DB0201EN-Week4-1-1-Analyzing-1-py

October 6, 2018

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

1 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 1. Store the dataset in an Db2 database on IBM Cloud instance 1. Retrieve metadata about tables and columns and query data from mixed case columns 1. Solve example problems to practice your SQL skills including using built-in database functions

1.1 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

Now download a static copy of this database and review some of its contents: https://ibm.box.com/shared/static/0g7kbanvn5l2gt2qu38ukooatnjqyuys.csv

1.1.1 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.

1.1.2 Connect to the database

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

```
In [65]: %load_ext sql
```

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

```
In [66]: # 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://dash6842:P2OEre7h_H_t@dashdb-entry-yp-dal09-09.services.dal.bluemix.ne
```

```
Out[66]: 'Connected: dash6842@BLUDB'
```

1.1.3 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

```
In [67]: # type in your query to retrieve list of all tables in the database for your db2 schema
         %sql select TABSCHEMA, TABNAME, CREATE_TIME from SYSCAT.TABLES where TABSCHEMA='dash684
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[67]: []
```

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

1.1.4 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?

```
In [68]: # type in your query to retrieve the number of columns in the SCHOOLS table
         %sql select COUNT(*) from syscat.columns where tabname = 'SCHOOL'
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[68]: [(Decimal('79'),)]
```

```
In [69]: %sql select COUNT(*) from syscat.columns where tabname = 'CRIME'
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[69]: [(Decimal('22'),)]
```

```
In [70]: %sql select COUNT(*) from syscat.columns where tabname = 'SOCIOECONOMIC'
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[70]: [(Decimal('9'),)]
```

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 [137]: %sql select distinct(name), coltype, length from sysibm.syscolumns where tbname = 'SOC
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[137]: [('COMMUNITY_AREA_NAME', 'VARCHAR ', 22),
           ('Community_Area_Number', 'SMALLINT', 2),
           ('HARDSHIP_INDEX', 'SMALLINT', 2),
           ('PERCENT_AGED_16__UNEMPLOYED', 'DECIMAL ', 4),
           ('PERCENT_AGED_25__WITHOUT_HIGH_SCHOOL_DIPLOMA', 'DECIMAL ', 4),
           ('PERCENT_AGED_UNDER_18_OR_OVER_64', 'DECIMAL ', 4),
           ('PERCENT_HOUSEHOLDS_BELOW_POVERTY', 'DECIMAL ', 4),
           ('PERCENT_OF_HOUSING_CROWDED', 'DECIMAL ', 4),
           ('PER_CAPITA_INCOME', 'INTEGER ', 4)]
```

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

1.1.5 Questions

1. Is the column name for the "SCHOOL ID" attribute in upper or mixed case?
2. What is the "Community Area Name" field called in your table? Have the spaces " " between the words been replaced by some other character?
3. Have the paranthesis (round brackets) in the "College Enrollment (number of students)" attribute been replaced by some other character?
4. Are there any columns in whose names the spaces and paranthesis (round brackets) have not been replaced by the underscore character "_"?

1.2 Problems

1.2.1 Problem 1

Find the total number of records for each of the tables

```
In [162]: %sql SELECT * FROM SCHOOL LIMIT 2
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[162]: [(610038, 'Abraham Lincoln Elementary School', 'ES', '615 W Kemper Pl ', 'Chicago', 'I
           (610281, 'Adam Clayton Powell Paideia Community Academy Elementary School', 'ES', '75
```

```
In [76]: %sql SELECT COUNT(*) FROM CRIME
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[76]: [(Decimal('533'),)]
```

```
In [82]: %sql SELECT COUNT(*) FROM SOCIOECONOMIC
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[82]: [(Decimal('78'),)]
```

Double-click **here** for a hint

```
In [77]: , COUNT("Community_Area_Name")  GROUP BY "Community_Area_Name"
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[77]: [(Decimal('78'),)]
```

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

1.2.2 Problem 2

Find average college enrollments by community area

```
In [105]: %sql SELECT "Healthy_Schools_Certified_"  FROM SCHOOL LIMIT 3
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[105]: [('Yes',), ('No',), ('No',)]
```

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

1.2.3 Problem 3

Find the number of schools that are healthy school certified

```
In [106]: %sql SELECT COUNT("Healthy_Schools_Certified_") FROM SCHOOL WHERE "Healthy_Schools_Cer
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[106]: [(Decimal('16'),)]
```

```
In [107]: %sql SELECT * FROM CRIME LIMIT 1
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[107]: [(3512276, 'HK587712', '08/28/2004 05:50:56 PM', '047XX S KEDZIE AVE', '0890', 'THEFT'
```

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

1.2.4 Problem 4

How many observations have a Location Description value of GAS STATION?

```
In [110]: %sql SELECT COUNT("Location_Description") FROM CRIME WHERE "Location_Description" LIKE
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[110]: [(Decimal('6'),)]
```

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

1.2.5 Problem 5

Retrieve a list of the top 10 community areas which have most number of schools and sorted in descending order

```
In [113]: %sql SELECT "Community_Area_Name", COUNT("Community_Area_Name") as NUMBER_OF_SCHOOL FR
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[113]: [('AUSTIN', Decimal('23')),
           ('SOUTH LAWNDALE', Decimal('22')),
           ('WEST TOWN', Decimal('20')),
           ('ENGLEWOOD', Decimal('17')),
           ('NORTH LAWNDALE', Decimal('16')),
           ('NEAR WEST SIDE', Decimal('16')),
           ('HUMBOLDT PARK', Decimal('13')),
           ('ROSELAND', Decimal('13')),
           ('WEST ENGLEWOOD', Decimal('13')),
           ('EAST GARFIELD PARK', Decimal('13'))]
```

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

1.2.6 Problem 6

How many observations have value MOTOR VEHICLE THEFT in the Primary Type variable (this is the number of crimes related to Motor vehicles)

```
In [116]: %sql SELECT COUNT("Primary_Type") FROM CRIME WHERE "Primary_Type" LIKE 'MOTOR%'

 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.


Out[116]: [(Decimal('24'),)]
```

```
In [43]: %sql SELECT "Name_of_School", REPLACE("Average_Student_Attendance", '%','') FROM CHICAG

 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.


Out[43]: [('Velma F Thomas Early Childhood Center', ''),
          ('Richard T Crane Technical Preparatory High School', '57.9'),
          ('Barbara Vick Early Childhood & Family Center', '60.9'),
          ('Dyett High School', '62.5'),
          ('Wendell Phillips Academy High School', '63.0')]
```

```
In [148]: %sql SELECT * FROM SOCIOECONOMIC LIMIT 1

 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.


Out[148]: [(1, 'Rogers Park', Decimal('7.7'), Decimal('23.6'), Decimal('8.7'), Decimal('18.2'),
```

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

1.2.7 Problem 7

Using INNER JOIN, find the minimum "Average Student Attendance" for community are where hardship is 96

```
In [171]: %sql SELECT Min(S."Average_Student_Attendance")\
          FROM SCHOOL S INNER JOIN SOCIOECONOMIC O \
          ON S."Community_Area_Number" = O."Community_Area_Number" \
          WHERE O.hardship_index = 96
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[171]: [('86.1%',)]
```

Double-click **here** for a hint
Double-click **here** for another hint
Double-click **here** for the solution.

1.2.8 Problem 8

Get the total College Enrollment (number of students) for each Community Area Double-click **here** for a hint
Double-click **here** for another hint
Double-click **here** for the solution.

1.2.9 Problem 9

Get the 5 Community Areas with the least total College Enrollment (number of students) sorted in ascending order

```
In [64]: %sql SELECT Distinct("Community_Area_Name"), SUM("College_Enrollment__number_of_student
         FROM CHICAGO_DATA \
         GROUP BY "Community_Area_Name"\
         ORDER BY sum_no LIMIT 5
```

```
 * ibm_db_sa://dash6842:***@dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB
Done.
```

```
Out[64]: [('OAKLAND', 140),
          ('FULLER PARK', 531),
          ('BURNSIDE', 549),
          ('OHARE', 786),
          ('LOOP', 871)]
```

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

1.3 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 names. You also used built in database functions. Copyright ľ 2018 cognitiveclass.ai. This notebook and its source code are released under the terms of the MIT License.