NSF Award 1261721
J. Ray Scott, PI
Nick Nystrom, Co-PI
Ralph Roskies, Co-PI

The Data Exacell

INTRODUCTION:

The Data Exacell (DXC) is a development pilot project to create, deploy, and test software and hardware building blocks that support data-analytic capabilities for data intensive scientific research. Building blocks created during this project include a hardware resource for supporting big data projects, the SLASH2 file system, which is designed for wide-area operation, a production resource to support XSEDE allocations oriented towards big data, and collaboration with big data projects by supporting them on the DXC hardware resources. The DXC hosted a workshop, Best Practices in Data Infrastructure, in early 2016, bringing together leaders from other data projects, including DIBBs and DataNet awardees. A report of the outcomes of this workshop is available.

A full description of the DXC project is available on the PSC website at: https://www.psc.edu/index.php/research/data-handling-analytics/data-exacell

MOST SIGNIFICANT CHALLENGE

It has been a significant challenge to identify research projects in diverse fields of science to work with us while at the same time researching and developing new big data tools for these projects. The researchers expect stability. Deployments of new tools often come with instability. This is particularly true when one of the major components is a file system. It is further complicated by attempting to deploy the file system across geographic and administrative domains.

During the project and as a part of the workshop, we have identified specific needs of big data researchers. These needs include the preservation of existing tools as well as the adoption of new tools, administrative and geographic domain independence, and data security, version consistency, discovery, and publishing.

THE SOLUTION

We used the projects to provide scientific and technological drivers and system validation for our approach. We assigned user support staff to each project and worked closely with them. We put a premium on system integrity monitoring and usage statistics gathering. We worked very closely with the systems administration staffs of the project sites to help them with issues related to large storage and networking.

REMAINING CHALLENGES

Providing these advanced data services to a new user community remains a challenge. We are beginning to look at ways to sustain the effort after this award ends.

FUTURE DIRECTIONS

We continue to seek big data analytics projects that will work with us and help inform our direction. We are approaching science gateway developers and facilitators such at the Science Gateways Community Institute, to incorporate our technology and practices into their material. We have recently demonstrated an integration of SLASH2 and the Galaxy workflow system. We plan to integrate SLASH2 with an existing advanced metadata management system. Other directions include more support for federated identity mapping, integration with cluster tools such as OpenHPC and OpenStack, and more support for virtual machine and containerized tool distributions.