UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT

DATA WAREHOUSE AND INTEGRATION COURSE

TOPIC:

OPTIMIZING SALES PERFORMANCE THROUGH DATA WAREHOUSING: A CASE STUDY OF THE ADVENTUREWORKS2019 DATABASE

Lecturer: Le Ba Thien, MSc.

Group LIVE LAUGH LOVE:

1. **K214162140 – Le Quoc Dan An**
2. **K214161243 – Tran Hoang Anh**
3. **K214162148 – Cao Nguyen Hai Nhu**
4. **K214160989 – Tran Thi Minh Hien**

Ho Chi Minh City, December 5th 2023

Members of Group

NO.	**Full Name**	**Student ID**	**Role**
1	Le Quoc Dan An	K214162140	Leader
2	Tran Hoang Anh	K214162143	
3	Cao Nguyen Hai Nhu	K214162148	
4	Tran Thi Minh Hien	K214160989	

Table of Work assignment

No.	Task	Assignment	Percentage of completion
1	**Document**		
1.1	*Chapter 1*	Hien	100%
1.2	*Chapter 2*	An, Nhu	100%
1.3	*Chapter 3*	Anh	100%
1.4	*Chapter 4*	All	100%
1.5	*Chapter 5*	All	100%
1.6	*Chapter 6*	Hien	100%
2	**Technique**		
2.1	*Building data warehouse*	All, Hoang Anh (main)	100%
2.2	*Building OLAP cube*	All	100%

Table of Contents

List of Figures

List of Tables

List of Acronyms

AWC	Adventure Works Cycles
OLAP	Online Analytical Processing
SSMS	SQL Server Management Studio
SSAS	SQL Server Analysis Services
SSIS	SQL Server Integration Services
ETL	Extract – Transform – Load
ERD	Entity-Relationship Diagram
PK	Primary Key
FK	Foreign Key

Chapter 1: Overview

1.1. Business problem

Adventure Works Cycles is a large, multinational manufacturing company. The company manufactures and sells metal and composite bicycles to North American, European, and Asian commercial markets. Operating on a large scale, the company needs to control its sales performance related to customers and products over time. However, data from each department are stored in separate places with different formats, making it difficult to overview the company's sales performance. Moreover, transactional databases often store current data, making it challenging to analyze historical trends. Finally, historical analysis requires businesses to track the changes in the database.

1.2. Objectives

Generally, this project aims to create an environment where needed data to perform sales performance such as data about customers, products, and locations are stored in a data warehouse. Also, to facilitate the process of decision-making, the following tasks should be conducted as the target of this project:

- A data warehouse;
- A multidimensional cube;
- Valuable information about sales performance.

Specifically, the following is the detailed target for each aspect:

Table 1: Table Define targets for each aspect of the project

Aspect	Target	Description
Data Warehouse	Well data integration	Integrate data from disparate sources into a centralized repository to ensure a unified and consistent view of data across the organization, reducing data silos.
	Able to perform historical data analysis	Enable analysis of historical data to identify trends and patterns, supporting long-term strategic planning and trend analysis
	Comprehensive metadata management	Facilitate document and track data lineage, enhancing understanding of the data source
Multidimensional cube	Successfully be deployed	
Sales performance investigation	Analyze trend	Recognize the changes over time
	Make comparison	Recognize the difference among attributes
	Identify target customers, products, etc	Recognize which objects to focus on

Scope

Time scope: 17/11/2023 – 5/12/2023

Space scope: Adventure Works 2019 data.

Structure of report

This work includes 57 pages, 49 figures, and charts.

In addition to the acronym catalog, the table and chart catalog, and the bibliography of references, the topic is divided into 6 sections:

- Chapter 1: Project Overview
- Chapter 2: Theoretical Basis
- Chapter 3: Requirements Analysis
- Chapter 4: Experimental Modeling
- Chapter 5: Discussion
- Chapter 6: Conclusion

Chapter 2: Theoretical Basis

2.1. Data warehouse concept

2.1.1. Data warehouse

Data Warehouse is a system for storing structured data, designed to support data querying and analysis for statistical and analytical purposes (Inmon et al., 2008). It is utilized by several businesses in the financial, e-commerce, retail, healthcare, and other industries to store data, facilitate analysis, and make decisions. It is a collection of data and information that abide by four characteristics such as subject-oriented, integrated, time-variant, and non-volatile. The data warehouse is organized based on the subject-oriented characteristic, which focuses on specific business issues and objectives of the enterprise. This makes data analysis easier, allowing users to quickly grasp all information about a specific topic. The integrated characteristic means that data from various applications is consistently integrated and stored in the data warehouse. The time-variant characteristic implies that data is labeled with a corresponding timestamp at the time of input, enabling users to analyze data over time. The non-volatile nature of the traditional data warehouse requires that once data is entered into the data repository, it must remain unchanged. All data must be kept in a read-only mode, and previous data should not be deleted when new data is entered.

With its capability to store structured data effectively, a data warehouse is utilized across various fields for diverse purposes. According to Shalid et al (2018), data warehouse is used to assist construction managers in making decisions and monitoring the effectiveness of construction; it also tracks the remaining inventory, trends related to inventory of materials, quantities, and amounts of each type of material, as well as the prices of all types of materials. In Marketing, data warehouse helps firms store and analyze data for main purposes such as trend Analysis, support web marketing, and market segmentation. Another important application of data warehouse is in the finance industry, data warehouse supports decision-making by consolidating diverse financial data, offering a centralized view for analysis and reporting. They enhance operational efficiency, support regulatory compliance, and enable historical trend analysis, risk assessment, and performance evaluation, fostering informed and competitive decision-making.

2.1.2. Data warehouse architecture

According to Inmon (2005), building your data warehouse architecture will go through 5 processes, including data source, staging area, data warehouse, data mart, and visualization. However, over the years, concepts and approaches to building a data warehouse have gradually evolved to be more suitable for optimal processing, but at its core, it still follows these 5 processes. The figure below illustrates the data warehouse architecture.

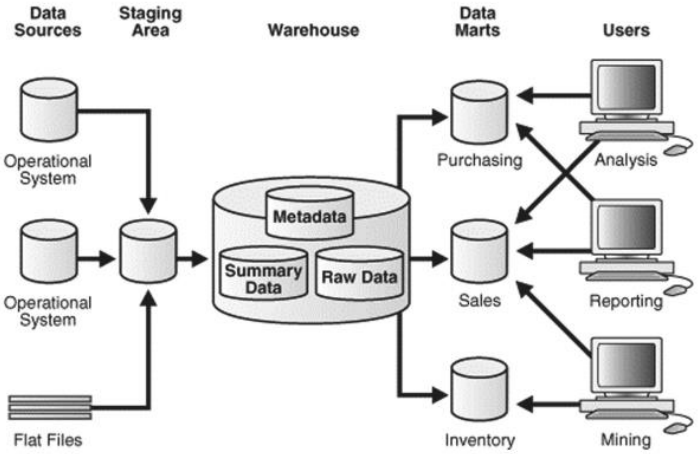

Figure 2- 1: Data warehouse architecture (Iqbal et al., 2020)

In the data source process, data is collected and prepared according to the objectives of building the business data warehouse. Data can come from various sources, including flat files, databases, etc., and is prepared for storage in the data warehouse.

The Staging area is a location for storing and processing data files to minimize errors before loading into the data warehouse (Inmon et al., 2014). In cases where data from two or more files must be merged and there is a timing issue, for example, when data from file ABC is ready for merger at 9:00 am, while data from file BCD is not ready until 5:00 pm, the data from file ABC must be “staged” until the merge is ready to occur.

After the data is prepared, it is used for ETL (Extract-Transform-Load), specifically extracting necessary data from the data source, transforming it into a consistent format, and loading it into the data warehouse. In the data warehouse construction process, data is organized into dimensional and fact tables. Tables are structured based on star schema, snowflake schema, or galaxy schema,

depending on the organization and business requirements of the enterprise.

Subsequently, data from the data warehouse is built into data marts to provide easy access and use for different departments. This approach enables departments to access and use data easily.

Finally, the data mining process visualizes data and results, supporting businesses in gaining a visual perspective on data for analysis and decision-making based on a suitable strategy.

2.1.3. Data warehouse modeling

Data stored in the data warehouse will be organized into dimensional and fact tables. These tables will be linked according to data warehouse models such as star schema, snowflake schema (Krishnan, 2013). Each model type will bring different benefits, depending on the purpose of building the data warehouse, and the choice of a specific model will be suitable accordingly.

Star schema

A star schema is a fundamental data modeling technique employed in the realm of data warehousing, designed to organize and represent data in a structured and intuitive manner. In a star schema, data is organized into a central fact table that contains the measures of interest, surrounded by dimension tables that describe the attributes of the measures. The figure below illustrates the star schema model.

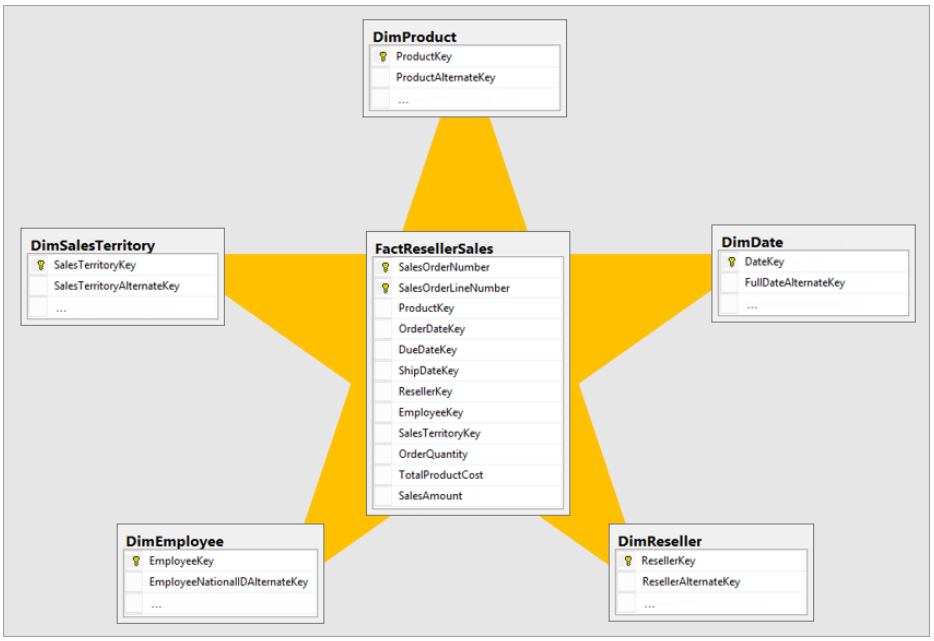

Figure 2- 2: Star schema model (Microsoft, 2023)

The basic components of a star schema include Facts, Dimensions, and Measures. The fact table in a star schema contains the measures or metrics that are interesting to the user or organization. Fact tables include the Primary Keys (PKs) found in dimension tables together with the term "Foreign Keys" (FK), which serves as a connection between the two types of tables. To illustrate, in a sales data warehouse, dimension tables could encompass categories like product, customer, time, and location. These dimensions offer diverse perspectives for scrutinizing the data, allowing users to glean insights and patterns from various angles. In a star schema, the dimension tables have details about the measures in the fact table. These details help chop up the data in the fact table, so people can study it from various angles. The column containing the key data representing a dimension table will be chosen as the PK. These Primary Keys will be linked to the Fact table to associate and retrieve data from the table, and the information will be displayed through the Fact table. For instance, in a sales data warehouse, the dimension tables might have info about products, customers, time, and location and these are used to map to the fact table to show data from these tables. This way, users can look at the data in different ways to understand things better (Giovinazzo, 2000; Iqbal et al., 2020).

Snowflake schema

According to Husemann (2000), the snowflake schema is a modified version of the star schema. In this schema, the main fact table is linked to various dimensions. Unlike the star schema, the snowflake schema represents dimensions in a normalized structure spread across multiple interconnected tables. There is a choice between using either of the two schemas. Snowflake schema, in contrast to the Star schema, is made up of three different kinds of tables: fact tables, dimension tables, and subdimension tables. The fact table is still located at the center of the schema, surrounded by the dimension tables and each dimension table undergoes further division into several connected tables, forming a hierarchical arrangement that takes on the appearance of a snowflake. The figure below illustrates this concept.

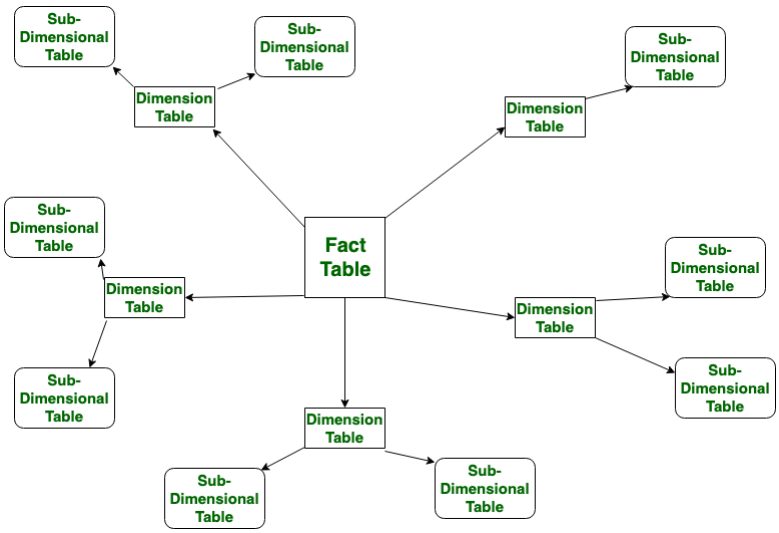

Figure 2- 3: Snowflake schema model(Jain, 2023)

The snowflake structure offers significant advantages, particularly in enhancing query performance by minimizing disk storage requirements and consolidating lookup tables (Hüsemann et al., 2000; Levene et al., 2003). This design facilitates a higher level of scalability in connecting dimensional levels and elements, contributing to a more adaptable and efficient system. Additionally, the absence of redundancy simplifies maintenance processes, streamlining the overall management of the schema. These characteristics make the snowflake structure a favorable choice for databases where optimizing query efficiency, scalability, and ease of maintenance are paramount considerations.

Comparison between the star and snowflake schema

According to Iqbal et al (2020), there are differences between the star and snowflake schema to be considered:

- ***Query Complexity:*** The star schema is simpler for querying compared to the snowflake schema. The star schema's denormalized form with fewer tables makes it easier to retrieve data, while the normalized tables in the snowflake schema result in more complex queries, especially when dealing with a large number of tables and joins.

- ***Execution Time:*** The execution time differs between star and snowflake schemas. Star schema, with its redundant data and larger tables due to lack of splitting, experiences some delays during traversal, making it somewhat time-consuming. In contrast, the normalized tables in the snowflake schema, with no redundancy and smaller, split tables, contribute to faster query execution.

- ***Effect on the size of Data Warehouse:*** Due to the redundancy of data, the Star schema requires more space to store the data than the Snowflake schema:

- ***Result:*** When we apply star schema, execute the query, and get results, the results of snowflake schema has some differences from star schema because of its queries.

- ***Query Optimization:*** Both star and snowflake schemas can benefit from performance improvements through the addition of bitmap indexing. While snowflake schemas generally exhibit good performance figures, the example can be extended to both schemas to explore and enhance their respective efficiencies further.

Each schema serves a specific purpose and has distinct benefits. However, in this report, we have chosen the snowflake schema as the main model for building the data warehouse because it aligns with the structure of the sales table and the selected component tables by our team. Besides that, measures and subdimension tables will also be suitable for querying with snowflake schema style.

2.1.4. ETL

According to Vassiliadis et al (2002), ETL stands for Extract, Transform, Load and it is a data warehousing process employed to extract information from diverse sources, convert it into a format suitable for integration into a data warehouse, and subsequently load it into the warehouse. The ETL process can be segmented into the following three phases (Albrecht et al., 2008):

Extract

The initial phase of the ETL process involves extraction. During this stage, data is gathered from diverse source systems, which may exist in various formats such as relational databases, NoSQL, XML, and flat files, and is transferred to the staging area. It is crucial to extract data from different source systems and deposit it into the staging area before directly loading it into the data warehouse. This precaution is taken because the extracted data comes in various formats and is susceptible to corruption. Directly loading it into the data warehouse could potentially compromise its integrity, making rollback procedures more challenging. Therefore, this step stands out as one of the pivotal elements in the ETL process.

Transform

The next stage in the ETL process is transformation. During this phase, a series of rules or functions is implemented on the extracted data to unify it into a standardized format. This step encompasses various processes and tasks, including:

- ***Filtering:*** Selectively loading specific attributes into the data warehouse.
- ***Cleaning:*** Addressing NULL values by replacing them with default values, and mapping variations.
- ***Joining:*** Merging multiple attributes into a single entity.
- ***Splitting:*** Dividing a singular attribute into multiple attributes.
- ***Sorting:*** Arranging tuples based on a specific attribute, usually a key attribute.

Load

The final phase of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The specific rate and timing of loading are entirely contingent on the system's requirements and can differ from one system to another.

2.2. OLAP Cube

Online Analytical Processing (OLAP) is a category of computer programs and software tools that empower users to interactively analyze multidimensional data from various perspectives. OLAP systems are designed to support complex queries and facilitate dynamic reporting, allowing users to gain insights into data trends, relationships, and patterns. The primary goal of OLAP is to provide a fast and interactive way to explore and analyze large volumes of data, promoting a more intuitive and user-friendly approach to data analysis.

OLAP leverages a multidimensional approach to data analysis, enabling users to efficiently explore and understand complex datasets. By organizing data into a cube structure, OLAP facilitates intuitive navigation through various dimensions, allowing users to perform ad-hoc queries and generate dynamic reports. The multidimensional nature of OLAP provides a holistic view of data, fostering a deeper understanding of relationships and trends within the information.

An OLAP cube, also known as a multidimensional cube or hypercube, is a key component of OLAP systems. It organizes data into a multidimensional structure, enabling users to navigate through different dimensions and levels to analyze data interactively. The cube represents data in an easily comprehensible and manipulable format, typically featuring measures (numeric data) along with dimensions (categories or attributes). These dimensions can include various aspects such as time, geography, products, and customers, providing a comprehensive view of the data.

The OLAP cube serves as the foundation for this multidimensional analysis, structuring data in a way that aligns with the inherent structure of the business. Measures within the cube represent the quantitative aspects of the data, while dimensions provide context and categorization. Users can easily drill down, roll up, or pivot within the cube, gaining insights into specific areas of interest

without the need for complex SQL queries or programming skills. This user-friendly and interactive approach positions OLAP as a valuable tool for decision-makers and analysts seeking to extract meaningful insights from large and intricate datasets.

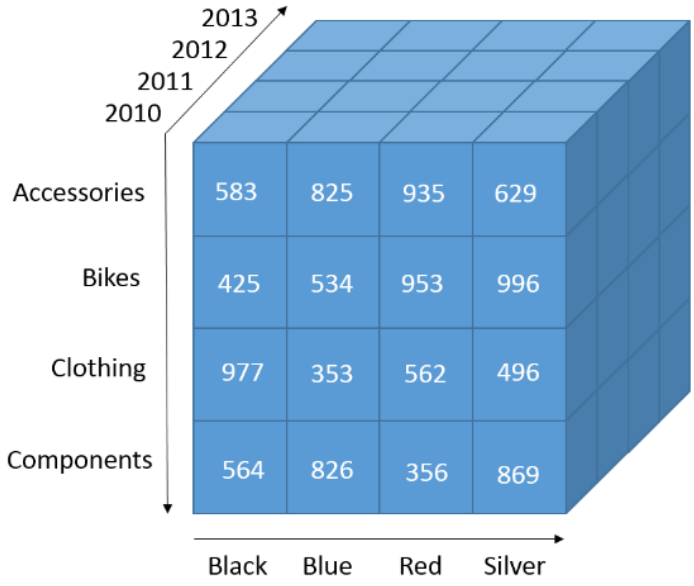

Figure 2- 4: OLAP Cube (Source: grapecity.com)

Chapter 3: Requirements Analysis

Business process and challenges

Adventure Works Cycles (AWC) is facing a significant challenge in their overall sales performance. Despite efforts to enhance their market presence, AWC is struggling to integrate and manage multi-source data effectively, which is impacting their sales achievements across all channels. This issue is particularly critical for the sales and marketing department, tasked with optimizing sales strategies and improving overall performance.

To effectively address this challenge, it is imperative to construct a data warehouse that consolidates and organizes data from Adventure Works Cycles (AWC)'s two vital sales channels: Resellers and Internet/online sales. In addition, a detailed examination of the sales data for cycling accessories and components, a significant contributor to AWC's financial success, is crucial. This data warehouse will serve as a foundational tool for gaining insights and driving sales performance improvements across all channels.

In this project, we will assume that the sales data comes from a sophisticated database integrated within the system. The Production data will be obtained from the Production Department, which is responsible for sending the necessary files. On the other hand, the Purchasing data will be sourced from a distinct database. This methodology ensures a comprehensive and precise data set, which is crucial for effectively addressing the multi-source data management challenges at AWC and enhancing their overall sales performance.

Chapter 4: Experimental modeling

4.1. Data warehouse designing concept

4.1.1. Overall architecture

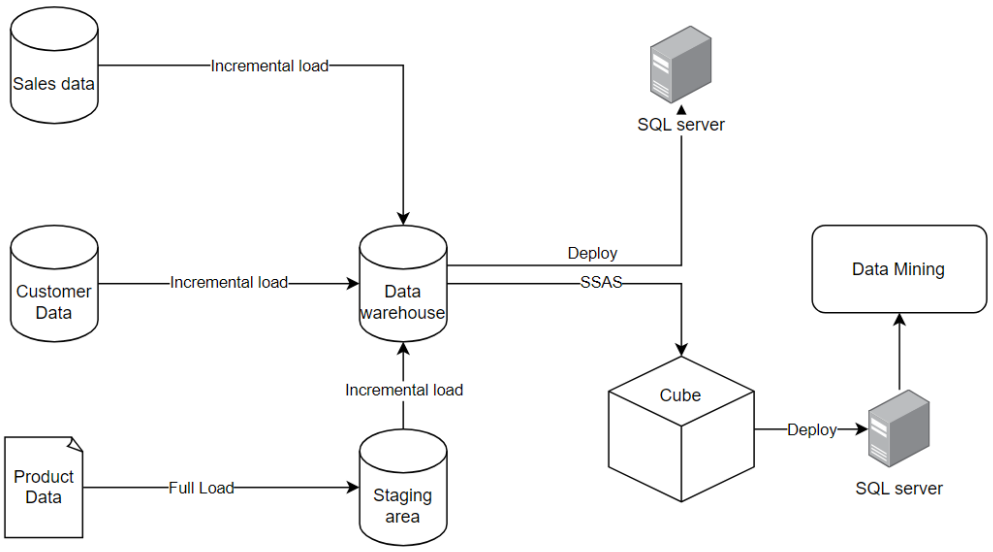

Figure 4- 1: Overal data warehouse architecture

```
Data/
    Category/
        Product_Category_data.tsv
    Culture/
        Product_Culture_data.xlsx
    Description/
        Product_Description_data.xlsx
    ModelDescriptionCulture/
        Product_Description_Model_data.xlsx
    Product/
        Product_data.xlsx
    ProductModel/
        ProductModel_data.xlsx
    Subcategory/
        Product_Subcategory_data.xml
    Subcategory_Schema/
        Product_Subcategory_data.xsd
```

Figure 4- 2: Folder hierarchy stucture

In the context of data organization, the initial focus lies on the data supplied by the Product Department. The process begins by loading the provided data into the staging area. This staging area serves as an interim repository for the incoming data before it undergoes further processing and integration. The data comprises seven files related to Category, Culture, Description, ModelDescriptionCulture (middle table), Product, Product Model, and Product Subcategory and organized follow a structure. In the event of receiving a new file, a systematic classification process is in place to ensure its placement into the appropriate folder (***Figure 4-2***) within the existing structure. The loading mechanism employs a full-load process, encompassing both the schema and content of each file. Subsequently, the staged data is transitioned into the Data Warehouse for comprehensive storage and accessibility.

In tandem with the product data, the data integration process extends to encompass customer data and sales data. This involves the extraction of customer information from the system and the collection of sales data from OLAP sources.

Similar to the product data workflow, the customer data and sales data undergo an initial full-load process. This process ensures the comprehensive incorporation of all relevant information. Subsequent to the initial load, the system is configured for incremental loading. This approach facilitates the ongoing capture and integration of new and updated data, allowing for real-time insights and analysis.

The integrated customer and sales data contribute to a holistic understanding of business performance. The combination of product, customer, and sales data sets the stage for a comprehensive evaluation of sales performance, enabling informed decision-making. The coordinated effort between the different data sources ensures a unified and up-to-date representation of the business landscape.

4.1.2. Data staging architecture

The Data Staging area is purpose-built for handling files from the Product Department. Utilizing a database on the server, schemas are predefined based on the file structure.

Table 4- 1: TabTbl_logs table

Attribute	Datatype
ID	int
StepName	varchar (200)
FlatFileName	varchar (200)
RecordsInserted	int
FileComponent	varchar(200)
Date	datetime

Table 4- 2: ProductModelDescription table

Attribute	Datatype
ProductModelID	int
ProductDescriptionID	int
CultureID	varchar (55)
LoadedDate	datetime

Table 4- 3: Product model

Attribute	Datatype
ProductID	int
Color	nvarchar(50)
ProductName	nvarchar(255)
ReOrderPoint	int
StandardCost	decimal(18,2)
Size	nvarchar(255)
Class	nvarchar(255)
ProductLine	nvarchar(50)
Style	nvarchar(50)
ProductNumber	nvarchar(50)
MakeFlag	bit
FinishedGoodsFlag	bit
SafetyStockLevel	int
ListPrice	decimal(18,2)
SizeUnitMeasureCode	nvarchar(3)
WeightUnitMeasureCode	nvarchar(3)
Weight	decimal(18,2)
DaysToManufacture	int

ProductSubcategoryID	int
ProductModelID	int
DiscountinuedDate	datetime
rowguid	nvarchar(255)
ModifiedDate	datetime
LoadedDate	datetime

Table 4- 4: Culture table

Attribute	Datatype
CultureID	nvarchar(255)
Name	nvarchar(255)
ModifiedDate	datetime
LoadedDate	datetime

Table 4- 5: ProductDescription table

Attribute	Datatype
ProductDescriptionID	int
Description	nvarchar(MAX)
rowguid	nvarchar(255)
ModifiedDate	datetime
LoadedDate	datetime

Table 4- 6: ProductSubcategory table

Attribute	Datatype
ProductSubcategoryID	int
ProductCategoryID	int
Name	nvarchar(255)
rowguid	nvarchar(255)
ModifiedDate	datetime
LoadedDate	datetime

Table 4- 7: ProductModel table

Attribute	Datatype
ProductModelID	int
Name	nvarchar(255)
CatalogDescription	nvarchar(255)
Istructions	nvarchar(max)
ModifiedDate	datetime
LoadedDate	datetime

Table 4- 8: ProductCategory table

Attribute	Datatype
ProductCategoryID	int
Name	nvarchar(255)
rowguid	nvarchar(255)
ModifiedDate	datetime
LoadedDate	datetime

4.1.3. Data warehouse architecture

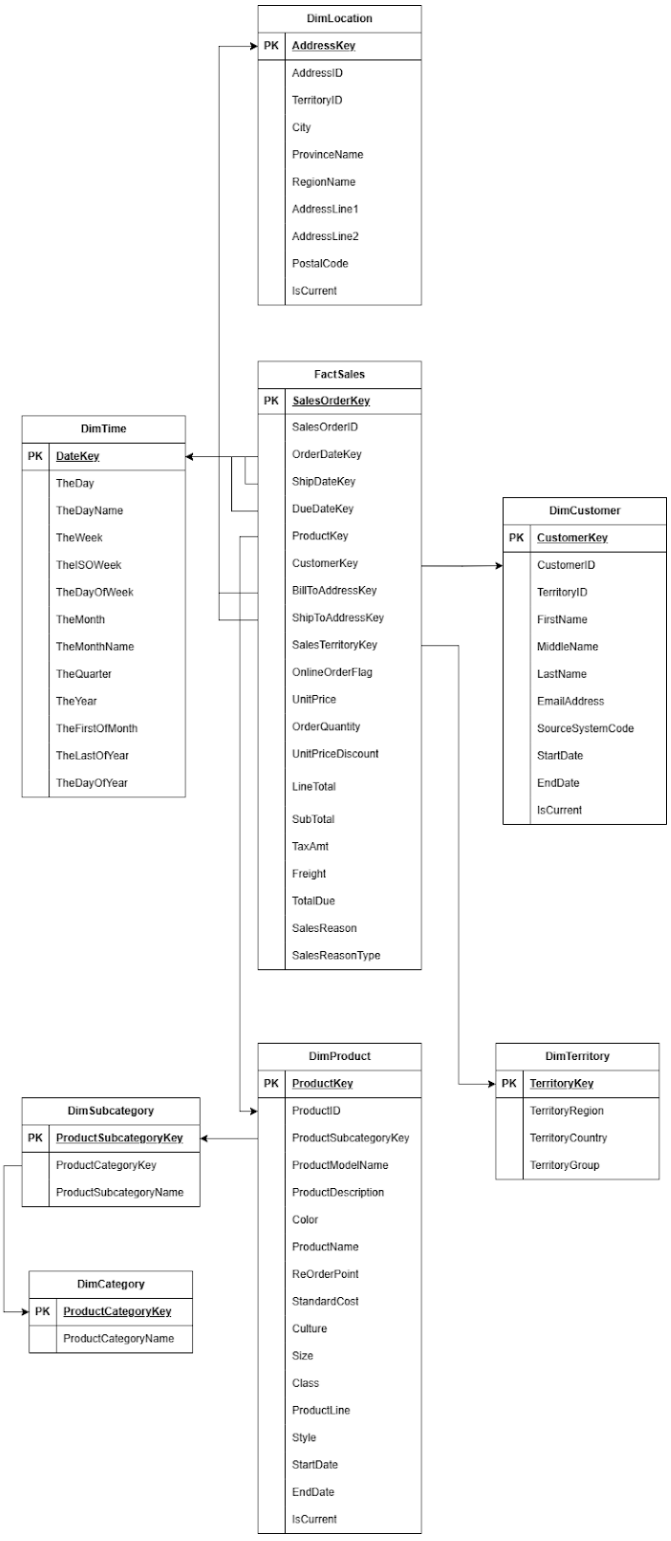

Figure 4- 3: Entity-Relationship Diagram

As shown in the above Entity-Relationship Diagram (ERD), our team decided to construct a robust Data Warehouse. Along with the primary FactSales table, this warehouse has seven dimension tables: DimLocation, DimTerritory, DimCustomer, DimProduct, DimCategory, DimSubcategory, and DimTime. This selection has been carefully constructed to considerably improve the analytical capabilities for in-depth sales and business performance analysis.

The **DimLocation** table plays a pivotal role by offering comprehensive geographic insights, enabling us to dissect sales performance across distinct regions and addresses. Attributes like TerritoryID, ProvinceName, and RegionName enhance our ability to recognize regional patterns, which is useful for firms with a wide range of geographic operations.

DimTerritory serves as a temporal lens into the evolution of sales territories over time. The inclusion of variables such as TerritoryRegion, TerritoryCountry, and TerritoryGroup enables us to study historical trends and adjust strategies to changing territorial dynamics.

DimCustomer appears as an important dimension for fine-grained customer analysis. Incorporating information such as FirstName, LastName, PhoneNumber, and EmailAddress enables for a more in-depth insight of client behavior and preferences. The TerritoryKey connection to DimTerritory enables region-specific customer information.

The **DimProduct**, **DimCategory**, and **DimSubcategory** tables capture the dynamic nature of items and their classification. These dimensions are derived from flat files and sheets that have been loaded into the staging area. ProductModelName, ProductDescription, ProductCategoryName, and ProductSubCategoryName attributes provide a thorough grasp of product qualities and classifications, which is critical for adapting marketing and sales tactics.

The **DimTime** table, which was built solely to increase query efficiency by giving critical time-based properties such as TheDay, TheMonth, and TheYear, adds to the temporal dimension. This temporal lens allows for faster assessments of seasonality, trends, and general sales patterns throughout time.

The **FactSales** table is at the core of our system, unifying transactional data and connecting it to the dimension tables. Foreign key constraints, such as FK_FactSales_DimTime and FK_FactSales_DimProduct, ensure data integrity by assuring proper dimensions-to-fact linkages.

Our rationale for selecting these specific tables and attributes is rooted in their direct relevance to sales and business performance analysis. The aforementioned collection of dimensions covers a broad range of business aspects, while the central fact table consolidates the transactional data required for a comprehensive picture. The architecture is built for versatility and scalability, allowing for future changes to meet evolving company demands.

Regarding the FactSales measures: When building the OLAP Cube based on the FactSales database, we included a set of essential measures to allow for a thorough examination of sales transactions. Fact Sales Count, Freight, Line Total, Order Quantity, Sub Total, Tax Amt, Total Due, Unit Price, and Unit Price Discount are among these measurements. Each metric has a specific purpose, from calculating transaction counts to assessing the impact of discounts, and together they provide a comprehensive picture of sales performance. The inclusion of these metrics is consistent with our goal of building a data model that provides sophisticated business intelligence and reporting capabilities.

Table 4- 9: Dimension and Fact tables

Type	Name	Attributes
Dimension	***DimTime***	DateKey, TheDay, TheDayName, TheWeek, TheISOWeek, TheDayOfWeek, TheMonth, TheMonthName, TheQuarter, TheYear, TheFirstOfMonth, TheLastOfYear, TheDayOfYear.
	DimProduct	ProductKey, ProductID, ProductSubcategoryKey, ProductModelName, ProductDescription, Color, ProductName, ReOrderPoint, StandardCost, Culture, Size, Class, ProductLine, Style, StartDate, EndDate, IsCurrent.
	DimCustomer	CustomerKey, CustomerID, TerritoryID, FirstName, MiddleName, LastName,

		EmailAddress, SourceSystemCode, StartDate, EndDate, IsCurrent.
	DimLocation	AddressKey, AddressID, TerritoryID, City, ProvinceName, RegionName, AddressLine1, AddressLine2, PostalCode, IsCurrent.
	DimTerritory	TerritoryKey, TerritoryRegion, TerritoryCountry, TerritoryGroup.
	DimCategory	ProductCategoryKey, ProductCategoryName.
	DimSubcategory	ProductSubcategoryKey, ProductCategoryKey, ProductSubcategoryName.
Fact	***FactSales***	SalesOrderKey, SalesOrderID, OrderDateKey, ShipDateKey, DueDateKey, ProductKey, CustomerKey, BillToAddressKey, ShipToAddressKey, SalesTerritoryKey, OnlineOrderFlag, UnitPrice, OrderQuantity, UnitPriceDiscount, LineTotal, SubTotal, TaxAmt, Freight, TotalDue, SalesReason, SalesReasonType.

4.1.4. Data warehouse modeling type

From the selected dim and fact tables, we have chosen the snowflake schema model as the modeling type to build the data warehouse. With this model, the data warehouse for sales performance that we construct will have advantages in terms of saving storage space and greater flexibility in changing the structure (if necessary).

4.2. ETL process

4.2.1. ETL process for Data Staging

In the Data Staging Area's ETL process, we employ a meticulous sequence to seamlessly integrate new files. Instead of a direct confirmation check, a foreach loop is initiated to iterate through all file names. Within the loop, the “Execute SQL Command” component is utilized to verify the loading status of each file. Once confirmed, the data flow is triggered to load the respective file into the designated table. Simultaneously, vital information, specifically the file name, is logged into the tbl_logs table. This iterative approach, encapsulated within a foreach loop, significantly enhances the adaptability and efficiency of the overall data integration workflow. For example, figure 4-4 and 4-5 is the ETL process of Culture table in the Data Staging Area.

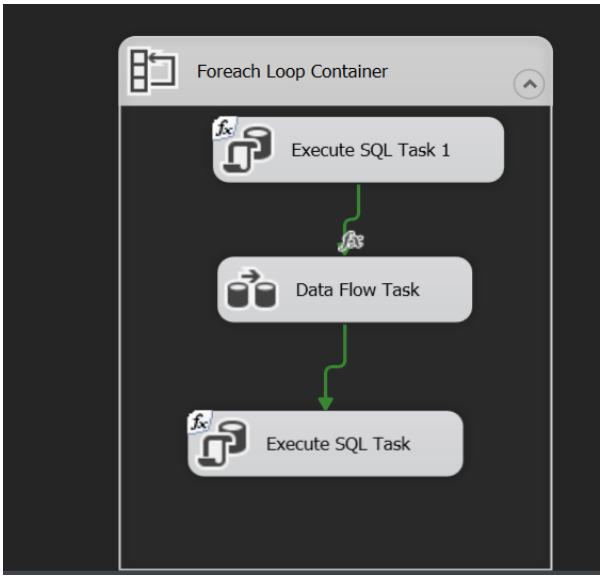

Figure 4- 4: Control flow of staging area

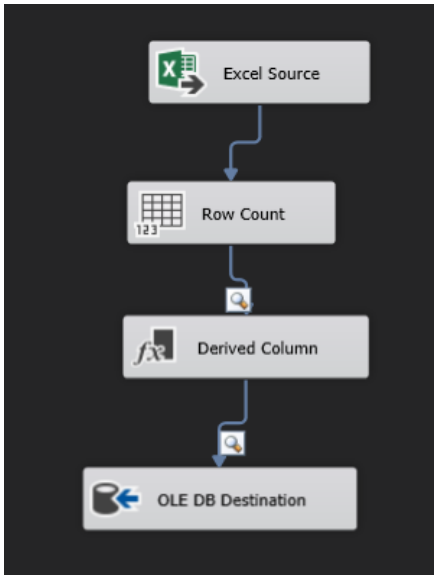

Figure 4- 5: Data flow of staging area

4.2.2. ETL process for Data Warehouse

DimCustomer

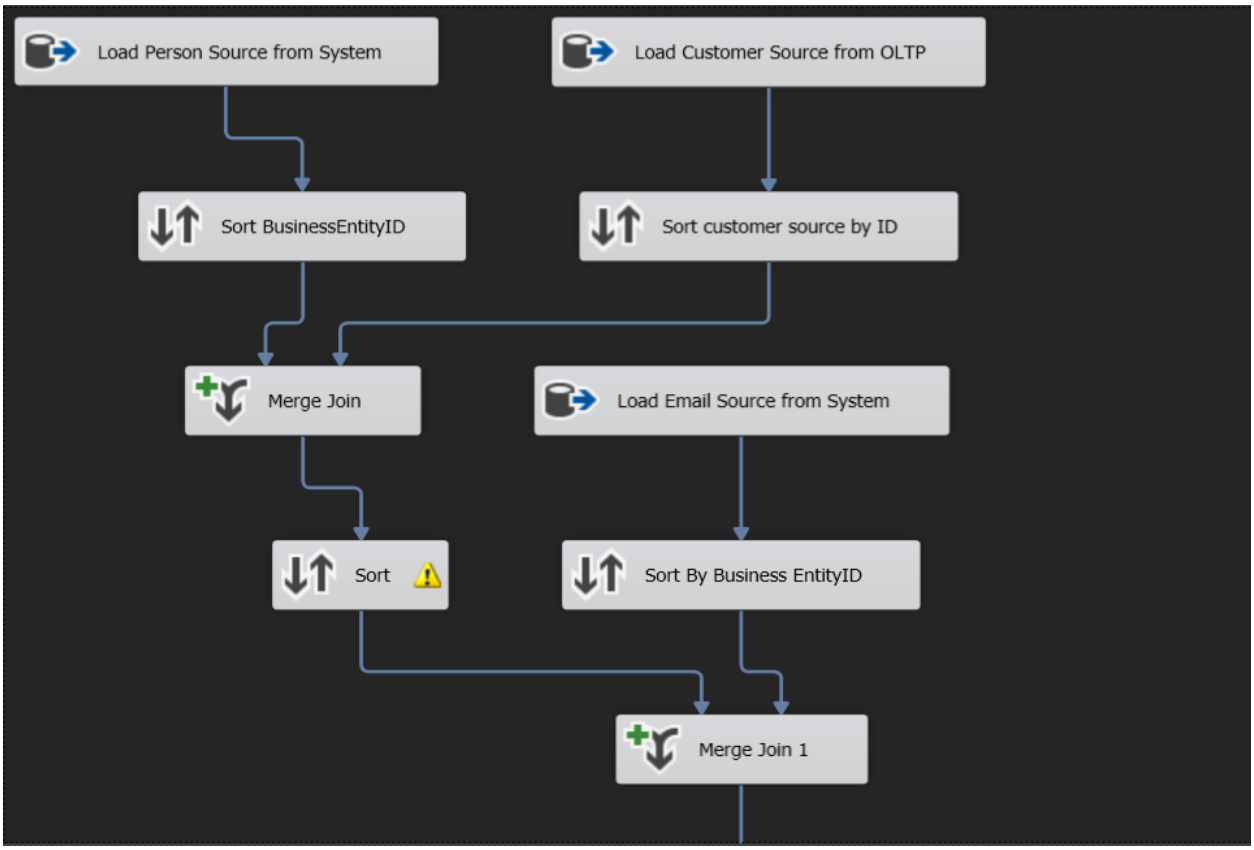

Figure 4- 6: Customer dimension ETL process (1)

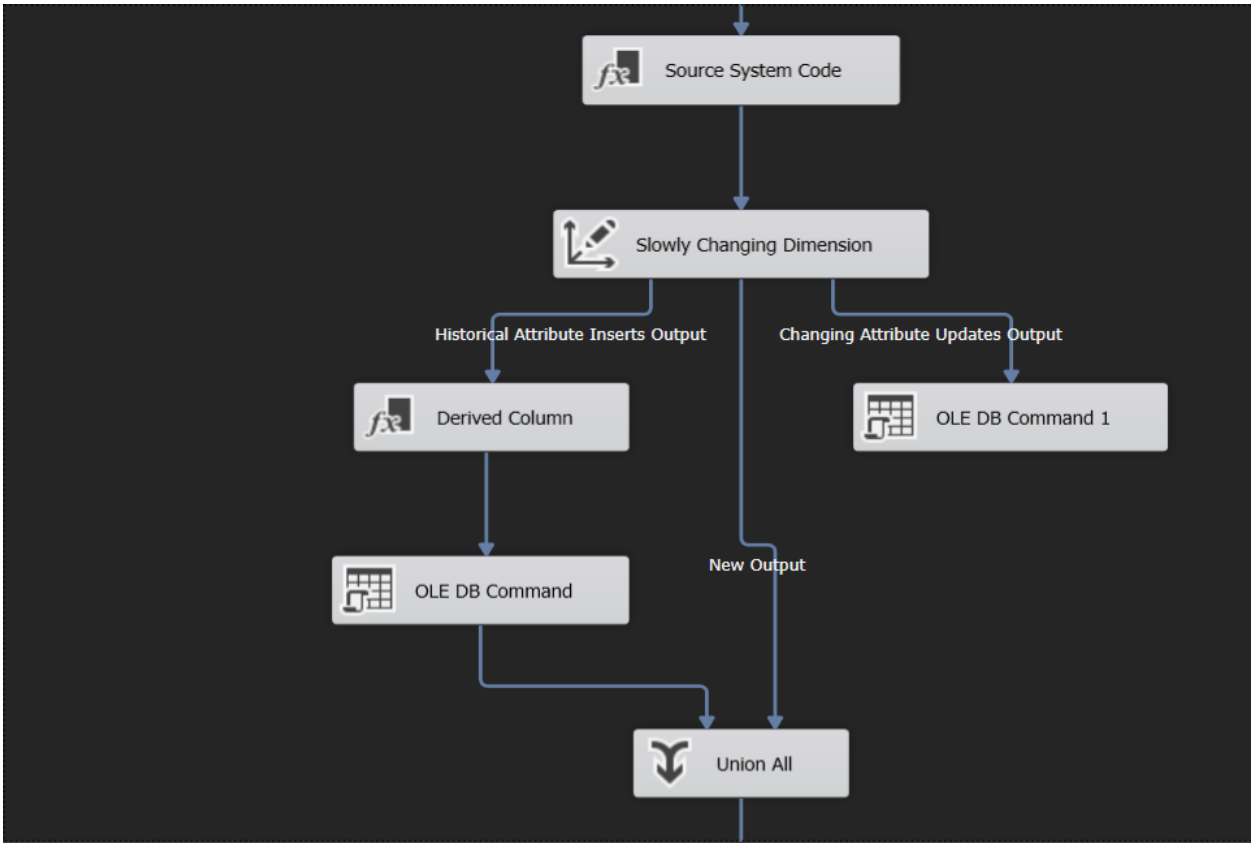

Figure 4- 7:Customer dimension ETL process (2)

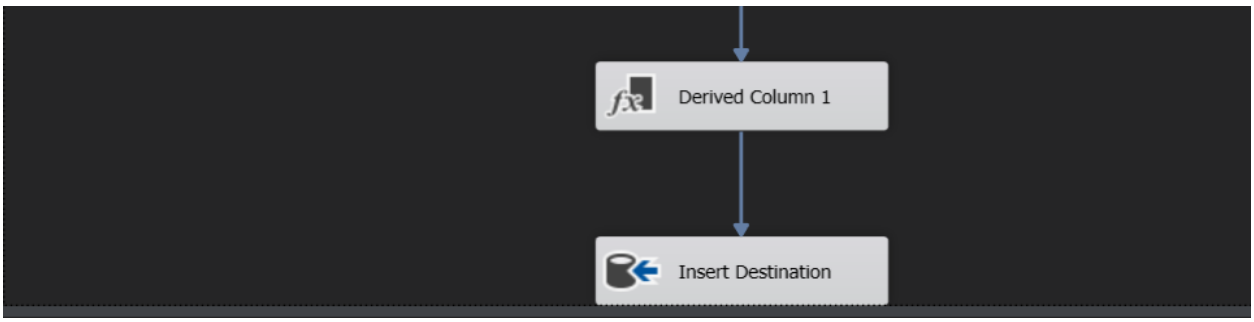

Figure 4- 8: Customer dimension ETL process (3)

The ETL process to populate the DimCustomer table initiates with data extraction from AdventureWorks2019's Sales.Customer and Person.Person tables joined on the BusinessEntityID column. This step is followed by a transformation where specific columns are selected and mapped to construct the target table's structure. Key transformations include creating a CustomerKey from CustomerID and populating TerritoryKey with TerritoryID. We then apply Slowly Changing Dimension (SCD) Type 2 logic, marking new values as 'current' in the IsCurrent column. The final

stage involves loading the transformed, SCD-adjusted data into the DimCustomer table, effectively managing the evolution of customer information.

DimLocation

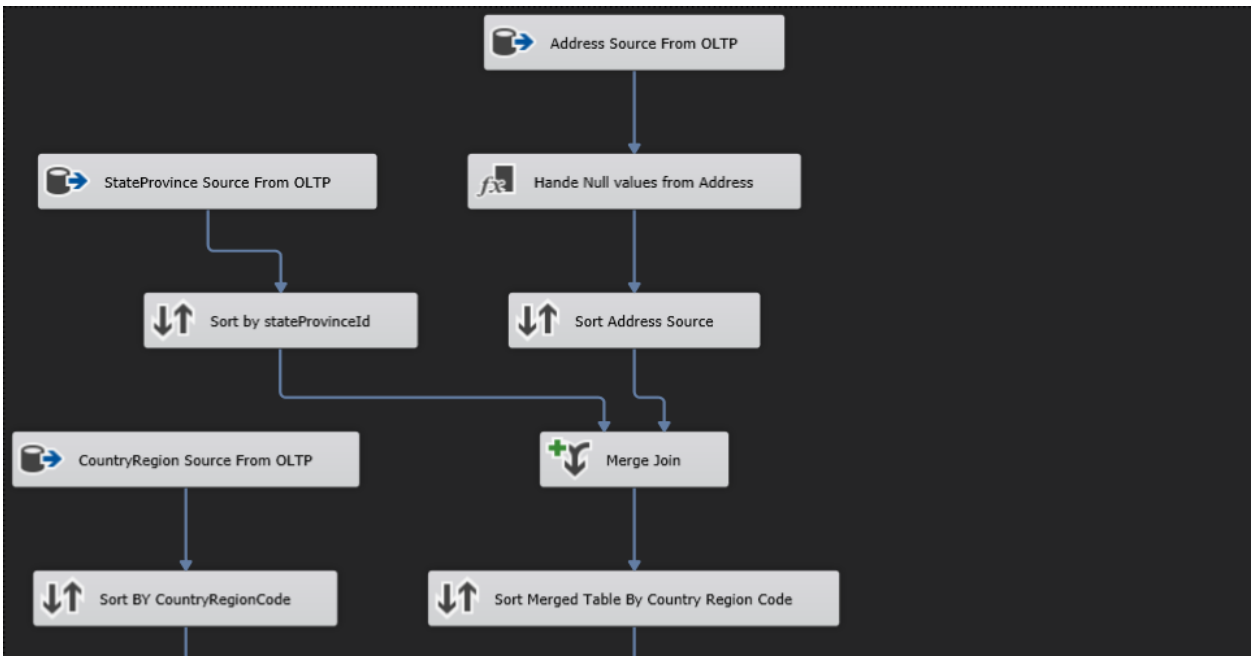

Figure 4- 9: Location dimension ETL process (1)

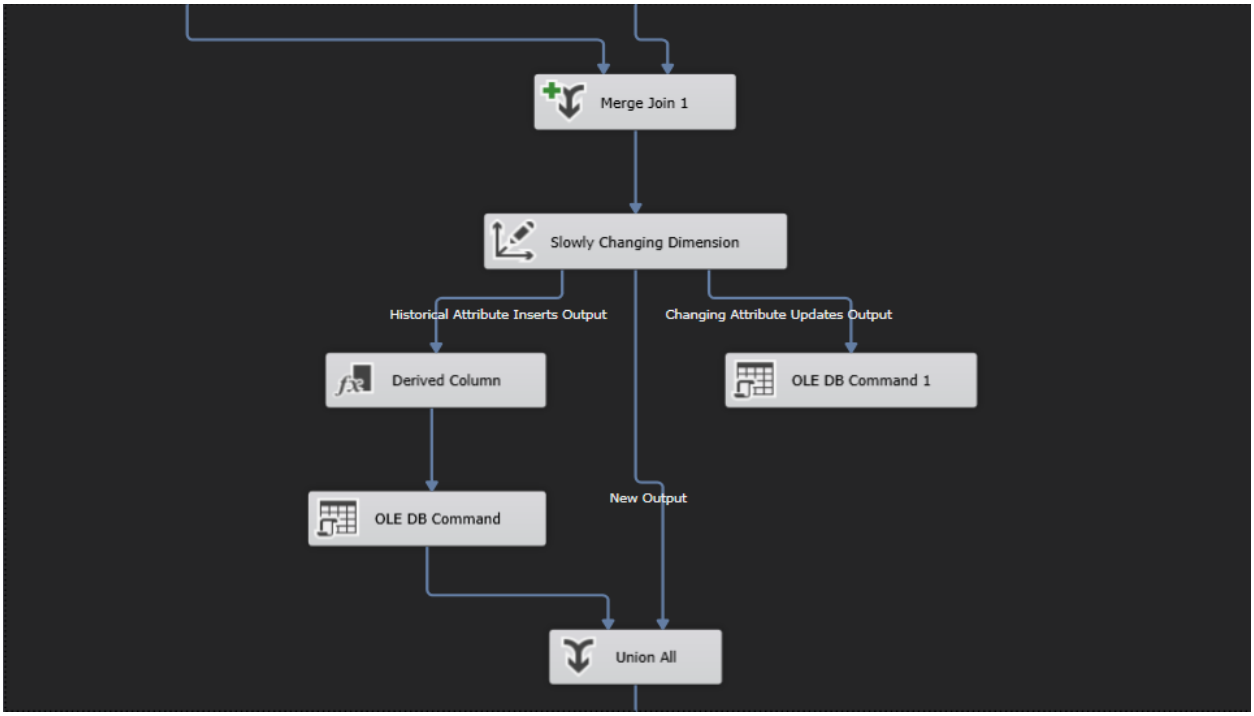

Figure 4- 10: Location dimension ETL process (2)

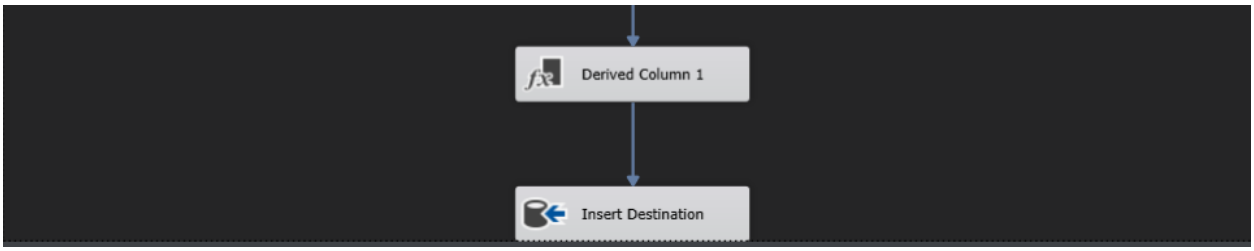

Figure 4- 11: Location dimension ETL process (3)

For the DimLocation table, data extraction occurs from AdventureWorks2019's Person.Address, Person.StateProvince, and Person.CountryRegion tables. The process involves JOIN operations on StateProvinceID and CountryRegionCode. We transform and map columns such as AddressID to AddressKey, along with TerritoryID, ProvinceName, and RegionName. After applying SCD Type 2 logic for historical data management, marked by IsCurrent, we load the transformed data into the DimLocation table, ensuring comprehensive historical location information.

DimTerritory

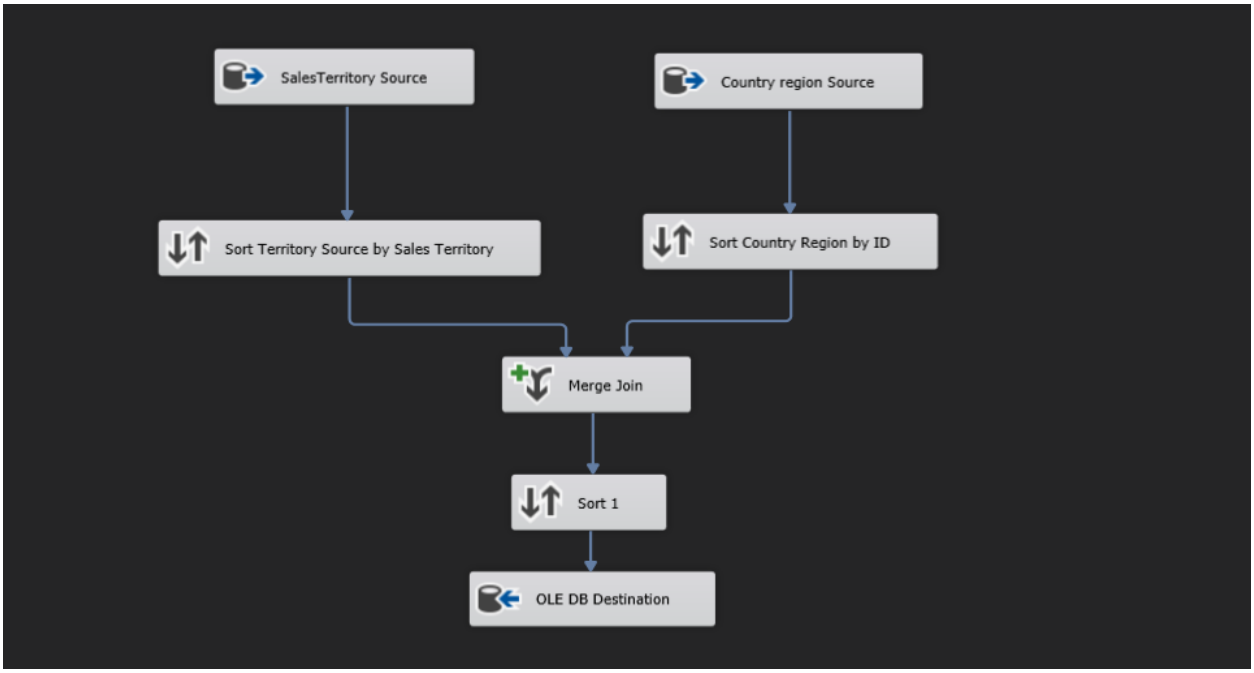

Figure 4- 12: Territory dimension ETL process

In constructing the DimTerritory table, data extraction involves the Sales.SalesTerritory and Person.CountryRegion tables from AdventureWorks2019, joined on CountryRegionCode. The transformation phase includes mapping TerritoryID to TerritoryKey, and aligning TerritoryRegion, TerritoryCountry, and TerritoryGroup from the respective source tables. This process culminates in populating the DimTerritory table with these transformed values.

DimCategory and DimSubcategory

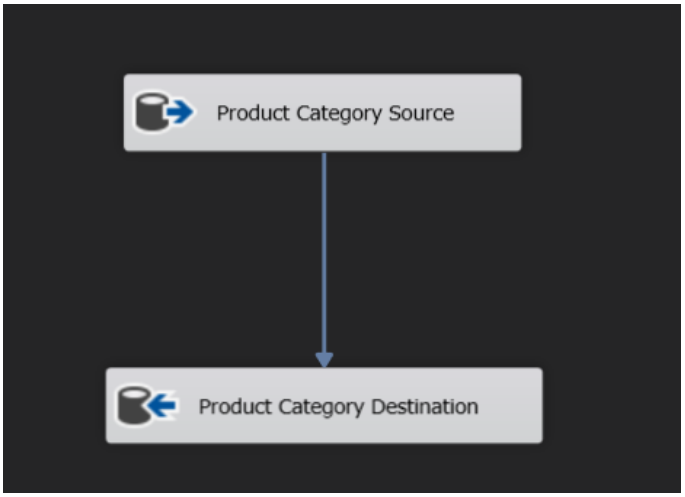

Figure 4- 13: Category dimension ETL process (1)

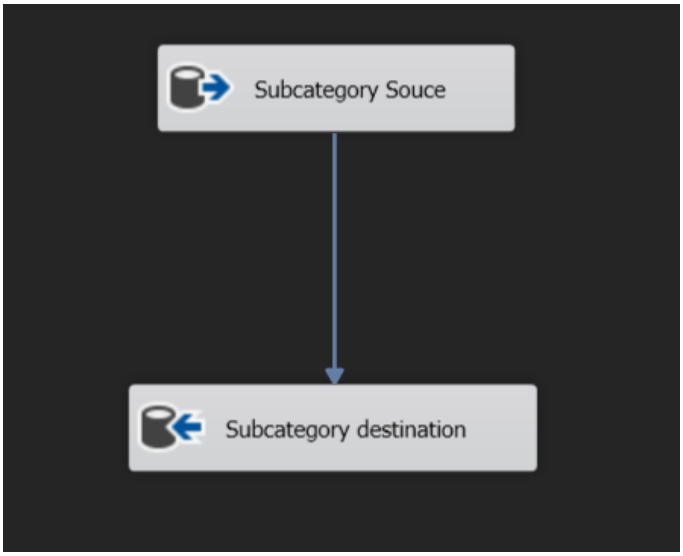

Figure 4- 14: Category dimension ETL process (2)

The process for creating both the Category and Subcategory dimension tables is straightforward. Data is extracted directly from the respective Staging.Category and Staging.Subcategory source tables into the destination. However, we ensure that only the newest data is loaded into each table.

DimProduct

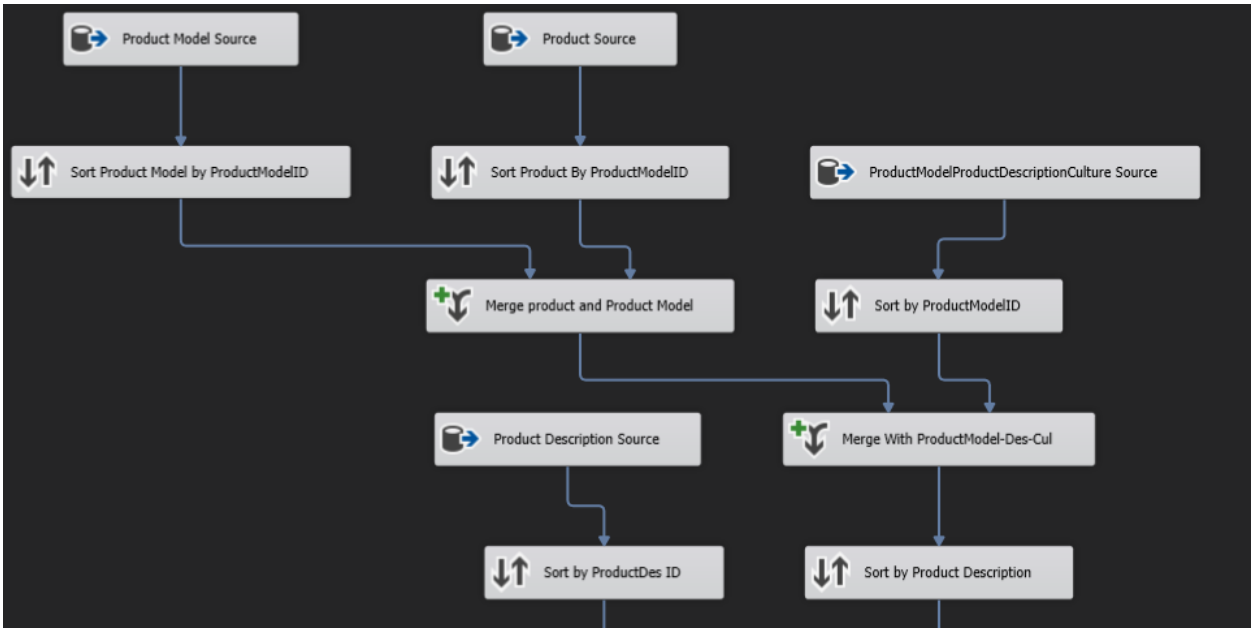

Figure 4- 15: Product dimension ETL process (1)

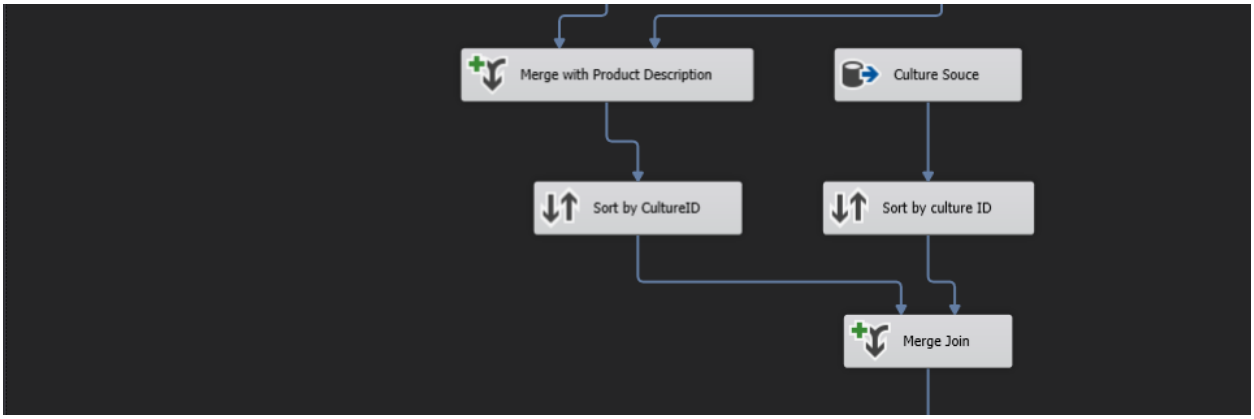

Figure 4- 16: Product dimension ETL process (2)

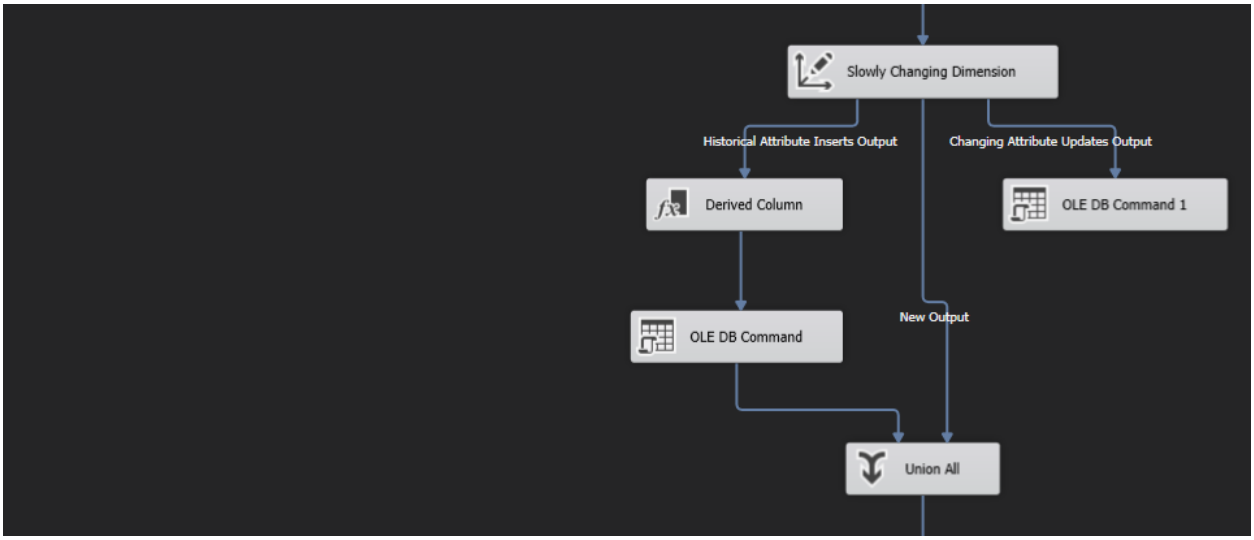

Figure 4- 17: Product dimension ETL process (3)

Figure 4- 18: Product dimension ETL process (4)

For the DimProduct table, the ETL process involves extracting data from multiple source tables in Staging: Product, Production.ProductModel, ProductModelProductDescriptionCulture, ProductDescription, and Culture. The extraction uses multiple JOIN operations on keys like ProductModelID, ProductDescriptionID, and CultureID. After applying SCD Type 2 logic to manage historical changes, marked by IsCurrent, the transformed data is loaded into the DimProduct table, ensuring it contains comprehensive and historical product information.

FactSales

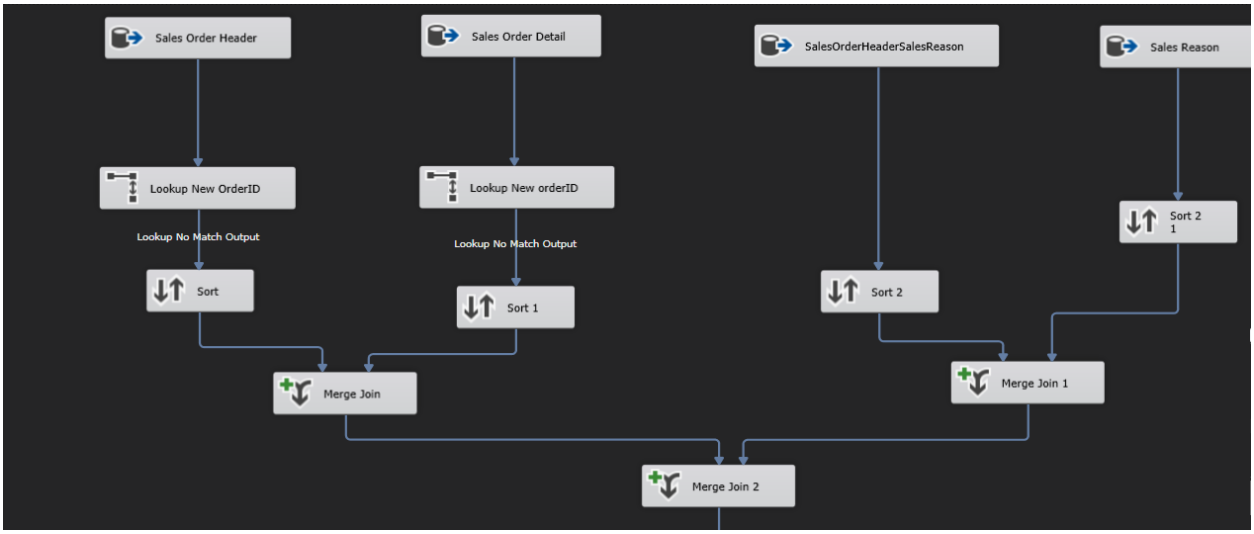

Figure 4- 19: Fact Sales ETL process (1)

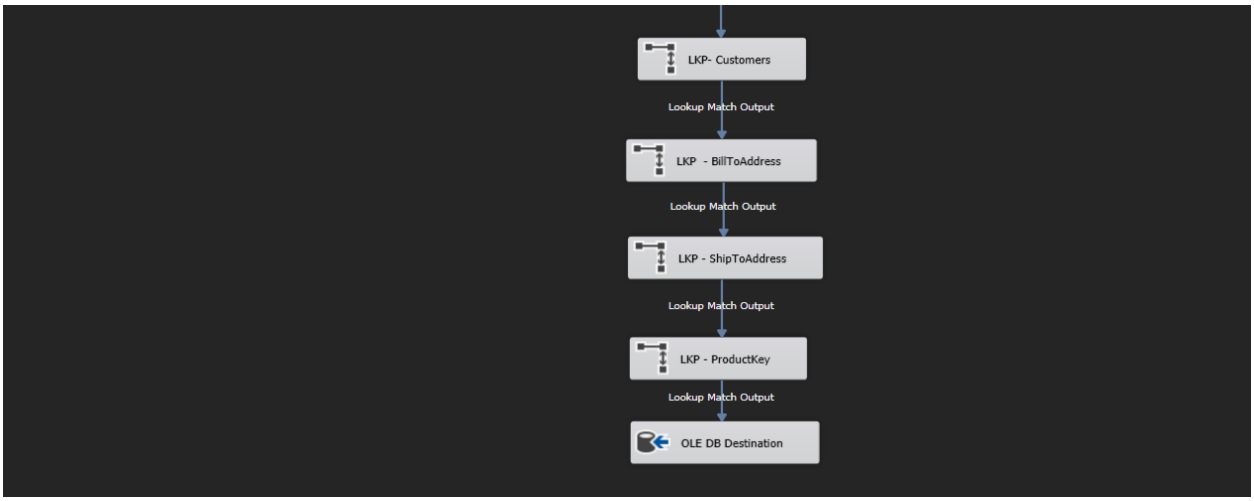

Figure 4- 20: Fact Sales ETL process (2)

Finally, for the FactSales table, data extraction comes from AdventureWorks2019's Sales.SalesOrderHeader, Sales.SalesOrderDetail, and Production.Product tables. The JOIN operations link these tables on SalesOrderID and ProductID. The selected columns are transformed and mapped to create the structure of the FactSales table, completing the ETL process for this critical component of the data warehouse.

4.3. SSAS OLAP Cube

From the listed tables and procedures, we proceed to build the cube. With this cube, querying and related analytical processes will become more straightforward. Below is an illustration depicting our cube.

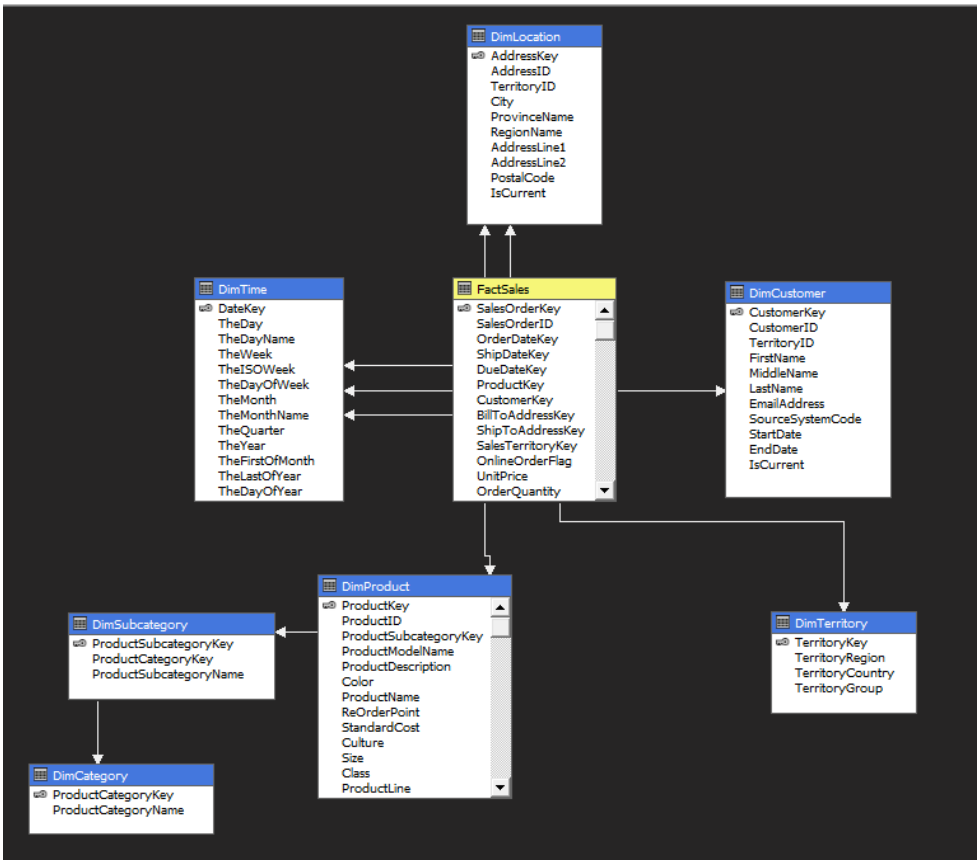

Figure 4- 21: Our OLAP cube is displayed by snowflake schema

Chapter 5: Discussion

5.1. Overall sales problem

What is the trend in sales growth or decline over the past quarter or year?

5.1.1. Overall

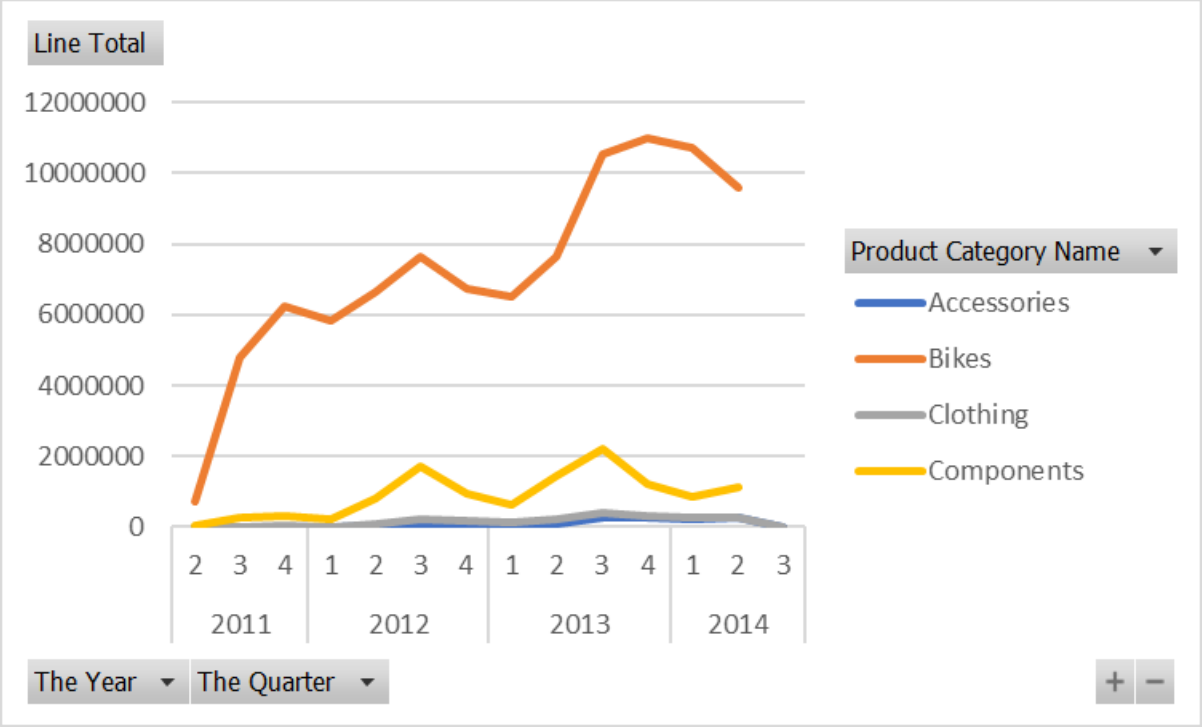

Figure 5- 1: Sales of 4 categories from 2011 to 2014

Over the past four years, the Bikes category has demonstrated significant sales growth, as evidenced by the Line Total measure reaching an impressive $11,000,000 by the end of 2013. The overall trend for Bikes appears positive, with a minor dip in the first quarter of 2013. This suggests the Bikes category as a key contributor to the overall revenue.

In contrast, the Clothing and Accessories categories display relatively stable trend lines, with Line Total values indicating more modest revenue contributions. While the Bikes and Components categories showcase significant sales figures, the Clothing and Accessories segments contribute to the overall revenue with more restrained Line Total values, suggesting a comparatively smaller

but still notable financial impact. The observed minor downturns in 2014 for Clothing and Accessories call for a closer examination of potential factors influencing these categories during that specific timeframe. Understanding these nuanced trends in revenue generation across different product categories is crucial for informed decision-making and strategic planning.

5.1.2. Accessories

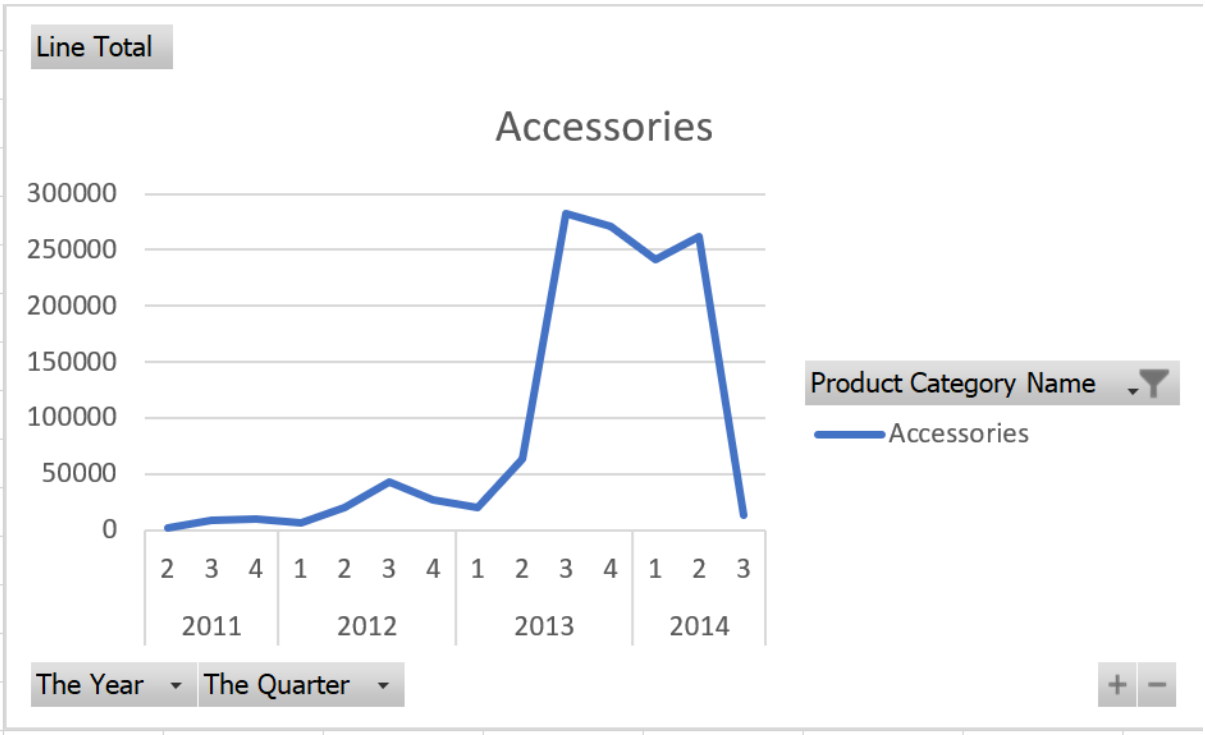

Figure 5- 2: Sales of accessories from 2011 to 2014

	TotalRevenue	MinRevenue	MaxRevenue	AvgRevenue	MedianRevenue
All	1272052.23999982	1.37	2549.81	65.153259577946	34.99
2011	20818.72	20.18	593.88	147.650496453901	121.11
2012	102432.99	11.99	1207.88	252.920962962963	201.85
2013	675015.769999931	1.37	2549.81	69.8051468459081	34.99
2014	473784.75999996	1.37	2071.43	50.900812204551	34.99

Figure 5- 3: Details of accessories's sales from 2011 to 2014

The Accessories category witnessed a notable increase in total revenue from 2011 to 2013, signaling a positive trend and heightened consumer interest. However, a distinct downturn became apparent in 2014, suggesting a potential shift or decline in market dynamics. This category, known for its diverse range of products, demonstrated versatility in meeting varied customer preferences,

as indicated by both its low minimum and maximum revenue values. The moderate average and median revenue further underscore the balanced distribution of sales within the Accessories category.

The intriguing sales pattern within Accessories becomes apparent when examining the Line Total figures. Experiencing a significant surge in 2013 with a peak revenue of over $250,000, the subsequent trend displayed a pronounced downgrade, dropping to below $10,000 in 2014. This abrupt decline raises questions about the influencing factors and market conditions during that specific period, prompting a deeper investigation to gain comprehensive insights into the observed fluctuations in revenue.

5.1.3. Bikes

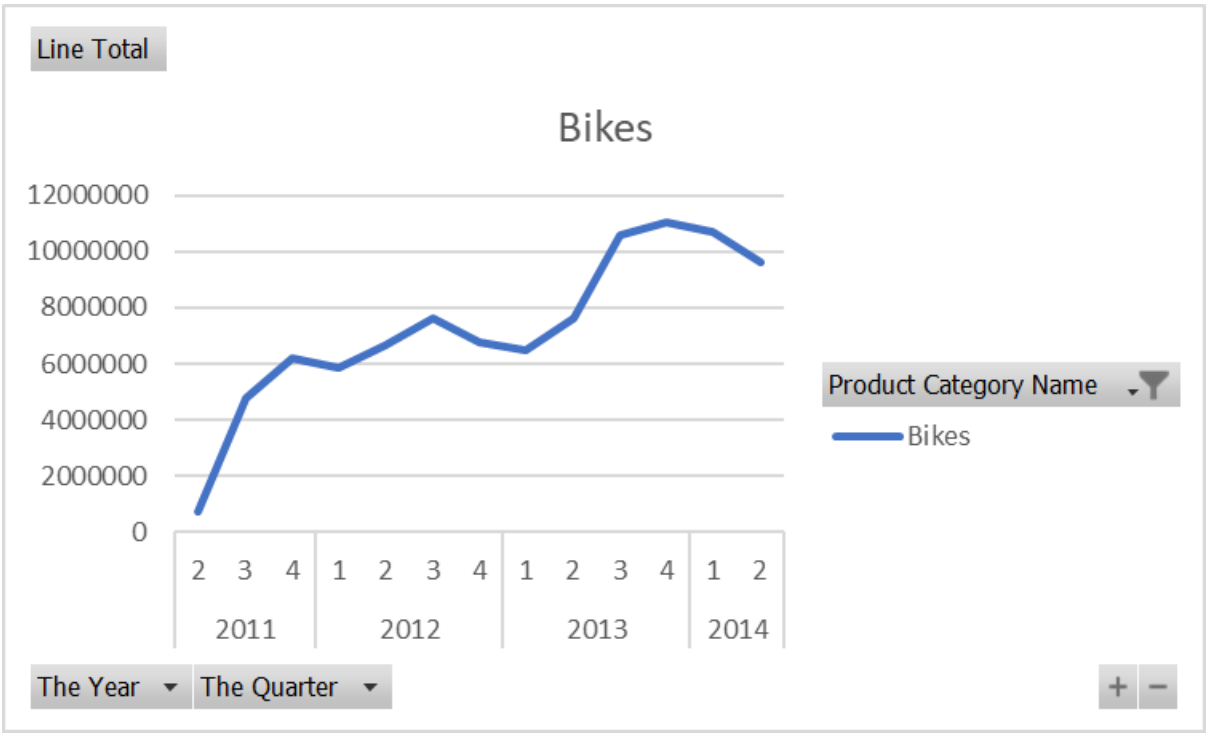

Figure 5- 4: Sales of bikes from 2011 to 2014

The Year	TotalRevenue	MinRevenue	MaxRevenue	AvgRevenue	MedianRevenue
(null)	94651051.0...	67.79	139803.28	5153.0406...	2294.99
2011	11945630.3...	419.45	117411.69	7531.9233...	3578.27
2012	28985470.0...	164.42	119643	7789.6990...	2443.35
2013	36266782.5...	283.94	139803.28	4794.0228...	2071.41
2014	17453168.0...	67.79	102397	3175.6128...	1700.99

Figure 5- 5: Detail of bikes's sales from 2011 to 2014

Bikes stand out as a substantial contributor to the overall revenue, marked by a notably high maximum revenue value. The relatively high average and median revenue values imply that sales are inclined towards higher-priced bikes within this category. The consistent increase in total revenue from 2011 to 2013 reinforces a positive trend, reflecting sustained consumer interest. However, a significant drop in revenue in 2014 compared to the peak in 2013 suggests a potential decline or shift in market conditions.

Examining the trend line for Bikes reveals a relatively stable pattern with slight fluctuations. There's a modest drop in early 2013, reaching $6,000,000, followed by an upward spike later in 2013, peaking at $11,000,000. The subsequent slight decline in 2014, dropping to below $10,000,000, underscores the need for a closer examination of the factors influencing consumer behavior and market dynamics during these specific periods. This nuanced analysis can provide valuable insights into the observed variations in revenue within the Bikes category.

5.1.4. Clothing

Figure 5- 6: Sales of clothing from 2011 to 2014

The Year	TotalRevenue	MinRevenue	MaxRevenue	AvgRevenue	MedianRevenue
(null)	2120495.04...	5.18	7128.3	214.69019...	53.99
2011	36030.07	5.18	823.43	137.51935...	91.79
2012	555571.64	5.18	7128.3	746.73607...	281.915
2013	1067666.82...	5.18	5139.71	227.55047...	53.99
2014	461226.509...	5.39	3785.88	110.36767...	49.99

Figure 5- 7: Detail of clothing's sales from 2011 to 2014

Clothing emerges with the lowest total revenue among the various categories, with both average and median revenue values suggesting that clothing items are generally priced lower compared to other product categories. Despite the lower pricing, Clothing exhibits a consistent increase in total revenue from 2011 to 2013, indicative of a positive trend. While there is a decline in revenue in 2014 compared to the peak in 2013, the drop is not as pronounced as observed in some other categories, indicating a more moderate downturn.

Analyzing the trend line for Clothing unveils two notable peaks in 2012 and 2013, reaching revenue values of $250,000 and $400,000, respectively. However, there is a significant drop in 2014, reflecting a decrease in sales. This trend underscores the need for a thorough investigation into the factors influencing consumer behavior and market dynamics during these specific years, shedding light on the observed fluctuations in revenue within the Clothing category.

5.1.5. Components

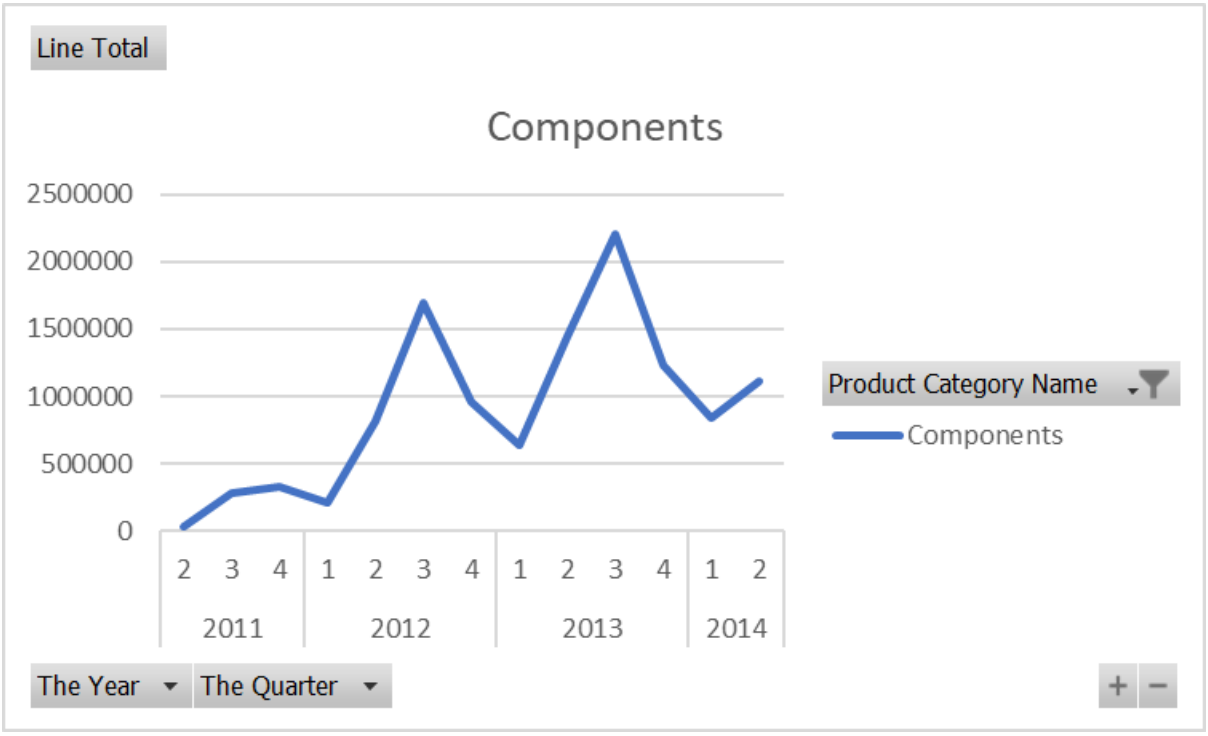

Figure 5- 8: Sales of components from 2011 to 2014

	TotalRevenue	MinRevenue	MaxRevenue	AvgRevenue	MedianRevenue
All	11802515.58	12.14	48097.11	4453.77946415095	2097.865
2011	639168.86	178.58	17177.72	2959.11509259259	1985.005
2012	3880731.67999999	20.52	35836.2	4856.98583229036	2638.32
2013	5612899.94	12.14	48097.11	4752.66718035563	1844.48
2014	1669715.1	12.14	28079.77	3677.78656387665	1482.915

Figure 5- 9: Detail of components's sales from 2011 to 2014

Components play a significant role in contributing to the overall revenue, with both average and median revenue reflecting a moderate but meaningful financial impact. The range between minimum and maximum revenue values suggests a diversity of component prices, showcasing a

varied product offering within this category. Components exhibit a consistent increase in total revenue from 2011 to 2013, indicating a positive and sustained trend. However, aligning with broader market patterns, there is a substantial decrease in revenue in 2014 compared to the peak in 2013, reflecting a noteworthy decline.

Delving into the trend line for Components reveals two distinct peaks in 2012 and 2013, reaching revenue values of $1,700,000 and $2,200,000, respectively. The subsequent drop to below $1,000,000 in the first quarter of 2014 suggests a significant downturn, followed by a slight recovery to above $1,000,000. This nuanced trend highlights the need for a detailed exploration into the factors influencing consumer behavior and market dynamics during these specific timeframes, offering insights into the observed fluctuations in revenue within the Components category.

```
WITH
 MEMBER [Measures].[TotalRevenue] AS
  SUM(
   FILTER(
    [Fact Sales].[Sales Order ID].[Sales Order ID].MEMBERS,
    [Measures].[Line Total] <> NULL
   ),
   [Measures].[Line Total]
  )

 MEMBER [Measures].[MinRevenue] AS
  MIN(
   FILTER(
    [Fact Sales].[Sales Order ID].[Sales Order ID].MEMBERS,
    [Measures].[Line Total] <> NULL
   ),
   [Measures].[Line Total]
  )

 MEMBER [Measures].[MaxRevenue] AS
  MAX(
   FILTER(
    [Fact Sales].[Sales Order ID].[Sales Order ID].MEMBERS,
    [Measures].[Line Total] <> NULL
   ),
   [Measures].[Line Total]
  )

 MEMBER [Measures].[AvgRevenue] AS
  AVG(
   FILTER(
    [Fact Sales].[Sales Order ID].[Sales Order ID].MEMBERS,
    [Measures].[Line Total] <> NULL
   ),
   [Measures].[Line Total]
  )
```

```
 MEMBER [Measures].[MedianRevenue] AS
  MEDIAN(
   FILTER(
    [Fact Sales].[Sales Order ID].[Sales Order ID].MEMBERS,
    [Measures].[Line Total] <> NULL
   ),
   [Measures].[Line Total]
  )

SELECT
 NON EMPTY {
  [Measures].[TotalRevenue],
  [Measures].[MinRevenue],
  [Measures].[MaxRevenue],
  [Measures].[AvgRevenue],
  [Measures].[MedianRevenue]
 } ON COLUMNS,
 NON EMPTY [Order Date].[The Year].MEMBERS ON ROWS
FROM [FinalLiveLaughLove]
WHERE [Dim Product].[Product Category Name].&[Accessories]
---
/*
WHERE [Dim Product].[Product Category Name].&[Bikes]
WHERE[Dim Product].[Product Category Name].&[Clothing]
WHERE[Dim Product].[Product Category Name].&[Components]
WHERE[Dim Product].[Product Category Name].[All].UNKNOWNMEMBER
*/
```

Figure 5- 10: MDX queries of 4 categories's sales from 2011 to 2014

5.2. Product-related problem

Which products have the highest sales throughout 4 years (from 2011 to 2014)?

```
SELECT
  NON EMPTY { [Measures].[Fact Sales Count] } ON COLUMNS,
  TOPCOUNT(
    [Order Date].[The Year].[The Year].MEMBERS * [Order Date].[The Quarter].[The Quarter] * [Dim Product].[Product Hierarchy].[Product].MEMBERS,
    50,
    [Measures].[Fact Sales Count]
  ) ON ROWS
FROM [DW Final Project]
```

The Year	The Quarter	Product Category	Product Subcategory	Product	Fact Sales Count
2014	1	Accessories	Bottles and Cages	Water Bottle - 30 oz.	1265
2013	4	Accessories	Bottles and Cages	Water Bottle - 30 oz.	1162
2013	3	Accessories	Bottles and Cages	Water Bottle - 30 oz.	1051
2014	2	Accessories	Bottles and Cages	Water Bottle - 30 oz.	1008
2014	1	Accessories	Tires and Tubes	Patch Kit/8 Patches	860
2013	4	Accessories	Tires and Tubes	Patch Kit/8 Patches	838
2013	3	Accessories	Tires and Tubes	Patch Kit/8 Patches	804
2013	4	Accessories	Tires and Tubes	Mountain Tire Tube	804
2014	1	Accessories	Tires and Tubes	Mountain Tire Tube	771
2013	3	Accessories	Tires and Tubes	Mountain Tire Tube	758
2014	2	Accessories	Tires and Tubes	Patch Kit/8 Patches	748
2014	2	Accessories	Tires and Tubes	Mountain Tire Tube	730
2014	1	Clothing	Caps	AWC Logo Cap	700
2014	1	Accessories	Helmets	Sport-100 Helmet, Red	664
2013	4	Accessories	Helmets	Sport-100 Helmet, Red	649
2014	1	Accessories	Helmets	Sport-100 Helmet, Blue	643
2013	4	Clothing	Caps	AWC Logo Cap	639
2014	1	Accessories	Helmets	Sport-100 Helmet, Black	621
2013	4	Accessories	Helmets	Sport-100 Helmet, Blue	607
2013	3	Clothing	Caps	AWC Logo Cap	605
2013	4	Accessories	Tires and Tubes	Road Tire Tube	604
2013	4	Accessories	Helmets	Sport-100 Helmet, Black	599
2013	3	Accessories	Helmets	Sport-100 Helmet, Black	587
2014	2	Accessories	Tires and Tubes	Road Tire Tube	582
2014	2	Accessories	Helmets	Sport-100 Helmet, Blue	578

Figure 5- 11: Top 50 highest sales products from 2011-2014

At first glance, the Accessories category is clearly prominent in Fact Sales Counts, indicating that consumers preferred to purchase accessories over other products in other categories. Additionally, a majority of these items are categorized under subgroups such as Bottles and Cages, Tires and Tubes, and Helmets.

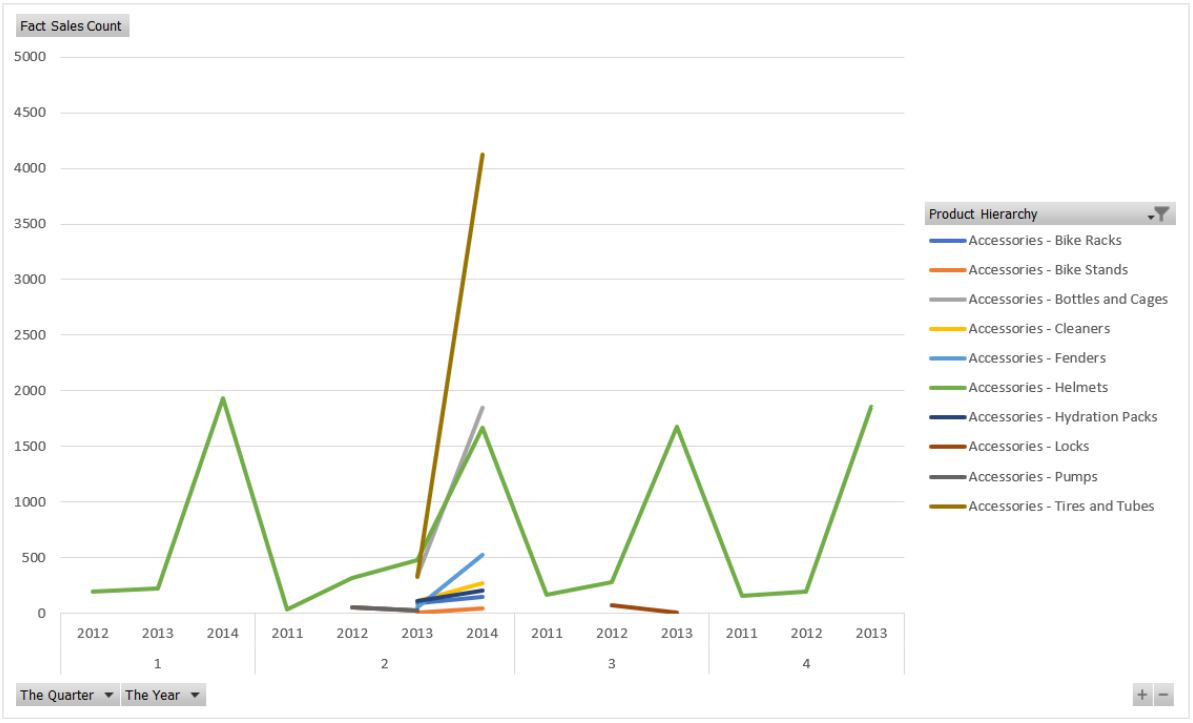

Figure 5- 12: Accessories sales

The majority of the Accessories category items were purchased in the second quarter of 2013. Notably, sales in the Tires and Tubes segment soared significantly during the same period, reaching over 4000 units purchased.

Helmets fluctuated predictably throughout a four-year period, exhibiting a specific cyclic pattern. These items were especially popular in 2014, with an annual average of 1700 sales—triple the average sales seen in the previous three years.

Other subcategories of products witnessed a concentration of purchasing, mainly in the second quarter of 2013. This temporal trend indicates a notable consumer preference or increased demand during that period. The unusual rise in buying activity for these subcategories during the second quarter of 2013 could be attributed to seasonality, promotional activities, or other external factors that influenced consumer behavior during that timeframe. A further look into the causes of this concentrated purchase pattern may yield significant information for organizations looking to match their strategy with customer preferences and market dynamics.

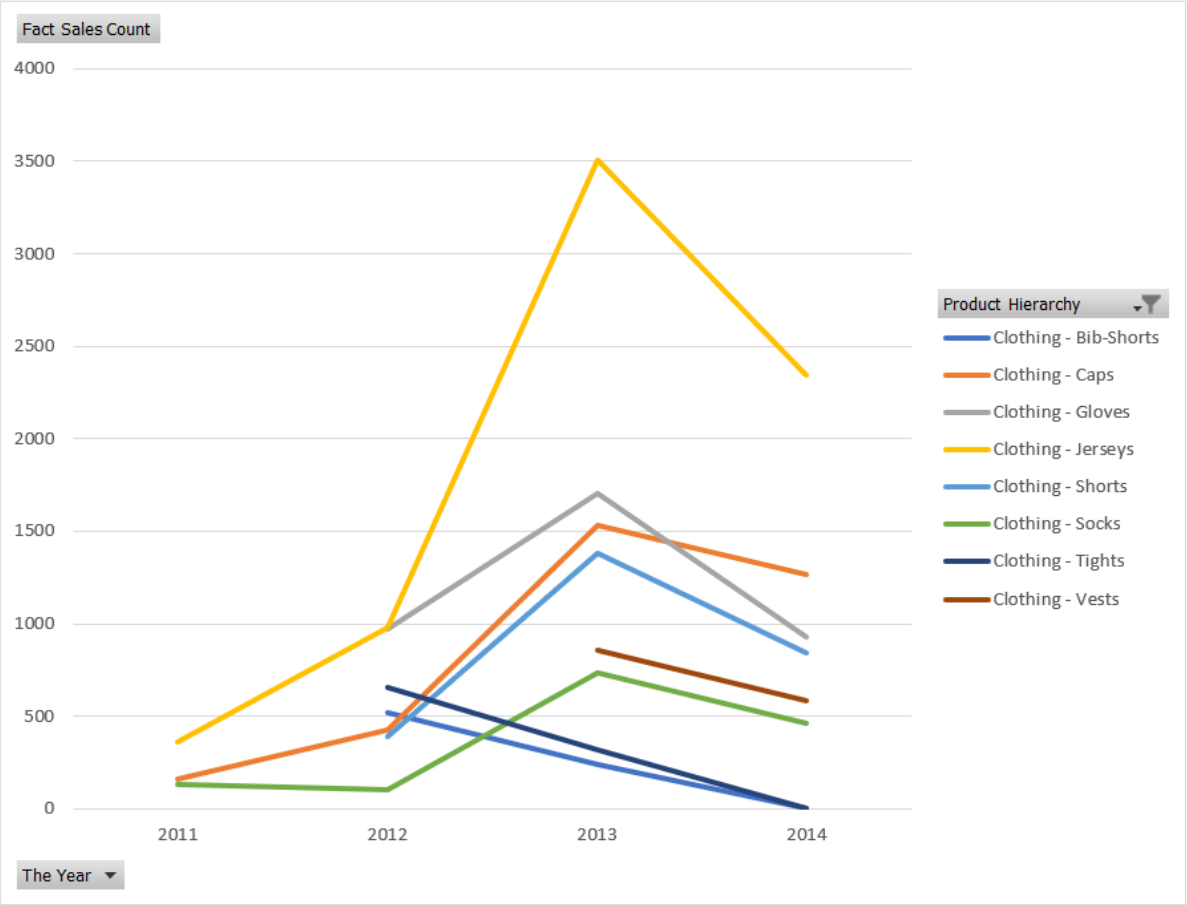

Figure 5- 13: Clothing sales

There is a notable upward trajectory in sales observed from 2011 to 2013, followed by a subsequent decline in sales across all categories. The Clothing category, in particular, experienced a peak in purchases during 2013, with Jerseys reaching a pinnacle at 3500 sales. Concurrently, sales for other segments within the Clothing category fell within the range of 500 to 1500 units during the same year.

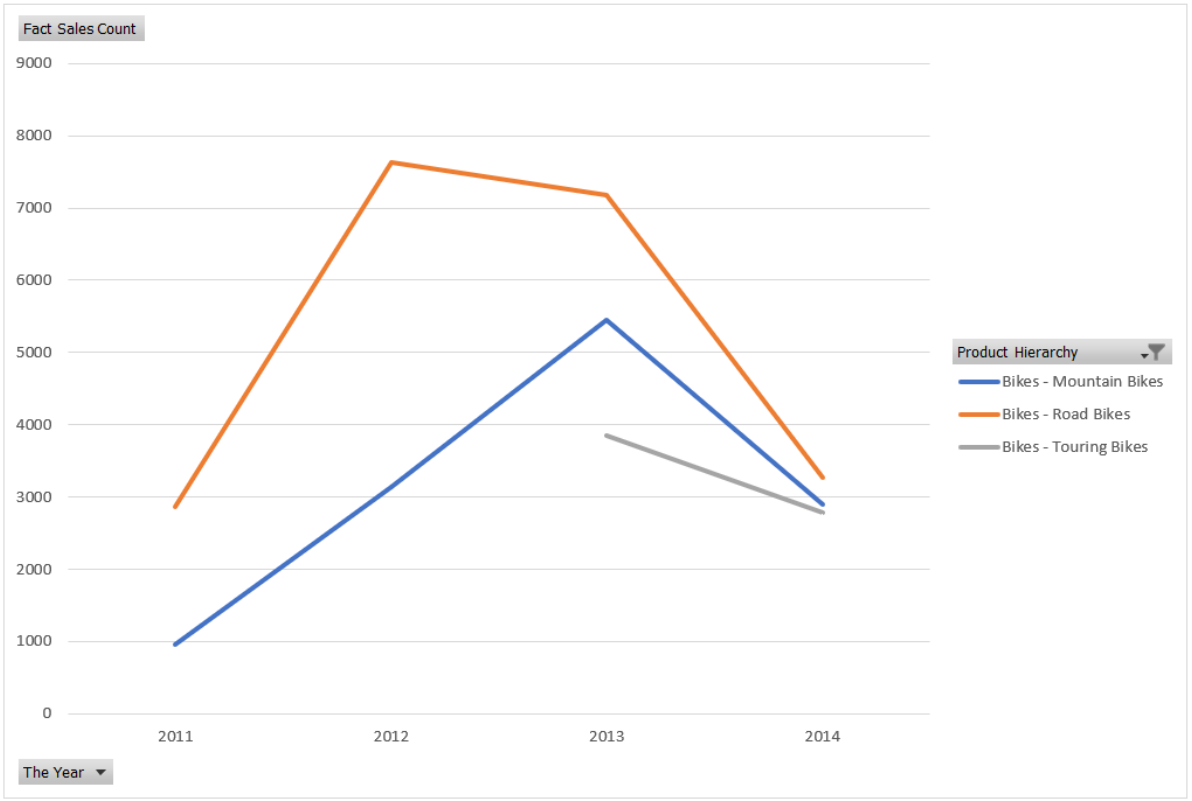

Figure 5- 14: Bikes sales

The Bikes category saw a significant increase in sales, increasing from around 3000 to 6000 on average between 2011 and 2013. However, sales fell significantly in the following year, 2014, sliding down to around 3000 units. This change in sales numbers suggests a major shift in consumer demand or market dynamics throughout the stated time period.

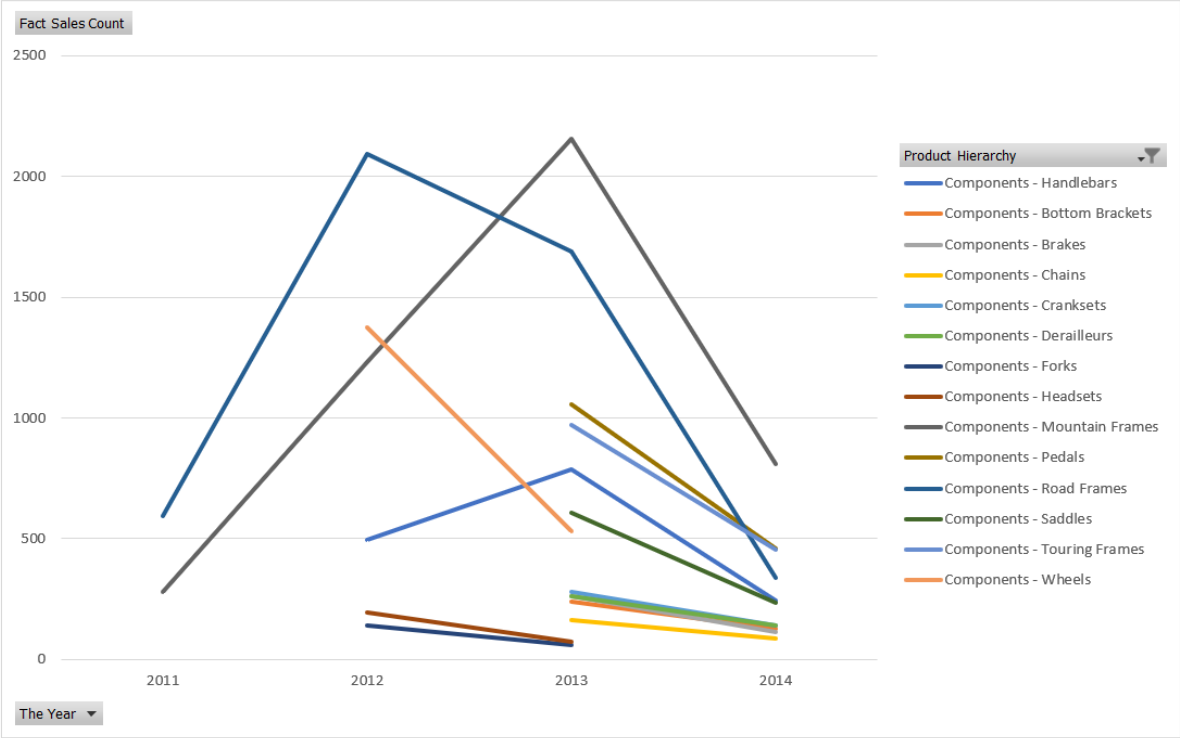

Figure 5- 15: Components sales

The observed sales pattern appears to be a constant trend across all categories, with an increase from 2011 to 2013 followed by a fall in 2014. Whether it's Bikes, Components, or other sectors, the majority of items in each category saw a high in purchasing in 2013, followed by a fall in sales in succeeding years.

This consistency in the sales trajectory points to a bigger market phenomenon or external factors influencing customer behavior during this time period. Understanding the underlying causes of this trend across several product categories can provide significant information for firms looking to adapt their strategy to changing market circumstances. Further investigation of consumer preferences, economic situations, or industry-wide factors may reveal significant determinants influencing this typical sales pattern.

5.3. Location-related problem

How does "Total due" vary across different locations over time? Are there specific regions where certain products perform better?

Perform a basic MDX query in SSAS, we receive the following result:

```
SELECT NON EMPTY { [Measures].[Total Due] } ON COLUMNS,
NON EMPTY { ([Dim Territory].[Territory Group].[Territory Group].ALLMEMBERS ) } ON ROWS
FROM [FinalLiveLaughLove]
```

Territory Group	Total Due
Europe	4780247...
North America	2378370...
Pacific	7057384...

Figure 5- 16: Total Due of 3 Territory Groups

Using Pivot Chart to visualize the statistics, we received the following line chart:

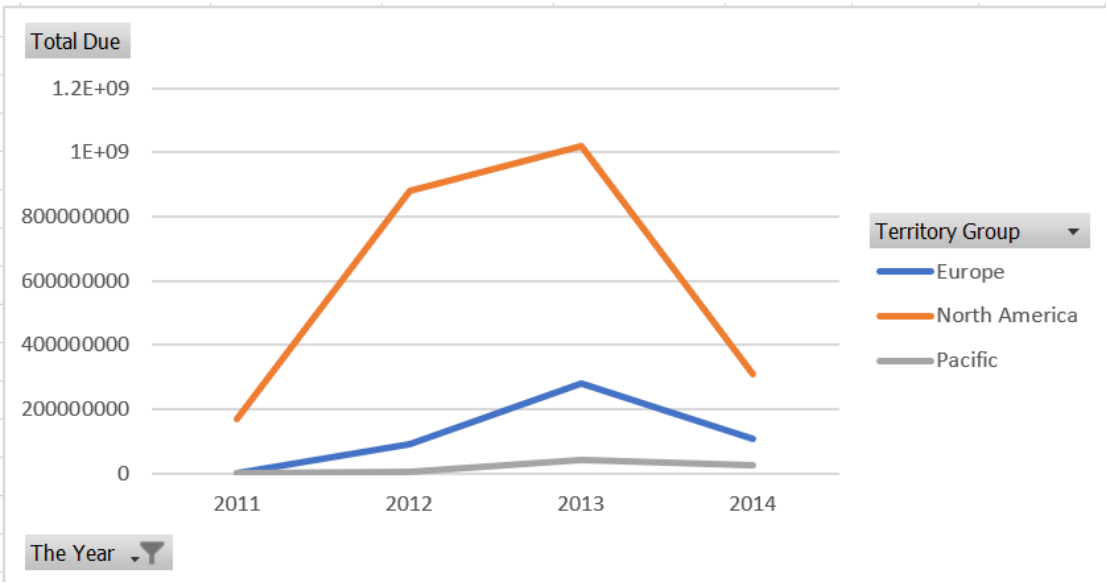

Figure 5- 17: Total Due of 3 Territory Groups over time

It can be seen that North America is the dominant group at all times, with the total due always above 20 million, while the highest in Europe is only around 30 million in 2013.

To have a deeper understanding of the difference in total due among countries in Europe, we then execute the following query:

```
SELECT NON EMPTY { [Measures].[Total Due] } ON COLUMNS,
NON EMPTY { ([Dim Territory].[Territory Country].[Territory Country].ALLMEMBERS ) } ON ROWS
FROM ( SELECT ( { [Dim Territory].[Territory Group].&[North America] } ) ON COLUMNS
FROM [FinalLiveLaughLove]) WHERE ( [Dim Territory].[Territory Group].&[North America] )
```

Territory Country	Total Due
Canada	5269693...
Central	2630991...
Northeast	2537712...
Northwest	4112071...
Southeast	2264274...
Southwest	6968964...

Figure 5- 18: Total Due of countries in North America

In order to recognize the result more easily, we visualize the result through the Pivot function of the cube:

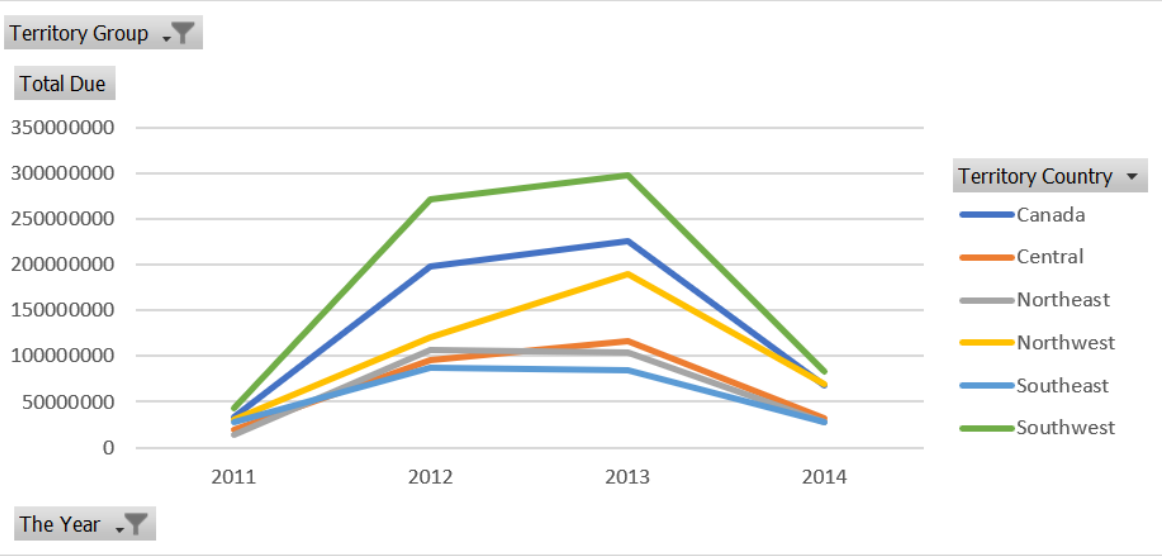

Figure 5- 19: Total due in North America over time

Although having different total due, all the territories of North America show the same trend and pattern, the ranking for each territory also remains the same with the highest of all time is Southwest.

5.4. Customer-related problem

Which individual customers contribute the best to the total revenue in each year?

```
WITH
  MEMBER [Measures].[TotalRevenue] AS
    [Measures].[Line Total] + [Measures].[Tax Amt] + [Measures].[Freight]

SET TopCustomers AS
  TopCount(
    [Dim Customer].[Customer ID].[Customer ID].MEMBERS,
    5,
    [Measures].[TotalRevenue]
  )

SELECT
  {[Measures].[TotalRevenue]} ON COLUMNS,
  {[Order Date].[The Year].[2011]:[Order Date].[The Year].[2014]} * TopCustomers ON ROWS
FROM [DW Final Project]
CELL PROPERTIES VALUE, FORMAT_STRING, LANGUAGE, BACK_COLOR, FORE_COLOR, FONT_FLAGS;
```

The Year	Customer ID	TotalRevenue
2011	29722	628910.44
2011	29614	569852.470...
2011	29701	(null)
2011	30103	(null)
2011	29957	(null)
2012	29722	2025416.47
2012	29614	2743909.09
2012	29701	2207893.37...
2012	30103	2456087.62
2012	29957	1892814.74
2013	29722	2123686.49
2013	29614	1596577.99
2013	29701	2284335.91
2013	30103	2177073.48...
2013	29957	2335658.48
2014	29722	892201.15
2014	29614	267916.1
2014	29701	570160.92
2014	30103	415027.33
2014	29957	750234.58

Figure 5- 20: Top 5 best contributing customers from 2011-2014

The analysis of the top 5 contributing customers by total revenue provides interesting insights into the financial performance across the years.

In 2011, there were only 2 top customers, yet they contributed significantly, generating approximately $600,000 in total revenue. The subsequent years, 2012 and 2013, stood out with the highest total revenue among the top 5 customers, reaching an impressive total of around $2,400,000. This indicates a substantial growth or increased business engagement during this period. However, in 2014, there was a noticeable drop in total revenue from the top 5 customers, decreasing to around $540,000. This decline might suggest shifts in customer preferences, market conditions, or other factors impacting the business landscape.

Understanding the specific factors influencing these fluctuations in revenue among top customers for each year can be crucial for adapting strategies, retaining valuable clients, and sustaining or improving financial performance.

5.5. Order-related problem

How does the amount of orders fluctuate according to the months of the year (from 2011 to 2014)?

Perform a basic MDX query in SSAS, we receive the following result:

```
SELECT
  NON EMPTY { [Measures].[Order Quantity] } ON COLUMNS,
  NON EMPTY {
    (
      [Order Date].[The Year].[The Year].ALLMEMBERS *
      [Order Date].[The Month].[The Month].ALLMEMBERS
    )
  } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM [Destination Database view]
```

The Year	The Month	Order Quantity
2012	10	6549
2012	11	3606
2012	12	5405
2012	2	1442
2012	3	3184
2012	4	2405
2012	5	7723
2012	6	11295
2012	7	9142
2012	8	5567
2012	9	8294
2013	1	4154
2013	10	14984
2013	11	9667
2013	12	11049
2013	2	5651

Figure 5- 21: Total orders by month from 2011 to 2014

To inquire more specifically, we can choose a particular year (here is 2011) to view the order quantity for each month based on the category name.

```
SELECT
  NON EMPTY { [Measures].[Order Quantity] } ON COLUMNS,
  NON EMPTY {
    (
      [Order Date].[The Year].[2011] *
      [Order Date].[The Month].[The Month].ALLMEMBERS *
      [Dim Product].[Product Subcategory Name].[Product Subcategory Name].ALLMEMBERS
    )
  } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM [Destination Database view]
CELL PROPERTIES
  VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

The Year	The Month	Product Subcategory Name	Order Quantity
2011	10	Caps	240
2011	10	Helmets	484
2011	10	Jerseys	457
2011	10	Mountain Bikes	1114
2011	10	Mountain Frames	249
2011	10	Road Bikes	2073
2011	10	Road Frames	512
2011	10	Socks	253
2011	11	Mountain Bikes	38
2011	11	Road Bikes	192
2011	12	Caps	25
2011	12	Helmets	29
2011	12	Jerseys	44
2011	12	Mountain Bikes	206

Figure 5- 22: Total orders by month for each category from in 2011

Additionally, these results can be visualized through “Analyze in Excel” features in the cube:

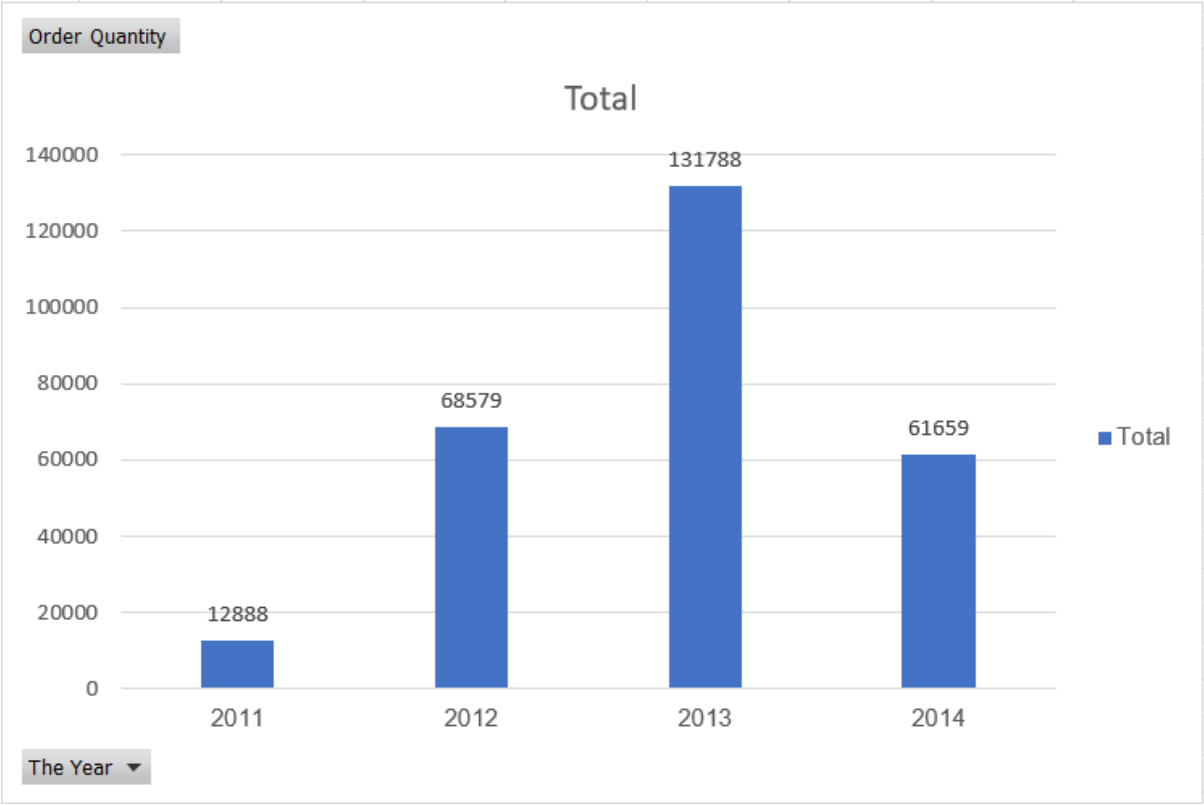

Figure 5- 23: Total order quantity by each year

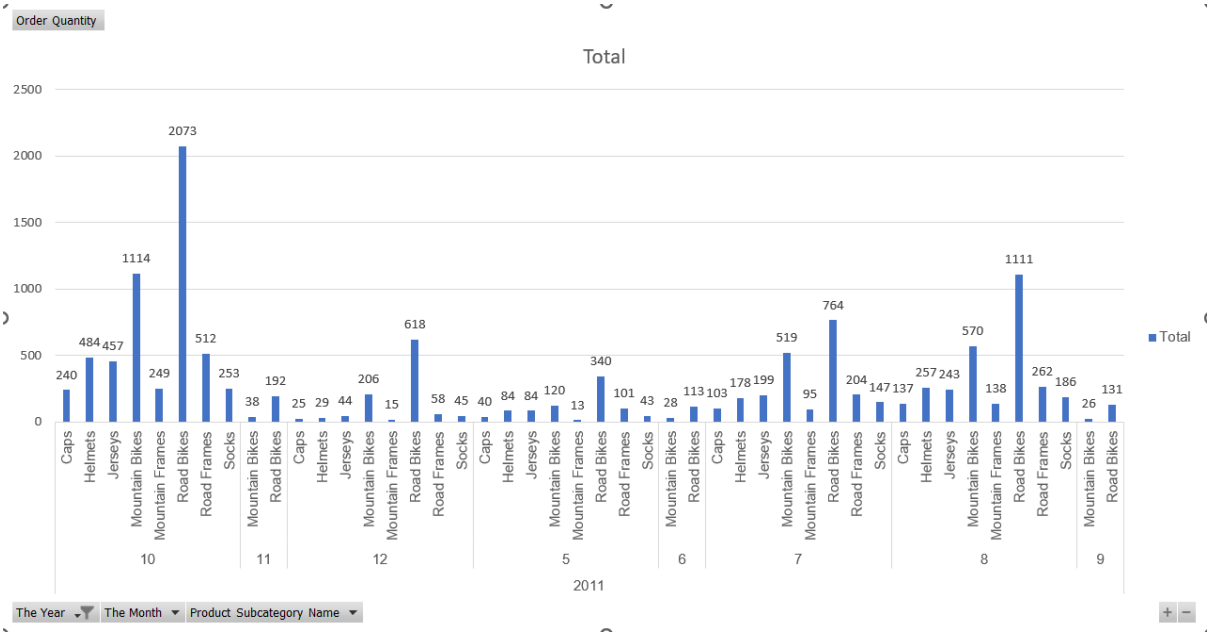

Figure 5- 24: Total order quantity for each category name across the months (2021)

Through these results, we can observe that in 2011, most items in the “Road Bikes” category consistently achieved the highest values. As a result, businesses can consider strategies to enhance the sales of items in this category to boost revenue. Therefore, businesses can choose which items in which category are popular at different times throughout the year to develop appropriate business strategies.

5.6. Evaluation

The project, as outlined in Table 1, aimed to address key business problems and targets. Remarkably, a significant portion of these challenges has been successfully tackled, showcasing a high percentage of completion.

Data Warehouse Implementation

The project introduced a robust data warehouse structure designed to provide a comprehensive overview of sales performance. Users now have the capability to query perspectives ranging from product and customer to sales order and location. This enables multifaceted analysis, allowing businesses to delve into sales data based on specific timeframes, locations, or product categories.

Effectiveness of the Data Warehouse

In terms of the project's business targets, the proposed data warehouse solutions have demonstrated their effectiveness. They facilitate the analysis of sales trends using historical data, allowing for comparisons among various attributes such as countries and years. This analytical capability empowers businesses to identify focal points and formulate strategic initiatives.

Chapter 6: Conclusion

6.1. Conclusion

The project has shown its effectiveness in addressing the challenge of fragmented data through the implementation of a comprehensive solution. A structured data warehouse was established using a snowflake schema, acting as a centralized repository for customer, product, and location information. Additionally, the integration of a multidimensional cube enhanced analytical capabilities, providing a clearer understanding of sales performance metrics.

This initiative facilitates more informed decision-making, empowering stakeholders to discern intricate relationships among customers, products, and locations. The insights gleaned from the data warehouse and multidimensional cube enable proactive responses to market dynamics. The proposed solutions, grounded in actionable intelligence, aim to optimize sales strategies, resource allocation, and overall business performance.

In essence, this project not only resolves immediate data challenges but also lays a foundation for sustained success in the competitive global market through data-driven decision-making.

6.2. Limitations

The data warehouse project faces significant limitations, notably the absence of auto-loading and the lack of quick query measures. The manual data loading process, due to the absence of auto-loading, introduces delays and potential inconsistencies, hampering real-time data availability. Additionally, the project's failure to establish quick query measures impacts user efficiency. These limitations highlight the need for improvements in automation and query optimization to enhance the project's overall performance and responsiveness to business demands.

6.3. Future Works

In order to enhance the quality of this project, our next steps will be incorporating additional data sources and implementing advanced data cleansing techniques. Furthermore, we aim to explore possibilities for real-time data updates within the data warehouse. This could involve implementing technologies that allow for continuous data integration, ensuring that decision-makers have access to the most current information.

Reference

Inmon, W.H., Strauss, D., & Neushloss, G. (2008). DW 2.0: The Architecture for the Next Generation of Data Warehousing. Elsevier Science.

Shahid, M. B., Sheikh, U., Raza, B., & Javaid, Q. (2016). Application of data warehouse in real life: State-of-the-art survey from user preferences' perspective. International Journal of Advanced Computer Science and Applications, 7(4).

Inmon, W. H. (2005). Building the data warehouse. John wiley & sons.

Iqbal, M. Z., Mustafa, G., Sarwar, N., Wajid, S. H., Nasir, J., & Siddque, S. (2020). A review of star schema and snowflakes schema. In Intelligent Technologies and Applications: Second International Conference, INTAP 2019, Bahawalpur, Pakistan, November 6–8, 2019, Revised Selected Papers 2 (pp. 129-140). Springer Singapore.

Inmon, W. H., & Linstedt, D. (2014). Data architecture: a primer for the data scientist: big data, data warehouse and data vault. Morgan Kaufmann.

Krishnan, K. (2013). Data warehousing in the age of big data. Newnes.

Microsoft. (2023, February 26). Understand star schema and the importance for Power BI - Power BI. Microsoft Learn. Retrieved December 2, 2023, from https://learn.microsoft.com/en-us/power-bi/guidance/star-schema

Giovinazzo, W. A. (2000). Object-oriented data warehouse design: building a star schema. Prentice Hall PTR.

Iqbal, M. Z., Mustafa, G., Sarwar, N., Wajid, S. H., Nasir, J., & Siddque, S. (2020). A review of star schema and snowflakes schema. In Intelligent Technologies and Applications: Second International Conference, INTAP 2019, Bahawalpur, Pakistan, November 6–8, 2019, Revised Selected Papers 2 (pp. 129-140). Springer Singapore.

Hüsemann, B., Lechtenbörger, J., & Vossen, G. (2000). Conceptual data warehouse design (Vol. 168). Universität Münster. Angewandte Mathematik und Informatik.

Levene, M., & Loizou, G. (2003). Why is the snowflake schema a good data warehouse design?. Information Systems, 28(3), 225-240.

IBM. (n.d.). What is a Data Mart | IBM. Retrieved November 17, 2023, from https://www.ibm.com/topics/data-mart

Databricks. (n.d.). What is a Data Mart? Retrieved November 17, 2023, from https://www.databricks.com/glossary/data-mart

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002, November). Conceptual modeling for ETL processes. In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP (pp. 14-21).

Albrecht, A., & Naumann, F. (2008). Managing ETL Processes. NTII, 8, 12-15.

Salley, C., & Codd, E.F. (1998). Providing OLAP to User-Analysts: An IT Mandate.