DB0201EN-Week4-1-1-RealDataPractice-v5-Solved

July 2, 2022

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

Estaimted time needed: 30 minutes

1.1 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

1.2 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.

1.2.1 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.

1.2.2 Connect to the database

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

The following modules are pre-installed in the Skills Network Labs environment. However if you run this notebook commands in a different Jupyter environment (e.g. Watson Studio or Ananconda) you may need to install these libraries by removing the # sign before !pip in the code cell below.

```
[1]: # These libraries are pre-installed in SN Labs. If running in another environment please uncomment lines below to install them:

# !pip install --force-reinstall ibm_db==3.1.0 ibm_db_sa==0.3.3

# Ensure we don't load_ext with sqlalchemy>=1.4 (incompadible)

# !pip uninstall sqlalchemy==1.4 -y & pip install sqlalchemy==1.3.24

# !pip install ipython-sql
```

[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://lqh43420:lb76AfXMQnkkX6lw@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.

bs2io90l08kqblod8lcg.databases.appdomain.cloud:31505/BLUDB?security=SSL
```

1.2.3 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

```
[4]: # type in your query to retrieve list of all tables in the database for your_

db2 schema (username)

%sql select TABSCHEMA, TABNAME, CREATE_TIME from SYSCAT.TABLES where

TABSCHEMA='lqh43420'
```

* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.

[4]: []

Double-click **here** for a hint

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

1.2.4 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?

```
[5]: # 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://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.

[5]: [(78,)]

Double-click here for a hint

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

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

```
[6]: # 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'
```

* $ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

```
[6]: [('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),
      ('ADEQUATE_YEARLY_PROGRESS_MADE_', 'VARCHAR ', 3),
      ('AVERAGE_STUDENT_ATTENDANCE', 'VARCHAR', 6),
      ('AVERAGE_TEACHER_ATTENDANCE', 'VARCHAR', 6),
      ('CITY', 'VARCHAR', 20),
      ('COLLABORATIVE_NAME', 'VARCHAR', 40),
      ('COLLEGE_ELIGIBILITY__', 'VARCHAR ', 4),
      ('COLLEGE_ENROLLMENT', 'SMALLINT', 2),
      ('COLLEGE_ENROLLMENT_RATE__', 'VARCHAR ', 4),
      ('COMMUNITY_AREA_NAME', 'VARCHAR', 22),
      ('COMMUNITY_AREA_NUMBER', 'SMALLINT', 2),
      ('CPS_PERFORMANCE_POLICY_LEVEL', 'VARCHAR', 15),
      ('CPS_PERFORMANCE_POLICY_STATUS', 'VARCHAR', 30),
      ('EMHS', 'VARCHAR', 2),
      ('ENVIRONMENT_ICON', 'VARCHAR', 11),
      ('ENVIRONMENT SCORE', 'SMALLINT', 2),
```

```
('FAMILY_INVOLVEMENT_ICON', 'VARCHAR', 20),
('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),
('GRADUATION_RATE__', 'VARCHAR', 4),
('GR_8_EXPLORE_MATH__', 'VARCHAR ', 4),
('GR_8_EXPLORE_READ__', 'VARCHAR ', 4),
('HEALTHY_SCHOOL_CERTIFIED', 'VARCHAR', 3),
('INDIVIDUALIZED EDUCATION PROGRAM COMPLIANCE RATE', 'VARCHAR', 7),
('INSTRUCTION_ICON', 'VARCHAR', 11),
('INSTRUCTION_SCORE', 'SMALLINT', 2),
('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),
('LATITUDE', 'DECIMAL', 18),
('LEADERS_ICON', 'VARCHAR', 4),
('LEADERS_SCORE', 'VARCHAR', 3),
('LINK', 'VARCHAR', 100),
('LOCATION', 'VARCHAR', 27),
('LONGITUDE', 'DECIMAL', 18),
('NAME_OF_SCHOOL', 'VARCHAR ', 100),
('NETWORK_MANAGER', 'VARCHAR', 40),
('NET_CHANGE_EXPLORE_AND_PLAN', 'VARCHAR', 3),
('NET_CHANGE_PLAN_AND_ACT', 'VARCHAR', 3),
('PARENT_ENGAGEMENT_ICON', 'VARCHAR', 7),
('PARENT ENGAGEMENT SCORE', 'VARCHAR', 3),
('PARENT_ENVIRONMENT_ICON', 'VARCHAR', 7),
('PARENT_ENVIRONMENT_SCORE', 'VARCHAR', 3),
('PHONE_NUMBER', 'VARCHAR', 20),
('PK_2_LITERACY__', 'VARCHAR ', 4),
('PK_2_MATH__', 'VARCHAR', 4),
('POLICE_DISTRICT', 'SMALLINT', 2),
('RATE_OF_MISCONDUCTS_PER_100_STUDENTS_', 'DECIMAL ', 5),
('SAFETY_ICON', 'VARCHAR', 20),
('SAFETY_SCORE', 'INTEGER ', 4),
('SCHOOL_ID', 'INTEGER', 4),
```

```
('STATE', 'VARCHAR', 2),
('STREET_ADDRESS', 'VARCHAR', 40),
('STUDENTS_PASSING__ALGEBRA__', 'VARCHAR', 4),
('STUDENTS_TAKING__ALGEBRA__', 'VARCHAR', 4),
('TEACHERS_ICON', 'VARCHAR', 11),
('TEACHERS_SCORE', 'VARCHAR', 3),
('TRACK_SCHEDULE', 'VARCHAR', 20),
('WARD', 'SMALLINT', 2),
('X_COORDINATE', 'DECIMAL', 13),
('Y_COORDINATE', 'DECIMAL', 13),
('ZIP_CODE', 'INTEGER', 4)]
```

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

1.2.5 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 "_"?

1.3 Problems

1.3.1 Problem 1

How many Elementary Schools are in the dataset?

```
[7]: %sql select count(*) from SCHOOLS group by "EMHS"
```

* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.

```
[7]: [(462,), (93,), (11,)]
```

Double-click here for a hint

Double-click here for another hint

Double-click here for the solution.

1.3.2 Problem 2

What is the highest Safety Score?

```
[8]: %sql select MAX(Safety_Score) AS MAX_SAFETY_SCORE from SCHOOLS
```

```
* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.
```

```
[8]: [(99,)]
```

Double-click **here** for a hint

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

1.3.3 Problem 3

Which schools have highest Safety Score?

 $* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

```
[9]: [('Abraham Lincoln Elementary School', 99),
      ('Alexander Graham Bell Elementary School', 99),
      ('Annie Keller Elementary Gifted Magnet School', 99),
      ('Augustus H Burley Elementary School', 99),
      ('Edgar Allan Poe Elementary Classical School', 99),
      ('Edgebrook Elementary School', 99),
      ('Ellen Mitchell Elementary School', 99),
      ('James E McDade Elementary Classical School', 99),
      ('James G Blaine Elementary School', 99),
      ('LaSalle Elementary Language Academy', 99),
      ('Mary E Courtenay Elementary Language Arts Center', 99),
      ('Northside College Preparatory High School', 99),
      ('Northside Learning Center High School', 99),
      ('Norwood Park Elementary School', 99),
      ('Oriole Park Elementary School', 99),
      ('Sauganash Elementary School', 99),
      ('Stephen Decatur Classical Elementary School', 99),
      ('Talman Elementary School', 99),
      ('Wildwood Elementary School', 99)]
```

Double-click here for the solution.

1.3.4 Problem 4

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

```
* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.
```

```
('Orozco Fine Arts & Sciences Elementary School', '97.60%'), ('Rachel Carson Elementary School', '97.60%'), ('Annie Keller Elementary Gifted Magnet School', '97.50%'), ('Andrew Jackson Elementary Language Academy', '97.40%'), ('Lenart Elementary Regional Gifted Center', '97.40%'), ('Disney II Magnet School', '97.30%'), ('John H Vanderpoel Elementary Magnet School', '97.20%')]
```

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

1.3.5 Problem 5

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

```
[11]: %sql SELECT Name_of_School, Average_Student_Attendance from SCHOOLS order by_Average_Student_Attendance LIMIT 5
```

```
* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.
```

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

1.3.6 Problem 6

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

```
[12]: %sql SELECT Name_of_School, REPLACE(Average_Student_Attendance, '%', '') from______SCHOOLS order by Average_Student_Attendance LIMIT 5
```

```
* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.
```

Double-click here for a hint

Double-click here for the solution.

1.3.7 Problem 7

Which Schools have Average Student Attendance lower than 70%?

```
[13]: %sql SELECT Name_of_School, Average_Student_Attendance from SCHOOLS where □
□DECIMAL ( REPLACE(Average_Student_Attendance, '%', '') ) < 70 order by □
□Average_Student_Attendance
```

 $* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

Double-click **here** for a hint

Double-click here for another hint

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

1.3.8 Problem 8

Get the total College Enrollment for each Community Area

[14]: %sql select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT_

ofrom SCHOOLS group by Community_Area_Name

 $* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

```
('CALUMET HEIGHTS', 1568),
('CHATHAM', 5042),
('CHICAGO LAWN', 7086),
('CLEARING', 2085),
('DOUGLAS', 4670),
('DUNNING', 4568),
('EAST GARFIELD PARK', 5337),
('EAST SIDE', 5305),
('EDGEWATER', 4600),
('EDISON PARK', 910),
('ENGLEWOOD', 6832),
('FOREST GLEN', 1431),
('FULLER PARK', 531),
('GAGE PARK', 9915),
('GARFIELD RIDGE', 4552),
('GRAND BOULEVARD', 2809),
('GREATER GRAND CROSSING', 4051),
('HEGEWISCH', 963),
('HERMOSA', 3975),
('HUMBOLDT PARK', 8620),
('HYDE PARK', 1930),
('IRVING PARK', 7764),
('JEFFERSON PARK', 1755),
('KENWOOD', 4287),
('LAKE VIEW', 7055),
('LINCOLN PARK', 5615),
('LINCOLN SQUARE', 4132),
('LOGAN SQUARE', 7351),
('LOOP', 871),
('LOWER WEST SIDE', 7257),
('MCKINLEY PARK', 1552),
('MONTCLARE', 1317),
('MORGAN PARK', 3271),
('MOUNT GREENWOOD', 2091),
('NEAR NORTH SIDE', 3362),
('NEAR SOUTH SIDE', 1378),
('NEAR WEST SIDE', 7975),
('NEW CITY', 7922),
('NORTH CENTER', 7541),
('NORTH LAWNDALE', 5146),
('NORTH PARK', 4210),
('NORWOOD PARK', 6469),
('OAKLAND', 140),
('OHARE', 786),
('PORTAGE PARK', 6954),
('PULLMAN', 1620),
('RIVERDALE', 1547),
```

```
('ROGERS PARK', 4068),
('ROSELAND', 7020),
('SOUTH CHICAGO', 4043),
('SOUTH DEERING', 1859),
('SOUTH LAWNDALE', 14793),
('SOUTH SHORE', 4543),
('UPTOWN', 4388),
('WASHINGTON HEIGHTS', 4006),
('WASHINGTON PARK', 2648),
('WEST ELSDON', 3700),
('WEST ENGLEWOOD', 5946),
('WEST GARFIELD PARK', 2622),
('WEST LAWN', 4207),
('WEST PULLMAN', 3240),
('WEST RIDGE', 8197),
('WEST TOWN', 9429),
('WOODLAWN', 4206)]
```

Double-click here for a hint

Double-click here for another hint

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

1.3.9 Problem 9

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

```
[15]: %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
```

 $* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

Double-click **here** for a hint

Double-click here for the solution.

1.3.10 Problem 10

List 5 schools with lowest safety score.

```
[16]: %sql SELECT name_of_school, safety_score FROM schools ORDER BY safety_score_
```

* $ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.$

Double-click here for the solution.

1.3.11 Problem 11

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

```
[17]: %sql SELECT hardship_index from chicago_socioeconomic_data CD, schools CPS_ where CD.ca = CPS.community_area_number and college_enrollment = 4368
```

* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.

[17]: [(6,)]

Double-click here for the solution.

1.3.12 Problem 12

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

* ibm_db_sa://lqh43420:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB Done.

```
[18]: [(5, 'North Center', 6)]
```

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

1.4 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.

1.5 Author

Rav Ahuja

1.6 Change Log

Date			
(YYYY-MM-DD)	Version	Changed By	Change Description
2021-07-09	2.4	Malika	Updated connection string
2021-05-19	2.3	Lakshmi Holla	Updated question
2021-04-20	2.2	Malika	Added the libraries
2020-11-27	2.1	Sannareddy	Modified data sets and added new problems
		Ramesh	
2020-08-28	2.0	Lavanya	Moved lab to course repo in GitLab

##

© IBM Corporation 2020. All rights reserved.