DB0201EN-Week4-2-2-PeerAssign-2-py

October 6, 2018

Assignment: Querying Real World Data Sets with SQL

1 Introduction

Using this Python notebook you will: 1. Understand 3 Chicago datasets
1. Load the 3 datasets into 3 tables in a Db2 database 1. Execute SQL queries to answer assignment questions

1.1 Understand the datasets

To complete the assignment problems in this notebook you will be using three datasets that are available on the city of Chicago's Data Portal: 1. Socioeconomic Indicators in Chicago 1. Chicago Public Schools 1. Chicago Crime Data

1.1.1 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.

For this assignment you will use a snapshot of this dataset which can be downloaded from: https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv

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

1.1.2 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. This dataset is provided by the city of Chicago's Data Portal.

For this assignment you will use a snapshot of this dataset which can be downloaded from: https://ibm.box.com/shared/static/0g7kbanvn5l2gt2qu38ukooatnjqyuys.csv

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

1.1.3 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.

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 sample of this dataset which can be downloaded from: https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv

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

1.1.4 Store the datasets in database tables

In many cases the dataset to be analyzed is available as a .CSV (comma separated values) file, perhaps on the internet. To analyze the data using SQL, it first needs to be stored in the database.

While it is easier to read the dataset into a Pandas dataframe and then PERSIST it into the database as we saw in Week 3 Lab 3, it results in mapping to default datatypes which may not be optimal for SQL querying. For example a long textual field may map to a CLOB instead of a VARCHAR.

Therefore, **it is highly recommended to manually load the table using the database console LOAD tool, as indicated in Week 2 Lab 1 Part II**. The only difference with that lab is that in Step 5 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".

Now open the Db2 console, open the LOAD tool, Select / Drag the .CSV file for the first dataset, Next create a New Table, and then follow the steps on-screen instructions to load the data. Name the new tables as folows:

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

1.1.5 Connect to the database

Let us first load the SQL extension and establish a connection with the database

```
In [ ]: %load_ext sql
```

```
In [ ]: # Remember the connection string is of the format:
        # %sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name
        # Enter the connection string for your Db2 on Cloud database instance below
        %sql ibm_db_sa://
```

1.2 Problems

Now write and execute SQL queries to solve assignment problems

1.2.1 Problem 1

How many rows are in each dataset?

```
In [ ]: # Rows in Census Data (Socieconimic Indicators)
```

```
In [ ]: # Rows in Public Schools
```

```
In [ ]: # Rows in Crime Data
```

1.2.2 Problem 2

Find average college enrollments by community area

1.2.3 Problem 3

Find the number of schools that are healthy school certified

1.2.4 Problem 4

How many observations have a Location Description value of GAS STATION

1.2.5 Problem 5

Retrieve a list of the top 10 community areas which have most number of schools and sorted in descending order.

1.2.6 Problem 6

How many observations have value MOTOR VEHICLE THEFT in the Primary Type variable (this is the number of crimes related to Motor vehicles)

1.2.7 Problem 7

Find the minimum "Average Student Attendance" for community are where hardship is 96. Hint: use INNER JOIN