| | |
|---|---|
| **Students Details** | |
| **Module Name:** | Software Engineering |
| **Module Lecturer/ Course Coordinator:** | l.S. Chathurika |
| **Department:** | School of Computing |
| **Submission Due on:** | 13th July ,2025 |
| **Type of Coursework:** | Report |
| **Title of the Coursework:** | Data Warehouse and Business Intelligence |

***Students Details:***

| | **Student No.** | **Student Name** |
|---|---|---|
| 01 | COHNDSE24.2F - 026 | P H N Sethulya |
| 02 | COHNDSE24.2F - 005 | M K D Gangadara |
| 03 | COHNDSE24.2F - 040 | W A U S Weerasooriya |
| 04 | COHNDSE24.2F - 008 | A H V M Jayaratine |
| 05 | COHNDSE24.2F - 085 | P K P C P Kapuge |

| **Office use only:** |
|---|
| ***Date Stamp Required of the Department*** |

# Table of Contents

**Chapter 1: Introduction** ..... 4

**1.1 Overview of the Business** ..... 4

**1.2 Importance of having Data Warehousing and BI in Retails** ..... 4

**1.3 Project Goal** ..... 5

**1.4 Tools and Technologies** ..... 5

**Chapter 2: Data Sources** ..... 6

**2.1 CSV Files** ..... 6

**2.2 SQL Files** ..... 7

**2.3 JSON Files** ..... 7

**2.4 Sample Tables and Screenshots** ..... 7

**Chapter 3: Data Warehouse Design** ..... 9

**3.1 Facts and Dimensions Tables** ..... 9

**3.2 Galaxy Schema** ..... 10

**Chapter 4: ETL Process** ..... 11

**4.1 Data Extraction** ..... 11

**4.2 Data Transformation** ..... 11

**4.3 Data Loading** ..... 12

**4.4 ETL Sample Workflows** ..... 12

**Chapter 5: Data Visualization** ..... 15

**5.1 Connection between the Data Warehouse & Power BI** ..... 15

**5.2 Dashboard Design** ..... 15

**Chapter 6: Analysis & Insights** ..... 18

**Chapter 7: Conclusion** ..... 18

## Roles and Contributions

| Index No | Contribution |
|---|---|
| **COHNDSE24.2F – 005** | • Found Data sources and gathered Data, suitable for the business.<br>• Using Apache Hop, extracted data and added validations, transformed data |
| **COHNDSE24.2F – 008** | • Found Data sources and gathered Data, suitable for the business.<br>• Designed the Galaxy Schema,<br>• Load data to Power bi and Transform data and Add connections |
| **COHNDSE24.2F – 026** | • Created the report and presentation<br>• Design Power BI Dashboards<br>• Identifying Business Problems in the domain and Solutions using the Data (Dashboard) |
| **COHNDSE24.2F – 040** | • Found Data sources and gathered Data, suitable for the business.<br>• Designed the Galaxy Schema,<br>• Using Apache Hop, extracted data and added validations, transformed data.<br>• Data warehouse implementation. |
| **COHNDSE24.2F – 085** | • Using Talend, did the ETL transformation and loaded data. |

# Chapter 1: Introduction

## 1.1 Overview of the Business

E - commerce products are amongst the most dynamic product categories in the modern retail business due to the high consumer demand and rapid development of technology. A retail establishment that sells electrical products, such as televisions, refrigerators, washing machines, mobile devices, and related accessories, is the business domain that we have selected. Customers from all around the world purchase these goods, which are available for purchase both online and in store.

The business generates massive amounts of data every day, including customer information, product details, reviews, ratings, and sales transactions. With careful analysis, this data's informative information can help guide strategic decisions regarding product performance, inventory management, marketing, and customer targeting.

## 1.2 Importance of  having Data Warehousing and BI in Retails

Business intelligence (BI) and data warehousing (DW) are important for turning operational data into insights that can be used in the retail industry. Data from multiple sources are collected, transformed, and stored in a structured format for analysis and reporting in a data warehouse, that acts as a central repository.

DW and BI allow electrical retail shops to:

- Analysis of sales trends by product, area, or time
- Monitoring consumer behavior for tailored advertising
- Analyzing product performance with sales and review data
- Forecasting inventory and demand

Retailers may improve decision-making and performance by visualizing these insights through interactive dashboards and reports using BI.

## 1.3 Project Goal

This project's objective is to use industry level tools and techniques to build and implement a working data warehousing solution.

- Collecting data from multiple sources (CSV files, Excel, SQL, JSON)
- Designing a Galaxy Schema with fact and dimension tables
- Implementing ETL processes using Talend to extract, transform, and load data into the data warehouse
- Creating a Power BI dashboard for visualizing business insights and trends.

## 1.4 Tools and Technologies

- MySQL - For creating and managing the data warehouse schema (database structure and queries)
- Apache Hop & Talend Open Studio - For performing ETL operations (data extraction, transformation, and loading)
- Microsoft Power BI - For data visualization, dashboard creation, and insight analysis
- Excel / SQL / JSON / CSV files - Used as source data formats for products, customers, reviews, and sales

# Chapter 2: Data Sources

We are using three types of data sources to get data for this data warehouse . The three types of data are:

- CSV Files
- Json Files
- SQL Files

## 2.1 CSV Files

The data that was taken from CSV Files

| File Name | Description | Key Fields Included |
|---|---|---|
| **Customers.csv** | Customer data with IDs and demographics | CustomerID, First Name, Last Name, City, Country, Phone, Email |
| **Geography.csv** | Location info for customers and stores | GeographyID, Country, City, Postal Code |
| **Product. Csv** | Product master list | ProductID, Name, Category, Price, Discounts, Decriptions, Images |
| **Rating. Csv** | Customer product ratings | RatingID, ProductID, CustomerID, RatingValue, RatingCount |
| **Shipping Carrier. csv** | Shipping service providers | CarrierID, CarrierName, Hotline, Website, Email |

## 2.2 SQL Files

The following are the SQL Data Sources that we gathered

| File Name | Description | Key Fields Included |
|---|---|---|
| **Inventory.sql** | Stock levels for products in warehouses | InventoryID, ProductID, StockLevel, LastUpdateDate |
| **Sales.sql** | Transaction-level sales data | OrderID, ProductID, CustomerID, Quantity, Price, Date, BillAmount |
| **Shipping data.sql** | Shipment records | ShipmentID, ShippingCost, Quantit |
| **Supplier data.sql** | Supplier information | SupplierID, Name, Address, City, PostalCode, Email, Phone |

## 2.3 JSON Files

The JSON formatted data we gathered

| File Name | Description | Key Fields Included |
|---|---|---|
| **Reviews.json** | Product reviews by customers | ReviewID, ProductID, CustomerID, ReviewContent, ReviewTitle |

## 2.4 Sample Tables and Screenshots

Customer Data

| | Customer Id | First Name | Last Name | City | Country | Phone 1 | Phone 2 | Email |
|---|---|---|---|---|---|---|---|---|
| 2 | EB54EF1154C3A78 | Heather | Callahan | Lake Jeffborough | Norway | 043-797-5229 | 915.112.1727 | urangel@espinoza-francis.net |
| 3 | 10dAcafEBbA5FcA | Kristina | Ferrell | Aaronville | Andorra | 932-062-1802 | (209)172-7124x36 | xreese@hall-donovan.com |
| 4 | 67DAB15Ebe4BE4a | Briana | Andersen | East Jordan | Nepal | 8352752061 | (567)135-1918 | haleybraun@blevins-sexton.c |
| 5 | 6d350C5E5eDB4EE | Patty | Ponce | East Kristintown | Northern Mari | 302.398.3833 | 196-189-7767x770 | hohailey@anthony.com |

Geography Data

| | country | city | postal code |
|---|---|---|---|
| 2 | Malaysia | Pulau Pinang | 10740 |
| 3 | Iran | KalÄ‌t-e NÄderÄ« | |
| 4 | Malaysia | Ipoh | 30200 |
| 5 | Philippines | Estaca | 1123 |

## Product Data

| | product_id | product_name | category | discounted_price | actual_price | discount_percentage | about_product | img_link | product_link |
|---|---|---|---|---|---|---|---|---|---|
| 2 | B07JW9H4J1 | Wayona Nylon Br | Computers&Accessories \| | $399 | $1,099 | 64% | High Compatibil | https://m.i | https://www.amazon.in/Wayona-Braided-WN3LG1 |
| 3 | B098NS6PVG | Ambrane Unbrea | Computers&Accessories \| | $199 | $349 | 43% | Compatible with | https://m.i | https://www.amazon.in/Ambrane-Unbreakable-Ch |
| 4 | B096MSW6CT | Sounce Fast Pho | Computers&Accessories \| | $199 | $1,899 | 90% | â€ Fast Charger | https://m.i | https://www.amazon.in/Sounce-iPhone-Charging- |
| 5 | B08HDJ86NZ | boAt Deuce USB | Computers&Accessories \| | $329 | $699 | 53% | The boAt Deuce | https://m.i | https://www.amazon.in/Deuce-300-Resistant-Tan |
| 6 | B08CF3B7N1 | Portronics Konn | Computers&Accessories \| | $154 | $399 | 61% | [CHARGE & SYN | https://m.i | https://www.amazon.in/Portronics-Konnect-POR-1 |

## Rating Data

| | rating_id | rating | rating_count | customer_id | product_id |
|---|---|---|---|---|---|
| 2 | RA1 | 4.2 | 24,269 | 8aA54b9BFBA0aD2 | B07JW9H4J1 |
| 3 | RA2 | 4 | 43,994 | Ca907A9DDdE2126 | B098NS6PVG |
| 4 | RA3 | 3.9 | 7,928 | BeB4EE902dcE991 | B096MSW6CT |
| 5 | RA4 | 4.2 | 94,363 | Eaf80AAe54E31EA | B08HDJ86NZ |
| 6 | RA5 | 4.2 | 16,905 | ab7b2f257c7B3f3 | B08CF3B7N1 |

## Shipping Carrier Data

| | shipping Carriers Id | shipping Carriers Name | Hotline | Website | Email |
|---|---|---|---|---|---|
| 2 | 1 | Maersk | 1-800-321 | https://www.maersk.com/ | |
| 3 | 2 | Mediterranean Shipping | 1-212-764 | https://www.msc.com/ | |
| 4 | 3 | CMA CGM Group | 1-877-556 | https://www.cma-cgm.com/ | |
| 5 | 4 | COSCO Shipping Lines | 1-281-765 | https://lin | cs.hq@coscon.com |

## Review Data


```
{
  "review_id": "R3HXWT0LRP0NMF,R2AJM3LFTLZHFO,R6AQJGUP6P86,R1KD19VHEDV0OR,R3C02RMYQMK6FC,R39GQRVBUZBWGY,R2K9EDOE15QIRJ,R3OI7YT648TL8I
  "review_title": "Satisfied,Charging is really fast,Value for money,Product review,Good quality,Good product,Good Product,As of now seems good",
  "FIELD3": "",
  "FIELD4": "",
  "customer_id": "8aA54b9BFBA0aD2",
  "product_id": "B07JW9H4J1",
  "review_content": "Looks durable Charging is fine tooNo complains,Charging is really fast, good product. Till now satisfied with the quality. This is a good product . The charging sp
}
{
```

# Chapter 3: Data Warehouse Design

## 3.1 Facts and Dimensions Tables

The data warehouse is designed using a Galaxy Schema that includes multiple fact tables and shared dimension tables.

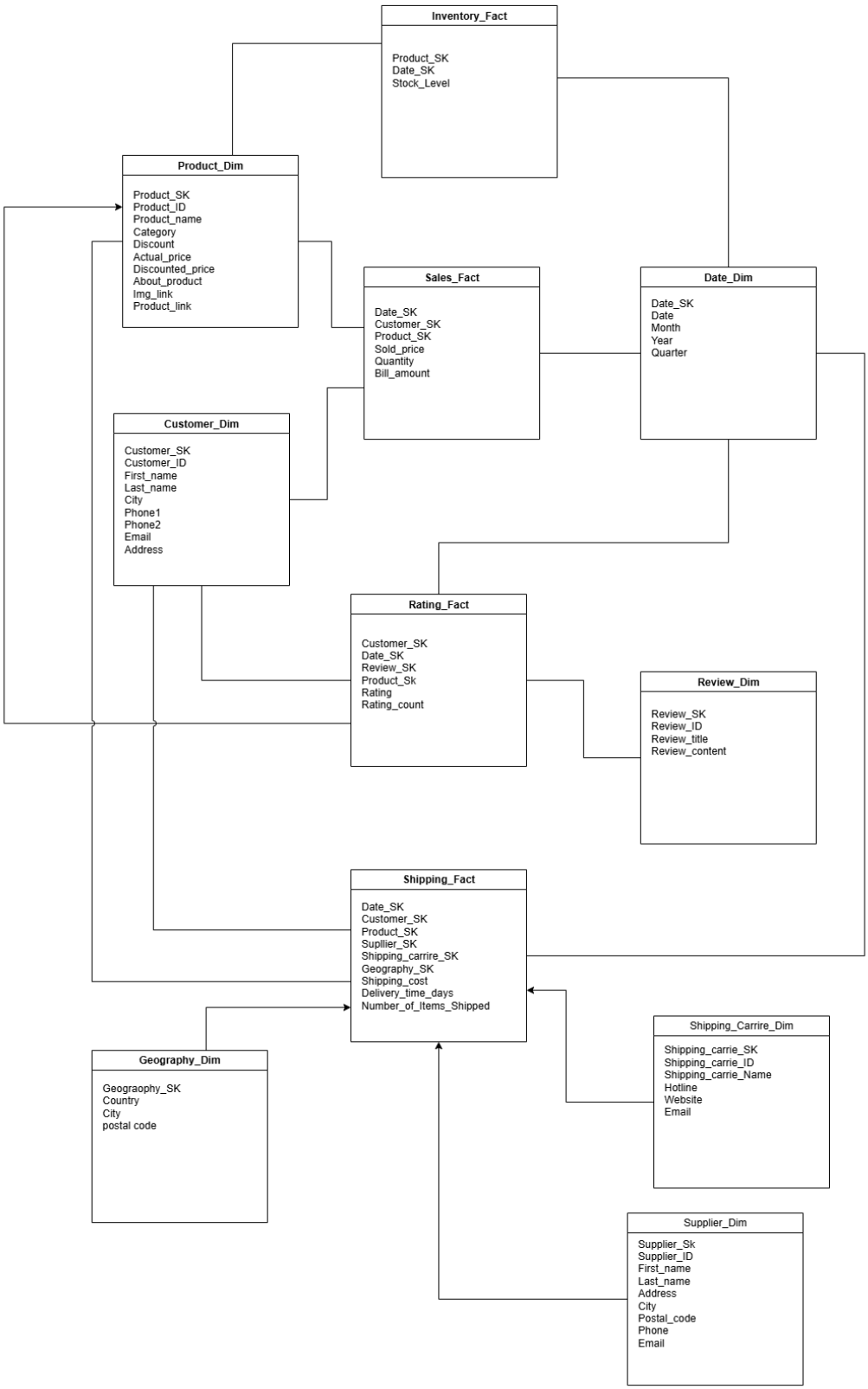
## Fact Tables

- **Sales Fact** – Stores information about customer purchases, including quantity sold and total amount.
- **Rating Fact** – Records customer ratings for products along with the date and review link.
- **Inventory Fact** – Tracks product stock movement (in/out) and current stock levels over time.
- **Shipping Fact** – Contains shipment details such as delivery status, shipping date, and associated customer/product.

## Dimensions Table

- **Customer Dimension** – Contains customer profile data such as name, contact info, address, and city.
- **Product Dimension** – Includes product attributes like name, category, pricing, discounts, and descriptions.
- **Date Dimension** – Stores time-related attributes like day, month, year, quarter, weekday, and weekend flag.
- **Review Dimension** – Contains the title and content of customer reviews linked to ratings and products.
- **Supplier Dimension** – Includes supplier details such as name, contact, and geographical location.
- **Geography Dimension** – Captures geographic data such as city, state/province, region, and country for location-based analysis.
- **Shipping Carrier Dimension** – Contains information about third-party shipping services used for delivering products.

## 3.2 Galaxy Schema



# Chapter 4: ETL Process

## 4.1 Data Extraction

For the ETL process, we used Talaend Open Studio, an open-source data integration tool.

**Tools & Components**

- File Input - To extract data from CSV and SQL sources (e.g., customer, product, review datasets).
- Table Input - Used to extract data from existing relational databases.
- Data Grid - Used to create sample static data for testing small segments (e.g., date_dim values).

**Data Sources**

- CSV Files – customer.csv, product.cs
- SQL Files – sales.sql
- Json Files – review.json
- Oracle Database: Used for storing the final data warehouse schema

## 4.2 Data Transformation

**Cleaning & Formatting**

- Handled null values with default values.
- Corrected data types (e.g., converting string dates to date format).
- Formatted dates to match the warehouse schema using Talend’s date functions.

**Business Rules**

- Discounted Price calculation using Calculator or Modified JavaScript Value transform.
- Customer Full Name applied using String Operations or Modified JavaScript Value step.

**Extra Logic**

- Duplicate removal with Unique Rows step
- Column renaming using Select Values
- Trimming whitespace

## 4.3 Data Loading

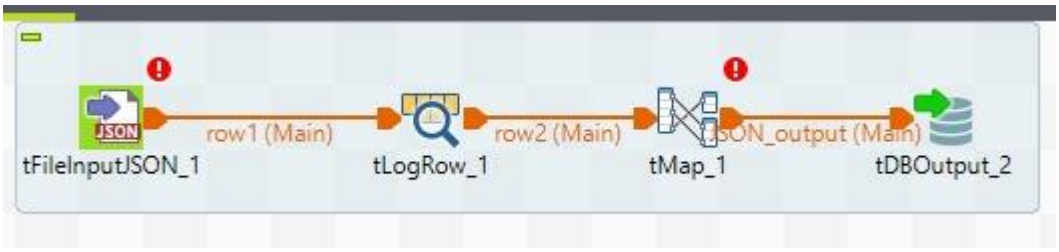
After data transformation, data will be loaded to MySQL datawarehouse.

- First, the Dimension tables will load
- Then the Fact tables will be loaded.

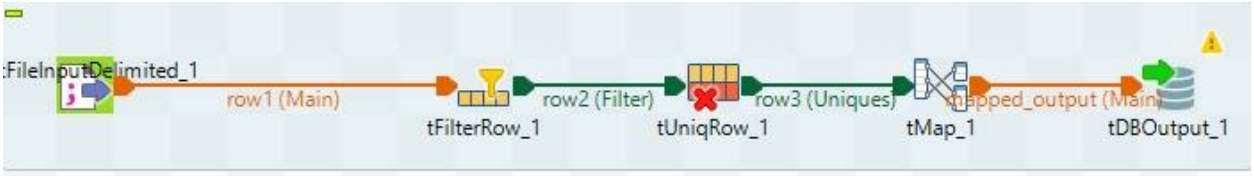
Reason: Integrating by loading referenced data first into the data warehouse.

## 4.4 ETL Sample Workflows

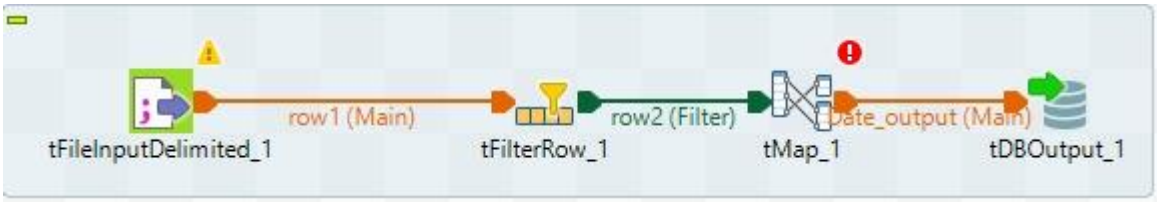
Reviews Data



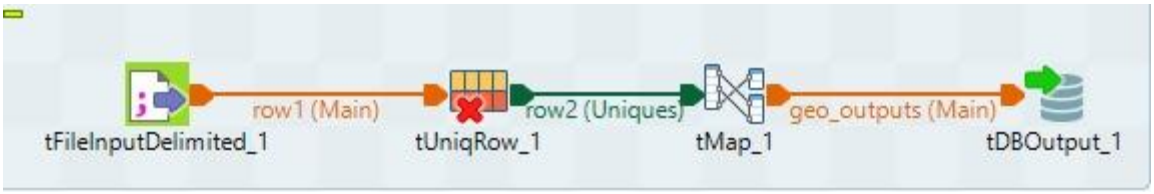

Customer Data



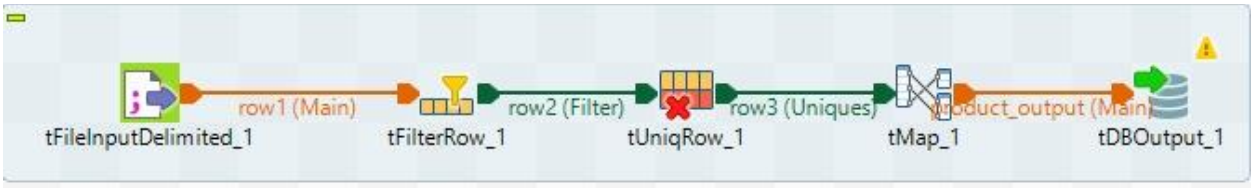

Date Data



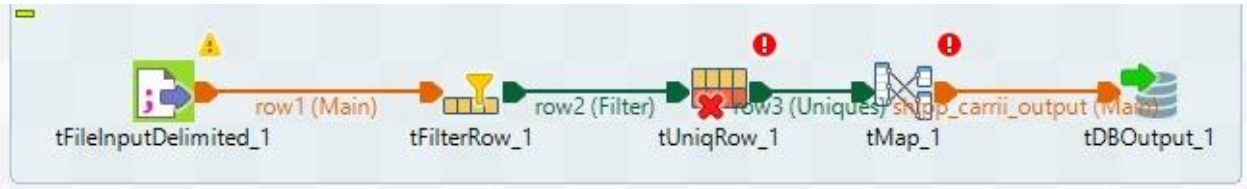
Geography Data



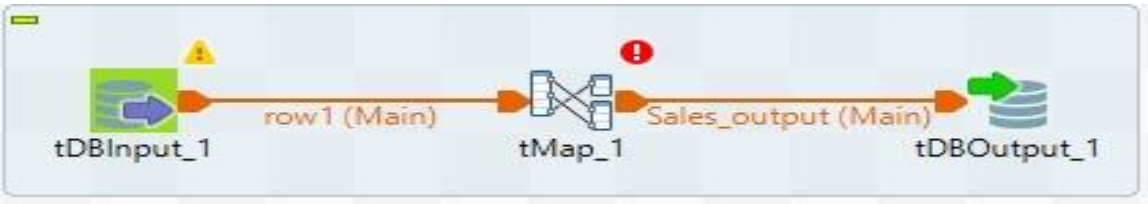

Product Data



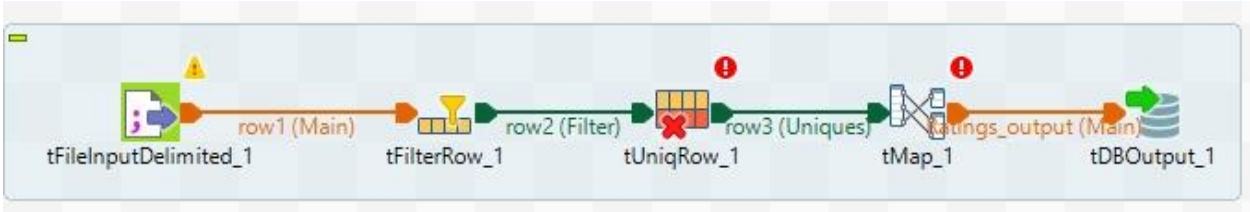

Shipping Carrier Data



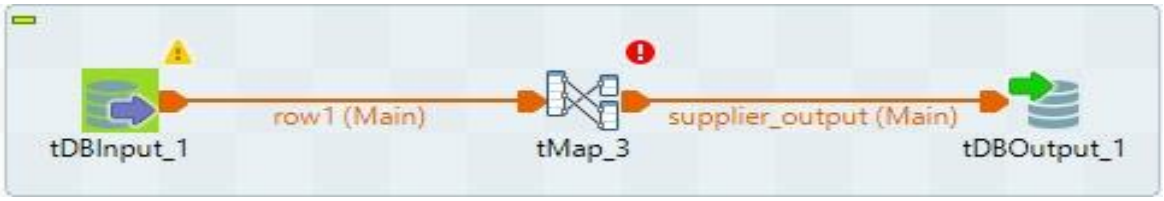

Sales Data



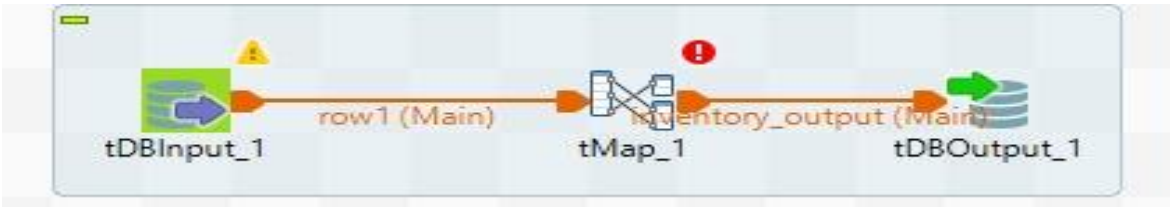

Ratings Data



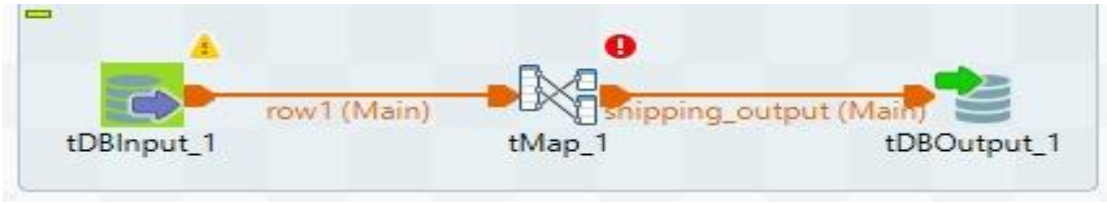
Supplier Data




Inventory Data




Shipping Data



# Chapter 5: Data Visualization

## 5.1 Connection between the Data Warehouse & Power BI

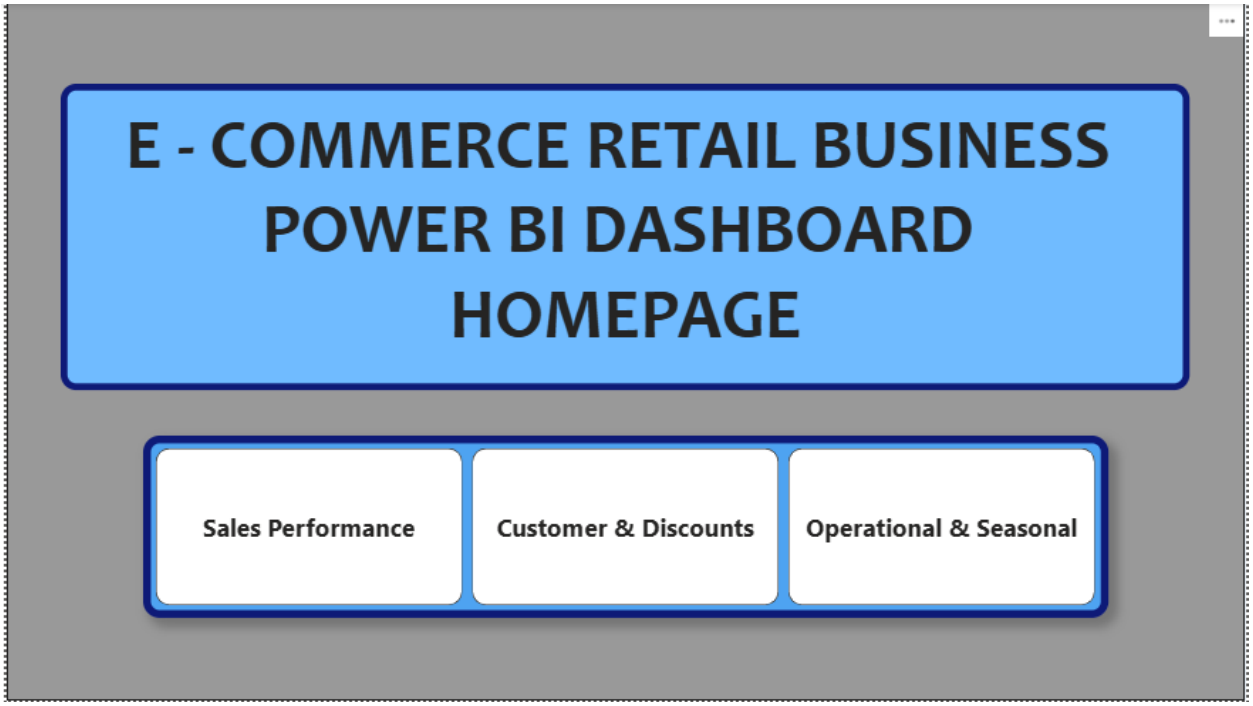
We used Power BI to vosialize the data to get analysis, therefore, we connected it to the MySQL data warehouse that we created for the e-commerce retail company.

- First, the data source selection was done, and the database credentials were added to congifured with read access.
- The necessary facts and dimensions tables were connected.
- In the data model view, we mapped the relationship between the tables based on foreign keys.

## 5.2 Dashboard Design

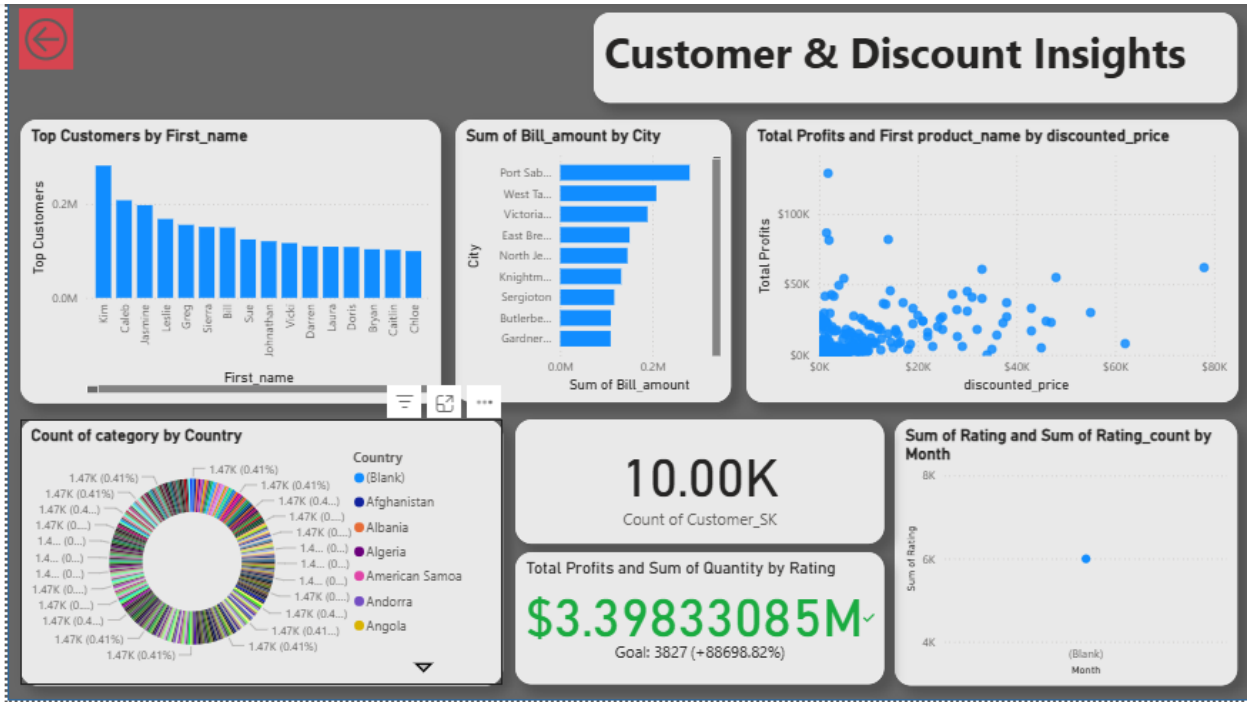
The following is our Power BI dashboard

1. Homepage

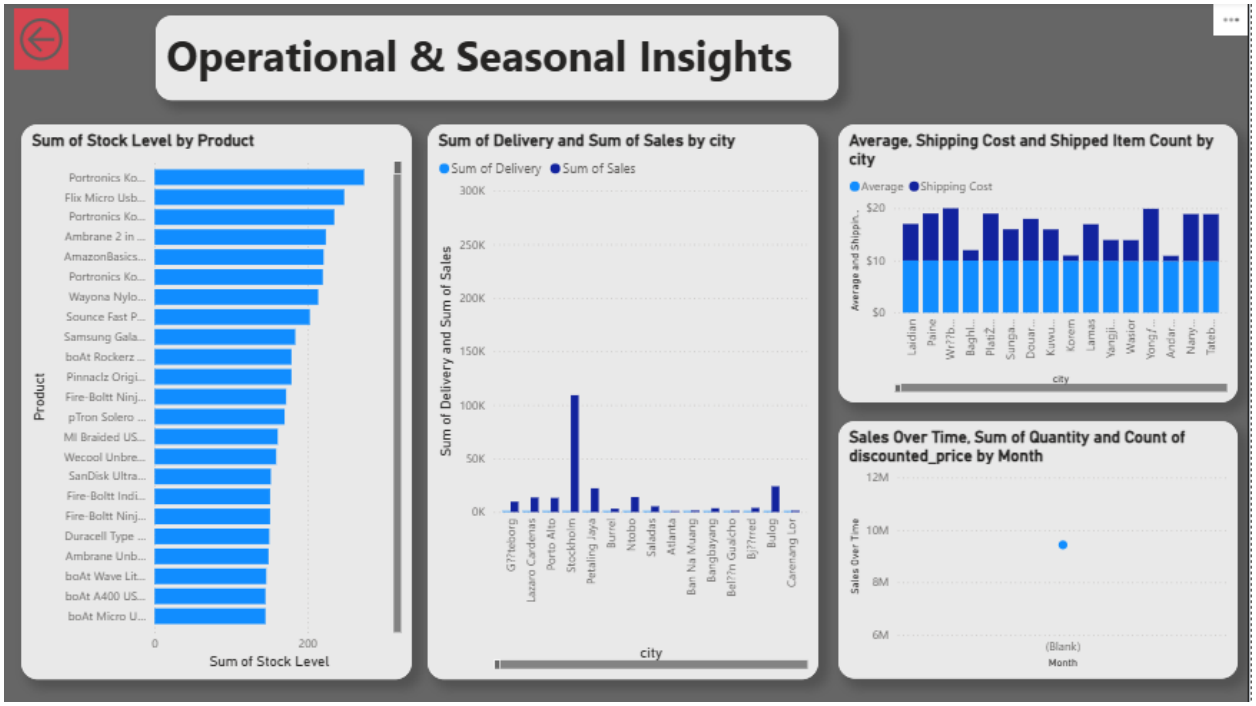


2. Sales Performance & Profitability Overview


3. Customer & Discount Insights



4. Operational & Seasonal Insights



# Chapter 6: Analysis & Insights

This chapter represents the main insights discovered by the dashboard.

- Able to figure out top selling and lowest sales products
    - Helps managements to focus on restocking based on the customer demand, manage pricing and promotions.
- Find which regions have more demand
    - Helps with target marketing, according to lowest sales regions and improve rest and have more strategies.
- Comparison between sales and customer ratings
    - High ratings might lead to better performance but may not as well. Analytics will help to figure pros and cons.
- Uncovering top performing products, categories and vice versa, identifying profit margins and support better pricing.

# Chapter 7: Conclusion

The design and construction of a data warehouse for an e-commerce company, a growing retailer of electrical items, is well demonstrated by this project. By merging data from multiple sources, including Excel spreadsheets, SQL dump files, and a JSON document, we were able to mimic real corporate data infrastructures.

The ETL procedure, which included extracting, cleaning, transforming, and loading the data into a structured Galaxy Schema, was carried out using Talend Open Studio. This schema architecture allowed for the effective organization of many fact tables, such as sales, inventory, and shipping, by utilizing shared dimension tables for customers, goods, location, and time.
We were able to overcome obstacles including inconsistent data formats, missing values, and the difficulty of schema mapping by using appropriate.