

Dealing with TB-sized datasets

Introduction

This session will explore issues surrounding distribution and efficient access to datasets that are 100GB to TBs scale – datasets too big for ScienceBase. How does a USGS modeler publish these large datasets? How can USGS researchers and colleagues gain effective and efficient access? Could we move to a new paradigm of storing data on a private or public cloud, and doing our analysis and visualization close to the data, using tools such as Jupyterhub, where your browser is the client?

- This engineering, while part of "Scientific Computing," requires an expertise that most project-level scientists do not have (or want to have).
- Even if one does know how to do this, this requires substantial infrastructure, which is also likely well out of reach of projects and even our current institutional/enterprise systems (like ScienceBase)
- Can we develop guidelines, tools, and human resource for this kind of data management to help the project-level scientist deal with this?
- Might need to come up w/recommendations or ideas to submit to folks like the SB development or HPC teams about future in-house technology and about leveraging infrastructure from outside the USGS or Federal government. Might also need to establish protocols to use the infrastructure options we currently have as well as possible.
- Probably need to build on existing scientific data life cycle management ideas. The organization and functionality associated with published archives w/things like live-access web services are likely very different than the data models used by the project scientist during the data development phase of the life cycle.

Possible Events

- maybe HPC person talks for a bit
- followed by a panel, where members do some lightening talks
- and then settle down to an open discussion w/crowd?

Possible outcomes

- identify folks w/interest
- begin to identify scale of likely data products and types of access and services around them (polling science data producers)
- begin to develop BMPs for physical organization of data content