IBM Developer SKILLS NETWORK

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

Estimated time needed: **30** minutes

Objectives

After completing 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.

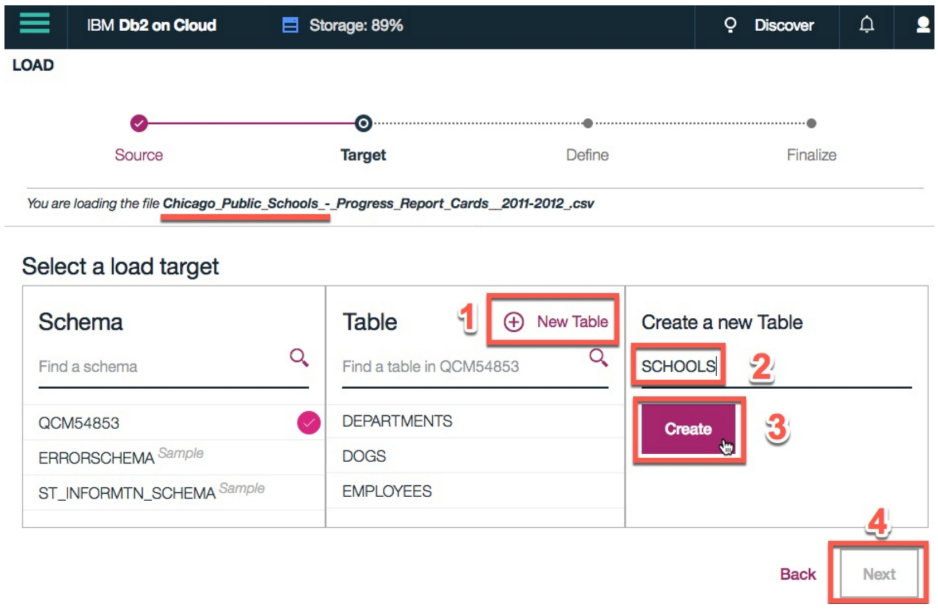
Store the dataset in a Table

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**.



Connect to the database

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

```
[ ]: %load_ext sql
```

```
[ ]: # 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://
```

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)
```

▼ Click here for the solution

```
#In Db2 the system catalog table called SYSCAT.TABLES contains the table metadata

%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
```

▼ Click here for the solution

```
#In Db2 the system catalog table called SYSCAT.COLUMNS contains the column metadata

%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
```

▼ Click here for the 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?

```
[ ]:
```

▼ Click here for the hint

```
Which column specifies the school type e.g. 'ES', 'MS', 'HS'?

Does the column name have mixed case, spaces or other special characters?
If so, ensure you use double quotes around the "Name of the Column"
```

▼ Click here for the solution

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

Correct answer: 462
```

Problem 2

What is the highest Safety Score?

```
[ ]:
```

•••

▼ Click here for the solution

```
# Use the MAX() function

%sql select MAX(Safety_Score) AS MAX_SAFETY_SCORE from SCHOOLS

Correct answer: 99
```

Problem 3

Which schools have highest Safety Score?

```
[ ]:
```

•••

▼ Click here for the 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"?

```
[ ]:
```

•••

▼ Click here for the 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

[ ]:

•••

▼ Click here for the 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

[ ]:

•••

▼ Click here for the solution

```
#Use the REPLACE() function to replace '%' with ''
#See documentation for this function at:
https://www.ibm.com/support/knowledgecenter/en/SSEPGG_10.5.0/com.ibm.db2.luw.sql.ref.doc/doc/r0000843.html

%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%?

[ ]:

•••

▼ Click here for the hint

```
The datatype of the "Average_Student_Attendance" column is varchar.
So you cannot use it as is in the where clause for a numeric comparison.
First use the CAST() function to cast it as a DECIMAL or DOUBLE
e.g. CAST("Column_Name" as DOUBLE)
or simply: DECIMAL("Column_Name")

Don't forget the '%' age sign needs to be removed before casting
```

▼ Click here for the 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

[ ]:

•••

▼ Click here for the hint

```
Verify the exact name of the Enrollment column in the database
Use the SUM() function to add up the Enrollments for each Community Area

Don't forget to group by the Community Area
```

▼ Click here for the 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

[ ]:

•••

▼ Click here for the solution

```
# Order the previous query and limit the number of rows you fetch

%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

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

[ ]:

▼ Click here for the solution

```
# 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 11

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

[ ]:

▼ Click here for the solution

```
# 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.

Author

Rav Ahuja

Change Log

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