Python – A Force Multiplier

Russell Robertson

December 14, 2022

What's so special about Python?

Why Python? (and not Excel)

- Designed for easily functional integration with production data stores
- Easier to learn and use than Excel's VBA
- Python code is maintainable and re-useable
- Scalability
- Math & analytics libraries available
- Free, state-of-the-art development platform available from Microsoft
- 100s of thousands of extensions developed and can be downloaded for free

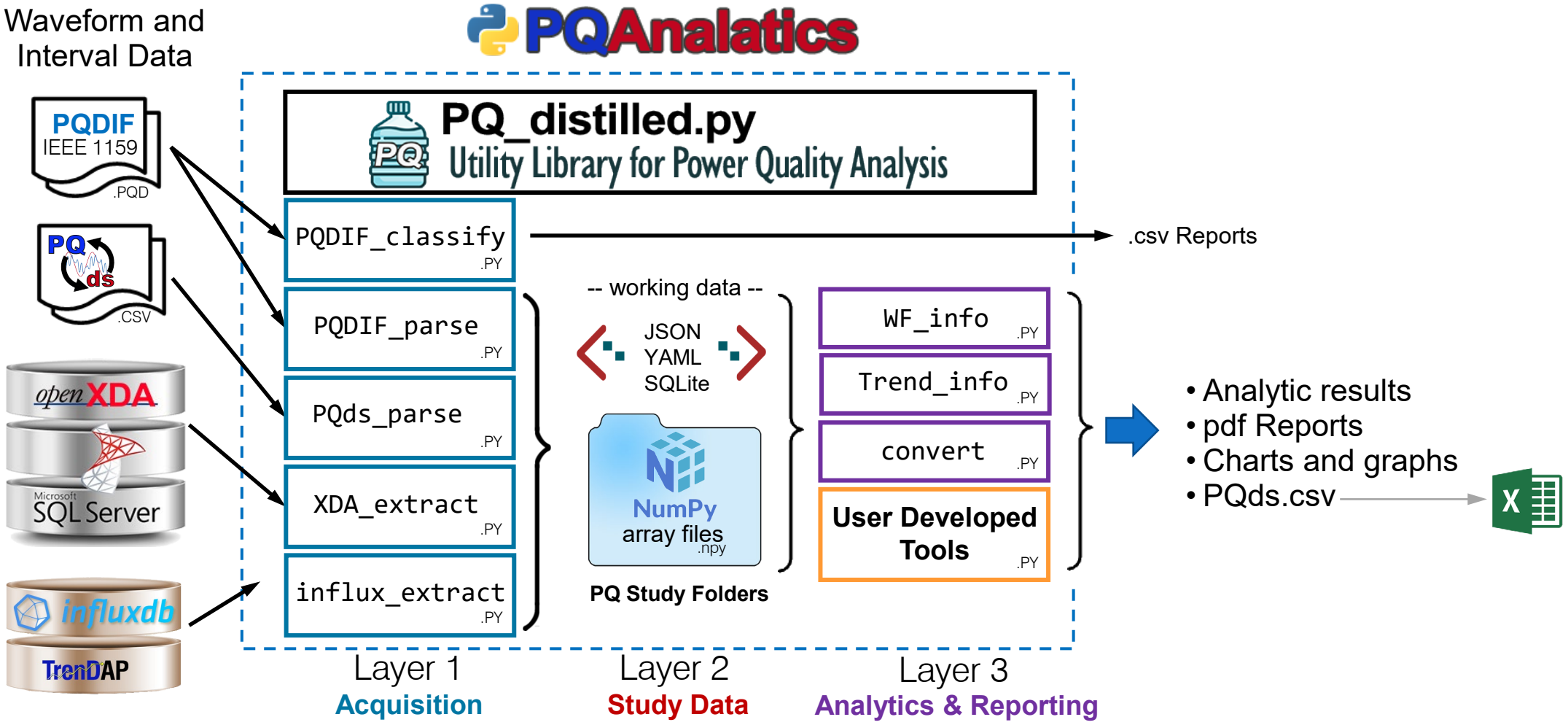
Quick History of Python

- 30+ years old
- Multi-paradigm language with dynamic typing
- Highly extensible via modules – NumPy, SciPy, Matplotlib
- Explosion in use occurred after release of Version 3.0 in 2008
- Core of multiple, huge code bases – Netflix, Spotify, Dropbox, reddit, …
- Most popular language. More that 6 million active developers
- Microsoft now funding language improvements

Python is a good choice for user-developed tools.

- Engineering Studies
  - Root-cause analysis
  - Comprehensive investigations
  - Rarely used analytic techniques
- Local Automation
  - File management and maintenance
- Production Tool Prototyping

The Vision – Create Robust Core Libraries



Flat Files Can Be A Good Choice for Prototyping

Advantage – Quick and Easy

- Python excels at reading data from and saving data to flat files
- Easy to understand and typically human readable
- Easy to implement
- Portable to multiple tools and languages

Disadvantage – Fragile & Inefficient

- Often requires building intelligence into file naming and/or folder structure
- Metadata often duplicated to facilitate human readability
- Filtering and building new relationships among data elements often requires building new flat files to augment the existing ones

Value of a High-Quality Data Layer

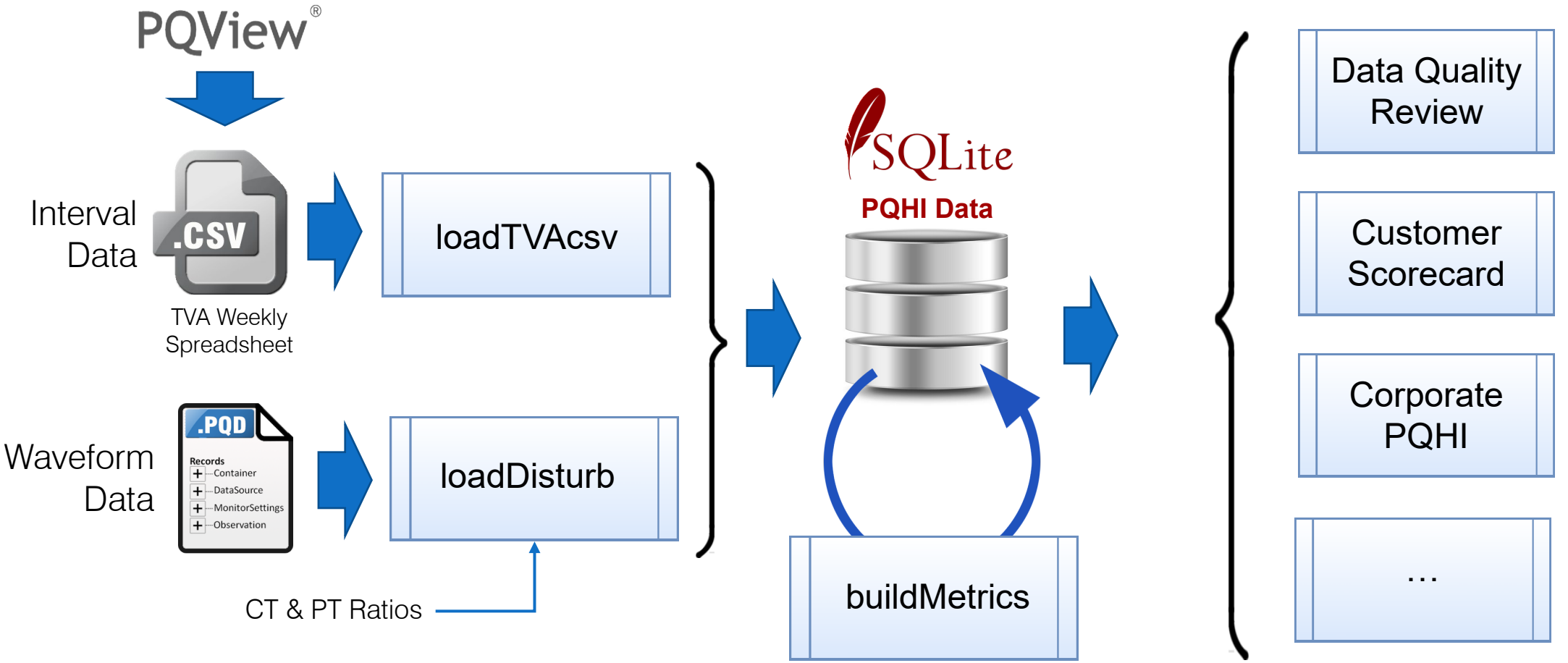
It's all about the data!

- A well-structured and open data layer reduces total cost and is functionally enabling
- Especially important for "big data" and PQ data is big.
- Not only requires good design but the right technology.

What's an open data layer?

- Not the same "open" as used in open source.
- Open data means:
  - Easy programmatic access by authorized systems and individuals
  - The data interface includes views that facilitate data use
  - Data access is possible via multiple techniques and technologies

PQHI Study Dataflow



80+% of the effort for the PQHI Study

Jupyter Notebooks

- First released in 2014. 10+ million notebooks now on GitHub
- Interactive tool that integrates text, rich media and can execute code
- Completely free and built into Microsoft's free VS Code editor. (40+ million downloads)
- Multiple languages supported -- R, F#, Python, & more
- A 'notebook' is powerful and flexible one column wide spreadsheet with cells that contain information
- In 2021, Jupyter notebooks was named as one of top 10 codes that transformed science
- Ubiquitous among data scientists
- Multiple web-based deployments now available

The Demo – Process a PQDIF File