Practice Assignment: Working with real world data-sets using SQL and Db2 on IBM Cloud

Estimated time needed: **45** minutes

Pre-requisites:

- IBM Cloud Account
- IBM Db2 service

  NOTE: If you don't have an IBM Cloud account or Db2 service, follow this link and go through the steps given in the Hands-on Lab: Create Db2 service instance and Get started with the Db2 console

Objectives

After completing this lab you will be able to:

- Describe the datasets for Chicago Public School and Chicago Socioeconomic Data
- Load the datasets in an Db2 instance database on IBM Cloud
- Retrieve metadata about tables and columns from system catalogs
- Write SQL queries to filter, order, group result sets and utilize nested queries and 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

Socioeconomic Indicators in Chicago

This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at:
https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

NOTE: Do not download the dataset directly from portal. Instead download a static copy which is a more database friendly version from this

- Chicago Public School
- Chicago Census Data

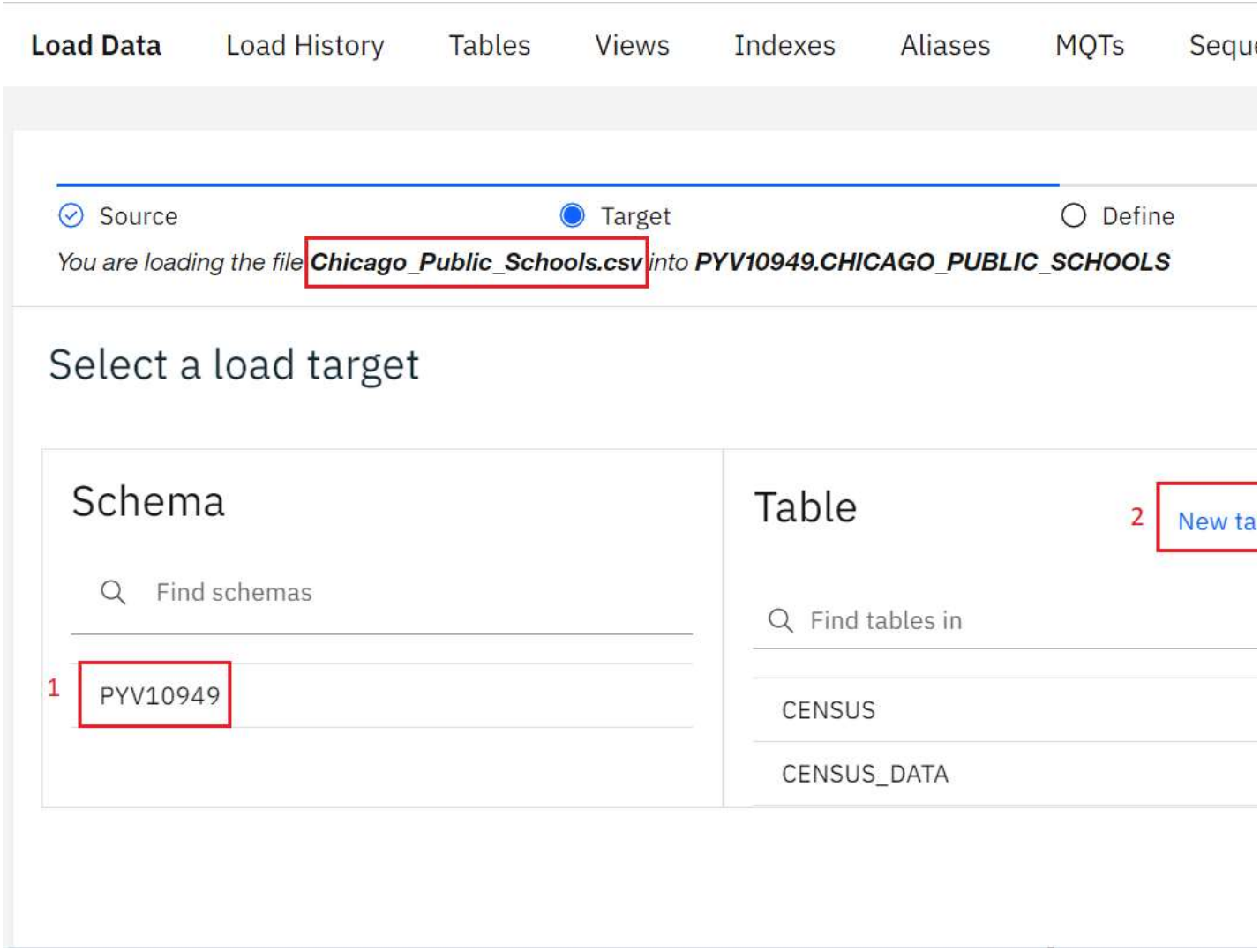
NOTE: If you find the timestamp error while loading the data, then you need to change/overwrite the default Timestamp format of **YYYY-MM-DD HH:MM:SS to MM/DD/YYYY HH:MM:SS TT**. You can also go through the link how to update/modify the timestamp format.
Now we will review some of its contents.

Store the dataset in a Table

In many cases the dataset to be analyzed can be found on the internet and is available as a .CSV (comma separated values) file. To analyze the data using SQL, it first needs to be stored in the database.

We highly recommend that you manually load the table using the database console LOAD tool, as indicated in Week/Module 2 (Optional) Db2 Lab- Create and Load Tables using SQL Scripts- Exercise 2. The only difference with that lab is that in Step 6 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".

- Open Db2 console
- Open the **LOAD** tool
- Select / Drag the .CSV file for the CHICAGO PUBLIC SCHOOLS dataset
- Load the dataset into a new table called **CHICAGO_PUBLIC_SCHOOLS**.



Similary, load the Chicago Socioeconmic Indicators Census Data into a new table called **CENSUS_DATA**

NOTE: If Chicago Socioeconomic Indicators Census Data has been loaded previously, you can skip loading it.

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

▶ Click **here** for a hint
▼ Click **here** for the solution

```
1. 1
2. 2
```

```
1. Solution:
2. select TABSCHEMA, TABNAME, CREATE_TIME from SYSCAT.TABLES where TABSCHEMA='YOUR-DB2-USERNAME';
```

Copied!

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

```
1. 1
2. 2
```

```
1. select TABSCHEMA, TABNAME, CREATE_TIME from SYSCAT.TABLES
2.       where TABSCHEMA not in ('SYSIBM', 'SYSCAT', 'SYSSTAT', 'SYSIBMADM', 'SYSTOOLS', 'SYSPUBLIC');
```

Copied!

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

```
1
```

```
select * from SYSCAT.TABLES where TABNAME = 'CHICAGO_PUBLIC_SCHOOLS';
```

Copied!

Query the database system catalog to retrieve column metadata

The CHICAGO_PUBLIC_SCHOOLS table contains a large number of columns. How many columns does SCHOOL table have?

▶ Click **here** for a hint
▼ Click **here** for the solution

```
1
2
3
```

```
::page{title="Solution:"}

select count(*) from SYSCAT.COLUMNS where TABNAME = 'CHICAGO_PUBLIC_SCHOOLS';
```

Copied!

Retrieve the list of columns in SCHOOLS table and their column type (datatype) and length

▼ Click **here** for the solution

```
1
2
3
4
5
6
7
```

```
::page{title="Solution:"}

select COLNAME, TYPENAME, LENGTH from SYSCAT.COLUMNS where TABNAME = 'CHICAGO_PUBLIC_SCHOOLS';

::page{title="or"}

select distinct(NAME), COLTYPE, LENGTH from SYSIBM.SYSCOLUMNS where TBNAME = 'CHICAGO_PUBLIC_SCHOOLS';
```

Copied!

Questions

1. Is the column name for the “SCHOOL ID” attribute in upper,lower 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 a hint
▶ Click **here** for another hint
▼ Click **here** for the solution

```
1
2
3
4
5
```

```
::page{title="Solution:"}

select count(*) from CHICAGO_PUBLIC_SCHOOLS where "Elementary, Middle, or High School" = 'ES';

::page{title="Correct answer: 462"}
```

Copied!

Problem 2: What is the highest Safety Score?

▶ Click **here** for a hint
▼ Click **here** for the solution

```
1
2
3
4
5
```

```
 # Hint:

select MAX(SAFETY_SCORE) AS MAX_SAFETY_SCORE from CHICAGO_PUBLIC_SCHOOLS;

::page{title="Correct answer: 99"}
```

Copied!

Problem 3: Which schools have the highest Safety Score?

▼ 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:

```
1
```

```
select NAME_OF_SCHOOL, SAFETY_SCORE from SCHOOLS where SAFETY_SCORE = 99;
```

Copied!

or, a better way:

```
1
2
```

```
select NAME_OF_SCHOOL, SAFETY_SCORE from CHICAGO_PUBLIC_SCHOOLS where
  SAFETY_SCORE= (select MAX(SAFETY_SCORE) from CHICAGO_PUBLIC_SCHOOLS);
```

Copied!

Correct answer: There are several schools with a Safety Score of 99.

Problem 4: What are the top 10 schools with the highest “Average Student Attendance”?

▼ Click **here** for the solution

```
1
2
3
4
```

```
::page{title="Solution:"}

select NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE from CHICAGO_PUBLIC_SCHOOLS
    order by AVERAGE_STUDENT_ATTENDANCE desc nulls last limit 10;
```

Copied!

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

```
1
2
3
4
5
6
```

```
::page{title="Solution:"}

SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE
     from CHICAGO_PUBLIC_SCHOOLS
     order by AVERAGE_STUDENT_ATTENDANCE
     fetch first 5 rows only;
```

Copied!

Problem 6: Now remove the ‘%’ sign from the above result set for Average Student Attendance column.

▶ Click **here** for a hint
▼ Click **here** for the solution

```
1
```

```
2
3
4
5
6
```

```
::page{title="Hint:"}

SELECT NAME_OF_SCHOOL, REPLACE(AVERAGE_STUDENT_ATTENDANCE, '%', '')
     from CHICAGO_PUBLIC_SCHOOLS
     order by AVERAGE_STUDENT_ATTENDANCE
     fetch first 5 rows only;
```

Copied!

Problem 7: Which Schools have Average Student Attendance lower than 70%?

▶ Click **here** for a hint
▶ Click **here** for another hint
▼ Click **here** for the solution

```
1
2
3
4
5
6
7
8
9
10
11
12
13
```

```
::page{title="Solution:"}

SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE
     from CHICAGO_PUBLIC_SCHOOLS
     where CAST ( REPLACE(AVERAGE_STUDENT_ATTENDANCE, '%', '') AS DOUBLE ) < 70
     order by AVERAGE_STUDENT_ATTENDANCE;

::page{title="or,"}

SELECT NAME_OF_SCHOOL, AVERAGE_STUDENT_ATTENDANCE
     from CHICAGO_PUBLIC_SCHOOLS
     where DECIMAL ( REPLACE(AVERAGE_STUDENT_ATTENDANCE, '%', '') ) < 70
     order by AVERAGE_STUDENT_ATTENDANCE;
```

Copied!

Problem 8: Get the total College Enrollment for each Community Area.

▶ Click **here** for a hint
▶ Click **here** for another hint
▼ Click **here** for the solution

```
1
2
3
4
5
6
```

```

::page{title="Solution:"}

select COMMUNITY_AREA_NAME, sum(COLLEGE_ENROLLMENT) AS TOTAL_ENROLLMENT
   from CHICAGO_PUBLIC_SCHOOLS
   group by COMMUNITY_AREA_NAME;
```

Copied!

Problem 9: Get the 5 Community Areas with the least total College Enrollment sorted in ascending order.

▶ Click **here** for a hint
▼ Click **here** for the solution

```
1
2
3
4
5
6
7
8
```

```
1.
2. ::page{title="Solution:"}
3.
4. select COMMUNITY_AREA_NAME, sum(COLLEGE_ENROLLMENT) AS TOTAL_ENROLLMENT
5.    from CHICAGO_PUBLIC_SCHOOLS
6.    group by COMMUNITY_AREA_NAME
7.    order by TOTAL_ENROLLMENT asc
8.    fetch first 5 rows only;
```

Copied!

Problem 10: List 5 schools with lowest safety score.

▼ Click **here** for the solution

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
```

```
1.
2. ::page{title="Solution:"}
3.
4. select NAME_OF_SCHOOL, SAFETY_SCORE
5. from CHICAGO_PUBLIC_SCHOOLS
6. order by SAFETY_SCORE
7. limit 5;
```

Copied!

Problem 11: Get the hardship index for the community area which has College Enrollment of 4368.

▼ Click **here** for the solution

NOTE: For this solution to work the CENSUS_DATA table
as created in the last lab of Week 3 should already exist

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
```

```
1. ::page{title="Solution:"}
2.
3. select HARDSHIP_INDEX
4.    from CENSUS_DATA CD, CHICAGO_PUBLIC_SCHOOLS CPS
5.    where CD.COMMUNITY_AREA_NUMBER = CPS.COMMUNITY_AREA_NUMBER
6.       AND COLLEGE_ENROLLMENT = 4368;
```

Copied!

Problem 12: Get the hardship index for the community area which has the school with the highest enrollment.

▼ Click **here** for the solution
NOTE: For this solution to work the CENSUS_DATA table as created in the last lab of Week 3 should already exist

```
1. 1
2. 2
3. 3
4. 4
5. 5
```

```
1. ::page{title="Solution:"}
2.
3. select COMMUNITY_AREA_NUMBER, COMMUNITY_AREA_NAME, HARDSHIP_INDEX FROM CENSUS_DATA
4.    where COMMUNITY_AREA_NUMBER in
5.    ( select COMMUNITY_AREA_NUMBER FROM CHICAGO_PUBLIC_SCHOOLS ORDER BY COLLEGE_ENROLLMENT DESC LIMIT 1 );
```

Copied!

Summary

In this lab you learned how to work with a real word dataset using SQL and Db2 on IBM Cloud. You have learned how to use built in database functions and practiced how to sort, limit, and order result sets. You also used sub-queries and worked with multiple tables.

Author

Rav Ahuja

Other Contributor(s)

Malika Singla

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2023-05-11	1.7	Eric Hao & Vladislav Boyko	Updated Page Frames
2023-05-10	1.6	Eric Hao & Vladislav Boyko	Updated Page Frames
2023-05-10	1.5	Eric Hao & Vladislav Boyko	Updated Page Frames
2023-05-10	1.4	Eric Hao & Vladislav Boyko	Updated Page Frames
2022-10-27	1.3	Appalabhaktula Hema	Updated instructions
2021-01-06	1.2	Rav Ahuja	Edits and corrections.
2021-11-23	1.1	Malika	Forked from Original.

© IBM Corporation 2023. All rights reserved.