

DATA2001

Data Science, Big Data
and Data Variety

Week 3

Introduction to SQL



THE UNIVERSITY OF
SYDNEY

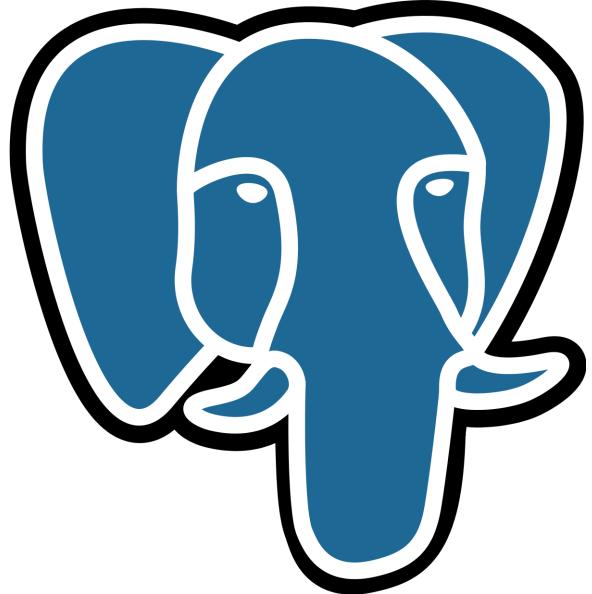
Outline

- Introduction – 10 mins
- Installing and familiarising with pgAdmin – 20 mins
- SQL exercises – 1 hour 15 minutes

pgAdmin

pgAdmin

- Development platform for PostgreSQL, which is an SQL-compliant relational database management system (RDBMS)
- Install through the [instructions on Canvas](#)
([download link here](#))

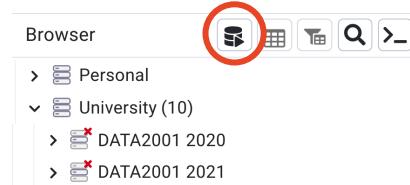


Did You Know?

- The elephant logo is named [Slonik](#)
- In Japan, the logo is a [turtle](#), originally because "its slow but it gets there"
- Early iterations involved a [cheetah](#) (to represent something fast) and a [devil penguin](#) (as a spoof of Linux)

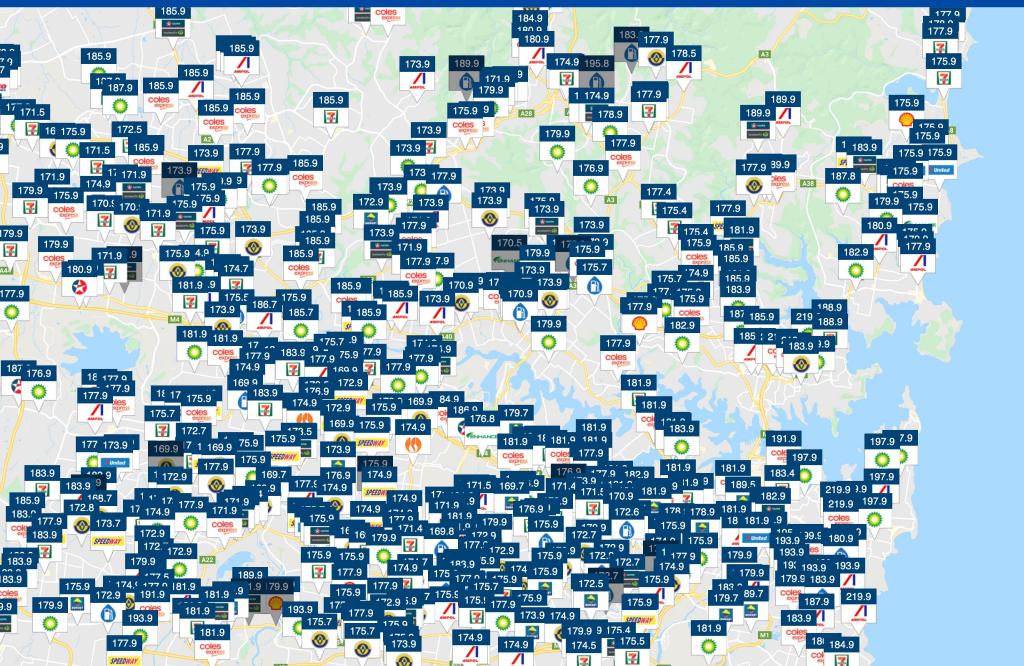
Setup

- Once installed, open pgAdmin and select "Add New Server"
 - Name: Your choice (e.g. "DATA2001")
 - Switch to the "Connection" tab
 - Host/name address: localhost
 - Username is postgres, and password is the one you set on installation
- Scroll down and click on your tab, then press the "Query Tool" button to get started!



Two screenshots of the pgAdmin interface. On the left, a 'Register - Server' dialog is open, showing the 'Connection' tab. It has fields for Host name/address (localhost), Port (5432), Maintenance database (postgres), Username (postgres), and Password (redacted). The 'Save password?' and 'Role' fields are also present. On the right, the main pgAdmin dashboard is shown. It features a 'Welcome' section with the pgAdmin logo and a brief description. Below it is a 'Quick Links' section with a 'Add New Server' button, which is circled in red. Other links include PostgreSQL Documentation, pgAdmin Website, Planet PostgreSQL, and Community Support. The bottom of the dashboard shows a navigation bar with icons for browser, personal, university, and search.

FuelCheck



Explore Metadata

15

[JSON](#)[RDF](#)[ISO19115/ISO19139 XML](#)

Data and Resources



FuelCheck 🔥

FuelCheck provides real-time information about fuel prices at service...

[Explore ▾](#)

FuelCheck FAQ 🔥

[Explore ▾](#)

FuelCheck Data Quality Statement - PDF 🔥

[Explore ▾](#)

FuelCheck Data Quality Statement - XML 🔥

[Explore ▾](#)

Fuelcheck Price History September 2016 🔥

[Explore ▾](#)

Fuelcheck Price History August 2016 🔥

[Explore ▾](#)

Fuelcheck Price History October 2016 🔥

[Explore ▾](#)

Fuelcheck Price History November 2016 🔥

[Explore ▾](#)

Fuelcheck Price History December 2016 🔥

[Explore ▾](#)

Fuelcheck Price History January 2017 🔥

[Explore ▾](#)

Publicly available through web API, or in monthly data dumps

Real-time data of fuel prices "at every station across NSW"

We'll be using a 'normalised' version of February 2023's data

Thought Questions



What benefits are there in analysing this data?

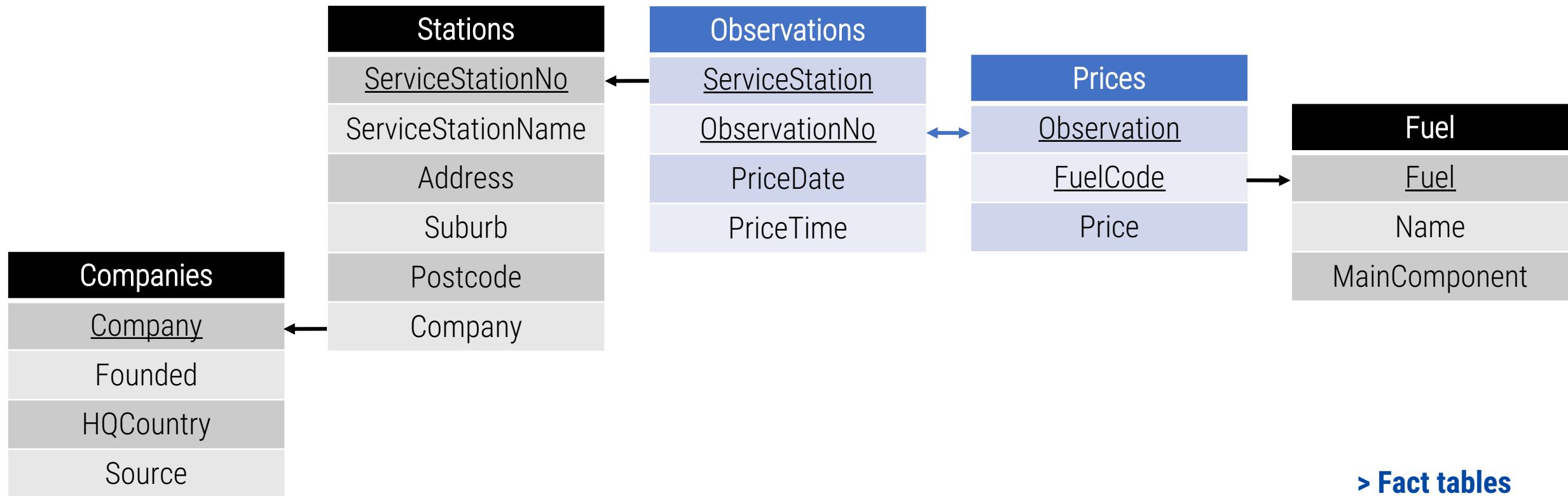


What other information would improve your analysis?



What are the potential implications of this data being publicly available?

Data Schema



> Fact tables

> Dimension tables

Populating the Data

- Download the 5 CSV files on Canvas (Wk 3 Tutorial section)
- Also download the “NSWFuelSchema.sql” file
 - This defines the data types and structure of our tables
- Run the SQL file in your pgAdmin window
 - This can be done by either opening the file, or copy-pasting its contents
- Right-click your server in the left sidebar and “Refresh”
- Find the tables (“Databases > postgres > Schemas > nswfuel > Tables”)
- Right-click each one at a time and import the data
 - Select “Import/Export Data” and load in the matching CSV

The screenshot shows the pgAdmin interface. On the left, there's a tree view of database objects under 'Tables (5)'. One item, 'companies', is highlighted with a blue selection bar. A context menu is open over this item, with 'Import/Export Data...' highlighted in a dark blue bar at the bottom.

On the right, there's a vertical sidebar with various options: 'Materialized Views', 'Operators', 'Procedures', 'Sequences', 'Tables (5)', 'companies', 'fuel', 'observations', 'prices', 'stations', and 'Trigger Functions'. Below these, there are buttons for 'Count Rows', 'Create', 'Delete/Drop', 'Refresh...', 'Restore...', 'Backup...', 'Drop Cascade', and 'Import/Export Data...'. The 'Import/Export Data...' button is also highlighted in a dark blue bar.

```
1 CREATE SCHEMA IF NOT EXISTS nswfuel;
2 SET search_path to nswfuel;
3
4 DROP TABLE IF EXISTS Prices;
5 CREATE TABLE Prices(
6     Observation INTEGER,
7     FuelCode VARCHAR(3),
8     Price NUMERIC
9 );
10
11 DROP TABLE IF EXISTS Observations;
12 CREATE TABLE Observations(
13     ServiceStation INTEGER,
14     ObservationNo INTEGER,
15     PriceDate DATE,
16     PriceTime TIME,
17     PRIMARY KEY (ServiceStation, ObservationNo)
18 );
19
20 DROP TABLE IF EXISTS Stations;
21 CREATE TABLE Stations(
22     ServiceStationNo INTEGER PRIMARY KEY,
23     ServiceStationName VARCHAR(100),
24     Address VARCHAR(100),
25     Suburb VARCHAR(50)
```

Populating the Data

- ALTERNATIVELY: There's also a "NSWFuelData.sql" on Canvas, which can be run in pgAdmin instead of loading in the CSVs sequentially
 - The NSWFuelSchema.sql file must be run first to establish the table structures
 - It's a larger file, filled with "INSERT" values that represent the CSV rows from each file
 - This can be executed by copy-pasting in its contents, or using the "Open file" button
- Run the query using the "Execute/Refresh" button

Query Query History

```
76376 INSERT INTO Companies VALUES ('United',1993,'Australia','https://en.wikipedia.org/wiki/United');
76377 INSERT INTO Companies VALUES ('Westside',2011,'Australia','https://www.lir.com.au/');
76378 INSERT INTO Companies VALUES ('Woodham Petroleum',NULL,'Australia','https://www.woodhampetroleum.com.au');
76379
76380
76381 INSERT INTO Fuel VALUES ('E10','Ethanol 94','Unleaded');
76382 INSERT INTO Fuel VALUES ('U91','Unleaded 91','Unleaded');
76383 INSERT INTO Fuel VALUES ('PDL','Premium Diesel','Diesel');
76384 INSERT INTO Fuel VALUES ('P98','Premium 98','Unleaded');
76385 INSERT INTO Fuel VALUES ('P95','Premium 95','Unleaded');
76386 INSERT INTO Fuel VALUES ('DL','Diesel','Diesel');
76387 INSERT INTO Fuel VALUES ('LPG','Liquid Petroleum Gas','Gas');
76388 INSERT INTO Fuel VALUES ('E85','Ethanol 105','Unleaded');
76389 INSERT INTO Fuel VALUES ('B20','Biodiesel 20','Diesel');
```

Data Output Messages Notifications

Using the Data

- The SQL file establishes a schema called "NSWFuel" to hold the data
- First define a 'search_path', to indicate which schema will be used, then try a simple query from one of the five tables
 - Note some are fairly large!

Query Editor Query History

```
1 set search_path to NSWFuel;
2
3 select * from Fuel;
```

Data Output Explain Messages Notifications

	fuel [PK] character varying (3)	name character varying (50)	maincomponent character varying (50)
1	E10	Ethanol 94	Unleaded
2	U91	Unleaded 91	Unleaded
3	PDL	Premium Diesel	Diesel
4	P98	Premium 98	Unleaded
5	P95	Premium 95	Unleaded
6	DL	Diesel	Diesel

1. Dimension Tables



a) List all recorded stations in your **surrounding suburbs**.

e.g. for stations near Redfern (2016), try searching for stations between postcodes 2014 and 2018

b) What **fuel types** are available in Premium?

If you drive, note the fuel you typically use from the list, we'll use it later.

c) Name the **five oldest** Australian fuel companies listed.

The oldest – Ampol – is itself an interesting case study of potential data quality issues, given its merge, and eventual departure from, Caltex.

2. Prices



- a) Return all occurrences where the price exceeded 235 cents per litre.
- b) Provide the distinct list of fuel codes that exceeded this price.
- c) Of these fuel codes, how many times were each recorded at above this price?
- d) While excluding premium fuels, list the 10 most expensive occurrences of fuel prices.

3. Observations



- a) List all observations made between **2am and 3am**.
- b) Of stations in the suburbs you investigated in Q1a), what was the **earliest and latest time** for which observations were made?
- c) For these stations and for a selected fuel type (e.g. 'PDL'), list all observations in ascending order of price, with a column indicating whether each observation was made on a **weekday or weekend**.
- d) **BONUS:** Order each of these stations by **average price** for your selected fuel type, and also note the total observations for each.

Break 10 minutes

