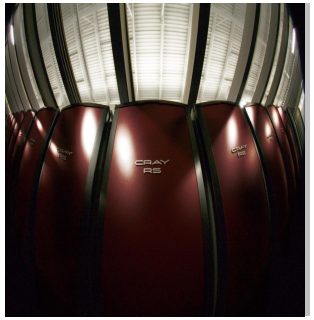

Metadata driven data management

Jay Lofstead

gflofst@sandia.gov

Data-Centric Compute BoF ISC 2019

Sandia National Laboratories

Exceptional service in the national interest

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Problem Space

Sandia National Laboratories

1. Long duration projects with LOTS of different kinds of files and data sets

2. What to do when people leave or retire?

3. How do we select data for machine learning regression?

Initial Work

Sandia National Laboratories

- EMPRESS focuses on rich, custom metadata for simulation runs
  - Optimized storage with rich query capabilities
  - Limited to particularly structured sims and data sets
- Labs looking at the archive issue
  - TagIt from ORNL
  - GUFI from LANL
  - ECMWF has data formatting and archive system fully integrated into their workflow
  - SNL, LLNL, and NERSC both have other efforts underway
- HPSS offers archive management, but it is simplistic
  - Extended attributes do not offer the complexity nor do they necessarily migrate properly

Desired Operations

Sandia National Laboratories

- Run regression on all data from SPPARKS am_weld simulations. What data is from that?
- What are the observational data sets related to this simulation configuration?
- Where is the final project report for project ‘B61’?
- Where are the data files used to generate the analysis in the project report for project ‘B61’?

Approach

Sandia National Laboratories

- EMPRESS is a fixed schema so it is not the right starting point
  - Embedded database is a good idea, but the current schema is not sufficient.
- Collecting requirements from various project groups
- Collecting requirements at a BoF at SC19 on usable archives
- Investigating how users want to interact with the system
- Determining how HPSS will work with the system

Benefits

Sandia National Laboratories

- Archives frequently write once, read never—this can make the data less opaque
- Opens stored data to long term data trending and analysis
- Meets organizational and regulatory requirements for reproducibility