



(<https://cognitiveclass.ai>)

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](https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2) (<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2>)
2. [Chicago Public Schools](https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t) (<https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t>)
3. [Chicago Crime Data](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2) (<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>)

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>
(<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> (<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:

<https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv>
(<https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.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>
(<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:

<https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv>
(<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>

(<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:** <https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv>
(<https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv>)
2. **CHICAGO_PUBLIC_SCHOOLS** <https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv>
(<https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv>)
3. **CHICAGO_CRIME_DATA:** <https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv>
(<https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv>)

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".

LOAD

Source Target Define Finalize

You are loading the file **Chicago_Public_Schools - Progress_Report_Cards_2011-2012.csv**

Select a load target

Schema	Table	Create a new Table
Find a schema	Find a table in QCM54853	SCHOOLS
QCM54853	DEPARTMENTS	Create
ERRORSCHEMA <i>Sample</i>	DOGS	
ST_INFORMTN_SCHEMA <i>Sample</i>	EMPLOYEES	

Back 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 follows:

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

In [1]:

```
%load_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://`



The screenshot shows the IBM Cloud console interface. On the left, there is a sidebar with 'Manage', 'Service credentials', and 'Connections'. The main area displays 'Db2-fk' with details: Location: Dallas, Org: rsahuja@ca.ibm.com, Space: dev. Below this, there is a table of service credentials. A red arrow points to the 'uri' field of a credential, which contains the string: 'db2://fbv67412:xxxxx@dashdb-txn-sbox-yp-dal09-03.services.dal.ibm.com:50000/BLUDB'. The entire string is highlighted with a red box.

In [2]:

```
# 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://vwr23078:h7cpqz4wfmh-0xl8@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.blue
mix.net:50000/BLUDB
```

Out[2]:

```
'Connected: vwr23078@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

In [6]:

```
%sql select count(*) from CHICAGO_CRIME_DATA
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB  
Done.
```

Out[6]:

```
1  
533
```

Problem 2

Retrieve first 10 rows from the CRIME table

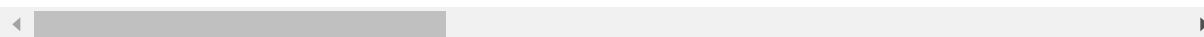
In [10]:

```
%sql select * from CHICAGO_CRIME_DATA fetch first 10 rows only
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[10]:

id	case_number	DATE	block	iucr	primary_type	description	location_
3512276	HK587712	2004-08-28 17:50:56	047XX S KEDZIE AVE	890	THEFT	FROM BUILDING	SMALL RET
3406613	HK456306	2004-06-26 12:40:00	009XX N CENTRAL PARK AVE	820	THEFT	\$500 AND UNDER	
8002131	HT233595	2011-04-04 05:45:00	043XX S WABASH AVE	820	THEFT	\$500 AND UNDER	HOME/RETIREM
7903289	HT133522	2010-12-30 16:30:00	083XX S KINGSTON AVE	840	THEFT	FINANCIAL ID THEFT: OVER \$300	F
10402076	HZ138551	2016-02-02 19:30:00	033XX W 66TH ST	820	THEFT	\$500 AND UNDER	
7732712	HS540106	2010-09-29 07:59:00	006XX W CHICAGO AVE	810	THEFT	OVER \$500	LOT/GARAGE(N
10769475	HZ534771	2016-11-30 01:15:00	050XX N KEDZIE AVE	810	THEFT	OVER \$500	
4494340	HL793243	2005-12-16 16:45:00	005XX E PERSHING RD	860	THEFT	RETAIL THEFT	GROCERY FC
3778925	HL149610	2005-01-28 17:00:00	100XX S WASHTENAW AVE	810	THEFT	OVER \$500	
3324217	HK361551	2004-05-13 14:15:00	033XX W BELMONT AVE	820	THEFT	\$500 AND UNDER	SMALL RET



Problem 3

How many crimes involve an arrest?

In [12]:

```
%sql select count(*) from CHICAGO_CRIME_DATA where ARREST = 'TRUE'
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[12]:

```
1
163
```

Problem 4

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

In [18]:

```
%sql select DISTINCT PRIMARY_TYPE from CHICAGO_CRIME_DATA where LOCATION_DESCRIPTION = 'GAS STATION'
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[18]:

```
primary_type
CRIMINAL TRESPASS
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'.

In []:

```
%sql select COMMUNITY_AREA_NAME from CENSUS_DATA where COMMUNITY_AREA_NAME like 'B%'
```

Problem 6

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

In [34]:

```
%%sql
select NAME_OF_SCHOOL,COMMUNITY_AREA_NUMBER,HEALTHY_SCHOOL_CERTIFIED from CHICAGO_PUBLIC_S
CHOOLS
      where COMMUNITY_AREA_NUMBER BETWEEN 10 AND 15 AND HEALTHY_SCHOOL_CERTIFIED = 'Yes';
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluem
i
x.net:50000/BLUDB
Done.
```

Out[34]:

name_of_school	community_area_number	healthy_school_certified
Rufus M Hitch Elementary School	10	Yes

Problem 7

What is the average school Safety Score?

In [38]:

```
%sql select avg(SAFETY_SCORE) from CHICAGO_PUBLIC_SCHOOLS
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluem
i
x.net:50000/BLUDB
Done.
```

Out[38]:

1
49.504873

Problem 8

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

In [39]:

```
%%sql

select COMMUNITY_AREA_NUMBER,COMMUNITY_AREA_NAME,COLLEGE_ENROLLMENT from CHICAGO_PUBLIC_SCH
HOOOLS
    where COLLEGE_ENROLLMENT > (select AVG(COLLEGE_ENROLLMENT) from CHICAGO_PUBLIC_SCHOOL
S)
    order by COLLEGE_ENROLLMENT desc
    limit 5;
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluem
x.net:50000/BLUDB
Done.
```

Out[39]:

community_area_number	community_area_name	college_enrollment
5	NORTH CENTER	4368
57	ARCHER HEIGHTS	3320
10	NORWOOD PARK	2922
58	BRIGHTON PARK	2883
16	IRVING PARK	2366

Problem 9

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

In [44]:

```
%sql select COMMUNITY_AREA_NAME,SAFETY_SCORE from CHICAGO_PUBLIC_SCHOOLS where SAFETY_SCOR
E = (select MIN(SAFETY_SCORE) from CHICAGO_PUBLIC_SCHOOLS);
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluem
x.net:50000/BLUDB
Done.
```

Out[44]:

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.

In [45]:

```
%%sql
```

```
select CD.PER_CAPITA_INCOME,CPS.COMMUNITY_AREA_NAME,CPS.SAFETY_SCORE  
      from CENSUS_DATA CD, CHICAGO_PUBLIC_SCHOOLS CPS  
      where CD.COMMUNITY_AREA_NUMBER = CPS.COMMUNITY_AREA_NUMBER and CPS.SAFETY_SCORE= '1';
```

```
* ibm_db_sa://vwr23078:***@dashdb-txn-sbox-yp-lon02-01.services.eu-gb.bluemix.net:50000/BLUDB  
Done.
```

Out[45]:

per_capita_income	community_area_name	safety_score
13785	WASHINGTON PARK	1

Copyright © 2018 cognitiveclass.ai ([cognitiveclass.ai?utm_source=bducopyrightlink&utm_medium=dswb&utm_campaign=bdu](https://cognitiveclass.ai/?utm_source=bducopyrightlink&utm_medium=dswb&utm_campaign=bdu)). This notebook and its source code are released under the terms of the [MIT License](https://bigdatauniversity.com/mit-license/) (<https://bigdatauniversity.com/mit-license/>).