| **ASSIGNMENT – 30%** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Module | | **IBM 3202 DATAWAREHOUSE AND MULTIDIMENSIONAL MODELLING**  **BDS3402 DATA WAREHOUSE** | | |
| Session | | **JANUARY 2025** | | |
| Programme | | **BITI/BCSI/BTDS** | | |
| Section | | 1H1/1G1/1DS1 | | |
| Lecturers | | **Chong Fong Kim** | | |
| Coursework Type | | **Assignment** | | |
| Percentage | | **30% out of 100%** | | |
| Hand-out Date | | **Week 4** | Due Date | **Week 14** |
| **Students’ Declaration:** | | | | | |
| ***We declare that:***   * ***We understand what is meant by plagiarism, self-plagiarized and academic dishonesty.*** * ***This project is all our own work and we have acknowledged any use of the published or unpublished works of other people.*** * ***We hold a copy of this project which we can produce if the original is lost or damaged***  | **Name** | | **ID** | **Section** | **Signature** | **Date** | | --- | --- | --- | --- | --- | --- | | **1** | **Christo Tonio** | **I25032053** | **1G3** | **Christo** | **18/04/2025** | | **2** | **Parvendan Rangasamy** | **I25032142** | **1G3** | **Pary** | **18/04/2025** | | **3** | **Bhulakshmi Ainala** | **I25032054** | **1G3** | **Bhu** | **18/04/2025** | | **4** | **Ashish Sree Vatsav** | **I25032140** | **1G3** | **Ashish** | **18/04/2025** | | **5** |  |  |  |  |  | | | | | | |
| **Learning Outcomes Assessed:** | | | | | |
| LO2 | Perform the physical implementation of OLAP Data Warehousing modelling. (P4, PLO3) | | | | |

**Note: Plagiarism & Copying**

**PLAGIARISM AND COPYING ARE STRICTLY PROHIBITED AND CONSIDERED A SERIOUS OFFENCE. THE FACULTY HAS THE FULL RIGHTS TO TAKE DISCIPLINARY ACTIONS AGAINST ANY STUDENT WHO PLAGIARISE OR COPY THE WORK OF OTHERS.**

The Faculty reserves the right to use any means to identify plagiarism and copying. If you reference the work of other authors, you must ensure that you properly reference them to avoid the charge of plagiarism.

Copying the work of others (current students, past students or the works of others) automatically earn you a Fail grade. If two or more students copied from each other, both students (regardless of who copies from whom) will earn a Fail grade.

**ASSIGNMENT RUBRIC**

| **Criterion** | **Weak** | **Moderate** | **Good** | **Excellent** | **Given Marks** |
| --- | --- | --- | --- | --- | --- |
| **Part A:**  **Business specification / business rules**  **(20 marks)** | Poor discussion  related to the scenario. Include minimal/no discussion on activities related to the particular domain. **(0-4 marks)** | Basic/Minimal discussion related to the scenario. Includes incomplete description of activities related to the particular domain.**(5-9 marks)** | Good introduction and clear definition and description that are related to the particular domain.  **(10-14 marks)** | Excellent description of the business specification/ business rules with precise description of overall activities of a particular domain. **(15-20 marks)** |  |
| **Part B:**  **Multidimensional**  **Model (Schema)**  **(20 marks)** | None or wrong model have been constructed with none/minimum number of primary keys, attribute and foreign keys in place. None/wrong justification on the model chosen.  **(0-4 marks)** | Incomplete multidimensional model star/snowflake/starflake) schema with some primary keys, foreign keys and attributes missing. There is insufficient justification on the model chosen. **(5-9 marks)** | Design a complete multidimensional model (star/snowflake/starflake) schema with primary keys, foreign keys and attributes in place. Good justification on the type of the model chosen. **(10 – 14 marks)** | Successfully design a complete and accurate multidimensional model (star/snowflake/starflake) schema with conformed dimensions, primary keys, foreign keys and attributes in place. Excellent justification on the type of model chosen.  **(15 – 20 marks)** |  |
| **Part C:**  **Data Integration (ETL)**  **(25 marks)** | Raw data is not feasible and some or most does not reflect Part B. Most of the ETL processes is failed to be completed. No or wrong explanation on the processes included. **(0 –5 marks)** | There are raw data that reflects Part B with some are incomplete. ETL process is partially done successfully which a few parts show error. None or few explanation on the processes. **(6 - 12 marks)** | Complete raw data that reflects Part B. ETL process is done successfully with some of the processes shown in detailed as well as some explanation on the process.  (**13 - 18 marks)** | Complete and feasible raw data that reflects Part B very well. ETL process is completed successfully and the processes is shown in detailed. Included meaningful explanation on the process.  **(19 - 25 marks)** |  |
| **Part D:**  **Queries (25 marks)** | Have 0-2 queries with same operator used with no OLAP operations used in the queries.  **(0-5 marks)** | Have 3-6 queries with very few operators used. One or two queries is using the OLAP operation.  **(6-12 marks)** | Have 7 - 10 queries with various operator used. Used multiple tables. Use at least 5 queries with OLAP operations. Queries able to extract desired information from the tables.  **(13 - 18 marks)** | Have more than 10 queries with various operators used.  Use of multiple tables. Use more than 5 queries with OLAP operations. Queries fits data well and easy to be interpret.  **(19 - 25 marks)** |  |
| **(I) TOTAL MARKS (90 marks)** | | | | |  |

| **Part E:** | **INDIVIDUAL ASSESSMENT** | | | **Student 1** | | **Student 2** | | **Student 3** | | **Student 4** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Presentation**  **(10 marks)** | Not clear communication. Ideas and information not communicated well and hampers understanding. Presentation is un-organized. Unable to answer most of the questions with low confidence level.  **(0-3 mark)** | Somewhat clear  Communication.  Ideas and information are  somewhat unclear and  hampers understanding.  Presentation is sufficiently  Organized. Able to answer questions with some confidence level  **(4-6 marks)** | Ideas and information communicated fairly clearly. Presentation is organized well. Able to answer questions with correct answer and high confident  **(7 - 10 marks)** |  | |  | |  | |  | |
| **Self-Contribution**  **report** | Few or lack of contribution made for the assignment. Have few or some knowledge on the describe tasks. | Describe the contribution made for the assignment. Have fair knowledge on the describe tasks. | Successfully describe the contribution made for the assignment. Highly knowledgeable on the describe tasks. |  | |  | |  | |  | |
| **(II) Total marks (10 marks)** | | | | |  | |  | |  | |  | |
| **(I) + (II) Total (100 marks)** | | | | |  | |  | |  | |  | |
| **Total Marks (30%)** | | | | |  | |  | |  | |  | |

| **Lecturer Comment:** |
| --- |

**Table of Contents**

[**1.0. INTRODUCTION**](#_mqikwp81yc6h) **8**

[1.1 Objectives of the Project](#_ojbvdgwl4a4t) 8

[1.2. Scope of the Project](#_txmem9l99nmu) 9

[1.3. Tools and Requirements](#_ojbvdgwl4a4t) 9

[**2.0. PART A: BUSINESS DESCRIPTION**](#_m2ebo3lw8u8f) **10**

[Business Description of the E-commerce Shopping System](#_dwti9zav9qfu) 10

[2.1. Business Rules:](#_h7fok439ohhu) 10

[2.2. Relationships and Cardinalities:](#_mzazo278xkai) 11

[2.3 Entity-Relationship Diagram (ERD)](#_kkh71lm2o773) 11

[**3.0. PART B: DATA WAREHOUSE DESIGN**](#_g8sl12x128zt) **12**

[3.1. Schema Type: Star Schema:](#_oqz8aw4i1nn) 12

[3.2. Fact and Dimension Tables](#_luf5qgtmxno5) 12

[3.3. Justification for Using Star Schema:](#_gsxkhkh0tlqb) 14

[**4.0. PART C: ETL (EXTRACT, TRANSFORM, AND LOAD) PROCESS**](#_74uaez9denzo) **16**

[4.1. DATA EXTRACTION (CSV File Input)](#_1121og1ak4t5) 17

[4.2. DATA CLEANSING](#_1ezhuy2xhlup) 17

[4.3. DATA TRANSFORMATION](#_yag4pml8x69g) 19

[4.4. Data Loading](#_jusvig2niz65) 26

[**5.0. PART D: DATA ANALYSIS USING SQL QUERIES**](#_qzl3u223p1zp) **30**

[**6.0. CONCLUSION**](#_irk2sfhxm8k1) **42**

[**TASK DISTRIBUTION TABLE**](#_r5f50t2j49zd) **43**

[**SELF-REFLECTIVE REPORT**](#_futzds9vkndm) **44**

[**References:**](#_t348deeytl1v) **46**

[**Appendix:**](#_36f9451e4tj5) **47**

[Appendix 1: 47](#_36f9451e4tj5)

[Appendix 2:](#_8kcevzkc9dg5) 48

# 1.0. INTRODUCTION

## At present time, the importance of data has increased in today’s dynamic business scenario, specifically in the e-commerce sector, which is growing rapidly within the e-commerce sector. Ability to gather, process, and analyze huge volumes of customer and transactional data well is pivotal in comprehending market trends, improving business performance and taking intelligent strategic decisions. The addressed gap in this project is the rising demand for appropriate data analysis within the e-commerce realm, specifically through providing the capability to transform the raw customer shopping data into meaningful business intelligence through a data warehouse solution.

## A case study centered around an e-commerce company in which this report provides a practical case of using these dataset to derive insights into its customer shopping data. As a team, our goal is to create, implement and analyze a dedicated data warehouse to the business intelligence needs of this company. With our data warehouse constructed well, we intend to create a one stop shop where the information is extracted from raw data and help the company in taking decisions based on the data throughout the various aspects of their business.

## 1.1 Objectives of the Project

## Understand the E-Commerce Business: Be able to fully understand the ways of shopping in e-commerce and the business rules that guide the operation of the system.

## Create an Entity Relationship Diagram (ER-D): An Entity Relationship Diagram is constructed using the business rules obtained from the tasks and exercises.

## Data Warehouse Schema Design: Design an appropriate data warehouse schema (star or snowflake) for analyzing the dataset based on which E-commerce ER is derived.

## Develop an ETL Pipeline: Design and implement an ETL pipeline to facilitate the loading of the data warehouse on MySQL database based on the CSV provided.

## Data Analysis using SQL: Outlining and executing SQL queries like OLAP functions to get answers to such business questions and also extract meaningful information from the data warehouse.

## Report the Entire Process: Include the report for the entire process, i.e. business rules, data models, ETL processes, SQL queries as well as the generated results.

## Discuss the Work: Outline the ideas and implementation of the project, the methodology used, the key findings and other work behind the project during a formal presentation.

## 1.2. Scope of the Project

## The scope of this project encompasses the following key areas:

## Data Source: The source of the provided customer\_shopping\_data.csv file is an e-commerce company.

## Data Modeling:Both ER-D for the relational database model and star/snowflake dimensional data warehouse schema design problems.

## ETL Implementation: Building an entire ETL pipeline through the Pentaho Data Integration (PDI) for 2 rows to convert the data from the CSV file to the chosen data warehouse schema and finally to load in the MySQL database.

## Target Database: The implementation of the data warehouse will be made using the MySQL database environment.

## Data Analysis: Developing SQL formulations that can answer specific business questions pertaining to the e-commerce data using OLAP functions and execute it.

## Reporting and Presentation: Writing a comprehensive report documenting all phases of the project and presenting the results to show that the team has a good understanding and analytical capabilities.

## Team Collaboration: Finishing all aspects of the assignment working in a group of no greater than four students.

## 1.3. Tools and Requirements

## The following tools and requirements were utilized for the completion of this project:

## Data: customer\_shopping\_data.csv (provided e-commerce customer shopping data).

## Data Modeling Tools: A suitable tool for creating Entity-Relationship Diagrams (e.g., draw.io, Lucidchart) and data warehouse schema diagrams.

## ETL Tool: Pentaho Data Integration (PDI) for designing and executing the ETL pipeline.

## Database Management System: MySQL for hosting the data warehouse.

## SQL Client: MySQL Workbench

# 2.0. PART A: BUSINESS DESCRIPTION

## Business Description of the E-commerce Shopping System

An e-commerce shopping system facilitates online deals of products between customers and the business they buy the products from. There are product categories, products with price and product name, displaying the product in the virtual shopping cart, and the possibility of going shopping. Payment can be done via multiple methods, each order with the details of concern including invoice number, company information (ID, gender, age), purchasing items (category, quantity), payment method, invoice date, and the mall of shopping lead (i.e. indicating the physical store or online platform). Its intention is to provide the customers with a user-friendly and practical shopping on the internet, while helping the business be able to oversee the sales, track customer behavior, and studies on purchasing habits from various types of products, paying methods and store locations.

## 2.1. Business Rules:

* Each transaction has its own invoice\_no for a given customer.
* There is only one registered customer\_id per each order.
* Each product has a specific quantity and it is possible to have one or more products in the same invoice.
* There are only products in one category.
* Each shopping\_mall has only one recognized by it.
* A single payment\_method is used in every transaction (invoice).
* There is a specific invoice\_date that orders are recorded.
* It has distinct customer\_id and personal demographic information (age, gender).
* It has different potential shopping\_mall locations (physical or online) where the organization operates.
* We can accept the customers’ payment via different modes, as specified by the payment\_method.

## 2.2. Relationships and Cardinalities:

1. A Customer can have one or many Invoices. (One-to-Many relationship)
2. Each Customer has only one Invoice. (Many-to-One relationship)
3. Exactly one Shopping Mall is linked with each Invoice. (Invoice is related to one Shopping Mall but one Shopping Mall can have multiple Invoices - one to many from Shopping Mall to Invoice).
4. A Product has an Invoice one or more. (The Invoice and Product are a many-to-many relationship, so we use the associative entity named 'Order Item' to resolve this issue.)
5. A Product can be included in any number of Invoices. (Product and Invoice (many to many) relationship, solved by an associative entity Order Item).
6. There’s one to many relationship between Product and Category. (Many-to-One relationship from Product to Category).
7. Products belong to the Category and a Category may have zero or many Products. (One-to-Many relationship from Category to Product).

## 2.3 Entity-Relationship Diagram (ERD)

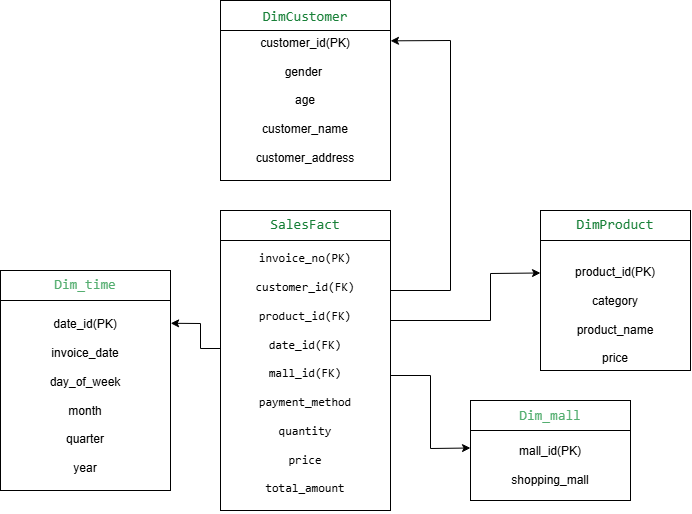
****

*Figure 1: Entity-Relationship Diagram (ERD)*

The ER diagram shows the core data elements of the e-commerce shopping system presented here. It illustrates the respective features of Customer (have such personal info),Product (belonging to them defined Category), the transaction part and their Invoices which are tied to the particular Shopping Mall and also to those order\_item details for each purchase.

# 3.0. PART B: DATA WAREHOUSE DESIGN

## 3.1. Schema Type: Star Schema:

****

*Figure-2: Star Schema*

SalesFact is the central table with respect to the schema, representing a type of sales event. It is surrounded by dimension tables describing customers, products, as well as transaction time and the shopping mall location for multi dimensional analysis.

## 

## 

## 

## 3.2. Fact and Dimension Tables

Based on the customer\_shopping\_data.csv dataset and the business rules, we have identified the following fact and dimension tables for our data warehouse:

**Fact Table:**

* **Sales Transaction Fact Table:** The system maintains this central table to track each specific purchase process.

**Dimension Tables:**

* **Customer Dimension:** The table contains descriptive customer-related data.
* **Product Dimension:** The table includes information about the products which will be sold.
* **Time Dimension:** This table will provide temporal context for the transactions.
* **Shopping Mall Dimension:** The table contains data for the shopping locations.

**Product Dimension:**

* **Primary Key:** product\_id (Product Identifier) - A unique identifier for each product (to be generated during ETL).
* **Attribute:** category (Product Category) - The broad category to which the product belongs (e.g., Clothing, Electronics).
* **Attribute:** product\_name (Product Name) - The specific name of the product (to be added during ETL).
* **Attribute:** price (Product Selling Price) - The price of a single unit of the product at the time of the data extraction.

**Time Dimension:**

* **Primary Key:** date\_id (Date Identifier) - A unique identifier for each distinct date (to be generated during ETL).
* **Attribute:** invoice\_date (Transaction Date) - The specific date on which the transaction occurred.
* **Attribute:** day\_of\_week (Day of the Week) - The day of the week for the transaction date.
* **Attribute:** month (Month of the Year) - The month of the year for the transaction date.
* **Attribute:** quarter (Quarter of the Year) - The quarter of the year for the transaction date.
* **Attribute:** year (Year of the Transaction) - The year of the transaction date.

**Shopping Mall Dimension:**

* **Primary Key:** mall\_id (Shopping Mall Identifier) - A unique identifier for each shopping mall (to be generated during ETL).
* **Attribute:** shopping\_mall (Shopping Mall Name) - The name of the shopping mall where the transaction took place.

## 3.3. Justification for Using Star Schema:

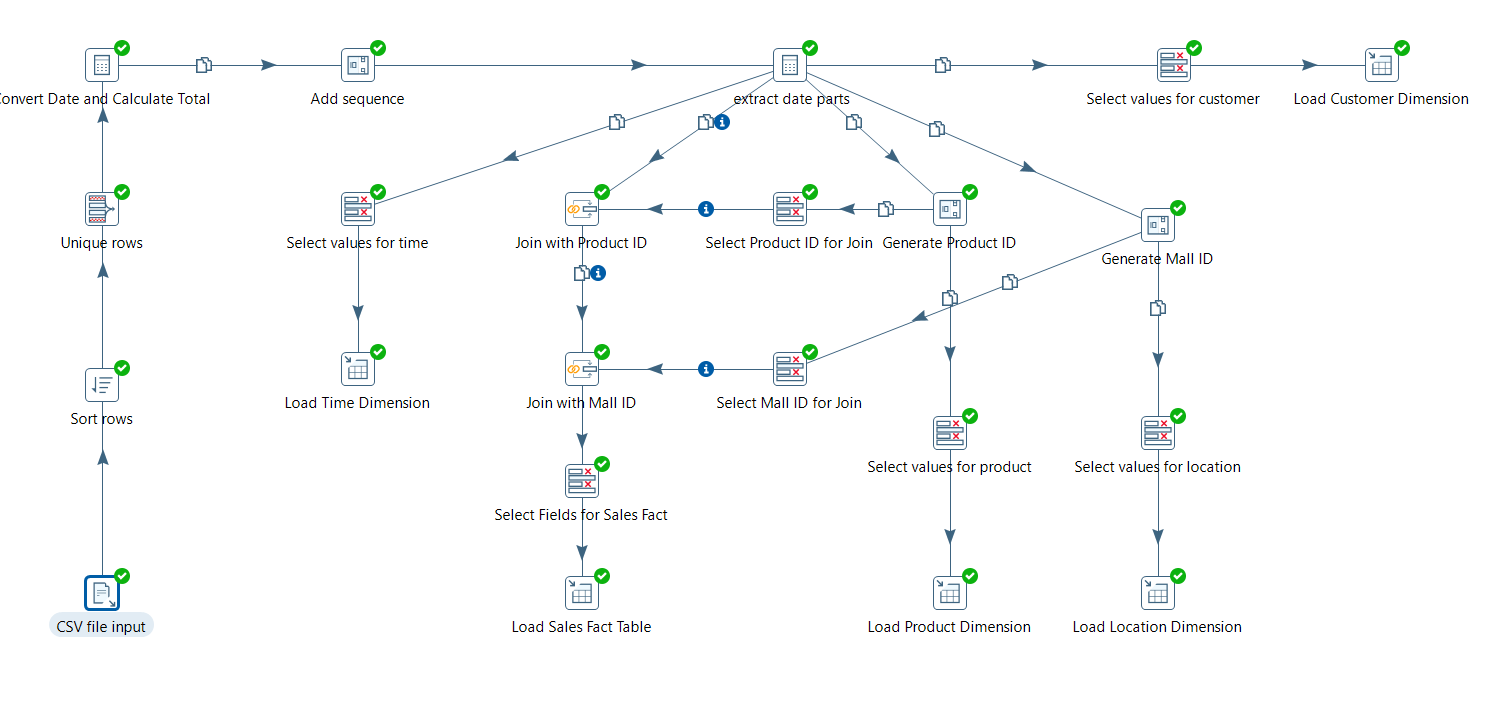
1. **Simplicity and Ease of Understanding:**Basic and intuitive star schema, a star schema based on the central fact and dimension tables. This is simpler for the business users to understand the data model and formulate queries for analysis. With the objective of the project to equip the e-commerce company to make data driven decisions, a model as a whole should be easy to understand.
2. **Optimized for Query Performance:** Denormalization occurs in star schemas in that the data can be redundant in the dimension tables. This denormalization however, allows fewer joins in the queries on the data. The fewer the joins, the faster the query execution time, which is important to query execution time for interactive business intelligence and timely reports of sales trends, customer behavior, and product performance.
3. **Direct Mapping to Business Dimensions:** Star schema’s dimension tables directly represent the key business entities (Customers, Products, Time, and Shopping Malls) that e-commerce companies are interested to analyse on its sales data. This direct mapping reduces the hassle of slicing and dicing of sales data with these dimensions so as to provide insight such as:  
   * Our analysis informs about commercial activity through customer population groups according to age and gender divisions.
   * Sales performance of different product categories.
   * Sales performances covering daily, monthly and yearly as well as quarterly and daily patterns.
   * Sales performance across different shopping mall locations.
4. **Clear Separation of Facts and Context:** By using the star schema, you can see clearly the quantitative data (measures: quantity and total\_amount in SalesFact table) is separated away from the descriptive context (attributes in dimension tables). By separating out what is being measured and the views through which these can be viewed, the measurement is clear and easy to distinguish what is being measured.
5. **Suitability for OLAP (Online Analytical Processing):** The Star schema exists as a database design specifically made to handle operations of OLAP because it enables analytical queries involving numerous dimensional aggregations and comparisons. The design matches exactly with the initiative to transform database information into practical business intelligence tools.
6. **Conformed Dimensions for Consistency:** Consistency between different analyses becomes possible because Dim\_time and Dim\_mall work as conformed dimensions. Through dimensional definitions the company obtains simultaneous access to time-based sales patterns and shopping mall performance analysis across consistent parameters of time and mall.

# 4.0. PART C: ETL (EXTRACT, TRANSFORM, AND LOAD) PROCESS

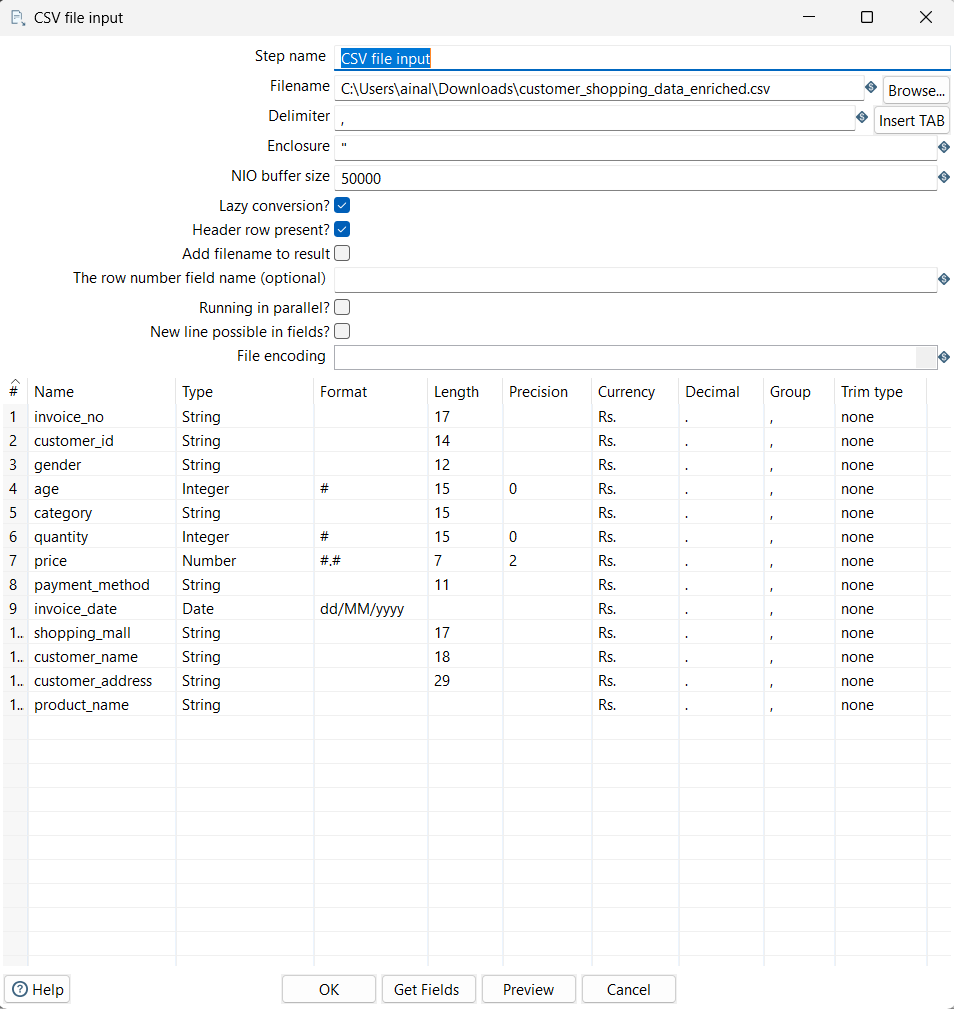
**Introduction to the ETL Pipeline**

In this section, we are loading the data from the customer\_shopping\_data.csv file into our data warehouse schema we want to use in MySQL using Pentaho Data Integration. It is data extracted, transformed (transformed to match the star schema) and loaded to the respective dimension and fact tables. In relation to this, before the execution of this ETL process, the raw customer\_shopping\_data.csv file had been improved with Python scripts in Google Colab to bring three additional attributes: customer\_name, customer\_address and product\_name. They are more contextual information to our analysis.

**Complete Pentaho Transformation Setup**

Below is the visualization of the whole Pentaho transformation designed for this project, shown above. This shows the flow of data from the first CSV input through the outlined cleaning, forming, loading steps in order to clean, shape, and load data into the dimension and fact tables of the e-commerce data warehouse.

## 4.1. DATA EXTRACTION (CSV File Input)

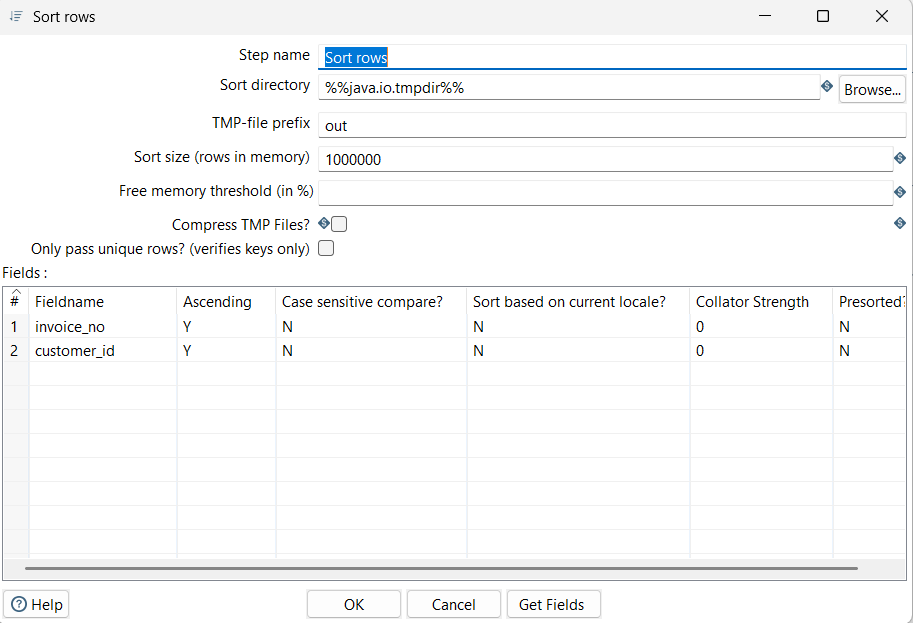
The ETL process begins with the "CSV file input" step. This step is configured to read the customer\_shopping\_data.csv file, which now includes the generated customer\_name, customer\_address, and product\_name columns alongside the original data. The configuration specifies details such as the file path, field separator, text enclosure, and whether the file contains a header row.

## 4.2. DATA CLEANSING

The next phase involves data cleansing to ensure data quality and consistency. Our pipeline incorporates the following cleansing steps:

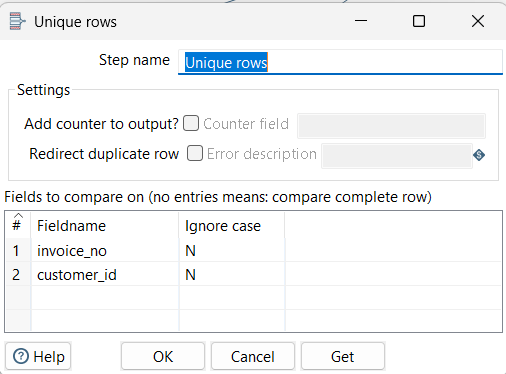
**4.2.1. Sorting Rows:**

The "Sort rows" step is used to order the incoming data based on specific fields. This sorting facilitates the subsequent duplicate removal process.

****

**4.2.2. Removing Duplicate Rows:**

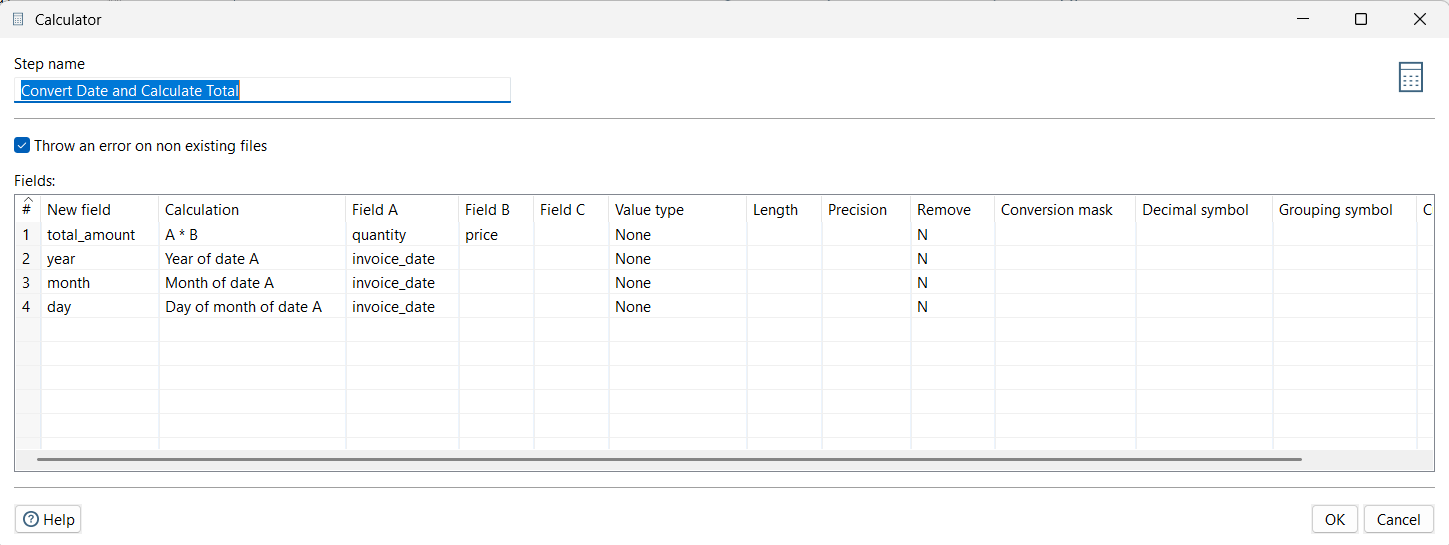
The "Unique rows" step processes the sorted data and removes any duplicate records based on all the input fields. This ensures that only unique transactions are processed further.

****

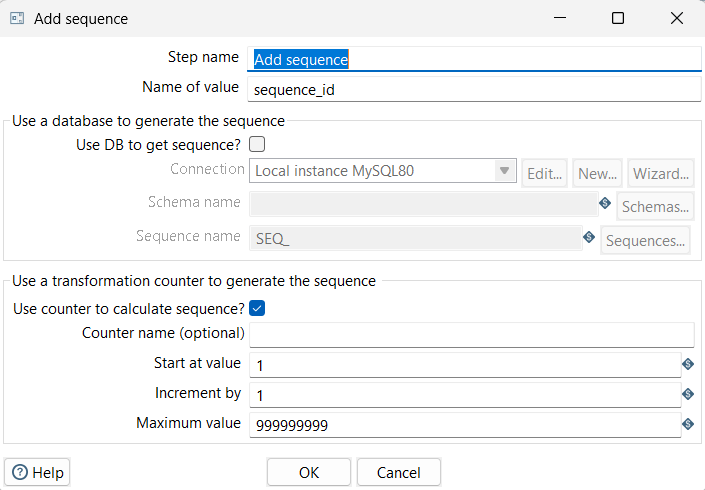
## 4.3. DATA TRANSFORMATION

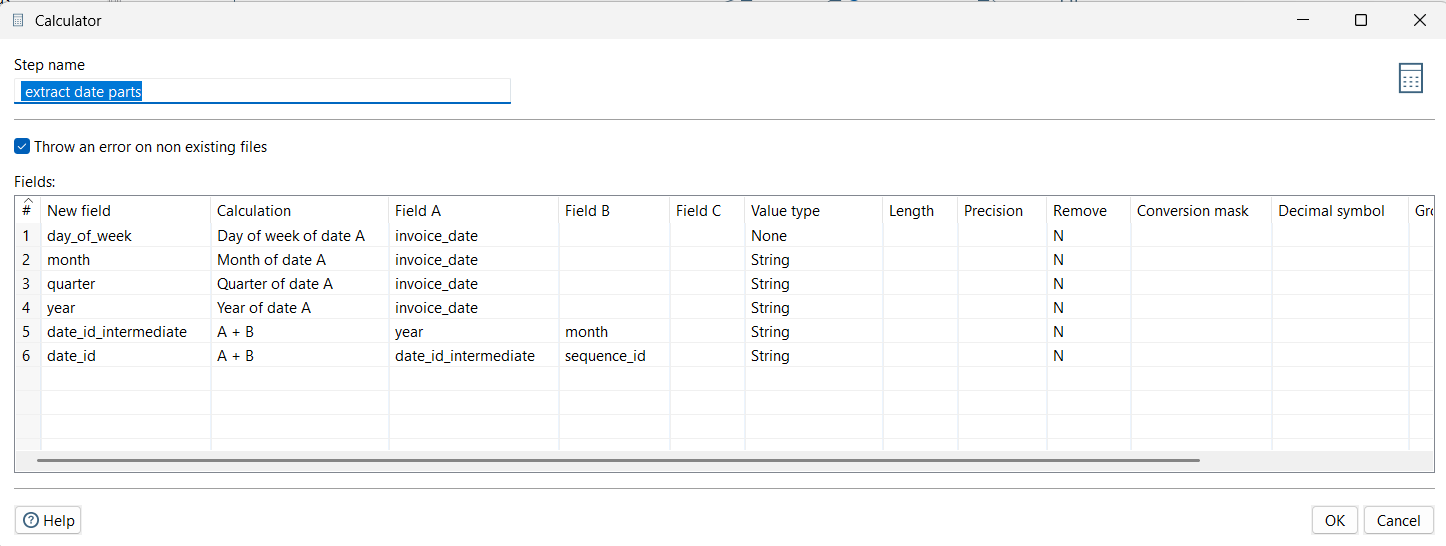
This phase aligns the extracted and the cleansed data to our star schema design. There are several steps involved in making the end keys and attributes for our dimension and fact tables.

**4.3.1. Time Dimension Population:**

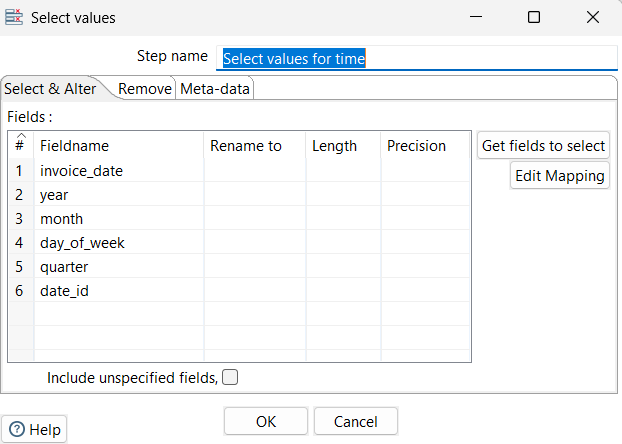
**4.3.1.1. Convert Date and Calculate Total:** This step likely involves converting the invoice date/time field into a consistent date format suitable for the Time Dimension. It also includes the calculation of total\_amount by multiplying quantity and price.

**4.3.1.2. Add Sequence:** The "Add sequence" step generates a unique sequential integer for each distinct date, which serves as the date\_id (primary key) for the Time Dimension.

**

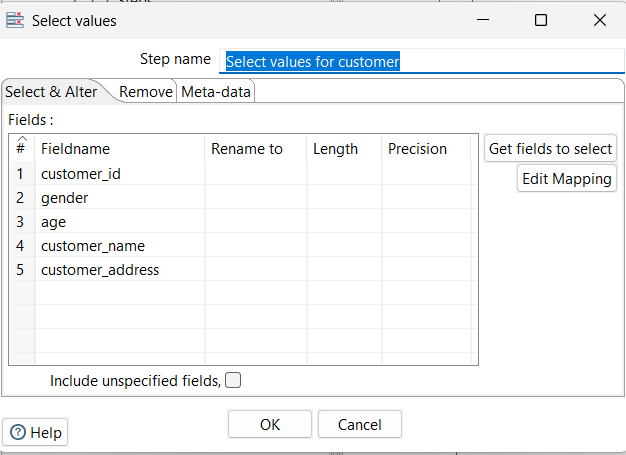
**4.3.1.3. Extract Date Parts:** The "Extract date parts" step takes the invoice date and separates it into its constituent parts: day\_of\_week, month, quarter, and year. These become attributes in the Time Dimension.

**4.3.1.4. Select values for time:** This "Select values" step chooses and potentially renames the necessary fields (date\_id, invoice\_date, day\_of\_week, month, quarter, year) for loading into the Time Dimension.

**

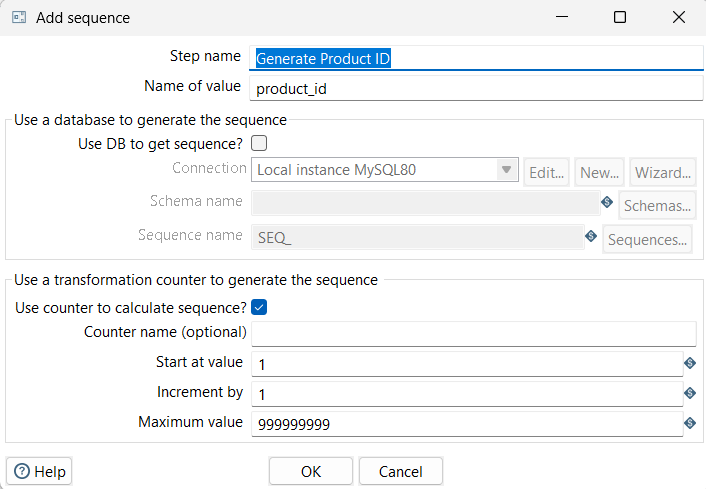
**4.3.2. Customer Dimension Population:**

**4.3.2.1. Select values for customer:** This "Select values" step selects the relevant customer attributes (customer\_id, gender, age, customer\_name, customer\_address) for loading into the Customer Dimension.

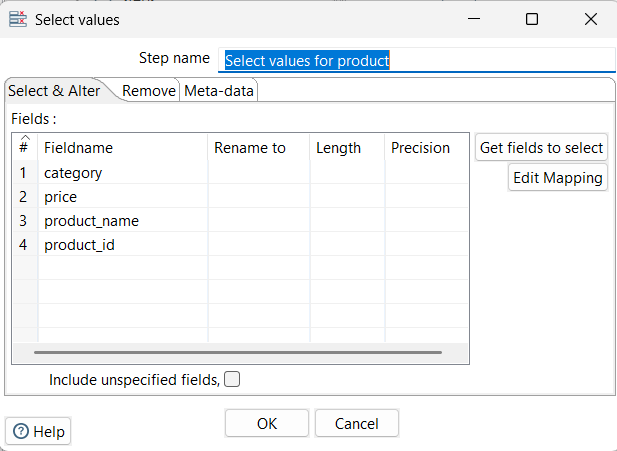
**

**4.3.3. Product Dimension Population:**

**4.3.3.1. Generate Product ID:** The "Generate Product ID" step creates a unique product\_id (primary key) for each unique product.

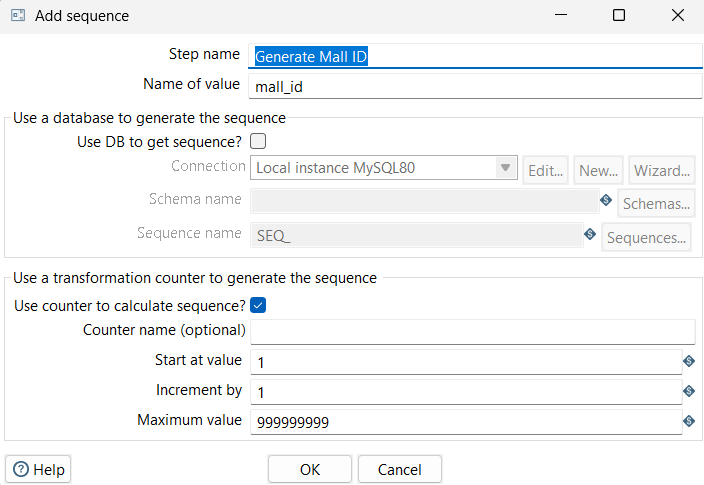
**

**4.3.3.2. Select values for product:** This "Select values" step selects the necessary product attributes (product\_id, category, product\_name, price) for loading into the Product Dimension.

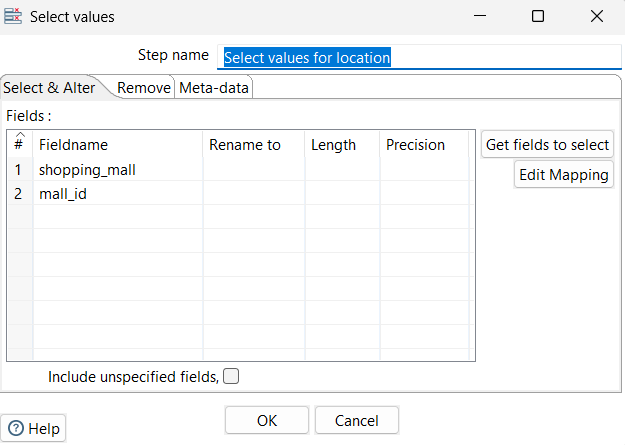
**

**4.3.4. Location Dimension Population:**

**4.3.4.1. Generate Mall ID:** The "Generate Mall ID" step creates a unique mall\_id (primary key) for each unique shopping mall.

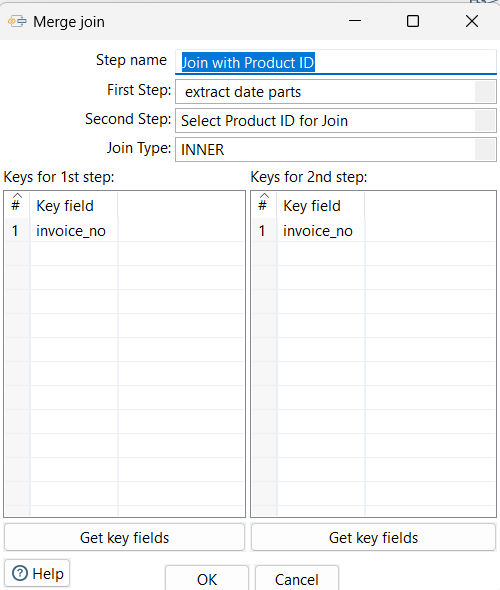
**

**4.3.4.2. Select values for location:** This "Select values" step selects the relevant location attributes (mall\_id, shopping\_mall) for loading into the Location Dimension.

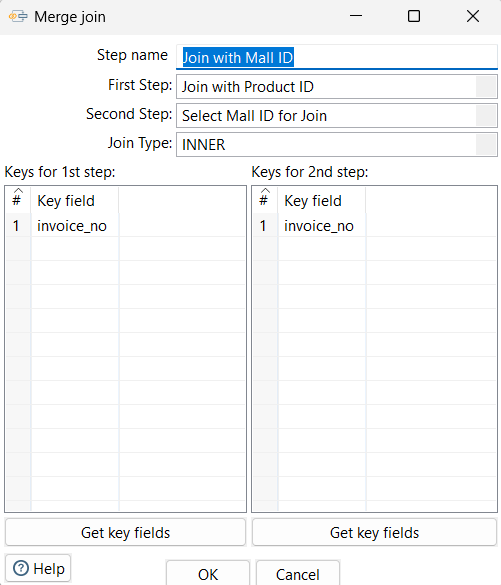
**

**4.3.5. Sales Fact Table Population:**

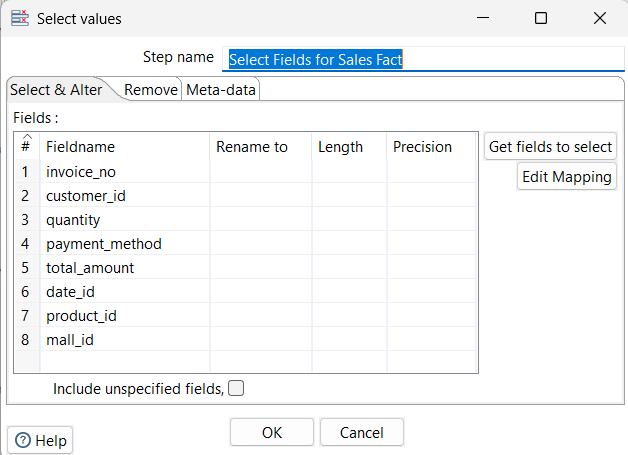
**4.3.5.1. Join with Product ID:** This "Join" step performs an inner join between the main transaction stream and a stream containing invoice\_no and product\_id (likely selected earlier) to retrieve the generated product\_id surrogate key.

**

**4.3.5.2. Join with Mall ID:** An additional inner join operation occurs in the "Join" step to read the invoice\_no and mall\_id values which were initially selected for retrieving the surrogate key from the mall\_id table.

**

**4.3.5.3. Select Fields for Sales fact:** This "Select values" step chooses the necessary fields for the Sales\_Transaction\_Fact table: invoice\_no, customer\_id, product\_id (from the join), date\_id (from the Time Dimension branch), mall\_id (from the join), quantity, payment\_method, and total\_amount.

**

## 

## 

## 

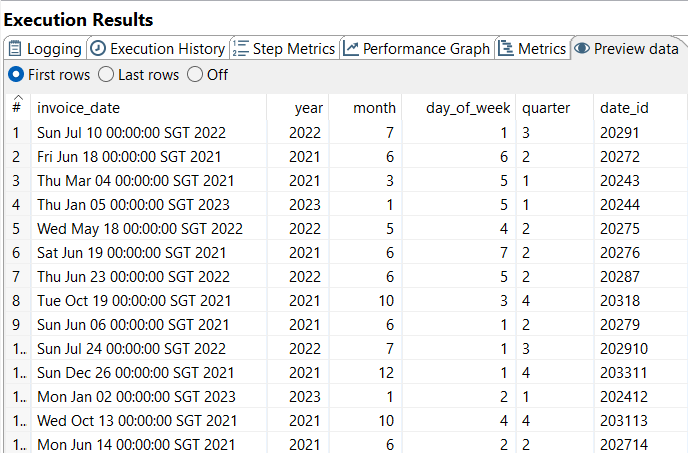
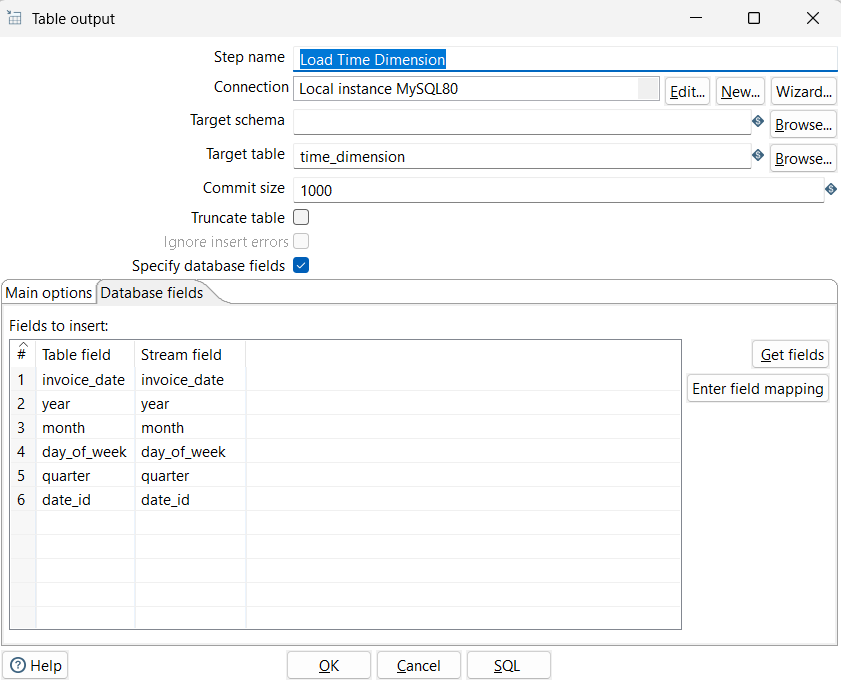
## 

## 

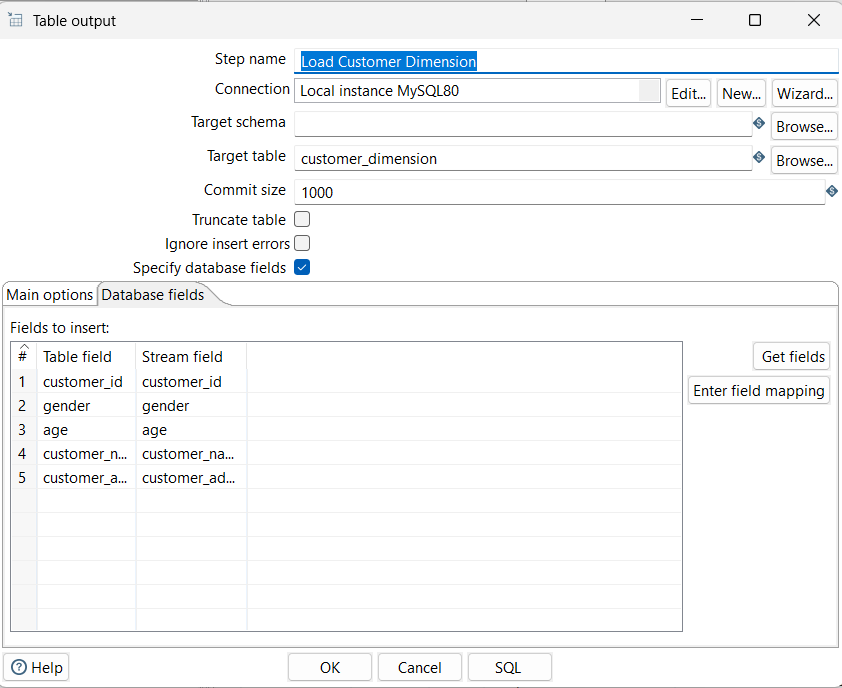
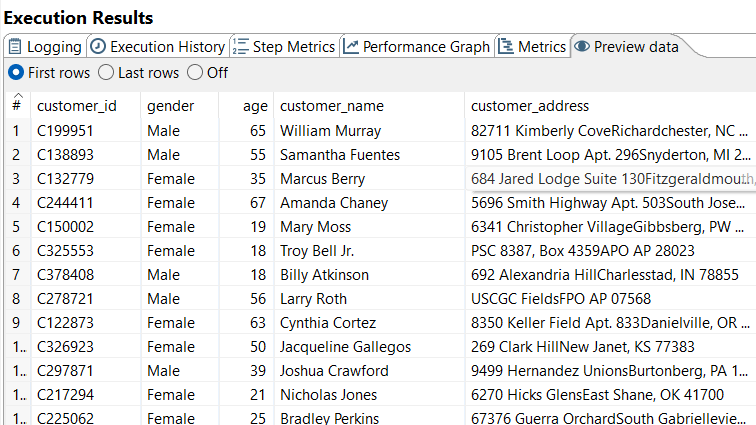
## 4.4. Data Loading

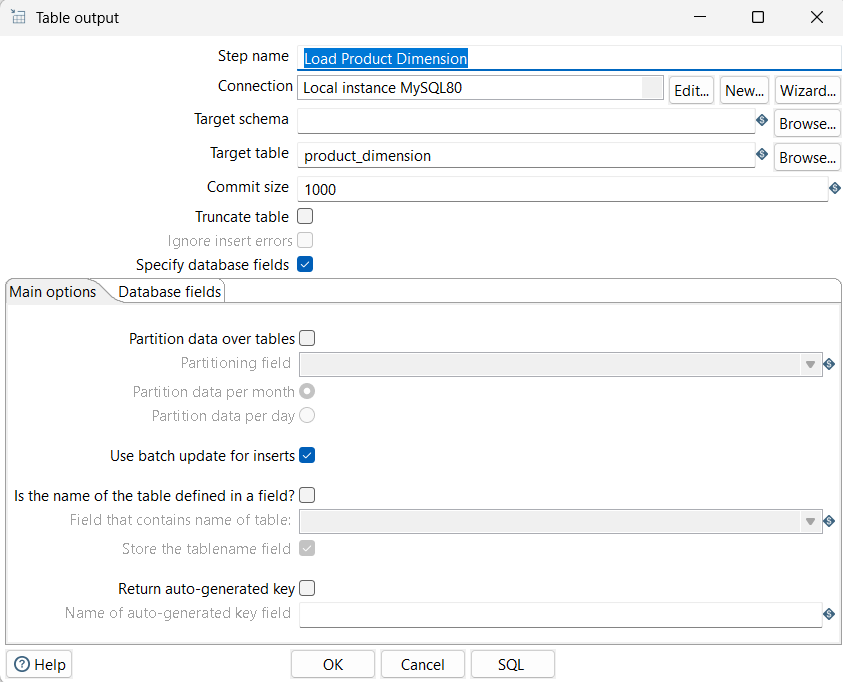
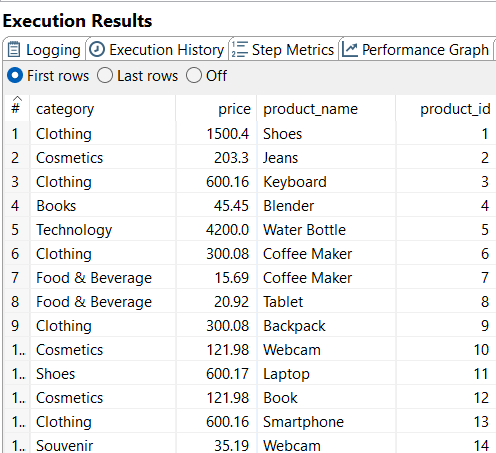
The final phase involves loading the transformed data into the respective tables in the MySQL database.

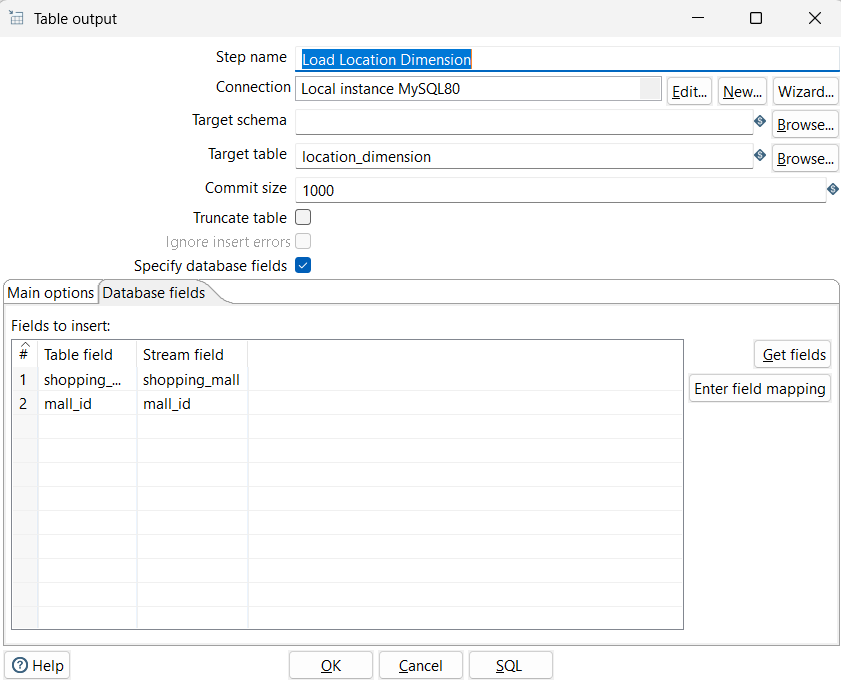
**4.4.1. Load Time Dimension:** The "Load Time Dimension" step is configured to write data into the Time\_Dimension table in MySQL.

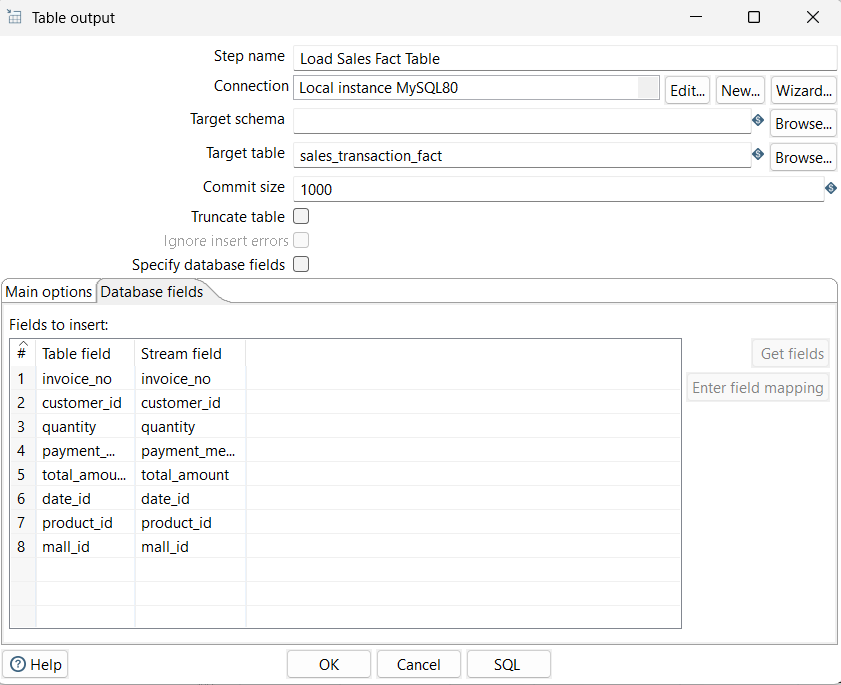


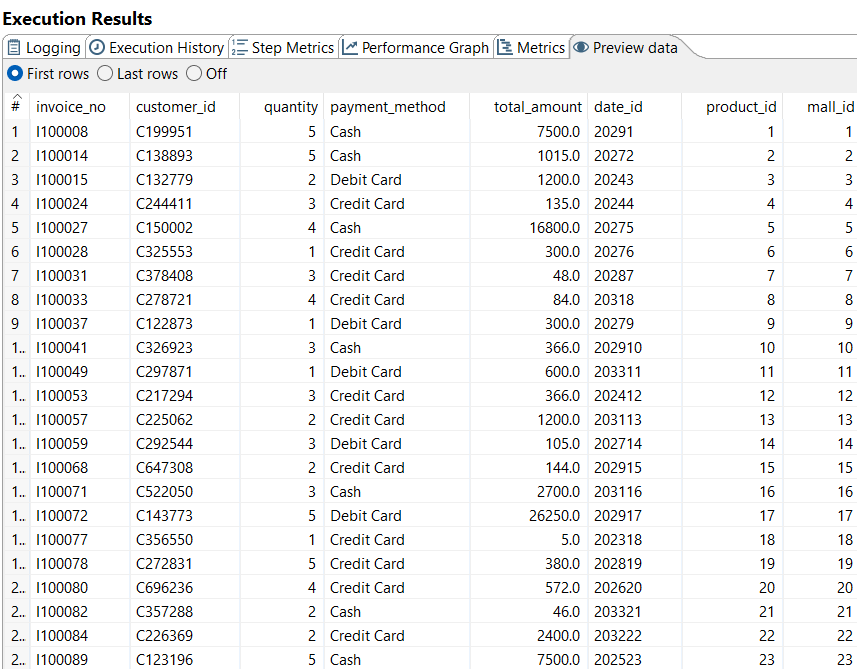
**4.4.2. Load Customer Dimension:** Similarly, the "Load Customer Dimension" step writes data to the Customer\_Dimension table in MySQL.



**4.4.3. Load Product Dimension:** The "Load Product Dimension" step writes data to the Product\_Dimension table in MySQL.

**4.4.4. Load Location Dimension:** The "Load Location Dimension" step writes data to the Location\_Dimension table in MySQL.

**4.4.5. Load Sales Fact Table:** The "Load Sales Fact Table" step writes data to the Sales\_Transaction\_Fact table in MySQL.

**FINAL SALES\_TRANSACTION\_FACT PREVIEW**

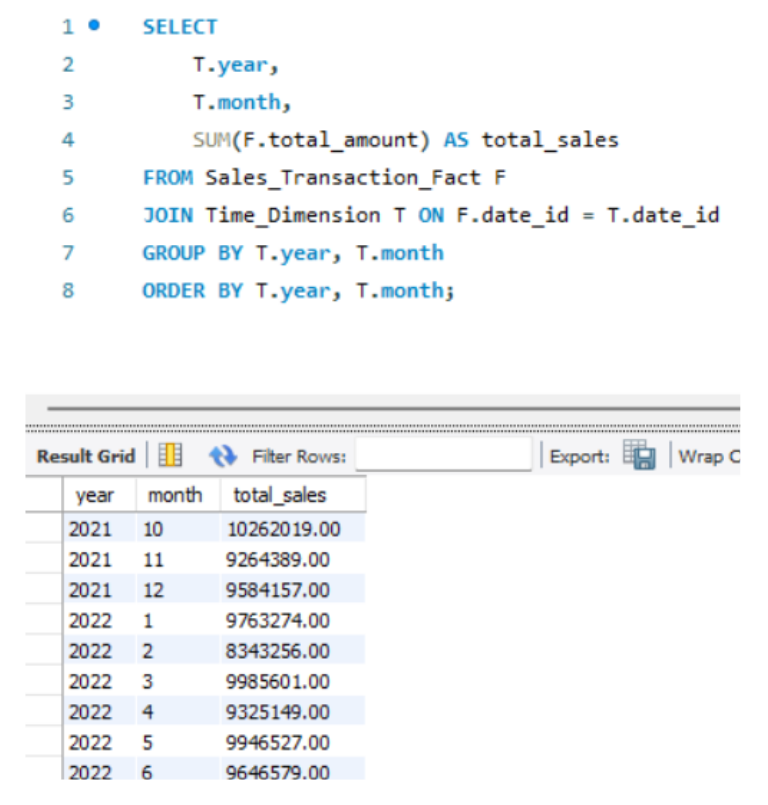
# 

# 5.0. PART D: DATA ANALYSIS USING SQL QUERIES

Through this section readers will discover how SQL queries process e-commerce information to produce valuable data about sales patterns and customer conduct together with product effectiveness in different measurement areas.

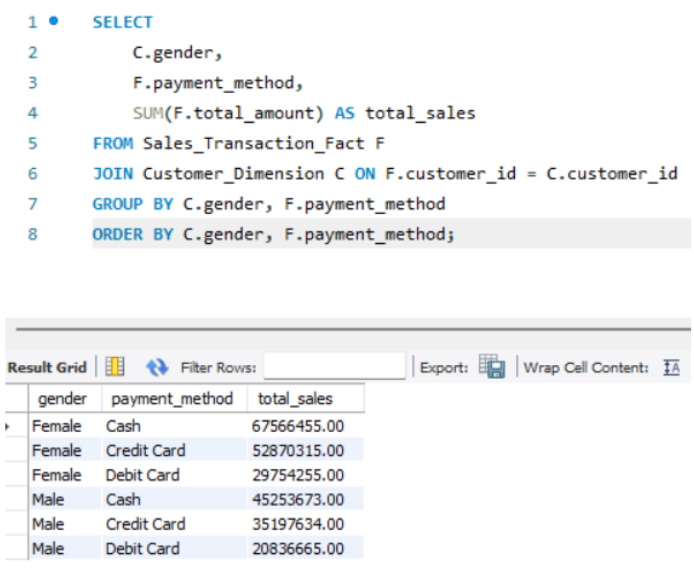
**Queries by Ashish:**

**Query 1**: Total Sales by Month and Year (using GROUP BY and ordering):

**Business Need:** Understand monthly sales trends for better resource allocation and forecasting.  
  


**Explanation:** Sales data is joined with time dimension, grouped on year and month and grouped by sum of the total sales per month. The output provides the total revenue obtained in each specific month in each specific year.

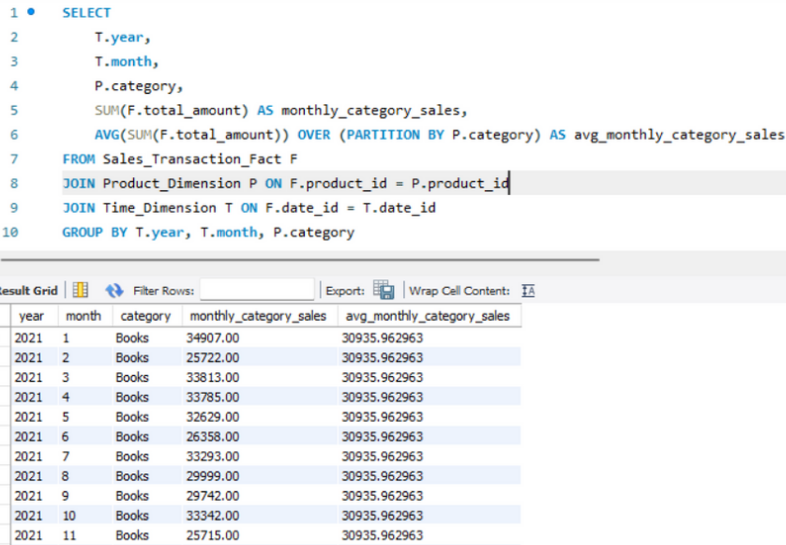
**Query 2:** Total Sales by Customer Gender and Payment Method (using GROUP BY for basic aggregation)



**Business Need:** Identify which customer genders prefer specific payment methods and their corresponding total spending

**Explanation:** By joining the customer dimension and sales data, it queries the group of total sales by the customer’s gender and the payment method used in the transaction. The result is the revenue obtained by each gender per distinct payment method.

**Query 3:** Average Monthly Sales per Product Category (using AVG with OVER):

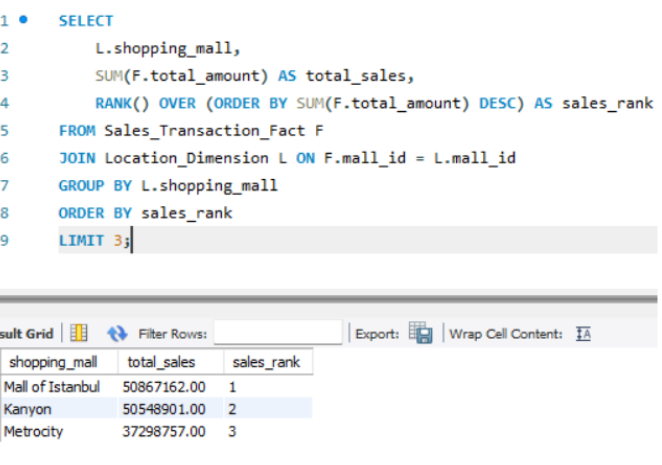


**Business Need:** Understand the monthly sales performance of different product categories and identify how each month's sales compare to the category's overall average monthly performance.

**Explanation of Answer:** The query calculates the total sales per month for each category and also the average monthly sales for each category overall, allowing for a direct comparison of monthly performance against the average.

**Queries by Bhulakshmi:**

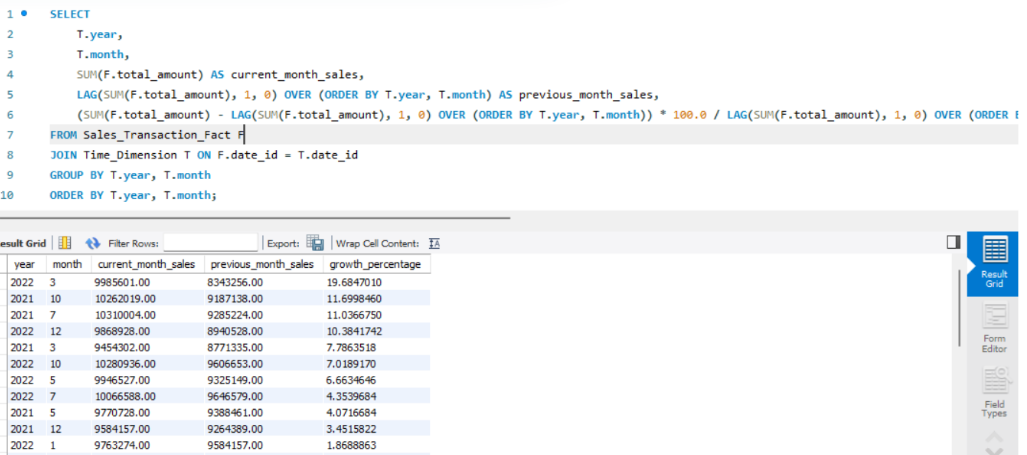
**Query 4:** Top 3 Shopping Malls by Total Sales (using RANK and LIMIT):



**Business Need:** Research the shopping centers which bring in the most revenue because they reveal vital sales factors and potential investment prospects.

**Explanation of Answer:** The query demonstrates total sales calculations for each mall followed by sales ranking and lastly displays only the three highest-selling malls.

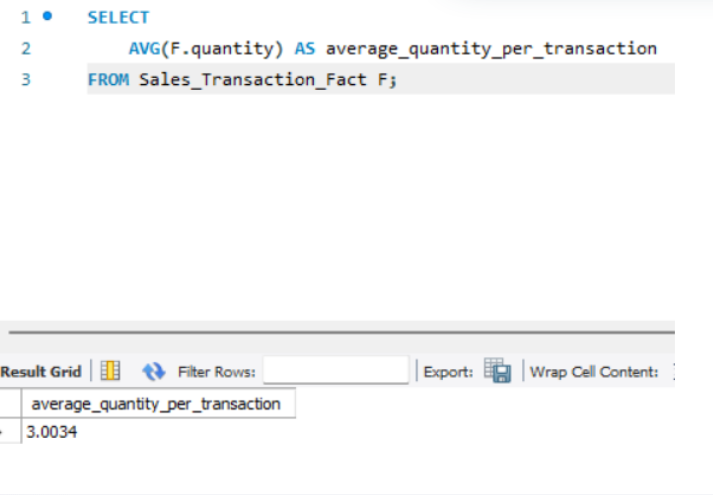
**Query 5**: Month-over-Month Sales Growth (using LAG):



**Business Need:** A tracking system should monitor sales figures from previous months in order to observe performance changes and detect developing issues.

**Explanation of Answer:** The report displays the overall monthly sales amounts which are matched with prior month figures alongside their sales percentage comparison.

**Query 6:** Average Quantity Purchased per Transaction:

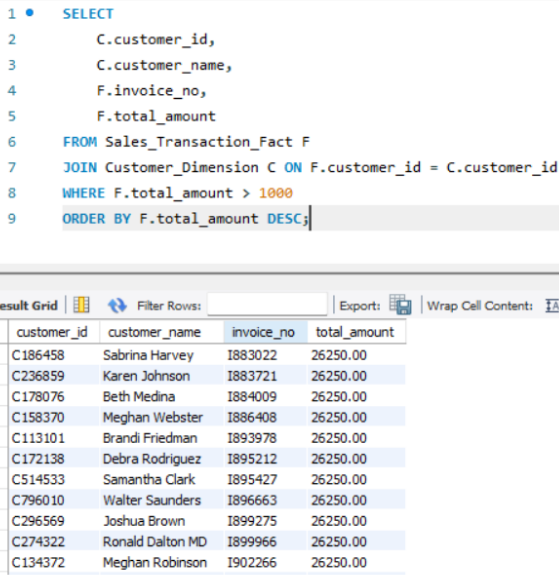


**Business Need:** A business needs to understand what number of products customers normally purchase during each transaction.

**Explanation of Answer:** This query uses the Sales\_Transaction\_Fact table to determine an average value of the quantity column for total product purchase frequency per transaction.

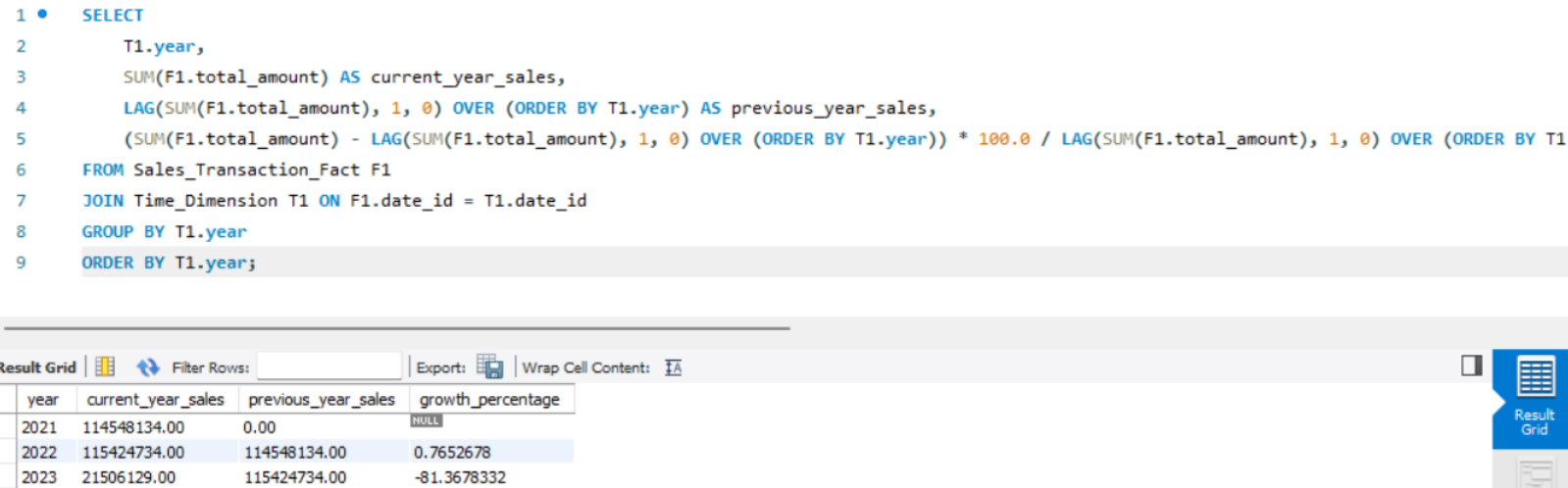
**Queries by Christo Tonio:**

**Query 7:** Customers with Transactions Exceeding $1000:



**Business Need:** Potential targeted marketing and customer relationship management activities need to focus on high-value customers with their major transactions.

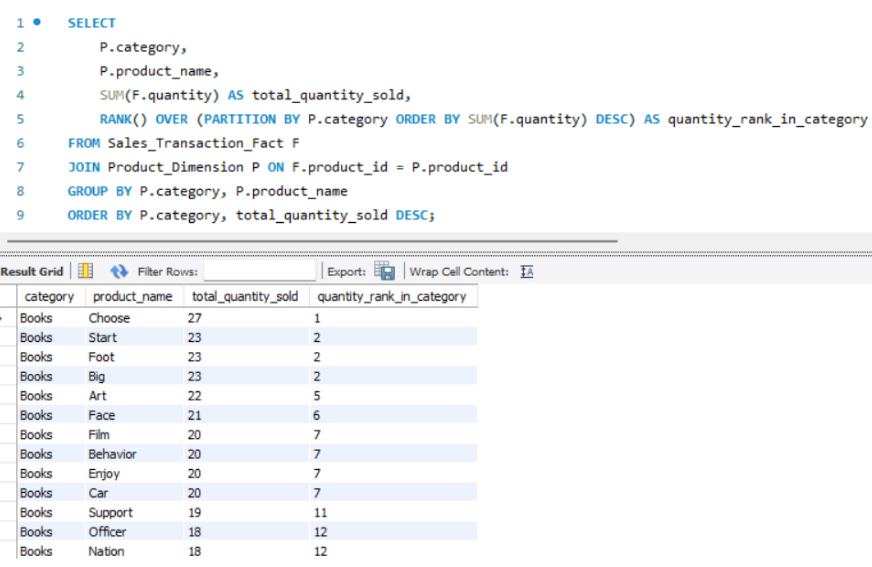
**Explanation of Answer:** The query joins sales transactions to customer information then displays customers with transactions larger than 1000 total amount while showing IDs, names, invoice numbers and total amounts in descending purchase value order among those high-value transactions.

**Query 8:** Year-over-Year Total Sales Comparison:  
  


**Business Need:** An evaluation of yearly sales data enables the surveillance of long-term development patterns and the detection of substantial changes.

**Explanation of Answer:** The query determines the yearly sales total at current\_year\_sales. The LAG() window function retrieves preceding year sales data which is stored in previous\_year\_sales. The query determines the growth\_percentage by calculating the annual sales evolution.

**Query 9:** Rank Products by Total Quantity Sold within Each Category:

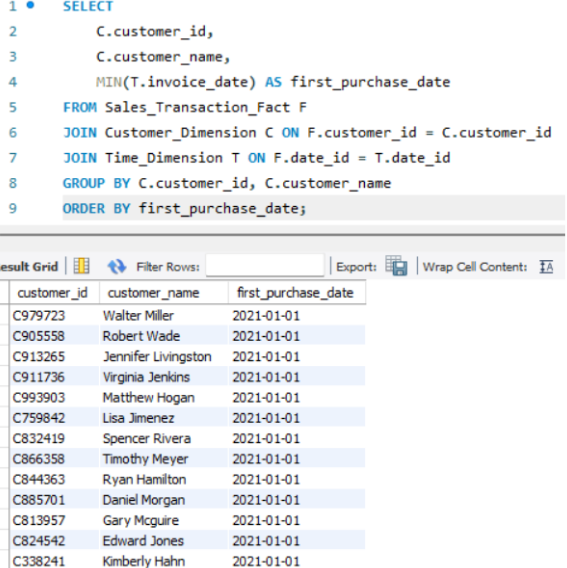


**Business Need:** The best-selling products of each category should be identified to gain knowledge about popular items which helps in both inventory planning and marketing campaigns.

**Explanation of Answer:**It joins sales data with product information and groups all the total quantities sold by product category and product name. Then it will use RANK() window function with P.category as partition and sorting on sum of F.quantity in descending order, to rank each product in its respective category according to the total quantity sold.

**Queries by Parvendan**

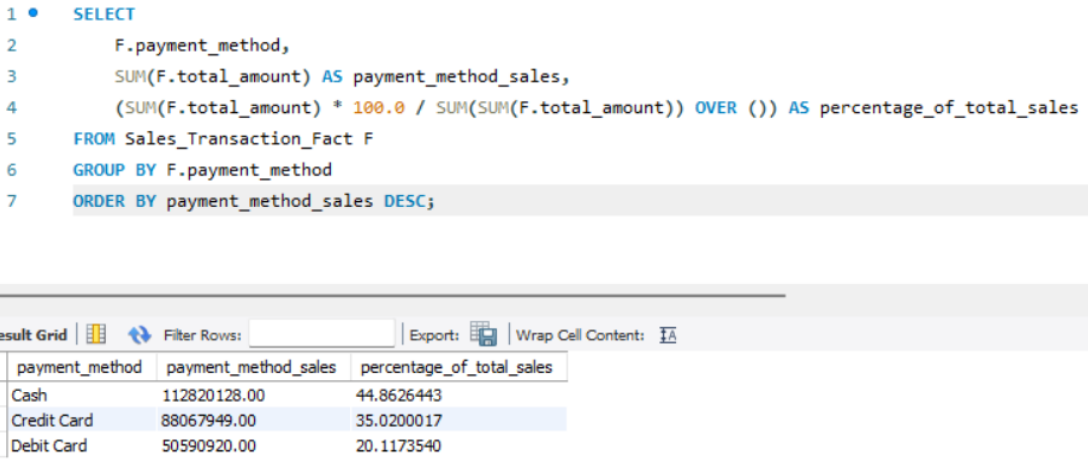
**Query 10**: Find the First Purchase Date for Each Customer:



**Business Need:** Currently exploring the date of acquisition across customers because it can help understand the customer acquisition timeline or discover new segments from joining periods.

**Explanation of Answer:** The query combines sales information with customer and time data to generate customer ID and name-based groupings while fetching the earliest purchase dates through MIN() aggregation on T.invoice\_date. The query returns results that include each customer ID with their name and their earliest purchase date according to the first purchase date order.

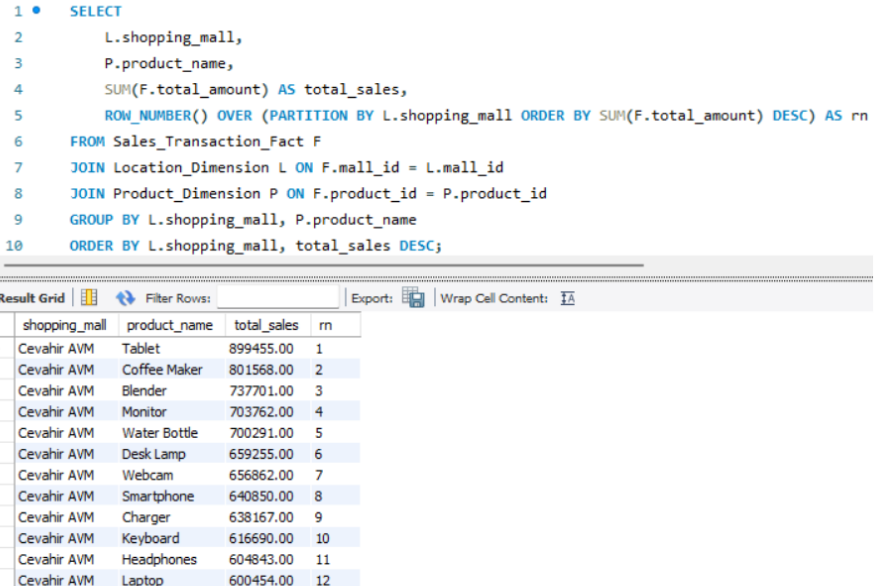
**Query 11**: Calculate the Percentage of Total Sales for Each Payment Method:



**Business Need:** Know what payment method makes up the largest part of the overall revenue, so as to determine if a change in payment gateway could be beneficial as well as what the customers prefer.

**Explanation of Answer:** It displays the total sales of each payment type and how much of all sales are payments by that payment type. The list is ordered by most used payment methods.

**Query 12**: Identify the Top Selling Product in Each Shopping Mall:



**Business Need:** Take advantage of what is inside each mall to find out what sells best to focus inventory and marketing efforts for each individual mall.

**Explanation of Answer:** Total sales of each product in each shopping mall is calculated using the query. Then it ranks products in a given mall according to the total sales, the top selling product is ranked first.

# 6.0. CONCLUSION

The following project shows that a data warehouse can be created for an ecommerce business’ analytical needs based on raw transactional data and provide structured business insights. The first step was to obtain a complete understanding of business domain and rules, then designing the Entity Relationship Diagram (ERD), then the Star Schema Data Warehouse model.

**The core achievements include:**

Schema Design: In this part, a sales fact table scoped as a central representative of the whole multidimensional analysis is implemented as a central sales fact table combined with four major dimension tables: customer, product, time and shopping mall, which support the multidimensional analysis.

ETL Pipeline: A good ETL process was built to use PDI for extracting, cleansing, transforming, and loading the enhanced shopping data (including contextual fields such as customer\_name, product\_name, etc.) into a MySQL data warehouse.

SQL Analysis: The structured data supports OLAP style queries for the team to answer business questions related to sales trends, customer behavior, product performance etc.

Business Intelligence Ready: It serves as a foundation for the generation of reports and dashboards that can have an impact on strategic decision-making for the enterprise at the level of the departments like marketing, operations and finance.

# TASK DISTRIBUTION TABLE

| **Member Name** | **Part of Assignment** | **Specific Tasks** |
| --- | --- | --- |
| ASHISH | Part A: Business Description & Traditional Database Model (ER-D) | Conducted research on the e-commerce shopping system. Wrote the business description and identified all relevant business rules. Based on the business rules, constructed the Entity-Relationship Diagram (ER-D) to represent the traditional database model. |
| BHULAKSHMI | Part B: Data Warehouse Design (Star/Snowflake Schema) | Designed the star or snowflake schema for the e-commerce dataset, ensuring dimensions were conformed and keys were clearly labeled. Identified the fact and dimension tables based on the dataset and ER-D. Defined appropriate measures and attributes. Explained and justified the chosen schema. |
| CHRISTO | Part C: ETL (Extract, Transform, Load) Process | Designed and developed the complete ETL pipeline in Pentaho Data Integration (PDI). Configured all steps for data extraction from the CSV file, data cleansing (handling duplicates), data transformation to align with the star schema, and loading data into the MySQL database. Implemented the loading of all dimension tables and the Sales Fact Table. |
| PARVENDAN | Part D: Data Analysis Using SQL Queries & Report Compilation | Created a minimum of three SQL queries (including OLAP functions) to answer various business questions from the data warehouse. Analyzed the query results and contributed to the interpretation of findings. Took a lead role in compiling and formatting the final project report. |

# SELF-REFLECTIVE REPORT

**Ashish (Part A):** "My initial focus was on understanding the e-commerce business domain, which allowed me to define the necessary business rules and subsequently develop the foundational ER-D."

**Bhulakshmi (Part B):** "Building upon the ER-D, I concentrated on designing the data warehouse schema, carefully considering the dimensions, facts, and their relationships to optimize for analysis."

**Christo (Part C):** "A significant portion of my effort was dedicated to the ETL process. I designed and implemented the Pentaho transformation to extract, cleanse, transform, and load the data into our MySQL data warehouse, ensuring data integrity and alignment with the designed schema."

**Parvendan (Part D):** "My contribution involved leveraging the populated data warehouse to answer key business questions through SQL queries, including the application of OLAP functions. I also took responsibility for the final report compilation."

Our four-member group consisting of Christo, Ashish, Bhulakshmi, Parvendan achieved project success through continuous teamwork and group expertise. The group photo from one of our numerous working meetings showcases our dedication to resolve multiple assignment obstacles and complete tasks.



# References:

Kimball, R. and Ross, M., 2013. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. Indianapolis, IN: Wiley.

Inmon, W. H., 2005. *Building the Data Warehouse*. 4th ed. Indianapolis, IN: Wiley.

Date, C. J., 2003. *An Introduction to Database Systems*. 8th ed. Boston, MA: Addison Wesley.

Connolly, T. and Begg, C., 2015. *Database Systems: A Practical Approach to Design, Implementation, and Management*. 6th ed. Harlow: Pearson Education.

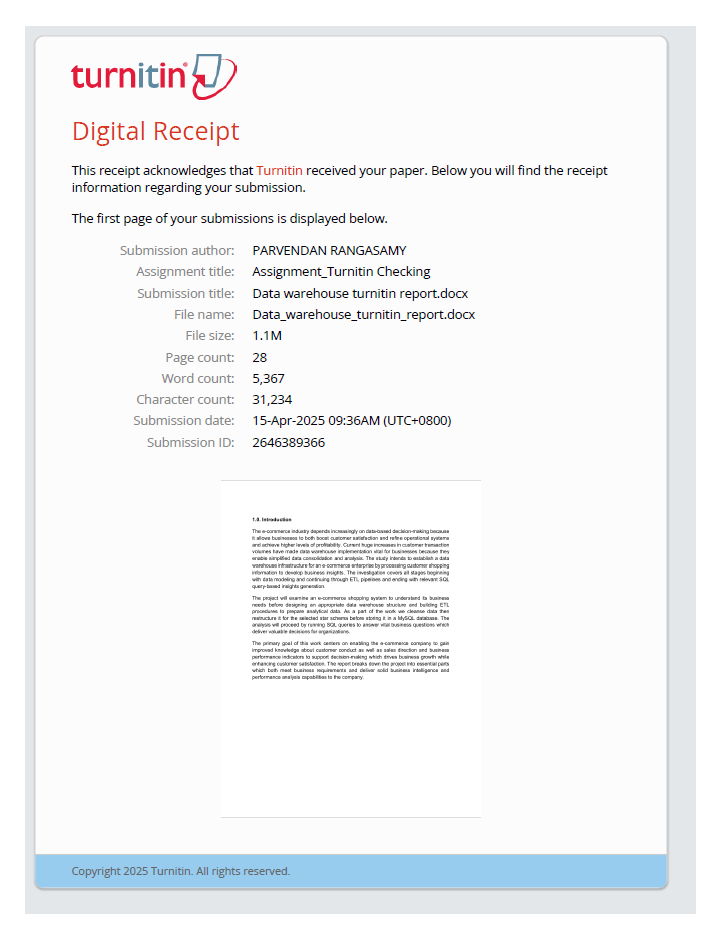
Vassiliadis, P. and Sellis, T., 2011. A Survey of Data Warehouse Process Modeling. *International Journal of Data Warehousing and Mining*, 7(1), 1-24.

**Websites/Online Resources:**

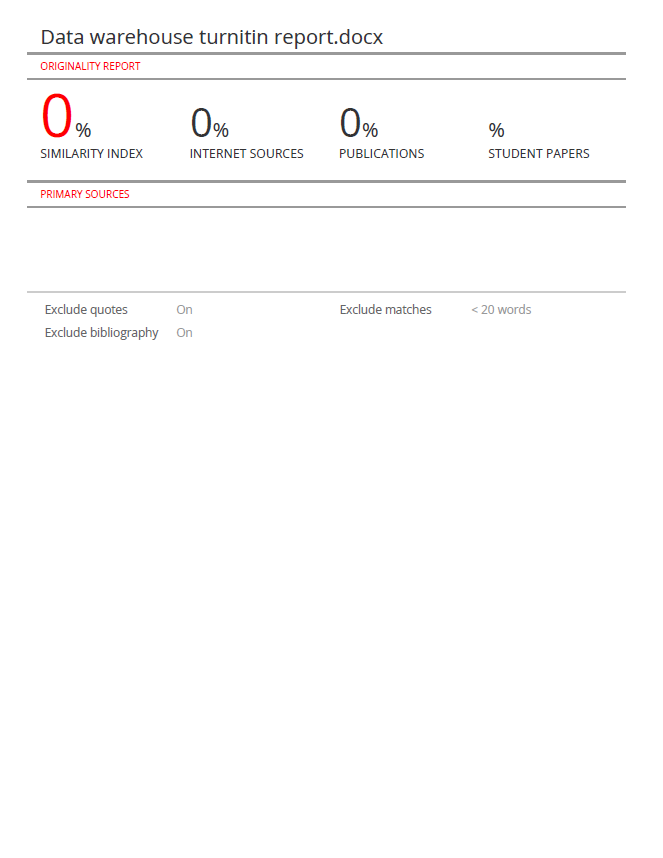
Pentaho Community. *Pentaho Documentation*. Available at:[Hitachi Vantara Documentation Portal](https://docs.hitachivantara.com/p/pentaho-dia)

MySQL. *MySQL Reference Manual*. Available at:<https://dev.mysql.com/doc/refman/8.0/en/>

# Appendix 1:

****

# Appendix 2:

****