Pangeo: A Community Platform for Big Data Geoscience

Kevin Paul
Computational & Information Systems Lab
National Center for Atmospheric Research

American Meteorology Society 99th Annual Meeting
Phoenix, AZ – January 9, 2019
A (very) little about Pangeo...

• Why does it exist?
• What is it?
• How can you get involved?
Data Analysis is complicated in the Era of Big Data

- Data is too big to analyze
  - ...or even move!
- Analysis is different for every...
  - ...scientist
  - ...dataset
  - ...question
- Analysis has many different components
  - Data wrangling (formats, re-formatting, etc.)
  - Visualization
  - “Intermediate” computation

*How is it possible to solve such a complex problem?*
Consider One Possible Approach...
The Rise of the Unfortunate All-Purpose Tool (UAPT)

- A small team forms an idea and...
  - ...writes a grant, or
  - ...starts a company
- The team decides to develop a general visualization tool
  - Natural place to start
  - Visualization is *always* the desired end goal!
- The tool grows in popularity
  - Or, at least, we hope!
- The tool starts to reach many different communities...
  - ...with many different use cases and needs.
The Fall of the Unfortunate All-Purpose Tool (UAPT)

• The size of the user community outsizes the developer team!
• Users start demanding more!
  • “How can I get UAPT to read my data format?”
  • “How can I get UAPT to plot with my projection?”
  • “How can I get UAPT to compute my intermediate results?”
  • “I keep blowing out memory when I use UAPT with my data!”
• UAPT starts to get complicated...
  • ...and the developer team can’t keep up with all of the individual user issues.
• Developers add an interface that lets users “plug in their own code”
  • e.g., “Custom analysis with your own Python code”
Whoa! Wait!
I think we went off the rails...

• The developer team could not keep up with the user community!
• UAPT started as a “tool to make your work easier” but ended with the requirement to learn a language!
  • Hence, users need to learn how to use the tool and the language!
• Visualization tools are great (and extremely useful), but the end goal for analysis is complete customizability.
  • Start with a high-level language (Python!!!), not a tool!
• But even a high-level language is a lot to ask of the user!
• What else do you need?
What I used to think my job as a Software Engineer meant...
What I now think my job as a Software Engineer means...
But how?

• Every scientist...
  • Pessimistic: ...hates different parts of there work!
  • Optimistic: ...loves different parts of there work!
• Everything else can be “abstracted away”
  • i.e., write code that does the stuff you hate!
  • i.e., leave you only the stuff you love!
• And a user can choose to use those abstractions or not!
  • Let the user choose the level of complexity in their work!
Python’s Scientific Stack

StatsModels
SymPy
NetworkX
scikit-image
PyMC3
matplotlib
pandas
SciPy
NumPy
Jupyter
Python
IPython
xarray
Dask
Anaconda
Numba

air • planet • people
Abstraction: *Like the layers of an onion!*

- Each layer of abstraction in the stack brings you closer to your natural language!
- **The End Goal:** *The language you speak with your colleagues!*
- This means...
  - There’s less to learn initially!
  - *You can contribute anywhere (in the stack) you feel comfortable!*
- That’s **Open Source**!
- And that leads us to...
A community with the goal to:

• Foster **collaboration** around the **open source** scientific python ecosystem for ocean / atmosphere / land / climate science.

• **Support the development** with domain-specific geoscience packages.

• Improve **scalability** of these tools to handle petabyte-scale datasets on **HPC and cloud** platforms.
### Pangeo Stack Layers & Interoperability (Incomplete!!)

<table>
<thead>
<tr>
<th>User Interface &amp; Processing Mode</th>
<th>Analysis &amp; Modeling Tools</th>
<th>Data Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="jupyter" /> <strong>jupyter</strong></td>
<td><img src="image" alt="serverless" /> <strong>serverless</strong></td>
<td><img src="image" alt="xarray" /> <img src="image" alt="Iris" /> <img src="image" alt="pandas" /> <img src="image" alt="DASK" /></td>
</tr>
<tr>
<td><img src="image" alt="#!/bin/bash" /> #PBS -l nodes=1:ppn=2</td>
<td><img src="image" alt="GCM" /></td>
<td><img src="image" alt="NumPy" /> <img src="image" alt="python" /></td>
</tr>
<tr>
<td><img src="image" alt="N-Dimensional Arrays Foundation" /></td>
<td><img src="image" alt="Compute Platform" /></td>
<td><img src="image" alt="Compute Platform" /></td>
</tr>
<tr>
<td><img src="image" alt="Compute Platform" /></td>
<td><img src="image" alt="Compute Platform" /></td>
<td><img src="image" alt="Compute Platform" /></td>
</tr>
</tbody>
</table>
Pangeo supports Open Source

• Anyone can contribute (including especially you!)
  • And Pangeo is here to help you contribute!

• Pangeo is Agile
  • Comprised of self-organizing teams
  • Depends on close interaction between “user” and “developer”
    • ...and everywhere in-between!
  • Excellent educational model!
    • You get to work (and learn from) extremely bright people in both the geosciences and CS!

• By fostering Open Source, Pangeo grows the developer team along with the user community!

• Pangeo is inclusive
  • Everyone is welcome!
How can you contribute?

• Uses GitHub Issue trackers for communication
  • [https://github.com/pangeo-data/pangeo/issues](https://github.com/pangeo-data/pangeo/issues)
  • ...Along with GitHub Issue trackers for all of the other Pangeo Stack tools!

• Gitter for conversations
  • [https://gitter.im/pangeo-data/Lobby](https://gitter.im/pangeo-data/Lobby)

• *Jump in wherever you feel most comfortable!*

You are welcome here!!!
Thank you!

• The Pangeo Website:
  • https://pangeo.io

• The Pangeo Conversation:
  • https://gitter.im/pangeo-data/Lobby

• The Pangeo Organization (GitHub):
  • https://github.com/pangeo-data

• The Pangeo Forum (GitHub Issues):
  • https://github.com/pangeo-data/pangeo/issues