Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

SAM For Users

Robert Illingworth

FIFE Workshop

21 June 2016

FABRIC FOR FRONTIER EXPERIMENTS

What is SAM For Users?

- Utilities to assist individual users to make use of the SAM catalogue for their own data
  - User data tends to be relatively unstructured compared to the more organized production style data
- It's still useful to make use of a proper catalogue
  - Make use of existing tools and monitoring for SAM jobs
  - Simplify moving the files between different storage locations
- Assorted sanity checks
  - Won't remove only copies of the files (unless explicitly told to)

Command information

- All commands live in the fife_utils package
- Long names starting with ‘`sam_`’ so you can tab-complete and know what they do

- Some commands make assumptions about pnfs directory structure; this must be set up before use (one time per experiment), for example

```
/pnfs/$EXPERIMENT/archive/sam_managed_users/$USER/data/a/1/2/file
```

File ownership/security

- The metadata in SAM belongs to the user who declares it
  - This user can modify the metadata, add or remove locations, or retire the metadata
  - Other users can't change anything (except admins)
- The files in the storage system will be owned by the mapped credentials used to copy them
  - For dCache this means an individual user
  - For offsite locations this may vary; quite likely to be shared ownership

File metadata

- Unlike production files, user files often have limited metadata
- Basic assumption – all you need are unique filenames, and it can help by renaming existing files with unique names
- The metadata also contains a tag name so you can retrieve a whole group of files
- If you need more detail, you can add your own metadata for each file, using all the normal SAM features

Simple example

You have 1000 files in /nova/data/users/$USER/stuff

you want to run reconstruction on them.

So you can make a dataset, put it in dcache scratch, and start the jobs.

```
sam_add_dataset --name=${USER}_stuff --directory=/nova/data/user/$USER/stuff
sam_copy2scratch --name=${USER}_stuff
submit... --dataset=${USER}_stuff
sam_remove_location_dataset --name=${USER}_stuff --dest=/pnfs/$EXPERIMENT/scratch
```

Pengfei will give a more detailed example later

Simple example (2)

Then you might decide you want to archive the files on tape and add some other metadata to them

```
sam_move2archive_datset --name=${USER}_stuff
```

```
sam_modify_dataset_metadata --name=${USER}_stuff --metadata=file.json
```

Dataset commands

- `sam_add_dataset`
  - file list or directory
  - renames files to be unique
  - makes a dataset via a tag
- `sam_revert_names`
  - puts uniqified-file names back
- `sam_modify_dataset_metadata`
  - bulk-add/change metadata on a dataset of files
- `sam_validate_dataset`
  - are all the files still there?
  - good for scratch
- `sam_retire_dataset`
  - clean out files and/or dataset declaration

Dataset copy and move

- `sam_clone_dataset`
  - copy files to given location
- `sam_move_dataset`
  - copy files to given location, remove source after sam_archive_dataset
- `sam_move2archive_dataset`
  - copy files to default archive (tape) location, remove others
- `sam_copy2scratch_dataset`
  - copy files to default scratch location
- `sam_move2persistent_dataset`
  - copy files to default scratch location, remove others

Delete datasets

- `sam_unclone_dataset`
- `sam_remove_location_dataset`
  - get rid of copies of dataset files at a given place (refuses to remove last copy of a file..)

- `sam_retire_dataset`
  - clean out files and/or dataset declaration

Miscellaneous commands

- These commands are for archiving a directory tree into SAM/enstore – useful for unstructured data you might want to keep. They are ***not*** suitable for doing regular backups
- `sam_archive_directory_image`
  - make tarballs of places, place in archive loc, declare to SAM
- `sam_restore_directory_image`
  - find tarballs by metadata, unpack

- `sam_prestage_dataset`
  - Ensure a dataset is prestaged from tape to disk. (This is just a wrapper for the equivalent samweb command.)

Getting started

- See Pengfei's upcoming tutorial
- The commands may already be available in your experiment's environment
  - If not, "`setup fife_utils`" may work

- Ideally each experiment that uses this will be able to develop local expertise that can act as the first source of advice for new users
  - Until then, feel free to ask us for advice

- Report problems and bugs through the servicedesk (it has its own entry in the service catalog)

Summary

- SAM For Users assists with managing analysis datasets
  - Assist with dataset movement/archiving
  - No need for complex metadata (unless you want to)

- Once added to the catalogue use your favourite tools for making and analyzing datasets, just like with production data