Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

SAM For Users Tutorial

Pengfei Ding

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

- Advantages of using SAM for Users toolkit:
  - users' own data will be just like production data,
    - submitting grid jobs using SAM project,
    - making use of existing tools and monitoring for SAM jobs;
  - moving files between different storage locations are made simple;
  - only copies of files will not be removed (unless explicitly told to)

List of available tools in SAM for Users toolkit

- Dataset commands:
  - sam_add_dataset
  - sam_revert_names
  - sam_modify_dataset_metadata
  - sam_validate_dataset
- Dataset copy and move:
  - sam_clone_dataset
  - sam_move_dataset
  - sam_move2archive_dataset
  - sam_copy2scratch_dataset
  - sam_move2persistent_dataset
- Delete datasets:
  - sam_unclone_dataset
  - sam_remove_location_dataset
  - sam_retire_dataset
- Miscellaneous commands:
  - sam_archive_dataset
  - sam_archive_directory_image
  - sam_restore_directory_image
  - sam_prestage_dataset
  - sam_audit_dataset
  - sam_condense_dataset
  - sam_pin_dataset

* Examples can be found in this tutorial

Setup requirements

- On your experiments' GPVMs, where /grid/fermiapp or cvmfs is mounted:

```
# ups setup
source /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
# get valid proxy for sam
kx509
kxlist -p
voms-proxy-init --rfc --voms=fermilab:/fermilab/nova/Role=Analysis --noregen
# setup fife_utils
setup fife_utils
# set some environment variables for nova
export GROUP=nova
export EXPERIMENT=nova
export SAM_EXPERIMENT=nova
export IFDH_BASE_URI=http://samweb.fnal.gov:8480/sam/nova/api
# for minerva
#export IFDH_BASE_URI="http://samweb-minerva.fnal.gov:20004/sam/minerva/api"
```

Current default version of fife_utils is v3_0_1

```
novagpvm11: ~ $ echo $SETUP_FIFE_UTILS
fife_utils v3_0_1 -f NULL -z /cvmfs/fermilab.opensciencegrid.org/products/common/db
```

Add dataset to SAM (I)

- `sam_add_dataset --name=${USER}_stuff --directory=/nova/data/user/$USER/stuff`
- This example: `sam_add_dataset -n dingpf_tutorial_20160621  -d /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntuples`

```
novagpvm11: ~ $
novagpvm11: ~ $ ll /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntuples
total 22928
-rw-rw-r-- 1 dingpf nova 5396608 Jun 21 09:45 neardet_r00011358_s01_nuexsec.root
-rw-rw-r-- 1 dingpf nova 6424415 Jun 21 09:44 neardet_r00011361_s08_nuexsec.root
-rw-rw-r-- 1 dingpf nova 6082761 Jun 21 09:46 neardet_r00011382_s15_nuexsec.root
-rw-rw-r-- 1 dingpf nova 5572893 Jun 21 09:46 neardet_r00011382_s16_nuexsec.root
novagpvm11: ~ $
novagpvm11: ~ $
novagpvm11: ~ $ sam_add_dataset -n dingpf_tutorial_20160621  -d /pnfs/nova/scratch/users/dingpf/sam4use
_tutorial_fife2016/ntuples
renamed neardet_r00011382_s16_nuexsec.root to 19153c3f-fd03-492a-a688-d8d66e541ae0-neardet_r00011382_s1
nuexsec.root
renamed neardet_r00011382_s15_nuexsec.root to 1c6dcd22-1fcc-4bbd-a59e-77a2ae4ac4d1-neardet_r00011382_s1
nuexsec.root
renamed neardet_r00011358_s01_nuexsec.root to 62202739-3f63-4260-b7cc-cb5156dd640b-neardet_r00011358_s0
nuexsec.root
renamed neardet_r00011361_s08_nuexsec.root to ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neardet_r00011361_s0
nuexsec.root
the dataset with name dingpf_tutorial_20160621 was created
novagpvm11: ~ $
novagpvm11: ~ $
novagpvm11: ~ $
```

Add dataset to SAM (II)

- Note that, files in the directory have been renamed with uuid:
  - renamed neardet_r00011382_s16_nuexsec.root to 19153c3f-fd03-492a-a688-d8d66e541ae0-neardet_r00011382_s16_nuexsec.root
- Also frequently used exmaples:
  - `sam_addataset -n dingpf_tutorial_20160621 -f sam4users_tutorial_files.list`

```
novagpvm11: ~ $ ls /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntuples
19153c3f-fd03-492a-a688-d8d66e541ae0-neardet_r00011382_s16_nuexsec.root
1c6dcd22-1fcc-4bbd-a59e-77a2ae4ac4d1-neardet_r00011382_s15_nuexsec.root
62202739-3f63-4260-b7cc-cb5156dd640b-neardet_r00011358_s01_nuexsec.root
ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neardet_r00011361_s08_nuexsec.root
```

Files got renamed!

Add dataset to SAM (III)

- Dataset description:
  - `samweb describe-definition dingpf_tutorial_20160621`

```
novagpvm11: ~ $
novagpvm11: ~ $ samweb describe-definition dingpf_tutorial_20160621
Definition Name: dingpf_tutorial_20160621
  Definition Id: 523513
  Creation Date: 2016-06-21T14:47:38.747907+00:00
       Username: dingpf
          Group: nova
     Dimensions: dataset.tag dingpf_tutorial_20160621
```

Dataset has very simple metadata.

Add dataset to SAM (IV)

- Files in the dataset:
  - samweb list-definition-files dingpf_tutorial_20160621

```
novagpvm11: ~ $ samweb list-definition-files dingpf_tutorial_20160621
ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neardet_r00011361_s08_nuexsec.root
62202739-3f63-4260-b7cc-cb5156dd640b-neardet_r00011358_s01_nuexsec.root
1c6dcd22-1fcc-4bbd-a59e-77a2ae4ac4d1-neardet_r00011382_s15_nuexsec.root
19153c3f-fd03-492a-a688-d8d66e541ae0-neardet_r00011382_s16_nuexsec.root
```

  - samweb locate-file

```
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9fc5ab
t
dcache:/pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/r
```

Modify file metadata (I)

- File metadata:
  - `samweb get-metadata 43ccc572-d856-4413-8f41-535fd66755bf-neardet_r00011382_s15_nuexsec.root`

```
     File Name: 43ccc572-d856-4413-8f41-535fd66755bf-neardet_r00011382_s15_nuexsec.root
       File Id: 212505656
   Create Date: 2016-06-21T16:12:22+00:00
          User: dingpf
     File Type: unknown
   File Format: unknown
     File Size: 6082761
      Checksum: (none)
Content Status: good
   Dataset.Tag: dingpf_test_20160621
```

Suggestion for experiments' SAM admins:

- add metadata parameters for users' own data;
- ask users to only modify metadata for those parameters.

```
novagpvm11: ~ $ samweb add-parameter Users.AnalysisGroup string
```

Modify file metadata (II)

- Modify file metadata:
  - `samweb modify-metadata ${FILE_NAME} ${METADATA_JSON_FILE}`

```
novagpvm11: ~ $ cat ~/file.json
{
 "Users.AnalysisGroup": "NDXsec",
 "file_format": "root"
}
novagpvm11: ~ $
```

```
novagpvm11: ~ $ samweb get-metadata 3f2346bb-a886-415a-9515-74fd3b463c57-neardet_r00011361_s08_nuexsec.root
          File Name: 3f2346bb-a886-415a-9515-74fd3b463c57-neardet_r00011361_s08_nuexsec.root
            File Id: 212503465
        Create Date: 2016-06-21T16:12:23+00:00
               User: dingpf
        Update Date: 2016-06-21T16:33:18+00:00
        Update User: dingpf
          File Type: unknown
        File Format: root
          File Size: 6424415
           Checksum: (none)
     Content Status: good
        Dataset.Tag: dingpf_test_20160621
Users.AnalysisGroup: NDXsec
```

Modify file metadata (III)

- Modify file metadata using SAM python API:
  - recommended for large number of files

```
import samweb_client as swc
sam = swc.SAMWebClient("nova")
mfmetadata = {"Users.AnalysisGroup": "NDXsec"}
f="3f2346bb-a886-415a-9515-74fd3b463c57-neardet_r00011361_s08_nuexsec.root"
sam.modifyFileMetadata(f,mfmetadata)
```

```
novagpvm11: ~ $ samweb list-files "Users.AnalysisGroup NDXsec"
3f2346bb-a886-415a-9515-74fd3b463c57-neardet_r00011361_s08_nuexsec.root
```

Files can now be queried with the user defined metadata fields.

Move/copy dataset (I) – sam_clone_dataset

- `sam_clone_dataset -n dingpf_tutorial_20160621 -d /pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fife2016/`

```
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9
t
dcache:/pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fi
dcache:/pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2
```

Files now have two locations.

Move/copy dataset (II) - sam_unclone_dataset

- `sam_unclone_dataset -n dingpf_tutorial_20160621 -d /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntuples`
- Similar to `sam_remove_location_dataset`

```
novagpvm11: ~ $ sam_remove_location_dataset -n dingpf_tutorial_20160621 -d /pnfs/n
f/sam4users_tutorial_fife2016/ntuples/ -j
I would 'ifdh rm /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntupl
2a-a688-d8d66e541ae0-neardet_r00011382_s16_nuexsec.root'
```

Use '-j' option to list what the command is going to do, without actually performing the action.

```
novagpvm11: ~ $ sam_unclone_dataset -n dingpf_tutorial_20160621 -d /pnfs/nova
ers_tutorial_fife2016/ntuples/
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neard
t
dcache:/pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fife2016/6/f
novagpvm11: ~ $
novagpvm11: ~ $ ls /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife201
```

Note: this deletes the files (except the only copy).

Fermilab

Revert file names (I)

- `sam_revert_names -d /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016/ntuples/f/c/`

```
novagpvm11: ~ $ sam_revert_names -d /pnfs/nova/scratch/users/dingpf/sam
/c/
renamed 1c6dcd22-1fcc-4bbd-a59e-77a2ae4ac4d1-neardet_r00011382_s15_nuex
nuexsec.root
1 files renamed.
finished
```

Revert files names in a specified directory.

```
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9fc5
t
dcache:/pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fife2016
dcache:/pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fife2
```

Files names are reverted, but SAM still have the previous file location.

Revert file names (II)

- Remove missing locations:

```
sam_validate_names -n dingpf_tutorial_20160621 --prune
```

```
missing: /pnfs/nova/scratch/users/dingpf/sam4users_tutorial_fi
d8d66e541ae0-neardet_r00011382_s16_nuexsec.root
-- location removed
```

"sam_validate_name --prune" removes missing locations.

```
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb9
t
dcache:/pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fife2016/
```

Only one location is left for the file and it is available.

Fermilab

Retire dataset (I)

- `sam_retire_dataset -n dingpf_tutorial_20160621`
- Retire the dataset, remove file locations, but keep one copy of the files

```
novagpvm11: ~ $ sam_retire_dataset -n dingpf_tutorial_20160621
novagpvm11: ~ $
```

```
novagpvm11: ~ $ samweb locate-file ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb9
t
File 'ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neardet_r00011361_s08_nuexs
```

Files are removed from SAM.

```
novagpvm11: ~ $ ls /pnfs/nova/persistent/users/dingpf/sam4users_tutorial_fife
ccc7f4dd-7a4b-44c4-aa78-42b9fc5abb91-neardet_r00011361_s08_nuexsec.root
```

Last copy of the files are not deleted.

Summary

- With SAM for Users toolkit, one can:
  - Add own files to SAM
  - Copy/move dataset files between different storage locations
  - No accidents of deleting files

  - Most importantly: various tools for using production data are now available to users' own data.

- More details:
  - https://cdcvs.fnal.gov/redmine/projects/fife_utils/wiki/UserGuide
  - https://cdcvs.fnal.gov/redmine/projects/nova_sam/wiki/User_Datasets

Backup

Submit grid jobs using SAM dataset and project

- ifdh startProject
- jobsub_submit with shell script containing:
  - ifdh findProject
  - ifdh establishProcess
  - ifdh getNextFile

- https://cdcvs.fnal.gov/redmine/projects/ifdhc/wiki/Wiki
- https://cdcvs.fnal.gov/redmine/projects/ifdhc/repository/entry/demo.sh

```
novagpvm11: ~ $ launch_clone_jobs --help
launch_clone_jobs --group=GRP --jobs=njobs --name=dataset --dest=url
   slice dataset and run multiple sam_clone_dataset instances
   in grid jobs
```

Command information

- All commands live in the file_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
```

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.)