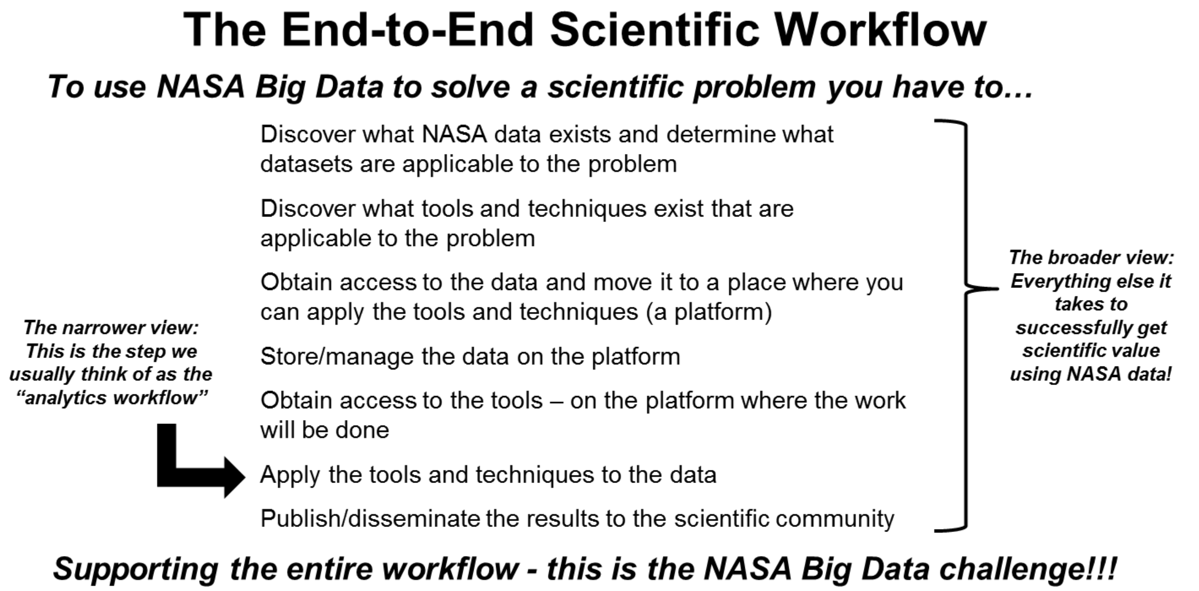
Next Steps in Supporting Big Data Analytics at the NASA Advanced Supercomputing (NAS) Division

Piyush Mehrotra, L. Harper Pryor
piyush.mehrotra@nasa.gov, harper.pryor@nasa.gov
NASA Advanced Supercomputing (NAS) Division, NASA Ames, Moffett Field CA 94035

At the NASA Advanced Supercomputing (NAS) Division we have been focusing on evaluating the requirements of NASA's big data users and implementing an infrastructure to meet these requirements. Last year in Fukuoka we had reported on some of our preliminary efforts highlighting the challenges posed by the complexity of our large scientific datasets, in particular NASA's vast repository of data from NASA's earth and space sensing satellites. In this paper, we describe our efforts to understand and support scientific workflows from a broader perspective. We also present some enhancements made to evolve the NASA Earth Exchange (NEX – our collaborative platform that brings together big data and big compute for the Earth science domain) to provide the infrastructure for big data at NAS. At the end, we will give some examples of the successes our users have had in applying the NEX big data infrastructure to important scientific problems.

Workflows and workflow tools generally address the work done within a specific computing environment – like accessing local datasets, setting up execution, running analyses, and then displaying the results. As we talked to users about how they do their work, it became clear that this is too narrow a view of the problem. A lot happens before a scientist gets to the point of executing this kind of workflow. So based on the interviews, we developed a use case for a true end-to-end scientific workflow, starting when the scientist poses a question all the way to publishing a result. The following figure shows the steps in this use case (workflow) and where the narrower workflow fits in.



The End-to-End Scientific Workflow

To use NASA Big Data to solve a scientific problem you have to…

Discover what NASA data exists and determine what datasets are applicable to the problem

Discover what tools and techniques exist that are applicable to the problem

Obtain access to the data and move it to a place where you can apply the tools and techniques (a platform)

Store/manage the data on the platform

Obtain access to the tools – on the platform where the work will be done

Apply the tools and techniques to the data

Publish/disseminate the results to the scientific community

The narrower view: This is the step we usually think of as the "analytics workflow"

The broader view: Everything else it takes to successfully get scientific value using NASA data!

Supporting the entire workflow - this is the NASA Big Data challenge!!!

Since our goal is to help scientists gain value from NASA datasets, to succeed we have to address this entire end-to-end workflow. Accordingly, we are developing a "User's Guide" that will consolidate in one place what a user needs to know at every step of the workflow. In doing this, we will be guided by the insights we have gained from our interviews.

We have been utilizing the NASA Earth Exchange (NEX – hosted at NAS http://nex.nasa.gov) to develop and implement some of these ideas. NEX, a platform for the Earth science community that provides a unique mechanism for scientific cooperation and knowledge sharing, comprises three parts: a web portal for collaboration, extensive high-resolution datasets and compute resources including a sandbox for prototyping models and algorithms along with HPC resources at NAS for large scale data analysis. In the last year, we have made several enhancements to the NEX architecture including: implementing an innovative data re-exporter that allows for sharing (on a read-only basis) of the dataset repository between systems operating at different security levels and implementing protocols to open the sandbox to Earth science users without NASA credentials in order to widen the accessibility of the data and compute resources.

One key insight from our user interviews is that we have to be careful not to assume too much about what potential users know about the datasets. As "insiders," it is easy to forget that a lot of what we take for granted is knowledge we have gained through many years working with this data. We have to find a way to bring this knowledge to the newcomers. Over the past couple years, NASA has been developing a semantic "front end" for searching earth science datasets as one tool to help users find applicable data starting with higher level queries and less initial knowledge about the datasets. A prototype of the system, called the Ontology-Driven Interactive Search Environment for Earth Sciences (ODISEES), is running at one of NAS's sister sites. The ODISEES ontology is suitable for describing - with high-precision - the data sets, instruments, missions, observations and model outputs in the atmospheric, land and ocean domains. As part of a recently funded effort, we at NAS will be deploying ODISEES within the infrastructure of the NASA Earth Exchange. We will extend the ontology to address our land remote sensing datasets. Then we will work with users to observe how they explore datasets using this new capability and provide feedback to the ODISEES developers so they can enhance the system.

Our HPC environment comprises a mix of heterogeneous resources including: Pleiades, our flagship distributed memory system; Endeavour, with two shared memory nodes (2TB and 4TB respectively); GPGPU nodes; and, our high-end visualization engine - the hyperwall. We are evaluating the use of these resources for data analytics including potentially adding a Hadoop cluster to the environment. We are also embarking on an evaluation project to enhance the Pleiades I/O infrastructure by adding local fast storage to the nodes for use as cache buffers or a local distributed file system.

The NEX and HPC resources are being used to address several petabyte scale big data "grand challenge" problems in Earth science. One example, the Web-enabled Landsat Data (WELD) project, jointly supported by NASA and the U.S. Geological Survey (USGS), is sifting the huge Landsat data record to select cloud-free pixels, creating unprecedented clear view composite Landsat mosaics of the conterminous U.S. and Alaska. Results will be used to document anthropogenic and natural changes over local to global extents. Another grand challenge solution enabled by NEX is the North American Forest Dynamics (NAFD) project, a collaborative effort among University of Maryland, Oregon State University, Goddard Space Flight Center and Forest Service scientists that is creating forest disturbance maps using the Landsat record since 1985, which are used to gain a spatial understanding of how forests are affected by harvest, fire, and other factors such as insects and disease. NAFD scientists recently used the NAS hyperwall to interact with animated datasets and reported that this immersive visualization experience gave them unprecedented insight into changes in forests made visible by these new maps.

The measures we have taken so far are only a beginning. We will continue to find ways to address the challenges users face in every step of the end-to-end scientific workflow, and we look forward to reporting on more of these next year.