IBM Developer SKILLS NETWORK

Assignment: Notebook for Peer Assignment

Introduction

Using this Python notebook you will:

1. Understand 3 Chicago datasets
2. Load the 3 datasets into 3 tables in a Db2 database
3. Execute SQL queries to answer assignment questions

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
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.

For this assignment you will use a snapshot of this dataset which can be downloaded from: Census Data

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. 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: Chicago Public School

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.

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: Chicago Crime Data

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

Download the datasets

In many cases the dataset to be analyzed is available as a .CSV (comma separated values) file, perhaps on the internet. Click on the links below to download and save the datasets (.CSV files):

1. **CENSUS_DATA:** Census Dataset
2. **CHICAGO_PUBLIC_SCHOOLS** Chicago Public School
3. **CHICAGO_CRIME_DATA:** Chicago Crime Data

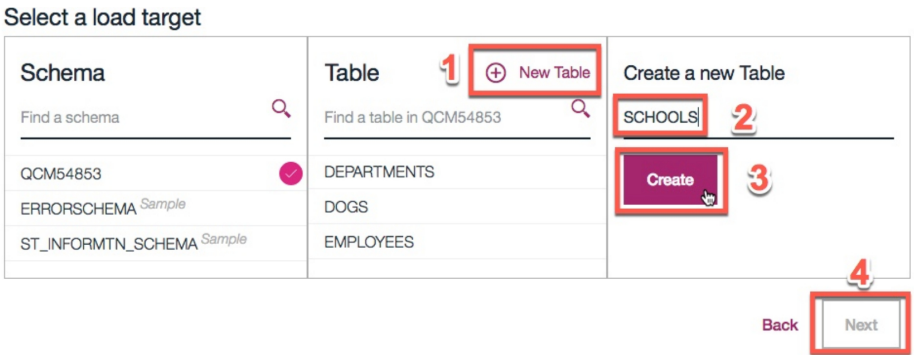
NOTE: Ensure you have downloaded the datasets using the links above instead of directly from the Chicago Data Portal. The versions linked here are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment.

Store the datasets in database tables

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:

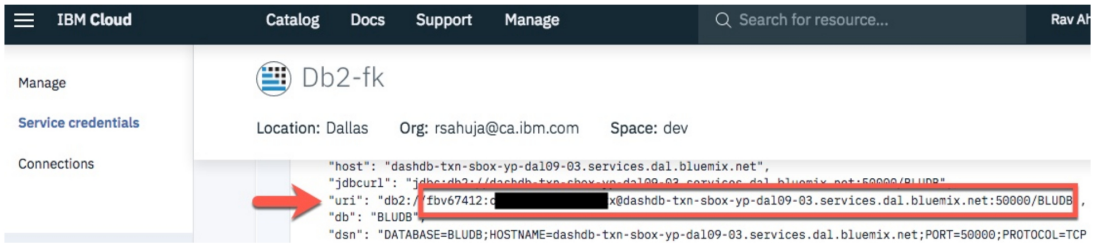
1. **CENSUS_DATA**
2. **CHICAGO_PUBLIC_SCHOOLS**
3. **CHICAGO_CRIME_DATA**

Connect to the database

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

[28]:
```
%reload_ext sql
```

In the next cell enter your db2 connection string. Recall you created Service Credentials for your Db2 instance in first lab in Week 3. From the **uri** field of your Db2 service credentials copy everything after db2:// (except the double quote at the end) and paste it in the cell below after ibm_db_sa://



[29]:
```
# 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://vhb98133:zv6p9nb0-z16p0xv@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
```

[29]:
```
'Connected: vhb98133@BLUDB'
```

Problems

Now write and execute SQL queries to solve assignment problems

Problem 1

Find the total number of crimes recorded in the CRIME table

[30]:
```
# Rows in Crime table
%sql select COUNT(*) as count_of_crimes FROM CHICAGO_CRIME_DATA1;
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[30]:

count_of_crimes
533

[31]:
```
### Problem 2

##### Retrieve first 10 rows from the CRIME table

%sql select * FROM CHICAGO_CRIME_DATA1 LIMIT 10;
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[31]:

id	case_number	DATE	block	iucr	primary_type	description	location_description	arrest	domestic	beat	district	ward	community_area_number	fbicode	x_coordinate	y_coordinate	YEAR	updat
3512276	HK587712	08/28/2004 05:50:56 PM	047XX S KEDZIE AVE	890	THEFT	FROM BUILDING	SMALL RETAIL STORE	FALSE	FALSE	911	9	14	58	6	1155838	1873050	2004	02/10/ 03:
3406613	HK456306	06/26/2004 12:40:00 PM	009XX N CENTRAL PARK AVE	820	THEFT	$500 AND UNDER	OTHER	FALSE	FALSE	1112	11	27	23	6	1152206	1906127	2004	02/28/ 03:
8002131	HT233595	04/04/2011 05:45:00 AM	043XX S WABASH AVE	820	THEFT	$500 AND UNDER	NURSING HOME/RETIREMENT HOME	FALSE	FALSE	221	2	3	38	6	1177436	1876313	2011	02/10/ 03:
7903289	HT133522	12/30/2010 04:30:00 PM	083XX S KINGSTON AVE	840	THEFT	FINANCIAL ID THEFT: OVER $300	RESIDENCE	FALSE	FALSE	423	4	7	46	6	1194622	1850125	2010	02/10/ 03:
10402076	HZ138551	02/02/2016 07:30:00 PM	033XX W 66TH ST	820	THEFT	$500 AND UNDER	ALLEY	FALSE	FALSE	831	8	15	66	6	1155240	1860661	2016	02/10/ 03:
7732712	HS540106	09/29/2010 07:59:00 AM	006XX W CHICAGO AVE	810	THEFT	OVER $500	PARKING LOT/GARAGE(NON.RESID.)	FALSE	FALSE	1323	12	27	24	6	1171668	1905607	2010	02/10/ 03:
10769475	HZ534771	11/30/2016 01:15:00 AM	050XX N KEDZIE AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE	1713	17	33	14	6	1154133	1933314	2016	02/10/ 03:

4494340	HL793243	12/16/2005 04:45:00 PM	005XX E PERSHING RD	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	TRUE	FALSE	213	2	3	38	6	1180448	1879234	2005	02/28/ 03:
3778925	HL149610	01/28/2005 05:00:00 PM	100XX S WASHTENAW AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE	2211	22	19	72	6	1160129	1838040	2005	02/28/ 03:
3324217	HK361551	05/13/2004 02:15:00 PM	033XX W BELMONT AVE	820	THEFT	$500 AND UNDER	SMALL RETAIL STORE	FALSE	FALSE	1733	17	35	21	6	1153590	1921084	2004	02/28/ 03:

[ ]:

•••

Problem 3

How many crimes involve an arrest?

[32]:
```
%sql SELECT COUNT(*) as count_of_arrest FROM CHICAGO_CRIME_DATA1 WHERE ARREST = 'TRUE';
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[32]:

count_of_arrest
163

Problem 4

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

[33]:
```
%sql SELECT PRIMARY_TYPE AS unique_types_of_crimes FROM CHICAGO_CRIME_DATA1 WHERE LOCATION_DESCRIPTION = 'GAS STATION' GROUP BY PRIMARY_TYPE;
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[33]:

unique_types_of_crimes
CRIMINAL TRESPA
NARCOTICS
ROBBERY
THEFT

Hint: Which column lists types of crimes e.g. THEFT?

Problem 5

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

[34]:
```
%sql SELECT COMMUNITY_AREA_NAME FROM CENSUS_DATA1 WHERE COMMUNITY_AREA_NAME LIKE 'B%';
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[34]:

community_area_name
Belmont Cragin
Burnside
Brighton Park
Bridgeport
Beverly

Problem 6

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

[35]:
```
%sql SELECT CPS.COMMUNITY_AREA_NUMBER, CPS.COMMUNITY_AREA_NAME, NAME_OF_SCHOOL, CPS.HEALTHY_SCHOOL_CERTIFIED FROM CHICAGO_PUBLIC_SCHOOLS1 as CPS JOIN CENSUS_DATA1 as CD ON CD.COMMUNITY_ARE
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[35]:

community_area_number	community_area_name	name_of_school	healthy_school_certified
10	NORWOOD PARK	Rufus M Hitch Elementary School	Yes

Problem 7

What is the average school Safety Score?

[36]:
```
%sql SELECT AVG(SAFETY_SCORE) AS average_school_safety_score FROM CHICAGO_PUBLIC_SCHOOLS1;
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[36]:

average_school_safety_score
49.504873

Problem 8

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

[37]:
```
%sql SELECT COMMUNITY_AREA_NAME, AVG(COLLEGE_ENROLLMENT) as AVG_COLLEGE_ENROLLMENT FROM CHICAGO_PUBLIC_SCHOOLS1 GROUP BY COMMUNITY_AREA_NAME ORDER BY AVG_COLLEGE_ENROLLMENT DESC LIMIT 5;
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[37]:

community_area_name	avg_college_enrollment
ARCHER HEIGHTS	2411.500000
MONTCLARE	1317.000000
WEST ELSDON	1233.333333
BRIGHTON PARK	1205.875000
BELMONT CRAGIN	1198.833333

Problem 9

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

[41]:
```
%sql SELECT COMMUNITY_AREA_NAME, SAFETY_SCORE FROM CHICAGO_PUBLIC_SCHOOLS1 WHERE SAFETY_SCORE = (SELECT MIN(SAFETY_SCORE) FROM CHICAGO_PUBLIC_SCHOOLS1);
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[41]:

community_area_name	safety_score
WASHINGTON PARK	1

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.

[39]:
```
%sql SELECT CPS.COMMUNITY_AREA_NAME, CPS.SAFETY_SCORE, CD.PER_CAPITA_INCOME FROM CENSUS_DATA1 as CD, CHICAGO_PUBLIC_SCHOOLS1 as CPS WHERE CPS.SAFETY_SCORE = 1 AND CPS.COMMUNITY_AREA_NUMBER
```

```
 * ibm_db_sa://vhb98133:***@dashdb-txn-sbox-yp-dal09-14.services.dal.bluemix.net:50000/BLUDB
Done.
```

[39]:

community_area_name	safety_score	per_capita_income
WASHINGTON PARK	1	13785

Author(s)

Rav Ahuja

Change log

Date	Version	Changed by	Change Description
2020-09-05	2.0	Malika Singla	Moved lab to course repo in GitLab