

# 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>

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>

**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](#).

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**.

## LOAD



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	

BackNext

## Connect to the database

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

```
In [1]: %load_ext sql
```

```
In [2]: # 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
%sql ibm_db_sa://vks14514:7vbht2mg%2Bfsr1tvj@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.ibmcloud.net:50000/BLUDB
```

```
Out[2]: 'Connected: vks14514@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 [8]: # type in your query to retrieve list of all tables in the database for your db2 schema (username)
%sql select * from SYSCAT.TABLES where TABNAME = 'SCHOOLS'
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.ibmcloud.net:50000/BLUDB
Done.
```

```
Out[8]: tabschema  tabname    owner  ownertype  TYPE  status  base_tabschema  base_tabname  rowtypeschema  rowtypename  create_time  ...
        VKS14514  SCHOOLS   VKS14514      U      T      N          None          None          None          None      2020-10-15  2
        13:06:17.149727  13:06:17.149727
```

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 [14]: # type in your query to retrieve the number of columns in the SCHOOLS table
%sql select count(*) from syscat.columns where TABNAME = 'SCHOOLS'
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.ibmcloud.net:50000/BLUDB
Done.
```

```
Out[14]: 1
```

78

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 [15]: # type in your query to retrieve all column names in the SCHOOLS table along with their datatypes and length
%sql select distinct(NAME), COLTYPE, LENGTH from SYSIBM.SYSCOLUMNS where TBNAME = 'SCHOOLS'
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

```
Out[15]:
```

	name	coltype	length
	10th Grade PLAN (2009)	VARCHAR	4
	10th Grade PLAN (2010)	VARCHAR	4
	11th Grade Average ACT (2011)	VARCHAR	4
	9th Grade EXPLORE (2009)	VARCHAR	4
	9th Grade EXPLORE (2010)	VARCHAR	4
	AVERAGE_STUDENT_ATTENDANCE	VARCHAR	6
	Adequate_Yearly_Progress_Made_	VARCHAR	3
	Average_Teacher_Attendance	VARCHAR	6
	COLLEGE_ENROLLMENT	SMALLINT	2
	COMMUNITY_AREA_NAME	VARCHAR	22
	COMMUNITY_AREA_NUMBER	SMALLINT	2
	CPS_Performance_Policy_Level	VARCHAR	15
	CPS_Performance_Policy_Status	VARCHAR	16
	City	VARCHAR	7
	Collaborative_Name	VARCHAR	34
	College_Eligibility__	VARCHAR	4
	College_Enrollment_Rate__	VARCHAR	4
	Elementary, Middle, or High School	VARCHAR	2

Environment_Icon	VARCHAR	11
Environment_Score	SMALLINT	2
Family_Involvement_Icon	VARCHAR	11
Family_Involvement_Score	VARCHAR	3
Freshman_on_Track_Rate__	VARCHAR	4
General_Services_Route	SMALLINT	2
Gr3_5_Grade_Level_Math__	VARCHAR	4
Gr3_5_Grade_Level_Read__	VARCHAR	4
Gr3_5_Keep_Pace_Math__	VARCHAR	4
Gr3_5_Keep_Pace_Read__	VARCHAR	4
Gr6_8_Grade_Level_Math__	VARCHAR	4
Gr6_8_Grade_Level_Read__	VARCHAR	4
Gr6_8_Keep_Pace_Math__	VARCHAR	4
Gr6_8_Keep_Pace_Read__	VARCHAR	4
Gr_8_Explore_Math__	VARCHAR	4
Gr_8_Explore_Read__	VARCHAR	4
Graduation_Rate__	VARCHAR	4
HEALTHY_SCHOOL_CERTIFIED	VARCHAR	3
ISAT_Exceeding_Math__	DECIMAL	4
ISAT_Exceeding_Reading__	DECIMAL	4
ISAT_Value_Add_Color_Math	VARCHAR	6
ISAT_Value_Add_Color_Read	VARCHAR	6
ISAT_Value_Add_Math	DECIMAL	3
ISAT_Value_Add_Read	DECIMAL	3
Individualized_Education_Program_Compliance_Rate	VARCHAR	7
Instruction_Icon	VARCHAR	11

Instruction_Score	SMALLINT	2
Latitude	DECIMAL	18
Leaders_Icon	VARCHAR	11
Leaders_Score	VARCHAR	3
Link	VARCHAR	78
Location	VARCHAR	27
Longitude	DECIMAL	18
NAME_OF_SCHOOL	VARCHAR	65
Net_Change_EXPLORE_and_PLAN	VARCHAR	3
Net_Change_PLAN_and_ACT	VARCHAR	3
Network_Manager	VARCHAR	40
Parent_Engagement_Icon	VARCHAR	7
Parent_Engagement_Score	VARCHAR	3
Parent_Environment_Icon	VARCHAR	7
Parent_Environment_Score	VARCHAR	3
Phone_Number	VARCHAR	14
Pk_2_Literacy__	VARCHAR	4
Pk_2_Math__	VARCHAR	4
Police_District	SMALLINT	2
Rate_of_Misconducts__per_100_students_	DECIMAL	5
SAFETY_SCORE	SMALLINT	2
Safety_Icon	VARCHAR	11
School_ID	INTEGER	4
State	VARCHAR	2
Street_Address	VARCHAR	30
Students_Passing__Algebra__	VARCHAR	4

Students_Taking__Algebra__	VARCHAR	4
Teachers_Icon	VARCHAR	11
Teachers_Score	VARCHAR	3
Track_Schedule	VARCHAR	12
Ward	SMALLINT	2
X_COORDINATE	DECIMAL	13
Y_COORDINATE	DECIMAL	13
ZIP_Code	INTEGER	4

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 [23]:

```
%sql
select *
from SCH00LS
limit 5
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[23]:

```
School_ID  name_of_school  Elementary, Middle, or High School  Street_Address  City  State  ZIP_Code  Phone_Number
```



610038	Abraham Lincoln Elementary School	ES	615 W Kemper Pl	Chicago	IL	60614	(773) 534-5720	http://schoolreports.cps.edu/SchoolProgressReport_Er
610281	Adam Clayton Powell Paideia Community Academy Elementary School	ES	7511 S South Shore Dr	Chicago	IL	60649	(773) 535-6650	http://schoolreports.cps.edu/SchoolProgressReport_Er
610185	Adlai E Stevenson Elementary School	ES	8010 S Kostner Ave	Chicago	IL	60652	(773) 535-2280	http://schoolreports.cps.edu/SchoolProgressReport_Er
609993	Agustin Lara Elementary Academy	ES	4619 S Wolcott Ave	Chicago	IL	60609	(773) 535-4389	http://schoolreports.cps.edu/SchoolProgressReport_Er
610513	Air Force Academy High School	HS	3630 S Wells St	Chicago	IL	60609	(773) 535-1590	http://schoolreports.cps.edu/SchoolProgressReport_Er

```
In [20]: %%sql
select count(*)
from SCH00LS
where "Elementary, Middle, or High School" = 'ES'

-- The Correct ans is 462
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

```
Out[20]: 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 [27]:

```
%%sql
```

```
select max(safety_score)
from SCH00LS
--The answer is 99
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[27]:

```
1
99
```

Double-click [here](#) for a hint

Double-click [here](#) for the solution.

## Problem 3

Which schools have highest Safety Score?

In [28]:

```
%%sql
```

```
select name_of_school
from SCH00LS
where safety_score = (select max(safety_score)
from SCH00LS)
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[28]:

```
name_of_school
```

```
Abraham Lincoln Elementary School
```

```
Alexander Graham Bell Elementary School
```

```
Annie Keller Elementary Gifted Magnet School
```

```
Augustus H Burley Elementary School
```

```
Edgar Allan Poe Elementary Classical School
```

Edgebrook Elementary School
Ellen Mitchell Elementary School
James E McDade Elementary Classical School
James G Blaine Elementary School
LaSalle Elementary Language Academy
Mary E Courtenay Elementary Language Arts Center
Northside College Preparatory High School
Northside Learning Center High School
Norwood Park Elementary School
Oriole Park Elementary School
Sauganash Elementary School
Stephen Decatur Classical Elementary School
Talman Elementary School
Wildwood Elementary School

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

## Problem 4

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

```
In [31]: %%sql
--select *
--from schools
--order by average_student_attendance desc
--limit 10
-- The above query is also correct but it includes the schools with no data like NULL
select *
from schools
order by average_student_attendance desc nulls last
limit 10
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[31]:

School_ID	name_of_school	Elementary, Middle, or High School	Street_Address	City	State	ZIP_Code	Phone_Number
-----------	----------------	------------------------------------	----------------	------	-------	----------	--------------

609959	John Charles Haines Elementary School	ES	247 W 23rd Pl	Chicago	IL	60616	(773) 534-9200	http://schoolreports.cps.edu/SchoolProgressReport_Er
610217	James Ward Elementary School	ES	2701 S Shields Ave	Chicago	IL	60616	(773) 534-9050	http://schoolreports.cps.edu/SchoolProgressReport_Er
610132	Edgar Allan Poe Elementary Classical School	ES	10538 S Langley Ave	Chicago	IL	60628	(773) 535-5525	http://schoolreports.cps.edu/SchoolProgressReport_Er
609842	Rachel Carson Elementary School	ES	5516 S Maplewood Ave	Chicago	IL	60629	(773) 535-9222	http://schoolreports.cps.edu/SchoolProgressReport_Er
610329	Orozco Fine Arts & Sciences Elementary School	ES	1940 W 18th St	Chicago	IL	60608	(773) 534-7215	http://schoolreports.cps.edu/SchoolProgressReport_Er
610084	Annie Keller Elementary Gifted Magnet School	ES	3020 W 108th St	Chicago	IL	60655	(773) 535-2636	http://schoolreports.cps.edu/SchoolProgressReport_Er
610060	Andrew Jackson Elementary Language Academy	ES	1340 W Harrison St	Chicago	IL	60607	(773) 534-7000	http://schoolreports.cps.edu/SchoolProgressReport_Er
610298	Lenart Elementary Regional Gifted Center	ES	8101 S LaSalle St	Chicago	IL	60620	(773) 535-0040	http://schoolreports.cps.edu/SchoolProgressReport_Er
610515	Disney II Magnet School	ES	3815 N Kedvale Ave	Chicago	IL	60641	(773) 534-3750	http://schoolreports.cps.edu/SchoolProgressReport_Er
610207	John H Vanderpoel	ES	9510 S Prospect Ave	Chicago	IL	60643	(773) 535-2690	http://schoolreports.cps.edu/SchoolProgressReport_Er

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 [36]: %%sql
select *
from (select * from schools order by average_student_attendance asc) as SCH
limit 5
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[36]:

School_ID	name_of_school	Elementary, Middle, or High School	Street_Address	City	State	ZIP_Code	Phone_Number	
609702	Richard T Crane Technical Preparatory High School	HS	2245 W Jackson Blvd	Chicago	IL	60612	(773) 534-7550	http://schoolreports.cps.edu/SchoolProgressReport_Er
609871	Barbara Vick Early Childhood & Family Center	ES	2554 W 113th St	Chicago	IL	60655	(773) 535-2671	http://schoolreports.cps.edu/SchoolProgressReport_Er
609736	Dyett High School	HS	555 E 51st St	Chicago	IL	60615	(773) 535-1825	http://schoolreports.cps.edu/SchoolProgressReport_Er
609727	Wendell Phillips Academy High School	HS	244 E Pershing Rd	Chicago	IL	60653	(773) 535-1603	http://schoolreports.cps.edu/SchoolProgressReport_Er
610389	Orr Academy High School	HS	730 N Pulaski Rd	Chicago	IL	60624	(773) 534-6500	http://schoolreports.cps.edu/SchoolProgressReport_Er

Double-click [here](#) for the solution.

## Problem 6

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

In [35]: 

```
%%sql
SELECT Name_of_School, REPLACE(Average_Student_Attendance, '%', '') -- REPLACE function replaces the % sign with null
from SCH00LS
order by Average_Student_Attendance
limit 5
```

\* ibm\_db\_sa://vks14514:\*\*\*@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB  
Done.

Out[35]:

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 [39]: 

```
%%sql
--DECIMAL(REPLACE(Average_Student_Attendance, '%', ''))
select name_of_school
from schools
where DECIMAL(REPLACE(Average_Student_Attendance, '%', '')) < 70
```

\* ibm\_db\_sa://vks14514:\*\*\*@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB  
Done.

Out[39]:

name_of_school
Barbara Vick Early Childhood & Family Center

Chicago Vocational Career Academy High School

Dyett High School

Manley Career Academy High School

Orr Academy High School

Richard T Crane Technical Preparatory High School

Roberto Clemente Community Academy High School

Wendell Phillips Academy High School

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 [42]:

```
%%sql
select SUM(college_enrollment)
from schools
group by community_area_name
limit 5
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[42]:

```
1
6864
4823
1458
6483
4175
```

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 [43]: %%sql
select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT
from SCHOOLS
group by Community_Area_Name
order by TOTAL_ENROLLMENT asc
limit 5
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

```
Out[43]: 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

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

```
In [44]: %%sql
select hardship_index
from chicago_socioeconomic_data CD
where CD.ca in (select CPS.community_area_number from schools CPS where college_enrollment = 4368)
```

```
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
```



Done.

Out[44]: **hardship\_index**

6.0

Double-click [here](#) for the solution.

## Problem 11

Get the hardship index for the community area which has the highest value for College Enrollment

```
In [46]: %%sql
select hardship_index
from chicago_socioeconomic_data CD
where CD.ca in (select CPS.community_area_number from schools CPS where college_enrollment = (select max(CPS.college_
* ibm_db_sa://vks14514:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[46]: **hardship\_index**

6.0

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](#)

## Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
-------------------	---------	------------	--------------------

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-08-28	2.0	Lavanya	Moved lab to course repo in GitLab

---

© IBM Corporation 2020. All rights reserved.