ARISTOTLE
CLOUD FEDERATION

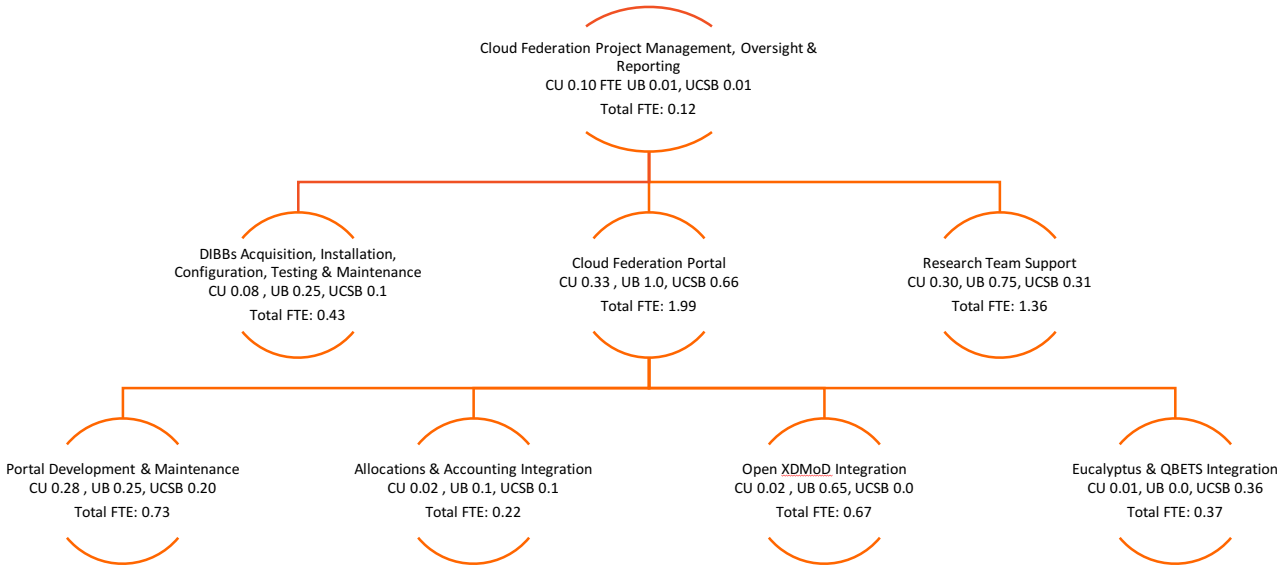
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 7/28/2016

Report 10 of 18

Submitted by David Lifka (PI)
lifka@cornell.edu

This is the tenth required monthly report of the Aristotle Cloud Federation team. We report on plans and activities for each area of the project Work Breakdown Structure (WBS).



#1541215

Contents

NSF
#1541215

1.0 Cloud Federation Project Management, Oversight & Reporting Report

1.1 Subcontracts
All subcontracts are in place. Nothing new to report.

1.2 Project Change Request
No new project change requests were made this month.

1.3 Project Execution Plan
The Project Execution Plan (PEP) was approved by NSF on 12/18/2015. We are operating as planned and continuously updating our PEP on a monthly basis.

1.4 PI Meetings
David Lifka and the Co-PIs are engaged in discussions with HPE regarding license agreements for Eucalyptus.

1.5 Status Call
7/5/2016 project status call topics included:

- Discussions regarding how long it should take for an instance store to fire up.
- Specific plans for portal additions are now available in a shared doc.
- Site-specific documentation should be placed in GitHub. Adam Brazier is the POC to provide your GitHub ID for access.
- Usage data graphs will be developed. CU-written scripts will collect the data and write to the database. Eventually, we will use the REST API.
- Discussions regarding how to manage instances across sites occurred. PCP will allow UB to collect information; sites will need to add it to their controller. Users won't have to do anything. We will inform them that this is running on their instances.

A second monthly call was cancelled due to the number of personnel participating in XSEDE16.

1.6 Project Planning and Preparation
The project web site went live at https://federatedcloud.org on 6/30/2016.

All of these efforts are described in more detail in this month's report.

2.0 DIBBs Acquisition, Installation, Configuration, Testing & Maintenance Report

2.1 Federation Resource Status Updates
We're happy to report that all Aristotle clouds—Cornell Red Cloud, Buffalo Lake Effect, and UC Santa Barbara Aristotle—are now in production with researchers running at each of the three sites.

- **UB**
  The UB team has been working with HPE on a license extension effective through October when all three sites plan to complete negotiations for a license for the remainder of the grant.

#1541215

- **UCSB**
  UCSB July activities included:

  Cloud Services:
  - Completed development of initial health/status tests and used them to successfully stress test the new installation.
  - Moved the Eucalyptus v4.2.2 cloud into production: Sedgwick researchers are starting to use the cloud.
  - Researching a backup solution for Ceph.

  Software:
  - Configured Nagios (OMD) for monitoring.
  - Developing provisioning process using LDAP Syncs "accounting groups" method and our Campus directory.

The infrastructure planning table was updated:

	Cornell	Buffalo	Santa Barbara
Cloud URL	https://euca4.cac.cornell.edu	https://console.ccr-cbls-1.ccr.buffalo.edu/	http://aristotle.ucsb.edu
Cloud Status	Production	Production	Production
Euca Version	4.2.2	4.2.2	4.2.2
Globus	Yes	Planned	Planned
InCommon	Yes	Yes	Yes
Hardware Vendor	Dell	Dell	Dell
# Cores	*168	**144	140
RAM/Core	4GB/6GB	up to 8GB	up to 9GB
Storage	SAN (226TB)	SAN (336TB)	CEPH (288TB)
10Gb Interconnect	Yes	Yes	Yes
Largest Instance Type	28 core/192GB RAM	24 core/192GB RAM	28 core/256GB RAM
	* 168 additional cores augmenting the existing Red Cloud (376 total cores)	** 144 additional cores augmenting the existing Lake Effect Cloud (312 total cores)	

2.2 Potential Tools: CloudLaunch & Supercloud
CloudLaunch is on hold; development efforts will resume during summer 2016. Supercloud: no updates to report.

#1541215

2.3 Industry Influence

Cornell and HPE had discussions regarding HPE support for Globus Auth. The HPE development team has confirmed that OAuth 2 support will be in Eucalyptus 4.4.

3.0 Cloud Federation Portal Report

There were content updates and additions this month to the new portal design that went live on 6/30/2016 at https://federatedcloud.org/. We are currently working on pulling in user documentation from the developer's pages located at GitHub.

The portal planning table below was unchanged this month.

Portal Framework			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 - End**	**1/2017 - End**
Gather portal requirements, including software requirements, metrics, allocations, and accounting. Install web site software.	Implement content/functionality as shown in following sections. Add page hit tracking with Google Analytics, as well as writing any site downloads to the database.	Implement content/functionality as shown in following sections. Add additional information/tools as needed, such as selecting where to run based on software/hardware needs and availability.	Release portal template via GitHub. Update periodically.
Documentation			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 – End**	**1/2017 - End**
Basic user docs, focused on getting started. Draw from existing materials. Available through CU doc pages.	Update materials to be federation-specific and move to portal access.	Add more advanced topics as needed, including documents on "Best Practices" and "Lessons Learned." Check and update docs periodically, based on ongoing collection of user feedback.	Release documents via GitHub. Update periodically.
Training			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 – 3/2017**	**4/2017 - End**
Cross-training expertise across the Aristotle team via calls and 1-2 day visits.	Hold 1 day training for local researchers. Offer Webinar for remote researchers. Use recording and materials to provide training asynchronously on the portal.	Add more advanced topics as needed. Check and update materials periodically, based on training feedback and new functionality.	Release training materials via GitHub. Update periodically.

#1541215

User Authorization and Keys			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 1/2016	**2/2016 – 5/2016**	**6/2016 – 9/2016**	**10/2016 – End**
Plan how to achieve seamless login and key transfer from portal to Euca dashboard.	Login to the portal using InCommon.	Switch to Globus Auth in order to better interface with the Euca web console Get 4.2.1 federated key.	Move seamlessly to Euca console after portal Globus Auth login.
Euca Tools			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 12/2016**	**1/2017 – End**	**1/2017 – End**
Establish requirements, plan implementation.	No longer relevant since Globus Auth will let us interface with Euca web console		Test access to Euca console.
Allocations and Accounting			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**3/2016 – 8/2016**	**9/2016 – 12/2016**	**1/2017 – End**
Plan requirements and use cases for allocations and account data collection across the federation. Design database schema for Users, Projects and collections of CPU usage and Storage Usage of the federated cloud.	Implement project (account) creation in the database and display on the portal. Integration hooks for user and project creation/deletion and synchronization across sites.	Automate project (account) creation by researcher, via the portal.	Report on usage by account, if the researcher has multiple funding sources. Release database schema via GitHub.
Metrics and Usage			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 7/2016	**7/2016 – 9/2016**	**10/2016 – 12/2016**	**1/2017 - End**
Implement graphs of basic usage data, including % utilization, available resources, and user balance, using scripts from Cornell and U Buffalo for basic data collection.	Provide documentation for installing XDMoD and SUPReMM at individual sites. Install Open XDMoD/SUPReMM at individual sites and begin data collection. This includes the installation of SUPReMM and the data collection piece at the federation sites. Begin integration with federated authentication providers.	Federated data collection across sites. Ship data from the individual sites to UB. We can summarize data remotely and send the summarized data or collect all raw data and summarize locally. Other job information will be federated as well using the prototype model under development with OSG. Display federated metrics in Open XDMoD at UB.	Release materials via GitHub. Update periodically.

#1541215

3.1 Software Requirements & Portal Platform

Work on implementing Globus authentication is still delayed due to a version problem; the widely-used league/oath2-client requires php 5.5 or higher, while 5.4.16 is the version provided with the currently available software stack.

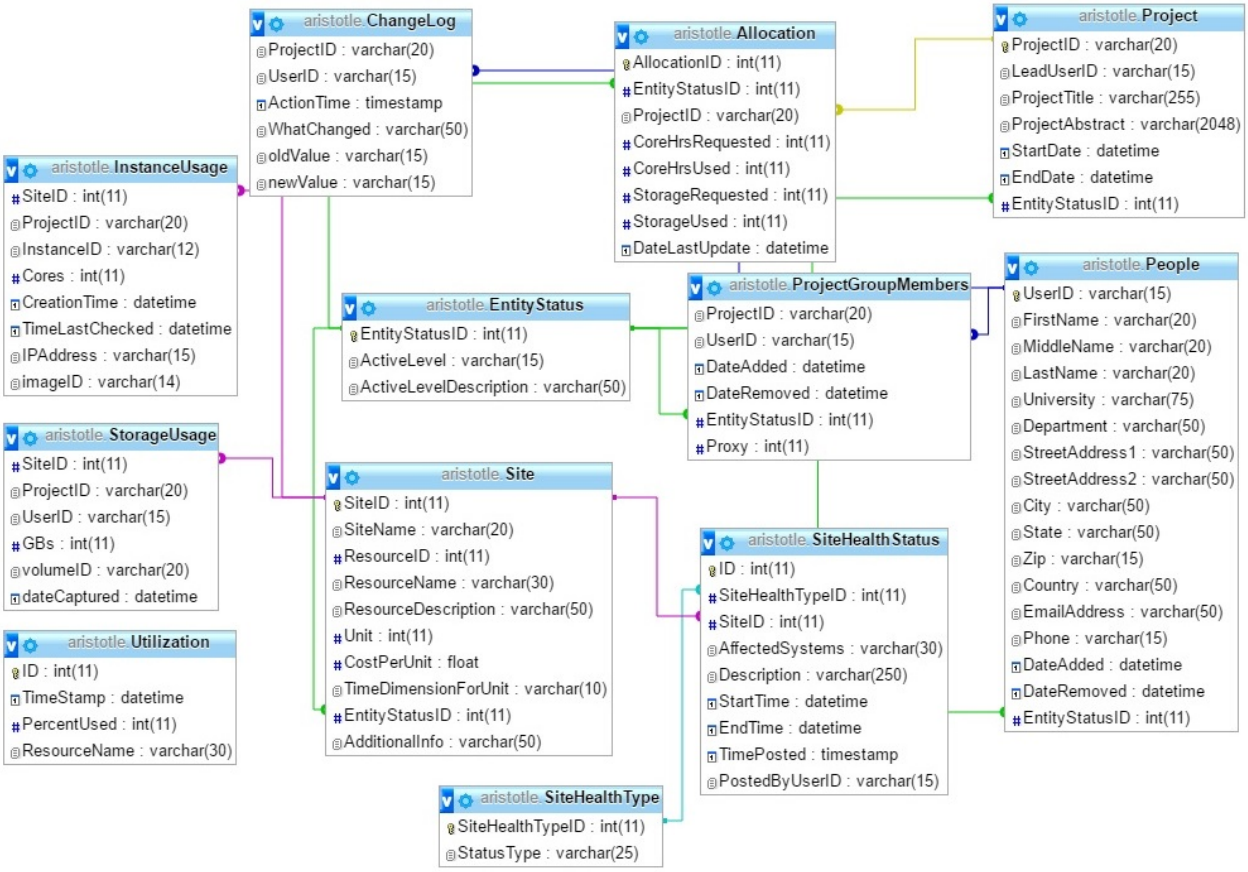
3.2 Integrating Open XDMoD and QBETs into the Portal

Usage graphs will be made available on the portal to show basic data until Open XDMoD and QBETS is implemented late this year. The code provided by UB via GitHub has been implemented at Cornell; Buffalo and UC Santa Barbara will provide usage data using the same mechanism.

Buffalo began data warehouse refactoring to support metrics for cloud and innovative resources. The existing XDMoD data warehouse was developed primarily to report on data generated by individual HPC jobs run on traditional HPC resources. With the advent of open source cloud solutions such as Eucalyptus and OpenStack, as well as non-traditional HPC resources such as Hadoop running alongside traditional HPC clusters at many centers, we must re-examine the infrastructure used to store and report on center utilization as well as the definition of an HPC job within XDMoD. The capabilities of the XDMoD data warehouse will be updated to support these new resource types and become more flexible to better manage new types developed in the future.

3.3 Allocations & Accounting

There were no changes to the database schema this month.



#1541215

4.0 Research Team Support

4.1. General Update

- Brazier and Barker (CU) have been producing user documentation for Aristotle which will be made available on the portal. Additional Use Case level documentation is being produced and will be made public at a later date when the work is completed and vetted by the science teams as suitable for public release; this will, in particular, include advice for others about installation of domain-specific software stacks in a cloud environment and will also cover some difficult installations (such as the parallelized version of WRF-Chem) in a thorough way useful to the science community. This will be valuable to scientists who are planning to migrate their computational operations and/or data analyses to the cloud.
- Some science teams are away on summer travel and consequently have no reports to make.
- The Aristotle Science Team Advisory Committee (STAC) quarterly meeting took place on 7/27/2016. Participation was high. Minutes will be posted on the Aristotle portal after review.

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
We have developed a Spark-based implementation of the Gaussian process-based change detection algorithm which is deployed on our Aristotle cloud (UB "Lake Effect" cloud). The algorithm allows for identifying changes from observational or simulation data in a distributed fashion. We are currently studying the scalability of the algorithm on the cloud. The results will be presented at *BigSpatial 2016 – the 5th International Workshop on Analytics for Big Geospatial Data* workshop in early November. The HTML5-based interface to the underlying Spark cluster on the Aristotle cloud is ready to be used by the general public. We are finalizing access control mechanisms to audit the cloud usage by users. PI Chandola will present the interface and its capabilities at the *Advances in Virtual Globe Technology Using NASA World Wind, an Open-Source Geobrowser* meeting at MITRE, McLean, VA, 8/10-11.

Use Case 2: Global Market Efficiency Impact
No update this month.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
No update this month.

Use Case 4: Transient Detection in Radio Astronomy Search Data
Work continues on the pipeline architecture and encoding, and how best to reduce the data by de-resolution in time and frequency.

Use Case 5: Water Resource Management Using OpenMORDM
No update this month.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
Barker met with one of the project postdocs to discuss and decide on possible algorithms to use for simulating symbiont metabolism.

NSF
#1541215

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production

- *Water-box experiment for California Live Oaks*: Last quarter, Sedgwick initiated an experiment to investigate the use of slow drip "water boxes" to help seedling development for California Live Oaks. There is little data on how drought conditions affect slow-growing trees in California, like the Live Oak, especially during the seedling stage. We are working on deploying a new moisture sensing infrastructure for this experiment. As part of that deployment, we attempted an end-to-end sensor test in the Live Oak test area that was to move data from the oak moisture sensors to the new UC Santa Barbara Aristotle machine. The test failed because of the sensor equipment, but the Aristotle back-end processing seems to be functional. We will attempt another test on 8/3/2016.
- *Agricultural soil moisture monitoring:* In the Sedgwick grape vineyard, the moisture sensor analysis indicates that the grapes are receiving too much water during a wetting event. We are working to understand how much less water we can use and with what frequency. Early analysis both of the soil moisture transfer rate and the irrigation system has resulted in two unsuccessful watering tests (that delivered too little water). A third experiment is scheduled for late July which will be followed by subsequent analysis. The goal of this work is to develop an Aristotle-based notification system that alerts Sedgwick personnel when irrigation is necessary and the amount of water to use.
- *Camera-trap animal survey:* Sedgwick is employing an undergraduate student to develop an animal-identification application based on Caffe that will be hosted on Aristotle. Caffe is an open-source machine-learning framework that has been used successfully for image identification. Using an open database of animal images to train a neural network, the student is developing the software framework necessary to process the camera-trap data for Sedgwick so that specific animal species can be identified automatically. This framework is running on the new Aristotle cloud system. Currently, the camera traps take approximately 200,000 photos per month that must be manually classified. Aristotle has already enabled the development of software to identify images that were triggered by non-animal movement (e.g., wind blowing leaves in front of the motion detector). This new system, when operational, will be able to provide additional classification indicating the likely species of animal.

  Also, as part of this effort, the student developed an OCR (optical character recognition) system to read the camera meta-data from the images. The Sedgwick camera traps print the meta-data on the images themselves (they do not have an alternative metadata store). Again, using machine-learning, the student's system (currently migrating from the old Aristotle hardware to the new) recovers the meta data so it can be used to index the images.

  Finally, the transfer of Sedgwick to eMammal (https://emammal.si.edu) continues. However, the estimated time for one-month's worth of image transfer is approximately 14 days because of the repository write speed. The repository is currently hosted in Box. The project attempted to use AWS S3 as an alternative and the speed increase was insignificant. The current plan is to use Aristotle as the alternative data repository (eMammal has an S3 interface and can interact with Aristotle without modification). The speed increase may be as high as two orders of magnitude.

5.0 Outreach Activities

5.1 Community Outreach

Tom Furlani presented federated XDMoD for Aristotle plans at the *XSEDE16 BOF—Coin of the Realm: Current Practices and Future Opportunities in Processing XSEDE Allocation Awards & Usage Data.*

#1541215