Report

ETL-Driven Data Warehouse Design and BI Reporting for Retail Analytics

Contents

LIST OF FIGURES

1. Data warehouse development

1.1 Proof of concept for data warehouse development

The data warehouse's viability and worth are demonstrated by the proof of concept. It demonstrates how decision-making may be enhanced and operations streamlined with a consolidated data source. The proof of concept describes technical viability, quantifiable objectives, and timescales utilizing a business case paradigm similar to bicycle sales. It demonstrates how crucial a data warehouse is for risk assessment, cost control, and enhancing business results.

1.2 Dimensional model for data warehouse

To facilitate effective querying and analysis, the data warehouse uses a star schema design, with a core fact table, FactOrder, encircled by supporting dimension tables. Important metrics like sales, profit, and other financial indicators are included in the FactOrder table. DimCustomer, which stores customer data like name, email, segmentation, and address details; DimProduct, which records product attributes like name, description, price, and category; DimDate, which provides temporal details like year, month, week, and day; DimLocation, which provides geographic and regional data filtered specifically for Australia, including city, state, and zip code; and DimDepartment, which contains operational data about departments and their identifiers. Detailed analysis of sales, consumer behavior, and regional performance specific to the Australian market is supported by this dimensional model.

1.3 Justification for selecting the subject area

The selection of sales and customer statistics was based on how well they support strategic decision-making. While consumer analytics provide a deeper understanding of purchasing patterns and retention tactics, sales data delivers insights into trends, geographical performance, and product profitability. To increase total business performance, these areas are essential.

1.4 Key stakeholders

Business analysts, data analysts, sales teams, managers, and IT staff are among the data warehouse's stakeholders. While data analysts concentrate on trends in customer and sales data, business analysts use the data for strategic planning. Managers make ensuring that employees and products are in line with client demand, while sales teams track performance to spot opportunities. The IT division oversees the technical setup and guarantees continuous system upkeep.

2. Schema for data warehouse

Multidimensional analysis is made possible by the schema's central fact table and auxiliary dimension tables. This design offers advanced analytics and guarantees query convenience. The connections and dependencies between these tables are displayed in the schema diagram included with the materials.

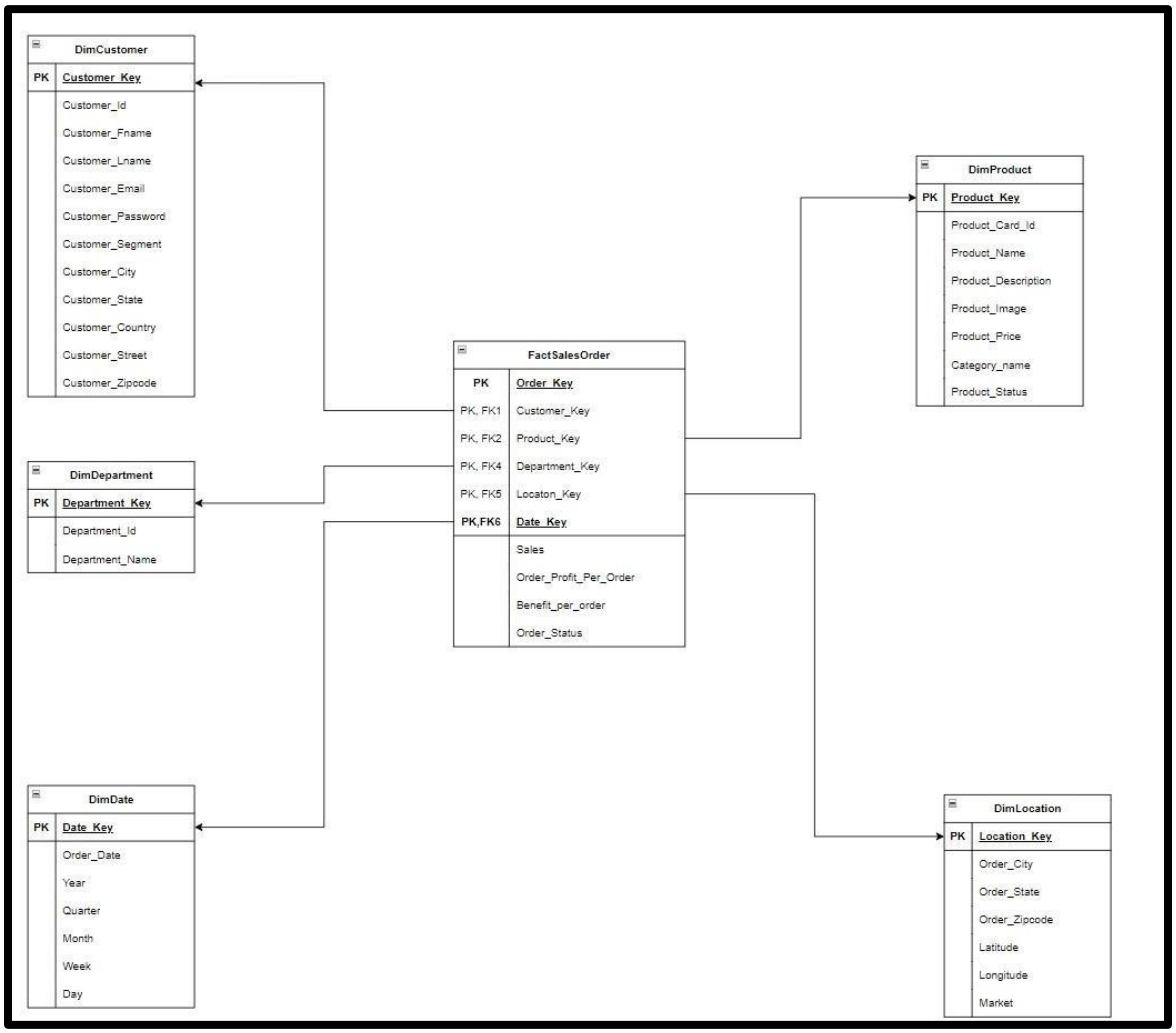

Figure 1: Data Warehouse Design

The FactSalesOrder table, which houses important metrics like sales, profit per order, and benefit per order and is connected to many dimension tables, is the focal point of the schema-based dimensional model shown in Figure 1. DimLocation contains geographic data, DimDate contains time-based analysis, DimCustomer has customer information, DimProduct contains product attributes, and DimDepartment contains organizational

units. Effective querying is made possible by this framework, which also facilitates thorough examination of sales, consumer behavior, product performance, and regional changes over time.

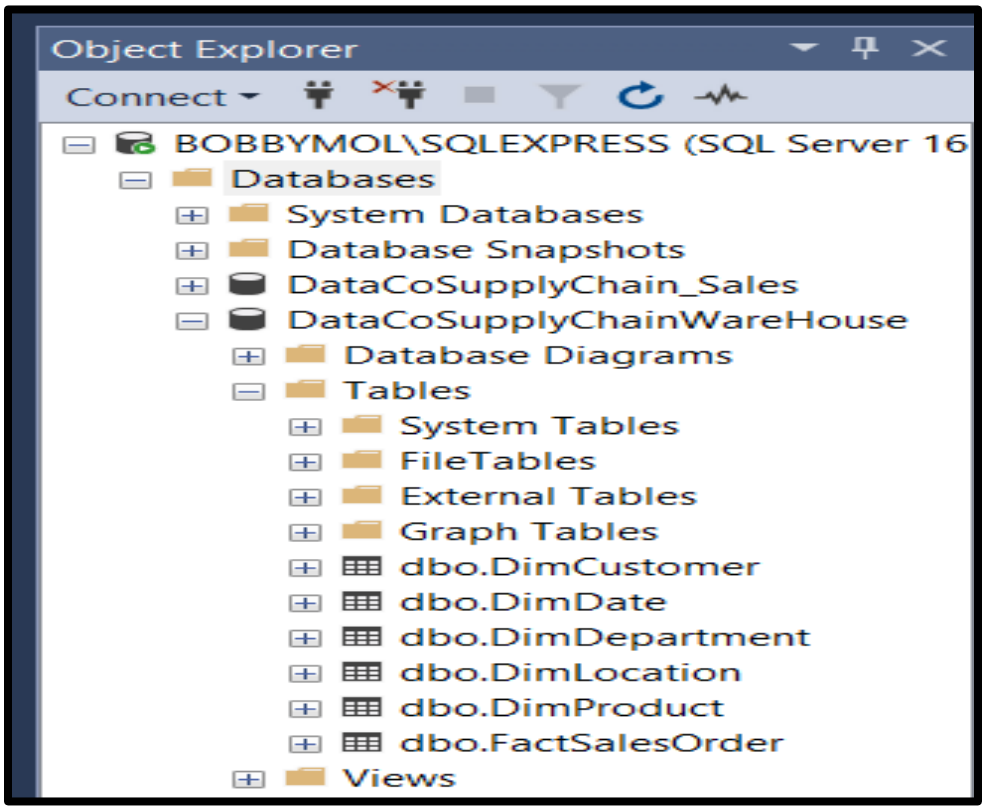

Figure 2: Data Warehouse Tables

The DataCoSupplyChainWareHouse database schema in SQL Server is displayed in Figure 2. The principal fact table, FactSalesOrder, and important dimension tables, including DimCustomer, DimDate, DimDepartment, DimLocation, and DimProduct, are arranged to support the star schema for effective data analysis and reporting.

```sql
-- Create the Data Warehouse Database
CREATE DATABASE DataCoSupplyChainWareHouse;
GO

-- Switch to the new database
USE DataCoSupplyChainWareHouse;
GO

-- Create Dimension Tables

-- DimCustomer
CREATE TABLE DimCustomer (
    Customer_Id INT PRIMARY KEY,
    Customer_Fname NVARCHAR(50) NOT NULL,
    Customer_Lname NVARCHAR(50) NOT NULL,
    Customer_Email NVARCHAR(100),
    Customer_Password NVARCHAR(100),
    Customer_Segment NVARCHAR(50),
    Customer_City NVARCHAR(50),
    Customer_State NVARCHAR(50),
    Customer_Country NVARCHAR(50),
    Customer_Street NVARCHAR(100),
    Customer_Zipcode NVARCHAR(20)
);
```

Figure 3: Creating a SQL Table

The SQL script in Figure 3 creates a data warehouse named DataCoSupplyChainWareHouse and defines a dimension table named DimCustomer to store customer information. The database includes columns like Customer_Key (primary key), Customer_Id, Customer_Fname, CustomThe SQL script in Figure 3 creates a data warehouse named DataCoSupplyChainWareHouse and defines a dimension table named DimCustomer to store customer information. The database includes columns like Customer_Key (primary key), Customer_Id, Customer_Fname, Customer_Lname, Customer_Email, Customer_Segment, Customer_City, and other parameters necessary for customer analysis to enable efficient data structuring and retrieval.er_Lname, Customer_Email, Customer_Segment, Customer_City, and other parameters necessary for customer analysis to enable efficient data structuring and retrieval.

3. ETL data into SQL server using SSIS package

SQL Server Integration Services (SSIS) is used to implement the ETL process. Data is imported into the data warehouse after being extracted from several sources and modified to conform to the schema. For reference, screenshots of every stage of the ETL process are included.

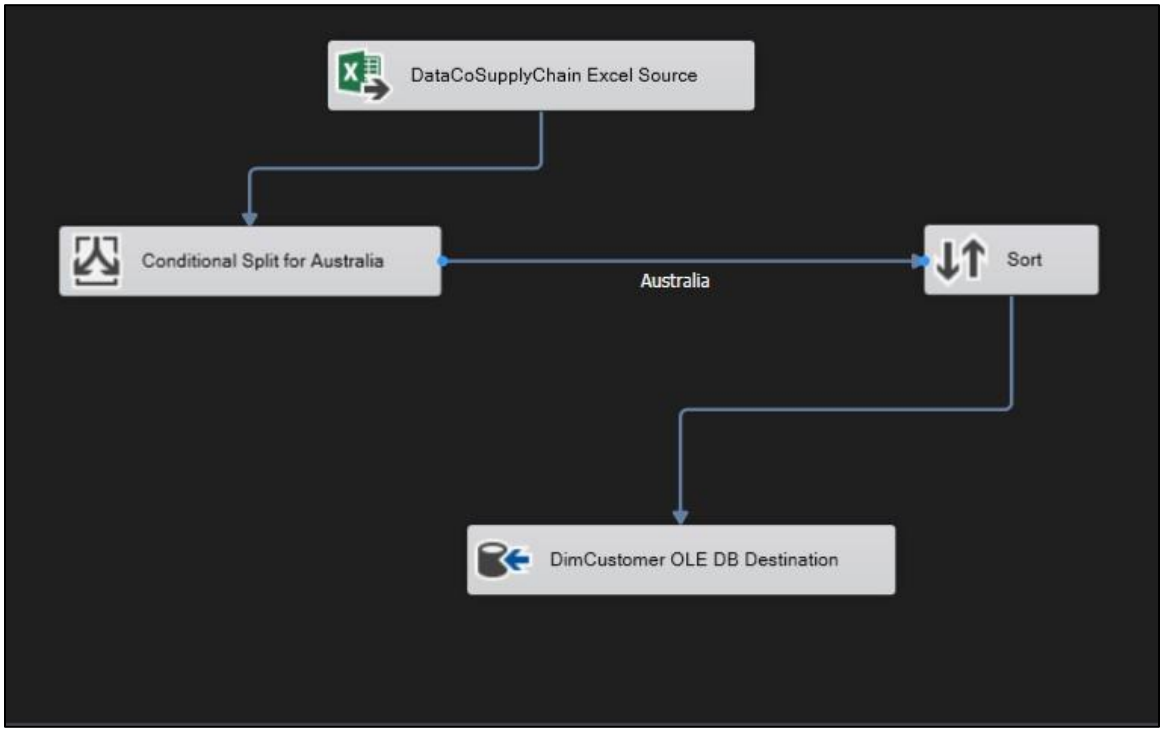

Figure 4: ETL Of Dim Customer using SSIS Package

After extracting data from the DataCoSupplyChain Excel Source, this SSIS data flow sorts the data, puts it into the DimCustomer OLE DB Destination table in the data warehouse, and applies a conditional split to filter records unique to Australia. Targeted data integration for regional analysis is ensured by this procedure.

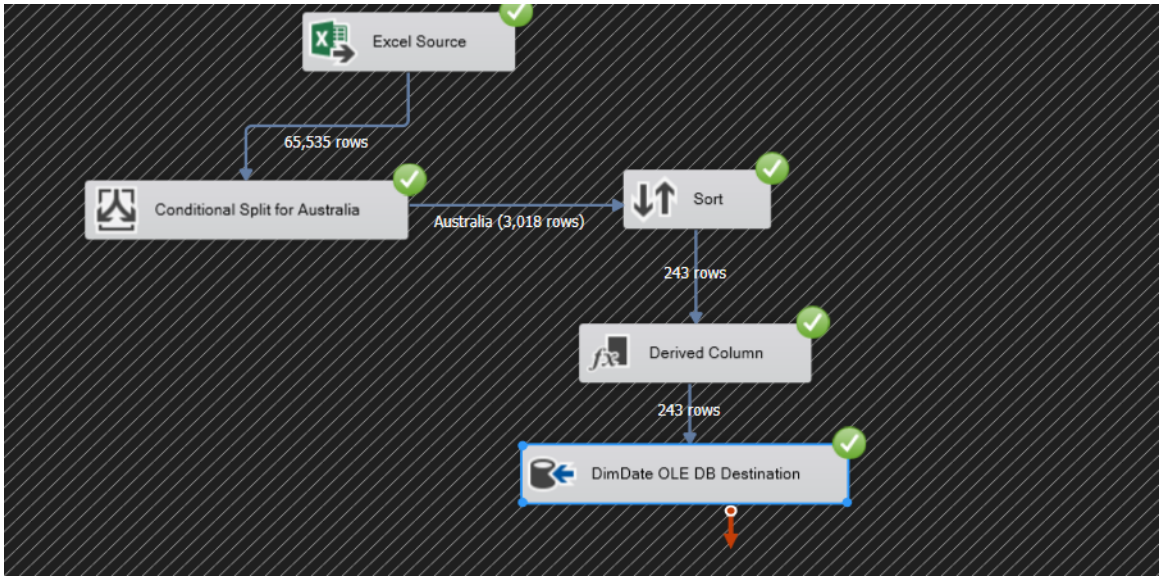

Figure 5: ETL Of DimDate using SSIS Package

An SSIS (SQL Server Integration Services) data flow procedure is depicted in this picture. After extracting the data from the DataCoSupplyChain Excel Source, it sorts the data, adds transformations like a conditional split to filter particular entries, and puts the data into the data warehouse's DimCustomer table. For analysis, our optimized ETL procedure guarantees effective and focused data integration.

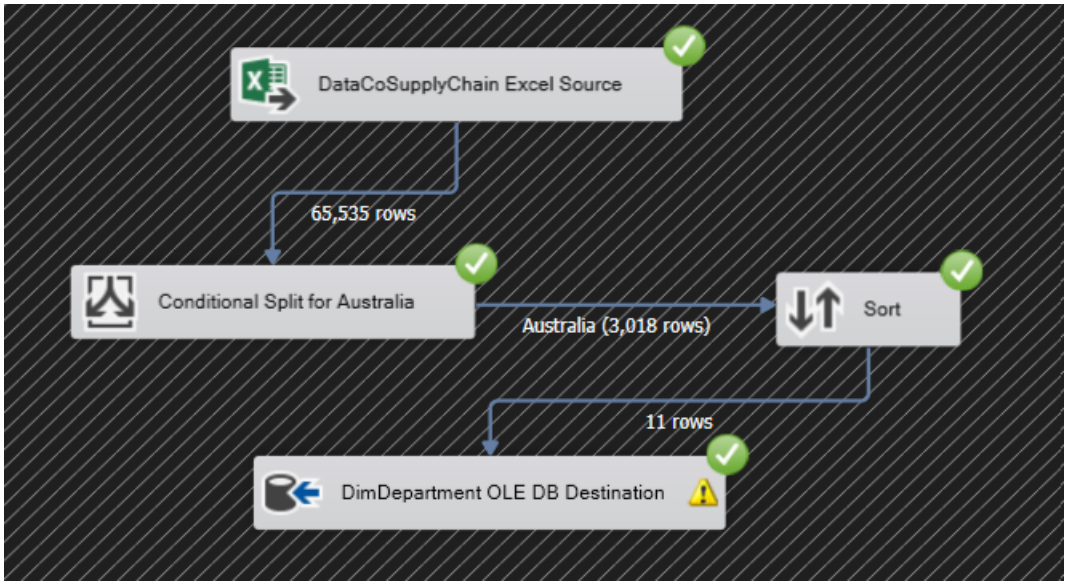

Figure 6: ETL of DimDepartment using SSIS Package

This SSIS data flow puts 11 entries into the DimDepartment OLE DB Destination table after extracting 65,535 rows from the DataCoSupplyChain Excel Source and using a conditional split to isolate 3,018 rows for Australia. This procedure guarantees the integration of structured and filtered data into the data warehouse.

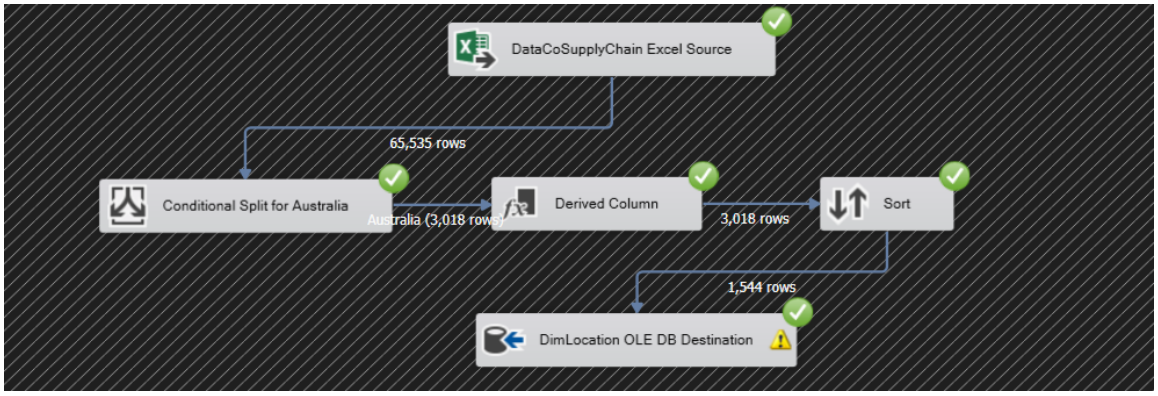

Figure 7:ETL of DimLocation using SSIS Package

This SSIS data flow ensures precise and focused data integration by extracting 65,535 rows from the DataCoSupplyChain Excel Source, sorting the data, applying transformations via a derived column, filtering 3,018 rows for Australia using a conditional split, and loading 1,544 rows into the DimLocation OLE DB Destination table.

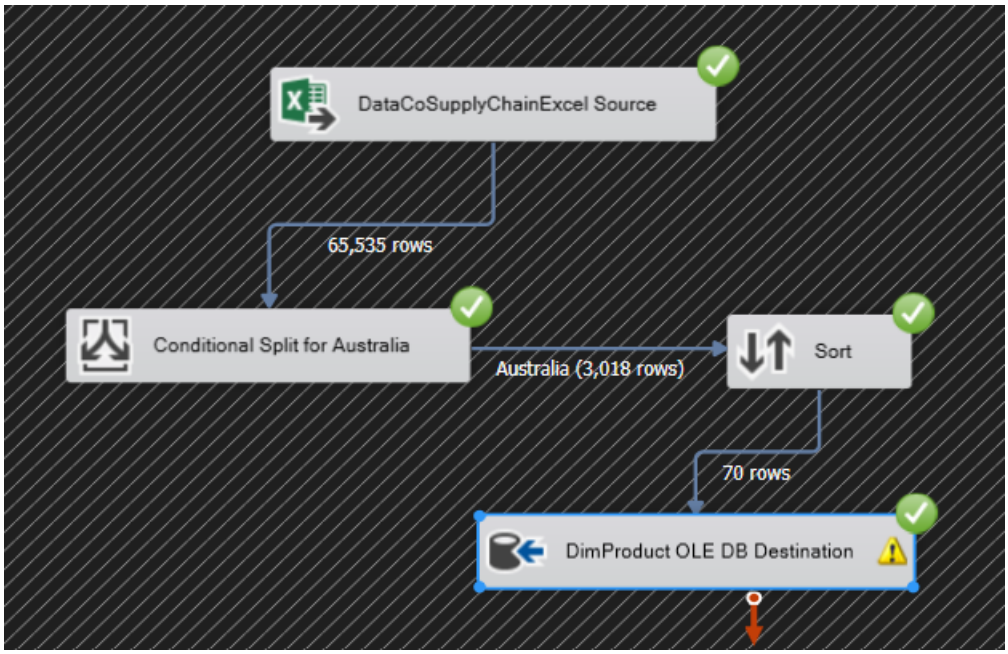

Figure 8: ETL Of DimProduct using SSIS Package

To ensure accurate product data integration for the Australian market, this SSIS data flow pulls 65,535 rows from the DataCoSupplyChain Excel Source, sorts the data, applies a conditional split to select 3,018 rows for Australia, and loads 70 rows into the DimProduct OLE DB Destination table.

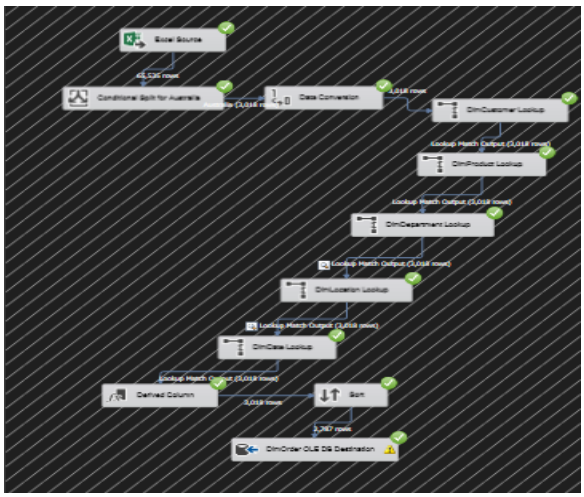

Figure 9: ETL of Order Dimension Tables in SQL using SSIS

In addition to extracting data from the Excel source, this SSIS data flow changes data types, applies a conditional split for Australia, and looks up information against dimension tables including DimCustomer, DimProduct, DimDepartment, and DimLocation. To ensure precise integration and alignment with current dimensions, the processed data is converted, sorted, and loaded into the DimOrder OLE DB Destination.

We created and deployed four thorough and perceptive reports as part of the development of our SSRS reports in order to provide vital business intelligence. Regional Sales, which assesses sales performance across various Australian states and helps identify high-performing and underperforming regions; Product Performance, which analyzes the performance of various products to determine top-selling and underperforming items; Monthly Sales Trends* from 2015 to 2018, which showcases seasonal patterns and temporal sales growth trends; and *Customer Sales Performance, which offers comprehensive insights into customer segmentation and their contribution. In order to guarantee consistency, dependability, and a single data source for all reports, we effectively extracted the necessary data from our data warehouse using stored procedures. Optimized performance was made possible by the incorporation of stored procedures, which simplified the data retrieval process. We made sure that the reports were based on precise and organized data by integrating SSRS with the data warehouse. We used data visualizations, filters, and sophisticated formatting in SSRS to create visually appealing and intuitive reports. Every report was customized to satisfy stakeholders'

analytical requirements while offering useful information for strategic planning and decision-making. With this strategy, we were able to produce reports that improved the reporting process overall while still being in line with business objectives

4. Visualisations

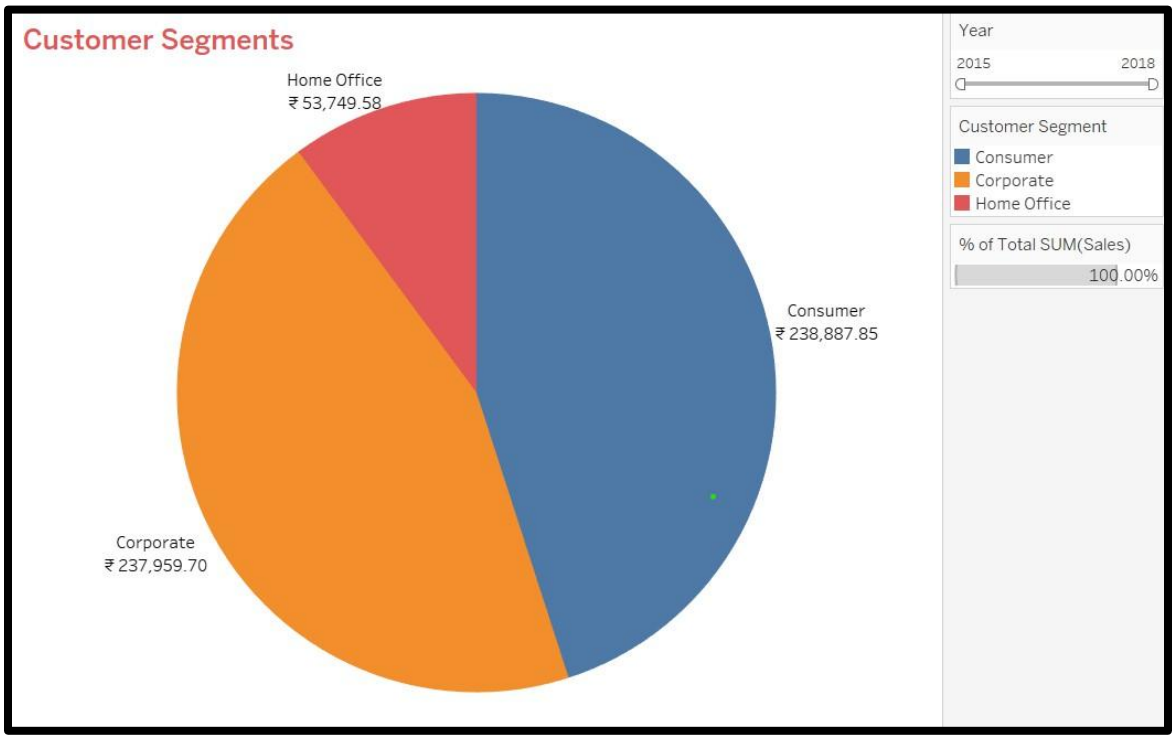

Figure 10:Customer Segmentation

This SSIS data flow pulls up information against dimension tables such as DimCustomer, DimProduct, DimDepartment, and DimLocation, applies a conditional split for Australia, and transforms data types in addition to pulling data from the Excel source. The processed data is transformed, sorted, and loaded into the DimOrder OLE DB Destination to guarantee accurate integration and alignment with current dimensions.

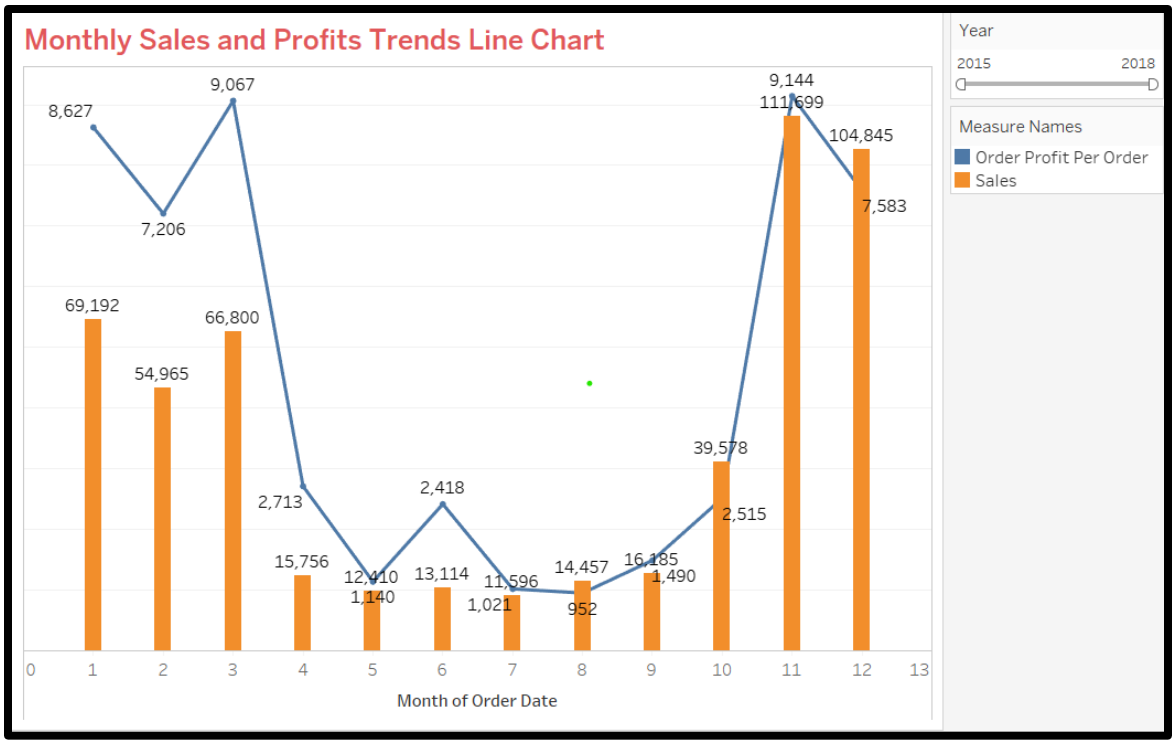

Figure 11: Monthly Sales and Profits trends line chart

The monthly trends in sales and earnings per order from 2015 to 2018 are shown in this line and bar chart. March (₹66,800) and November (₹111,699) have the highest sales, and these months also see the highest profits, with November seeing the highest profit of 9,144. On the other hand, July had the lowest sales and profits (₹11,596 in sales and 1,021 in profit), suggesting a seasonal variation. The graph shows notable fluctuations, with months with strong sales corresponded to profitability peaks.

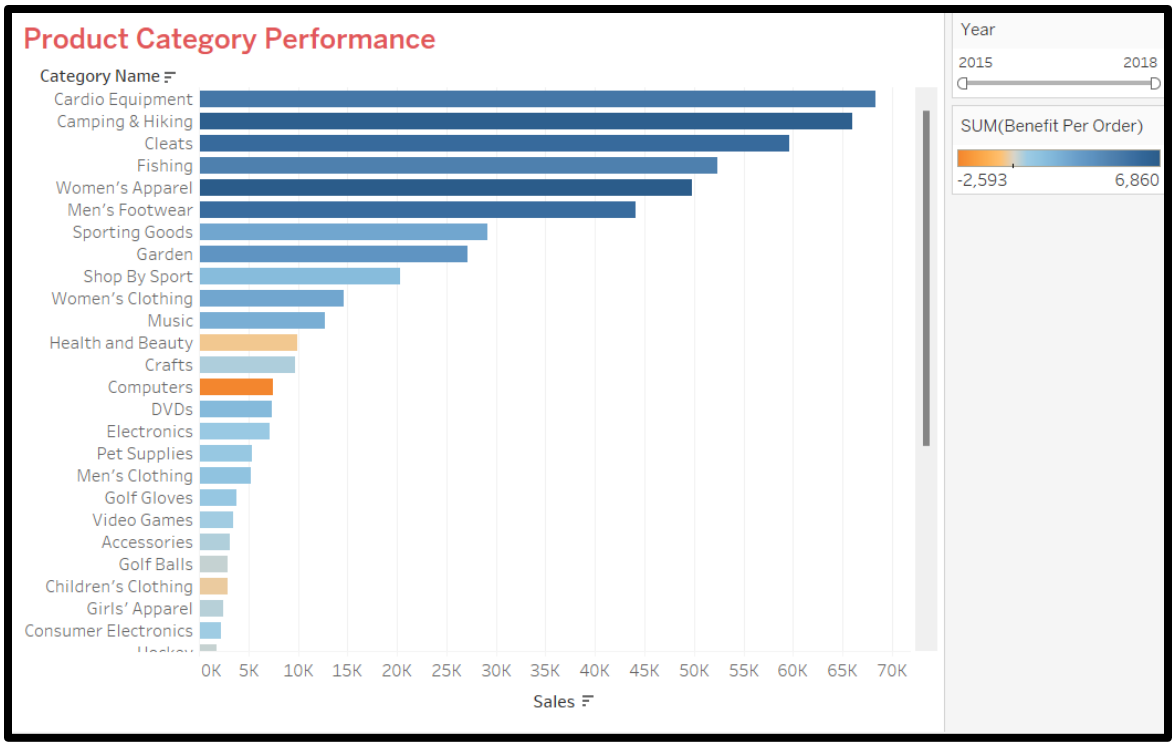

Figure 12: Category Performance

Product category performance from 2015 to 2018 is shown in this bar chart based on sales and benefits per order. While areas like Computers and Health and Beauty have lower sales and little profitability, Cardio Equipment and Camping & Hiking are the top sellers and bring in the most money. Notably, cardio equipment also offers significant benefits per order, while golf balls and children's clothing perform poorly and show low profitability. Strategic decisions are aided by the chart's clear insights into the product categories with the highest and lowest profitability.

.

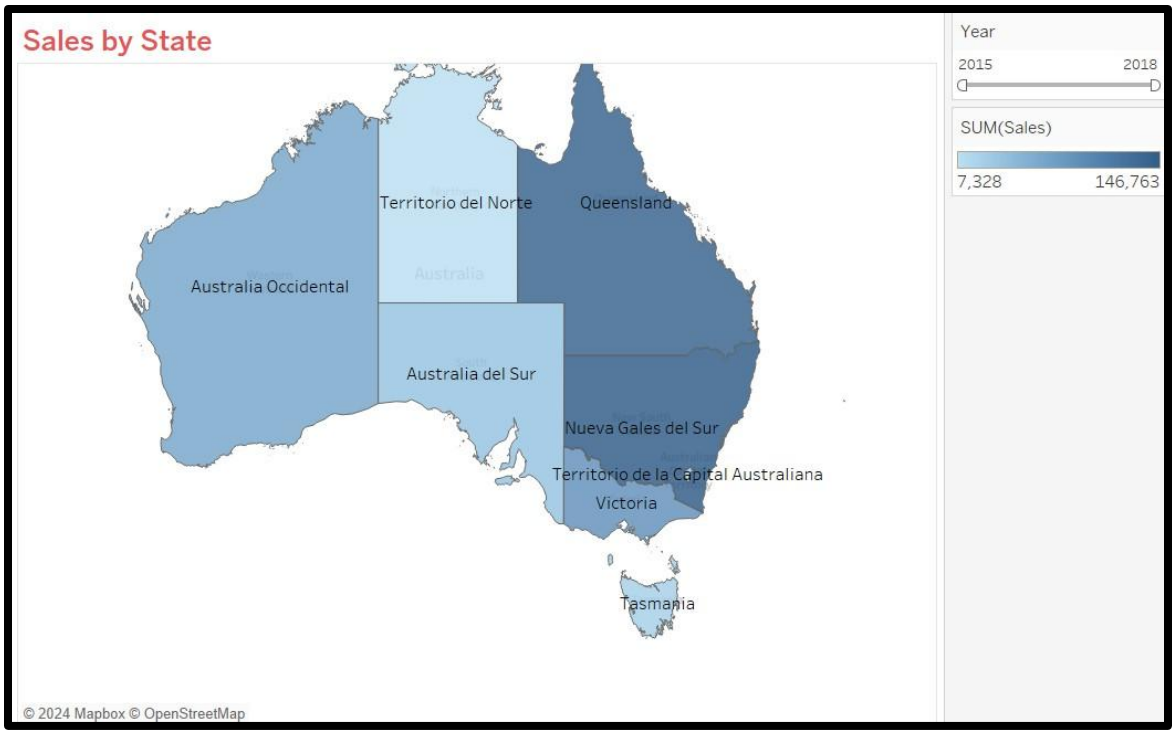

Figure 13:Sales by state

The distribution of sales by Australian state between 2015 and 2018 is depicted on this map. The Northern Territory and Tasmania record the lowest sales, as low as ₹7,328, while New South Wales and Queensland lead with the highest sales statistics, up to ₹146,763. Significant regional differences are highlighted in the chart, offering information for resource allocation and focused marketing.

Bibliography

Iqbal, M.Z., Mustafa, G., Sarwar, N., Wajid, S.H., Nasir, J. and 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.

Antunes, A.L., Cardoso, E. and Barateiro, J., 2022. Incorporation of ontologies in data warehouse/business intelligence systems-a systematic literature review. *International Journal of Information Management Data Insights*, *2*(2), p.100131.

Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M., 2021, January. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8, p. 28).

Yang, P., Xiong, N. and Ren, J., 2020. Data security and privacy protection for cloud storage: A survey. *Ieee Access*, *8*, pp.131723-131740.

MacLean, A., Young, R.M., Bellotti, V.M. and Moran, T.P., 2020. Questions, options, and criteria: Elements of design space analysis. In *Design rationale* (pp. 53-105). CRC Press.

Geetha, K., 2020. Param (2020). Data Analysis and ETL Tools in Business Intelligence. *International Research Journal of Computer Science (IRJCS)*, *7*, pp.127-131.