Final Project

Group 1:

Ahmed Ali • Anuphap Chansatit • Chotiros Srisiam • Halari Shanpru

Karthikeyan Jeyabalasuntharam • Nichapat Boonprasertsri  
Vitchaya Siripoppohn • Yat Chit Law

October 20, 2023

Contents

[**Business Assumption** 1](#_Toc148567836)

[**1.** **Data Quality** 2](#_Toc148567837)

[1.1. Criticality Principle 2](#_Toc148567838)

[1.1.1. Identification Policy 2](#_Toc148567839)

[1.1.1.1. Critical Data Identification Procedure 2](#_Toc148567840)

[1.1.1.2. Data Profiling and Scoring Procedure 2](#_Toc148567841)

[1.1.2. Quality Assurance Policy 2](#_Toc148567842)

[1.1.2.1. Quality Control Framework Procedure 2](#_Toc148567843)

[1.1.2.2. Data Cleansing and Enrichment Procedure 2](#_Toc148567844)

[1.2. Standards-driven Principle 2](#_Toc148567845)

[1.2.1. Standards Policy 3](#_Toc148567846)

[1.2.1.1. Standardization and Documentation Procedure 3](#_Toc148567847)

[1.2.1.2. Data Quality Training and Certification Procedure 3](#_Toc148567848)

[1.2.2. Continuous Monitoring and Adherence Policy 3](#_Toc148567849)

[1.2.2.1. Continuous Monitoring Framework Procedure 3](#_Toc148567850)

[1.2.2.2. Adherence Improvement Action Procedure 3](#_Toc148567851)

[**2.** **Data Integration & Interoperability** 3](#_Toc148567852)

[1.1. Business Accountability Principle 3](#_Toc148567853)

[2.1.1. Transparency Policy 3](#_Toc148567856)

[2.1.1.1. Data Integration Visibility Procedure 4](#_Toc148567857)

[2.1.1.2. Data Interaction Documentation Procedure 4](#_Toc148567858)

[2.1.2. Consistency and Efficiency Policy 4](#_Toc148567859)

[2.1.2.1. Data Integration Efficiency Procedure 4](#_Toc148567860)

[2.1.2.2. Data Consistency Validation Procedure 4](#_Toc148567861)

[1.2. Enterprise Perspective Principle 4](#_Toc148567862)

[2.2.1. Scalability and Extensibility Policy 4](#_Toc148567867)

[2.2.1.1. Scalability Assessment Procedure 4](#_Toc148567868)

[2.2.1.2. Extensibility Planning Procedure 4](#_Toc148567869)

[2.2.2. Compatibility and Interoperability Policy 4](#_Toc148567870)

[2.2.2.1. Interoperability Testing Procedure 5](#_Toc148567871)

[2.2.2.2. Compatibility Compliance Procedure 5](#_Toc148567872)

[**3.** **Data Security** 5](#_Toc148567873)

[2.1. Proactive Management Principle 5](#_Toc148567875)

[3.1.1. Risk Assessment and Mitigation Policy 5](#_Toc148567878)

[3.1.1.1. Security Risk Assessment Procedure 5](#_Toc148567879)

[3.1.1.2. Risk Mitigation Action Procedure 5](#_Toc148567880)

[3.1.2. Security Enhancement and Maintenance Policy 5](#_Toc148567881)

[3.1.2.1. Security Enhancement Planning Procedure 5](#_Toc148567882)

[3.1.2.2. Security Maintenance and Patching Procedure 5](#_Toc148567883)

[2.2. Clear Accountability Principle 5](#_Toc148567884)

[3.2.1. Roles and Responsibilities Policy 6](#_Toc148567890)

[3.2.1.1. Role-Based Access Control Procedure 6](#_Toc148567891)

[3.2.1.2. Access Review and Audit Procedure 6](#_Toc148567892)

[3.2.2. Security Compliance and Oversight Policy 6](#_Toc148567893)

[3.2.2.1. Security Compliance Monitoring Procedure 6](#_Toc148567894)

[3.2.2.2. Security Oversight Reporting Procedure 6](#_Toc148567895)

[**Lab Exercise 3’s Operational Report** 7](#_Toc148567896)

[**Lab Exercise 3’s Executive Report** 9](#_Toc148567897)

[**Annex** 10](#_Toc148567898)

[Lab Exercise 1 - Submission A 10](#_Toc148567899)

[Data Analysis 10](#_Toc148567900)

[Completeness 10](#_Toc148567901)

[Consistency 10](#_Toc148567902)

[Redundancies 10](#_Toc148567903)

[Duplicates 11](#_Toc148567904)

[Target Audience 11](#_Toc148567905)

[Operational Report 11](#_Toc148567906)

[Executive Report 11](#_Toc148567907)

[Context and Additional Assumptions 11](#_Toc148567908)

[Operational and Executive Reports 12](#_Toc148567909)

[Operational Report 12](#_Toc148567910)

[Executive Report 12](#_Toc148567911)

[Empty Templates for Reports 13](#_Toc148567912)

[Operational Report 13](#_Toc148567913)

[Executive Report 14](#_Toc148567914)

[Lab Exercise 2 - Submission B 15](#_Toc148567915)

[Analysis 15](#_Toc148567916)

[Entities 15](#_Toc148567917)

[Attributes 15](#_Toc148567918)

[Domains 15](#_Toc148567919)

[Data Cleansing 16](#_Toc148567920)

[Logical-level ERD 17](#_Toc148567921)

[Data Flow 18](#_Toc148567922)

[Database Schema 19](#_Toc148567923)

[SQL Scripts 21](#_Toc148567924)

[ETL Process 23](#_Toc148567925)

[Files 25](#_Toc148567926)

[Lab Exercise 3 - Submission C 26](#_Toc148567927)

[Operational Report 26](#_Toc148567928)

[Executive Report 26](#_Toc148567929)

[Files 26](#_Toc148567930)

# **Business Assumption**

* **Business Needs:** It is considered that prior to adopting any data governance initiatives, a thorough understanding of the organization's specific business goals and objectives is gained. Identifying vital data assets, business processes, and key performance indicators (KPIs) is part of this. One business process identified was that some sales are made with a negative profit, implying sales below cost. We believe this is likely due to the company’s desire to get rid of old stock, or as an occasional customer retention tactic, and possibly to undercut competitors. The customers’ names and addresses are important personally identifiable information which must be treated with strict and measured security.
* **Audience and Stakeholders:** The organization is considered to understand the need to identify and engage with all important stakeholders, both internal and external. Understanding their data needs and expectations is critical for effective data governance. The reports are addressed to Operational and Executive teams which is why they have different degrees of granularity and detail. We expect Executives to require a high-level overview of the company’s profitability by region, while Operations can examine various City and Branch metrics.
* **Technology Infrastructure:** A thorough examination of the existing technological infrastructure is carried out. This includes assessing the data management systems, tools, and platforms, as well as their compatibility with data governance initiatives. The primary data management tools used were MySQL Workbench and Python, and all the code is provided in the Annex section for review, verification, and replication.
* **Processes and Workflows:** It is expected that the organization understands the importance of mapping existing data processes and workflows. Understanding how data moves through the organization, who is responsible for it at each point, and where possible bottlenecks or issues may develop are all part of this. As mentioned earlier, the code is provided in the Annex section for transparency and reproducibility. Some unique Product Names were found to have duplicated Product IDs and treating them as such we defined new Product IDs. We achieved this by changing the Product ID by +1 for all applicable duplicates, which gave us a clean dataset.
* **Report Objectives:** The organization understands that reporting objectives must be clearly defined. This includes determining the types of reports needed, their frequency, the key metrics to be included, and the target audience for these reports. These reports are designed to advise Executive and Operations on metrics and KPIs, both current and historical, on appropriate time scales, to guide their decisions on improving business logic and processes.
  + - 1. **Data Quality**

Ensuring data accuracy, completeness, and consistency. This aspect of Data Governance is essential to the Superstore, as it ensures the metrics and progress tracked align with goals and benchmarks for employees and the whole company. Using Data Quality Principles, we can ensure customers can expect data-driven decisions from the company, such as customer segmentation, and pricing. This can increase customer loyalty, retention, and lifetime value.

* 1. Criticality Principle

Not all data is of equal importance. Focus on critical data to ensure it meets high-quality standards, which are essential for decision-making and business operations.

* + 1. Identification Policy

Identify critical data elements that significantly impact decision-making and business operations to ensure a clear understanding of the most important data, making it a top priority for quality control and protection.

* + - 1. Critical Data Identification Procedure

Collaborate with stakeholders to identify critical data elements to continuously review and validate the list of critical data to ensure its relevance and accuracy.

* + - 1. Data Profiling and Scoring Procedure

Implement data profiling techniques and scoring systems for critical data elements to regularly assess the quality of critical data and generate operational reports highlighting data quality scores.

* + 1. Quality Assurance Policy

Implement stringent quality control measures for critical data, ensuring it meets high-quality standards to guarantee that critical data is consistently accurate, reliable, and readily available for decision-making and operational needs.

* + - 1. Quality Control Framework Procedure

Establish a comprehensive quality control framework for critical data to continuously apply quality checks and audits to ensure data conforms to quality standards, generating executive reports on data quality trends.

* + - 1. Data Cleansing and Enrichment Procedure

Employ data cleansing and enrichment processes to enhance data quality. to regularly execute data cleansing procedures and report on the improvements in data quality achieved.

* 1. Standards-driven Principle

Consistency is key. Having well-defined standards ensures that data is consistently accurate and reliable, making it easier to maintain high data quality across the organization.

* + 1. Standards Policy

Standardization and Documentation Procedure: Develop and document data quality standards and guidelines to ensure adherence to standards and generate operational reports highlighting instances of compliance or deviations.

* + - 1. Standardization and Documentation Procedure

Develop and document data quality standards and guidelines to ensure adherence to standards and generate operational reports highlighting instances of compliance or deviations.

* + - 1. Data Quality Training and Certification Procedure

Provide training and certification programs on data quality standards to track employee certifications and generate executive reports showcasing the level of staff compliance with data quality standards.

* + 1. Continuous Monitoring and Adherence Policy

Ensure that data systems are compatible and interoperable across the enterprise to promote seamless data exchange and interaction between different systems, improving overall efficiency.

* + - 1. Continuous Monitoring Framework Procedure

implement a continuous monitoring framework for data quality to regularly assess and report on data quality deviations and areas that require corrective action.

* + - 1. Adherence Improvement Action Procedure

Develop action plans to address data quality deviations and non-compliance to generate operational and executive reports on the effectiveness of these actions in maintaining data quality standards.

1. **Data Integration & Interoperability**

Create a data integration plan to collect, transform, and load data from different sources, like sales, orders, and customer data, into the Superstore dataset. This might involve ETL processes.

* 1. Business Accountability Principle

Data integration and interoperability play essential roles in ensuring business accountability by promoting transparency, consistency, and efficiency in data management and system interactions.

* + 1. Transparency Policy

Promote transparency in data integration processes to ensure business accountability to enable clear visibility into data integration activities, fostering trust and accountability in data management.

* + - 1. Data Integration Visibility Procedure

Establish data integration tracking mechanisms for transparency to generate operational reports showing the status and progress of data integration projects and their impact.

* + - 1. Data Interaction Documentation Procedure

Document data interactions and integrations with external systems to create executive reports showcasing the efficiency and effectiveness of data interactions with partners and external systems.

* + 1. Consistency and Efficiency Policy

Design data integration processes for consistency and efficiency to support business accountability to streamline data integration to ensure that data flows smoothly and consistently, enhancing overall accountability.

* + - 1. Data Integration Efficiency Procedure

Optimize data integration processes for consistency and efficiency to monitor and report on the efficiency gains achieved through process optimization.

* + - 1. Data Consistency Validation Procedure

Develop data consistency validation checks during integration to generate operational reports on the number of data consistency checks performed and the outcomes.

2.2. Enterprise Perspective Principle

Design with an enterprise mindset to guarantee future scalability and extensibility.

* 1. Enterprise Perspective Principle
     1. Scalability and Extensibility Policy

Develop data integration solutions with an enterprise perspective, considering future scalability and extensibility to ensure that data integration solutions can adapt and grow with the organization's needs, minimizing disruptions.

* + - 1. Scalability Assessment Procedure

Periodically assess the scalability of data integration solutions to generate executive reports on the readiness and adaptability of data integration solutions to future business needs.

* + - 1. Extensibility Planning Procedure

Create plans for extending data integration solutions to generate operational reports on the progress of extending data integration capabilities and their impact.

* + 1. Compatibility and Interoperability Policy

Ensure that data systems are compatible and interoperable across the enterprise to promote seamless data exchange and interaction between different systems, improving overall efficiency.

* + - 1. Interoperability Testing Procedure

Conduct regular interoperability testing with external systems to generate operational reports on the results of interoperability testing and any issues encountered.

* + - 1. Compatibility Compliance Procedure

Ensure data systems adhere to compatibility standards to generate executive reports on the compatibility of data systems and any required actions to maintain interoperability.

* + - 1. **Data Security**

Protecting data from unauthorized access and breaches.

1. 1. Proactive Management Principle

Anticipate and address security risks before they become threats. Regularly assess and enhance security measures to protect data.

1. 1. 1. Risk Assessment and Mitigation Policy

Conduct regular risk assessments and proactively mitigate security risks to identify and address potential security threats before they can harm data integrity.

* + - 1. Security Risk Assessment Procedure

Perform regular risk assessments to identify potential security threats to generate operational reports highlighting identified risks and potential mitigation strategies.

* + - 1. Risk Mitigation Action Procedure

Implement actions to mitigate identified security risks to generate executive reports on the effectiveness of mitigation efforts and any remaining risks.

* + 1. Security Enhancement and Maintenance Policy

Continuously enhance and maintain security measures to protect data to ensure that security measures remain effective and up-to-date, safeguarding data from evolving threats.

* + - 1. Security Enhancement Planning Procedure

Develop plans for enhancing data security measures to generate operational reports on the progress of security enhancement projects and their impact.

* + - 1. Security Maintenance and Patching Procedure

Regularly maintain and update security measures to generate executive reports on the maintenance activities performed and their impact on the overall security posture.

* 1. Clear Accountability Principle

Ensure that individuals or teams have specific roles and are held accountable for safeguarding data.

3. 2. 1. Roles and Responsibilities Policy

Define clear roles and responsibilities for individuals or teams responsible for data security to ensure that specific entities are accountable for safeguarding data and understand their responsibilities.

* + - 1. Role-Based Access Control Procedure

Implement role-based access control for data and systems to generate operational reports on user access patterns and permissions compliance.

* + - 1. Access Review and Audit Procedure

Conduct regular access reviews and security audits to generate executive reports summarizing access review findings and audit results, along with recommendations for improvements.

* + 1. Security Compliance and Oversight Policy

Establish a compliance and oversight framework to monitor and enforce data security practices to enforce security practices consistently and transparently while complying with relevant regulations and standards.

* + - 1. Security Compliance Monitoring Procedure

Monitor and ensure compliance with security policies and regulations to generate operational reports on the level of compliance and areas requiring attention or improvement.

* + - 1. Security Oversight Reporting Procedure

Provide oversight reports to management and stakeholders to generate executive reports on the state of security compliance and measures taken to address any non-compliance issues.

# **Lab Exercise 3’s Operational Report**



一張含有 文字, 功能表, 數字, 文件 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

The operational report shows the sales data broken down by Region, State, City, Sub-Category, Product Name.

The first few rows provide data for the Central region, specifically Illinois and Arlington. The Sub-Category is Art, and the Product Name is Newell 332. The Quantity sold was 6, resulting in Sales of $14.11 and a Profit of $1.23.

There are also Sub-Total rows for the Art category in Arlington, Illinois. The Quantity sold was 6, resulting in Sales of $14.11 and a Profit of $1.23.

The third row provides a Sub-total for Arlington, Illinois with a total Quantity of 6, Sales of $14.11 and a Profit of $1.23.

The fourth row provides a total for Illinois with a total Quantity of 1845, Sales of 60185.77, but a negative Profit of -8537.52. This indicates a loss.

The final row provides a total for the Central region with a Quantity of 8780, Sales of $302,911.4 and a Profit of $32,100.7.

# **Lab Exercise 3’s Executive Report**



A screenshot of a computer screen

Description automatically generated

The executive report depicts the sales performance of three main categories in various regions during the second quarter of 2021.

Our goods are divided into three categories: "Office Supplier", "Furniture", and "Technology". The superstore is also divided into four regions, which are "Central", "East", "South", and "West".

The report has three primary sections: categorical sales performance by region and national, categorical profit margins, and a categorical sales ratio comparison of Q1 and Q2 performance in the last several years.

For the first section, there are sales and profit comparisons between three categories in one region and all regions at the same time.

To determine which category is the greatest seller in a region, the sums of all percentages of sales or profit must be 100% in one region.

To demonstrate which category in which area is the greatest seller, summaries of all percentages of sales or profit must be 100% for all regions.

The second section displays the profit margin ratios of the various categories. The equation is (Profit/Sales)x100%. It measures a company's ability to manage its costs in relation to its income. It is stated as a percentage, with larger percentages signifying more profit.

The last section displays the categorical sales ratios for the first quarters of 2019, 2020, and 2021, as well as a comparison between the first and second quarters of that year. It might illustrate the annual or seasonal selling pattern of several categories, whether it is decreasing or increasing.

# **Annex**

## Lab Exercise 1 - Submission A

### Data Analysis

#### Completeness

We identified a completeness issue in the dataset. Specifically, the Postal Code column contained 11 missing entries (Figure 1).  
  
A screenshot of a computer

Description automatically generated

Figure 1: Missing Values (Postal Codes)

#### Consistency

We discovered an inconsistency related to the format of Postal Codes. While examining the dataset, we observed that some Postal Code entries contained only four digits (Figure 2). Further investigation revealed that these entries should have included five digits, with a leading zero. Unfortunately, this leading zero was dropped during the conversion of the dataset to a .csv file.

A screenshot of a computer

Description automatically generated

Figure 2: Examples of 4 digits Postal Codes

Apart from the Postal Codes format issue, our analysis did not uncover any other inconsistencies within the dataset. The data appears to be consistent in terms of formatting, units, and other key attributes.

#### Redundancies

Our assessment did not reveal any instances of redundant data.

#### Duplicates

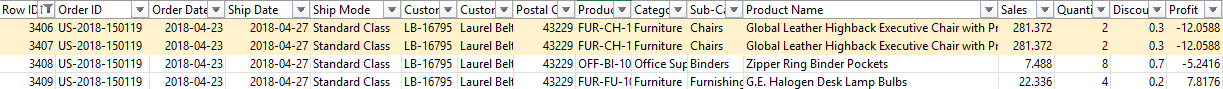
We identified a case of duplicate data, where the same order was recorded twice (Figure 3).  


Figure 3: Duplicate Data

### Target Audience

#### Operational Report

The operational report is primarily designed for the Sales Team. This report serves as a crucial resource for the dedicated members of the Sales Team, who are responsible for overseeing sales performance within their respective state in regional offices. Its primary objective is to facilitate efficient monitoring and control of sales activities. By providing key performance metrics and insights over a monthly timeframe, it empowers the Sales Team to ensure that sales targets are consistently met and that business processes run smoothly within the context of the defined period.

#### Executive Report

The executive report is directed towards Regional Managers. This specialized report caters to the informational needs of Regional Managers, who hold a pivotal role in strategic decision-making. By offering comprehensive insights into sales performance, profitability, and areas requiring improvement, such as returns, this report equips Regional Managers with the data required to make informed and impactful decisions. It serves as a valuable resource for enhancing overall sales operations and profitability, aligning the organization's goals with strategic actions at the regional level.

### Context and Additional Assumptions

* High-level management will consult executive reports when making strategic decisions.
* Employees immediately involved in sales and order processing will use operational reports.
* Executive and operational reports will be created for each designated region.
* The profits for some orders were negative because of the discounts on the products.
* Both reports will be produced on a regular basis, either monthly or quarterly.
* The data in the Sample Superstore spreadsheet is taken to be an accurate representation of the sales data for the business.
* Reports will be kept up to date with recent information on a regular basis.

### Operational and Executive Reports

#### Operational Report

The operational report is designed to provide a comprehensive overview of sales activities during the specified report period. The report format includes the following key elements:

* Report Period: The operational report captures data over a defined period (monthly), allowing frequent assessment of sales performance.
* Region, State, City, Sub-category, and Item Name: These categorical elements serve as the basis for a detailed breakdown of sales data, enabling a thorough analysis of product performance across different dimensions.
* Previous Sales ($): This column represents the total sales amount of the item in the state for the previous period (month).
* Current Sales ($): This column represents the total sales amount of the item in the state for the current period (month).
* Quantity: This column represents total item sold in the state in the current period (month)
* Profit ($): This column represents the financial outcome of each sale, accounting for expenses and discounts.
* Sales Growth Rate (%): This column calculates the percentage change in sales between the previous and current month, aiding in performance evaluation.   
  *Sales Growth Rate (%) = (Current Sales – Previous Sales / Previous Sales) \*100*
* Sub-Total Rows: Sub-total rows are included for each sub-category and item name, presenting aggregated figures for a quick summary of performance within specific product categories and items.
* Total Row: The total row displays cumulative figures for the entire report period, offering a holistic view of selected sales performance.

#### Executive Report

The executive report is focused on presenting quarterly gross sales data. The report structure encompasses the following elements:

* Region: The region section specifies the geographic scope of the report.
* State: This column represents insights into sales at the state level.
* Sales by Quarter ($): These columns represent the sales figures for the previous quarter, the current quarter, and the corresponding percentage change. Additionally, it presents sales figures for the same quarter in the previous year and the percentage change.
* Net Profit Margin Ratio (%): This column represents the percentage of profit earned from each dollar of sales. It is a key financial metric that measures the profitability of a business by indicating how much profit is generated for each dollar of sales.  
  *Net Profit Margin Ratio = (Net Profit/Total Sales) \* 100*
* Discount Effective Rate (%): This column represents the percentage of the total sales revenue that is attributed to discounts. It helps measure the impact of discounts on the overall sales revenue as a percentage.  
  *Discount Effective Rate = ((Total Sales with Discount - Total Sales without Discount) / Total Sales without Discount) \* 100*

### Empty Templates for Reports

#### Operational Report

**Sales Team Operations by Monthly**

Report Period: 03/01/2020 - 03/31/2020 {start date - end date}

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Region | State | City | Sub-category | Item Name | Previous Sales ($) | Current Sales ($) | Quantity | Profit ($) | Sales Growth Rate (%) |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Sub-total** | |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Sub-total** | |  |  |  |  |  |
| **Sub-total** | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Sub-total** | |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Sub-total** | |  |  |  |  |  |
| **Sub-total** | | |  |  |  |  |  |
| **Total** | | | |  |  |  |  |  |
| Total | | | | |  |  |  |  |  |

#### Executive Report

**Sales Team Quarterly Executive Report**

Region: East {Selected Region}

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| State | Sales Q1-2020 {previous quarter} ($) | Sales Q2-2020 {current quarter} ($) | Sales Comparison Q2-2020 vs. Q1-2020 (%) | Sales Q2-2019 {previous year} ($) | Sales Comparison Q2-2020 {current year} vs. Q2-2019 (%) | Net Profit Margin Ratio (%) | Discount Effective Rate (%) |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **Total** |  |  |  |  |  |  |  |

## Lab Exercise 2 - Submission B

### Analysis

#### Entities

Entities include:

* Customers
* Products
* Regional Managers
* Addresses
* Orders (invoices for example)

#### Attributes

Attributes are:

* Orders: Order ID (PK), Order Date, Ship Date, Ship Mode
* Customers: Customer ID (PK), Customer Name, Segment
* Addresses: City, State, Postal Code (PK), Region
* Products: Product ID (PK), Category, Sub-Category, Product Name
* Sales: Order No (PK), Sales, Quantity, Discount, Profit, Price
* Order Records: Order No (PK), Order ID, Customer ID, Product ID

#### Domains

Domains are:

* Order No – unique integer – VARCHAR(4)
* Order ID – Fixed length alphanumeric ID AA-YYYY-###### - VARCHAR(15)
* Order Date – Date DD/M/YYYY - TEXT
* Ship Date – Date DD/M/YYYY - TEXT
* Ship Mode – First Class, Second Class, Standard Class ONLY - TEXT
* Customer ID – Alphanumeric - Customer Initials (2)-5-digit ID AA-##### - VARCHAR(8)
* Customer Name – Character string: First Name + Last Name - TEXT
* Segment – Consumer, Home Office, Corporate ONLY - TEXT
* Country/Region - US ONLY - TEXT
* City – Valid US City ONLY - TEXT
* State – Valid State ONLY - TEXT
* Postal Code – 5 digits ##### ONLY - VARCHAR(5)
* Region - East, South, West, Central ONLY - TEXT
* Product ID – 3 char category + 2 char subcategory + 8-digit unique - VARCHAR(15)
* Category Furniture Office Supplies Technology ONLY - TEXT
* Sub-Category - TEXT
  + If Furniture: Bookcases Chairs Tables Furnishings ONLY
  + If Office Supplies: Appliances Art Binders Envelopes Fasteners Labels Paper Storage Supplies ONLY
  + If Technology: Accessories Copiers Machines Phones ONLY
* Product Name: Unique String - TEXT
* Sales: $ amount - TEXT
* Quantity: Integer - INT
* Discount: Decimal/fraction - DOUBLE
* Profit: $ value – TEXT

The referential integrity is maintained by holding the Primary Keys of each table in the Order Records Table as Foreign Keys. The Row No. is renamed Order No.

Order Records contains:

* Order No (PK) - This is the primary key for this table
* Order ID – This is a Foreign Key for this table, and Primary Key for Orders table
* Customer ID – This is a Foreign Key for this table, and Primary Key for Customers table
* Product ID - This is a Foreign Key for this table, and Primary Key for Products table

### Data Cleansing

* Rounded up Sales/Profit to 2 decimal places.
* Fixed Product IDs with multiple products
* Turning all postal codes to US Postal Code format which contains 5 digits.
* Fixed Postal Codes with multiple cities

### Logical-level ERD

A computer screen shot of a diagram

Description automatically generated

### Data Flow

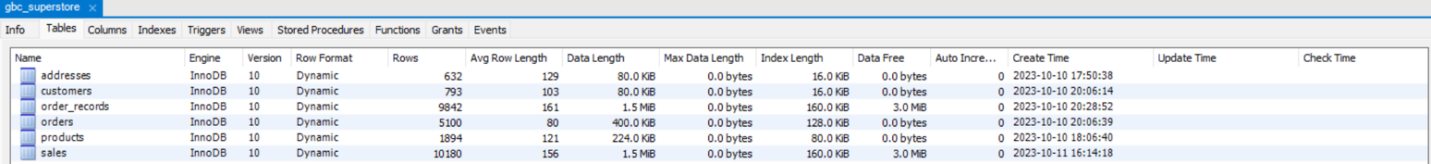
A screenshot of a diagram

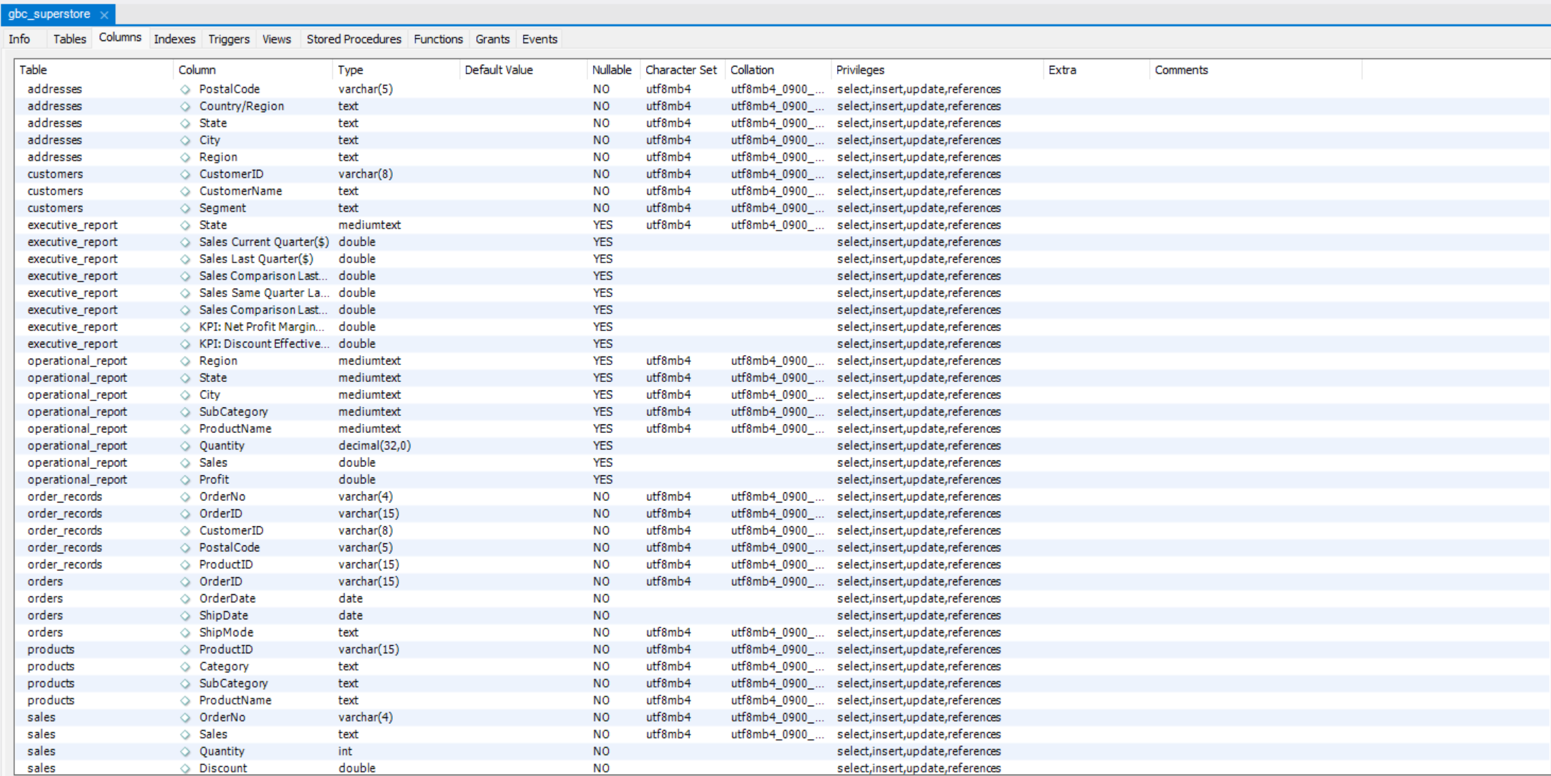
Description automatically generated

### Database Schema

A screenshot of a computer

Description automatically generated





A screenshot of a computer

Description automatically generated

#### SQL Scripts

|  |  |
| --- | --- |
| **Process** | **Code** |
| Create Database | CREATE DATABASE GBC\_Superstore; |
| Create Tables | USE GBC\_Superstore;    CREATE TABLE `order\_records` (  `OrderNo` varchar(4) NOT NULL,  `OrderID` varchar(15) NOT NULL,  `CustomerID` varchar(8) NOT NULL,  `PostalCode` varchar(5) NOT NULL,  `ProductID` varchar(15) NOT NULL,  PRIMARY KEY (`OrderNo`),  UNIQUE KEY `Order No\_UNIQUE` (`OrderNo`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci;    CREATE TABLE `customers` (  `Customer ID` varchar(8) NOT NULL,  `Customer Name` text NOT NULL,  `Segment` text NOT NULL,  PRIMARY KEY (`Customer ID`),  UNIQUE KEY `Customer ID\_UNIQUE` (`Customer ID`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci;    CREATE TABLE `addresses` (  `PostalCode` varchar(5) NOT NULL,  `Country/Region` text NOT NULL,  `State` text NOT NULL,  `City` text NOT NULL,  `Region` text NOT NULL,  PRIMARY KEY (`PostalCode`),  UNIQUE KEY `PostalCode\_UNIQUE` (`PostalCode`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci;    CREATE TABLE `orders` (  `Order ID` varchar(15) NOT NULL,  `Order Date` DATE NOT NULL,  `Ship Date` DATE NOT NULL,  `Ship Mode` text NOT NULL,  PRIMARY KEY (`Order ID`),  UNIQUE KEY `Order ID\_UNIQUE` (`Order ID`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci;    CREATE TABLE `products` (  `ProductID` varchar(15) NOT NULL,  `Category` text NOT NULL,  `SubCategory` text NOT NULL,  `ProductName` text NOT NULL,  PRIMARY KEY (`ProductID`),  UNIQUE KEY `Product ID\_UNIQUE` (`ProductID`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci;    CREATE TABLE `sales` (  `Order No` varchar(4) NOT NULL,  `Sales` text NOT NULL,  `Quantity` int NOT NULL,  `Discount` double NOT NULL,  `Profit` text NOT NULL,  `Price` text NOT NULL,  PRIMARY KEY (`Order No`),  UNIQUE KEY `Order No\_UNIQUE` (`Order No`)  ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci; |

### ETL Process

|  |  |
| --- | --- |
| **Process** | **Code** |
| Generate and export CSV file | import pandas as pd  # Read Excel file  xlsx = pd.ExcelFile('GBC\_Superstore.xlsx')  # Get a list of sheet names  sheet\_names = xlsx.sheet\_names  # Iterate through each sheet and save it as a CSV file  for sheet\_name in sheet\_names:  df = pd.read\_excel(xlsx, sheet\_name) # Read the sheet into a DataFrame  df.to\_csv(f'{sheet\_name}.csv', index=False) # Save as a CSV file |
| Load CSV files into MySQL tables | LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\orders.csv'  INTO TABLE orders  FIELDS TERMINATED BY ';' ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS;    LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\addresses.csv'  INTO TABLE addresses  FIELDS TERMINATED BY ';'ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS;    ALTER TABLE products MODIFY COLUMN ProductName VARCHAR(250) CHARACTER SET latin1;  SET NAMES utf8; -- or SET NAMES latin1; or the appropriate character set    LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\products.csv'  INTO TABLE products  FIELDS TERMINATED BY ';'ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS;    LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\customers.csv'  INTO TABLE customers  FIELDS TERMINATED BY ';'ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS;    LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\sales.csv'  INTO TABLE sales  FIELDS TERMINATED BY ';'ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS;    LOAD DATA INFILE 'C:\Users\yatch\Desktop\Foundation of Data Management\Ass 2\order\_records.csv'  INTO TABLE order\_records  FIELDS TERMINATED BY ';'ENCLOSED BY '"'  LINES TERMINATED BY '\r\n'  IGNORE 1 ROWS; |
| Verify data completeness, data integrity and referential integrity | SELECT COUNT(\*) FROM order\_records;  SELECT \* FROM order\_records WHERE OrderNo NOT IN (SELECT OrderNo FROM sales);  SELECT \* FROM order\_records WHERE CustomerID NOT IN (SELECT CustomerID FROM customers);  SELECT \* FROM order\_records WHERE OrderID NOT IN (SELECT OrderID FROM orders);  SELECT \* FROM order\_records WHERE ProductID NOT IN (SELECT ProductID FROM products);  SELECT \* FROM order\_records WHERE PostalCode NOT IN (SELECT PostalCode FROM addresses);    SELECT COUNT(\*) FROM customers;  SELECT \* FROM customer WHERE CustomerID IS NULL;  SELECT \* FROM customer WHERE CustomerID NOT IN (SELECT CustomerID FROM order\_records);    SELECT COUNT(\*) FROM addresses;  SELECT \* FROM addresses WHERE PostalCode IS NULL;  SELECT \* FROM addresses WHERE PostalCode NOT IN (SELECT PostalCode FROM order\_records);    SELECT COUNT(\*) FROM products;  SELECT \* FROM product WHERE ProductID IS NULL;  SELECT \* FROM product WHERE ProductID NOT IN (SELECT ProductID FROM order\_records);    SELECT COUNT(\*) FROM sales;  SELECT \* FROM sales WHERE OrderNo IS NULL;  SELECT \* FROM sales WHERE OrderNo NOT IN (SELECT OrderNo FROM order\_records);    SELECT COUNT(\*) FROM orders;  SELECT \* FROM orders WHERE OrderID IS NULL;  SELECT \* FROM orders WHERE OrderID NOT IN (SELECT OrderID FROM order\_records); |
| Create Foreign keys to relate different tables | ALTER TABLE `gbc\_superstore`.`order\_records`  ADD INDEX `FK\_OrderOrder\_idx` (`OrderID` ASC) VISIBLE,  ADD INDEX `FK\_OrderCustomer\_idx` (`CustomerID` ASC) VISIBLE,  ADD INDEX `FK\_OrderAddress\_idx` (`PostalCode` ASC) VISIBLE,  ADD INDEX `FK\_OrderProduct\_idx` (`ProductID` ASC) VISIBLE;  ;  ALTER TABLE `gbc\_superstore`.`order\_records`  ADD CONSTRAINT `FK\_OrderOrder`  FOREIGN KEY (`OrderID`)  REFERENCES `gbc\_superstore`.`orders` (`OrderID`)  ON DELETE NO ACTION  ON UPDATE NO ACTION,  ADD CONSTRAINT `FK\_OrderSales`  FOREIGN KEY (`OrderNo`)  REFERENCES `gbc\_superstore`.`sales` (`OrderNo`)  ON DELETE NO ACTION  ON UPDATE NO ACTION,  ADD CONSTRAINT `FK\_OrderCustomer`  FOREIGN KEY (`CustomerID`)  REFERENCES `gbc\_superstore`.`customers` (`CustomerID`)  ON DELETE NO ACTION  ON UPDATE NO ACTION,  ADD CONSTRAINT `FK\_OrderAddress`  FOREIGN KEY (`PostalCode`)  REFERENCES `gbc\_superstore`.`addresses` (`PostalCode`)  ON DELETE NO ACTION  ON UPDATE NO ACTION,  ADD CONSTRAINT `FK\_OrderProduct`  FOREIGN KEY (`ProductID`)  REFERENCES `gbc\_superstore`.`products` (`ProductID`)  ON DELETE NO ACTION  ON UPDATE NO ACTION; |

#### Files



## Lab Exercise 3 - Submission C

### Operational Report

A table with numbers and names

Description automatically generated

### Executive Report

A screenshot of a computer screen

Description automatically generated

#### Files

