**Task 1: Relational Database Design and Implementation**

Joseph Emmanuel Cayetano

Western Governors University

D597: Data Management Task 1

September 16, 2024

1. **Select one provided scenario**

I selected Scenario 2, which is about an online marketplace called EcoMart that has a platform for environmentally conscious consumers that want eco-friendly products.

**A1. Describe a business problem that can be solved with a database solution that aligns with Scenario 2**

One business problem about EcoMart’s scenario is having a single-table approach to store their customer orders data. One of the biggest drawbacks of creating a single table is that the “data management process become more complex” (Leonardo, 2023, par. 7). Since EcoMart is an emerging online marketplace, their single-table approach will slowly become problematic. For example, after examining their current data table, there is a higher risk of data entry errors because the same information is repeated for every row. The unit price and unit cost for each item type are repeatedly inputted for every order in the table. The unit price and unit cost are always the same for each item type so there is really no need for it to be inputted repeatedly. Also, the region values are written repeatedly, despite only having a handful of unique values. Manually inputting this same information could lead to data inconsistencies. Additionally, since there are more data in the table, performance issues may appear, and it can slow down the SQL software when querying big volumes of data. We could solve this problem by normalizing EcoMart’s database by creating separate tables for entities like Sales Order, Items and Region. Creating separate entities would reduce data redundancy and enhance scalability to ensure that EcoMart’s platform can handle increasing data volumes without harming speed and performance.

**A2. Propose a data structure to solve the identified business problem**

The data structure that I will use to solve the identified business problem for Ecomart will involve three tables/entities and they are all connected to each other. The three tables/entities are Sales Order, Items, and Region. For the Sales Order table, it will have 13 columns which are: order ID, country, item type, sales channel, order priority, order date, ship date, units sold, total revenue, total cost, total profit, and region ID. Order ID will the primary key. Item type is a foreign key that references the primary key of the Items table, which is also named item type. The Items table has three columns and they are item type, unit price, and unit cost. The Region table has two columns and they are region ID and region. Region ID is the primary key of the Region table.



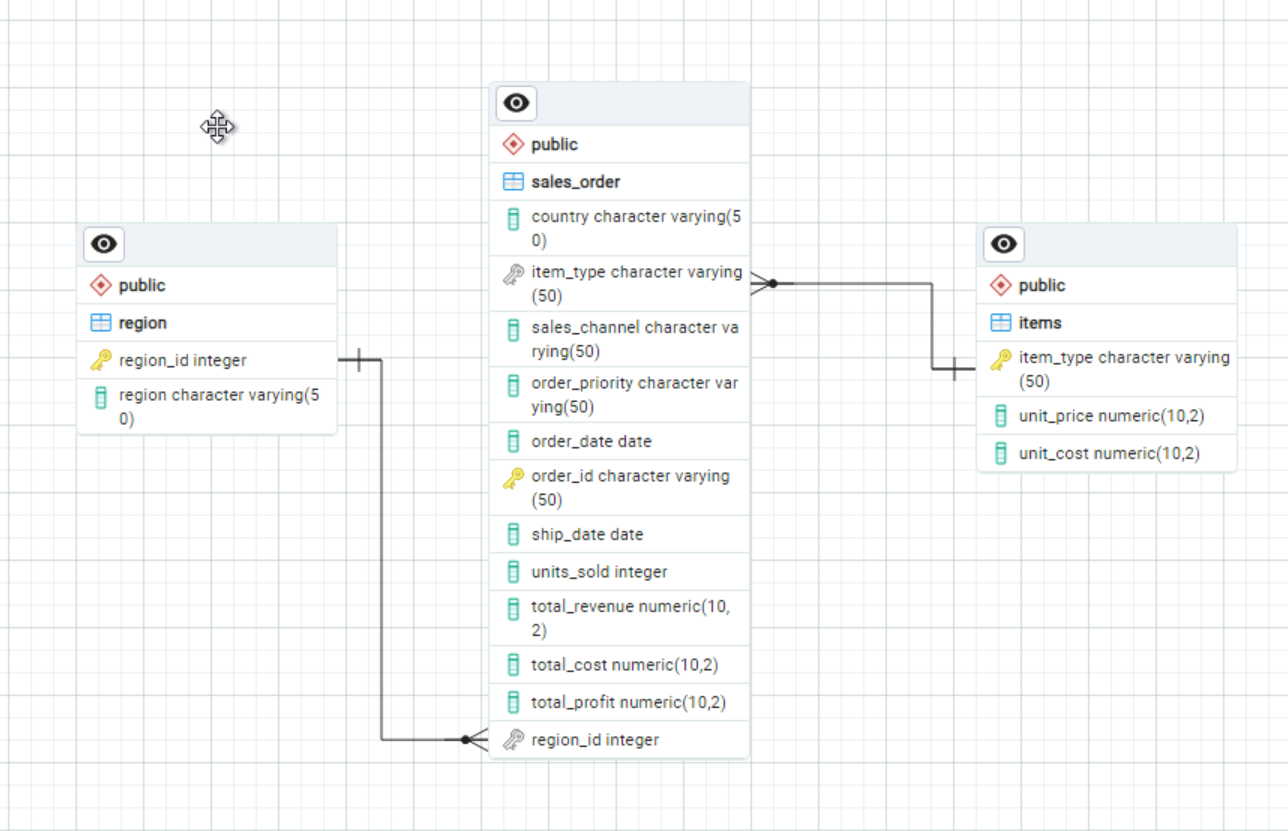
**A3. Justify why a database solution will solve the identified business problem**

By normalizing the database by creating separate tables for different entities, it will solve EcoMart’s business problem because it will reduce data redundancy, lowering the risk of inconsistencies. For example, having a separate Items table where it has all the different item types and their set unit prices and costs, it eliminates the need to include the unit price and cost for an item type in every order in the Sales Order table, which is the main table. Also, it makes it easy to update the unit price and cost for an item type because updates will be made in a smaller table, automatically reflecting across al related records. The Sales Order table will use the item type column as a foreign key to reference the primary key of Items table to retrieve the prices and costs for all item types. As for having a separate table for Region, It will remove the need to input the name of region for every order in the Sales Order table. Since there are only seven distinct region values, it is more practical to use numbers instead of typing out the whole region name for every order. This is the reason why I added a region ID column and used it as a foreign key to reference the primary key of Region table to retrieve the region names for each order. Implementing these data normalization methods will help improve the data integrity, consistency, scalability, and performance of EcoMart’s database.

**A4. Explain how the business data will be used within the database solution**

Here is how EcoMart’s business data will be used within the database solution. The Sales Order table will act as the main table in the database, and it will contain all sales orders including every order’s country of origin, order and shop date, region ID, and total revenue, cost, and profit. The Sales Order table can be used to track orders, analyze shipping times, view geographic trends in purchasing behavior, and perform profit analysis. As for the Items table, it will contain different item types and their unit prices and costs. The Items table will serve as a connection table to the Sales Order, and it can be valuable because it enables tracking of individual product types, their costs. And prices. This helps to make inventory management decisions. As for the Region table, it contains region names and like the Items table, it also serves as a connection table to the Sales Order table. The Region table is useful geographic trends in purchasing behavior of EcoMart’s customers.

1. **Logical data model in the database solution**

****

1. **Describe the database objects and storage, identifying the file attributes within the database solution**

There are three main objects or tables in the database, and they are Sales Order, Items, and Region. The Sales Order table is the main table in the database, and it contains all the details for every sales order. It has twelve attributes or columns. The first attribute is country with a character data type, and it is where the sales order was placed. The second attribute is item type with a character data type, and it is the type of item ordered. This attribute also serves as a foreign key to reference the Items table, which contains the unit price and cost for each item type. The third attribute is sales channel with a character data type, and it is the channel through which the sale was made. The fourth attribute is order priority with a character data type, and it is the priority level of the order. The fifth attribute is order date with a date data type, and it the date the order was placed. The sixth attribute is order ID with a character data type, and it is the primary key in the Sales Order table. It serves as a unique identifier for each order. The seventh attribute is ship date with a date data type, and it is the date the order was shipped. The eight attribute is units sold with an integer data type, and it is the number of units sold. The ninth attribute is total revenue with a numeric data type, and it is the total revenue from the order. The tenth attribute is total cost with a numeric data type, and it is the total cost of the order. The eleventh attribute is total profit with a numeric data type, and it is the total profit from the order. The twelfth and last attribute is region ID with an integer data type, and it serves as a foreign key to reference the Region table, which contains region names. The Items table contains the prices and costs for all different items sold. It has three attributes. The first attribute is item type with a character data type, and it is the type of item ordered. It also serves as the primary key in the table. The second attribute is unit price with a numeric data type, and it is the price per unit of the item. The third attribute is unit cost with a numeric data type, and it is the cost per unit of the item. The Region table stores information about different regions. It has two attributes. The first attribute is region ID and it is a primary key that serves as a unique identifier for each region. The second attribute is region and it is the name of the region. Regarding the file attributes, the size of the d597\_task\_1 database is 33 MB. For the tables, the size of the Items table is 24kb, the size of the Region table is 2232kb, and the size of the Sales Order table is 23 MB. The Sales Order table contains the most data. As for the permissions for the d597\_task\_1 database, postgres is the only user that can perform insert, select, and update commands in the database. PostgreSQL data files are stored in this path: "C:/Program Files/PostgreSQL/16/data."

**D. Discuss how the proposed database design addresses scalability concerns, including strategies that align with the chosen scenario**

The proposed database design will solve EcoMart’s scalability concerns in multiple ways. First, having a separate Items table will make it easy to add new product items or update unit costs or prices because it can be done in a smaller table with less data, as opposed to the main table that has thirteen columns. This is useful for a business like EcoMart that has a wide range of products. Second, dividing a large table like the Sales Order table into smaller, more manageable parts can reduce query load times and resource usage, which is important for handling big volumes of data. Third, since the tables are well-defined and organized, it would be effortless to locate any data inconsistency regardless of how large the data is.

**E. Outline the privacy and security measures that should be implemented in the proposed database design**

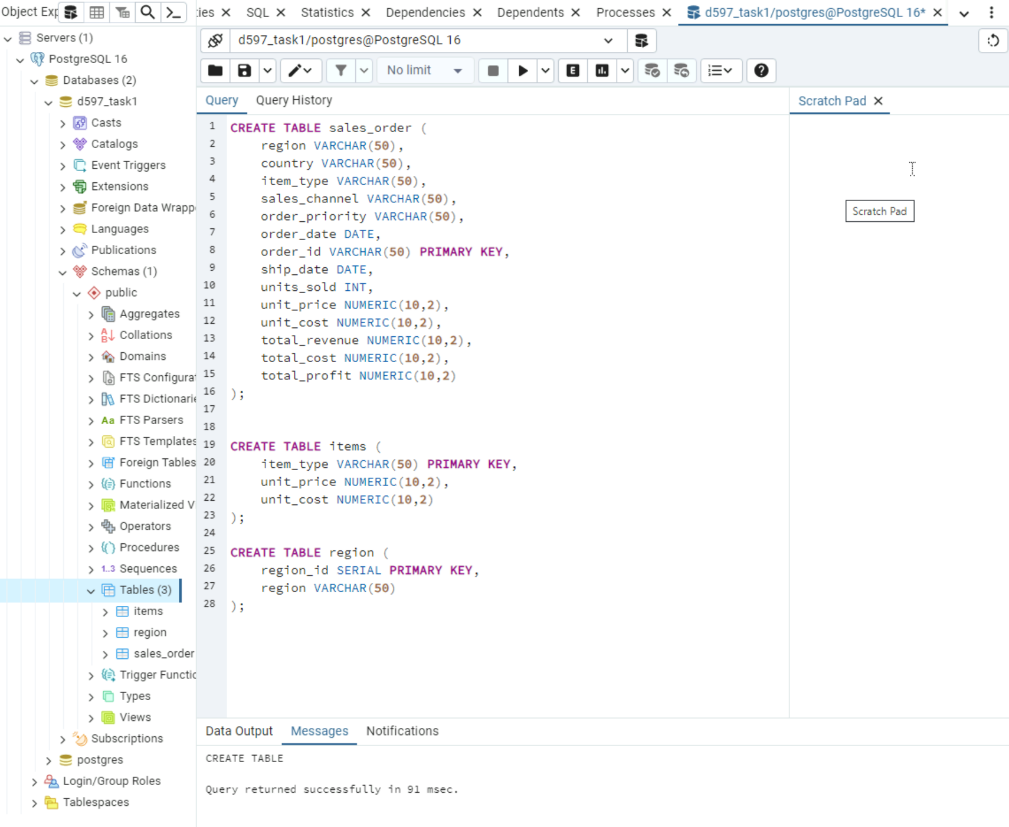
There are security and privacy measures that should be implemented in the database design for EcoMart. One measure is using Multi-Factor Authentication (MFA). MFA adds a second layer of protection for database access. Another measure is using data encryption protocols like Transport Layer Security (TLS), which provides encryption for sensitive data in the database. Performing regular backups of the database also helps in case of technical issues or breaches.

**F1. Script to create a database instance named “D597 Task 1.” Provide screenshot**

A screenshot of a computer

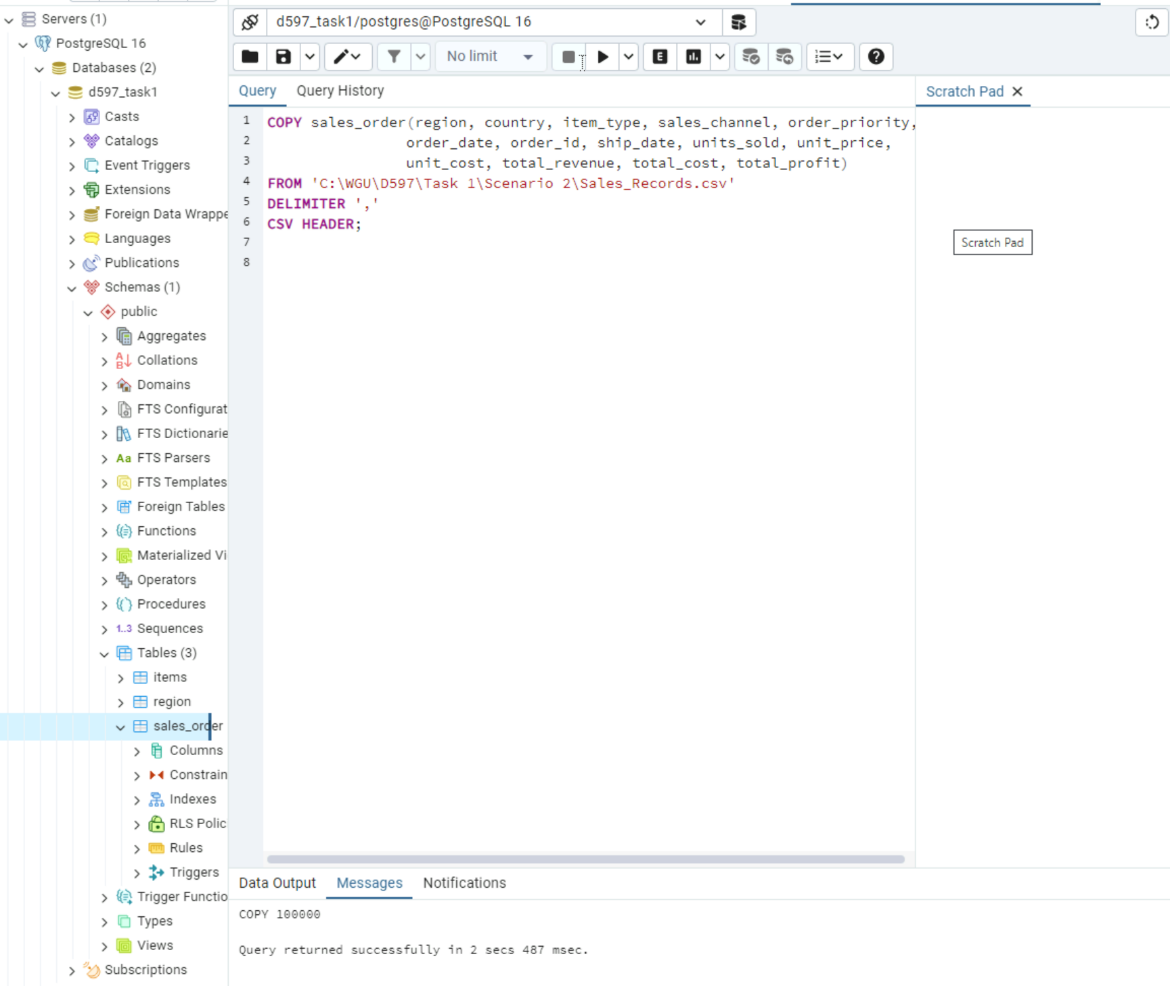
Description automatically generated

This is the SQL script to create the D597\_TASK1 database.

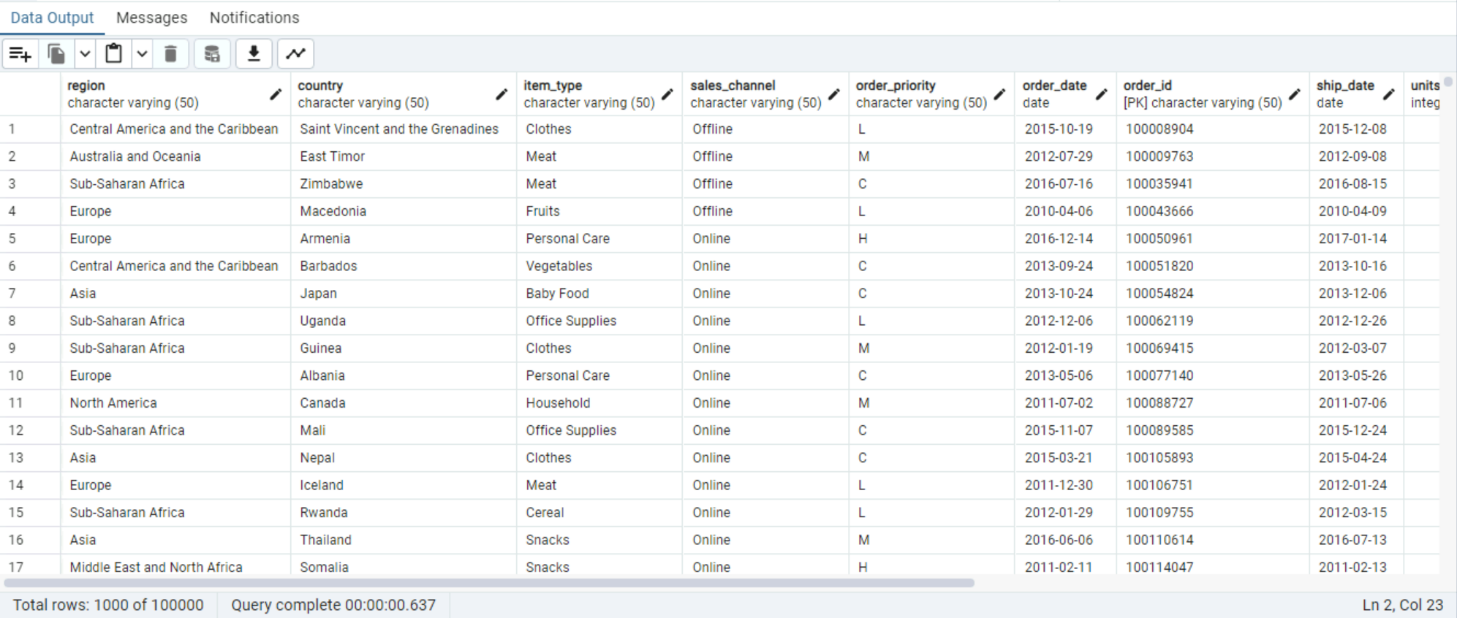


This is the SQL script to create the tables in the D597\_TASK1 database.

**F2. Script to import the data records from the chosen scenario CSV files into the database instance. Provide a screenshot**



This is the script that imports data records from a CSV file into the database.



A screenshot of a computer

Description automatically generated

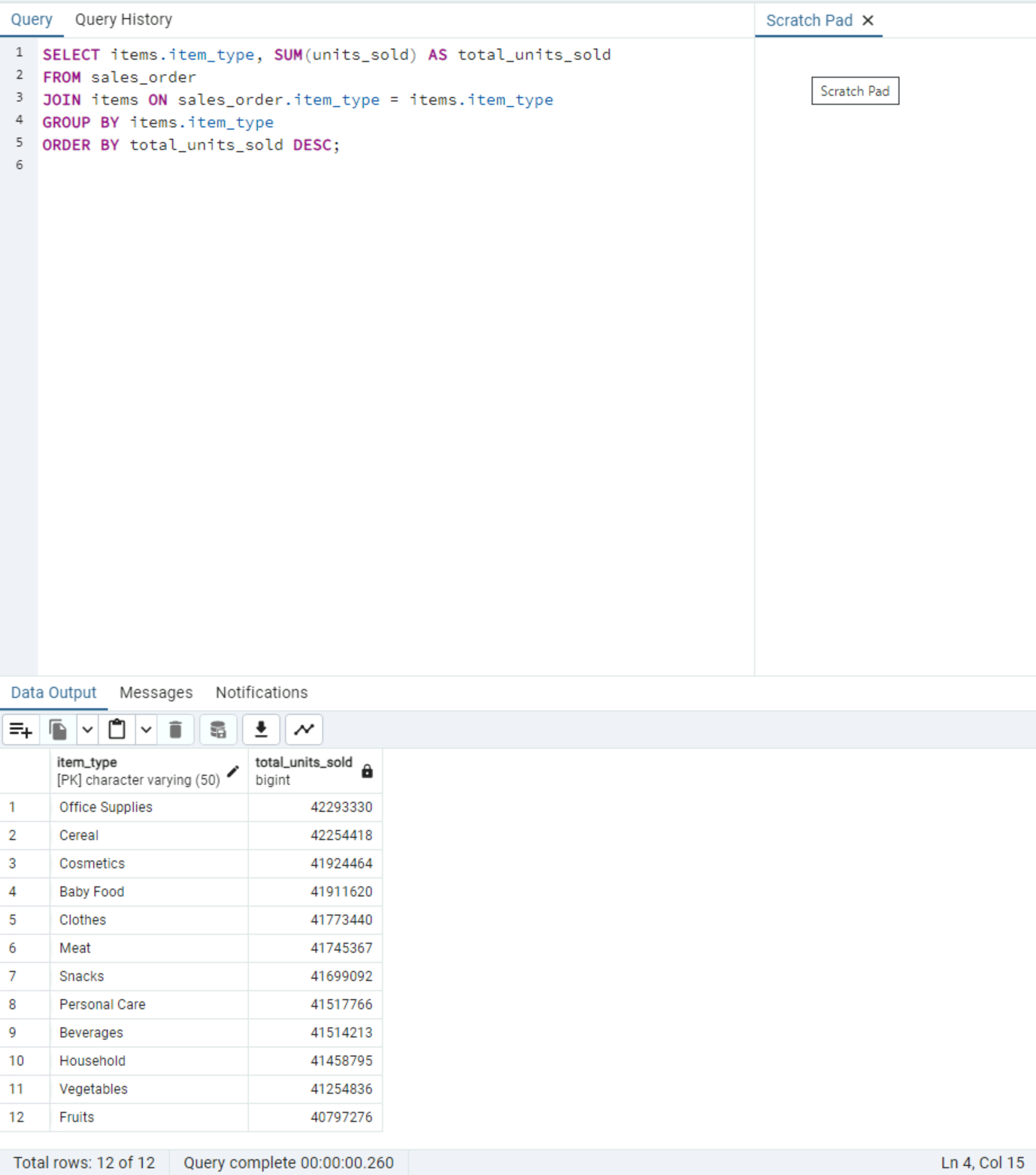
After importing the data records, this is what the table Sales\_Order looks like. The import statement inserted all the data from the CSV file.

**F3. Script for three queries to retrieve specific information from the database that solve the business problem.**

A screenshot of a computer

Description automatically generated

This is the first query. This query helps identify which regions generate the most revenue. This query returns region names and the total revenue for each region.



This is the second query. This query identifies the most sold items, aiding inventory management. This query returns all the item types, and the units sold for each item type.

A screenshot of a computer

Description automatically generated

This is the third query. This query identifies regions with the most shipping delays. This query returns all the region names and the number of delayed orders for each region.

**F4. Apply optimization techniques to improve the run time of your queries from part F3, providing output results via a screenshot**

Before applying indexes to improve query performance, here are the execution times for all the three queries.

A screenshot of a computer

Description automatically generated

The first query has an execution time of 103 milliseconds.

A screenshot of a computer

Description automatically generated

The second query has an execution time of 85 milliseconds.

A screenshot of a computer

Description automatically generated

The third query has an execution time of 85 milliseconds.

A screenshot of a computer

Description automatically generated

This is a screenshot of two indexes being created. The first index is in the region id column, and the second index is in the item type column. This is in the Sales Record table. Here are the execution times for the three queries after applying the indexes.

A screenshot of a computer

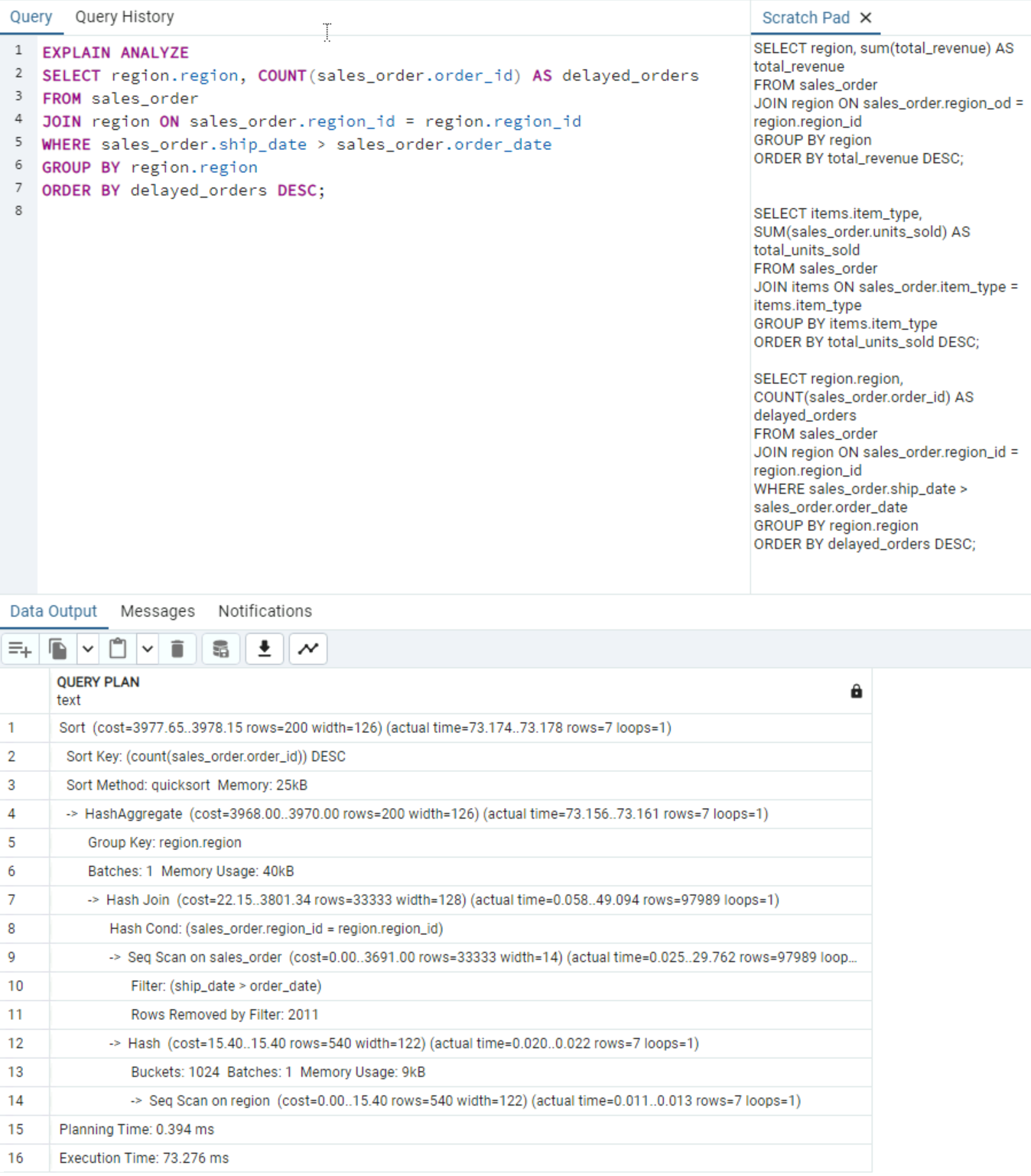
Description automatically generated

The execution time of the first query improved because it went from 103 to 99 milliseconds.

A screenshot of a computer

Description automatically generated

The execution time of the second query improved because it went from 85 to 74 milliseconds.



The execution time of the third query improved because it went from 82 to 73 milliseconds.

**G. Panopto Presentation**

Here is the link to my Panopto Presentation: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=50c012f6-eee9-45b3-9826-b1ed002ded95>

**References**

Leonardo, P. (2023, April 4). Why creating a single table in SQL is not always the best approach. Medium. https://medium.com/nerd-for-tech/why-creating-a-single-table-in-sql-is-not-always-the-best-approach-da179decfe50