitive environment, and increase sales and market share.

# References

| [1] | Dataedo, [Online]. Available: https://dataedo.com/samples/html/AdventureWorks/doc/AdventureWorks\_2/home.html. |
| --- | --- |