CDI Monthly Meeting 20200212

February 12, 2020: Pangeo: A flexible open-source framework for scalable, data-proximate analysis and visualization

Meeting Recording and Slides

Recordings and slides are available to CDI Members that are logged in to the wiki.

These are the publicly available introductory slides. Log in as a CDI member to see the other presentation slides.

If you would like to become a member of CDI, join at https://listserv.usgs.gov/mailman/listinfo/cdi-all.

CDI_20200212_OpeningSlides.pdf, 200212-cdi-group-announcements.pdf

Agenda (in Eastern time)

11:00 am Welcome and Announcements

11:05 am CDI Group Learning Topics

11:15 am What's happening around the CDI

11:25 am **Pangeo: A flexible open-source framework for scalable, data-proximate analysis and visualization** - Rich Signell and Renee Pieschke, USGS

12:30 pm Adjourn

Abstracts

Pangeo: A flexible open-source framework for scalable, data-proximate analysis and visualization

Pangeo provides a complete and flexible framework for data analysis and visualization, with scalable computations next to the data. The framework uses components already in widespread use in the community:Dask, Xarray, Pandas and Jupyter. The user needs only a web browser locally to interact with the data. We will give demonstrations of using Pangeo for several different types of USGS workflows, involving large simulation model results and remote sensing applications.

Rich Signell is a Research Oceanographer at the Coastal and Marine Science Center in Woods Hole. He has worked with the Integrated Ocean Observing System (IOOS), served on the CF Standards Committee, served on the Unidata Users Committee and is a member of the Pangeo Steering Council.

Renee Pieschke is a Technical Specialist for the Technical Services Support Contract at the Earth Resources Observation and Science Center in Sioux Falls, SD. She has been supporting the Landsat Product Improvements Project, which builds the infrastructure to move and process Landsat data at scale in the CHS AWS Cloud as well as improve access through tools such as Pangeo.

Highlights

1. **Opening**
    a. CDI is open to anyone! You don't have to be USGS. Remember our diversity and work on communicating in plain language.
    b. Why is CDI valuable to you? Why do you participate? (Slido poll)
    c. How can you get more involved in CDI?
        i. CDI playbook - CDI Playbook: How#How:Bestpracticesforcommunityparticipation
2. **CDI Group Learning Topics**

    a. Set a goal > Gather a group > Make a deadline > Share your progress
    b. Group learning - goal 1
        i. Usability resource reviews from the Usability wiki page.
    c. Group learning - goal 2
        i. NetCDF - how and why?
    d. Look for future email/wiki page to participate in these or propose your own.
3. **Group Announcements**
    a. **Leslie missed this slide on the call** - ICEMM (Interagency Collaborative for Environmental Modeling and Monitoring) group. contact pglynn@usgs.gov. Annual ICEMM meeting March 17-18, 2020 in Reston, VA. Theme: integrated modeling, monitoring, and working with nature. Agenda posted on the group wiki page.
    b. Semantic Web working group - discussion of a journal article inspired research on semantic web basics (DataONE webinar, books, and websites). Talk with Sky Bristol on this topic will be rescheduled.
    c. Data Management working group - Presentation on Monday with Claire from the Science Gateways Community Institute. Created sample value propositions on the value of CDI to data management.
    d. Software development cluster - January - talked about software policies you should know. This month, demo of APIs and swagger.
    e. Usability group - Usability alternates between town hall meetings and resource review postings. Even month = town hall; odd month = resource review. Topic for this month is choosing usability techniques. This topic is inspired by questions and posts by usability group members.
    f. Risk - January meeting, Sophie discussed applying usability to a web-tool, and stakeholder interviews for the DOI RISK project. Received 29 proposals to RFP. Anticipating announcing rewards by end of the month.
        i. February 20th - kickoff for training in Human-Centered Design and Inclusive Problem Solving.
    g. Open Innovation
        i. Next monthly meeting February 26th. Sharee Watson will talk about three different new scientific projects she's working on in Hawaii. Update on USGS Open Innovation Strategy. Demo of collaboration tools & resources to share/edit, with a team of 3 teams - guidance, policy, and catalog & toolkit.
    h. Tech Stack working group - meeting tomorrow, February 13th.
4. **Pangeo: A flexible open-source framework for scalable, data-proximate analysis and visualization**
    a. Pangeo framework is funded by EarthMAP - situational awareness driven by sensors & models.
    b. New form of model data analysis - nothing installed locally, working on the cloud.
    c. Pangeo is an open community of people trying to build software for science that is community-driven, flexible, and collaborative. Slowly building out set of tools with a common philosophy. GitHub contains all Pangeo-related files.
    d. i. 1. http://support.chs.usgs.gov/
        a. Click Request New Service
        b. Click "I am interested in using Pangeo."
        c. Test notebooks available here:
            i. https://code.chs.usgs.gov/earthmap/notebooks

Q&A

1. Two way tight coupled atmosphere-Ocean? ie. feedback and convergence at every timestep?
    a. Every few minutes, there is a chance for some elements to affect others.
2. Where can I go to learn more about EarthMAP?
    a. USGS Director's Science Planning Strategy (EarthMap!)
    b. More information rolling out in March
    c. The CHS Help Center article for Pangeo is located at this link https://support.chs.usgs.gov/x/OwA1Ag
3. Why do you think that the Pangeo instance of community-driven software has been successful? Isn't it hard to get contributors usually?
    a. Ryan Abernathy has championed Pangeo. Started with big ideas and goals, got government people and commercial people on board from the beginning. Approachable and open. Not developing something to try and sell.
4. I've heard that you can use Pangeo even if you don't work with big data - is that true? why? how?
    a. If you're interested in machine-learning, or using Google Earch engine, Pangeo provides a cloud-agnostic way of doing the same kinds of workflows.
5. you mentioned costs are free for moving data within region. Who do we talk to about that type of issue?
    a. Reach out to CHS. report.chs.ugs.gov - request support
    b. EROS will be developing some help documentation for accessing Landsat as we start publicizing data availability from USGS
6. Is there a good start up guide to Pangeo for Matlab users who're interested in transitioning to some of these python tools?
    a. Most Matlab users find the Python language structure fairly similar/accessible. Of course the vocab is different, which is tough. There are some sites that try to help: https://realpython.com/matlab-vs-python/
7. Or even better, can these tools be tapped through using Matlab commands via Jupyter notebooks?
    a. Octave, the free version (and reduced functionality version) of Matlab can work with Jupyter. We would have to enable that capability on pangeo.chs.usgs.gov if there was sufficient interest.
8. What is level 1 vs level 2 processing?
    a. For Landsat, a Level-1 product is a top of atmosphere product, where a Level-2 will apply atmospheric corrections to get closer to surface reflectance for better scientific analysis. For more information on Landsat Collections and Levels, see the Landsat Missions Website.
9. When is the next Pangeo crash course/training available?
    a. Workshop in Flagstaff at the end of ~~February~~ March - 10 spaces left. Contact ianeece@usgs.gov for details and to sign up (or get on wait list). Will cover Jupyter, DASK, xarray, cloud-optimized storage formats, remote sensing, machine learning. "Test" session to see how well it works.
    b. The CHS Help Center article for Pangeo is located at this link https://support.chs.usgs.gov/x/OwA1Ag
10. What happens when we build on this and in the future the costs get passed down to us? Is somebody tracking the costs that are currently paid to plan 4 this?
    a. We are tracking costs that Pangeo will generate; cannot distribute costs down to a user. If the current subsidy goes away, we can work with users to see what the cost would be. Will have to be some investment from investigators/users.
11. What is a simple explanation of how USGS can use Pangeo?
    a. A common software environment; you can just log in and start using this without downloading anything. Eliminates problems in replicating software environment. Low barrier of entry to try these tools.
    b. As ScienceBase moves to a cloud-based platform, that data will be easily accessible through Pangeo.

12. is the CHS Pangeo public or internal only, meaning if you upload data or create a notebook can someone external access it?
    a. It is tied to your active directory, so you get dedicated space to work on things.
    b. Pangeo Binders is something to explore if you want to share a notebook environment with a team of individuals. Consult CHS for more information about this.
13. can you clarify what is meant by a "cloud optimized geotiff?
    a. Regular geotiffs are read line by line, cloud optimized chunks it into squares and gives you overlays; allows you to get just your area of interest easier than reading line by line.
    b. For more information, see this medium.com article.
14. To reproduce research, versioning of the input data would be needed in addition to versioning of the source code. Is this happening?
    a. There's a versioning proposal in STAC that could use more input if you would like to reach out to the STAC community through their gitter chat or through the GitHub issues board.
15. Will Pangeo be independent of the Cloud Service chosen. ie, Google vs Amazon vs Microsoft?
    a. It will need to be stood up on the different cloud infrastructures, but the tools themselves are pretty agnostic. It's all python.
16. Is it fair to think of pangeo as an 'open source' alternative to google earth engine with more analysis capabilities ?
    a. Yes. The data is accessible and reproducible on any cloud.
17. How hard would it be to set up an R environment built on top of the Pangeo stack?
    a. Pangeo is Jupyter, so you can use Jupyter, Python, or R in this deployment.

Poll - Why is CDI valuable to you? Why do you participate?

Poll administered on sli.do

- I like to hear about (and share) the cool work folks are doing throughout the USGS! The Communities are valuable because they allow folks to share innovative research and discuss ways we can do so while following Department, Bureau, Mission Area policy.
- new ideas and new networking
- CDI provides relevant, useful, and timely data management related issues, projects, and tools.
- I learn about new technology applications and learn of colleagues I might collaborate with.
- To learn more about the technology and approaches other members are using to support science and decision making.
- The CDI helps me to get my work done in my daily job! I find the people who are part of the CDI are amazing to interact with - they are engaged, enthusiastic, and interested in making things better at USGS. CDI has made me feel like I am more in touch with the USGS - there is so much going on in this Bureau, and CDI keeps me informed and makes me feel like I am part of something bigger than just my daily job. Thanks for all you do to make this Community so vibrant!
- Learn new tools and techniques
- Provides valuable information on data management, tools that others are developing, issues that others are having, support of workgroups that have been great resources (metadata workgroup, data management). I have no other venue for this and had been floundering solo as there doesn't appear to be anyone else interested in this at my center (or, if they are, they don't say anything;). :(
- To learn more about data management and tools to help the staff at CAWSC.
- To find partners in building USGS capabilities to serve the nation with our data products.
- Agree with Leslie's statement but would also add that we can support others with reaching their goals by sharing our expertise with them.
- Because keeps me updated about what others are doing
- demonstrate that best practices in data sci/software/etc. is important to colleagues
- Melding of Earth and Computer Science
- Network of knowledgeable peers
- Knowledge sharing
- prof dev
- Data projects and data management collaboration with EVERYONE across the country.
- Learning
- Near and dear to my cold heart
- Connecting with others with similar interests across the USGS and beyond.
- information sharing
- Anyone
- Let's me know what is happening with data integration across other parts of the USGS.
- Feeling part of a larger community. Leveraging what others are doing
- Diverse community, wide range of experience and expertise