Stats Can to Pandas

Harold Henson - Hensky Consulting

Sept 22, 2016

Codie's Café

Shopify

Large Volumes of Data to be Updated on a Regular Basis

- Many areas in government invest in current data on an ongoing basis
- Data provided by Statistics Canada is free and updated on a regular basis
- Downloading process is too some degree stable
  - No official batch process
  - Data is organized as Matrices
  - Each name individual timeseries has a stable name

Python-Pandas Perfect Tool

- Can store data in Pandas file that can drive Business Intelligence
  - Briefing Notes
  - Q's and A's
  - Analytical Reports
- One version of database
  - Ensure consistent definitions used
  - Reduce Errors

Emphasis on Documentation and Audit Trails

- Python has Log objects that can be used
- Can keep multiple versions of database
- Source code will be saved

HDF5 file format is perfect

- Write once read many times is perfect for a database where one data analysts supports many policy analysts
- HDF5 allows for maintenance of the tags
  - Matrix number
- Lack of security not an issue for Stats Can data

Can Drive Automated Reports

- Automated reports can be driven off the HDF5 format
- Python programs can be written to load into spreadsheets
  - Drives most business reporting in government
- Both Stata and SAS have capacity to read HDF5 files
  - A bit tricky but possible

Future Directions

- Will go up on GitHub when reasonably complete
  - Not waiting for perfection
- May try to automate interactive download
  - Stat Can position not stable on this
- Will move on to automating tables in Django
- A world in which all tables are automatically updated is possible

References

- Python for Data Analysis – Wes McKinney
  - Core document for project
  - Crucial details are online
- Python and HDF5 – Andrew Colette
  - Intuition behind the HDF5 file structure
- Learning the Pandas Library – Matt Harrison
  - Tutorial style book