

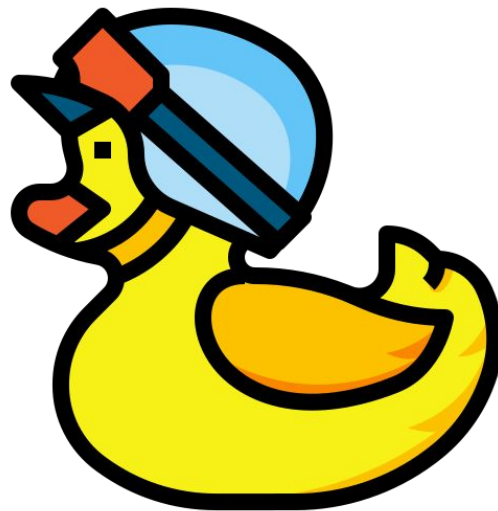
DuckDB



20 problems, 20 minutes -
one solution?

  Simon Aubury @SimonAubury

  Ned Letcher @nletcher



DE Meetup Syd

Slides





Hello!



Simon Aubury

Associate Director –
Data Platforms

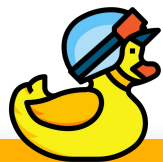
Simple
Machines



Ned Letcher

Lead Data Science
Engineer

/thoughtworks

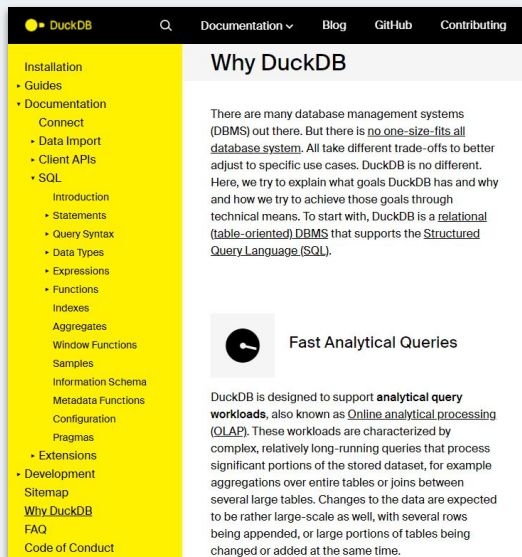


DuckDB – why?

Why do I care about another database?

Why DuckDB

DuckDB is free / open-source



https://duckdb.org/why_duckdb.html



RDBMS / SQL / OLAP - Extensive SQL with a large function library, window functions etc. ACID transactional



No external dependencies - compiled into two files, embedded within a host process. DuckDB uses **PostgreSQL's SQL parser**, Google's RE2 regular expression engine and SQLite's shell

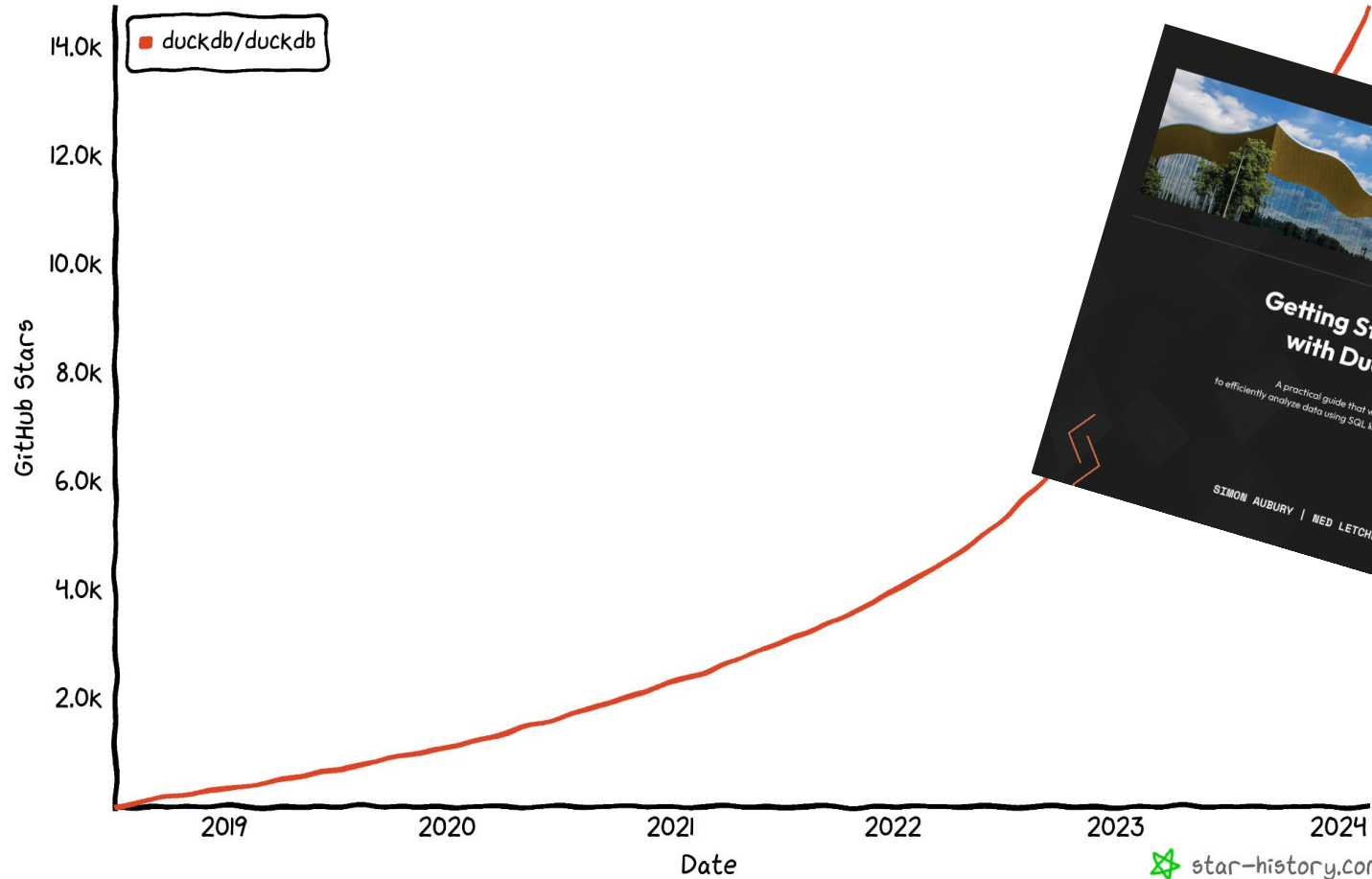


Process foreign data without copying, run queries directly on Pandas data, Python and R, APIs for Java, C, C++



Extensions - Httpfs, s3, parquet, json, FTS, geospatial, iceberg

Star History



Comparing DuckDB

in-process



DuckDB

DuckDB is
MIT-licensed in
perpetuity!



chDB

standalone



PostgreSQL



ClickHouse



pinot



snowflake



druid



MySQL



Microsoft
SQL Server



Google BigQuery



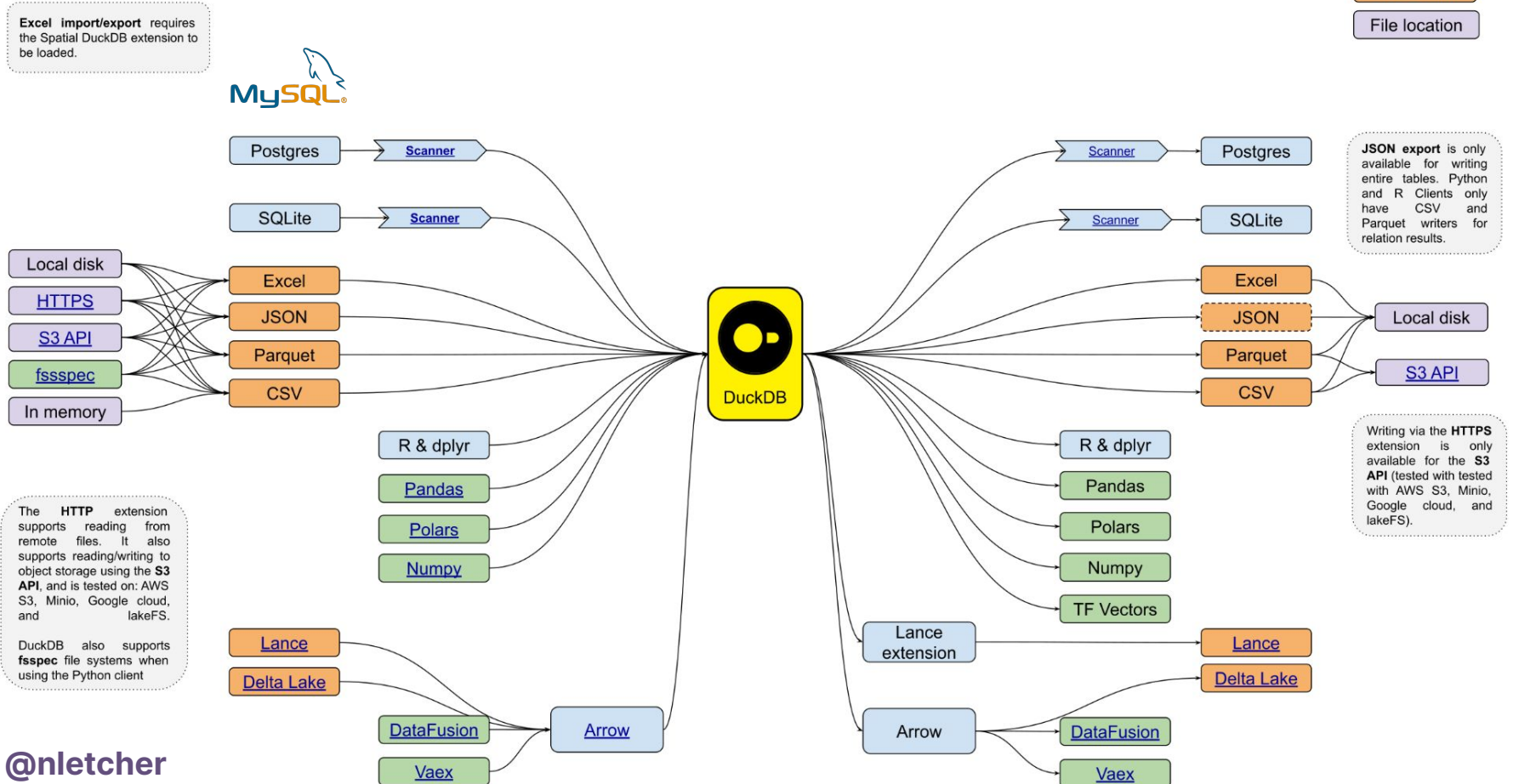
amazon
REDSHIFT

OLTP

OLAP

DuckDB Data I/O Pathways

This diagram represents user interface flows through API calls, rather than underlying data flow between formats etc.



Installation

<https://duckdb.org/docs/installation/index>

```
brew install duckdb
```

CLI / macOS

```
./duckdb
```

```
pip install duckdb
```

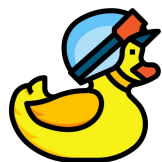
Python

```
import duckdb
cursor = duckdb.connect()
print(cursor.execute('SELECT 42').fetchall())
```

```
npm install duckdb
```

node.js

```
var duckdb = require('duckdb');
var db = new duckdb.Database(':memory:');
db.all('SELECT 42 AS fortytwo', function(err, res) {
  console.log(res[0].fortytwo)
});
```

Demo

**Hands-on DuckDB
(with the DuckDB CLI)**



https://github.com/saubury/duckdb_meetup

Demo CSV files

pizza_n.csv

```
food_name,color,calories,is_healthy
Margherita,mixed,60,FALSE
Chicago,yellow,70,FALSE
Pepperoni,mixed,80,FALSE
Hawaiian,yellow,60,FALSE
Neapolitan,red,45,FALSE
BBQ,red,62,FALSE
Veggie,mixed,32,TRUE
```

fast_food.csv

```
food_name,color,calories
burger,mixed,60
fries,yellow,35
sandwich,white,22
```

salad.csv

```
food_name,calories,is_healthy,color
sushi,60,TRUE,mixed
pho,70,TRUE,yellow
```

Demo FitBit files

exercise-nnnn.json

```
{
  "logId" : 52683872653,
  "activityName" : "Walk",
  "activityTypeId" : 90013,
  "activityLevel" : [{
    "minutes" : 0,
    "name" : "sedentary"
  },{
    "minutes" : 6,
    "name" : "lightly"
  },{
    "minutes" : 0,
    "name" : "fairly"
  },{
    "minutes" : 10,
    "name" : "very"
  }],
  "averageHeartRate" : 83,
  "calories" : 126,
  "duration" : 921000,
  "activeDuration" : 921000,
  "steps" : 1112,
  "logType" : "auto_detected",
  "manualValuesSpecified" : {
    "calories" : false,
    "distance" : false,
    "steps" : false
  },
  "heartRateZones" : [{
    "name" : "Out of Range",
    "min" : 30,
    "max" : 101,
    "minutes" : 15,
    "caloriesOut" : 125.697
  },{
    "name" : "Fat Burn",
    "min" : 101,
```

Parquet files

The screenshot shows the NYC Taxi & Limousine Commission website. The browser address bar displays `nyc.gov/site/tlc/about/tlc-trip-record-data.page`. The website header includes the NYC logo, the text 'Taxi & Limousine Commission', and a search bar. The main navigation menu has links for 'About', 'Passengers', 'Drivers', 'Vehicles', 'Businesses', and 'TLC Online'. Below this is a secondary navigation bar with buttons for 'About TLC', 'Data and Reports', 'TLC Initiatives', and 'Contact TLC'. The 'Data and Reports' section is active, showing a sidebar with links for 'Data', 'Pilot Programs', 'Reports', 'TLC Trip Record Data', and 'Request Data'. The main content area is titled 'TLC Trip Record Data' and contains two paragraphs of text. The first paragraph describes yellow and green taxi trip records, and the second paragraph describes For-Hire Vehicle (FHV) trip records. Below the text are social media sharing icons and a 'Print' button. At the bottom, there is an 'ATTENTION!' section with a date and a list of changes to trip record files.

nyc.gov/site/tlc/about/tlc-trip-record-data.page

NYC Taxi & Limousine Commission

311 Search all NYC.gov websites

한국어 Translate Text-Size

About Passengers Drivers Vehicles Businesses TLC Online Search

About TLC Data and Reports TLC Initiatives Contact TLC

Data

Pilot Programs

Reports

[TLC Trip Record Data](#)

Request Data

f x t Share

Print

TLC Trip Record Data

Yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

ATTENTION!

On 05/13/2022, we are making the following changes to trip record files:

1. All files will be stored in the PARQUET format. Please see the 'Working With PARQUET'

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

iMessage SQLite

iMessage uses a local SQLite database to store a copy of message conversations

Optional - if you want to get your own iMessage database

```
cp ~/Library/Messages/chat.db .
```



iMessage



SQLite

Demo book review files

Books_rating.csv - a collection of 300,000 book reviews

Title,description,authors,image,previewLink,publisher,publishedDate,infoLink,categories,ratingsCount

A Walk in the Woods: a Play in Two Acts,,['Lee Blessing'],,http://books.google.nl/books?id=6HD0wAEACAAJ&dq=A+Walk+in+the+Woods:+a+Play+in+Two+Ac
source=gb_s_api,,1988,http://books.google.nl/books?id=6HD0wAEACAAJ&dq=A+Walk+in+the+Woods:+a+Play+in+Two+Acts&hl=&source=gb_s_api,,3.0

Saint Hyacinth of Poland,"The story for children 10 and up of St. Hyacinth, the Dominican who planted the Faith in Poland, Lithuania and Russia
He went to Rome, where he met St. Dominic, and was one of the first to receive at his hands the habit of the newly established Order of Friars P
novitiate he made his religious profession, and was made superior of the little band of missionaries sent to Poland to preach. Impr. 189 pgs 16
Windeatt'],http://books.google.com/books/content?id=lmLqAAAACAAJ&printsec=frontcover&img=1&zoom=1&source=gb_s_api,http://books.google.nl/books?id
+Hyacinth+of+Poland&hl=&cd=1&source=gb_s_api,Tan Books & Pub,2009-01-01,http://books.google.nl/books?id=lmLqAAAACAAJ&dq=Saint+Hyacinth+of+Poland&
['Biography & Autobiography'],

Rising Sons and Daughters: Life Among Japan's New Young,"Wardell recalls his experience as a foreign student in Japan, providing a comparison be
['Steven Wardell'],,http://books.google.nl/books?id=rbLZugEACAAJ&dq=Rising+Sons+and+Daughters:+Life+Among+Japan%27s+New+Young&hl=&cd=1&source=gb
1995,http://books.google.nl/books?id=rbLZugEACAAJ&dq=Rising+Sons+and+Daughters:+Life+Among+Japan%27s+New+Young&hl=&source=gb_s_api,['Social Scien

Muslim Women's Choices: Religious Belief and Social Reality (Cross Cultural Perspectives on Women),"Counters the Western views and stereotypes o
presenting a cross-cultural perspective of their experiences and choices in contemporary Muslim communities. The main theme of these papers is th
women manipulate religious belief to negotiate their gender roles, within the context of their lives.",['Camillia Fawzi El-Solh', 'Judy Mabro']
books/content?id=o7izAAAAIAAJ&printsec=frontcover&img=1&zoom=1&source=gb_s_api,http://books.google.nl/books?id=o7izAAAAIAAJ&q=Muslim+Women%27s+Ch
+and+Social+Reality+(Cross+Cultural+Perspectives+on+Women)&dq=Muslim+Women%27s+Choices:+Religious+Belief+and+Social+Reality+(Cross+Cultural+Pers
cd=1&source=gb_s_api,Berg Pub Limited,1994-02-17,http://books.google.nl/books?id=o7izAAAAIAAJ&dq=Muslim+Women%27s+Choices:+Religious+Belief+and+S
+Cultural+Perspectives+on+Women)&hl=&source=gb_s_api,['Religion'],

Dramatica for Screenwriters,"Dramatica for Screenwriters by Armando Saldana Mora is a must for any writer who wishes to get the most out of the
development software. Written by a working screenwriter, this book gives real world, practical applications for using Dramatica to develop and w
topics include: * What Dramatica offers the screenwriter * Dramatica in thirty seconds * Characters: Beyond the archetypes * Character Relations
and sketch story development * How to get a complete plot * Four Dramatica Acts, Three Classical Acts * Narrative flow & writing the screenplay
['Armando Salda A-Mora'],http://books.google.com/books/content?id=iTueuAAACAAJ&printsec=frontcover&img=1&zoom=1&source=gb_s_api,http://books.goog
id=iTueuAAACAAJ&dq=Dramatica+for+Screenwriters&hl=&cd=1&source=gb_s_api,,2005-07,http://books.google.nl/books?id=iTueuAAACAAJ&dq=Dramatica+for+Sc
source=gb_s_api,['Reference'],

Demo Excel XLSX file



AutoSave ... houses ▾

Home Insert Draw Page Layout Formulas Data Review View Automate Tell me

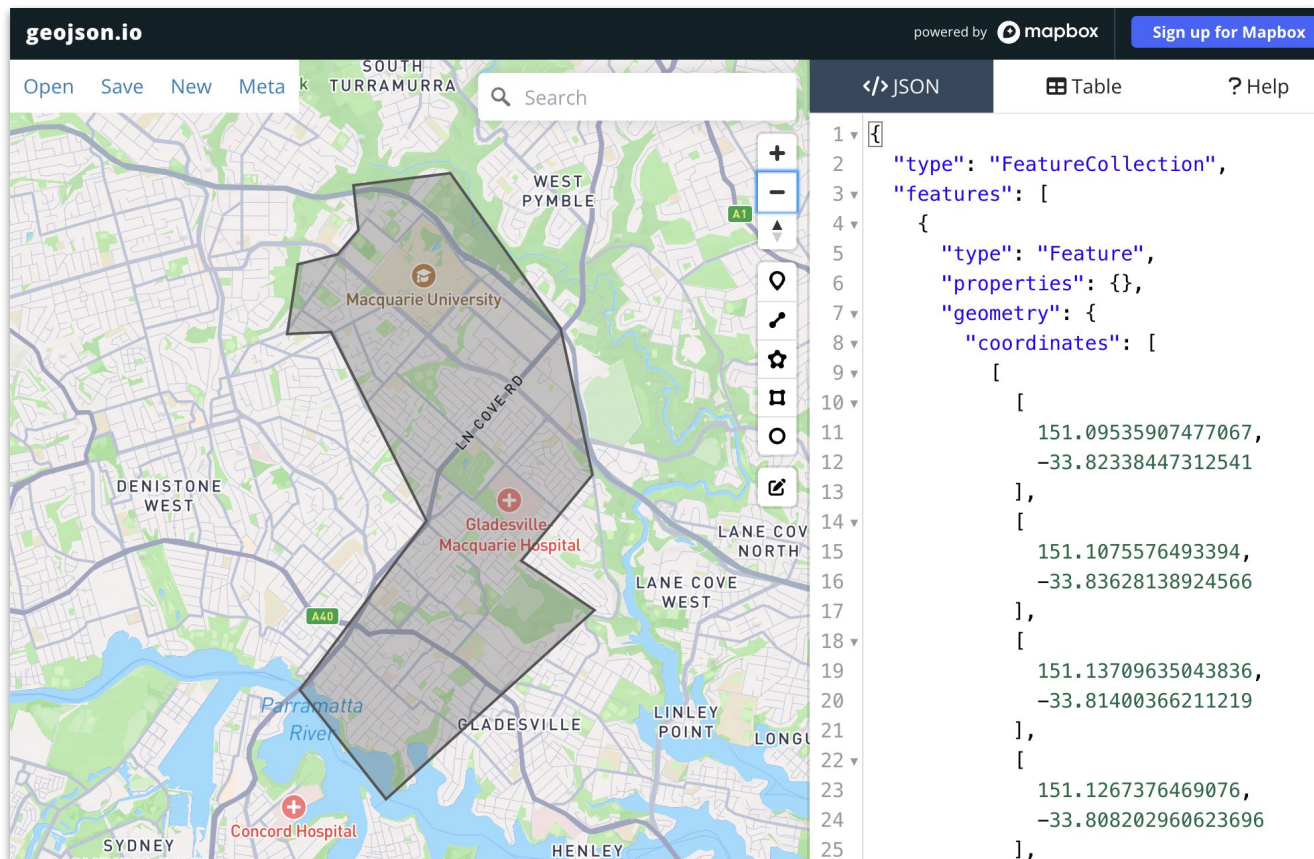
Paste Arial 10 A⁺ A⁻ B I U Wrap Text General \$ % '

F17

	A	B	C	D	E	F
1	house	longitude	latitude			
2	200 Raymond St Marsfield	151.1180957	-33.81005862			
3	6a Herring Rd Marsfield	151.1185	-33.8132			
4	200 Balaclava Rd Eastwood	151.0953261	-33.77973288			
5	1 Lincoln St Eastwood	151.0949594	-33.78044307			
6	100 Lincoln St Eastwood	151.0977619	-33.78204917			
7	Sydney Opera House	151.215256	-33.856159			
8	Sydney Harbour Bridge	151.210556	-33.852222			
9						
10						
11						

Houses.xlsx - A Microsoft Excel file

Geo



The screenshot shows the geojson.io interface. The map on the left displays a shaded polygon representing a school boundary in the Gladesville area of Sydney. Key landmarks visible include Macquarie University, Macquarie Hospital, and the Parramatta River. The right panel shows the JSON data for the selected feature, which is a FeatureCollection containing a single Feature with a polygon geometry.

```
1 {
2   "type": "FeatureCollection",
3   "features": [
4     {
5       "type": "Feature",
6       "properties": {},
7       "geometry": {
8         "coordinates": [
9           [
10            [
11              151.09535907477067,
12              -33.82338447312541
13            ],
14            [
15              151.1075576493394,
16              -33.83628138924566
17            ],
18            [
19              151.13709635043836,
20              -33.81400366211219
21            ],
22            [
23              151.1267376469076,
24              -33.808202960623696
25            ],
26            [
27              151.09535907477067,
28              -33.82338447312541
29            ]
30          ]
31        ]
32      }
33    ]
34  }
```


Public API

The screenshot shows the TVMaze website interface. At the top, there's a navigation bar with the TVMaze logo, a search bar, and links for Login and Register. Below the navigation bar, there's a section for "Popular shows airing tonight" featuring five show cards: CSI: Vegas, The Simpsons, Tracker, The Walking Dead: The Ones Who Live, and Curb Your Enthusiasm. To the right of this section is a "Schedule for Mar 24" table listing shows and their air times. Below the popular shows section is a "More shows" button. Further down is a section for "Upcoming Season Premieres" featuring five show cards: Mean Girl Murders, Narco Wars, We Were the Lucky Ones, Jerrod Carmichael Reality Show, and 48 Hours to Buy. To the right of this section is a "Countdown" button.

Popular shows airing tonight

Show	Network	Air Time
CSI: Vegas	CBS	20:00
The Simpsons	FOX	20:00
Tracker	CBS	20:00
The Walking Dead: The Ones Who Live	AMC	21:00
Curb Your Enthusiasm	HBO	21:00

More shows »

Upcoming Season Premieres

Show	Network	Air Time
Mean Girl Murders	Netflix	25
Narco Wars	Netflix	27
We Were the Lucky Ones	Hulu	28
Jerrod Carmichael Reality Show	Netflix	29
48 Hours to Buy	Netflix	30

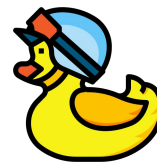
Countdown »

Schedule for Mar 24

Time	Show	Network
20:00	Tracker	CBS
20:00	The Simpsons	FOX
20:00	American Idol	ABC
20:30	Krapopolis	FOX
21:00	The Walking Dead: The Ones Who Live	AMC
21:00	CSI: Vegas	CBS
21:00	In Restless Dreams: The Music of Paul Simon	MGM+
21:00	The Regime	HBO
21:00	The Great North	FOX
21:30	Grimsburg	FOX
22:00	What Would You Do?	ABC

<https://www.tvmaze.com/api>

Thanks!



[Pre-order link](#)



  **Simon Aubury** @SimonAubury

  **Ned Letcher** @nletcher

 https://github.com/saubury/duckdb_meetup

❤️ Presentation template by [SlidesCarnival](#), Icons [Flaticon](#) photos [Unsplash](#) and [Pixabay](#)