Introduction to DataLab: Jupyter meets BigQuery

Experiments with Data Wrangling in the Google Cloud
Katharine Jarmul (@kjam)
DevFest 2015 - Berlin

Who am I?

- Python Developer since 2008
- Originally from Los Angeles, now I live in Berlin (since July 2013)
- Co-founder PyLadies
- O’Reilly book: Data Wrangling with Python
- @kjam / kjamistan.com

What is Data Wrangling?

- Answering questions by collecting, analyzing and studying data
- Using code (Python, R, MatLab, etc) to help answer those questions for large, complex or unclean datasets
- Using your own brain to evaluate good questions, answers and sources

Why Google Cloud?

- New Tool: DataLab
- Ability to use BigQuery
- $300 voucher to try
- Reasonably priced instances that are great for data analysis (ComputeEngine)

Getting Started with DataLab

- One-click deploy (also Docker image avail)
- Initial issues with deployment (it *is* Beta)
- Currently only available in US Cloud
- Pricing not yet released
- Really responsive with bugs on StackOverflow

BigQuery

- Access any databases you have in Google Cloud / Storage
- Super fast, append only
- Can load from Google Storage, Google Datastore and stream
- Simple SQL syntax

Jupyter Notebooks

- Replacing IPython notebooks
- Tornado-based server
- Can also run bash, R and Julia
- Share code, documentation, MarkDown and charting
- Iterative and agile approach to Data Science / Wrangling

Let's take a look at it in action

Sample Datasets

- Annual weather data
- Github commit data
- Sample HTTP logs
- US Birth Data
- Shakespeare’s works
- Wikipedia edits (2003-2009)

Sample Code

- Each Notebook Server comes with code examples using the sample data
- DataLab Guide introducing BigQuery and Storage Examples
- Easy to get started!

Pandas

- Data Analysis for Python, especially useful for larger datasets
- Easy imports and exports
- Matrix and Series calculations
- Split, apply and combine
- Built-in charting and statistical methods

Let's take a look at it in action

Google Charts

- Simple (and slightly more complex) charts available from Google Charting API
- Already integrated with user actions, zoom, Google Maps and other pre-built features
- Great to use if you have no front-end skills!

NumPy

- Advanced mathematical library
- Uses matrices and arrays
- Statistical analysis for correlation, deviations, covariance and dataset description
- Extremely fast at large computations

GitHub Repositories

- Source code version control
- Every DataLab notebook connects with Google Source Repositories
- Each instance has it’s own branch
- Can keep a local synced copy

Let's take a look at it in action

Adding Outside Data

- Imports via any data files on Google Storage
- Example notebook:

```
cars2 = storage.Item('cloud-datalab-samples',
                     'cars2.csv').read_from()

df2 = pd.read_csv(StringIO(cars2))
```

Natural Language Processing

- Ability to use computers and machine learning to identify language
- Can be used to help with synonyms / antonyms, translation, sentiment analysis
- Google Compute Engine NLP to power your notebook (along with nlp Python libraries)

Sharing Your Notebooks / Repo

- Download as notebook or as a Python script
- Share the code via the Source Repository
- Share the DataLab notebook via Google Authentication

Questions?

- Now?
- Ask away!
- Later?
- @kjam on Twitter / Freenode
- Tonight’s chat
- kjamistan.com