ARISTOTLE
CLOUD FEDERATION

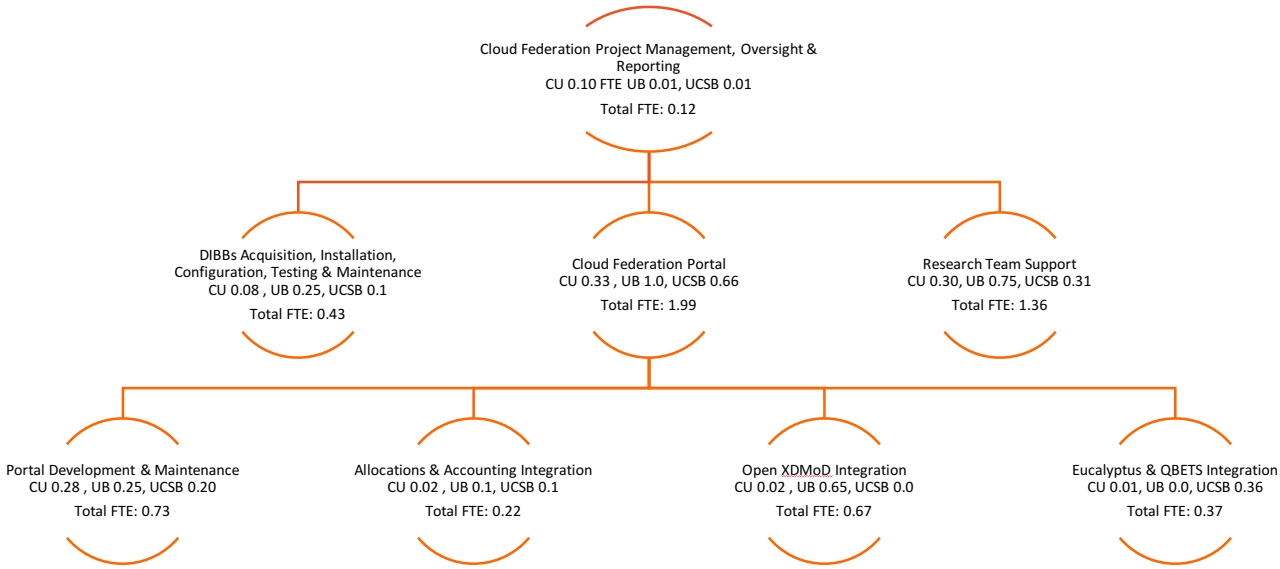
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 11/30/2016

Report 14 of 18

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

This is the fourteenth 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

#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
Lifka had correspondence with Amy Walton regarding the 1st NSF DIBBs PI Workshop (DIBBs17) that is scheduled for January 11-12, 2017 in Arlington, VA. Lifka is Chair for this event. The Cornell team developed the workshop website and 37 out of the 38 currently funded NSF DIBBs projects plan to participate.

1.5 Status Calls

11/8/2016 project status call topics:

- Aristotle NSF reports are available at a restricted access page. For access, send email to help@federatedcloud.org.
- Resolved polling glitch on the cloud usage graph (3 sites were polling at different intervals).
- Allocations dash board is coming.
- Cornell is investigating deploying a Ceph storage back end with their year 2 funding.
- UCSB is investigating what to buy with their year 2 funding. They are leaning towards more CPU since their cloud is running at 70% utilization!
- Cornell has a working Docker container with MPI ready for testing by Pryor's aerosol particles use case team.
- A UCSB graduate student is trying to get the solar panels to work properly on the sensor platform. Another graduate student is making progress on the EC soil analysis and has a new idea for building a SaaS for doing EC mapping.

11/22/2016 project status call topics:

- Discussed what to show on the cloud usage graph (Aristotle vs. local usage).
- Discussed the use of Aristotle for education, possibly providing some limited access provided it is within reason (so we are not swamped). Would need to capture this usage through XDMoD.
- UCSB will provide a local seminar on using Ceph.
- Discussed submitting REUs.

NSF
#1541215

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

2.1 Hardware Acquisition

- UB ordered 10G network equipment for their public connection; it should be installed, tested, and in production by the end of December. They are also investigating how to spend the remainder of year 2 funds (compute vs. storage). Current thoughts are to add more node controllers (compute) and use their development cloud to test Ceph. They would then invest in Ceph storage in year 3.
- Cornell is preparing Ceph storage requirements to send out to bid to Dell, HP, and other vendors.
- UCSB submitted quote requests for year 2 hardware to Dell, HP, and Iron Systems (Supermicro). They will be adding compute to their clouds. UCSB is also working on incorporating an HPE equipment donation into their cloud infrastructure for bursting (this indirectly affects Aristotle, i.e. bursting potential).

2.2 Software Installation, Configuration, and Testing

- UB migrated their production cluster from 4.2 to 4.3. They continue to troubleshoot with the HPE Eucalyptus team a S3 bucket problem that they encountered after the migration. They also had to rebuild all the instance-store images in order to be able to spin up instance-store VMs.
- Cornell migrated their test cluster to 4.3 and encountered a known bug where instances with attached EBS volumes wouldn't boot. They are now testing early 4.4 code which has the promise to support OAuth 2 authentication. They are working closely with the 4.4 developers.

2.3 System Maintenance

- Cornell/UCSB continue to diagnose network bottlenecks between southern CA and central NY.

The infrastructure planning table was updated this month:

	Cornell (CU)	Buffalo (UB)	Santa Barbara (UCSB)
Cloud URL	https://euca4.cac.cornell.edu	https://console.ccr-cbls-2.ccr.buffalo.edu/	https://console.aristotle.ucsb.edu
Cloud Status	Production	Production	Production
Euca Version	4.2.2	4.3	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)

#1541215

10Gb Interconnect	Yes	10Gb inter-cluster; 1Gb external, 10Gb external planned	Yes
Largest Instance Type	28 core/192GB RAM	24 core/192GB RAM	16 core/16GB RAM
	* 168 additional cores augmenting the existing Red Cloud (376 total cores)	** 144 additional cores augmenting the existing Lake Effect Cloud (312 total cores)	

2.4 Potential Tools

- **CloudLaunch**
  The Cornell team continues to work on deploying a virtual cluster in Red Cloud with a generic compute node image for functional testing, including running sample jobs.

- **HPE Helion Eucalyptus**
  Cornell began OAuth 2 testing (4.4) and is working with the HPE developers on the implementation.

- **Supercloud**
  Nothing new to report this month.
  .

3.0 Cloud Federation Portal Report

Content updates to the project are ongoing: https://federatedcloud.org.

The usage graph https://federatedcloud.org/using/federationstatus.php was completed last month; it shows basic early usage data from all 3 sites. For ease of conformity between federated sites, we have requested that the REST API code provided by UB via GitHub be modified to report time in UTC/GMT. We expect to implement Open XDMoD and QBETS in first quarter 2017.

We completed putting a process in place to request and approve access to online NSF project reports by individual.

The portal planning table (pages 5-7) was updated this month..

Portal Framework			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 12/2016**	**1/2017 - 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	Implement content/functionality as shown in following sections. Add additional information/tools as needed, such as selecting	Release portal template via GitHub. Update periodically.

#1541215

	writing any site downloads to the database.	where to run based on software/hardware needs and availability.	
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 – 12/2017**	**4/2017 – 12/2017**	**1/2018 - 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/materials to provide asynchronous training 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.
User Authorization and Keys			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 1/2016	**2/2016 – 5/2016**	**6/2016 – 3/2017**	**1/2017 – End**
Plan how to achieve seamless login and key transfer from portal to Euca dashboard.	Login to the portal using InCommon.	Beta testing Euca 4.4 with Euca console supporting Globus Auth. Will deploy and transition to Euca 4.4. on new Ceph-based cloud.	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	N/A	N/A
Allocations and Accounting			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**3/2016 –3/2017**	**3/2017 – 6/2017**	**6/2017 – End**
Plan requirements and use cases for allocations and account data collection across the federation. Design database schema for	Implement project (account) creation in the database and display on the portal. Integration hooks for user and project creation/deletion	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.

#1541215

Users, Projects and collections of CPU usage and Storage Usage of the federated cloud.	and synchronization across sites.		
Metrics and Usage			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 7/2016	**7/2016 – 9/2016**	**10/2016 – 3/2017**	**10/2017 - End**
Buffalo team utilize Cornell scripts to design a REST API for basic cloud data and deploy at 3 sites and publish usage data to project portal (completed). Buffalo currently standardizing the API by using UTC across the sites and refactoring the code efficiency. Buffalo also completed a redesign of the XDMoD data warehouse to support cloud metrics and is moving this into the testing phase.	Provide documentation for installing XDMoD and SUPReMM at individual sites. Install Open XDMoD/SUPReMM at individual sites and begin data collection. This includes installation of SUPReMM and the data collection piece at the federation sites. Begin integration with federated authentication providers. Currently waiting for latest release of Open XDMoD (v.6.5.1) which will be available at year end at http://open.xdmod.org/. Note: this version does not support cloud metrics but will give sites an opportunity to get infrastructure in place for a future version that does.	Federated data collection will ship data from XDMoD instances at the individual sites to a master XDMoD instance at UB where overall cloud data will be displayed. This is in alpha testing at UB with completion planned for 3/2017.	A prototype cloud realm using Euca data is planned for 10/2017. When completed, federated data from all 3 sites will be available at the master XDMoD instance. Release materials via GitHub. Update periodically.

3.1 Software Requirements & Portal Platform

We continue our work to implement Globus OAuth2 authentication. The next release of Eucalyptus is expected near the end of 2016 which will allow us to incorporate OAuth to facilitate seamless support from the portal to the Eucalyptus console.

3.2 Integrating Open XDMoD and QBETs into the Portal

UB continues work on benchmarking proposed changes.

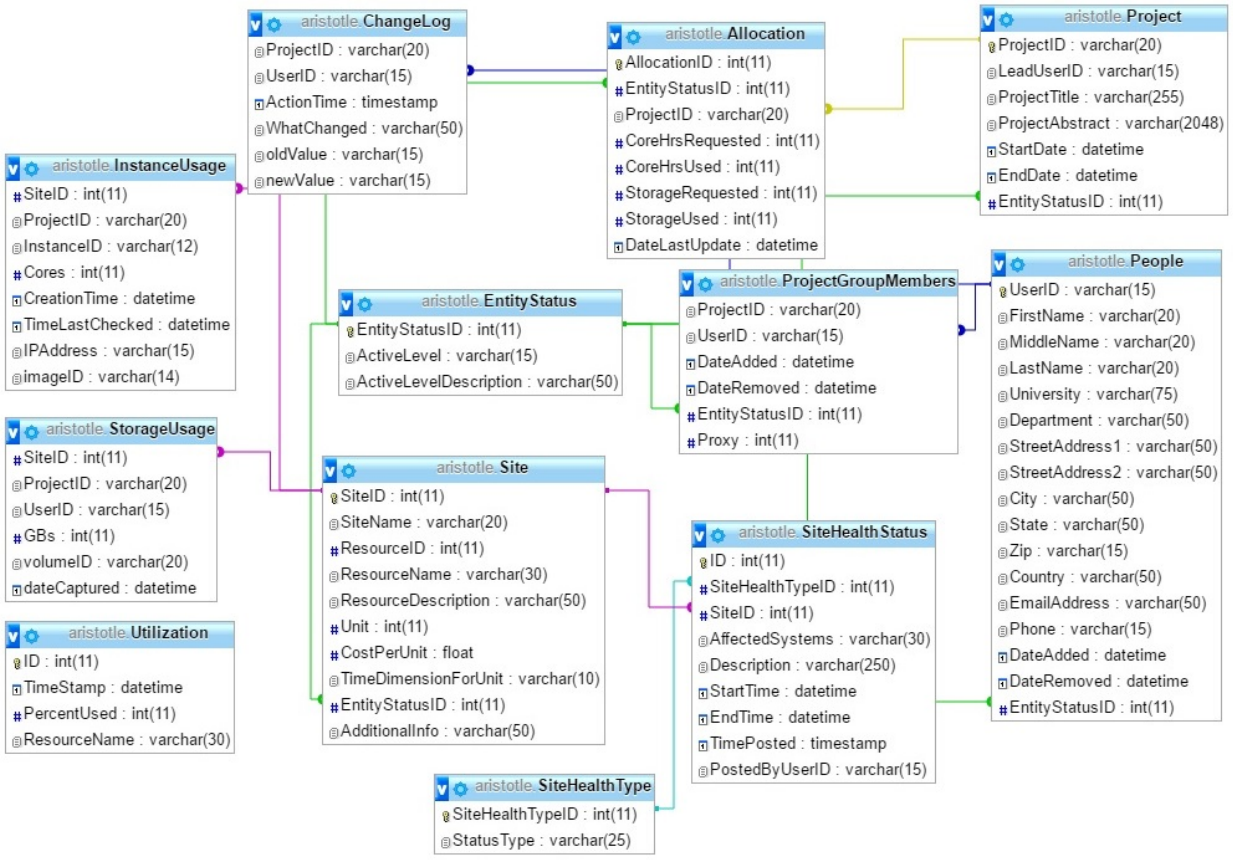
3.3 Allocations & Accounting

There were no changes to the database schema (page 8) this month.

#1541215

We continue our work to implement Globus OAuth 2 authentication; it has now been added for portal authentication. Users can now use their credentials from InCommon, Google, and other identify providers supported by Globus Auth to log in. In addition to using this method of authentication to access portal content, the next release of Eucalyptus, expected near the end of 2016, will allow us to incorporate OAuth 2 to facilitate seamless support from the portal to the Eucalyptus console. The Cornell team has been working with the HPE team on testing Globus Auth support for the Euca Console in the upcoming Eucalyptus 4.4 release. Testing and debugging on nightly builds are ongoing on a test cloud at Cornell.

Development on accounting and allocations is proceeding. The database and tables with test data are complete, and interface implementation is starting. We are also setting up scripts to import project usage data into the database.



4.0 Research Team Support

4.1. General Update

- Docker scripts are now able to mount volumes. This is necessary for receiving output from MPI and is also useful for high-speed data transfer from containers.
- Performed initial testing of Docker Swarm on Red Cloud. Prepackaged demos for Swarm are working, but our MPI demo isn't working yet on Docker Swarm.

#1541215

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
Based on our results from running the change detection model on 200 years of climate simulation outputs, we have started our interpretation of the results and corroborating them with expert knowledge. Additionally, UB ran the same analysis on reanalysis data (from 1948-2006) to validate against known results. While the distributed framework provides excellent scalability on the Lake Effect cloud, the overall time is still long for interactive analysis. We are developing an approximation of the change detection algorithm which can significantly reduce the analysis run time.

Use Case 2: Global Market Efficiency Impact
We now have a license for using OneTick software which is the core of the framework we developed to analyze high frequency financial data. This framework is now setup on the UB Lake Effect cloud. The long-term goal is to make this framework accessible to other researchers. To achieve this goal, we need more than 7TBs storage space. This request was recently approved by the Aristotle cloud management team at UB. The next plan is to start providing access to this framework to a Graduate Research Assistant and eventually demonstrate to other researchers within UB and potentially researchers at Cornell.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
A Docker file and image for WRF has been built and testing is underway. Tristan Shepherd, a new Cornell postdoc from the Pryor group, is being onboarded into Aristotle.

Use Case 4: Transient Detection in Radio Astronomy Search Data
No new updates this month.

Use Case 5: Water Resource Management Using OpenMORDM
Adam Brazier and Dave Hadka (Penn State collaborator) will be meeting to define a deliverables plan. This discussion will be broadened to include the entire Reed group.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
A manuscript continues to be written by CU's Nana Ankrah that is informed by simulation data generated on Aristotle resources.

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

- *"Where's The Bear?" (formerly the camera trap analysis application)* – There is a new verification effort underway for this application involving a larger set of manually tagged images. UCSB plans to go into a full production run with 240,000 images soon.
- *Agricultural Food Security Project* – The grapes element of this project is suspended until early sprint due to grape vine dormancy. The oaks investigation continues.

NSF
#1541215

5.0 Outreach Activities

5.1 Community Outreach

- Cornell featured Aristotle at the SC16 conference exhibit in Salt Lake City on November 14-17. Three use cases, one from each site (UB Big Geospatial Data; CU Gut Microbiota; and UCSB Multi-Sourced Data Analytics to Improve Food Production and Security) were highlighted in a presentation at Cornell's booth. Cornell briefed AWS, NSF staff, universities, and OEMs and ISVs on the Aristotle concept and year 1 status. The presentation is posted on the Aristotle portal at https://federatedcloud.org/papers/SC16AristotleCloudFederationOverviewAndUseCases.pdf.
- Cornell designed and launched the DIBBs17 website at https://dibbs17.org and is managing communications, registrations, and other details for the NSF's first DIBBs PI workshop.

#1541215