(https://cognitiveclass.ai)

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

Introduction

This notebook shows how to work with a real world dataset using SQL and Python. In this lab you will:

1. Understand the dataset for Chicago Public School level performance
2. Store the dataset in an Db2 database on IBM Cloud instance
3. Retrieve metadata about tables and columns and query data from mixed case columns
4. 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)

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?download=true)

NOTE: Do not download the dataset directly from City of Chicago portal. Instead download a more database friendly version from the link below. Now download a static copy of this database and review some of its contents: https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv (https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv)

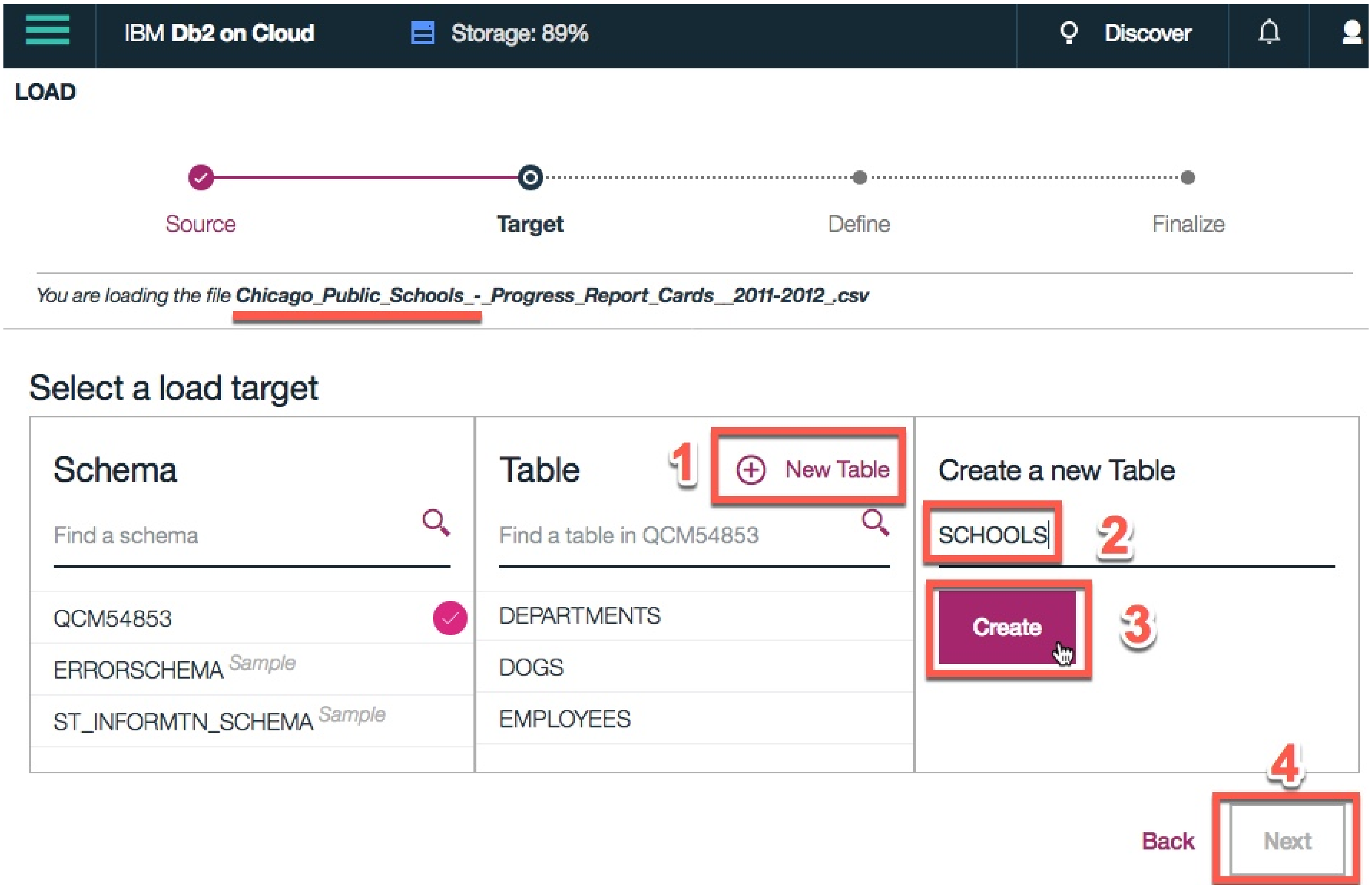
Store the dataset in a Table

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.



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

Select a load target

[(https://cognitiveclass.ai)](https://cognitiveclass.ai)

Connect to the database

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

In [18]:

```
%load_ext sql
%reload_ext sql
```

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

In [19]:

```
# 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://sdp03042:5df-tzzcfp3mx76d@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/B
LUDB
```

Out[19]:

```
'Connected: sdp03042@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 [20]:

```
# 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://sdp03042:***@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/BL
UDB
Done.
```

Out[20]:

tabschema	tabname	owner	ownertype	TYPE	status	base_tabschema	base_tabname	rowtype

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 [21]:

```
# 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://sdp03042:***@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/BL
UDB
Done.
```

Out[21]:

1
0

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 [22]:

```
# 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://sdp03042:***@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/BL
UDB
Done.
```

Out[22]:

colname	typename	length

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 [24]:

```
%sql select count(*) from SCHOOLS where "Elementary, Middle, or High School" = 'ES'
```

```
 * ibm_db_sa://sdp03042:***@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/BL
UDB
(ibm_db_dbi.ProgrammingError) ibm_db_dbi::ProgrammingError: SQLNumResultCols failed: [IBM]
[CLI Driver][DB2/LINUXX8664] SQL0204N  "SDP03042.SCHOOLS" is an undefined name.  SQLSTAT
E=42704 SQLCODE=-204
[SQL: select count(*) from SCHOOLS where "Elementary, Middle, or High School" = 'ES']
(Background on this error at: http://sqlalche.me/e/f405)
```

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 [ ]:

```
%sql select MAX(Safety_Score) AS MAX_SAFETY_SCORE from SCHOOLS
```

Double-click **here** for a hint

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

Problem 3

Which schools have highest Safety Score?

In [ ]:

```
%sql select Name_of_School, Safety_Score from SCHOOLS where Safety_Score = 99
```

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

Problem 4

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

In [ ]:

```
%sql select Name_of_School, Average_Student_Attendance from SCHOOLS \
  order by Average_Student_Attendance desc nulls last limit 10
```

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 [ ]:

```
%sql SELECT Name_of_School, Average_Student_Attendance  \
   from SCHOOLS \
   order by Average_Student_Attendance \
   fetch first 5 rows only
```

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

Problem 6

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

In [ ]:

```
%sql SELECT Name_of_School, REPLACE(Average_Student_Attendance, '%', '') \
   from SCHOOLS \
   order by Average_Student_Attendance \
   fetch first 5 rows only
```

Double-click **here** for a hint

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

Problem 7

Which Schools have Average Student Attendance lower than 70%?

In [ ]:

```
%sql SELECT Name_of_School, Average_Student_Attendance  \
    from SCHOOLS \
    where CAST ( REPLACE(Average_Student_Attendance, '%', '') AS DOUBLE ) < 70 \
    order by Average_Student_Attendance
```

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 [ ]:

```
%sql select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
   from SCHOOLS \
   group by Community_Area_Name
```

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 [ ]:

```
%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
```

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 4638

In [ ]:

```
%%sql
select hardship_index
  from chicago_socioeconomic_data CD, schools CPS
  where CD.ca = CPS.community_area_number
    and college_enrollment = 4368
```

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 [28]:

```
%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 )
```

```
 * ibm_db_sa://sdp03042:***@dashdb-txn-sbox-yp-lon02-02.services.eu-gb.bluemix.net:50000/BL
UDB
(ibm_db_dbi.ProgrammingError) ibm_db_dbi::ProgrammingError: SQLNumResultCols failed: [IBM]
[CLI Driver][DB2/LINUXX8664] SQL0204N  "SDP03042.SCHOOLS" is an undefined name.  SQLSTAT
E=42704 SQLCODE=-204
[SQL: select ca, community_area_name, hardship_index from chicago_socioeconomic_data     where c
a in     ( select community_area_number from schools order by college_enrollment desc limit 1 )]
(Background on this error at: http://sqlalche.me/e/f405)
```

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.