ESIP Lab

SUPPORTING DEVELOPMENT OF EARTH SCIENCES CYBER-INFRASTRUCTURE

PROJECT SUMMARY

Name of project: *ESIPhub: Developing and promoting an ESIP community resource for sharing and running scientific workflows via JupyterHub.*

Project lead and contact details: Sean Gordon (scgordon@hdfgroup.org)

Project partners (if applicable) and contact details:

- NCAR Library: Keith Maull (kmaull@ucar.edu), Matt Mayernik (mayernik@ucar.edu)
- UCAR/UCP/UNIDATA: Ryan May (rmay@ucare.edu)
- HDF Group: Aleksandar Jelenak (ajelenak@hdfgroup.org)
- USGS: Rich Signell(rsignell@usgs.gov)

Proposed start and end date: Jan 2018 - July 2018

Budget Requested: $7000

Budget Summary: To support AWS cloud costs(~$1000) Conference registration costs for each project partner for workshops, presentations, and reports at 2018 ESIP Meetings(~$3000) and conference registration costs for two at JupyterCon 2018 to present the findings(~$3000).

This project relates to the following key priority areas (please indicate all that apply):
X Earth Science Cyber-infrastructure
☐Semantic Technologies
☐Socioeconomic value of data
☐Other (explain)

PROJECT OUTLINE

Jupyter Notebooks have emerged as a powerful way to document and enact scientific workflows. They combine richly documented and executable code with the resulting analysis and interactive visualizations in a single sharable document. Notebooks are being adopted by teachers, scholars, scientists and working practitioners alike because they can provide direct evidence of processes and executable procedures through code and exposition, but more importantly provide a recipe for repeating, reproducing and repurposing those processes and procedures so that colleagues, collaborators and community members alike can understand, explore and ultimately benefit from the work. These Notebooks have the potential to *dramatically accelerate adoption of* ***community approaches*** *and* ***best practices*** *within the ESIP community,* ***but more infrastructure is needed to realize this potential***.

Jupyter Notebooks, important as they are to the 21st century scientific workbench, are not without challenges. Running Notebooks often require non-trivial software environment setups and configuration, imposing significant challenges on the end user to download software and customize their local environment in ways that may be difficult to reproduce by others. Python users, for example, may install various versions of the core language (e.g. 2.x, 3.x) with libraries that may require additional components or compilation steps if the target operating system is different. Some users, particularly those new to the platform or with limited technical expertise, may experience many hours or even days of additional work to create an environment consistent with the requirements of the run-time environment of the target Notebook, cancelling many of the benefits Jupyter Notebooks provide.

To address this problem several solutions have been created that are gaining attention. Binder (http://www.mybinder.org/) is a free service that creates an executable target Jupyter environment on the Cloud by allowing its users to link Github repositories with Notebooks automatically. Users can specify the dependencies and requirements of their Notebook environment and Binder will build the environment in the cloud and allow the user to execute the Notebook directly in their browser. Though Binder is built on top of JupyterHub (https://jupyterhub.readthedocs.io/en/latest/) and is a free service, it is still under development. Thus it has limitations that make it unstable for high availability, high performance, multi-user contexts, such as multi-user training or high performance scientific computing.

JupyterHub, upon which Binder is built, is another solution that provides cloud-based serving of Notebooks. JupyterHub allows for an entire Jupyter environment to be configured on a target server, so that multiple users can remotely access the resources of that server to spawn, store, share and execute their Notebooks on a consistent resource with appropriate library, computational and data configurations for their needs.

We therefore think the best solution for ESIP would be to install a JupyterHub instance on the ESIP Cloud resources, allowing multi-user login with ESIP authentication, reliable resources, and shared environments. This incubator project will allow us to investigate and scope this problem by interacting with the Jupyter community, developing a pilot instance, and reporting on the results at ESIP meetings.

In this proposal, we would like to accomplish the following:

1. Install JupyterHub on the ESIP Cloud (ESIPhub) to explore a new paradigm of scientific analysis: data-proximate computing with reproducible workflows and environments, using a modern browser as interface.
2. Gather best practices and lessons learned for using JupyterHub for ESIP community.
3. Use ESIPhub for training during the 2018 ESIP Summer Meeting, and evaluate success on the basis of user feedback.
4. Develop plan for sustainability of ESIPhub.

PROJECT OBJECTIVES

The primary objective of this proposal is to provide an open computing environment that facilitates the sharing and reuse of code used across ESIP.

Project Activity	**Expected Completion**	**Anticipated Outcome(s)**
Review ESIP provided Cloud resources	Nov 2017	Documented review of ESIP cloud resources
Review of JupyterHub solutions already implemented by Harvard, UC Berkeley, and Lawrence Berkeley National Lab	Nov 2017	Documented review of existing solutions at other institutions
Draft project plan for ESIP Hub implementation	Dec 2017	Documented project plan
Present project plan at 2018 ESIP Winter Meeting	Jan 2018	Community feedback and interest; plan approval
Implement project plan	Jan 2018 - June 2018	JupyterHub and Notebooks implemented on ESIP cloud; ESIP credentials used for authentication and authorization; ESIP members surveyed for kernels and modules needed; ESIP members try out the service; workshop style usage tested
Present project outcomes at ESIP Summer Meeting	July 2018	Completed presentation
Conclude project	Aug 2018	Project documentation; completed installation and plan for next steps
Present project outcomes as talk/poster at JupyterCon'18	Aug 2018	Completed presentation

Benefits and advancements of key Earth sciences priority areas:

Cyberinfrastructure providing a shared computing environment with the potential to dramatically accelerate adoption of community approaches and best practices within the ESIP community, through actual examples that can be easily reproduced and repurposed.

Project Partners (as applicable):

Description of project partners and their involvement:

- Keith Maull: Technical Lead, workshop at ESIP Summer
- Matt Mayernik: Community Development, stewardship, survey, ESIP Summer report/presentation
- Ryan May: Technical Development, workshop at ESIP Summer
- Aleksandar Jelenak: Technical Development, techdive presentation
- Rich Signell: Community Development, ESIP winter 2018 presentation, Continuing support
- Sean Gordon: Project lead, Notebook tutorials, cluster presentations, ESIP workshop

Involvement of ESIP collaboration areas:

Documentation Cluster, IT & I Committee, Data Stewardship Committee

What groups/audiences will be engaged in the project? Any member that wants to share their work in a repeatable way using a Jupyter Notebook. Any member that wants to explore and understand through running a Notebook.

How will you judge that project has had impact? Use of the resource (users and Notebook counts) and frequency of reuse of Notebooks (count of copies made and downloads of the file). Use of resource for telecons, presentations, workshops, etc at ESIP Meetings.

How will you share the knowledge generated by the project? Presentations at JupyterCon 2018 and workshops and demonstrations at ESIP 2018 Summer Meeting. Creation of containers facilitating members to recreate the environment and We also intend to present during cluster meetings and the monthly Tech Dive telecon. Engagement with the Documentation Cluster, IT & I Committee, Data Stewardship Committee for guidance and perspective on carrying the project on after July 2018.