DATA2001: Data Science: Big Data and Data Diversity

Tutorial 3

Welcome to the Tutorial 3 of DATA2001. In this tutorial you will learn how to access and analyse data in a RDBMS such as PostgreSQL.

Using pgAdmin version III or IV connect to your local or SIT database server. SIT database server address is `soit-db-pro-2.ucc.usyd.edu.au`, and your username follow this format `y18s1d2001_yourUnikey` where password is your SID. For example, if your unikey is abcd1234 then your username would be y18s1d2001_abcd1234.

To generate waterways database perform the follow steps:

1. Create a new database.
2. Open a new SQL query window.
3. Download waterwaysdata_simple_ddl.sql and waterwaysdata_simple_dml.sql files from Canvas.
4. Open and run waterwaysdata_simple_ddl.sql
5. Open and run waterwaysdata_simple_dml.sql

This will generate a new schema, `waterwaysdata` and four tables: measurements, oganisations, sensors and stations.

Syntax of most commonly used SQL statements:

SQL Command	*Meaning*
CREATE TABLE T *(...)*	creates a new table *T*; list the attributes in brackets in the form `attribute type`
DROP TABLE T	if needed - removes an existing table *T*
INSERT INTO T *VALUES (..)*	inserts a new row into table T
DELETE FROM T	deletes *all* rows from table *T*
SELECT COUNT(*) FROM T	count how many tuples are stored in table *T*
*SELECT * FROM* T	list the content of table *T*

To execute SQL queries against water ways data, we need to first set our search path to waterwaysdata schema.

```
set search_path to waterwaysdata;
```

Execute and understand the following queries:

```
select * from stations
```

```
select * from measurements where sensorid='temp';

select * from measurements where date<'2008-01-01';

select * from measurements as m inner join stations as s on
m.stationid=s.id limit 10;

select *  from measurements m INNER JOIN stations s ON
m.stationid=s.id WHERE stationid = 409204;
```

Then

- Find all sensors from the sensor table
- Find all the stations where sampling is continuing (i.e., sampling not ceased)
- Display all organisation
- Find all non-zero measurements from the Measurements table
- Find all measurements related to sensor ‘temp’
- Find all measurements related to sensor ‘disc’ before 01-01-2011
- Find all measurements related to sensor ‘level’ between 01-01-2008 and 31-12-2010
- Find top ten measurements related to sensors those have ceased operations
- Find the total number of measurements performed by the organisation named ‘NSW Department of Water and Energy’
- Find all measurements by the organisation ‘Queensland Department of Natural Resources and Water’ between the date 01-01-2008 and 31-12-2011

Advanced Users

We understand some students have studied databases before, therefore, we would like to give them advanced tasks.

- Advance users can try working with a postgresql database using the 'psql' command instead of using pgAdmin.
- Download `stockdump.csv` and `stockrecommendation.csv` files from Canvas.
- Bring this `stocks` information data into two respective tables.
- As a first step you can import all the data as text data-type.
- Once you have imported all the data successfully, covert price of stock fields to decimal type and date fields to date-type.
- Perform queries to work-out if recommendation of stocks has correlation with their price increase in future.
- Work out which were the best/worst performing stocks in each month.
- Work out how change in EPS (earning per share), shortratio or pricetarget affect change in the value of stock in future.