THE UNIVERSITY OF SYDNEY

School of Computer Science
Uwe Roehm

DATA2001/2901: Data Science, Big Data, and Data Diversity Semester 1, 2021

Tutorial Week 4: Declarative Data Analysis with SQL and Python

This week's tutorial is about how to load data into a relational database, such as PostgreSQL, from a CSV file, as well as how to transform and clean the data using Python code and Jupyter notebooks. We will also further train SQL and the data analytical operations which we covered in this week's lecture.

The details for logging into the university hosted database are as follows (same as Week 03):

- **hostname**: soitpw11d59.shared.sydney.edu.au
- **username**: `y21s1d2x01_`*yourUnikey*
- **password**: your_SID
- **port**: 5432

(**Note**: this is not your normal unikey login/password!). For example, if your unikey is abcd1234 then your username would be y21s1d2x01_abcd1234.

For today's exercises - you should modify the **db2x01.json** file to match your appropriate credentials.

Off-campus Access: To be able to access either Jupyter or the database off-campus, you need the University's CISCO VPN connected. (vpn.sydney.edu.au) and the FortiClient VPN if you are connecting from China. Details available here: https://edstem.org/courses/5592/discussion/399227

If you are connecting to the university database using a local installation of Jupyter - You will still need to connect the the University network via VPN.

Exercise 1. SQL+Python - Resources for today

We have provided a Jupyter notebook, 1 credentials file and 4 data files for todays tutorial. You should download all of these from Canvas. You should store all the files in the same folder when working on your Jupyter notebook. If you are using the university servers - please make sure to **edit your credentials in the db2x01.json file!**

Exercise 2. Data Loading and Database Creation with Python

This weeks Jupyter notebook has some instructions on how to connect to a database using Python. These will guide you in how to load the Organisations.csv file into the database.

Follow the instruction there except now load the Measurements.csv into a new table `'MeasurementsWk4'` in PostgreSQL.
(**Note:** the data provided is slightly simpler than last week to make this task easier; every attribute has its own column here.)

Exercise 3. Data Analysis in SQL + Querying a Database from Python

Still working in Jupyter, your task is to answer each of the following questions with an SQL query which you are issuing from Python, and whose result you give out here in the Jupyter notebook. The idea here is to train both SQL analytical queries, and how to query an existing database from a Python program.

(a) List the average water temperature per year.

(b) Find the minimum and the maximum water temperature per year.

(c) List the average water flow per station and year.

(d) List the number of temperature measurements per station, with the stations given by name and in descending order of the number of measurements.

(e) How many stations does each organisation have? List the organisations by name and in descending order of the number of associated stations.

Exercise 4. Data Importing via command line (ADV)

Follow the instructions on the jupyter notebook to load the data via psql.

This will be a similar login process to last weeks advanced exercises.

Exercise 5. Prescriptive Statistics with SQL (ADV)

The following set of SQL questions are for students in the advanced stream (DATA2901). They refer back to the advanced SQL content covered in the advanced seminar.

(a) Find the average water temperatures per year and per station, as well as the averages per station and the overall temperature values per year – using a single SQL query.

(b) Find the five statistical values needed for multiple Tukey Boxplots on the value distributions of the water temperature measurements at station 'Murray River at Corowa' per year.

(c) Are there any outliers of water temperature measurements at 'Murray River at Corowa' per year? If yes, list them (per year).

(d) Is there a correlation between the annual water temperature measurements at 'Murray River at Albury' and at 'Murray River at Barham'?

Exercise 6. SQL Grok Exercises

SQL Exercises 5-9 have been released. You should complete these at your own place. We recommend completing these before your SQL Quiz in Week 07. We will begin dedicated tutor SQL help desk hours from Week 05 (Wednesdays and Thursdays to start).

Please wait for the announcement on Ed.