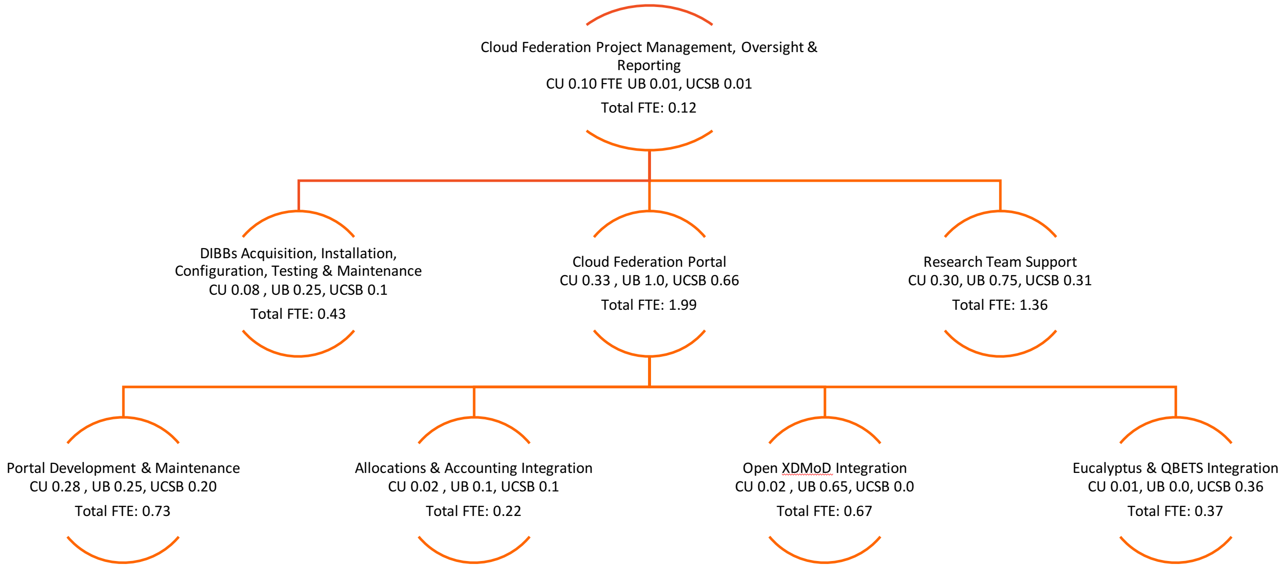
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 8/30/2016

Report 11 of 18

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

This is the eleventh 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).


Cloud Federation Project Management, Oversight & Reporting
CU 0.10 FTE UB 0.01, UCSB 0.01
Total FTE: 0.12
DIBBs Acquisition, Installation, Configuration, Testing & Maintenance
CU 0.08 , UB 0.25, UCSB 0.1
Total FTE: 0.43
Cloud Federation Portal
CU 0.33 , UB 1.0, UCSB 0.66
Total FTE: 1.99
Research Team Support
CU 0.30, UB 0.75, UCSB 0.31
Total FTE: 1.36
Portal Development & Maintenance
CU 0.28 , UB 0.25, UCSB 0.20
Total FTE: 0.73
Allocations & Accounting Integration
CU 0.02 , UB 0.1, UCSB 0.1
Total FTE: 0.22
Open XDMoD Integration
CU 0.02 , UB 0.65, UCSB 0.0
Total FTE: 0.67
Eucalyptus & QBETS Integration
CU 0.01, UB 0.0, UCSB 0.36
Total FTE: 0.37

#1541215

Contents

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

The PI and Co-PIs continue Eucalyptus maintenance support with HPE. HPE is very pleased to be part of this highly visible activity in the national community and we are hopeful that they will support the Aristotle team for the lifetime of the grant at a significant discount.

Lifka had discussions with Amy Walton and Bob Chadduck about leading a DIBBs PI workshop in early 2017. Lifka is working on a proposal to NSF based on guidance from Amy and Bob.

1.5 Status Calls

8/16/2016 project status call topics included:

- Discussions regarding how detailed usage graphs should be.
- Preparation for the 4.3 HPE Helion Eucalyptus upgrade. Cornell will take the lead.
- UCSB experiencing slow downs in speed between East and West coast. Cornell is checking on internal speeds and potentially NYSERNet speeds or other possible factors.
- UB plans to scale their geo use case on CU's Red Cloud using 32 and 64 cores.

8/2/2016 status call topics included:

- Science Team Advisory Committee (STAC) meeting minutes were posted at https://federatedcloud.org/science/advisorycommittee.php. Amy Walton kicked off the meeting. Participation was high.
- Discussions regarding support implications if we decide to use CEPH as our storage platform.

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

Topics discussed this month included:

- **Log History**
  UB noticed that instance history is deleted from log files if the instance is started, stopped, and then deleted. We decided that this is not a problem because it will be rare for a researcher to delete their instance.

- **OAuth2 Support for Single Sign-In Credentials**
  HPE Helion Eucalyptus has announced support for OAuth2 in version 4.4 which is slated for release in September 2016. Eucalyptus 4.3 is out now. The biggest 4.3 release change is it requires CentOS 7. Cornell has a test cluster and will use it to test the 4.3 upgrade procedures and then share results with UB and UCSB. UB is setting up a small test cluster as well.

- **CEPH Storage**
  Cornell and UCSB discussed recent progress investigating CEPH. Cornell shared iozone results that compared reads and writes to CEPH vs. reads and write to the SAN. The CEPH numbers look great. Wolski noted that if we decide to standardize on CEPH (vs. SAN), we will minimize our need for an HPE Helion Eucalyptus support contract. A contract is required if we continue to use HPE's SAN driver. This discussion is ongoing.

The infrastructure planning table was updated this month:

	Cornell	**Buffalo**	**Santa Barbara**
Cloud URL	https://euca4.cac.cornell.edu	https://console.ccr-cbls-2.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**
  The Cornell team is working on deploying a virtual cluster in Red Cloud with a generic compute node image for functional testing, including running sample jobs.

- **Supercloud**
  No updates to report.

3.0 Cloud Federation Portal Report

There were content updates and additions this month to the project portal: https://federatedcloud.org/. We added a basic status graph at https://federatedcloud.org/using/federationstatus.php that will show the percentage of resource used at each site. Currently only Cornell's early usage is displayed.

We are working on pulling in user documentation from the developer's pages located at GitHub. To facilitate document ingestion by the portal, a public repository was created to house the production documents. Documents under development or with sensitive project data are in a private repository.

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/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 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.

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.

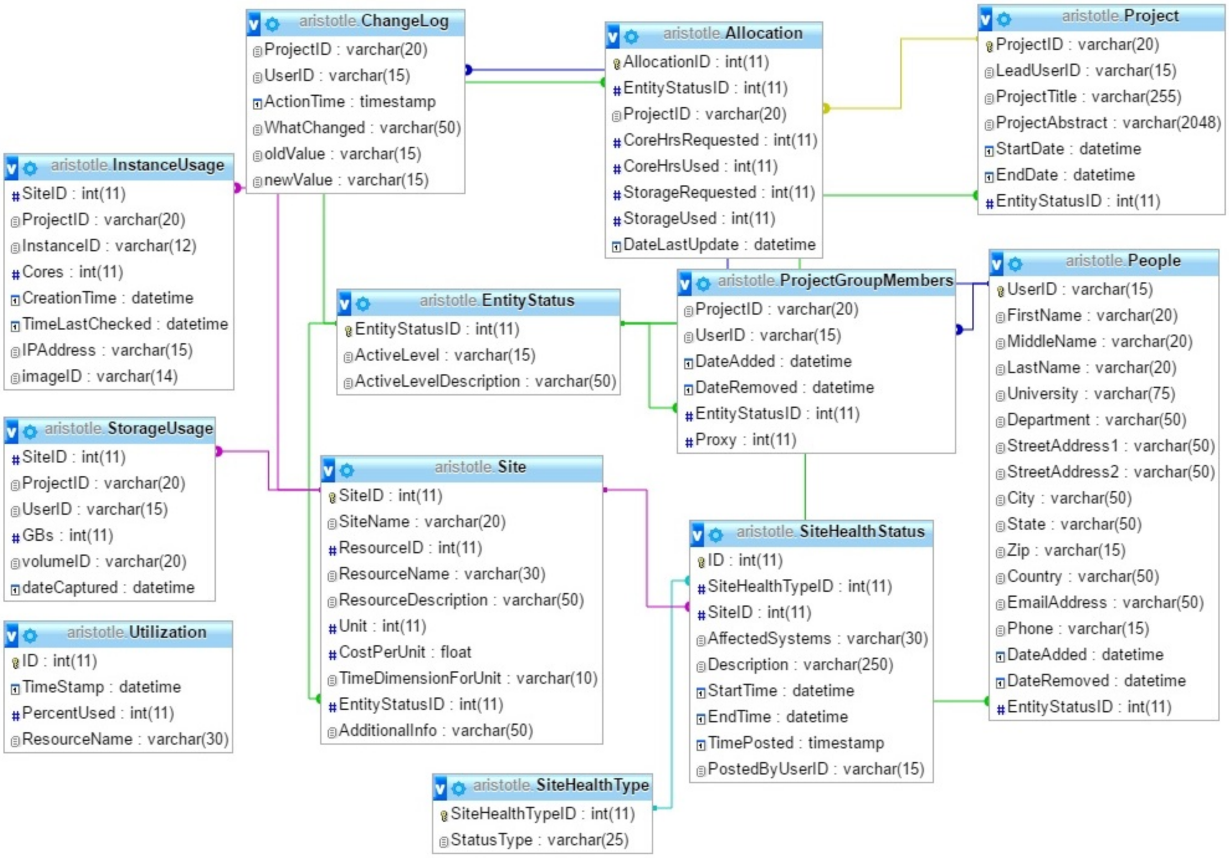
3.2 Integrating Open XDMoD and QBETs into the Portal

UB completed a review of the existing data warehouse to analyze the impact of any changes. They are working on benchmarking the existing warehouse and making the changes necessary to support cloud reporting.

3.3 Allocations & Accounting

There were no changes to the database schema this month.

The usage graph mentioned in Section 3.0 is intended to show basic data until Open XDMoD and QBETS are implemented (~late 2016). As part of the graph creation, GetUtilization, a new stored procedure, was added to pull in the current usage data when the page is loaded. The REST API code provided by UB via GitHub will be used by each site to share this data with the portal. All three sites are working on installing the API. UCSB is working on the installation and troubleshooting a log ingestion glitch. Cornell continues to work on implementing the Aristotle Usage REST API for the Ithaca site; the instance was moved to the Aristotle Infrastructure account and usage will eventually be available at http://aristotle-usage.cac.cornell.edu. UB plans to provide this data after migrating existing HPC jobs in September.



aristotle.ChangeLog
ProjectID : varchar(20)
UserID : varchar(15)
ActionTime : timestamp
WhatChanged : varchar(50)
oldValue : varchar(15)
newValue : varchar(15)

aristotle.Allocation
AllocationID : int(11)
EntityStatusID : int(11)
ProjectID : varchar(20)
CoreHrsRequested : int(11)
CoreHrsUsed : int(11)
StorageRequested : int(11)
StorageUsed : int(11)
DateLastUpdate : datetime

aristotle.Project
ProjectID : varchar(20)
LeadUserID : varchar(15)
ProjectTitle : varchar(255)
ProjectAbstract : varchar(2048)
StartDate : datetime
EndDate : datetime
EntityStatusID : int(11)

aristotle.InstanceUsage
SiteID : int(11)
ProjectID : varchar(20)
InstanceID : varchar(12)
Cores : int(11)
CreationTime : datetime
TimeLastChecked : datetime
IPAddress : varchar(15)
imageID : varchar(14)

aristotle.EntityStatus
EntityStatusID : int(11)
ActiveLevel : varchar(15)
ActiveLevelDescription : varchar(50)

aristotle.ProjectGroupMembers
ProjectID : varchar(20)
UserID : varchar(15)
DateAdded : datetime
DateRemoved : datetime
EntityStatusID : int(11)
Proxy : int(11)

aristotle.People
UserID : varchar(15)
FirstName : varchar(20)
MiddleName : varchar(20)
LastName : varchar(20)
University : varchar(75)
Department : varchar(50)
StreetAddress1 : varchar(50)
StreetAddress2 : varchar(50)
City : varchar(50)
State : varchar(50)
Zip : varchar(15)
Country : varchar(50)
EmailAddress : varchar(50)
Phone : varchar(15)
DateAdded : datetime
DateRemoved : datetime
EntityStatusID : int(11)

aristotle.StorageUsage
SiteID : int(11)
ProjectID : varchar(20)
UserID : varchar(15)
GBs : int(11)
volumeID : varchar(20)
dateCaptured : datetime

aristotle.Site
SiteID : int(11)
SiteName : varchar(20)
ResourceID : int(11)
ResourceName : varchar(30)
ResourceDescription : varchar(50)
Unit : int(11)
CostPerUnit : float
TimeDimensionForUnit : varchar(10)
EntityStatusID : int(11)
AdditionalInfo : varchar(50)

aristotle.SiteHealthStatus
ID : int(11)
SiteHealthTypeID : int(11)
SiteID : int(11)
AffectedSystems : varchar(30)
Description : varchar(250)
StartTime : datetime
EndTime : datetime
TimePosted : timestamp
PostedByUserID : varchar(15)

aristotle.Utilization
ID : int(11)
TimeStamp : datetime
PercentUsed : int(11)
ResourceName : varchar(30)

aristotle.SiteHealthType
SiteHealthTypeID : int(11)
StatusType : varchar(25)

4.0 Research Team Support

4.1. General Update

- A Science Team Advisory Committee meeting occurred 7/27/2016. Minutes were distributed and are available on the project portal.
- Brazier is currently preparing Science Team goals for the next quarter.
- Barker (CU) is researching Docker use in Eucalyptus and has begun documenting how to run Docker on Windows instances. Varun Chandola's UB student (Dinh Tran) has created documents on how to create a Spark cluster in the cloud and how to port OpenNebula VM's to the Aristotle cloud.
- Barker is assessing how to carry out MPI on Aristotle's Eucalyptus platform; Docker is among the options being investigated.

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
We created general purpose scripts to dynamically generate Spark clusters on the Aristotle cloud. We're currently migrating their virtual machines to the UB cloud. We also developed a revamped browser-based interface (webglobe) to allow users to drive the sustainability use case analysis on the cloud.

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 building a pipeline to process the data. Unfortunately, PALFA data is temporarily unavailable due to a machine failure (of a non-Aristotle file server), but work is ongoing to restore access to the data.

Use Case 5: Water Resource Management Using OpenMORDM
We are making plans for building the software stack and testing it. This project will be a candidate for the work Barker is conducting on implanting MPI in Aristotle.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
We established some initial guidelines for file access between the host OS and Docker containers, which will be useful given the large amount of data that would otherwise need to be transferred; these guidelines were added to the documentation under "Docker/Data Transfer." We also established that the sequence analysis currently being targeted will be massively parallel and a good candidate for testing HPC with Docker by starting multiple Docker containers to run identical processes on different data. To this end, we updated a library that abstracts over the Docker REST API and implemented test scripts that successfully start multiple containers in parallel, waits for them to finish (or until a time limit is reached), and then tries to shut down the containers gracefully. This script should be extendible to the sequence analysis program once the container and requisite data are in place.

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

- *Wildlife Survey*: The work to automatically identify animals in the camera trap imagery is proceeding. The project has switched from Caffe to Google's TensorFlow machine learning system. TensorFlow provides a better success rate for camera trap pictures using an existing animal imagery data base; however, the rate is not high. Part of the problem is that there are relatively few pictures of some species of animals. For example, in almost 200,000 images from last month, there are only 6 pictures of bears (although many pictures of deer). The project is now looking at "mashing up" images from Google's image search with backgrounds taken by the camera traps. That is, by superimposing images of different animals from Google on the fixed and empty background taken from the camera traps, the hope is that the training data will become more effective.

- *Precision Agriculture:* Using the sensing technology that is in place, Sedgwick personnel have started to do irrigation scheduling. Part of the current issue is that they do not know what the "drain rate" is for the vineyard under study. That is, the sensors are generating an alert for when to put water on the grapes, but the current soil content map generated by the agronomists seems to be making inaccurate predictions of when to turn the water off. The project (a collaboration with Fresno State) is looking at directly correlating Electrical Conductivity (EC) with the different soil types in the vineyard. With this correlation, it should be possible to create alerts for both water start and shutoff. Looking ahead, the goal is to "close the loop," i.e., to make the process completely automated.

- *Drought Ecology of Valley Oaks:* The project installed WATERMARK sensors in each of three test seedlings to measure "slow drip" effects from water boxes; however, the sensor control platform (new for this project) is not yet functional. Initial, short-term tests show good results, but the robustness necessary for a long-term study is not yet possible. The team is actively developing the reliability engineering required.

5.0 Outreach Activities

5.1 Community Outreach

Rich Wolski et al. wrote a technical report on the *Probabilistic Guarantees of Execution Duration for Amazon Spot Instances* that is available at the Aristotle portal: https://federatedcloud.org/about/publications.php.

A Securities and Exchange Commission Branch Chief FCW presentation on *Challenges and Opportunities in the Cloud* highlighted the NSF Aristotle Cloud Federation as "what's next" in cloud computing: http://www.digitalgovernment.com/media/Downloads/asset_upload_file34_5802.pdf.