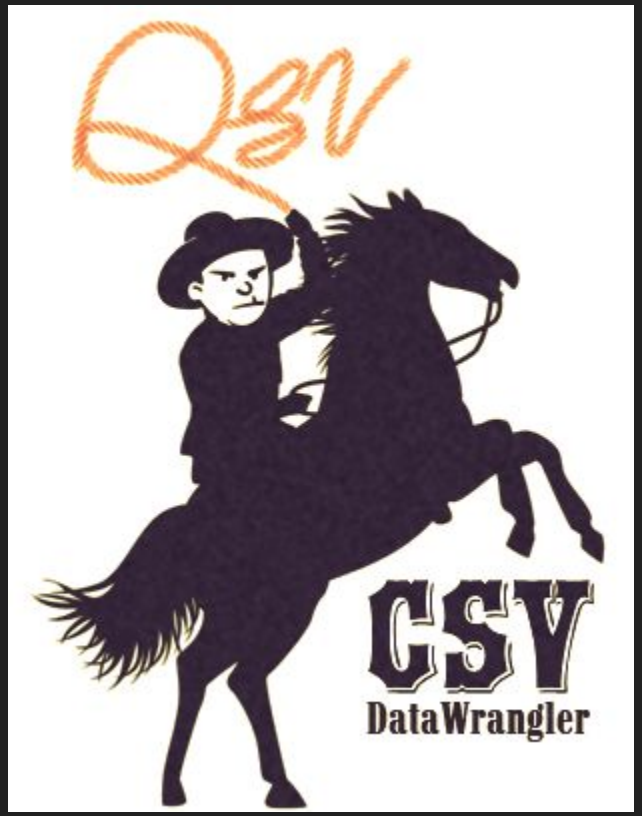
qsv

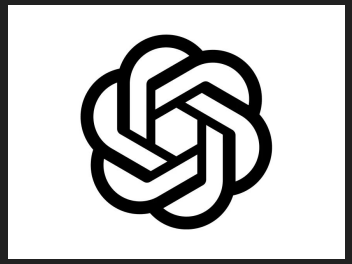
A blazing-fast, multi-platform, command-line, Data-wrangling toolkit

Joel Natividad
csv,conf,v8
May 2024

datHere
Data Infrastructure Engineering


The Looney Tunes Show
BOOMERANG


CSV
DataWrangler



https://github.com/jqnatividad/qsv/releases/tag/0.128.0

A little history....

Born of Open Data

ckan

1st US Professional Services Partner

2013

NYC 311

Reinvent 311
Winner
Jan 2014

IBM WATSON

Finalist
IBM Watson Mobile Challenge
May 2014

2016
Apr

OPENGOV

Search Login Request a Demo

OpenGov Acquires Open Data Leader Ontodia

Combination of Ontodia's Open Data and Performance Management Capabilities with OpenGov's leading Financial Intelligence and Transparency Platform Ushers in Next Wave of Open and Efficient Government

REDWOOD CITY, Calif. –April 13, 2016 – Today OpenGov, the world leader in government financial intelligence, planning, and transparency, expands its platform with the acquisition of Ontodia. Ontodia is the leading provider of Open Data and performance management solutions using CKAN, the premier open-source data-portal for governments around the world. CKAN powers Data.gov, the home of U.S. Government's Open Data initiative

~50 installations across the US and there was one recurring problem…

Data Quality

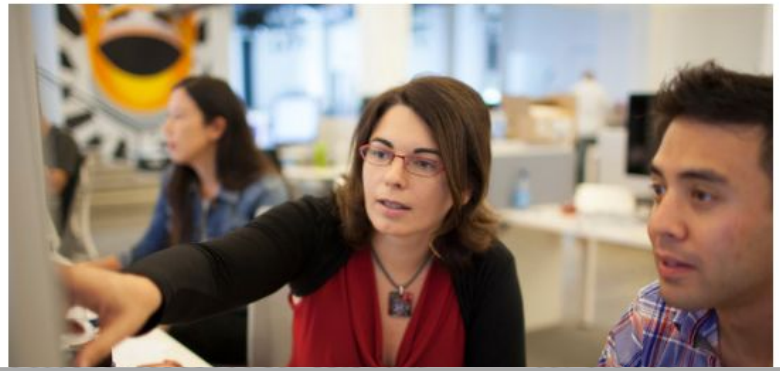
The New York Times

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.



The field known as "big data" offers a contemporary case study. The catchphrase stands for the modern abundance of digital data from many sources — the web, sensors, smartphones and corporate databases — that can be mined with discoveries and insigh smarter, data-driven every field. That is wh economy's hot new jo

Yet far too much hand data scientists call "da munging" and "data

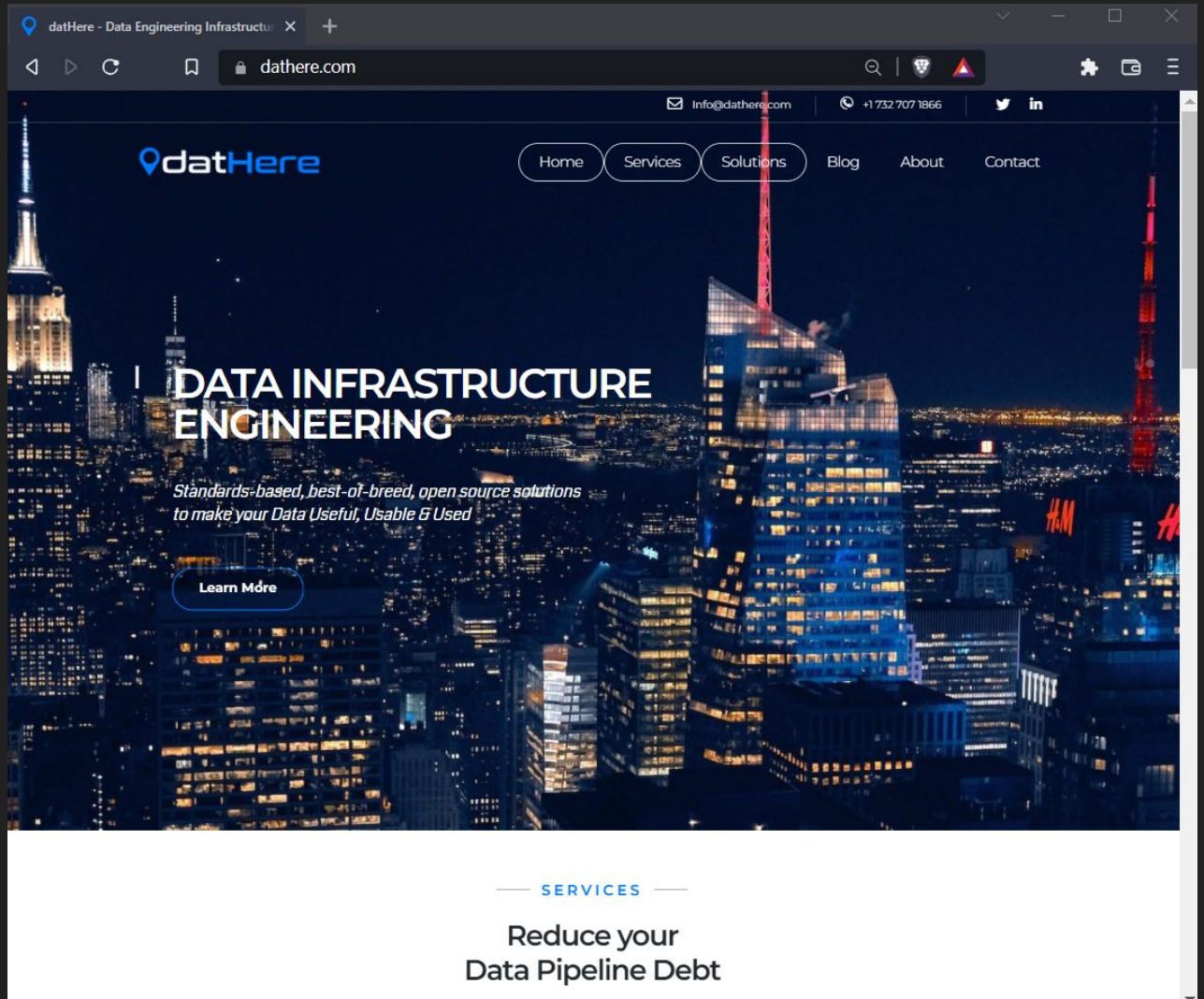
red. Data scientists, according to interviews a
l from 50 percent to 80 percent of their time
lane labor of collecting and preparing unruly

“Data Wrangling” Challenges

- Brittle data pipelines
- Larger & larger datasets
- “Regular” tools cannot scale (i.e. Excel)
- Specialized tools
  - Platform specific
  - Expensive
- Specialized “data science” skills
- Slow
  - Ramp-up time
  - Preparation time
  - Execution time

datHere launching 2020

The band gets back together to take on Data Quality…



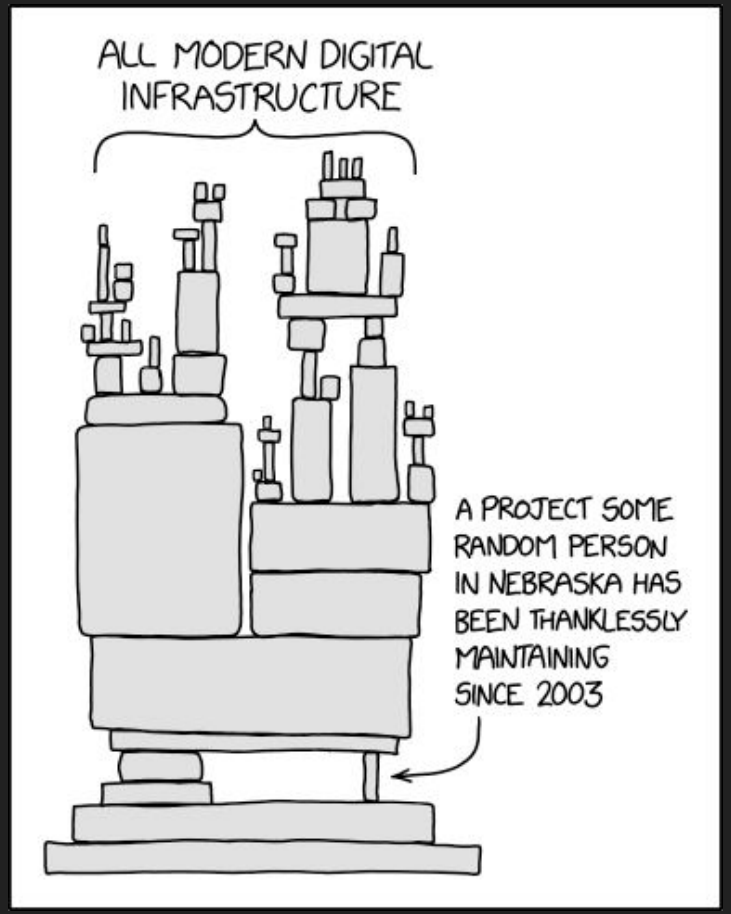
datHere

Standards-based, best-of-breed, open source solutions to make your Data Useful, Usable & Used

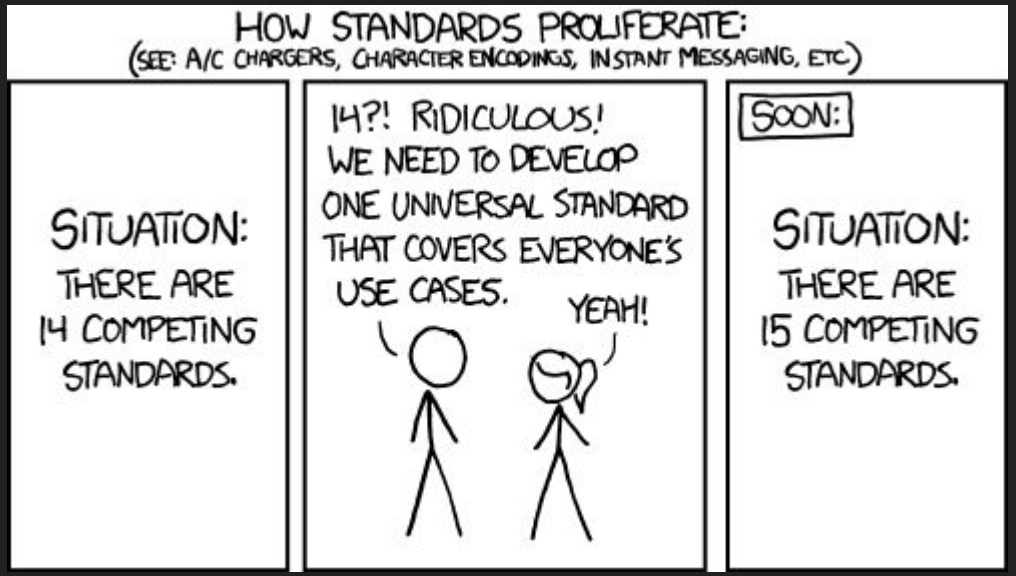
We deploy & co-create Data Infrastructure:

- Open Data Portals
- Internal Data Exchange
- Data Libraries
- Data Pipelines
- Water Data Practice

Open Source



Open Standards



But then

2020

happened…

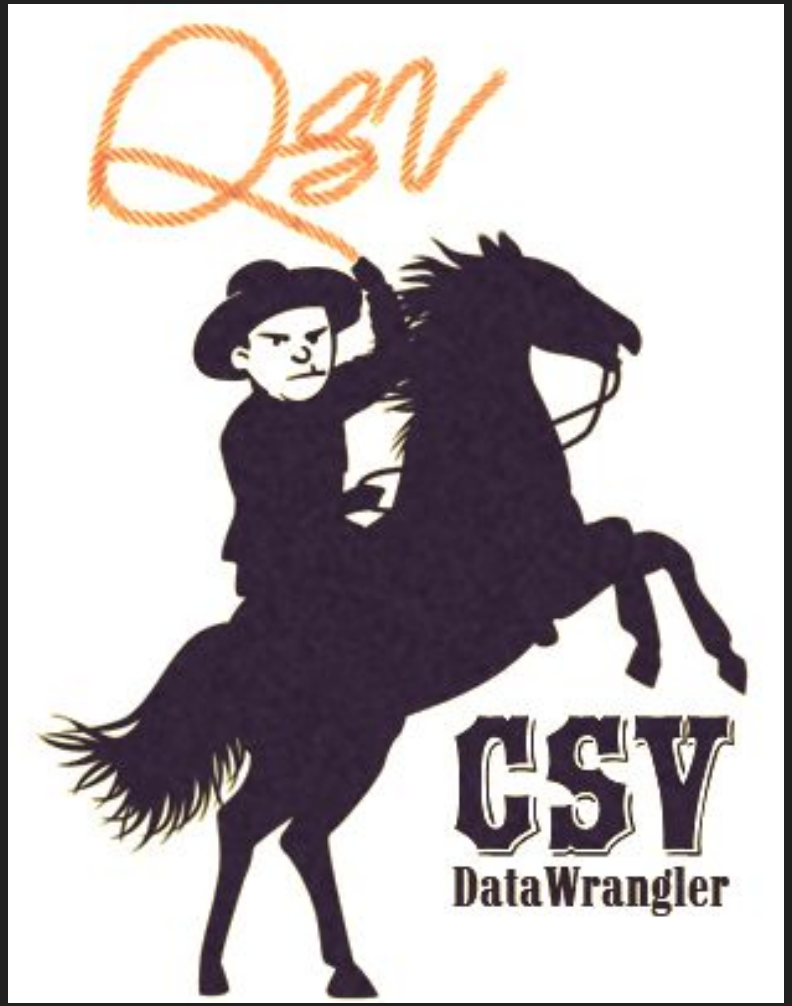
When Satan Met 2020

4,684,366 views • Dec 3, 2020

196K 4.2K SHARE SAVE

We needed a "Data Wrangler"

- Works with a universal data format
- Cross-platform
- Fast, blazing Fast!
- Open Source
- Easy to Learn
- Easy to Use for initial investigations
- But powerful enough to integrate into mission-critical data pipelines


CSV
CSV
DataWrangler

Origins

It all started with a **failed pilot** with a Hedge Fund to build an Internal Data Portal

- Brand new startup during COVID
- Data Portals anybody?
- An Internal Data Catalog Pilot, populated with latest metadata
- Traditional metadata ingestion pipeline (csvkit) was too slow
- Forked xsv to start qsv…

qsv **"Data Wrangler"** Goals

- Works with a universal data format
- Cross-platform
- Open Source
- Easy to Learn
- Easy to Use for initial investigations
- But powerful enough to integrate into mission-critical data pipelines

CSV, Excel, JSON, JSONL, PostgreSQL, SQLite, Parquet, Data Package, AVRO + recognizes 130 file formats

Linux, macOS + Windows

Fast! Blazing "Speedy Gonzales" Fast!!!

How fast is Blazing, “Speedy Gonzales” fast?

For a **1 million row sample of NYC’s 311 data (41 columns, 520 mb)**:

- 11 “streaming” summary statistics in **0.204 secs**
- 21 more statistics & infer dates(19 formats recognized) in **1.97 secs**
- Frequency table in **1.129 secs**
- Count rows in **0.05 secs**
- Validate against RFC 4180 CSV standard in **0.5 secs**
- Validate against a JSON Schema in **1.266 secs**
- Run a simple SQL query in **0.05 secs**, a SQL aggregation in **0.082 secs** & a very inefficient SQL aggregation in **0.144 secs**
- Reverse geocode WGS84 coordinate against Geonames in **3.59 secs**
- And more…

https://qsv.dathere.com/benchmarks

field	type	sum	min	max	range	min_length	max_length	mean	stddev	variance	nullcount	max_precision	sparsity
Unique Key	Integer	32687965858032	11465364	48478173	37012809	8	8	32687965.86	9013895.336	81250309125279	0		0
Created Date	DateTime		2010-01-01T00:00:00+00:00	2020-12-23T01:25:51+00:00	4009.05962			2015-11-10T18:05:22.615+00:00	1155.01606	1334062.092	0		0
Closed Date	DateTime		1900-01-01T00:00:00+00:00	2100-01-01T00:00:00+00:00	73049			2015-11-14T10:16:16.743+00:00	1314.70016	1728436.508	28619		0.0286
Agency	String		3-1-2001	TLC		3	42				0		0
Agency Name	String		3-1-2001	Valuation Policy		3	82				0		0
Complaint Type	String		../../WEB-INF/web.xml;x=	ZTESTINT		3	41				0		0
Descriptor	String		1 Missed Collection	unknown odor/taste in drinking water (QA6)		0	80				3001		0.003
Location Type	String		1-, 2- and 3-Family Home	Wooded Area		0	36				239131		0.2391
Incident Zip	String		*	XXXXX		0	10				54978		0.055

```sql
SELECT
    A.Agency,
    A.Borough,
    COUNT(*) AS total_incidents,
    SUM(
        CASE
            WHEN A."Complaint Type" LIKE 'Noise%' THEN 1
            ELSE 0
        END
    ) AS noise_related_incidents,
    SUM(
        CASE
            WHEN A.Status = 'Closed' THEN 1
            ELSE 0
        END
    ) AS closed_incidents,
    SUM(
        CASE
            WHEN A.Status != 'Closed' THEN 1
            ELSE 0
        END
    ) AS open_incidents,
    SUM(
        CASE
            WHEN POSITION('Water' IN A."Complaint Type") > 0 THEN 1
            ELSE 0
        END
    ) AS water_related_incidents,
    MAX(LENGTH (A."Complaint Type")) AS max_complaint_type_length,
    SUM(
        CASE
            WHEN UPPER(A."Complaint Type") = UPPER(A."Complaint Type")
            THEN LENGTH (A."Complaint Type")
            ELSE 0
        END
    ) AS sum_complaint_type_lengths,
    COUNT(DISTINCT A."Complaint Type") AS distinct_complaint_types
FROM
    read_csv ('NYC_311_SR_2010-2020-sample-1M.csv') A
GROUP BY
    A.Agency,
    A.Borough
ORDER BY
    total_incidents DESC;
```

ANSWERED IN 0.144 SECONDS!

How is it so Fast?

by standing on the Shoulders of Giants & the Ecosystem

- Rust
- Multi-threaded, Multi-I/O
- Performance architecture
  - Indexed access
  - Various caching techniques
  - Performance oriented memory allocator
- Built on a solid foundation (xsv)
- Polars Dataframes Engine
- Vibrant Rust & Polars Ecosystems

Why the Obsessive Need for Speed?

What does it unlock?

- Big Data is getting Bigger
- Embedding into other Systems
- Quicker Data Investigations
- Enables new Data Workflows
  - Preemptive metadata inferencing
  - Compile Extended Data Dictionaries
  - Leverage AI

Datapusher+

Embedded use case

- Next-gen Data Ingestion extension for CKAN
- Guaranteed Data Type inferences
- Data Validation
  - Dedupe
  - PII screening
  - As context for AI - “describeGPT”
  - Extended Data Dictionary
  - Pre-calculate metadata (spatial extent, date range for time-series data, etc.)
  - Pre-populate DCAT 3 recommended metadata fields

https://ckan.org/events/ckan-datapusher-plus-automagical-metadata

Standards-based, best-of-breed, open source solutions to make your Data Useful, Usable & Used

Data that is Useful, Usable & Used

}

We have a solution for this with DP+ & qsv

?

*But what about actually **Using the Data***

*to gain **Actionable Insight**,*

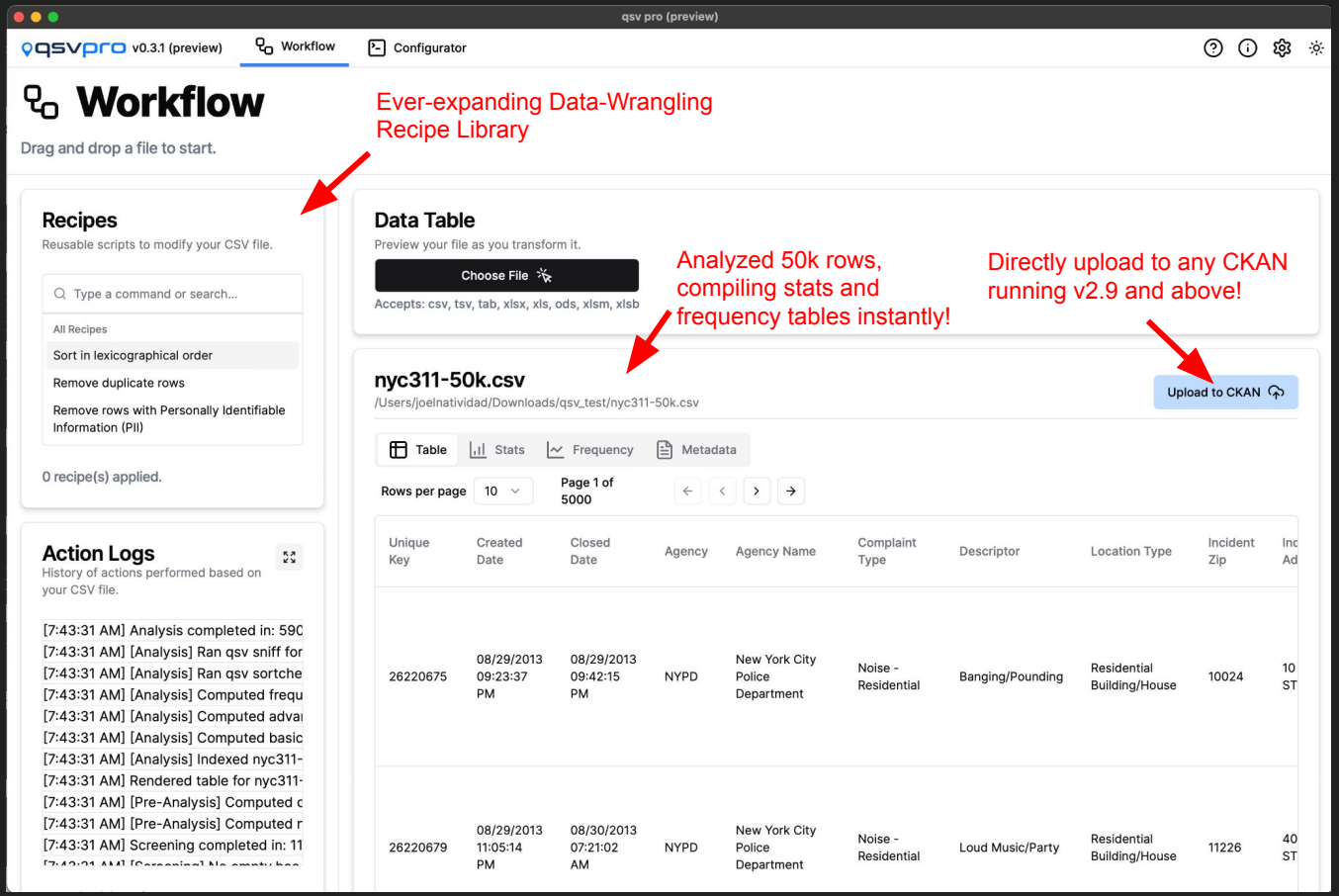
*to drive **Evidence-based Decisions**?*

qsv pro

Cross-Platform Desktop Data-Wrangling & Query tool for the Rest of Us

https://qsvpro.dathere.com

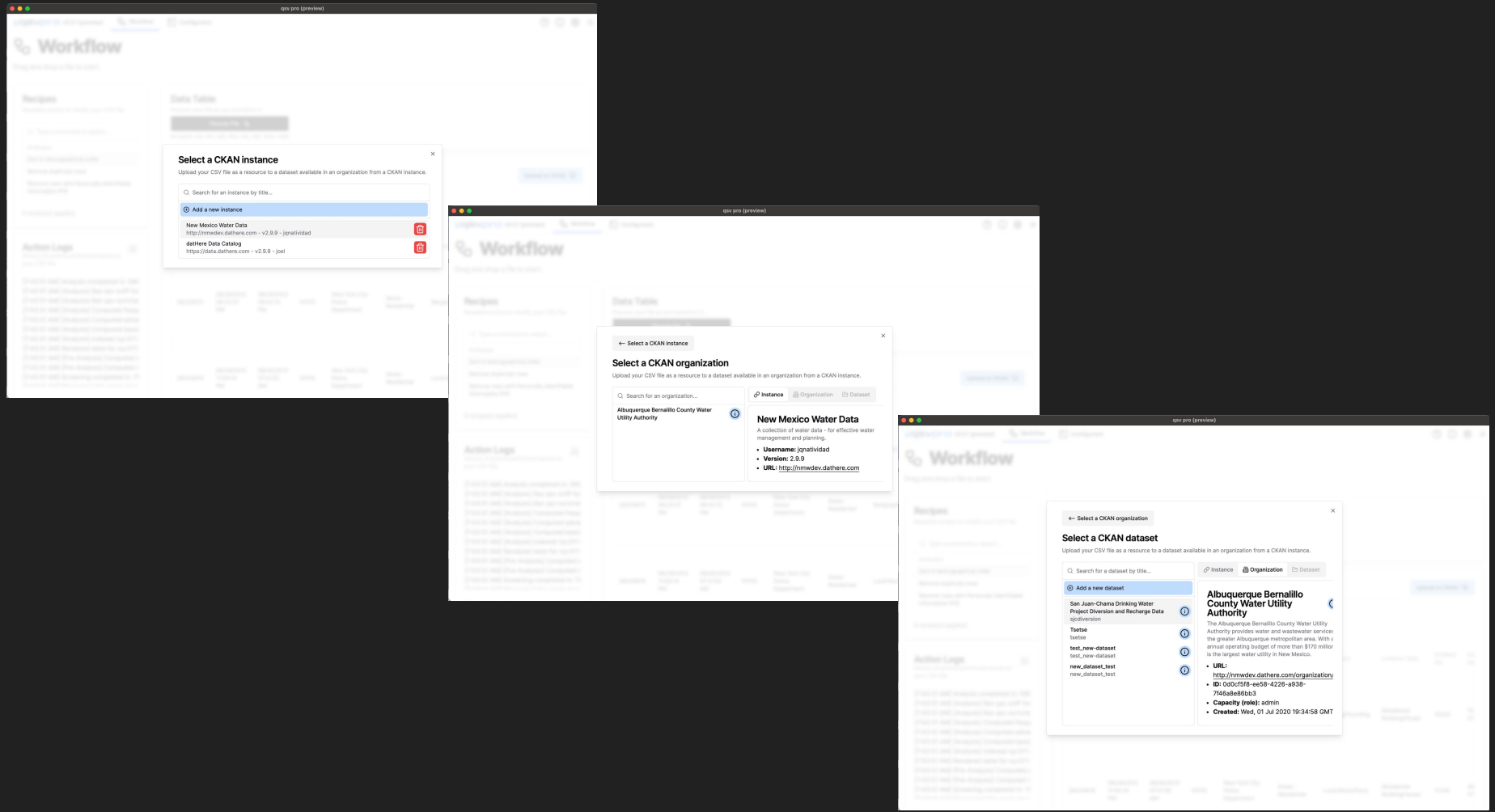
- OpenRefine + Excel + qsv + CKAN + recipes + High Value Curated Data = qsv pro
- Familiar spreadsheet interface
- No need to know complex Command Line Interface (CLI) commands
- FAST! Blazing Fast!
- Recipes! (desktop ETL)
- Integration with datHere's upcoming cloud-based services
  - High Value Data Feeds
  - Data Enrichment
  - Data Normalization
  - Geocoding
- Natural Language Interface



qsvpro

Cross-platform Desktop Data Wrangling & Query tool for the Rest of Us

- For a Data Analyst Audience
- You don't need to be a Developer
- Use ready-made Recipes for common tasks (e.g. Scan for PII, geocode, deduplicate records, etc.)
- Create/modify/combine Recipes using either Luau or Python
- Share your Recipes on the datHere Recipe Catalog
- Pre-process security-sensitive data on your desktop without uploading it first
- Enrich your data with datHere's ever-expanding corpus of High Value Data like the Census, Bureau of Labor Statistics, etc.
- Use the "Answering People Interface" on your data or of other CKAN portals
- Upload to your CKAN or to datHere's Data Catalog to share your data with the world!


qsv pro (preview)
Workflow
Select a CKAN instance
Upload your CSV file as a resource to a dataset available in an organization from a CKAN instance.
Search for an instance by title...
Add a new instance
New Mexico Water Data
http://nmwdev.dathere.com - v2.9.9 - jqnatividad
datHere Data Catalog
https://data.dathere.com - v2.9.9 - joel
qsv pro (preview)
Workflow
← Select a CKAN instance
Select a CKAN organization
Upload your CSV file as a resource to a dataset available in an organization from a CKAN instance.
Search for an organization...
Albuquerque Bernalillo County Water Utility Authority
Instance
Organization
Dataset
New Mexico Water Data
A collection of water data - for effective water management and planning.
Username: jqnatividad
Version: 2.9.9
URL: http://nmwdev.dathere.com
qsv pro (preview)
Workflow
← Select a CKAN organization
Select a CKAN dataset
Upload your CSV file as a resource to a dataset available in an organization from a CKAN instance.
Search for a dataset by title...
Instance
Organization
Dataset
Add a new dataset
San Juan-Chama Drinking Water Project Diversion and Recharge Data
sjcdiversion
Tsetse
tsetse
test_new-dataset
test_new-dataset
new_dataset_test
new_dataset_test
Albuquerque Bernalillo County Water Utility Authority
The Albuquerque Bernalillo County Water Utility Authority provides water and wastewater service
the greater Albuquerque metropolitan area. With a
annual operating budget of more than $170 millior
is the largest water utility in New Mexico.
URL: http://nmwdev.dathere.com/organization/
ID: 0d0cf5f8-ee58-4226-a938-7f46a8e86bb3
Capacity (role): admin
Created: Wed, 01 Jul 2020 19:34:58 GMT

qsv pro (preview)

qsvpro v0.5.0 (preview) | Workflow | Configurator

Successfully ran SQL query!
Estimated elapsed time: 1139ms

Ran query in 1139ms!

Upload to CKAN | Choose File

Recipes

Reusable scripts to modify your CSV file.

Search for a recipe...

All Recipes

Sort in lexicographical order

Remove duplicate rows

Remove rows with Personally Identifiable Information (PII)

0 recipe(s) applied.

Action Logs

History of actions performed based on your CSV file. Durations and timestamps are estimates.

[12:30:18 PM] Completed SQL query run ir
[12:27:35 PM] Completed SQL query run ir
[12:25:00 PM] Estimated total processing
[12:25:00 PM] Analysis completed in: 4717
[12:25:00 PM] [Analysis] Ran qsv sniff for
[12:24:59 PM] [Analysis] Ran qsv sortchec
[12:24:58 PM] [Analysis] Computed limitec
[12:24:48 PM] [Analysis] Computed freque
[12:24:35 PM] [Analysis] Computed advan
[12:24:17 PM] [Analysis] Computed basic s
[12:24:12 PM] Rendered table for NYC_311

18 action(s) performed.

NYC_311_SR_2010-2020-sam

D:\Work\datHere\Projects\sample\NYC_311_SR_2010-2020-sample-1M.csv

Table | Stats | Frequency | Metadata | SQL Query

Run a SQL Query

Run a Polars SQL query on your CSV file. Refer to your CSV file as a table named _t_1.

Natural language query, along with summary stats, frequency & metadata sent to preferred LLM…

Enter your natural language query. It should be converted to a SQL query below.

What are the most common complaint types by borough?

Ask

Enter your SQL query.

```
SELECT
    Borough,
    "Complaint Type",
    COUNT(*) AS Complaint_Count
FROM
    _t_1
GROUP BY
    Borough,
    "Complaint Type"
ORDER BY
    Borough,
    Complaint_Count DESC;
```

… an LLM we prompt to create a SQL query based on the Natural Language query & the context we provided

Recent query's estimated elapsed time: **1139ms**

Run SQL query | Reset SQL query | Save output to file | Decrease code size | Increase code size

Reproducible, hallucination-free answers

Rows per page 10 | Page 1 of 129 | Borders | Wrap Rows

Borough	Complaint Type	Complaint_Count
BRONX	Noise - Residential	24284
BRONX	HEAT/HOT WATER	18584
BRONX	Street Light Condition	8354

DMS Applications

- Open Data Portal
- Internal Data Exchange
- Enterprise Data Catalog
- Regional Data Center
- Data Library
- Water Data Hubs

datHere DMS Framework

datHere DMS Distribution

High Value Data Sources

United States Census Bureau · BLS · Federal Election Commission · FBI · data datHere · ...

Data Management System (DMS) Platform

Metadata

ckan

Analytics

analyze datHere powered by Superset

plotly

Data Wrangling

qsv

qsvpro datHere

Polars

Data Enrichment

- Curated High Value Reference Data
- Geocoded Data
- Political Contexts
- Census Contexts

datHere

DMS Framework

more than a Data Portal, a
Data Management System Framework
you can build on

- Built around CKAN
- Certified CKAN Extensions
- Bundled with other Best-of-Breed open source tooling
- Integrated Data Enrichment
- Build DMS applications like
  - Water Data Hubs
  - Open Data Portals
  - Internal Data Exchange
  - Data Library
  - Enterprise Data Catalog
  - and more…

Pathways to Open Source Ecosystems (POSE)

- NSF initiative that *”aims to harness the power of **open-source development** for the creation of **new technology solutions to problems of national and societal importance**”.*
- In 2023, University of Pittsburgh and datHere conducted Phase 1 study on how to scale up Civic Data Ecosystem around CKAN and other open source **Data Infrastructure** initiatives.
- Summer 2024, we anticipate we’ll announce new initiatives to implement our Phase 1 scale up proposal
- More info at **civicdataecosystems.org**

datHere

Data Infrastructure Engineering

Standards-based, best-of-breed, open source solutions to make your Data Useful, Usable & Used

https://datHere.com