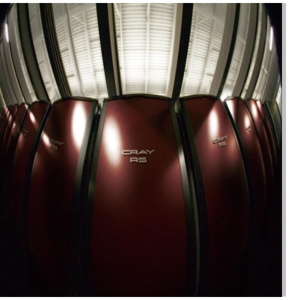
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2023-02747C



Metadata Management to Aid Data Discovery

Jay Lofstead

Sandia National Laboratories

Exceptional service in the national interest

NSDF All Hands

12 April 2023

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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

- Long term data archiving
  - Which data set(s) contain what I want to study?
- Short/medium term data set identification
  - What does each set contain and which ones to save?
- Provenance largely focuses on environment
  - Critical, but insufficient/inefficient

Three current generations of tools with a fourth being developed

First Generation Tools

- File-level metadata tagging
- POSIX extended attributes
- HPSS
- Starfish (https://starfishstorage.com/)
- JAMO – Joint Genome Institute Archive and Metadata Organizer

Second Generation Tools

- Raw data indexing
  - Exact value or binning
- FastBit
  - Does a value in the bin range exist in this file/dataset?
- SciDB
  - Multi-dimensional array data model with rich query facilities
- IO libraries
  - HDF5, ADIOS, NetCDF attribute capabilities

Third Generation Tools

- Feature tagging
  - Region/var/run
  - Bounding box, simple tag

- Key-value based
  - SoMeta - encode tag information into key and use value for data location (focused on object stores)
  - TagIt – distributed, shared nothing storage integrated for faster data searching

Third Generation Tools

- RDMBS-based
  - EMPRESS – independent database(s) of tags related to run, timestep, or var; flexible query interface into data directly via logical locations
  - BIMM – image database with tags

Fourth Generation Tool

- Derived quantity information-based tagging
  - For a climate model, where is the pressure gradient greater than a particular value

- Problems!
  - Derived quantities take time to compute
  - Derived quantities can be as large as the original data

Thank you

- gflofst@sandia.gov