ARISTOTLE
CLOUD FEDERATION

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

## Science Team Advisory Committee (STAC)

## 7/27/2016 Meeting Minutes

**Adam Brazier, Aristotle science team lead**
**brazier@cornell.edu**

**Invited to 7/27/2016 Science Team Advisory Committee Meeting** *(attendees italicized)*

**Cornell University (CU):**
*James Cordes* cordes@astro.cornell.edu
*Shami Chatterjee* shami@astro.cornell.edu
*Adam Brazier* brazier@cornell.edu
*Angela Douglas* aes326@cornell.edu
Bessem Chouaia bc335@cornell.edu
*Nana Ankrah* na423@cornell.edu
*Brandon Elam Barker* brandon.barker@cornell.edu
Sara C. Pryor sp2279@cornell.edu
*Patrick Michael* Reed pmr82@cornell.edu
*Bernardo Carvalho Trindale* bct52@cornell.edu
*Julianne Dorothy Quinn* jdq8@cornell.edu
*Resa Reynolds* rda1@cac.cornell.edu
Susan Mehringer shm7@cornell.edu
*Paul Redfern* red@cac.cornell.edu
*David Lifka* lifka@cornell.edu

**University at Buffalo (UB):**
*Tom Furlani* furlani@ccrbuffo.edu
*Varun Chandola* chandola@buffalo.edu
Cristian Tiu ctiu@buffalo.edu
*Dominik Roesch* drosch@buffalo.edu
Brian A. Wolfe bawolfe@buffalo.edu

**University of California, Santa Barbara (UCSB)**
*Rich Wolski* rich@cs.ucsb.edu
*Andreas Boschke* andreas@cs.ucsb.edu
Kate McCurdy kate.mccurdy@lifesci.ucsb.edu

**National Science Foundation**
*Amy Walton* awalton@nsf.edu



#1541215

**Meeting Introduction -** *Adam Brazier, science team lead; David Lifka, PI (CU)*
Adam opened the Aristotle Science Team Advisory Meeting (STAC) by emphasizing that science is a key factor that the project will be judged on, but it's not the only thing. Dave explained that the project goal is to improve time to science. We want to demonstrate that the elasticity of federated cloud resources (extending between sites but also to Jetstream, NSF clouds, and AWS) and enable the science teams to get their work done faster. All 3 Aristotle clouds are up and running: Cornell Red Cloud, University at Buffalo Lake Effect, and UC Santa Barbara Aristotle. We've been working very hard in getting all the parts and pieces ready to go—testing the sharing of resources across sites, developing the Aristotle portal, and working on OpenXDMoD with cloud support and QBETS for metrics collection, analysis, and graphical display. Our goal is to have all these pieces in place (V.1) for our 18 month review and the science use cases running with initial insights into what works well in the federated cloud paradigm and what are the challenges. At the end of the day, we want to demonstrate that Aristotle is a complementary resource to existing Cyberinfrastructure and that we can build a resource over time that benefits the national research community.

**NSF Introduction -** *Amy Walton, NSF program manager*
The entire Foundation (e.g., CISE, GEO, Mathematical and Physical Sciences) is very interested in the Aristotle project and in making data more transportable, accessible, and meaningful to interested communities. They are watching this project closely with a keen interest in the social experiment of sharing resources and bursting. Each science use case team should drive hard over the next 6 months to see what you can learn while using the federated cloud and make a list of what works well and what doesn't Adam added that we will see the fruits of this project occurring elsewhere if this goes well.

**Infrastructure Update -** *Resa Reynolds, infrastructure team lead (CU)*
- We're happy to report that all 3 sites have production clouds running with initial users.
- Some cross campus experiments have occurred.
- Users should relay any questions they have to their local infrastructure people or to Adam Brazier at brazier@cornell.edu.

**Portal Update –** *Adam Brazier on behalf of Susan Mehringer, portal team lead (CU)*
- The portal was launched in July at https://federatedcloud.org with information on the project goals, leadership, advisors, partners, news & events, and publications as well as the federation overview, emerging technologies, and science use cases. Future features will include allocations and usage info, systems status (how busy each node is), user documentation/training, and eventually OpenXDMoD and QBETs for advanced resource prediction.
- Users can run at their own sites (CU, UB, or UCSB). If you need more hours, let us know. There should be no barriers to using your onsite machine as the federated capability continues to be developed.
- CU is working on instructions on how to use the machine. Users can send their GitHub ID to Adam so you can see the original wiki and can make contributions which would be helpful to all. After vetting by the science teams, this documentation will be cloned and displayed in the portal.
- We encourage users to document the software and tools they use or develop as well as their experiences. For example CU (Steven Lee) recently developed and documented a parallel version of WRF-Chem built for the cloud.
- We encourage user feedback on the portal as it evolves.

**Science Team Update** - *Adam Brazier, science team lead (CU)*
- All Science Use Cases have allocations on their campus clouds. We want to see science being done now and would like everyone to produce something by the end of the quarter.



#1541215

**Individual Science Team Updates**

**Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data -** *Varun Chandula update (UB)*
- We're creating a framework so scientists can interact with large amounts of geospatial data, e.g., remote sensing and climate data. This framework will potentially allow users to analyze and visualize geo data, and run advanced machine learning codes on top of it. We're currently using the cloud to write a paper on climate data analysis, with historical and future simulation outputs. Spark has been very valuable and we will provide documentation on how to use it.
- Our goals include figuring out how to add other user communities without overloading the federation. Dave Lifka suggested that we might consider developing a science gateway to make our tools available to the broader community and to provide allocations on Aristotle, and perhaps even Jetstream.

**Use Case 2: Global Market Efficiency Impact** - *Dominik Roesch (UB)*
- Using the TRTH (Thomson Reuter Tick History) database, we plan to construct a way to measure efficiency in stock prices and derivations from fundamentals across international markets. "Efficiency" means how well prices reflect the fundamental value of a company.
- We're currently looking at sample stock trades. In a few months, we plan to run an analysis on the sample time series and then extend it to different countries.

**Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties** - *Adam Brazier on behalf of Sara Pryor (CU)*
- The Pryor team is developing a model to better understand atmospheric aerosol particle concentrations, one of the largest uncertainties in understanding climate change.
- This research required WDF-Chem so CU built a parallel version for the cloud.
- The team plans to run a limited model next quarter.

**Use Case 4: Transient Detection in Radio Astronomy Search Data** - *Shami Chatterjee (CU)*
- Our team plans to use PALFA data acquired by Arecibo, scale it down to a manageable size, and then look for single pulses in a sea of noisy data that is very fast in time and frequency.
- The goal is to run on demand. We also want to introduce new search algorithms fairly easily and run them on partial or full datasets to test them out, and then use them on other radio telescopes.
- We're currently examining the data sets to understand exactly what we have and will then decide what specific approach to take.

**Use Case 5: Water Resource Management Using OPENMORDM** – *Patrick Reed (CU)*
- We're interested in evaluating Aristotle largely for the analysis of data. We completed large scale Blue Water runs on NC Research Triangle and Vietnam Red River data. We'd plan to test our analytic model in the cloud.
- Our next quarter goal is to get the software stack up and develop a simple usage guide.
- Ultimately, we're interested in using many cores effectively. We'd like to, over the course of the project, see if we can find a way to spin up many containers across various clouds to do similar things. We may not run as fast or as efficient as an HPC system, but if we can run at scale and get access quicker that would be valuable to analysts. We have an algorithm that would be amenable to that type of environment. If we can demonstrate scalability, we'd like to begin planning for an AWS service for US municipalities.



#1541215

**Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota –** *Angela Douglas (CU)*
- Our goal is to develop metabolic models using algorithms with widely varying run times.
- We've been doing a lot of prep work and have started to conduct these simulations.
- There's a new National Microbiome Initiative to advance microbiome science. We'll be at that meeting in September where we plan to link up with other micro bio researchers. We're optimistic there will be a lot activity over the next 6 months.
- Cornell CAC collaborator Brandon Barker added that he's working on the transcriptomic pipeline using Docker and anticipates it will meet 90-95% of their requirements. This is being documented on the wiki. NixOS might be a future option.

**Use Case 7: Multi-Sourced Data Analytics to Improve Food Production –** *Rich Wolski (UCSB) on behalf of Kate McCurdy (Sedgwick Reserve/UCSB)*
- Sedgwick is developing an animal-identification app using Caffe. The Caffe framework is running on the new Aristotle cloud. Currently, camera traps take 200,000 photos per month that must be manually classified. The goal is to automate this system and improve image identification. In addition, the images are being transferred to eMammal (https://emammal.si.edu/) for the broader community use. The plan is to use Aristotle as the data repository (emammal.com has an S3 interface). The image transfer speed may increase as much as 2x with the use of Aristotle.
- In the agricultural soil moisture monitoring project, we're working to develop an Aristotle-based notification system that alerts Sedgwick personnel when irrigation is necessary for the grape vineyard and the amount of water to use.
- In a new project, we're attempting to use data to better understand how drought conditions are affecting California Life Oaks and other slow-growing trees, especially during the seedling stage. We plan to use Aristotle for back-end processing of data from new moisture sensors.

**Closing -** *Amy Walton, NSF program manager; Adam Brazier, science team lead*
Amy concluded by saying she likes the meeting format and that sharing progress and technologies used (such as Docker) between use case teams is beneficial. Adam will distribute the meeting minutes and then iterate with the Science Use Case leads to establish goals for the next quarter so that his team can prioritize their support efforts.



#1541215