

Reference ID: 11227586049_Abernathey

Reference ID: 11227586049_Abernathey

Submission Date and Time: 12/16/2019 9:54:21 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Consent Statement: "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

Consent answer: I consent to NSF's use and display of the submitted information.

Author Names & Affiliations

Submitting author: Ryan Abernathey - Columbia University

Additional authors: Joseph Hamman, National Center for Atmospheric Research; Deepak Cherian, National Center for Atmospheric Research; Matthew Rocklin, NVIDIA; Ryan May, University Corporation for Atmospheric Research; Aneesh Subramanian, University of Colorado Boulder;

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

Oceanography Climate Science Meteorology Neuroscience Parallel computing

Title of Response

Give Me All Your Data: CI Needs for Interactive Petascale Scientific Data Analytics

Abstract

We argue that, across scientific domains, scientists have a growing need to interactively process, analyze, and visualize datasets of petabyte size and beyond. Rather than extracting only small subsets,

many of the most exciting data-driven involve analyzing the WHOLE record. Example applications include calculating statistics across simulation ensembles, analyzing global patterns of spatiotemporal variability, and training machine learning models. Crucially, we contend that scientific discoveries are most often made through interactive, “human-in-the-loop” analysis. NSF’s CI priorities should reflect this fact. Since these datasets are often shared among thousands of researchers, we encourage the development of cloud-style data-proximate computing facilities which can bring together communities of researchers with their petascale datasets, improving reproducibility along the way. These facilities could leverage commercial cloud platforms or, with some cultural changes, be built around existing HPC systems. Scientists also need software that can accelerate interactive data analysis with parallel computing without drastic changes to scientists’ workflows. The open source scientific python ecosystem, including tools such as Jupyter Lab and the Dask parallel computing library, offers a promising and sustainable foundation for building the next generation of data analytics environments

Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

In diverse scientific fields--from astronomy to geoscience to neuroscience--community efforts have produced valuable datasets whose size is measured in Petabytes (PB). These datasets arise from both new observing technologies and also from simulations, which play an increasingly vital role in many fields. For example, in astronomy the Large Synoptic Scale Survey will produce 1 PB of imagery per year. In Climate Science, the Coupled Model Intercomparison Project phase 6 (CMIP6) of the World Climate Research Program will eventually comprise over 20 PB, with the next generation increasing by an order of magnitude. In neuroscience, the NIH’s BRAIN initiative will generate 5 - 20 TB of data per week. A common thread between these examples is that a single dataset is shared among a community of thousands of researchers and is anticipated to support thousands of publications. This emergent cross-disciplinary pattern should influence how NSF develops CI for Harnessing the Data Revolution. Most scientific data download portals assume that a scientist is searching for just a small piece of these large datasets. However, we contend that the most exciting and ambitious applications require the whole dataset, or a large subset, rather than a small piece. Examples include calculating statistics across simulation ensembles, analyzing global patterns of spatiotemporal variability, and training machine learning models. Perhaps the most challenging nature of these applications is their diversity; while traditional simulation applies known algorithms repetitively, in data analysis, the only limit to the nature of the calculation should be creativity and not the computational infrastructure. In such examples, scientists need to refine their analysis methodologies by iterating their calculations interactively, with the “human in the loop,” at scale. This is how new discoveries are made with large, complex datasets.

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).

Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

Our existing CI is woefully inadequate for this type of workload. The most common pattern is for researchers to download large data volumes to local data servers, employ graduate students to organize and make the data “analysis-ready,” and then apply ad-hoc strategies for data processing. This effort is often duplicated within each research group who use the data, often with multiple copies even at the same institution. This CI is often technically inadequate to enable petabyte-scale analytics. Without infrastructure to support interactive data analysis at scale, scientists’ creativity is stifled, leading them to look for safe, expected results. This approach is also extremely cost ineffective--NSF often funds both the data provider, local storage servers, and grad student / postdoc toil. Interactive data-driven computing has been greatly enhanced by Jupyter notebooks and related technologies for example; however, Jupyter notebooks only address the user interface. Computational tools (software) and infrastructure are needed to enable interactive data exploration and analysis at scale. On the hardware side, the need can be summarized as “bringing the compute to the data.” Rather than encouraging downloads (the current status quo), data facilities can collocate high-performance storage of petascale datasets together with massively scalable computing. This is already happening in various forms at many HPC facilities. However, there is some tension between the HPC centers’ desire to achieve full system utilization and the intermittent nature of interactive data analysis. Moreover, the potential audience for data analytics is many times larger than that for traditional HPC computing, and HPC centers’ security restrictions and quotas limit the possibility of broad-based access to their platforms. This hinders collaboration and limits the value of datasets stored there. NSF could encourage their HPC centers to prioritize interactive data analytics as a first-class HPC application. An alternative computing paradigm is found in the commercial cloud, which trades the homogeneous architecture of HPC for high flexibility, on-demand computing, massive scale, and high aggregate throughput from object storage. In our experience, commercial cloud architecture is an ideal platform for interactive data analytics at scale. It allows researchers to scale-out their data processing on-demand to large compute clusters for short time periods. It also gives scientists the ability to completely customize their analysis environment and tools. It facilitates collaboration by providing a global namespace for data and code sharing. Projects such as Jupyter’s Binder illustrates the potential of cloud computing to transform scientific reproducibility. NSF should find better ways to support cloud-style data analytics, either by partnering with commercial cloud providers or by supporting the development of cloud-style data and computing facilities and federations perhaps at existing HPC centers. The costs of storing a PB of data in commercial cloud (~\$250K per year) are currently too high for individual research groups to bear, but they might be

worthwhile at an aggregate scale, particularly coupled with effective cloud-computing infrastructure. Scientists also need software tools for data analytics at scale. Traditional scientific analysis software doesn't scale beyond a single machine. On the other hand, established "big data" tools used widely in industry (e.g. Hadoop, Spark) are often hard to adapt to scientific analysis, in which the datasets are often more complex and higher-dimensional than business data. We believe great promise lies in parallel computing frameworks such as the Python library Dask, which provides a familiar community-standard API for general-purpose numerical data analysis, but can distribute computations over heterogeneous computing clusters under the hood. Such tools make it possible for domain scientists to tackle petabyte-scale datasets without major changes to their analysis code or their data.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

The development of sustainable CI for data-intensive computing should seek to leverage open standards and open-source software in building a common foundation for applications across different domains. In our experience with the Pangeo project, we have learned that the basic CI needs for big-data astronomy, geoscience, econometrics, and neuroscience are not fundamentally different. All require mass storage of large arrays and dataframes, scalable compute, and interactive computing. The differences come at the last mile, in building high-level domain-specific toolkits. We believe the open-source scientific Python software ecosystem already contains many of the necessary building blocks for future CI. NSF should seek to nourish the foundations of this ecosystem while encouraging higher-level development on top of it.

-- End Submission --