Working with a real world data-set using SQL and Python

Estaimted time needed: **30** minutes

Objectives

After complting this lab you will be able to:

- Understand the dataset for Chicago Public School level performance
- Store the dataset in an Db2 database on IBM Cloud instance
- Retrieve metadata about tables and columns and query data from mixed case columns
- Solve example problems to practice your SQL skills including using built-in database functions

Chicago Public Schools - Progress Report Cards (2011-2012)

The city of Chicago released a dataset showing all school level performance data used to create School Report Cards for the 2011-2012 school year. The dataset is available from the Chicago Data Portal: 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?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork22-2022-01-01&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.Newsletter.M12345678&cvo_campaign=

This dataset includes a large number of metrics. Start by familiarizing yourself with the types of metrics in the database: https://data.cityofchicago.org/api/assets/AAD41A13-BE8A-4E67-B1F5-86E711E09D5F?download=true (https://data.cityofchicago.org/api/assets/AAD41A13-BE8A-4E67-B1F5-86E711E09D5F?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork22-2022-01-01&download=true&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.Newsletter.M12345678&cvo_campaign=

NOTE:

Do not download the dataset directly from City of Chicago portal. Instead download a static copy which is a more database friendly version from this link (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoPublicSchools.csv).

NOTE:

For the learners who are encountering issues with loading from .csv in DB2 on Firefox, you can download the .txt files and load the data with those: link (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoPublicSchools.txt).

Now review some of its contents.

Store the dataset in a Table

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 the previous lab, 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 CHICAGO PUBLIC SCHOOLS dataset and load the dataset into a new table called SCHOOLS.


(https://cognitiveclass.ai/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork22-2022-01-01)

Connect to the database

Let us now load the ipython-sql extension and establish a connection with the database

The following modules are pre-installed in the Skills Network Labs environment. However if you run this notebook commands in a different Jupyter environment (e.g. Watson Studio or Ananconda) you may need to install these libraries by removing the `#` sign before `!pip` in the code cell below.

In [44]:
```
# These libraries are pre-installed in SN Labs. If running in another environment please uncomment lines below to
# !pip install --force-reinstall ibm_db==3.1.0 ibm_db_sa==0.3.3
# Ensure we don't load_ext with sqlalchemy>=1.4 (incompadible)
# !pip uninstall sqlalchemy==1.4 -y && pip install sqlalchemy==1.3.24
# !pip install ipython-sql
```

In [45]:
```
%load_ext sql
```

```
The sql extension is already loaded. To reload it, use:
  %reload_ext sql
```

In [46]:
```
# Enter the connection string for your Db2 on Cloud database instance below
# %sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name?security=SSL
%sql ibm_db_sa://svw77997:RB1ijcrd3WAzHN9V@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.app
%sql ibm_db_sa://
```

Out[46]:
```
'Connected: svw77997@BLUDB'
```

Query the database system catalog to retrieve table metadata

You can verify that the table creation was successful by retrieving the list of all tables in your schema and checking whether the SCHOOLS table was created

In [47]:
```
# type in your query to retrieve list of all tables in the database for your db2 schema (username)
%sql SELECT TABSCHEMA, TABNAME , CREATE_TIME FROM SYSCAT.TABLES WHERE TABSCHEMA='SVW77997';
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3
0119/BLUDB
Done.
```

Out[47]:

tabschema	tabname	create_time
SVW77997	INSTRUCTOR	2023-04-19 18:19:24.274735
SVW77997	INTERNATIONAL_STUDENT_TEST_SCORES	2023-04-19 18:48:16.687849
SVW77997	CHICAGO_SOCIOECONOMIC_DATA	2023-04-19 20:01:33.651798
SVW77997	SCHOOLS	2023-04-20 02:10:03.058189

Double-click **here** for a hint

Double-click **here** for the solution.

Query the database system catalog to retrieve column metadata

The SCHOOLS table contains a large number of columns. How many columns does this table have?

In [48]:
```
# type in your query to retrieve the number of columns in the SCHOOLS table
%sql SELECT * FROM SYSCAT.COLUMNS WHERE TABNAME='SCHOOLS';
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:30119/BLUDB
Done.
```

Out[48]:

tabschema	tabname	colname	colno	typeschema	typename	length	scale	typestri
SVW77997	SCHOOLS	SCHOOL_ID	0	SYSIBM	INTEGER	4	0	
SVW77997	SCHOOLS	NAME_OF_SCHOOL	1	SYSIBM	VARCHAR	64	0	C
SVW77997	SCHOOLS	ELEMENTARY__MIDDLE__OR_HIGH_SCHOOL	2	SYSIBM	VARCHAR	2	0	C
SVW77997	SCHOOLS	STREET_ADDRESS	3	SYSIBM	VARCHAR	29	0	C
SVW77997	SCHOOLS	CITY	4	SYSIBM	VARCHAR	7	0	C
SVW77997	SCHOOLS	STATE	5	SYSIBM	VARCHAR	2	0	C
SVW77997	SCHOOLS	ZIP_CODE	6	SYSIBM	INTEGER	4	0	
SVW77997	SCHOOLS	PHONE_NUMBER	7	SYSIBM	VARCHAR	14	0	C

Double-click **here** for a hint

Double-click **here** for the solution.

Now retrieve the the list of columns in SCHOOLS table and their column type (datatype) and length.

In [49]:
```
# type in your query to retrieve all column names in the SCHOOLS table along with their datatypes and length
%sql SELECT colname,typename, length FROM SYSCAT.COLUMNS WHERE TABNAME='SCHOOLS';
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:30119/BLUDB
Done.
```

Out[49]:

colname	typename	length
SCHOOL_ID	INTEGER	4
NAME_OF_SCHOOL	VARCHAR	64
ELEMENTARY__MIDDLE__OR_HIGH_SCHOOL	VARCHAR	2
STREET_ADDRESS	VARCHAR	29
CITY	VARCHAR	7
STATE	VARCHAR	2
ZIP_CODE	INTEGER	4
PHONE_NUMBER	VARCHAR	14
LINK	VARCHAR	78

Double-click **here** for the solution.

Questions

1. Is the column name for the "SCHOOL ID" attribute in upper or mixed case?
2. What is the name of "Community Area Name" column in your table? Does it have spaces?
3. Are there any columns in whose names the spaces and paranthesis (round brackets) have been replaced by the underscore character "_"?

Problems

Problem 1

How many Elementary Schools are in the dataset?

In [50]:
```
%sql SELECT COUNT(*) FROM SCHOOLS WHERE "ELEMENTARY__MIDDLE__OR_HIGH_SCHOOL"= 'ES';
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[50]:

1
462

Double-click **here** for a hint

Double-click **here** for another hint

Double-click **here** for the solution.

Problem 2

What is the highest Safety Score?

In [59]:
```
%sql SELECT MAX(SAFETY_SCORE) AS MAX_SAFETY_SCORE FROM SCHOOLS;
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[59]:

max_safety_score
99

Double-click **here** for a hint

Double-click **here** for the solution.

Problem 3

Which schools have highest Safety Score?

In [52]:
```
%sql select Name_of_School, Safety_Score from SCHOOLS where Safety_Score = 99
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[52]:

name_of_school	safety_score
Abraham Lincoln Elementary School	99
Alexander Graham Bell Elementary School	99
Annie Keller Elementary Gifted Magnet School	99
Augustus H Burley Elementary School	99
Edgar Allan Poe Elementary Classical School	99
Edgebrook Elementary School	99
Ellen Mitchell Elementary School	99
James E McDade Elementary Classical School	99
James G Blaine Elementary School	99
LaSalle Elementary Language Academy	99
Mary E Courtenay Elementary Language Arts Center	99
Northside College Preparatory High School	99
Northside Learning Center High School	99
Norwood Park Elementary School	99
Oriole Park Elementary School	99
Sauganash Elementary School	99
Stephen Decatur Classical Elementary School	99
Talman Elementary School	99
Wildwood Elementary School	99

Double-click **here** for the solution.

Problem 4

What are the top 10 schools with the highest "Average Student Attendance"?

In [56]:
```
%sql SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE FROM SCHOOLS ORDER BY AVERAGE_STUDENT_ATTENDANCE DESC LIMIT
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[56]:

name_of_school	average_student_attendance
Velma F Thomas Early Childhood Center	None
John Charles Haines Elementary School	98.40%
James Ward Elementary School	97.80%
Edgar Allan Poe Elementary Classical School	97.60%
Rachel Carson Elementary School	97.60%
Orozco Fine Arts & Sciences Elementary School	97.60%
Annie Keller Elementary Gifted Magnet School	97.50%
Andrew Jackson Elementary Language Academy	97.40%
Lenart Elementary Regional Gifted Center	97.40%
Disney II Magnet School	97.30%

Double-click **here** for the solution.

Problem 5

Retrieve the list of 5 Schools with the lowest Average Student Attendance sorted in ascending order based on attendance

In [65]:
```
%sql SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE FROM SCHOOLS ORDER BY \
    AVERAGE_STUDENT_ATTENDANCE  \
    fetch first 5 rows only
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[65]:

name_of_school	average_student_attendance
Richard T Crane Technical Preparatory High School	57.90%
Barbara Vick Early Childhood & Family Center	60.90%
Dyett High School	62.50%
Wendell Phillips Academy High School	63.00%
Orr Academy High School	66.30%

Double-click **here** for the solution.

Problem 6

Now remove the '%' sign from the above result set for Average Student Attendance column

In [66]:
```
%sql SELECT NAME_OF_SCHOOL, REPLACE(AVERAGE_STUDENT_ATTENDANCE, '%', '') \
    FROM SCHOOLS ORDER BY AVERAGE_STUDENT_ATTENDANCE\
    FETCH FIRST 5 ROWS ONLY
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[66]:

name_of_school	2
Richard T Crane Technical Preparatory High School	57.90
Barbara Vick Early Childhood & Family Center	60.90
Dyett High School	62.50
Wendell Phillips Academy High School	63.00
Orr Academy High School	66.30

Double-click **here** for a hint

Double-click **here** for the solution.

Problem 7

Which Schools have Average Student Attendance lower than 70%?

In [69]:
```
%sql SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE  FROM SCHOOLS WHERE CAST (REPLACE(AVERAGE_STUDENT_ATTENDANC
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3
0119/BLUDB
Done.
```

Out[69]:

name_of_school	average_student_attendance
Barbara Vick Early Childhood & Family Center	60.90%
Chicago Vocational Career Academy High School	68.80%
Dyett High School	62.50%
Manley Career Academy High School	66.80%
Orr Academy High School	66.30%
Richard T Crane Technical Preparatory High School	57.90%
Roberto Clemente Community Academy High School	69.60%
Wendell Phillips Academy High School	63.00%

Double-click **here** for a hint

Double-click **here** for another hint

Double-click **here** for the solution.

Problem 8

Get the total College Enrollment for each Community Area

In [70]:
```
%sql SELECT COMMUNITY_AREA_NAME, SUM(COLLEGE_ENROLLMENT) AS TOTAL_ENROLLMENT FROM SCHOOLS\
    GROUP BY COMMUNITY_AREA_NAME
```

LAKE VIEW	7055
LINCOLN PARK	5615
LINCOLN SQUARE	4132
LOGAN SQUARE	7351
LOOP	871
LOWER WEST SIDE	7257
MCKINLEY PARK	1552
MONTCLARE	1317
MORGAN PARK	3271
MOUNT GREENWOOD	2091
NEAR NORTH SIDE	3362
NEAR SOUTH SIDE	1378
NEAR WEST SIDE	7975

Double-click **here** for a hint

Double-click **here** for another hint

Double-click **here** for the solution.

Problem 9

Get the 5 Community Areas with the least total College Enrollment sorted in ascending order

In [71]:
```
%sql SELECT COMMUNITY_AREA_NAME, SUM(COLLEGE_ENROLLMENT) AS TOTAL_ENROLLMENT FROM SCHOOLS \
    GROUP BY COMMUNITY_AREA_NAME ORDER BY TOTAL_ENROLLMENT ASC FETCH FIRST 5 ROWS ONLY
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[71]:

community_area_name	total_enrollment
OAKLAND	140
FULLER PARK	531
BURNSIDE	549
OHARE	786
LOOP	871

Double-click **here** for a hint

Double-click **here** for the solution.

Problem 10

List 5 schools with lowest safety score.

In [74]:
```
%sql SELECT NAME_OF_SCHOOL, SAFETY_SCORE FROM SCHOOLS ORDER BY SAFETY_SCORE ASC\
    FETCH FIRST 5 ROWS ONLY
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[74]:

name_of_school	safety_score
Edmond Burke Elementary School	1
Luke O'Toole Elementary School	5
George W Tilton Elementary School	6
Foster Park Elementary School	11
Emil G Hirsch Metropolitan High School	13

Double-click **here** for the solution.

Problem 11

Get the hardship index for the community area which has College Enrollment of 4368

In [76]:
```
%sql SELECT HARDSHIP_INDEX FROM CHICAGO_SOCIOECONOMIC_DATA CD, SCHOOLS CPS\
    WHERE CD.ca=CPS.COMMUNITY_AREA_NUMBER AND COLLEGE_ENROLLMENT=4368;
```

```
 * ibm_db_sa://svw77997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

Out[76]:

hardship_index
6.0

Double-click **here** for the solution.

Problem 12

Get the hardship index for the community area which has the school with the highest enrollment.

In [ ]:
```
%sql SL
```

Double-click **here** for the solution.

Summary

In this lab you learned how to work with a real word dataset using SQL and Python. You learned how to query columns with spaces or special characters in their names and with mixed case names. You also used built in database functions and practiced how to sort, limit, and order result sets, as well as used sub-queries and worked with multiple tables. ¶

Author

Rav Ahuja (https://www.linkedin.com/in/ravahuja/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork22-2022-01-01)

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-07-09	2.4	Malika	Updated connection string
2021-05-19	2.3	Lakshmi Holla	Updated question
2021-04-20	2.2	Malika	Added the libraries
2020-11-27	2.1	Sannareddy Ramesh	Modified data sets and added new problems
2020-08-28	2.0	Lavanya	Moved lab to course repo in GitLab