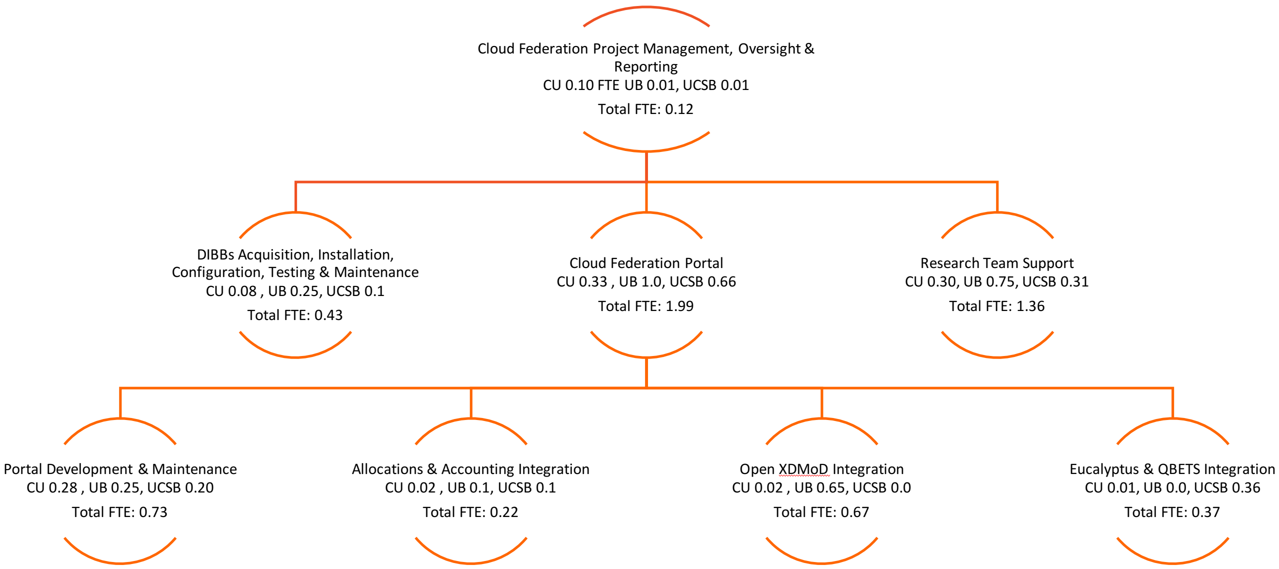
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 6/30/2016

Report 9 of 18

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

This is the ninth 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

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

David Lifka participated in the XSEDE 1 PY5 annual review at the National Science Foundation on 6/13-16/2016 as the Level 2 lead of Campus Infrastructure (XCI). This was a review of PY5 accomplishments and XSEDE 1 lessons learned that can be applied to XSEDE 2. In several sessions questions regarding Cloud use cases and interoperability were raised by the review panel. It is clear that Aristotle will bring important lessons learned in Cloud usage modalities and federation models to the national community.

1.5 Status Calls

Project status calls were held on 6/7/2016 and 6/21/2016. Topics included:

- Portal development and plans to launch the portal during June 2016.
- XSEDE interest in QBETS as a queueing management system.
- Plans to analyze Cornell's and Buffalo's allocations and accounting models for design synergies.
- Discussions regarding Ceph which is working well for both S3 and volumes at UCSB with a few error handling and timing issues. UCSB is stress testing their Aristotle cloud before declaring it production ready.
- Aristotle team review of Supercloud to better understand its capabilities and limitations.

1.6 Project Planning and Preparation

The initial portal design was completed and the project web site will be public on 6/30/2016. The project URL is federatedcloud.org.

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

- **CU**

  CU's Red Cloud infrastructure is in production with researchers running.

- **UB**

  UB's cloud is online and Center for Computational Research (CCR) staff has completed testing and will add research accounts the week of 6/27/2016. CCR staff also implemented automated image building using Packer.

- **UCSB**
  UCSB made significant progress this month in both services and software.

  Cloud Services:
  - Eucalyptus 4.2.2 running > currently stress testing.
    - Added secondary Availability Zone (AZ) for testing.
    - Back up database for service integrity with Barman.
  - Ceph v.0.94.7 (Hammer) > added pools for new AZ.
  - Planning UCSB's website (aristotle.ucsb.edu).

  Software:
  - Integrating Ansible for provisioning and orchestration.

The infrastructure planning table has been 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	Internal User Testing	Installation Testing
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
	* 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.

2.3 Industry Influence

Cornell and HPE had discussions regarding HPE support for Globus Auth. The HPE development team is working on a plan and will provide Cornell with an estimated implementation date as soon as they have one.

3.0 Cloud Federation Portal Report

The portal planning table below was updated this month. A notable change was in the User Authorization section; we will be using Globus Authentication instead of InCommon for a better interface with both XSEDE and the Euca Console. This also affected the Euca Tools section; development of minimal tools will not be necessary given the expectation of an improved interface to the Euca console.

Work on implementing Globus authentication has been 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.

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.

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

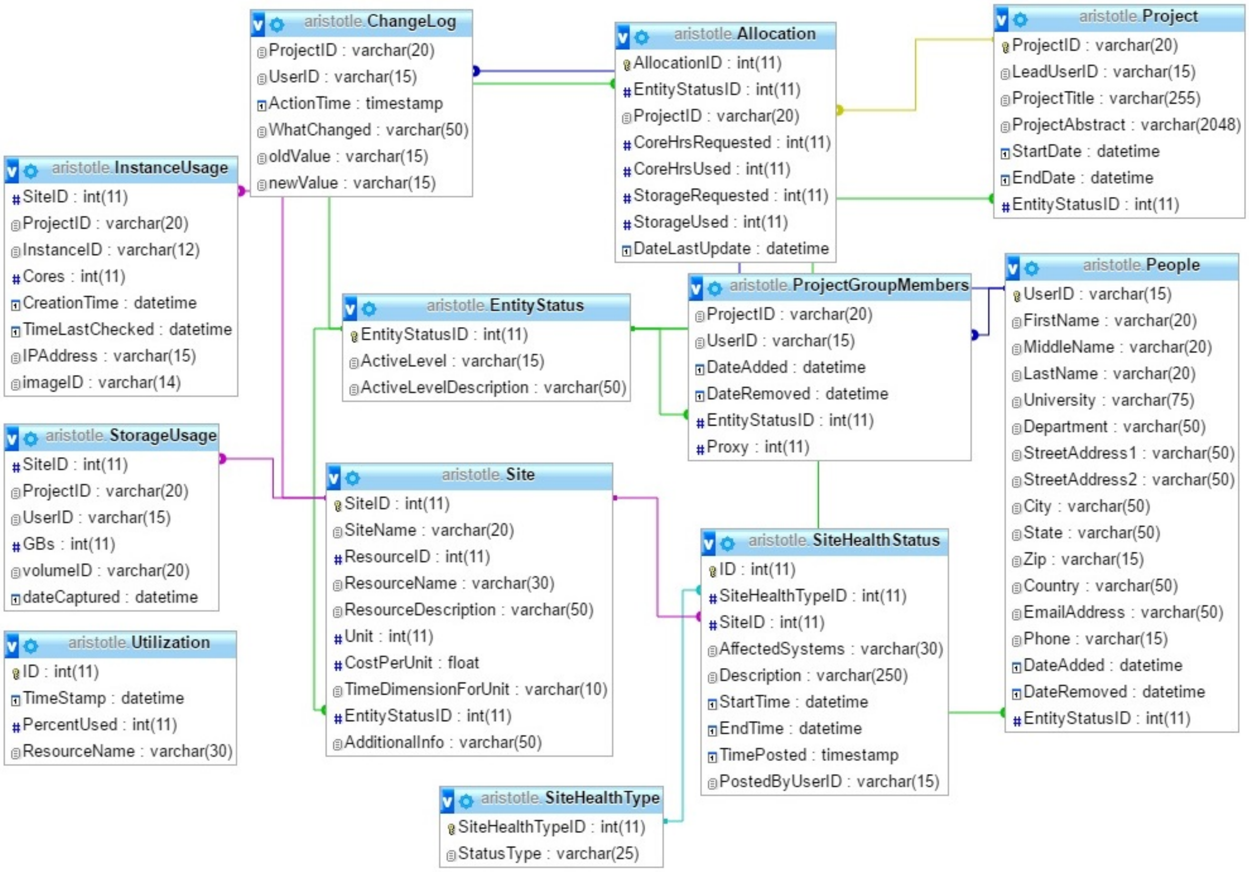
The new portal design implemented in May was tested, improved, and is now publicly available at https://federatedcloud.org/. Next month we will continue to add content and functionality.

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 downloaded at Cornell; Cornell and UB are working together to deploy this API, which will write usage data to the portal database, which will generate usage information.

3.3 Allocations & Accounting

Refinement continues of the database schema for allocations and usage data, included below. Changes this month include a new Utilization table for the portal utilization graph, and new tables for posting status and downtimes, called SiteHealthStatus and SiteHealthType.



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 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
The prototype that demonstrates the use of cloud to analyze large volumes of NetCDF data to understand the local impact of climate change using simulation data is ready. Chandola's team is currently preparing a paper for *BigSpatial 2016 – the 5th International Workshop on Analytics for Big Geospatial Data* that will summarize their results.

Use Case 2: Global Market Efficiency Impact
The migration of the OpenNebula image to the Aristotle cloud is now complete. Access to Thomson Reuters Tick History (TRTH) database has been secured and initial testing has begun, anticipating more work later this summer.

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

Use Case 4: Transient Detection in Radio Astronomy Search Data
Software requirements were identified and architecture planned by Brazier and Chatterjee. Modification of file-handling from Julia Deneva's psrfits2psrfits routines is underway to allow deresolution of data.

Use Case 5: Water Resource Management Using OpenMORDM
No update for June.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
Barker constructed a new Linux system that is specified by a Docker file. The purpose of the system is for genomic and transcriptomic data-preprocessing that will serve as input to metabolic modeling, and will be used primarily by postdoc Bessem Chouaia. Barker had a meeting with Bessem and another postdoc, Nana Ankrah, to discuss Docker and Aristotle. Nana and Bessem have completed construction of the initial metabolic models to be used for the analysis, and Nana plans to run a Monte Carlo flux sampling algorithm on an Aristotle instance as a first step very soon. To this end, the Gurobi convex optimization solver and the COBRA Toolbox for MATLAB have been installed on our previously configured Windows instance. In the process, Barker created a page on Docker and continued adding documentation to the "Metabolic Models" page on the Aristotle GitHub wiki.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production
Sedgwick is going to begin publishing its camera trap data to eMammal. eMammal collects, stores, and shares camera trap data for scientists and citizen scientists: https://emammal.si.edu. These data are useful for addressing important scientific and conservation questions. Sedgwick is gathering approximately 200,000 images per month so the formatting, metadata, and compression workload is significant. A test upload using UCSB and Box.com is projected to take 29 days (almost the whole next month) using the current infrastructure. They anticipate much better turn around once UCSB's new Aristotle resources become available. We are holding off from putting them up on the existing cloud for two reasons. First, the new hardware is just about ready for production (Andreas Boschke is working on it now). Second, the

current cloud needs maintenance which will happen in the next couple of weeks. We didn't want to go live and then to immediate migrate.

In addition, the agricultural analysis work is making progress. Sedgwick is now publishing its weather and moisture sensor data via GSN (a sensor network web service standard): http://128.111.84.213:22001. We will put in a DNS name for the site once we migrate to the new hardware.

5.0 Outreach Activities

5.1 Community Outreach

Initial planning is underway to produce a video on Aristotle and Eucalyptus in conjunction with HPE.