IBM Developer
SKILLS NETWORK

Akram Alzaghir

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)
2. Chicago Public Schools (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)

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 (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_SKO/data/Census_Data_-_Selected_socioeconomic_indicators_in_Chicago__2008___2012-v2.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?cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette
_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette

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 (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_SKO/data/Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012-v3.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?cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette

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 (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_SKO/data/Chicago_Crime_Data-v2.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?cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette

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 (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera/data/Census_Data_-_Selected_socioeconomic_indicators_in_Chicago__2008___2012-v2.csv)
2. **CHICAGO_PUBLIC_SCHOOLS** Chicago Public School (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera/data/Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012-v3.csv)
3. **CHICAGO_CRIME_DATA:** Chicago Crime Data (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera/data/Chicago_Crime_Data-v2.csv)

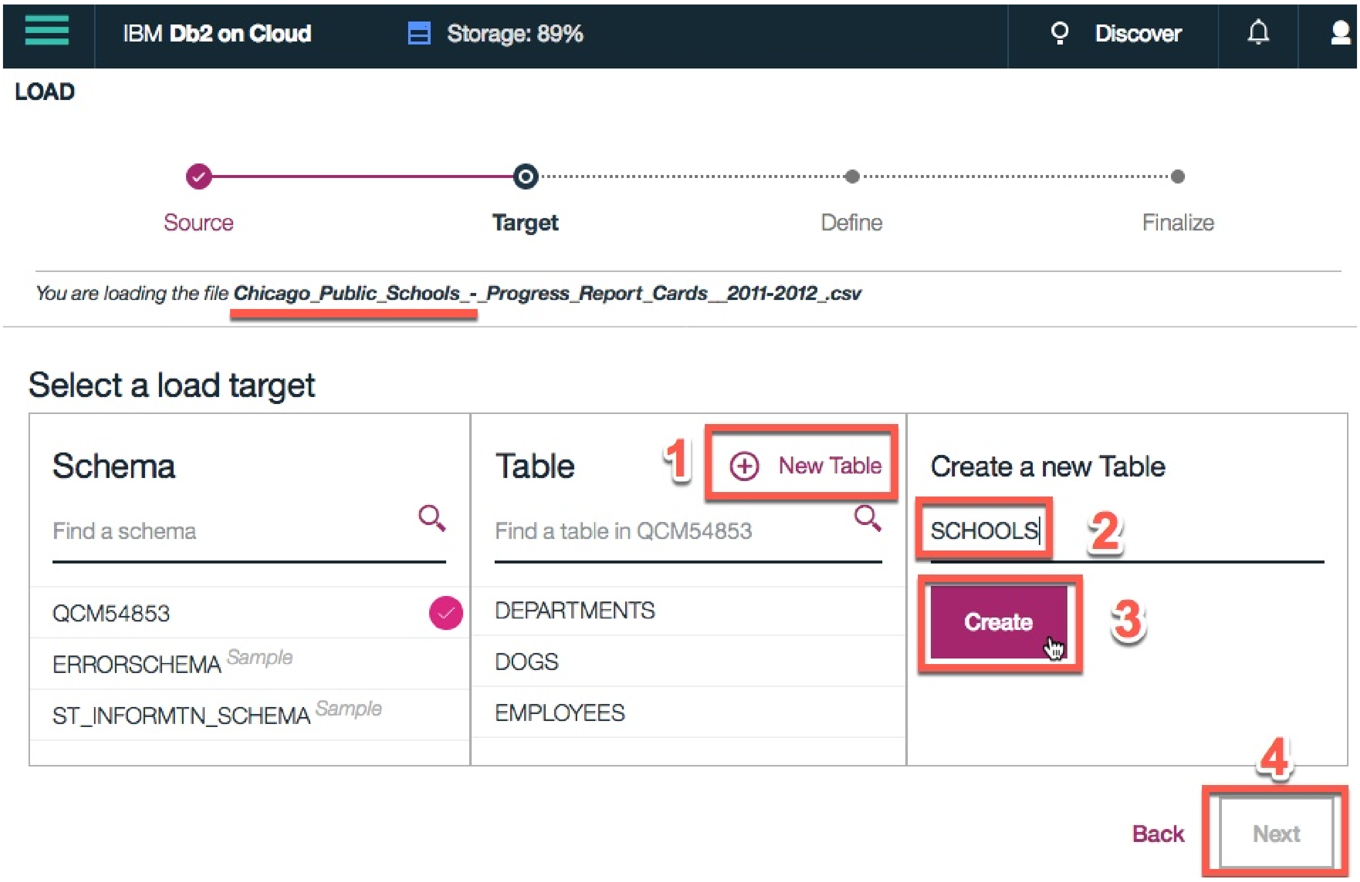
NOTE: Ensure you have downloaded the datasets using the links above instead of directly from the Chicago Data Portal. The versions linked here are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment.

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".



You are loading the file ***Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012_.csv***

Select a load target

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**

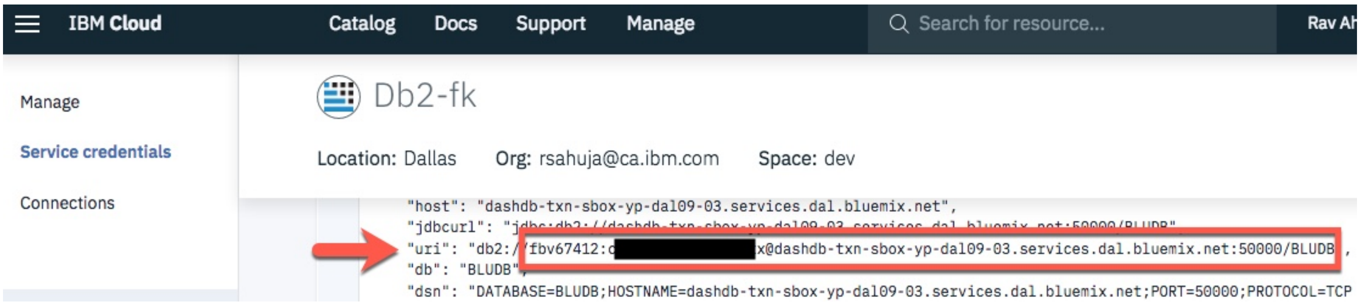
Connect to the database

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

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://



Location: Dallas Org: rsahuja@ca.ibm.com Space: dev

```
"host": "dashdb-txn-sbox-yp-dal09-03.services.dal.bluemix.net",
"jdbcurl": "jdbc:db2://dashdb-txn-sbox-yp-dal09-03.services.dal.bluemix.net:50000/BLUDB",
"uri": "db2://fbv67412:c[redacted]x@dashdb-txn-sbox-yp-dal09-03.services.dal.bluemix.net:50000/BLUDB",
"db": "BLUDB",
"dsn": "DATABASE=BLUDB;HOSTNAME=dashdb-txn-sbox-yp-dal09-03.services.dal.bluemix.net;PORT=50000;PROTOCOL=TCP
```

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://tnm91075:krrsj3j7zjn%40nphp@dashdb-txn-sbox-yp-dal09-12.services.dal.b
luemix.net:50000/BLUDB
```

Out[2]:

```
'Connected: tnm91075@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 [3]:

```
# we use count(*) function to count Rows in Crime table
%sql select count(*) as total_No_Crimes from CHICAGO_CRIME_DATA;
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[3]:

total_no_crimes
533

Problem 2

Retrieve first 10 rows from the CRIME table

In [4]:

```
%sql select * from CHICAGO_CRIME_DATA limit 10;
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[4]:

id	case_number	DATE	block	iucr	primary_type	description	locati
3512276	HK587712	2004-08-28 17:50:56	047XX S KEDZIE AVE	890	THEFT	FROM BUILDING	SMALL
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/RETI
7903289	HT133522	2010-12-30 16:30:00	083XX S KINGSTON AVE	840	THEFT	FINANCIAL ID THEFT: OVER $300	
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/GARAG
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	GROCER
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

Problem 3

How many crimes involve an arrest?

In [5]:

```
%sql select count(*) as Arrest_crimes from CHICAGO_CRIME_DATA where ARREST = True
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[5]:

arrest_crimes
163

Problem 4

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

In [8]:

```
# The SELECT DISTINCT statement is used to return only distinct (different) values.
# but in this problem 4, it asks to filter the crimes based on the location (place)
%sql select DISTINCT PRIMARY_TYPE from CHICAGO_CRIME_DATA WHERE LOCATION_DESCRIPTION
 ='GAS STATION';
# or
# %sql select DISTINCT (PRIMARY_TYPE) from CHICAGO_CRIME_DATA WHERE LOCATION_DESCRIPTIO
N ='GAS STATION';
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[8]:

primary_type
CRIMINAL TRESPASS
NARCOTICS
ROBBERY
THEFT

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

Problem 5

In the CENSUS_DATA table list all Community Areas whose names start with the letter ‘B’.

In [9]:

```
%sql select COMMUNITY_AREA_NAME from CENSUS_DATA WHERE COMMUNITY_AREA_NAME LIKE 'B%';
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[9]:

community_area_name
Belmont Cragin
Burnside
Brighton Park
Bridgeport
Beverly

Problem 6

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

In [10]:

```
%sql select S.NAME_OF_SCHOOL,C.COMMUNITY_AREA_NUMBER,C.COMMUNITY_AREA_NAME,S.healthy_sc
hool_certified from CENSUS_DATA as  C \
LEFT OUTER JOIN CHICAGO_PUBLIC_SCHOOL as S \
on UPPER(C.COMMUNITY_AREA_NAME) = UPPER(S.community_area_name) \
where C.COMMUNITY_AREA_NUMBER between 10 and 15 AND \
S.healthy_school_certified = 'Yes';
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[10]:

name_of_school	community_area_number	community_area_name	healthy_school_certified
Rufus M Hitch Elementary School	10	Norwood Park	Yes

Problem 7

What is the average school Safety Score?

In [11]:

```
%sql select AVG(safety_score) as AVERAGE_SCRORE from CHICAGO_PUBLIC_SCHOOL;
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[11]:

average_scrore
49.504873

Problem 8

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

In [24]:

```
# GROUP BY statement is used for grouping the data and it mainly uses with aggregate fu
nctions.
# we here use GROUP BY statement and AVG() function at the same time
# to calculate the average COLLEGE ENROLLMENT of each COMMUNITY_AREA_NAME
# always when ask to retreive specific row from one column by the average value in othe
r column
# we you group by function
%sql select COMMUNITY_AREA_NAME,AVG(COLLEGE_ENROLLMENT) as COLLEGE_ENROLLMENT_AVG \
from CHICAGO_PUBLIC_SCHOOL GROUP BY COMMUNITY_AREA_NAME order by COLLEGE_ENROLLMENT_AVG
desc LIMIT 5;
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[24]:

community_area_name	college_enrollment_avg
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?

In [30]:

```
#  sub-query mean two select statements
%sql select COMMUNITY_AREA_NAME,SAFETY_SCORE from CHICAGO_PUBLIC_SCHOOL \
WHERE SAFETY_SCORE = (SELECT MIN(SAFETY_SCORE) FROM CHICAGO_PUBLIC_SCHOOL);
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[30]:

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 [51]:

```
%sql select  COMMUNITY_AREA_NAME,per_capita_income \
from CENSUS_DATA \
where community_area_number = (select community_area_number \
from CHICAGO_PUBLIC_SCHOOL where safety_score = 1);

# or by using join operator as follow
%sql select Per_Capita_Income \
from CENSUS_DATA as CD full join CHICAGO_PUBLIC_SCHOOL as CPS \
on CD.COMMUNITY_AREA_NUMBER = CPS.COMMUNITY_AREA_NUMBER where safety_score = 1;
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[51]:

per_capita_income
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