ARISTOTLE
CLOUD FEDERATION

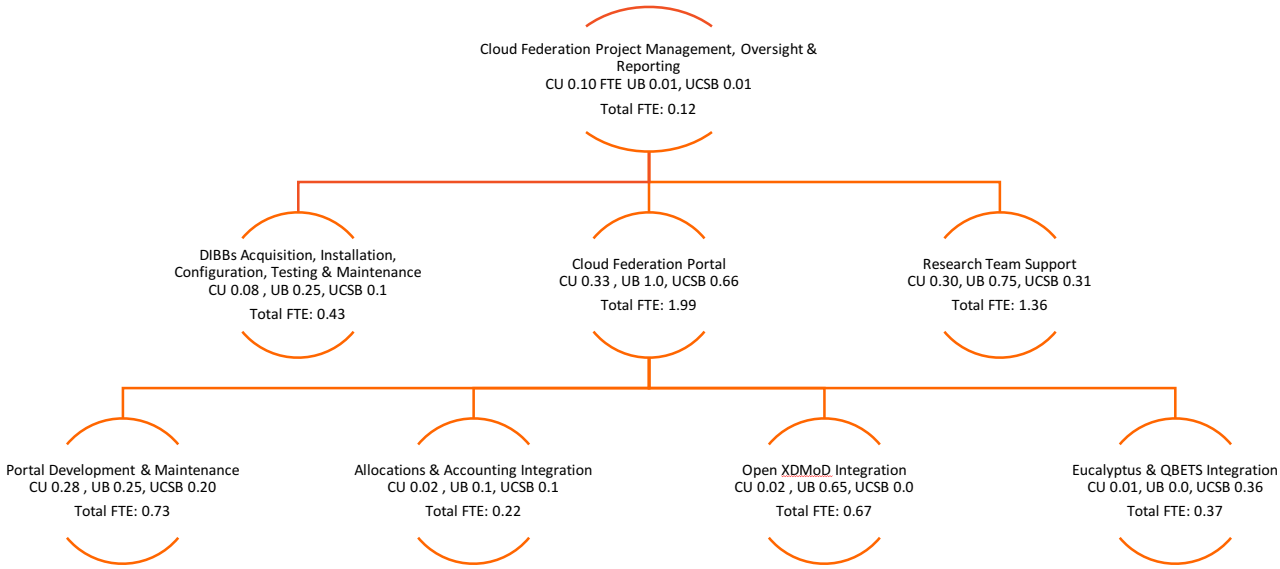
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 10/31/2016

Report 13 of 18

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

This is the thirteenth 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

1.0 Cloud Federation Project Management, Oversight & Reporting Report 3
1.1 Subcontracts 3
1.2 Project Change Request 3
1.3 Project Execution Plan 3
1.4 PI Meetings 3
1.5 Status Calls 3

2.0 DIBBs Acquisition, Installation, Configuration, Testing & Maintenance Report 4
2.1 Hardware Acquisition 4
2.2 Software Installation, Configuration, and Testing 4
2.3 System Maintenance 4
2.4 Potential Tools 5

3.0 Cloud Federation Portal Report 5
3.1 Software Requirements & Portal Platform 7
3.2 Integrating Open XDMoD and QBETs into the Portal 7
3.3 Allocations & Accounting 7

4.0 Research Team Support 8
4.1. General Update 8
4.2 Science Use Case Team Updates 9
Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data 9
Use Case 2: Global Market Efficiency Impact 9
Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties 9
Use Case 4: Transient Detection in Radio Astronomy Search Data 9
Use Case 5: Water Resource Management Using OpenMORDM 9
Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota 9
Use Case 7: Multi-Sourced Data Analytics to Improve Food Production 9

5.0 Outreach Activities 10
5.1 Community Outreach 10

#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 discussions with Amy Walton and Bob Chadduck about leading a DIBBs PI workshop. Cornell submitted a proposal on 9/27/2016 which was funded on 10/27/2016. The 1st NSF DIBBs PI Workshop (DIBBs17) will be January 11-12, 2017 in Arlington, VA. Lifka will Chair DIBBs17.

1.5 Status Calls
10/27/2016 project status call topics:

- Analysis of PY2 hardware needs underway
- Secured/received HPE support license
- Completed basic graph showing early use at all three sites
- Discussed new concept possibly worthy of future investigation: putting a virtual cluster in front of HPC cluster to improve time to science.

10/11/2016 project status call topics:

- Clarified on 2nd year funding for UB and UCSB
- UB sent their usage URL and CU implemented it into the portal
- UB developed a real-time monitoring tool for the cloud infrastructure which CU and UCSB will get from GitHub and evaluate.

10/2016 HPE OAuth2 WebX demo with David Kavangh, Tony Beckham (developer), Steve Jones (support), Jenny Loza (PM) and Dmitrii Calzago (Aristotle Executive Advisory Committee)

- Successfully authenticated Euca Console to Globus (using a Google account) and mapped the token to a role in an Eucalyptus account
- Completed Eucalyptus implementation per AWS standard with assume-role
- Aristotle team provided HPE the following feedback
    - Users might be on multiple projects/Eucalyptus accounts. On the login screen, users should be able to enter which Eucalyptus account to log in, and then go through Globus Auth. If Globus Auth returns success, the user then gets mapped to the role in the selected Eucalyptus account.
    - Once logged in, users should be able to log out of the current account on the Euca console and switch to a different account.
    - HPE says it is possible to come up with IAM policies for the mapped role that gives different levels of permissions according to the Globus Auth sub so we can give different users different permissions if desired.

#1541215

    - HPE says it is possible to do all the provisioning via command line/API so we can automate account management. They are using standard AWS tools to provision the role to which Globus Auth is mapping on their development system.
- Agreed on future plans
    - HPE wants to release Euca 4.4 in Dec. 2016 or Jan. 2017 with this feature. They will do this before the HPE split to make sure the release is not delayed due to organizational changes.
    - HPE will ship us beta code in a few weeks to test and give feedback. HPE want us to give the final green light to release this feature.
    - Once Cornell has the 4.4 beta up and running, testing will occur and Cornell will develop and share the account management code for provisioning Globus Auth.

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

2.1 Hardware Acquisition

- Each site is busy planning for PY2 hardware acquisitions. Sites will evaluate usage to determine if they need cores of storage or other infrastructure to support their clusters. Once each site has defined their requirements, they will request vendor quotes and pursue best possible pricing.

2.2 Software Installation, Configuration, and Testing

- HPE Helion Eucalyptus 4.3 testing has begun at CU, UB, and UCSB. Eucalyptus 4.4 will support OAuth2; sites must be at 4.3 prior to upgrading to 4.4.
- CU successfully migrated their test cluster to 4.3. Their next step is to upgrade each server to CentOS/7. The outcome of this testing will determine the path for migrating CU's production cluster. UB and UCSB also successfully upgraded their dev-cloud and test installation from 4.2 to 4.3. UB reports that the process went well with only a few minor issues.
- CU and UCSB continue to diagnose the network bottlenecks between southern California and central New York.

2.3 System Maintenance

- UB shared a locally developed cluster monitoring tool that monitors the status of the cluster infrastructure. Both CU and UCSB are looking forward to implementing it.

There were no updates to the infrastructure planning table 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	4.2.2	4.2.2
Globus	Yes	Planned	Planned

#1541215

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	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**
  The HPE Eucalyptus team gave Cornell a demo on how they plan to support OAuth2. Cornell will be doing OAuth2 testing as soon as the code is available.

- **Supercloud**
  The Aristotle and Supercloud teams collaborated and submitted an allocations proposal to Jetstream. Read more about this in Sec. 4.1, page 8.

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 is now complete, showing basic early usage data from all three sites. The REST API code provided by UB via GitHub has been implemented by each site to share this data with the portal. We expect to implement Open XDMoD and QBETS in ~late 2016/early 2017.

We are working on a process to request and approve access to online NSF project reports by individual; further minor improvements to the approval process are planned before implementation.

The portal planning table (pages 6-7) was unchanged this month.

#1541215

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/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 – 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		Test access to Euca console.

#1541215

	interface with Euca web 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.

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

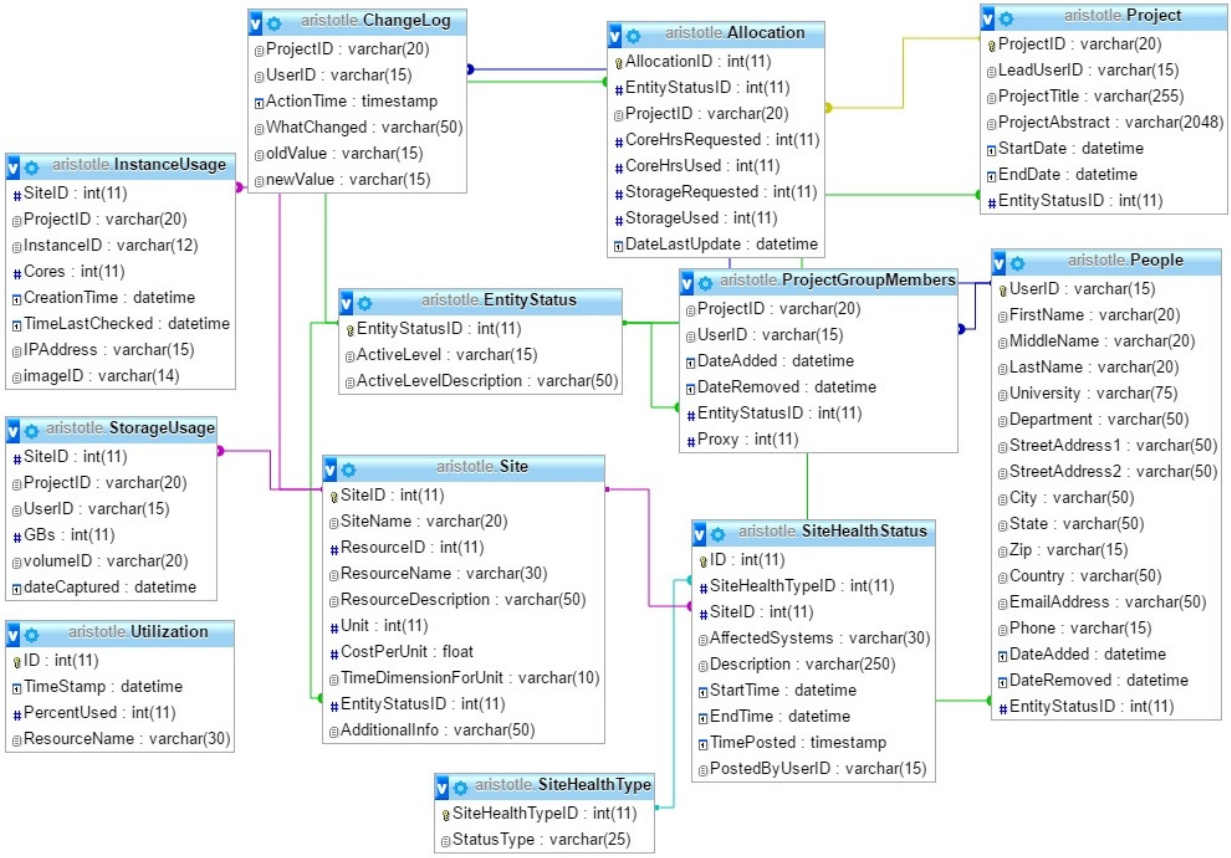
UB continues work on benchmarking proposed changes.

3.3 Allocations & Accounting

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

#1541215

Development on account and allocations is proceeding. The database and tables with test data are complete, and interface implementation will start this month.



4.0 Research Team Support

4.1. General Update

- We wrote a Jetstream allocation proposal for 2.4 million SUs (1 SU = 1 vCPU hour on Jetstream; for many applications this will be close to a core-hour in terms of utility) to provide extension/bursting work for each Use Case. Submitted to XSEDE for mid-December decision. Allocation would start January 2017. Where suitable, migration to Jetstream would be achieved using Supercloud. CU Prof. Hakim Weatherspoon, Supercloud PI, is Co-PI on this allocation request.
- Work continues on MPI using Docker containers. We have extended our framework to enable a script to launch MPI-enabled Docker containers and send "mpi run" commands to the container, and demonstrated that this is working. Initial targets are Use Cases 3 & 6.
- All projects have agreed upon goals for the next project year.

#1541215

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
UB finished first runs of their change detection model on 200 years of climate simulation outputs. They are currently interpreting the outputs, which correspond to anomalous changes in the global temperatures, and consulting with climate experts for validation. They have also performed scalability tests to show that their new distributed method exhibits strong scaling when increasing the number of cores.

Use Case 2: Global Market Efficiency Impact
UB financial researchers have the software license needed to setup the framework in the cloud (in particular, moving around 5TB data to the server and importing it). They have finished identifying the stocks needed for the project. Using the developed framework, they will start requesting the tick-by-tick data from the TRTH database shortly.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
New physics-only WRF instances are being built. The Pryor group at CU has imminent funding for personnel which will be used in part for Aristotle work.

Use Case 4: Transient Detection in Radio Astronomy Search Data
Smoothing code is being written at CU. The basic pipeline architecture was designed with an emphasis on pluggability (including selection of different smoothing, if preferred, although the initial use case is for a pre-prepared dataset with a simple sinc-based smoothing approach).

Use Case 5: Water Resource Management Using OpenMORDM
No major update this month as the Reed team is demonstrating their product at a forum sponsored by the World Bank. Contact has been established with the key graduate student who will be carrying this work forward; training of this student will be documented to facilitate training of the rest of the group.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
A manuscript is being prepared 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) - The science team completed a large scale TensorFlow training and classification run using Aristotle. It ran approximately 20 training runs, each using 2000 core hours, and a classification run, using 1800 core hours to classify a test sample of 10,000 images. Based on the strength of the results the team is preparing to run the classifier on 240,000 images from a single camera. This run will require a fully distributed TensorFlow deployment which is in progress now. A paper detailing the initial classification experiment and the verification of the training methodology is available as a UCSB Computer Science Technical Report at https://www.cs.ucsb.edu/research/tech-reports/2016-07. This technical report is a draft of a paper submitted to the IEEE International Conference on IoT Design and Implementation (http://conferences.computer.org/IoTDI/) for consideration.

#1541215

- *Agricultural Food Security Project* – This project that uses SmartFarm. The team deployed a new Arduino-based sensor platform that is measuring soil moisture content in the Sedgwick vineyard and the oak tree water boxes. Unfortunately, the solar-powered charging mechanism does not appear to be sufficiently weatherproofed to provide more than 2 weeks of continuous operation. Debugging of the hardening effort is underway, but data capture and analysis is stalled pending a fix for the hardware. The existing IRROMETER sensing, however, is currently being used to schedule vineyard irrigation. Sedgwick personnel use the moisture measurements reactively to irrigate the grapes saving approximately 300% of the water used previously. Irrigation scheduling research will suspend shortly due to grape vine dormancy.

5.0 Outreach Activities

5.1 Community Outreach

Cornell will feature Aristotle at their SC16 conference exhibit in Salt Lake City.

Dave Lifka received a supplement to this DIBBs award to Chair the 1st NSF DIBBs PI Workshop (DIBBs17) on January 11-12, 2017 in Arlington, VA. The goals of the workshop are to exchange results and lessons learned from the ~40 currently active projects that have been funded through DIBBs solicitations (NSF 12-557, NSF 14-530, NSF 15-534, and NSF 16-530), and to consider the implications of project results across the existing research portfolio for advances in the vision and goals for data cyberinfrastructure.