Build your own data warehouse for personal analytics with SQLite and Datasette

This document: https://bit.ly/pyconau-dogsheep

Friday September 4th, 4-4:25pm - https://2020.pycon.org.au/program/73uk8x/

Simon Willison - https://simonwillison.net/ - https://twitter.com/@simonw

(Feel free to add notes to this document as I go along)

Some links to get you started

- Datasette: https://docs.datasette.io/
- Dogsheep: https://dogsheep.github.io/ - https://github.com/dogsheep
- https://simonwillison.net/2020/May/21/dogsheep-photos/
- https://til.simonwillison.net/til/til/macos_find-largest-sqlite.md
- Dogsheep Beta: https://github.com/dogsheep/beta

Some demos:

- register-of-members-interests.datasettes.com/regmem/items
- register-of-members-interests.datasettes.com/regmem/items?_search=hamper
- https://australian-dunnies.now.sh/australian-dunnies/dunnies
- https://australian-dunnies.now.sh/australian-dunnies/dunnies?_facet=FacilityType&_facet=ToiletType

Notes go here

- https://writings.stephenwolfram.com/2019/02/seeking-the-productive-life-some-details-of-my-personal-infrastructure/
- https://sqlite-utils.readthedocs.io/en/stable/
- https://github.com/simonw/datasette-bplist is the plugin I used to decode bplist data in the Apple Photos database

Add your questions here (I'll address them in the Q&A)

- What's the cutest costume your dog has worn? (OK I fixed this one…)
    - She won best costume at DogFest dressed as the Golden Gate Bridge!

- I'm assuming if you're data is a json source, you need to convert that to sqllite before running datasette, but is that a fair assumption? (Geoff Crompton)

I have a bunch of tools I use for this. http://github.com/simonw/csvs-to-sqlite converts one or more CSV files to SQLite. https://github.com/simonw/sqlite-utils is my main tool: it's a combination CLI tool and Python library for ingesting data (as JSON, TSV or CSV) and loading it into SQLite, plus utilities for adding foreign keys, configuring full-text search and more.

- Is there support for styling the datasette interface? (Ned)

Yes! You can provide custom templates (Jinja) and custom CSS as well. Info on how to do that here: https://docs.datasette.io/en/stable/custom_templates.html

The best example I have of a styled instance in the wild is this project from the Baltimore Sun: https://salaries.news.baltimoresun.com/

- [EDIT: Simon just answered my question - he uses cron on a server. But were any other alternatives considered?) What's your recommended way of getting the data out of source systems on a regular schedule? (Allan)

My personal Dogsheep runs off a big list of crons, and it works really well. I've thought about moving to a more sophisticated database-backed scheduling system for larger volumes of automation but that's just an idea I'm brewing at the moment. Cron works really well in my experience.

I also make extensive use of scheduled GitHub Actions for building public Datasette instances. https://github.com/simonw/covid-19-datasette is a repo that runs once an hour to build https://covid-19.datasettes.com/ for example.

- Do you know if we can extract Google Photos data in a similar way to Apple Photos? It does a similar ML tagging of photos

I tried to build my photos stuff against Google Photos first, but their API has one infuriating limitation: they don't provide access to the latitude and longitude of your photos, which is the data I am most interested in! There's a very long issue thread complaining about that here: https://issuetracker.google.com/issues/80379228

- Any plans to steal Palantir's market share?

They can keep the creepy high-end giant data warehouse business. I'm going after journalists and data enthusiasts who want to hack around with their own personal data in the privacy of their own network!

- Are there plans to make the source Db interchangeable? Eg use MySQL, Postgres in the future?

I've considered this, and even did a quick spike to see if it could work against PostgreSQL. It's definitely possible but I'm nervous about having to maintain yet another database compatibility layer, so I've not yet put that on my active roadmap. I do have https://github.com/simonw/db-to-sqlite which is a tool for converting an existing MySQL or PostgreSQL database into SQLite so you can use it with Datasette - I run that in my own Dogsheep periodically to suck down my blog's (https://simonwillison.net/) Django database from Heroku PostgreSQL.

- Has having control about your data like that changed your perspective on privacy or the way others aggregate data?

It hasn't changed my perspective really but I have been greatly enjoying seeing what's out there, and taking advantage of the fact that the European GDPR law means that companies all have to give you an export option these days. LinkedIn, Facebook, Google all have "export" buttons hidden deep in their preferences which email you a huge zip file full of XML and JSON. Part of the idea behind Dogsheep is to build open source tools to convert those to SQLite so people can actually start exploring them.

- Can you do queries with joins between datasets?

Yes, absolutely. You have to load the tables into the same SQLite database file at the moment (though SQLite can support cross-file joins, I just haven't exposed that capability to Datasette yet). Here's an example where I join the US census county population data with

the New York Times latest Covid-19 case counts per county:
https://covid-19.datasettes.com/covid/latest_ny_times_counties_with_populations