ST1501 : DATA ENGINEERING   
CA2 INDIVIDUAL REPORT

Student ID: 1939213

Name: Silviana

Class: DIT/FT/2A/14

# Table of Contents

[**Table of Contents**](#_dp6ogvyjx6i9) **2**

[**1. Importing Data**](#_u0n60jmamcom) **3**

[1.1. Exporting Data from Microsoft SQL Server](#_rh8upvy7wldp) 3

[1.2. Setting Up a Database in Databricks](#_suyec4k5b88x) 3

[1.3. Loading Data to Databricks](#_jg8zhm31prca) 3

[**2. Queries from Microsoft SQL Server**](#_jooqnbhqtrxb) **4**

[2.1. Customer OrderCount](#_mf7e7zgz0don) 4

[2.2. Office Sales](#_b0vfz34atzmt) 4

[**3. Additional Query in Databricks**](#_jdowu8ug83qg) **5**

[3.1. Offices 2004 Sales Summary](#_gj3uku5hir3b) 5

# 1. Importing Data

## 1.1. Exporting Data from Microsoft SQL Server

There are different ways of exporting data from Microsoft SQL Server and I have tried 3 methods.

* BCP SQL Script

This method requires the user to change sql server configurations and exporting to C drive requires extra system level permission that I did not explore. Thankfully, I have a D drive and I managed to export my table as csv using this method. However, this method does not export column names. The script is provided in this zip file as exporttocsv.sql.

* Export as a flat file (.csv) using SQL Server Import and Export Wizard.
* Export as an excel file using SQL Server Import and Export Wizard and converting excel file into csv in Microsoft Excel.

## 1.2. Setting Up a Database in Databricks

The script to create a database in databricks is the same as creating a database on SQL Server.

|  |
| --- |
| **CREATE** **DATABASE** CarSalesDW\_B; |

## 1.3. Loading Data to Databricks

Different ways of exporting data will sometimes produce a different result, hence, a different loading method. I have explored methods of importing a .csv file in Databricks because of different file formats.

* Importing .csv file with column names

|  |
| --- |
| **CREATE** **TABLE** **IF** **NOT** **EXISTS** customerDIM **USING** CSV OPTIONS(**path** "/FileStore/tables/carSalesCA2/customer.csv", header "true", inferSchema "true"); |

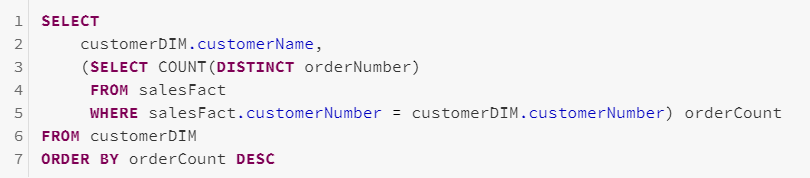
* Importing .csv file without column names

To import a csv file without column names, we would need to define our own column names and their corresponding data types.

|  |
| --- |
| **USE** CarSalesDW\_B; **CREATE** **TABLE** **IF** **NOT** **EXISTS**  employeesDIM (  employeeId INT,   employeeNumber INT,  lastName **STRING**,  firstName **STRING**,  extension **STRING**,  email **STRING**,  reportsTo INT,  jobTitle **STRING**  ) **USING** CSV OPTIONS(**path** "/FileStore/tables/carSalesCA2/employeesDIM.csv", inferSchema "true"); |

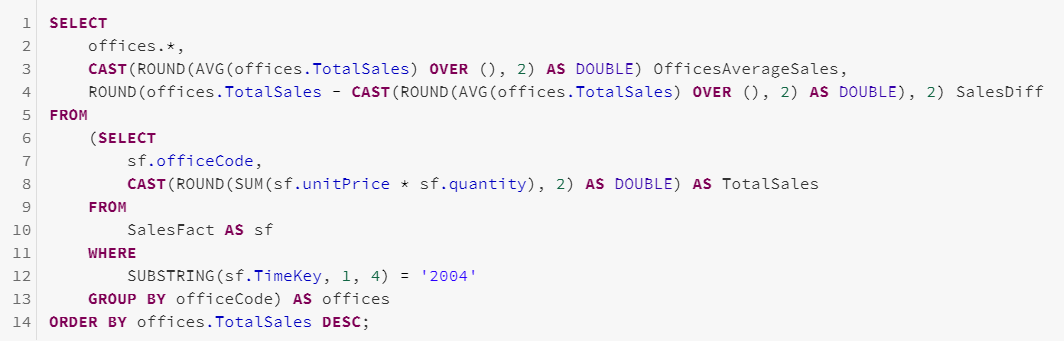
# 2. Queries from Microsoft SQL Server

## 2.1. Customer OrderCount



This query shows the customer name and the number of orders that the customer has made to ClassicCars, sorted by the number of orders. Using this query, ClassicCars can consider providing a membership or appreciation gift for customers that have exceeded a certain number of orders to make the customers feel valued and appreciated.

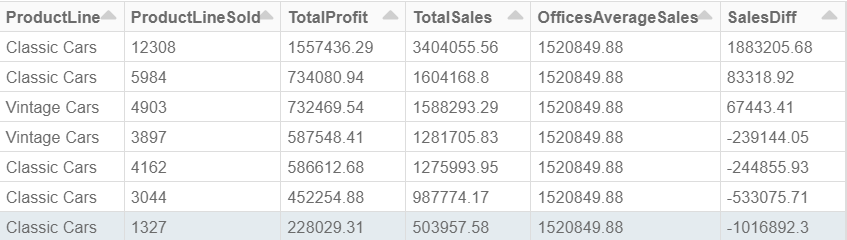
## 2.2. Office Sales



This query shows the office codes, total sales, average of total sales (by the 7 offices), and the difference between the total sales and average sales. With this, ClassicCars can see which office is doing better in terms of sales. The data acquired can be used for marketing planning purposes and future office locations.

# 3. Additional Query in Databricks

## 3.1. Offices 2004 Sales Summary



Expanding further from the Office Sales query, this query shows the summary of sales made by all offices in 2004 based on several metrics. This query can help ClassicCars plan, predict and manage their stores better in terms of the manpower needed, products or product lines to order per customer preferences, and which month would generate more sales.

The France and Japan branches are opposite of each other in terms of sales and employee count. This could be because the store in France is larger than the one in Japan. ClassicCars can use this data to assess if more manpower or a better marketing strategy is needed for the japan office.

As we can see from the data, the highest selling month is generally around the end and the beginning of the year. One factor that might contribute to this would be Christmas and New Year holidays, where people often buy gifts as a holiday tradition.

Another thing we can notice is that classic and vintage cars are the preferred product lines. ClassicCars can look into more classic and vintage car models to buy in the future.