**Working with a real world data-set using SQL and Python[¶](https://jupyterlab-3-labs-prod-jupyterlab-us-east-1.labs.cognitiveclass.ai/user/akulasamson/lab/tree/labs/DB0201EN/DB0201EN-Week4-1-1-RealDataPractice-v5%20(3).ipynb" \l "Working-with-a-real-world-data-set-using-SQL-and-Python" \t "_self)**

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-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01&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=000026UJ)

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-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01&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=000026UJ)

**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?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01).

Now review some of its contents.

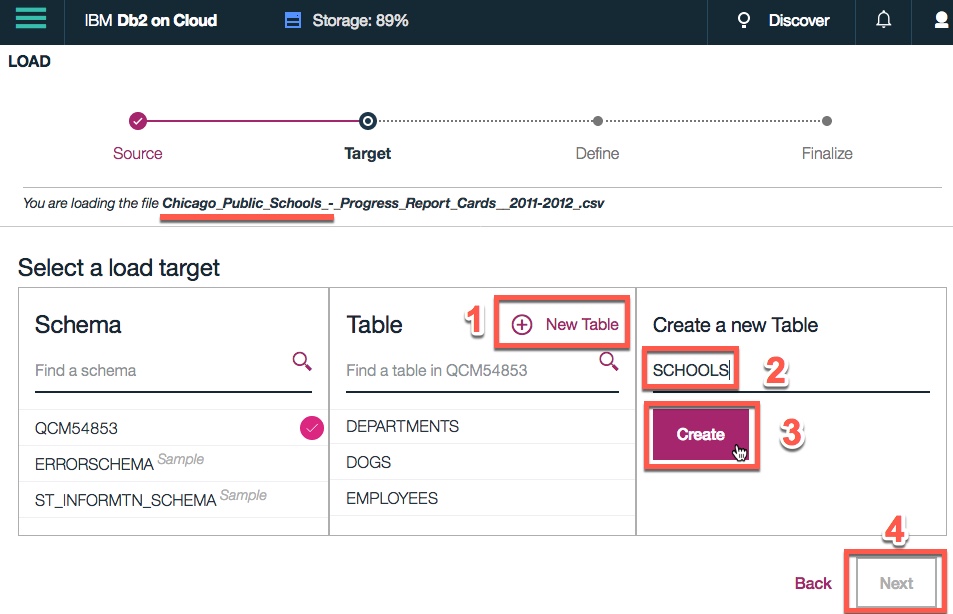
**Store the dataset in a Table**

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-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01)

**Connect to the database**

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

[1]:



**!**pip install sqlalchemy**==**1.3.9

**!**pip install ibm\_db\_sa

Requirement already satisfied: sqlalchemy==1.3.9 in /home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (1.3.9)

Requirement already satisfied: ibm\_db\_sa in /home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (0.3.3)

Requirement already satisfied: sqlalchemy>=0.7.3 in /home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from ibm\_db\_sa) (1.3.9)

[2]:



**%**load\_ext sql

[3]:



*# 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:**//**kvn77701:INISp92TCJkh15dQ**@**54a2f15b**-**5c0f**-**46df**-**8954**-**7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733**/**BLUDB**?**security**=**SSL

[3]:

'Connected: kvn77701@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**

[ ]:



*# type in your query to retrieve list of all tables in the database for your db2 schema (username)*

​

Double-click **here** for a hint



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

​

*<!-- Solution:*

​

*%sql select TABSCHEMA, TABNAME, CREATE\_TIME from SYSCAT.TABLES where TABSCHEMA='YOUR-DB2-USERNAME'*

​

*or, you can retrieve list of all tables where the schema name is not one of the system created ones:*

​

*%sql select TABSCHEMA, TABNAME, CREATE\_TIME from SYSCAT.TABLES \*

*where TABSCHEMA not in ('SYSIBM', 'SYSCAT', 'SYSSTAT', 'SYSIBMADM', 'SYSTOOLS', 'SYSPUBLIC')*

*or, just query for a specifc table that you want to verify exists in the database*

*%sql select \* from SYSCAT.TABLES where TABNAME = 'SCHOOLS'*

​

*-->*

​

**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?**

[ ]:



*# type in your query to retrieve the number of columns in the SCHOOLS table*

​

Double-click **here** for a hint



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

​

*<!-- Solution:*

​

*%sql select count(\*) from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'*

​

*-->*

​

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

[ ]:



*# type in your query to retrieve all column names in the SCHOOLS table along with their datatypes and length*

​



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

​

*<!-- Solution:*

​

*%sql select COLNAME, TYPENAME, LENGTH from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'*

​

*or*

​

*%sql select distinct(NAME), COLTYPE, LENGTH from SYSIBM.SYSCOLUMNS where TBNAME = 'SCHOOLS'*

​

*-->*

​

**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?**

[ ]:



​

Double-click **here** for a hint

Double-click **here** for another hint



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

​

*<!-- Solution:*

​

*%sql select count(\*) from SCHOOLS where school\_type = 'ES'*

​

*Correct answer: 462*

​

*-->*

​

**Problem 2**

**What is the highest Safety Score?**

[ ]:



​

Double-click **here** for a hint



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

​

*<!-- Hint:*

​

*%sql select MAX(Safety\_Score) AS MAX\_SAFETY\_SCORE from SCHOOLS*

​

*Correct answer: 99*

*-->*

​

**Problem 3**

**Which schools have highest Safety Score?**

[ ]:



​



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

​

*<!-- Solution:*

*In the previous problem we found out that the highest Safety Score is 99, so we can use that as an input in the where clause:*

​

*%sql select Name\_of\_School, Safety\_Score from SCHOOLS where Safety\_Score = 99*

​

*or, a better way:*

​

*%sql select Name\_of\_School, Safety\_Score from SCHOOLS where \*

*Safety\_Score= (select MAX(Safety\_Score) from SCHOOLS)*

​

​

*Correct answer: several schools with with Safety Score of 99.*

*-->*

​

**Problem 4**

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

[ ]:



​



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

​

*<!-- Solution:*

​

*%sql select Name\_of\_School, Average\_Student\_Attendance from SCHOOLS \*

*order by Average\_Student\_Attendance desc nulls last limit 10*

​

*-->*

​

**Problem 5**

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

**Did you know? IBM Watson Studio lets you build and deploy an AI solution, using the best of open source and IBM software and giving your team a single environment to work in.**[**Learn more here.**](https://cocl.us/ibm_watson_studio_infobox)

[ ]:



​



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

​

*<!-- Solution:*

​

*%sql SELECT Name\_of\_School, Average\_Student\_Attendance \*

*from SCHOOLS \*

*order by Average\_Student\_Attendance \*

*fetch first 5 rows only*

​

*-->*

​

**Problem 6**

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

[ ]:



​

Double-click **here** for a hint



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

​

*<!-- Hint:*

​

*%sql SELECT Name\_of\_School, REPLACE(Average\_Student\_Attendance, '%', '') \*

*from SCHOOLS \*

*order by Average\_Student\_Attendance \*

*fetch first 5 rows only*

​

*-->*

​

**Problem 7**

**Which Schools have Average Student Attendance lower than 70%?**

[ ]:



​

Double-click **here** for a hint

Double-click **here** for another hint



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

​

*<!-- Solution:*

​

*%sql SELECT Name\_of\_School, Average\_Student\_Attendance \*

*from SCHOOLS \*

*where CAST ( REPLACE(Average\_Student\_Attendance, '%', '') AS DOUBLE ) < 70 \*

*order by Average\_Student\_Attendance*

*or,*

​

*%sql SELECT Name\_of\_School, Average\_Student\_Attendance \*

*from SCHOOLS \*

*where DECIMAL ( REPLACE(Average\_Student\_Attendance, '%', '') ) < 70 \*

*order by Average\_Student\_Attendance*

​

*-->*

​

**Problem 8**

**Get the total College Enrollment for each Community Area**

[ ]:



​

Double-click **here** for a hint

Double-click **here** for another hint



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

​

*<!-- Solution:*

​

*%sql select Community\_Area\_Name, sum(College\_Enrollment) AS TOTAL\_ENROLLMENT \*

*from SCHOOLS \*

*group by Community\_Area\_Name*

​

*-->*

​

**Problem 9**

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

[ ]:



​



Double-click **\*\*here\*\*** for a hint

​

*<!--*

*Order the previous query and limit the number of rows you fetch*

*-->*

​



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

​

*<!-- Solution:*

​

*%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*

​

*-->*

​

**Problem 10**

**List 5 schools with lowest safety score.**

[ ]:



​



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

​

*<!-- Solution:*

​

*%sql SELECT name\_of\_school, safety\_score*

*FROM schools*

*ORDER BY safety\_score*

*LIMIT 5*

*-->*

​

**Problem 11**

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

[ ]:



​



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

​

*<!-- Solution:*

*NOTE: For this solution to work the CHICAGO\_SOCIOECONOMIC\_DATA table*

*as created in the last lab of Week 3 should already exist*

​

*%%sql*

*select hardship\_index*

*from chicago\_socioeconomic\_data CD, schools CPS*

*where CD.ca = CPS.community\_area\_number*

*and college\_enrollment = 4368*

​

*-->*

​

**Problem 12**

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

[ ]:



​



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

​

*<!-- Solution:*

*NOTE: For this solution to work the CHICAGO\_SOCIOECONOMIC\_DATA table*

*as created in the last lab of Week 3 should already exist*

​

*%sql select ca, community\_area\_name, hardship\_index from chicago\_socioeconomic\_data \*

*where ca in \*

*( select community\_area\_number from schools order by college\_enrollment desc limit 1 )*

​

*-->*

​

**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.**