Data Management Survey 2012

Data Management Survey 2012 - DRAFT

As you may know, Scientific Computing at SLAC has recently been reorganized. Scientific Computing Applications (SCA), a Division in PPA, is now under the leadership of Head of SLAC Scientific Computing, Amber Boehnlein, with a focus on lab wide scientific computing. As part of this, the newly created SCA Data Management Department is conducting a survey of Data Management at SLAC to better understand the current situation and future needs at the entire laboratory. Your contributions to the survey would be valuable and we would appreciate if we could meet at a time of your convenience.

The definition we use for Data Management:

Data Management is any content neutral interaction with the data and includes Data Storage/Archival, Access, Distribution and Curation. It includes technical solutions, procedures and policies and deals with the full life cycle of the data.

covers many very different aspects - from nuts and bolts of moving bits to data distribution policies. To make a meeting more productive we have prepared a list of potential questions. This list is not exhaustive and we would of course appreciate any additional input you may have. If there are presentations and documents already answering some of the technical questions, please feel free to forward them to us.

General questions

- What experiment/group are you responding for?
- Are there any science topics which are currently limited by your ability to manage data?
- Are there any legacy constraints on your data management?
- What resource constraints do you have? Is lack of qualified manpower an issue?
- What are your main concerns about data management?
- Do you foresee any future limitations on your ability to manage data?
- Are there any places you see where SLAC and SCA can contribute (more)?

Detailed questions

Data Storage

- Data volume:
  - What is the data volume per year?
  - What is the total data volume you have to manage?
  - Does the total data volume have to be easily accessible or is part of it archived?
  - Is the data rate constant or bursty?
  - Is your data all at SLAC or are there any mirrored copies offsite?

- Storage technology:
  - Is it all disk based storage? Do you also use SSD? Tape? Some combination of these?
  - Do you need high performant storage hardware?
  - Do you use a layered storage system i.e. high performant SSD in front of disks etc?
  - What fraction of the data is disk resident vs on tape?
  - Is tape only used for backups?
  - What are your uptime requirements?

- Data Life cycle Management:
  - Do you have a well defined and complete data life cycle management system?
    - "DLM products automate the processes involved, typically organizing data into separate tiers according to specified policies, and automating data migration from one tier to another based on those criteria. As a rule, newer data, and data that must be accessed more frequently, is stored on faster, but more expensive storage media, while less critical data is stored on cheaper, but slower media."
  - Do you have (or would like to have) automatic migration of older files from disk to tape?
  - Do you have (or would like to have) transparent access to data that are on tape only?

- Resource monitoring:
  - Do you use central SLAC systems for resource monitoring (Nagios, Ganglia)?
  - If not, have you developed your own monitoring?

- Data streams:
  - Is on-the-fly data reduction possible?
  - Do you have single or multiple (independent) data streams - like from different (simultaneous) experiments?

- File systems:
  - Which file systems do you use for storing data? NFS? xrootd? Gluster? Lustre? Anything else?
  - If you use multiple file systems, what fraction of the data is stored in each of them?
  - Are different file systems used for different type of files?
  - Do you need or would like to have a distributed file system?
  - Is POSIX compliance an absolute requirement?
  - Do you currently have performance issues when reading/writing data to disk?
  - Do you use quotas? If so, at what granularity (user, group, experiment)?

- File sizes:
  - What is the typical file size(s)?
  - Do you have small configuration/meta-data files and large data files?

    - How many size categories?

- File numbers:
    - Do you have a large number of files? Is the total number of files a problem?
    - How many files of each size category?
    - Do you use a database to track file locations/provenance? Something else than a database?

- File formats:
    - What file formats do you use? Root? HDF5? XTC? Fits? Other?
    - Is the data format/structure complex?
    - What tools to do you use to analyze the data? Root, Matlab, homegrown, experiment specific?

- Databases:
    - Do you use databases? If so, for what purpose? Storing data? Meta-data? Configuration information etc?
    - Do you use Oracle? MySQL? Other?
    - Is the database administration sufficient? Do you use local expertise or Computing Department resources (db-admin)?

- Security requirements:
    - What are your security requirements for the data?
    - At what granularity do you need security and permission rights? Individual user? Group? Experiment?
    - Do you need ACLs?

- Centralized storage:
    - SLAC is currently working on a centralized storage solution. Do you know about this initiative?
    - Is this something you would consider using? If so, for part of your storage or all your storage?
    - Would you like to get out of the storage business altogether and let SLAC deal with it?
    - Or is local group expertise needed to fulfill your specific storage needs?

Data Access

- How do the users find data? What meta-data do you have?
- Do you have homogenous or heterogenous data sets?
- If the latter, does it influence the type of data (storage and) access you provide?
- Do users only access their own data set, multiple data sets or everything? Are there permission issues?
- Are users systematically always going through their entire dataset every time? Are there "hot" data?
- Are there any latency requirements? Is most analysis interactive? Or batch?
- What is the access pattern?
    - Do the users copy over all the raw data to their home institution and do all analysis there?
    - Do users skim (reduce) the data and then copy it to their home institution?
    - What is involved in the data reduction? Can it be defined beforehand or does it involve extensive analysis of the entire data set (at SLAC)?
    - Do the users keep the data at SLAC and do (all) their analysis at SLAC?
    - When users analyse their data at SLAC, how often do they access their data set?
    - How do you deal with "hot" data sets?
- Is the current data access model limiting the science users can do?

Data Distribution

- How are data distributed? Is distribution web based? Script based?
- Is there any public data distribution? Or just each group copying over their own data set?
- How much data are distributed (downloaded) per year?
- What security needs are required for data distribution?
- Do you need to persist data provenance and ownership?
- How do the users find the data they want?
- Can you technically share data between groups if you need to?
- Is the current data distribution model limiting the science users can do? Does it limit cooperation between different groups?

Data Curation

- Do you need to provide long term data access?
- Do you expect to have to provide public (non-SLAC) access to some or all the data sets in the future?
- Have you looked into how to tag data sets for the long term - for example with Digital Object Identifiers?

Data Management Automation

- How much of your data management is automated?
- Do you have an automated data processing system?
- The Data Management department has an Applications group that provides applications for data processing, and (web based) data access and monitoring. Would you be interested in talking to them?