Final Project: Exploratory Data Analysis with SQL

Estaimted time needed: **40-60** minutes

Pre-requisites:

- IBM Cloud Account
- IBM Db2 service

> NOTE: If you don't have an IBM Cloud account or Db2 service, follow this link and go through the steps given in the Hands-on Lab: Create Db2 service instance and Get started with the Db2 console.

Objectives

After completing this lab you will be able to:

- Describe three Chicago datasets
- Load the three datasets into three tables in a Db2 database
- Write and Execute SQL queries to perform exploratory data analysis

Project Overview

Imagine you have been hired by a non-profit organization that strives to improve socio-economic conditions and educational outcomes for children and youth in the City of Chicago. Your job is to analyze the census, crime, and school data.

You will be asked questions that will help you understand the data just like a real world data professional would. You will be assessed on the correctness of both your SQL queries and results.

Step-By-Step Assignment Instructions

In this assignment, you will (I) Become familiar with the datasets, (II) load them into a database, and (III) write and execute SQL queries to perform exploratory analysis on the data.

Task I: Review and familiarize yourself with the datasets

To complete the assignment problems you will be using three datasets that are available on the city of Chicago's Data Portal:

1. Socioeconomic Indicators in Chicago
2. Chicago Public Schools
3. Chicago Crime Data

1. Socioeconomic Indicators in Chicago

This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

2. Chicago Public Schools

This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

3. Chicago Crime Data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

> This dataset is quite large - over 1.5GB in size with over 6.5 million rows. For the purposes of this assignment we will use a much smaller subset of this dataset.

Download and review the datasets

These datasets have been made availabel as .CSV (comma separated values) files.

> NOTE: Ensure you download the datasets using the links below instead of directly from the Chicago Data Portal. The versions linked below are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment. The CSV

> file provided above for the Chicago Crime Data is a very small subset of the full dataset available from the Chicago Data Portal. The original dataset is over 1.55GB in size and contains over 6.5 million rows. For the purposes of this assignment you will use a much smaller sample with only about 500 rows.

To download and save the datasets as .csv files on your device, click on each of the links below (or Right Click and Save Link As):

1. Chicago Socioeconomic Indicators

2. Chicago Public Schools

3. Chicago Crime Data

Once you have downloaded the above datasets, preview them using a text editor (e.g. Notepad / TextEdit) or a spreadhseet (Excel/Numbers/Google Sheets) to become familiar with them. Review the various fields/columns in each dataset and the type of data they contain.

> NOTE: While viewing the datasets on your device do NOT modify or save them. If you make any changes to the .CSV files, the data may not load properly in the database.

Task II: Load the datasets in database tables

To analyze the data using SQL, it first needs to be stored in the database. Perform this task using the LOAD tool in the Db2 console.

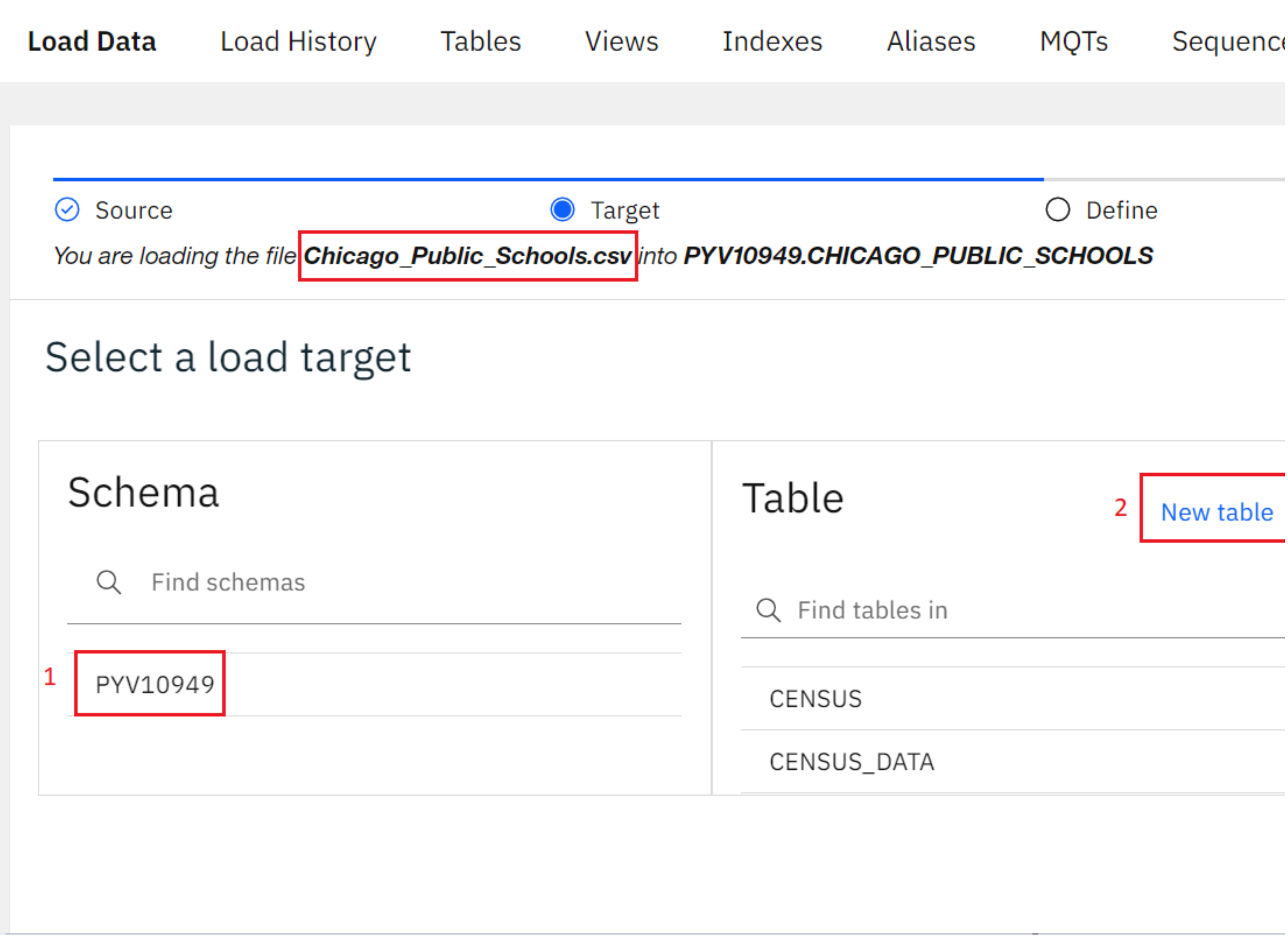
- Open Db2 console
- Open the **LOAD** tool
- Select / Drag the .CSV file for the dataset e.g. CHICAGO PUBLIC SCHOOLS
- Load the dataset into a new table

Then follow the steps on-screen instructions to load the data. This is similar to the Exercise 2 in the Module / Weelk2 Lab - Create and Load tables. The only difference with that lab is that in Step 6 of the instructions you will need to click on create *(+) New Table* and specify the name of the table you want to create and then click *Next*.

Name each of the new tables as follows:

1. **CENSUS_DATA**
2. **CHICAGO_PUBLIC_SCHOOLS**
3. **CHICAGO_CRIME_DATA**

> For reference review the Screenshot below illustrating loading of `Chicago_Public_School.csv` into a table called `CHICAGO_PUBLIC_SCHOOL`. The sequence of steps involved are numerically labelled and highlighted in the screenshot using red rectangles: (1) Your Db2 schema name e.g.: PYV10949 ; (2) New Table + (3) Table Name e.g. CHICAGO_PUBLIC_SCHOOL ; (4) Create button ; (5) Next button



You are loading the file ***Chicago_Public_Schools.csv*** *into* ***PYV10949.CHICAGO_PUBLIC_SCHOOLS***

Select a load target

NOTE: If any of the datasets such as the Chicago Socioeconomic Data (Census data) has already been loaded from a previous lab, you can skip loading it again.

NOTE: If you find the timestamp error while loading the data, then you may need to change/overwrite the default Timestamp format of **YYYY-MM-DD HH: MM: SS to MM/DD/YYYY HH: MM: SS TT**. You can also go through the link how to update/modify the timestamp format.

Task III: Write and execute queries to analyze the data

Carefully read and understand what is required for each query. Compose and execute the appropriate SQL queries in Db2 to answer each of the problems. Take a screenshot of each query and its results and save it as a .jpg (or .png) file to reference later in the project evaluation stage that comes next.

Problem 1

Find the total number of crimes recorded in the CRIME table.

Problem 2

Retrieve first 10 rows from the CRIME table.

Problem 3

How many crimes involve an arrest?

Problem 4

Which unique types of crimes have been recorded at GAS STATION locations?

▶ Click here for a hint

Problem 5

In the CENUS_DATA table list all Community Areas whose names start with the letter 'B'.

Problem 6

Which schools in Community Areas 10 to 15 are healthy school certified?

Problem 7

What is the average school Safety Score?

Problem 8

List the top 5 Community Areas by average College Enrollment [number of students]

Problem 9

Use a sub-query to determine which Community Area has the least value for school Safety Score?

Problem 10

[Without using an explicit JOIN operator] Find the Per Capita Income of the Community Area which has a school Safety Score of 1.

In case you get stuck, feel free to review the content and labs in previous modules and the practice assignment.

Good luck!

Author

Rav Ahuja

Other Contributor(s)

Malika Singla