IS 6480
Data Warehousing Group Project – Adventure Works Data

Group 5 - Azure

Vidya Kattyparambil Balakrishnan
Maniteja Kurukunda

04/23/2023

TABLE OF CONTENTS

1. Executive Summary:

This report outlines the design and implementation of a data warehouse that serves as a repository for various data sources. The warehouse is designed to support the generation of insights and analytics related to sales data from Adventure Works, a bicycle manufacturer. The purpose of the warehouse is to enable users to query and analyse large volumes of data efficiently and derive valuable insights for business decision-making.

In the current digital age, data plays a crucial role in providing actionable insights to businesses. A data warehouse is an essential tool that allows businesses to collect, store, and analyse vast amounts of data from various sources. In the case of Adventure Works, the warehouse is designed to support sales data from 2015 to 2017. The warehouse enables the company to identify trends and patterns in sales data, such as top-performing products, customer demographics, and geographic locations with the highest sales. These insights can be used to optimize marketing strategies, improve customer retention, and increase revenue.

The Adventure Works data warehouse is designed to support multiple fact and dimension tables, allowing the company to collect and analyse data from various sources. The warehouse consists of several dimension tables, such as customer, product, calendar, and territories, and fact tables for sales data and returns providing a more comprehensive view of the business operations. Overall, the Adventure Works data warehouse provides a powerful tool for the company to leverage its data assets and drive business value. By analysing and stay ahead of the competition.

2. About the Organization

Adventure Works is a fictitious bicycle manufacturing company that produces high-quality bicycles and accessories. The company's primary objective is to provide its customers with the best possible products and services by constantly innovating and improving its products. The company caters to diverse customers, including professional cyclists, biking enthusiasts, and casual riders.

Adventure Works offers a wide range of products and services, including bicycles, accessories, apparel, and services related to bicycle maintenance and repairs. The company's product line includes road bikes, mountain bikes, hybrid bikes, and electric bikes, among others. Additionally, the company also offers a variety of accessories, such as helmets, bike locks, and lights, to ensure the safety and convenience of its customers.

Adventure Works is an organization that has invested in a data warehousing and analytics platform, recognizing the importance of data-driven decision-making to achieve its objectives. The data warehouse integrates various data sources, such as sales, customers and products, to provide a unified view of the company's performance. Adventure Works can

analyse and visualize data with the analytics platform, gaining insights into customer behaviour, product performance, and market trends. These insights enable the company to identify growth opportunities, optimize its supply chain, and enhance customer experience. Moreover, the company aims to use data analytics to identify new market opportunities and improve product quality while striving to increase revenue, reduce costs, and improve customer satisfaction.

3. Dimensional Model Requirements

Business Process	Dimensions				
	Customer	Product	Territory	Date	Orders/Returns
Sales Performance					
Region with highest Sales		X	X	X	X
Income Level of Top Customers	X				X
Sales based on holiday				X	X
Sales based on product Category	X				X
Returned Products Performance					
Loss due to returns					X
Highest returned product		X			X
Duration between order & return				X	X
Region which returns most products		X	X		X

Table - 1

Requirement Name	Short Description	Status
Sales Performance		
Region vs sales data	Helps to shortlist regions to focus	Highest in North America
Customer Income vs sales data	Helps to understand customer needs	Customers with income level range $100k - $130K generate most revenue
Holiday vs sales data	Trends based on weekend vs weekdays	Most sales are during weekdays
Product category vs sales data	Product categories from which profit can be increases	Bikes generate highest revenue
Returns		
Amount vs returns data	Total loss due to returned goods	To be determined
Product vs returns data	Improve quality of highest returned product	Accessories are returned the most
Date vs order and return data	Improve quality of highest returned product	To be determined
Territory vs returns data	Focus more on the region to understand consumer's needs	North America has highest returns

Table – 2

4. Dimensional Model

Below is the Dimensional Model designed for Adventure Works corresponding to sales and returns.

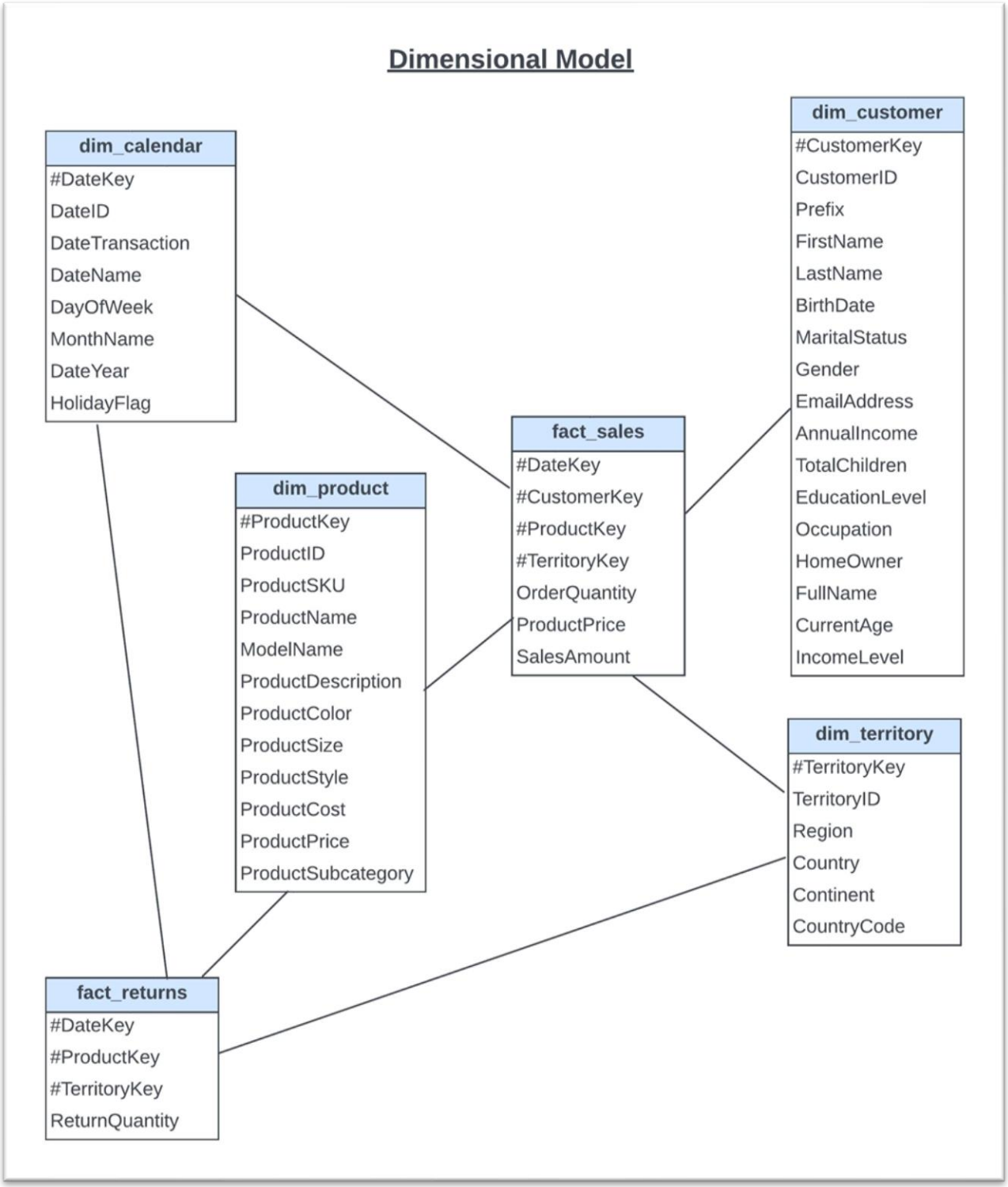


Figure 1

The dimensional model consists of 4-dimension tables and 2 fact tables. Fact tables contain attributes that are used as measurements for analysis. Dimension tables contains attributes that are constraints for measurement of the attributes of the fact table.

4.1 Dimension Tables:

dim_customer: The dim_customer table consists of customer data. It contains various common attributes of the customers such as name, birth date, gender, email address and also consists of specific attributes such as current age, income level, home owner, total children etc which will be helpful to group sales according to these categories and determine sales trends by customer attributes.
For example, customers between which age ranges generate most of the sales or how much sales is generated from customer with average income level. These key observations can be made from this data to improve sales by focusing on specific customers.

dim_product: The dim_product table consists of product data. The tables product_category and product_subcateory have been combined along with the already existing product table to eliminate snowflake structure. Important observations such as sales based on subcategory or category of products can be obtained to improve business outcomes.

dim_territory: The dim_terrirtory table contains information of territories where sales happen. Sales can be grouped under region, country or continent to obtain a perspective on which territory needs to be improved based on sales amount so that more money can be invested or targeted to improve sales based on territories.

dim_calendar: As the name suggests, dim_calendar table contains day name, month name, year, holiday flag etc to validate the period in which sales are more and when sales is poor. All the information needed for data analysts and data scientists to work on is pre-recorded into this table.

4.2 Fact Tables:

fact_sales: The fact_sales table consists of information such as price quantity and sales amount. The are the attributes that become contents in report. It consists of warehouse keys from its associated dimension tables such as product, customer, calendar and territory.

fact_returns: The fact_returns table consists of information of returned products of Adventure Works. It is important to know returns based on dim_tables such as product, calendar and territory. For example, highest returned quantity can be segregated by product subcategory and product name therefore helping business to improve on that product in the future.

5. Data Warehouse Architecture

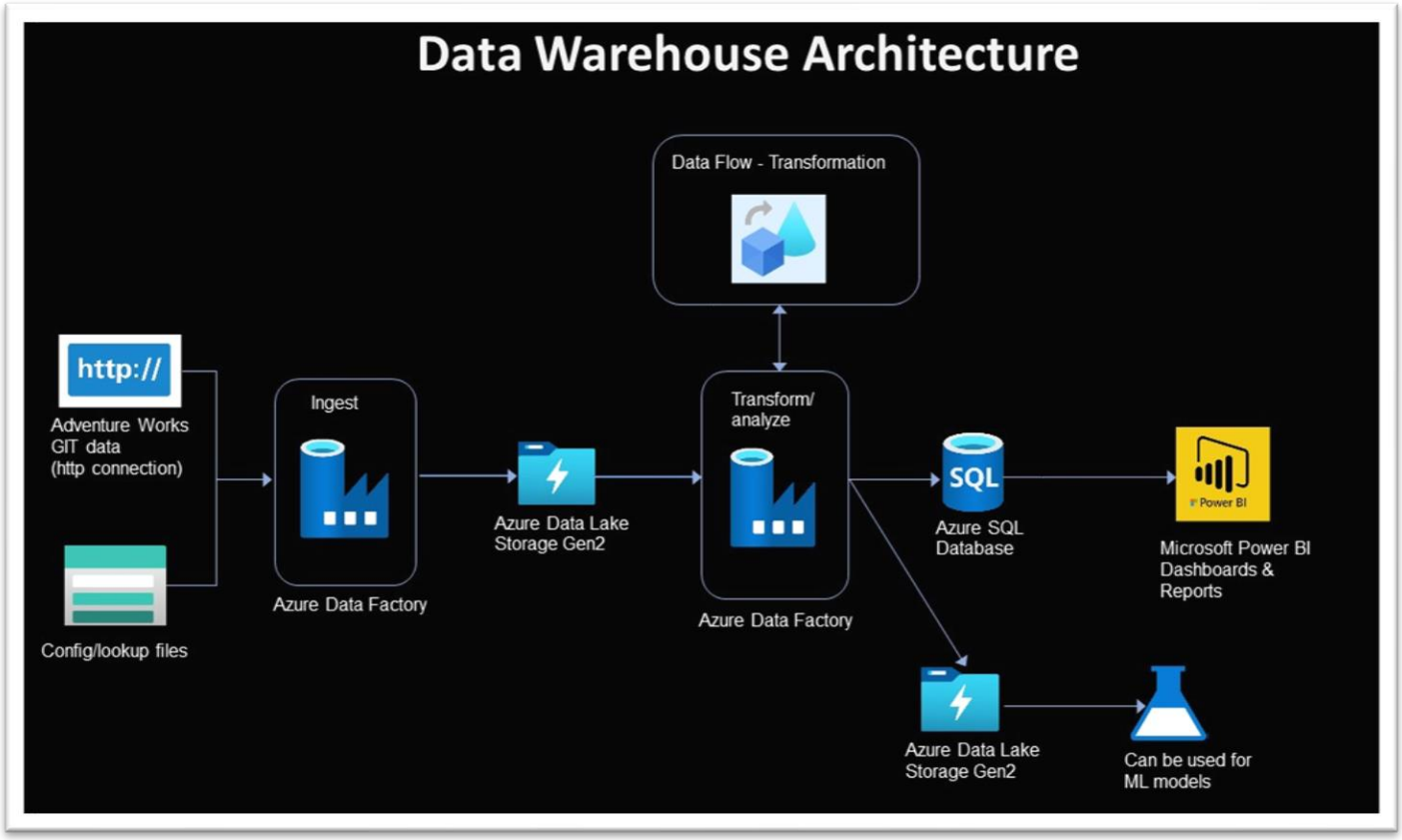


Figure - 2

Even small-scale business these days, generate a large amount of data. The data generated comes from a variety of sources such as IOT platforms, on-premises applications, SaaS Applications and so on. Real-time data is generated every millisecond at faster rates and in different formats ranging from unstructured, structured to semi-structured and there is a need for businesses to streamline this data into usable format to reap benefits in a consistent manner.

Azure cloud services have been used for Adventure Work's Data Warehouse solution. The following table provides a list of the pipelines used for the ETL process.

Pipelines	Operation	Pipeline name in Azure Data Factory
Pipeline 1	INGEST	01_pl_ingest_adv_works_data_from_git_to_raw_datalake
Pipeline 2	TRANSFORM	02_pl_tranform_and_move_raw_to_staging_datalake
Pipeline 3	LOAD	03_pl_load_staging_datalake_to_staging_sql
Pipeline 4	LOAD	04_pl_load_sql_staging_to_sql_presentation_dim_tables
Pipeline 5	LOAD	05_pl_load_sql_presentation_to_sql_presentation_fact_tables

Table - 3

5.1 DATA INGESTION/EXTRACTION

Azure Data Factory is a fully managed, cloud-based ETL data integration solution which can be used to ingest enormous data from different sources, prepare and also transform the ingested data at a large scale.

Data Source:

To resemble real-time data from websites storing data like covid data, population related data etc, Adventure works csv data has been uploaded to a GIT repository which can be accessed using this git link which is a HTTP website. A HTTP connector is used by the data factory to get this data. Pipeline 1 is created to ingest the data from the HTTP website into the Azure Data Lake Storage Gen2.

Pipeline 1 *- 01_pl_ingest_adv_works_data_from_git_to_raw_datalake*

This pipeline reads the config file which contains the baseURL, relativeURL and filename of the csv that is to be copied from the GIT repository into Azure Data Lake. A *'Lookup'* activity is used to read the config file which is stored in Azure Blob storage account. This config file is in JSON format. For every entry in the config file, a *'For Each'* activity is called which in turn invokes the *'Copy Data'* activity and copies .csv files from the http website into a folder named 'raw' in Azure Data Lake Storage Gen2. Required linked services and datasets are created for this pipeline to work successfully.
A Trigger named *"01_trigger_ingest_adventure_works_data"* has been created for this pipeline, such that it can be configured to trigger at a scheduled time repeatedly until a specified end time. This is to make sure that the real time data from the HTTP website is loaded on a daily basis or regular intervals to keep the data in the Data Lake updated.

5.2 DATA TRANSFORMATION

The raw data that is stored in the Data Lake has to be transformed according to the usage and stored into the staging area in Data Lake. To transform the raw data into processed data *Data Flows* have been used in Data Factory. Data Flows have a variety of operations that can be performed on the incoming data streams such as join, conditional split, union, lookup, derived columns, select, aggregate, pivot, filter, sort, sink etc to name a few.

The following are the data transformations that are performed on the raw data downloaded from git.

- **Customer data:**
  - A SELECT transformation is applied to alter the names of a few fields in the customer table.

- A DERIVED COLUMNS transformation is applied to derive new columns from the existing ones such as the column FullName, formed by combining columns - prefix, FirstName and LastName. Similarly, the column CurrentAge is derived by subtracting current year from the year of the BirthDate column. IncomeLevel column is derived from the AnnualIncome column with an IIF condition which states that if
    - If AnnualIncome <= $40,000 then IncomeLevel = Low
    - Else if $40,000 < AnnualIncome <= $90,000, then IncomeLevel = Average
    - Else if $90,000 < AnnualIncome <= $130,000, then IncomeLevel = High
    - Else if AnnualIncome > $130,000 then IncomeLevel = Very High
- This transformed data is then copied into the staging area in Data Lake

➢ **Product data:**

- Since there are 3 different tables related to Product – product table, product subcategory table and product category table, a JOIN is performed on the 3 tables to combine them into a single product table.
- A SELECT transformation is performed to edit column names and to remove any duplicate columns if any
- Finally, the transformed product data is copied to the staging area in Data Lake

➢ **Territories data:**

- The territories table consists of column – region, country and continent but lacks the country code column which will be useful down the lane. Therefore, a LOOKUP transformation is used to lookup the country code for each country and then add the country code column into the territories table from the lookup file which is stored in the Azure Blob storage.
- SELECT transformation is then applied to alter names and to select only needed columns.
- Finally, the transformed territories data is copied to the staging area in Data Lake

➢ **Calendar data:**

- In the calendar table, there is a column for order dates. To identify the data uniquely, a column was added called Date_id using window function transformation, propagated forward by select transformation, and the Date column was changed to Date_col.
- A derived column transformation is applied to the data to add Day_name, Day_of_week, Month_name, Year_col, and Holiday_flag columns.

- The Calendar data has been loaded into the staging area in Data Lake after all transformations.

- **Sales data:**
  - Since the sales data for Adventure works is broken into 3 tables based on the year 2015, 2016 and 2017, these 3 tables are combined using JOIN transformation
  - SELECT transformation is used to alter names of columns
  - Finally, the transformed sales data is copied to the staging area in Data Lake

- **Returns data:**
  - SELECT transformation is used to edit names of columns as needed

***Pipeline 2** - 02_pl_tranform_and_move_raw_to_staging_datalake*

In order for the transformation to take effect, a pipeline 2 is created to execute each data flow. On execution of this pipeline, the raw data is transformed into processed data and stored into the Staging folder of Azure Data Lake Storage Gen2.

5.3 DATA LOADING

- **Loading data into Staging database schema:**

***Pipeline 3** - 03_pl_load_staging_datalake_to_staging_sql*

The pipeline performs a COPY activity individually on all the above created data flows keeping the source as Azure Data Lake Storage Gen2 and sink as Azure SQL database with a schema for staging namely *aw_staging* and loads data from data lake to Azure SQL database.

- **Loading data into Presentation database schema:**

***Pipeline 4** - 04_pl_load_sql_staging_to_sql_presentation_dim_tables*
***Pipeline 5** - 05_pl_load_sql_presentation_to_sql_presentation_fact_tables*

Separate pipelines are created for 1) loading the dim tables from SQL staging database to SQL presentation database and 2) loading fact tables into SQL presentation database. A schema named *aw_presentation* is created in the SQL server to capture data moved to presentation area. The surrogate keys or warehouse keys are added to each table in this stage. As the fact tables depend on the warehouse keys of dim tables, pipeline 4 is run before running pipeline 5.

6. Future Reports, Dashboards and Analyses

6.1 Executive Dashboard:

Sales Trend Analysis: Quarterly Sales Performance:

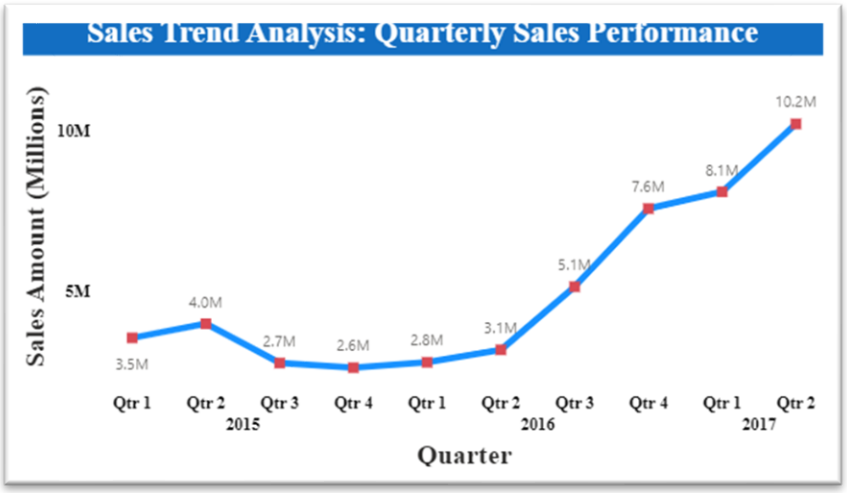


Figure - 3

The total sales amount in millions of dollars on the Y-axis is calculated based on order quantity and price to create this chart. On the X-axis, using the quarter for each year to create a timeline. Using the line chart type to show the trend helps analyse the overall direction and seasonal sales patterns.

The visualization shows the company's quarterly sales performance over a given period. Analysing the trend over the quarters allows us to identify periods when sales are high or low. From the above visualization, we can observe that company sales are high in quarter 2 of each year, and company sales are increasing since 2016 without any decline in performance This can help plan marketing and promotional activities for the upcoming quarters to maintain a positive trend.

Sales Performance by Product Category and Continent:

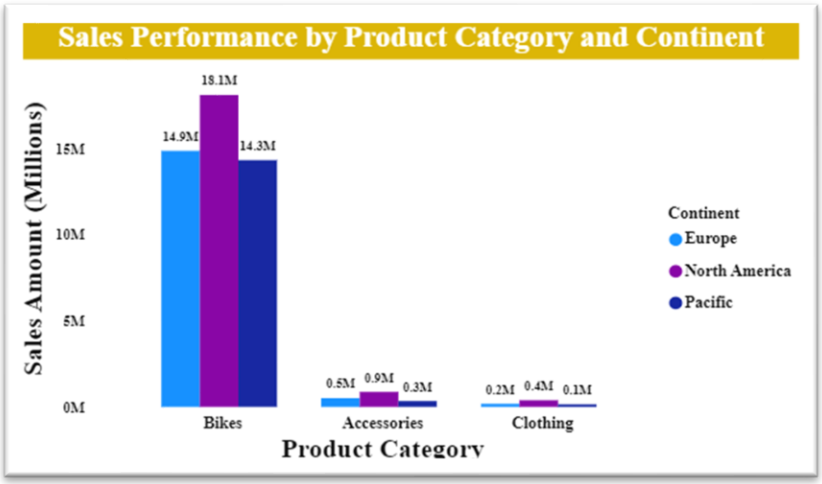


Figure - 4

The X-axis displays the product categories, while the Y-axis shows the total sales amount in millions of dollars. The legend indicates the continents where the products were sold.

To create this visualization, we used a Clustered column chart type, where the product categories are stacked on top of each other based on the total sales amount. In addition, each stack is color-coded by continent, which helps to compare the sales performance of each product category across continents.

The visualization shows the sales performance of each product category across different continents. It helps identify which product categories are performing well in which continents and not. The chart shows that Bikes are the top performing product on all continents and North America is the continent where the maximum number of sales happened. This can help determine each continent's product development and marketing strategies based on preferences and demands. In addition, it can identify potential growth areas in product categories that could be performing better in specific continents.

Sales Performance by Customer Demographics:

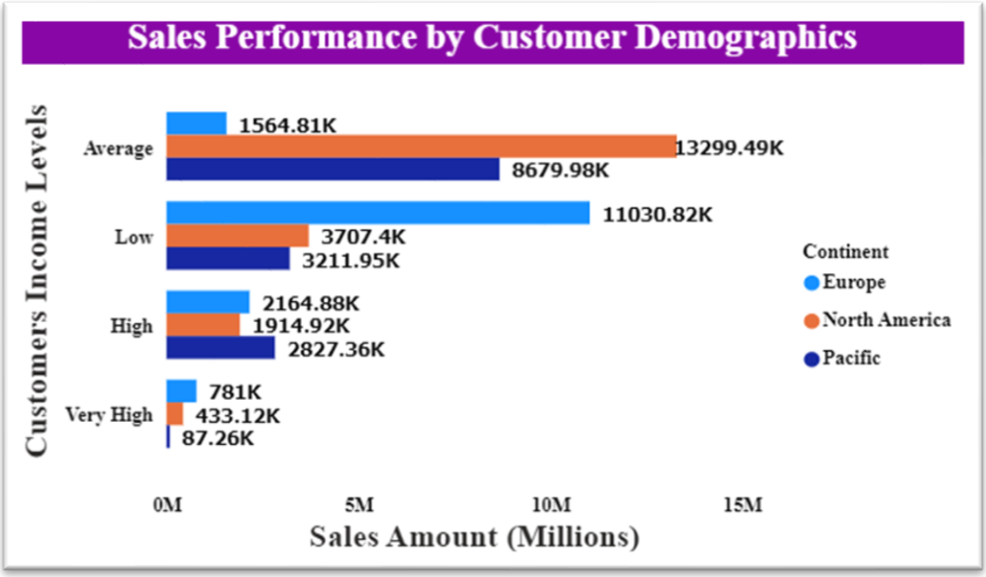


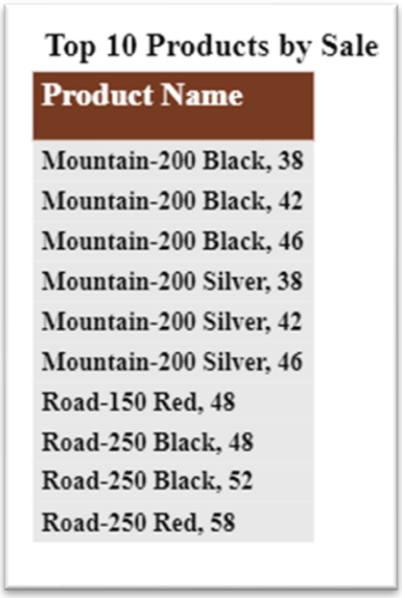
Figure - 5

The customer income levels are plotted on the Y-axis, and the total sales amount in millions on the X-axis, with the legend indicating the continent. This visualization is created using a Horizontal Clustered chart to represent sales distribution based on the income levels of customers from different continents.

The visualization shows the relationship between the income level of customers and the sales amount they generate. This can help identify which income groups contribute the most to sales and which continents are the most profitable. The visualization shows that customers with average and low-income levels are purchasing more products when compared to high

and very high levels. In addition, this insight can be used to tailor marketing campaigns and promotions to target specific income groups and continents to improve sales performance.

Top 10 Products by Sales:



Top 10 Products by Sale

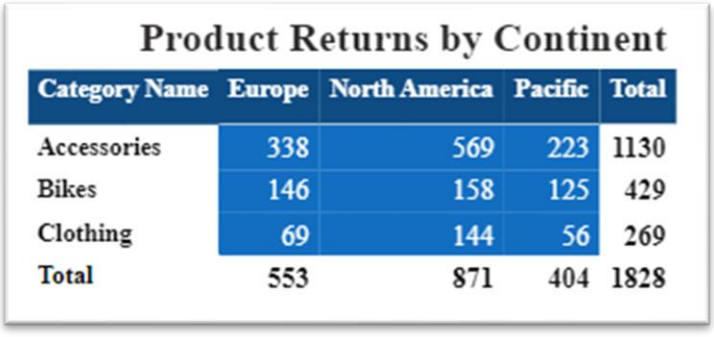
Product Name
Mountain-200 Black, 38
Mountain-200 Black, 42
Mountain-200 Black, 46
Mountain-200 Silver, 38
Mountain-200 Silver, 42
Mountain-200 Silver, 46
Road-150 Red, 48
Road-250 Black, 48
Road-250 Black, 52
Road-250 Red, 58

Figure - 6

The table is created by sorting the products based on sales amount in descending order and selecting the top 10 products. The product names are displayed in the first column.

This visualization helps identify the products generating the highest sales revenue for the company. From the above table, we can observe that Mountain Bikes are performing well and followed by Road Bikes. By analysing the top 10 products, the company can focus on improving the marketing and promotion strategies for these products to maintain or increase their sales performance. Additionally, the company can use this information to identify potential cross-selling or upselling opportunities to increase revenue.

Product Returns by Continent:



Product Returns by Continent

Category Name	Europe	North America	Pacific	Total
Accessories	338	569	223	1130
Bikes	146	158	125	429
Clothing	69	144	56	269
Total	553	871	404	1828

Figure - 7

This visualization uses a matrix with product categories as rows and continents as columns. The cell values display the quantity of returned products, and subtotals for each column and row provide a comprehensive summary of product returns by continent.

Analysing the returns by product category and continent can help identify patterns and specific products or regions more prone to returns. The above matrix shows that North America is the continent where more returns are happening, and the more returned product category is Accessories. This information can be used to improve the quality of the products, the shipping and handling processes, and the customer service policies to minimize returns and enhance customer satisfaction. Additionally, it can. It can provide valuable feedback to the product development team, allowing them to make necessary improvements and modifications to the products to reduce the likelihood of returns.

Filters:

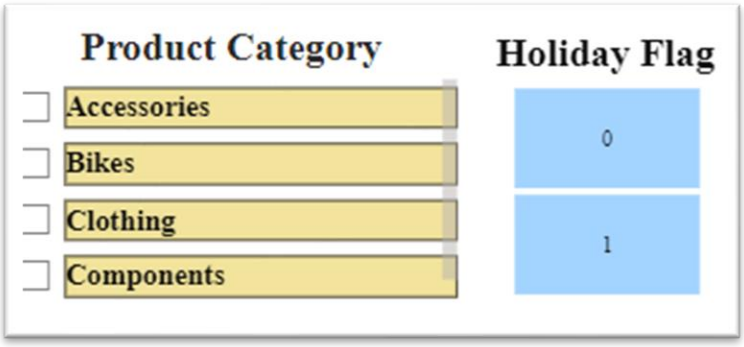


Figure - 8

Analysing sales by product category and holiday flag can help identify which products sell the most during holidays and adjust marketing and promotional activities accordingly. This information can aid in maximizing revenue during the holiday season.

6.2 Employee Dashboard:

Top Customers by Average Order Value:

Top Customers by Average Order Value

Full Name	Average Order value	Email Address	Income Level
MR. BENJAMIN SHAN	3,578.00	benjamin33@adventure-works.com	Low
MR. BRANDON ZHANG	3,578.00	brandon20@adventure-works.com	Average
MR. CAMERON HENDERSON	3,578.00	cameron0@adventure-works.com	Average
MR. CARSON JENKINS	3,578.00	carson5@adventure-works.com	Low
MR. CHRISTIAN BUTLER	3,578.00	christian29@adventure-works.com	Low
MR. COLE RICHARDSON	3,578.00	cole11@adventure-works.com	Low
MR. IAN EDWARDS	3,578.00	ian38@adventure-works.com	Average
MR. IAN HENDERSON	3,578.00	ian45@adventure-works.com	Average
MR. IAN WILSON	3,578.00	ian8@adventure-works.com	Low
MR. ISAIAH EDWARDS	3,578.00	isaiah23@adventure-works.com	Average
MR. JACK ZIMMERMAN	3,578.00	jack24@adventure-works.com	Low
MR. JAMES MILLER	3,578.00	james81@adventure-works.com	Low
MR. JARED MOYER	3,578.00	jared16@adventure-works.com	Average
MR. JONATHAN GONZALES	3,578.00	jonathan16@adventure-works.com	Low

Figure - 9

This visualization presents a table of the top customers based on their average order value, displaying the full name, email address, income level, and average order value. The data has been filtered by setting a threshold of $3500 for the average order value.

Businesses can determine their most valuable customers by identifying the top customers with the highest average order value. These customers are more likely to have a higher customer lifetime value (CLV) as they spend more on each purchase. Understanding and catering to the needs of these high-value customers can help increase their loyalty to the company, resulting in repeat purchases and positive word-of-mouth advertising.

Return Quantity by Product Category and subcategory:

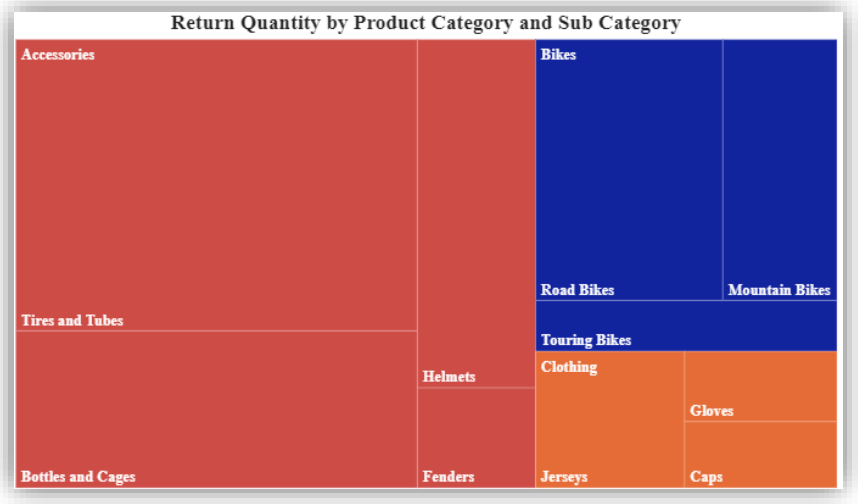
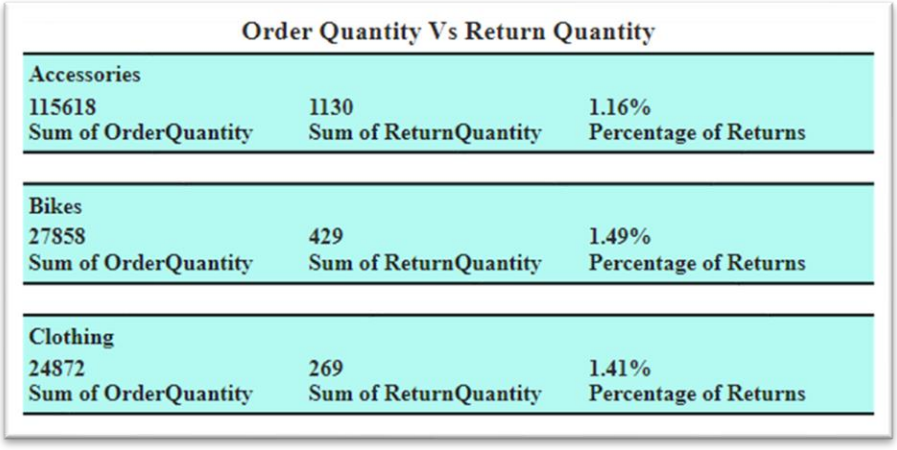


Figure - 10

This visualization uses a tree map to display return quantities by product category and sub-category. The size of each rectangle is proportional to the return quantity, with the colour indicating the product category and sub-category.

By analysing return quantity by product category and sub-category, businesses can identify areas where they need to improve product quality, customer service, or delivery to reduce returns. Accessories with subcategories Tires and Tubes and Bottles and Cages have more product returns. Additionally, can aid inventory management by highlighting the products and categories returned most frequently.

Order Quantity vs Return Quantity:



Order Quantity Vs Return Quantity

Accessories		
115618 Sum of OrderQuantity	1130 Sum of ReturnQuantity	1.16% Percentage of Returns
Bikes		
27858 Sum of OrderQuantity	429 Sum of ReturnQuantity	1.49% Percentage of Returns
Clothing		
24872 Sum of OrderQuantity	269 Sum of ReturnQuantity	1.41% Percentage of Returns

Figure - 11

This visualization presents the order quantity versus the return quantity for each category, displayed in a multi-card visualization. The categories are arranged in descending order based on the sum of order quantity and return quantity. The cards display the sum of the order quantity, the sum of the return quantity, and the percentage of returns from the ordered quantity.

Businesses can identify which products categories have a higher return rate by analysing each category's order quantity and return quantity. The multi-card visualization shows that Bikes has the highest percentage of product returns from the ordered quantity. This insight can help businesses optimize their return policies and improve product quality to reduce the number of returns.

Single Cards and Filters:

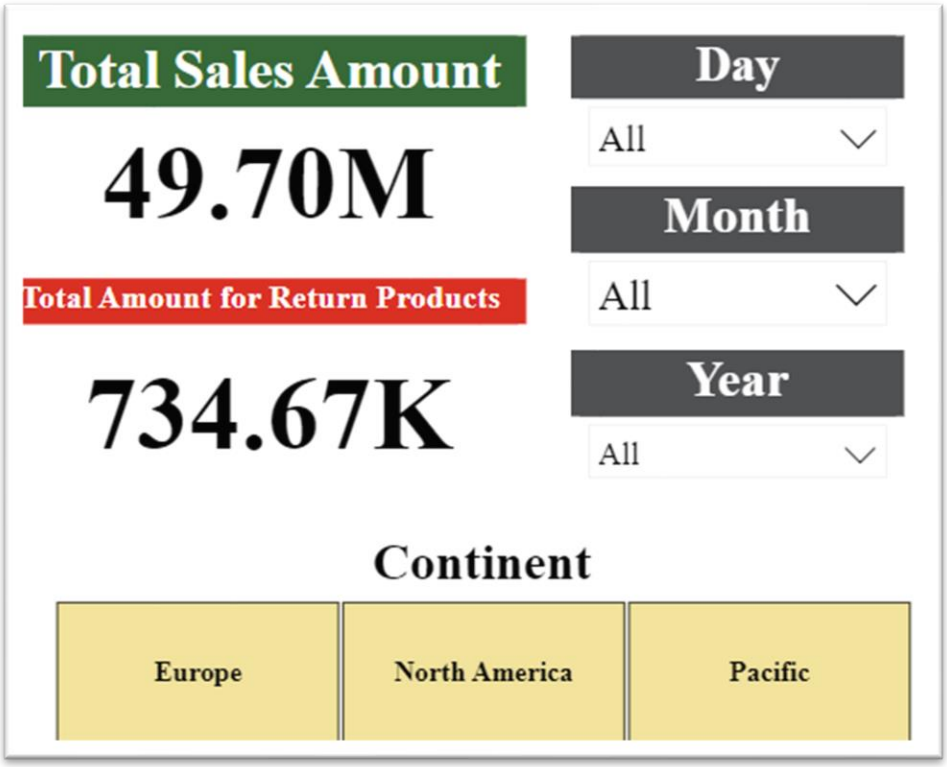


Figure - 12

This visualization consists of two card visuals that provide essential insights into the sales and refund data of a business. The first card visual displays the total sales amount across all regions, while the second card visual displays the total refund amount for returned products. What makes this visualization even more powerful is the ability to filter the data by day, month, year, and continent. By utilizing these filters, businesses can gain a more detailed understanding of their sales and refund data and identify trends that can be used to optimize their sales strategies and improve customer satisfaction.

Businesses can gain valuable insights into their sales and refund data by using the day, month, year, and continent filters in the two-card visuals of this visualization. These filters

allow for a more detailed analysis of the Order Quantity vs. Return Quantity, helping companies identify which products have a higher return rate and take appropriate measures to reduce them. Similarly, analysing the Return Quantity by Product Category and subcategory by applying these filters can help businesses understand which categories and subcategories of products are more prone to returns in specific regions and time frames.

7. Appendix

- Link to Presentation recording
- Summary of Time-tracking section

Team Member	Total Hours Spent for Project
Vidya	30
Mani	30

- Detailed Time-tracking section

Date	Team Member	Hours Spent	Description of work
04/10/2023 04/11/2023 04/12/2023	Vidya	10	- Explored and got familiar with Azure environment – pipelines, triggers, linked service, data flow transformations, datasets in Data Factory, storage accounts – data lake, blob storage, Azure SQL Database and Data Studio - Set up Azure development environment
04/14/2023 04/15/2023 04/16/2023	Vidya	10	- Gathered dataset & created dimensional model - Set up GIT as data source, created pipelines and triggers to get raw data from GIT and store it into data lake - Transformed and moved the processed data to staging area in data lake and SQL database - Resolved debug errors

04/17/2023 04/18/2023 04/19/2023	Vidya	10	-Captured all design and architecture details in Summary Report and Presentation
04/13/2023 04/14/2023 04/15/2023	Mani	8.5	- I got familiar with the Azure cloud environment by learning through a Udemy course that is connecting data from different sources and performing ETL and building pipelines to store the data in the destination
04/16/2023 04/17/2023 04/18/2023	Mani	12	-Created Tables in the presentation area and loaded the data from Staging to presentation area SQL database tables
04/19/2023 04/20/2023 04/21/2023	Mani	9.5	- Connect Azure to Power BI and created executive and employee dashboards using visualizations and created measures ad VLOOKUP's. - Documented Dashboard, Reports, Analysis and did PPT for the project presentation

8. References

- Udemy course - Azure Data Factory For Data Engineers - Project on Covid19
- Microsoft Documentation on Azure Data Factory
- Adventure Works – Data Dictionary