ARISTOTLE
CLOUD FEDERATION

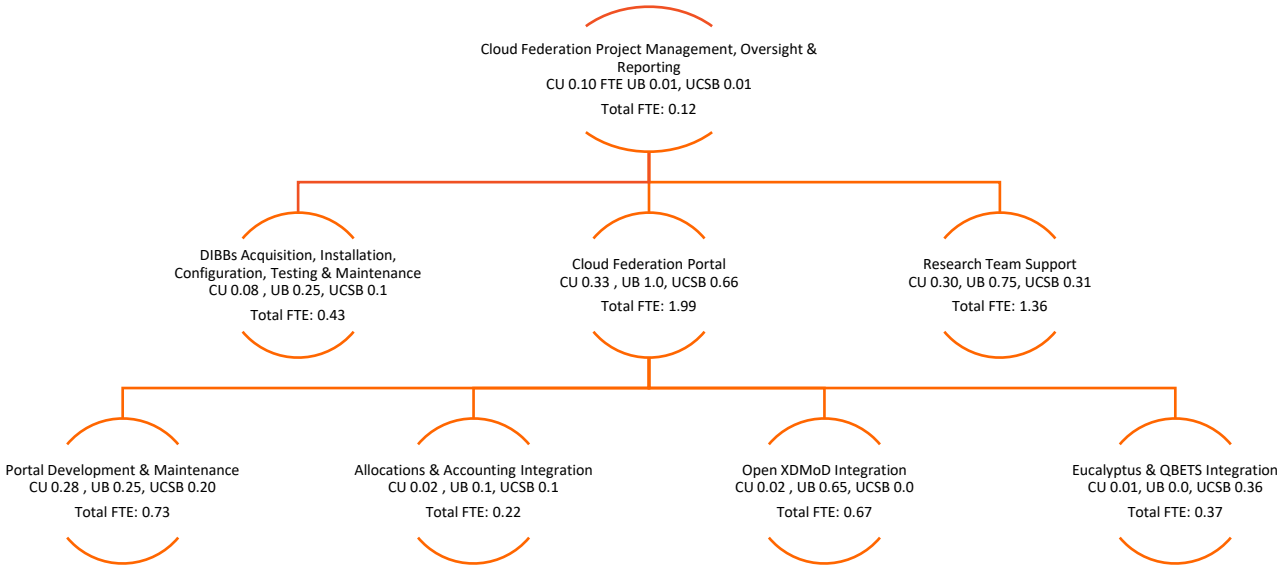
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 4/27/2016

Report 7 of 18: Part 1 of 2

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

This is the seventh 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

NSF
#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
Rich Wolski and the Globus Genomics team conducted several experiments using DrAFTS: a method for predicting spot market bid prices for AWS spot instances. DrAFTS (which is an acronym for "Durability Agreements from Time Series") uses QBETS to make several internal predictions which it then combines into a suggested bid price. The DrAFTS bid both ensures a specific time duration until the spot instance will be terminated and minimizes the maximum spend the user might incur. Early experimentation using a DrAFTS service hosted at UCSB with Globus Genomics test runs indicate that the bids are both correct (in terms of the duration they ensure) and tight (in terms of the spending control). We are preparing an initial paper describing these experiments for submission to the USENIX Symposium in the next two weeks. We also plan further integration work and to collaborate on possible scheduler enhancements for the Globus Genomics framework.

Tom Furlani and the UB science use case team led by Dominik Roesch received approval from Amy Walton on 4/22/2016 to purchase the Thomson Reuters Tick History (TRTH) database in order to investigate under- or over-pricing in common stocks. The main challenge is to estimate the unknown true value of a stock. One way to address this, is to compare prices of the same stock across different markets (e.g., Australia, Brazil, China, USA, etc.). For this to occur, it is crucial to have intraday price data for stocks trading around the globe. Having this data allows us to create time varying estimates of under- or overpricing in each specific market. The TRTH database is the only available source for international intraday price data and will allow us to extract price data from 1996 to now for 1,000 stocks around the world. The subscription will be active for one year during which we can update our data. The subscription also allows us to access a web interface that provides view-only access to several petabytes of data. After the subscription expires in one year, we will be allowed to continue to use the extracted data for existing projects for another 4 years. The subscription also allows other academic researchers to use the data for scientific investigations. Subscription pricing is $800/month and a $2,000 one-time fee. UB intends to cancel the subscription after the first year (total cost $11,600) and use the database for the remaining 4 years of the Aristotle award.

The first Science Team Advisory Committee (STAC) meeting was held on 4/1/2016. Program Director Amy Walton spoke first and emphasized that the Aristotle project is of great interest to NSF because it addresses NSF cyberinfrastructure sustainability and metrics concerns. Next, the Aristotle portal team and infrastructure team leads provided an overview of progress to date and future plans. The science team lead then introduced the science use case team members who described their research goals, instances created to date, and provided helpful feedback to the project team. The STAC meeting minutes are available in Appendix A (pages 11-14 of this report).

Lifka was invited by NSF to present an overview of Aristotle Q1/Q2 progress in Arlington which was also held on 4/1/2016. The presentation was well attended by NSF program directors including Irene Qualters, Amy Friedlander, Bob Chadduck, Amy Walton, and James Kurose. There was great interest from all in attendance with close to 30 minutes of Q&A. Afterwards, Amy Walton told Lifka and Furlani (who also attended to answer questions on behalf of the team) that she was very pleased with the presentation, turnout and our ability to stimulate discussion among NSF program directors well after the presentation was over.

That same week Lifka met with Chaitanya Baru, CISE Senior Science Advisor for Data Science. Baru and Kurose were planning a trip to Amazon to discuss interactions with NSF and the CI community. Lifka shared information on Cornell's use of AWS for research at Cornell and as a resource for Aristotle. The possibility of the Cornell Center for Advanced Computing being the institution responsible for XSEDE allocations of AWS resources was discussed as a logical extension to the Aristotle project. Discussions about next steps with Baru and Walton are ongoing.

1.5 Status Calls with Aristotle Program Director

Project status calls were held on 4/13/2016 and 4/26/2016. Topics included:

- Launch of the Aristotle portal is planned for May 2016 with these initial elements: home page highlights, project summary, project leadership, partners, news and outreach, presentations, publications, federation overview, user support, software and tools, emerging technologies, use case goals, and use case descriptions. Getting started and user guide information will be added in the coming months. We plan on adding the resource status, project dashboard, XDMoD and QBETS by the end of CY2016.
- Initial Supercloud testing was successful. An image at Cornell was run on Jetstream and vice versa. More testing will continue with possible demos.
- CU's Red Cloud is in production; UB's Aristotle cloud is installed and will enter test phase soon; and, UCSB's cloud hardware is racked and the UCSB team is working on tools to roll out the cloud environment.
- CU created 3 VMs for science use cases and has set up 5 accounts using Cornell's current allocations system.
- UCSB is planning test runs of recently collected soil electrical conductivity towed array data fused with soil density survey maps to better understand soil properties. The first image processing run with wildlife camera trap data is also expected to occur soon. Scientists, citizen scientists, and industry have expressed enthusiasm for both of these projects.
- Furlani plans to provide an Aristotle presentation at the Best Practices in Data Infrastructure Workshop, May 17-18, 2016, in Pittsburgh.

1.6 Project Planning and Preparation

Major portal design work was accomplished in April and the launch of initial portal elements will occur in May 2016.

A draft of the federated cloud requirements for Federated XDMoD was created this month. See Appendix B (pages 15-25 of this report).

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**
  Cornell's Red Cloud infrastructure is now in full production. The CU team also assisted UB with their cloud stack networking setup.

- **UB**
  The UB Aristotle cloud is installed. UB is working on refining the installation modules (Puppet) in order to automate the installation. They are trying out a few different configurations to ascertain which provides the best performance. The next step will be the testing phase when the UB team will ensure that all nodes perform consistently. UB's wildcard certificate has been requested and will be installed upon arrival.

- **UCSB**
  The UCSB team has received all of their year one hardware (Dell servers/storage and HPE switches). The hardware has been racked and a public IP address space (/23) has been assigned. Efforts continue towards setting up the build/configuration tools to roll out the cloud environment, along with planning the CEPH installation to provide EBS and S3 storage.

The CU/UB/UCSB infrastructure planning table has been updated:

	CU	UB	UCSB
Cloud URL	euca4.cac.cornell.edu	ccr-cbls-2.ccr.buffalo.edu**	TBD**
HPE Helion Eucalyptus Version	4.2.1	4.2.1	4.1.2
Migrate to 4.2.1	1/1/2016	Started 4.2.1	5/1/2016
Globus	Yes	Planned	Planned
InCommon	Yes	Yes	Yes
Hardware Quotes	Hardware deployed March 2016.	Cloud hdw and sw stack installed; currently testing deployment and management tools.	Hardware has arrived; installation in progress.
Hardware Vendor	Dell	Dell	Dell
# Cores	168*	144	140
Ram/Core	4GB/6GB/8GB	8GB	9GB
10Gb Interconnect	Yes	Yes	Yes

***** 168 additional cores augmenting the existing Red Cloud
******UB and UCSB will install Aristotle as new clouds and will not integrated with existing clouds.

2.2 Potential Tools: CloudLaunch & Supercloud

The Cornell team continues to massage the CloudLaunch code so that it can be run by outside clients. The Cornell Computer Science group reports that they have successfully migrated a virtual machine image (VM) between Red Cloud and Jetstream. They will provide a demonstration in the near future.

#1541215

3.0 Cloud Federation Portal Report

No changes were made this month to the portal planning table below:

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.	Update materials to be federation-specific.	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.	Get 4.2.1 federated key after InCommon login.	Move seamlessly to Euca console after portal InCommon login.

#1541215

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.	Implement minimal set of Euca Tools to bridge portal to Euca console.	Add/refine/update, based on ongoing collection of user feedback.	Release via GitHub. Update periodically.
Allocations and Accounting			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**3/2016 – 5/2016**	**6/2016 – 9/2016**	**10/2016 – 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 – 3/2016	**2/2016 – 5/2016**	**6/2016 – 10/2016**	**11/2016 - End**
Implement graphs of basic usage data, including % utilization, available resources, and user balance, using scripts from Cornell CAC 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.

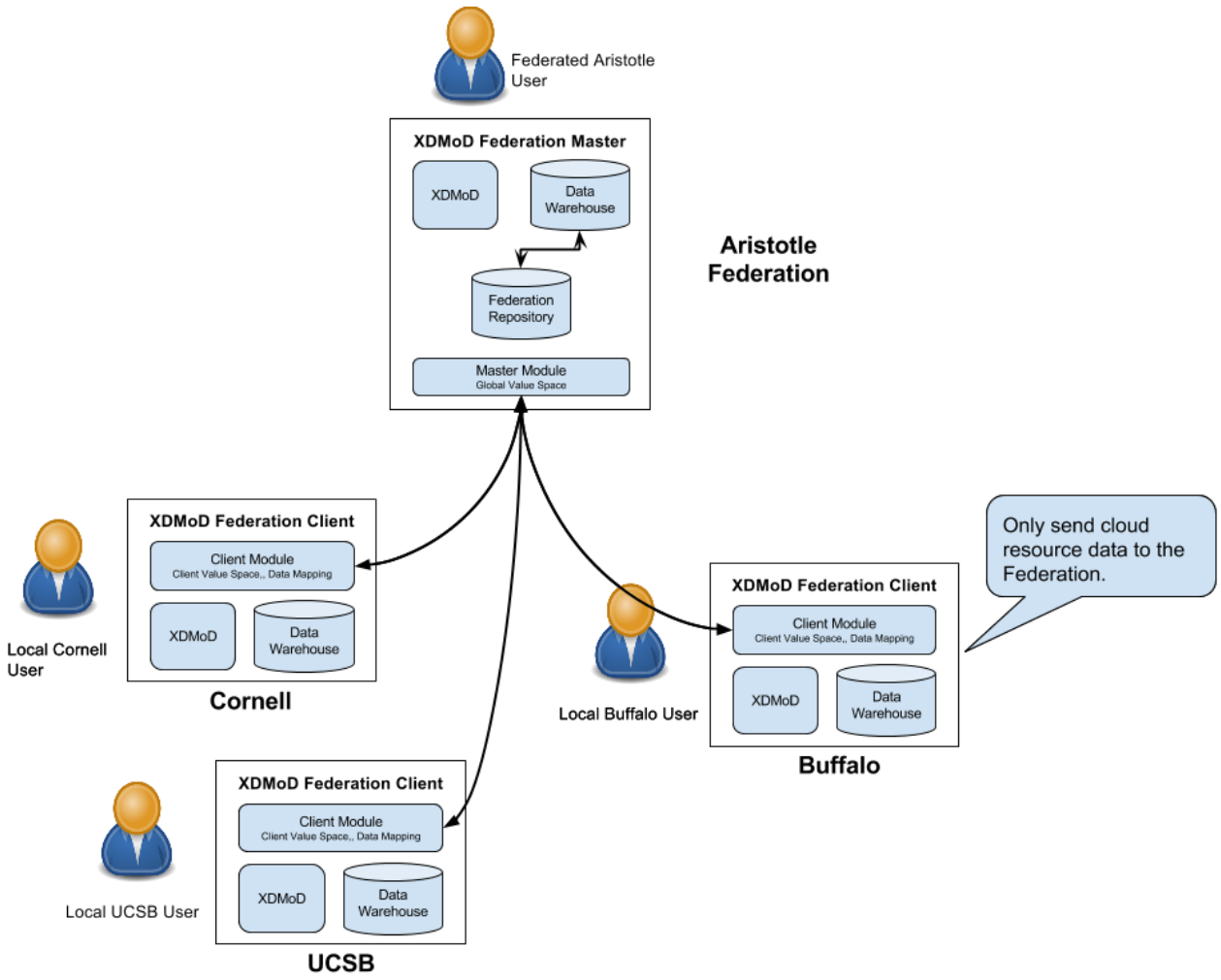
3.1 Software Requirements & Portal Platform

The portal design and content was revised this past month, incorporating and fleshing out the plan for design, content, and ease of use. Implementation of the new design began and is expected to go public in May 2016. This month InCommon authentication was implemented as well.

#1541215

3.2 Integrating QBETS into Open XDMoD

Federated XDMoD will support the collection and aggregation of data from individually managed HPC centers into a single federated instance of XDMoD for displaying federation-wide metrics (see figure below). The detailed requirements were completed this month and we are now moving on to the implementation. Data particular to an individual center will be available by applying filters. Each participating center will deploy an XDMoD instance through which local data will be collected and viewed. The Client will ship data for selected resources to a central repository for aggregation and display by a master XDMoD instance. In essence, we will create a central repository to support the display of data collected from multiple organizations (Service Providers) including jobs, users, projects (allocations), organizational hierarchies, etc. Data from individual organizations will be mapped into a common format where appropriate. This is similar to what XSEDE has done with the XDCDB but **without a single, XSEDE-provided set of resources, users, and allocations defined in common across all of the organizations**. See Appendix B for a draft of the federated cloud requirements for Federated XDMoD.

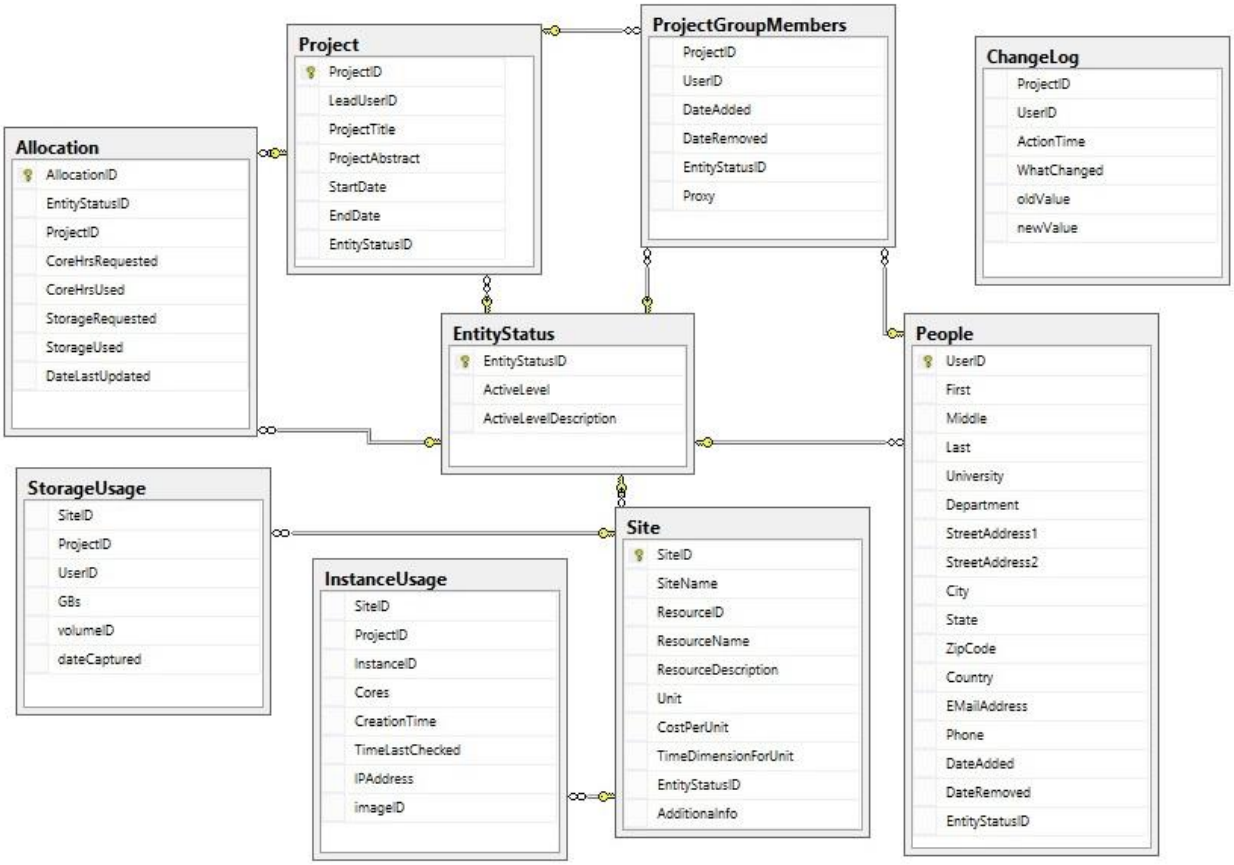


#1541215

3.3 Allocations & Accounting

The database schema for allocations and usage data (see figure below) is currently being expanded to include data on Site Health. This capability is for staff to post patch days, scheduled down times, unscheduled down times, updates, etc. The temporary database currently holds data on five project groups, increased from four project groups last month. The data currently collected includes project, user, and usage data, enabling the researchers to begin setting up instances.

Data from UB and UCSB will be ingested into the main database by use of a REST API developed by UB, which exposes Eucalyptus accounting data based on the scripts provided by Cornell. This month Cornell continued testing this API for use within the federation.



4.0 Research Team Support

4.1 General Update

- A guide to using Eucalyptus for Aristotle science teams was started on our GitHub wiki in Markdown. Initial topics include an overview of Eucalyptus, tips on image management for scientists, and an introduction to the euca2ools package. Arising from discussions between

#1541215

science and infrastructure teams, the guide will include "best practices" to allow scientists to minimize the time spent on image management and systems administration. Versioned releases to the portal are planned.

- The first quarterly meeting of Scientific Team Advisory Committee occurred 4/1/2016 (minutes available in Appendix A).

4.2 Science Use Case Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
We've created a fully functional stack on UB's Lake Effect cloud to analyze distributed NetCDF data stores. We extended the SciSpark library to handle individual massive NetCDF files (instead of a large collection of small files) by implementing spatial distribution into the SciSpark framework. Further refinements to the web-based client are being made to allow users to interact with large scale climate simulation data.

Use Case 2: Global Market Efficiency Impact
A UB PI is actively investigating open source image processing software to process images on the VM installed on the cloud. The PI for the second project is in the process of acquiring data needed for the use case. An image with a 2TB volume was created and is available on Cornell's Red Cloud; the PI, however, is interested in migrating an existing instance from an OpenNebula system. Varun Chandola asked a student to assess the feasibility of this transfer.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
Brazier attended a colloquium by Paola Crippa detailing their use case work, then met with Sara Pryor and Crippa to discuss implementation. Discussions included whether it might be possible to burst to Azure using Supercloud. An image with an MPI version of WRF-Chem was configured for user-level access and made available in an initial 4-core version for users.

Use Case 4: Transient Detection in Radio Astronomy Search Data
A machine image was created with the basic software and user access configured. Installation of the processing suite, and data access, is planned for May 2016.

Use Case 5: Water Resource Management Using OpenMORDM
An instance was built, user-level access created, and key software installed: it is ready for users. The science team is currently investigating additional software requirements.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
The users tested connectivity and are scaling their workload locally. They plan to use Aristotle when modeling multicompartment networks due to increased computational demands.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production
Progress has been made on two subprojects:

- Image Processing Project - the latest camera trap images from the last quarter have been gathered. A manual upload stage is underway and once that completes, we're planning a test

NSF
#1541215

of the image identification application. The early prototype appears to be working. This will be the first test with a complete image corpus.

- Soil Moisture Survey - We seem to have worked out many (but not all) of the bugs with the vendor hardware and software. Unbeknownst to the team, a soil temperature sensor needed to be added to the sensor mesh (we installed it last week). We also now have the data from the soil electrical conductivity (EC) towed array soil density survey. Work is now taking place to fuse the moisture sensor data, the EC map, SSURGO maps, and soil analysis probes.

5.0 Outreach Activities

5.1 Media Outreach

A 3/30/2016 "Efficient cloud reduces time to science" article in *Scientific Computing World* referenced the Aristotle Cloud Federation:
http://www.scientific-computing.com/news/news_story.php?news_id=2798

5.2 Industry Outreach

Cornell wrote/edited an HPE Helion Eucalyptus case study on Red Cloud and Aristotle this month. It is now published:
https://www.hpe.com/h20195/v2/GetPDF.aspx/4AA6-5086ENW.pdf

A 4/4/2016 feature in *CloudStories* provides additional details:
http://community.hpe.com/t5/Grounded-in-the-Cloud/Cornell-University-s-Red-Cloud-A-hybrid-cloud-model-for-academia/ba-p/6847772#.Vx-DkNIrJaQ

#1541215