# Lab Data extraction and transformation

## 

## SCENARIO

VanArsdel is a company that manufactures and sells sporting goods. The company has offices in the United States (US) and several other countries. Its sales comprise of US sales and International sales. VanArsdel’s sales come from its owned manufactured products, as well as other manufacturers’ products.

VanArsdel's US office stores the sales data on an Access database. VanArsdel International sales transactions are available as comma separated (CSV) files. They could be generated daily, either manually by someone, or automatically by an automated process. They are available in a dedicated folder. These CSV files have the same column structure as the sales table for the US sales that comes from the SQL Database.

You want to perform analysis on VanArsdel's worldwide sales data for the year 2000 to 2015. You need to bring all these data into Power BI Desktop before you can perform any analysis. Finally, you want to compare VanArsdel's country sales with the country population. You need to import the country population data from a less structured Excel report to Power BI.

## This lab comprises of three exercises:

* 1. In the first exercise, you will import data to Power BI Desktop from an Access database file.
  2. In the second exercise, you will import data from CSV files which resides in a file folder. You will append this new data to the corresponding existing data that comes from the Access Database.
  3. In the third exercise, you will import data to Power BI Desktop from an Excel file that is less structured.

## WHAT YOU'LL NEED

* + A computer with the latest version of Power BI Desktop installed on it.
  + A copy of the Access Database containing VanArsdel's US sales data.
  + 4 CSV files, containing VanArsdel’s international sales data:
    - CA Sales.csv
    - FR Sales.csv
    - DE Sales.csv
    - MX Sales.csv
  + An Excel file containing country population data.

## Connecting to an Access Database for the first time

If you are connecting to an Access database for the first time, you might need to install the Access redistributable package.

* + If you have a 32 bit machine, you need to install the 32 bit Power BI Desktop and the 32 bit redistributable of Access.
  + If you have a 32 bit Office installed (regardless of your machine), you need to install the 32 bit Power BI Desktop and the 32 bit redistributable of Access.
  + Otherwise, you can install the 64 bit Power BI Desktop and the 64 bit redistributable of Access.

Follow the link provided by Power BI Desktop when trying to connect to the Access database.

## Exercise 1: Import Data from Access Database

VanArsdel's US office stores the sales data on an Access database. You will need to perform analysis on that data, but before you can do so, you need to import the data to Power BI Desktop and perform some transformations.

**IMPORTANT!** Before you start, if your locale settings is not English (United States), you might want to change this, since the data you will import is based on this locale.

1. Extract the VanArsdel's Access database (AccessDatabasePowerBI.zip) to "C:\DAT207x". You can use other folders, but if you are going to use the starter file provided at each lab instead of your own file, then it is recommended to use the absolute path given above in order to successfully perform the schedule refresh in Lab 4.
2. Start with a blank Power BI Desktop file.
3. Connect to the Access database by using **Get Data** and select the Access database file.
4. Select the following tables to be imported: **bi\_date**, **bi\_geo**, **bi\_manufacturer**, **bi\_product**, and **bi\_salesFact**.
5. Transform data before loading to the data model.
6. Filter the rows on the **bi\_salesFact** query to include dates from January 1st, 2000. (Hint: Apply a **Date filter** in the **Date** column to import dates after December 31st, 1999. Before you can apply a date filter, you need to change the **Date** column's **Data Type** to **Date**).
7. Filter the rows on the **bi\_date** query to include dates from January 1st, 2000. (Hint: Apply a **Date filter** in the **Date** column to import dates after December 31st, 1999. Before you can apply a date filter, you need to change the **Date**column's **Data Type** to **Date**).
8. Rename the queries as follows:
   * bi\_date: **Date**
   * bi\_geo: **Locations**
   * bi\_manufacturer: **Manufacturers**
   * bi\_product: **Products**
   * bi\_salesFact: **Sales**
9. Load the data into Power BI Desktop. This might take a few minutes.
10. Explore the imported data in the **Data**View.

### **Lab Questions**

Answer the following question by reviewing the imported data in Power BI Desktop. (Hint: Use the Data view).

1. How many rows were imported from the Date query?
2. How many rows were imported from the Locations query?
3. How many rows were imported from the Manufacturer query?
4. How many rows were imported from the Products query?
5. How many rows were imported from the Sales table?

## Exercise 2: Import Data from a Folder Containing CSV Files

Continue with your Power BI file from the previous exercise. You want to create a Query for the International sales and append the Query to the Query from US Sales.

1. Extract VanArsdel's international sales data (InternationalSales.zip) to "C:\DAT207x". You can use other folders, but if you are going to use the starter file provided at each lab instead of your own file, then it is recommended to use the absolute path given above in order to successfully perform the schedule refresh in Lab 4. You should see 4 CSV files in the "C:\DAT207x\International" folder.
2. Import the data from the file folder by using the **Get Data**. To do this, click the**Get Data / More** option, select **Folder** and click **Connect**. Select the folder where you saved the 4 CSV files containing VanArsdel’s international sales data.
3. Transform data before loading to the data model.
4. Perform the following steps:
5. Select to combine (combined and transform) the content of those 4 files, leave the default settings, and click **OK**.
   * Name the query **International Sales**.
   * Filter the rows that are after **December 31st, 1999**.
   * Remove the **Source.Name** column.
6. Edit the **Sales** Query from the US Sales.
7. Append the **International Sales** table to the **Sales** query from the US Sales.
8. In the **Sales** query, add a custom column named **Country Name**which takes the value of the **Country** column when it is not null and the value of "USA" when the **Country** column is null.
9. Remove the **Country** column.
10. Load the data into Power BI Desktop. This might take a few minutes.
11. Explore the imported data in the **Data**View.
12. Hide the **International Sales** table from report view.

### **Lab Questions**

Answer the following question by reviewing the imported data in Power BI Desktop. (Hint: Use the Data view).

1. How many rows were imported from the International Sales query?
2. How many rows are now in the Sales query?

## Exercise 3: Import a Less Structured Data from an Excel File

Continue with your Power BI file from the previous exercise. You want to import an Excel report containing population data for the countries that VanArsdel operates.

1. Extract CountryPopulationByYear.zip to "C:\DAT207x". You can use other folders, but if you are going to use the starter file provided at each lab instead of your own file, then it is recommended to use the absolute path given above in order to successfully perform the schedule refresh in Lab 4.
2. Import the data from the file folder by using the **Get Data**. To do this, click the**Get Data / Excel**option. Select the Excel file and select the sheet containing the Country Population data. Transform data and perform the following steps:
   * Name the Query **Country Population**.
   * Remove the first three rows of the table.
   * Promote the row that represents the title of the column to the table headers.
   * Transform the yearly columns to rows and name the resulting columns appropriately. (Hint: You might find the **Unpivot Columns** or **Unpivot Other Columns** feature useful).
   * Rename the resulting columns to **Year** and **Population** respectively.
   * Set the **Data Type** of both the **Year** and **Population** columns to **Whole Number**.
   * Filter out the year **1999**.
3. Load the data into Power BI Desktop.
4. Explore the imported data in the **Data**View.

### **Lab Questions**

Answer the following question by reviewing the imported data in Power BI Desktop. (Hint: Use the Data view).

1. How many rows were imported from the Country Population query?
2. How many distinct values of the Year column were imported from the Country Population query?