ARISTOTLE
CLOUD FEDERATION

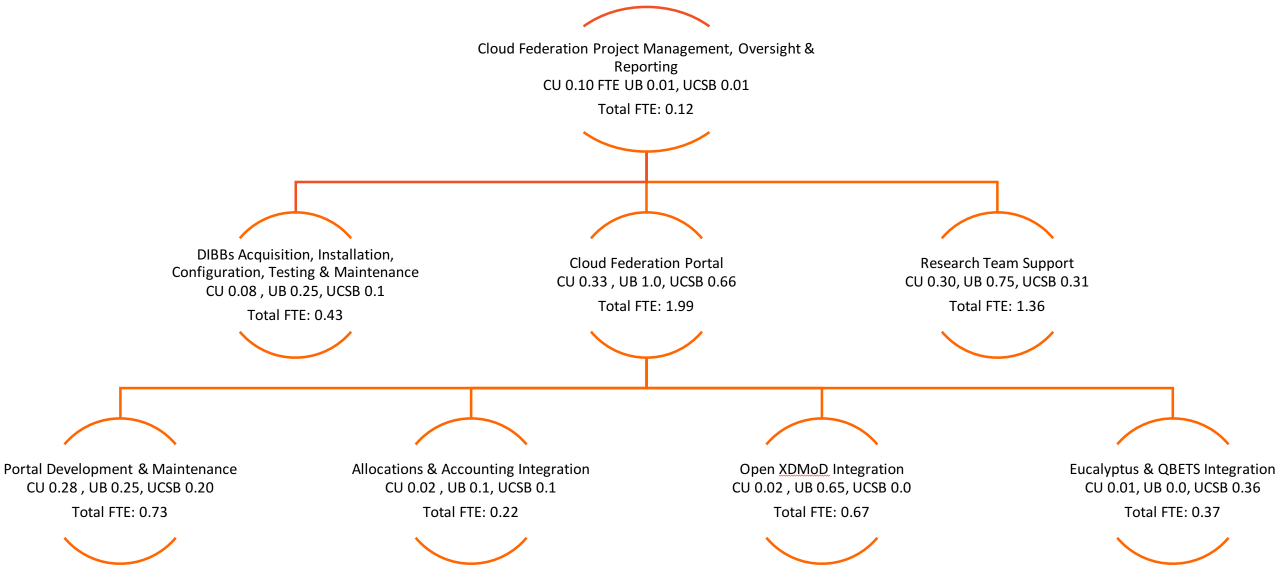
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 9/30/2016

Report 12 of 18

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

This is the twelfth 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

A budget change request was approved by Amy Walton on 9/15/2016 for Hewlett Packard Enterprise (HPE) support. The agreement provides the Aristotle partners with an HPE professional maintenance offering, including support for SAN drivers. As indicated by the deep discounts HPE is offering, HPE is strongly interested in learning more about the NSF cyberinfrastructure community, and how company capabilities can better support this community.

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 to lead a January 2017 DIBBs PI Workshop in Arlington, VA.

An Aristotle Executive Advisory Committee meeting occurred 9/20/2016:

- Attendees included Dmitrii Calzago (HPE), Ian Foster (ANL/U. Chicago), Steve Johnson (Weill Cornell Medicine), Sanjay Padhi (Amazon Web Services), Ben Rosen (Dell), Craig Stewart (Indiana U.), John Towns (XSEDE), Aristotle PIs and Co-PIs, and the Aristotle Portal, Infrastructure, and Science Use Case leads.
- Cornell CS (Zhiming Shen, Hakim Weatherspoon, and Robbert van Renesse) demonstrated Supercloud, a cloud architecture that enables application migration as a service across different cloud providers. Shen started a Supercloud instance and demonstrated how the user has full control over the location where the virtual machine runs with latency minimized by software defined networking. An instance was migrated from AWS to Google Compute Engine and between Red Cloud and AWS. Previous demos showed migration from Red Cloud to Jetstream and back. Discussions will continue on how this work might be further developed to bring value to the national community (John Towns expressed interest in this technology for XSEDE).
- Updates on the portal were provided, including plans for Aristotle to move to OAuth2 (the federated identity management used by XSEDE) which will be built into Eucalyptus 4.4 (Dmitrii Calzago, HPE, asked Cornell to alpha/beta OAuth2). The portal itself will eventually move as a package to AWS where it will be readily available as a template for any organization to use.
- The Infrastructure team explained that all 3 campus clouds are up with users running locally and one use case (UB and CU) ran cross-site via the manual sharing of accounts. This will eventually be automated when all core functionality, i.e., allocations and accounting, are in place.
- The Science Use Case lead provided updates on the 7 use cases which are documented in monthly reports, and results to date/lessons learned are shared at quarterly Science Team Advisory Committee meetings (see minutes at: https://federatedcloud.org/science/advisorycommittee.php).
- Wolski described how UCSB built a system for predicting the "bid price" that an AWS user should bid in the spot market to ensure a minimum duration of execution before AWS terminates the instance. Tests have shown substantial savings. This technology, called DrAFTS (Durability

Agreements from Time Series), was developed as part of the Aristotle project. It uses QBETs (Quantile Bound Estimation Time Series) internally.

- Furlani's team is developing a federated Open XDMoD capability which requires re-engineering the Open XDMoD data warehouse. Cloud metrics will be added to Open XDMoD; QBETS services will integrate its predictive capabilities with other metrics generated by Open XDMoD (bandwidth, VM duration, storage load, etc.) in an attempt to generate statistical bounds on guaranteed delivered performance levels.

1.5 Status Calls

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

2.1 Federation Resource Status Updates

Topics discussed this month included:

- **OAuth2 Support for Single Sign-In Credentials**
  Cornell and UB will use small development clusters to test the upgrade to HPE Helion Eucalyptus 4.3, a CentOS 7-based product. Once at 4.3, both sites will be able to alpha and beta test the OAuth2 code due to be released in v. 4.4.

- **Network Bandwidth**
  Further network testing between sites this month has provided more information but no conclusions. The sites will continue to analyze where the bottlenecks are (if any) and provide a report on the results at the conclusion of testing.

- **Ceph Storage**
  UB's development cluster will deploy Ceph storage in order to test its features and get some Ceph experience. Cornell and UCSB continued to investigate Ceph; thus far, the numbers look very good.

- **Storage Performance**
  UCSB brought a reserved node controller (NC) online and reconfigured to a separate NC to use SSD as ephemeral storage for the Sedgwick users with the goal of improving performance.

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	4.2.2	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	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)	

2.2 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 announced that they are close to having code in support of OAuth2 authentication ready for alpha testers. Cornell volunteered to do early testing on their test cluster.

- **Supercloud**
  A demo migrating a streaming video application between clouds was provided to the Aristotle Executive Advisory Committee.

3.0 Cloud Federation Portal Report

There were content updates and additions this month to the project portal: https://federatedcloud.org/. We added links to user documentation from the developer's pages located at GitHub. To facilitate document links from 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. We are currently implementing a process to request and approve access to online reports by individual. 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	Implement content/functionality as shown in following sections. Add page hit tracking with Google	Implement content/functionality as shown in following sections. Add additional information/tools as	Release portal template via GitHub. Update periodically.

site software.	Analytics, as well as writing any site downloads to the database.	needed, such as selecting 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 – 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	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.

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 – 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

A newer version of php (5.6) was installed this month so we can continue our work to implement Globus OAuth2 authentication. The instance for the main portal site was moved to a single-function project for security reasons. This will allow us to move forward with allocation and project management on the portal.

The next release of Eucalyptus, expected near the end of 2016, will allow us to incorporate OAuth2 support to facilitate seamless support from the portal to the Eucalyptus console.

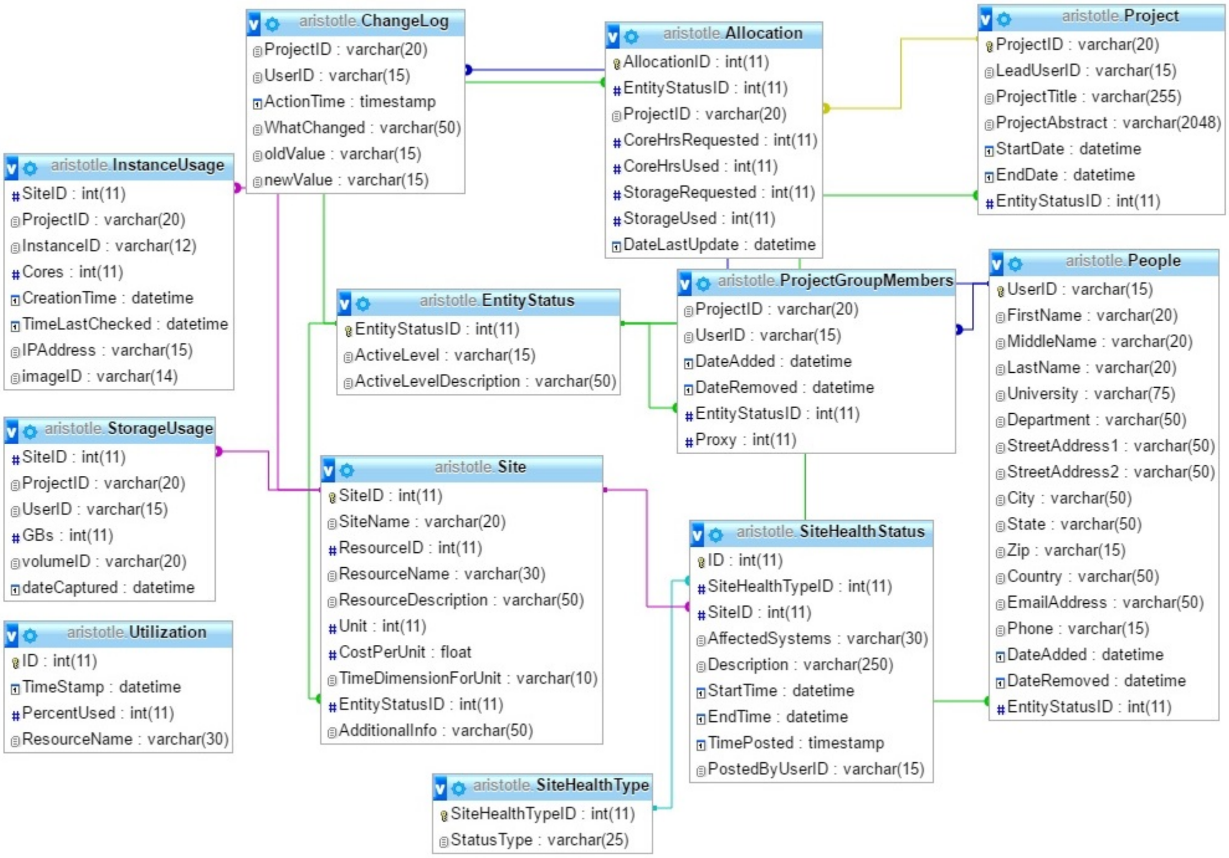
3.2 Integrating Open XDMoD and QBETs into the Portal

UB completed benchmarking the existing data warehouse and is currently working on benchmarking proposed changes.

3.3 Allocations & Accounting

There were no changes to the database schema (see page 8) 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/Q1 2017). 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 continues work on the installation. The log ingestion glitch was resolved by better understanding of the UB code. Cornell continues to work on implementing the Aristotle Usage REST API for the Ithaca site (http://aristotle-usage.cac.cornell.edu). UB plans to provide this data after migrating existing HPC jobs.



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

- Introductory documentation developed by the Science Use Case team was made available via the portal.
- Work continues on MPI using Docker containers.
- All projects have agreed upon goals for the next project year.

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
We ran scalability tests on Red Cloud at CU before migrating the application to Lake Effect at UB.

Use Case 2: Global Market Efficiency Impact
The software required is installed and the researchers plan to begin initial runs on the UB Lake Effect cloud using the TRTH database in the October/November time frame.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties

Paola Crippa tested the WRF-Chem installation running on a single instance and it passed tests. Brandon Barker is approaching this use case as the first target for the MPI/Docker project.

Use Case 4: Transient Detection in Radio Astronomy Search Data

Brazier met with Robert Wharton and Shami Chatterjee to discuss the project and characterize the data. Data reduction code is being prepared.

Use Case 5: Water Resource Management Using OpenMORDM

No update this month.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota

The research team developed results using Windows instances developed by the Science Use Case team at Cornell and running on Red Cloud. Nana Y.D. Ankrah (Cornell) reported that with metabolic modeling, we have been able to show that (1) the metabolism of whitefly symbionts are co-evolved to minimize the overlap of inputs (competition) and outputs (efficiency), (2) differences in the metabolic inputs from the whitefly bacteriocyte drive the difference in EAA released from symbionts although the relationship between substrate and product concentration is not always linear for all EAAs, and (3) the whitefly bacteriocyte acts as a sink for NH3 and facilitates symbiont N recycling for EAA production through shared biosynthetic pathways.

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

- *Wildlife Survey*: A grad student and undergraduate student developed the animal-identification system that runs Google's TensorFlow in parallel on Aristotle to auto-identify animals from camera-trap images (200,000 photos/month).

- *Drought Ecology of Valley Oaks:* Aristotle back end processing is functional for collecting moisture sensor data to improve seedling development of CA Live Oaks; testing continues.

5.0 Outreach Activities

5.1 Community Outreach

The Mapping Transcriptome Data to Metabolic Models of Gut Microbiota use case team presented research at an American Society for Microbiology (ASM) conference (http://conferences.asm.org/images/1905%20asm%20beneficial%20microbes%20web2.pdf) that was based on results from using a Windows instance in Red Cloud:

Ankrah, N.Y.D., Luan, J. & Douglas, A. Evolution of Metabolite Exchange in a Three-Partner Symbiosis. *6th ASM Conference on Beneficial Microbes*, September 10, 2016, Seattle, WA.

The Aristotle team configured a Windows instance with MATLAB and associated software (COBRA Toolbox, Gurobi, etc.) to support this research.