**Big Data Facility to Enable Data-Driven Science**

Shane Canon, Prabhat, Katie Antypas, Jeff Broughton, Sudip Dosanjh
NERSC, Lawrence Berkeley National Lab.

*NERSC's role in Big Data:* NERSC has a proven track record of running a production quality national user facility for over 40 years. We provide high performance computing resources and user support to thousands of scientists spanning hundreds of scientific projects. Our users range from Nobel Prize winners to distinguished scientists to emerging early career researchers. For more than a decade we have engaged with big data communities in High Energy Physics, Climate and Genomics in supporting their data-centric workloads. We import experimental and observational datasets from telescopes, light sources, particle physics detectors and genome sequencers, and support complex data workflows that handle data ingest, pre-processing, analytics and visualization phases, resulting in scientific insight.

*Suggestions for NITRD Vision and Priorities Document:* We share the broad goals outlined in the NITRD document. We would like to emphasize that sustained investment in production facilities, reusable and cross-domain tools and infrastructure, data representation and storage mechanisms will be critical for long-term impact in the Big Data space.

*Proposed High Impact Ideas*: The Department of Energy is a pioneer in building and operating national users facilities and applying advanced computing to enable scientific discovery. This is illustrated by its premier computing facilities such as NERSC. To date, these facilities have been primarily focused on enabling large-scale modeling and simulation and access has been generally limited to projects linked to the DOE mission. With the increasing importance of data in enabling scientific discovery and driving scientific leadership, the nation requires a comprehensive and coordinated initiative to expand these national capabilities to address data-intensive scientific challenges. We envision this initiative as an integrated, multi-agency effort that couples cutting edge cyber infrastructure, next generation software services, and intensive training and consulting to enable scientists to address problems of national importance and maintain leadership in science.

*Facilities*: Scientists are increasingly limited in the problems that they can tackle due to limits in storage, analysis resources, and networking connectivity. The available resources are often mismatched both in design and scale to the challenges at hand. We envision a limited set of federated resources, potentially multi-agency in design, that are architected and optimized for the needs of the scientific community. Access to these resources will remove many of the barriers that exists today. Centralizing these resources will provide both efficiencies in operation, access to unprecedented scale, and enable a level of data integration that can not be achieved to date.

*Services*: A set of of scalable, API-driven software services built on open standards is required to enable the hardware, ensure open-access, and foster collaboration across agencies. Investments are needed at all levels of the Data stack: from workflow and batch systems to co-ordinate data movement and job execution, to runtime analytics frameworks (e.g. BDAS, Hadoop), to user-facing, productive libraries (e.g. MLBase, GraphX). Finally, the ability to upload, host and share data nationally, and internationally will be critical; continued investment in portal and gateway technologies is required.

*Expertise and Training*: To enable scientists to effectively utilize these capabilities, they require increased access to experts and training. These training efforts should integrate with existing academic programs and leverage emerging online learning. Existing resource operators already posses significant skill in both data science and domain science and must be a part of the strategy. This includes developing and disseminating training material, but also providing more advanced consulting for critical challenges.

*Recommendations on fostering new cross-domain and cross-sector partnerships:* NITRD has done a commendable job of organizing various inter-agency working groups (Big Data, LSN, JET, MAGIC). Such working groups are a critical component in terms of gathering strategic input from a variety of perspectives and coordinating activities. We would recommend arranging in-depth town-hall meetings, or requirement gathering workshops which conduct deep-dive sessions into leading Big Data science use cases from various agencies, and explicitly target common research and production infrastructure related issues. Currently, there is significant redundancy in our collective approach towards exploring the Big Data space, and NITRD can play a leading role in proposing a common vision.

*Relevance of NERSC to NITRD strategic plan:* NERSC has over 40 years of demonstrated production experience in the High Performance Computing arena. Over 5000 users utilize NERSC facilities to produce breakthrough science results each year. While our primary scientific engagements stem from the Department of Energy, recent collaborations with with NIH and NSF have positioned us to respond well to a broader class of science use cases and inter-agency collaborations. We believe strongly that we have the right people, skills and infrastructure to both provide input to NITRD strategic plan, as well as execute a Big Data vision at the national scale.