Sequential Access with Metadata (SAM)

Tammy Walton

2021 Intensity Frontier Computing Summer School

June 18, 2021

Outline

- What is SAM?
- Using samweb
- Using SAM-for-Users

Fermilab

Sequential Access with Metadata (SAM)

- SAM is a data handling system that manages the storing and retrieve of files by accessing the file metadata
  - Metadata is a collection of information pertaining to the history, quality, and characteristic of the file.
    - Example of metadata
      - date, experiment, location, condition, data type, run type, etc.
  - The data are maintained in a database

- Resources
  - https://cdcvs.fnal.gov/redmine/projects/sam/wiki
  - https://cdcvs.fnal.gov/redmine/projects/sam/wiki/User_Guide_for_SAM
  - https://cdcvs.fnal.gov/redmine/projects/nova_sam/wiki
  - https://cdcvs.fnal.gov/redmine/projects/uboonecode/wiki/Sam#Data-Handling-and-SAM

- This is interactive tutorial / overview using SAM
- Interact using zoom chat feature and reaction emoji
- I apologize beforehand. I am only registered in the SAM database for g-2, DUNE, and MINERvA experiments. I do not have exercises for everyone. Hopefully, you can follow along using your experiment data.

Getting Started

1. Login your experiment general purpose virtual machine
   - ssh –XYK twalton@gm2gpvm02.fnal.gov

2. Setup your experiment software and common area
   1. source /cvmfs/gm2.opensciencegrid.org/prod/g-2/setup
   2. source /grid/fermiapp/products/common/etc/setups.sh

3. Setup SAM
   1. setup sam_web_client
   2. export SAM_EXPERIMENT=gm2

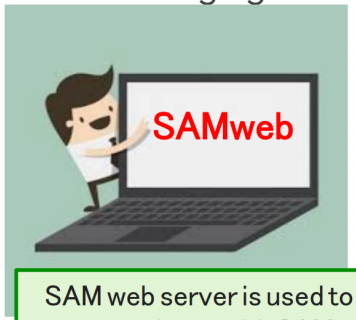
Outline

- What is SAM?
- Using samweb
- Using SAM-for-Users

- We use the database, **S**equential **A**ccess **M**etadata for managing and delivering the metadata.
- Simple diagram of the procedure.

DAQ writes midas file to /data1/gm2/ or /data2/gm2 on g2be2 machine.

FTS extracts the metadata, adds the metadata and file to SAM, and transfers the file to tape, /pnfs/GM2/daq.



SAM web server is used to communicate with SAM. The tool samweb gives everything YOU want to know about the data.

SAM Authentication

- Check you have a valid kerberos ticket on the gpvm machine

```
<gm2gpvm04.fnal.gov> klist
Ticket cache: FILE:/tmp/krb5cc_43472_VMqK41iG3X
Default principal: twalton@FNAL.GOV

Valid starting       Expires              Service principal
06/17/2021 20:40:57  06/18/2021 22:40:21  krbtgt/FNAL.GOV@FNAL.GOV
06/17/2021 20:40:57  06/18/2021 22:40:21  nfs/homesrv01.fnal.gov@FNAL.GOV
06/17/2021 20:40:57  06/18/2021 22:40:21  nfs/filesrv01.fnal.gov@FNAL.GOV
```

- Invalid tickets

```
<gm2gpvm04.fnal.gov> kinit twalton
Password for twalton@FNAL.GOV:
<gm2gpvm04.fnal.gov>
```

- To obtain authentication, a kx509 certificate is generated to communicate with the samweb server.

```
<gm2gpvm04.fnal.gov> kx509
Checking if /tmp/x509up_u43472 can be reused ... no
Authorizing ...... authorized
Fetching certificate ..... fetched
Storing certificate in /tmp/x509up_u43472
Your certificate is valid until: Thu Jun 24 21:46:26 2021
```


Fermilab

SAM Authentication

- To test your authentication

```
[<gm2gpvm04.fnal.gov> samweb -s server-info
 SAMWeb API for gm2
 SAMWeb version: 3.1.0-pre1-15-g4188d8d
 Connected to: https://samweb.fnal.gov:8483/sam/gm2/api

 Cherrypy version: 18.6.0
 Backend server hostname: samwebgpvm06.fnal.gov
 SQLAlchemy version: 1.3.19
 Database: postgresql+psycopg2://samdbs:***@sampgsdb03.fnal.gov:5437/sam_gm2_prd
 HTTP User-Agent: SAMWebClient/v3_0 (samweb.py) python/2.7.5

 User information:
   Untrusted identity: twalton@gm2gpvm04.fnal.gov
   Authenticated username: twalton
   Roles: dataset, project, metadata
```

The received message for an authorized access

Fermilab

samweb : Help Menu

```
[<gm2gpvm04.fnal.gov> samweb --help
Usage: samweb.py [base options] <command> [command options] ...

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --help-commands       list available commands

  Base options:
    -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server. If not set, defaults to
                        $SAM_EXPERIMENT if set, else $EXPERIMENT.
    --dev               use development server
    -s, --secure        always use secure (SSL) mode
    --cert=CERT         x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
    --key=KEY           x509 key for authentication (defaults to same as
                        certificate)
    --max-timeout=MAX_TIMEOUT
                        set the max timeout in seconds (After this period even
                        retriable errors will fail. This setting will not
                        interrupt an active connection; use socket-timeout for
                        that.)
    --socket-timeout=SOCKET_TIMEOUT
                        set the socket timeout in seconds (max time for data
                        to be sent or received)
    -r ROLE, --role=ROLE
                        specific role to use for authorization
    -z TIMEZONE, --timezone=TIMEZONE
                        set time zone for server responses
    -v, --verbose       Verbose mode
```

samweb : Help Menu

```
<gm2gpvm04.fnal.gov> samweb --help-commands
Available commands:
  Data file commands:
    add-file-location
    count-files
    declare-file
    file-lineage
    get-file-access-url
    get-metadata
    list-file-locations
    list-files
    locate-file
    modify-metadata
    remove-file-location
    retire-file
    validate-metadata
    verify-file-checksum

  Definition commands:
    count-definition-files
    create-definition
    delete-definition
    describe-definition
    list-definition-files
    list-definitions
    modify-definition
    snapshot-info
    take-snapshot

  Project commands:
    find-project
    get-next-file
    list-projects
    prestage-dataset
    project-recovery
    project-summary
    release-file
    run-project
    set-process-file-status
    set-process-status
    start-process
    start-project
    stop-process
    stop-project

  Utility commands:
    file-checksum
    server-info

  Admin commands:
    add-application
    add-data-disk
    add-parameter
    add-user
    add-value
    describe-user
    list-applications
    list-data-disks
    list-parameters
    list-users
    list-values
    modify-user
```

User available commands

Fermilab

SAM : Data Location

- Users do not need to keep track of the file locations or manage the movement of files
- SAM handles the booking and delivering of files
- Files are stored on tape systems:
  - dCache/Enstore at Fermilab
  - dCache/HPSS at BNL
  - CASTOR at CERN
  - GPFS/TSM at INFN

Fermilab

SAM : Data Location

- Given a file and the experiment, SAM provides the location

DUNE example

```
<gm2gpvm04.fnal.gov> samweb locate-file np04_raw_run001088_18_dl1.root -e dune
castor:/neutplatform/protodune/rawdata/np04/detector/None/raw/02/51/18/36
enstore:/pnfs/dune/tape_backed/dunepro/protodune/np04/detector/None/raw/02/79/62/23
rucio:protodune-sp
```

MINERvA example

```
<gm2gpvm04.fnal.gov> samweb locate-file -e minerva MV_00019167_0025_numib_v09_1602221314_RecoData_v21r1p1.root
enstore:/pnfs/minerva/data_reconstructed/reconstructed-pool/numibeam/v21r1p1/00/01/91/67(1144@vpk532)
<gm2gpvm04.fnal.gov>
```

g-2 example

```
<dunegpvm05.fnal.gov> samweb locate-file gm2dqc_ana_45712282_32313.00197.root -e gm2
dcache:/pnfs/GM2/scratch/users/gm2pro/preprod/InFillDQCWMetaData/5211P/Run3E-v9_64/runs_32000/32313
<dunegpvm05.fnal.gov> samweb locate-file gm2_run43552_00121.mid -e gm2
enstore:/pnfs/GM2/daq/run4/raw/runs_43000/43552(2093@fl8914l8)
```


Fermilab

SAM : Understanding the File via Metadata

- A file must have metadata
- Metadata allows the file to be cataloged and queried
- Metadata includes characteristics about the files and processing, configuration, and physics conditions about the data
- Metadata should be small and manageable
- It important to keep track of the file history. SAM stores the parent file and how it was derived.

SAM : Understanding the Files via Metadata

```
<gm2gpvm04.fnal.gov> samweb list-values --help_categories -e dune
Available database categories:

data_tiers
    Data tier of a file (raw, simulated, reconstructed, etc)
checksum_types
    Checksum type
run_types
    Type of a run
groups
    Work groups
data_streams
    Data stream names
file_formats
    Format of the file
file_types
    Type (ie general purpose) of a file
```


Fermilab

SAM : Understanding the Files via Metadata

```
<gm2gpvm04.fnal.gov> samweb list-values --help_categories -e dune
Available database categories:

data_tiers
    Data tier of a file (raw, simulated, reconstructed, etc)
checksum_types
    Checksum type
run_types
    Type of a run
groups
    Work groups
data_streams
    Data stream names
file_formats
    Format of the file
file_types
    Type (ie general purpose) of a file
```

```
<gm2gpvm04.fnal.gov> samweb list-values data_tiers -e dune
simulated
raw
hit-reconstructed
full-reconstructed
generated
detector-simulated
reconstructed-2d
reconstructed-3d
sliced
dc1input
dc1output
root-tuple
root-hist
dqm
decoded-raw
sam-user
pandora_info
reco-recalibrated
<gm2gpvm04.fnal.gov> samweb list-values data_tiers -e gm2
raw
calibration
fcl
log
reco
sim
analysis
nmr-test
mdc0
truth
processed
nearline
unpack
hists
sam-user
full
infilldqc
gain_corr
hadd
gaincorr
oof_corr
dqc_hadd
gain_hadd
qdb
subrun_dqc
qdb_dqc
```


Fermilab

SAM : Understanding the Files via Metadata

```
<gm2gpvm04.fnal.gov> samweb list-values --help_categories -e dune
Available database categories:

data_tiers
    Data tier of a file (raw, simulated, reconstructed, etc)
checksum_types
    Checksum type
run_types
    Type of a run
groups
    Work groups
data_streams
    Data stream names
file_formats
    Format of the file
file_types
    Type (ie general purpose) of a file
```

```
<gm2gpvm04.fnal.gov> samweb list-values file_formats -e minerva
unknown
dspack
root
ethereal
tar
gzipped-tar
run-1-sta
evpack
compressed evpack
root.tar
pool
pool.tar
binary
binary.tar
<gm2gpvm04.fnal.gov> samweb list-values file_formats -e gm2
unknown
art-root
root
text
midas
artroot
json
```


Fermilab

SAM : Understanding the Files via Metadata

```
<gm2gpvm04.fnal.gov> samweb list-values --help_categories -e dune
Available database categories:

data_tiers
    Data tier of a file (raw, simulated, reconstructed, etc)
checksum_types
    Checksum type
run_types
    Type of a run
groups
    Work groups
data_streams
    Data stream names
file_formats
    Format of the file
file_types
    Type (ie general purpose) of a file
```

```
<gm2gpvm04.fnal.gov> samweb list-values file_types -e minerva
unknown
nonPhysicsGeneric
physicsGeneric
importedDetector
derivedDetector
importedSimulated
derivedSimulated
<gm2gpvm04.fnal.gov> samweb list-values file_types -e dune
mc
test-data
unknown
fastmc
physics-data
calib-data
dc1simdaq
detector
binary
photon_detector
```


Fermilab

SAM : Understanding the File via Metadata

```
<gm2gpvm04.fnal.gov> samweb get-metadata -e minerva MV_00019167_0025_numib_v09_1602221314_RecoData_v21r1p1.root
                File Name: MV_00019167_0025_numib_v09_1602221314_RecoData_v21r1p1.root
                  File Id: 27739580
             Create Date: 2018-01-08T18:35:34+00:00
                    User: minervapro
                File Size: 806776976
                 Checksum: enstore:636545548
           Content Status: good
                File Type: importedDetector
             File Format: root
                   Group: minerva
               Data Tier: reconstructed-pool
             Application: reconstructed reco v21r1p1
             Event Count: 700
             First Event: 97574476
              Last Event: 97575175
              Start Time: 2018-01-05T19:03:52+00:00
                End Time: 2018-01-05T19:03:52+00:00
             Data Stream: numibeam
      Offline.executable: SystemTestsApp.exe
  Offline.executablepath: /minerva/app/users/minervapro/cmtuser/Minerva_v21r1p1/Tools/SystemTests/x86_64-slc6-gcc49-opt
        Offline.machine: minervapro-2404969-0-fnpc5013.fnal.gov
        Offline.options: Reco_fCalDigits_MV_00019167_0025_numib_v09_1602221314_v21r1p1.opts
     Offline.optionspath: /storage/local/data1/condor/execute/dir_21123/no_xfer/0/OPTS
           Offline.user: fermigli
    Online.triggerconfig: hvon.hwcfg
      Online.triggertype: numibeam
        Quality.MINERvA: good
    RawDigits.executable: SystemTestsApp.exe
RawDigits.executablepath: /minerva/app/users/minervadat/cmtuser/Minerva_testing/Tools/SystemTests/x86_64-slc6-gcc44-opt
       RawDigits.machine: fnpc3055.fnal.gov
       RawDigits.options: MV_00019167_0025_numib_v09_1602221314_RawDigits_raw24_notknown.root.opts
   RawDigits.optionspath: /minerva/data2/rawdata/minerva/scripts/numib/00/01/91/67
         RawDigits.user: fermigli
    SupDigits.executable: SystemTestsApp.exe
SupDigits.executablepath: /minerva/app/users/minervadat/cmtuser/Minerva_testing/Tools/SystemTests/x86_64-slc6-gcc44-opt
       SupDigits.machine: fnpc4117.fnal.gov
       SupDigits.options: MV_00019167_0025_numib_v09_1602221314_SupDigits_raw24_sup24_numibeam.root.opts
   SupDigits.optionspath: /pnfs/minerva/data_processing_rawdigits/rawdigits/numibeam/24/00/01/91/67
         SupDigits.user: fermigli
                   Runs: 19167.0025 (minerva)
         File Partition: 1
        Lum Block Ranges: 97574476-97575175
                 Parents: MV_00019167_0025_numib_v09_1602221314_IDODDigits_v21r1p1.root
```

SAM specific

Experimental specific

Provenance

SAM : Understanding the File via Metadata

- Use SAM webserver to find your experiment metadata parameters.
  - https://samweb.fnal.gov:8483/sam/nova/api/files/list/dimensions
  - https://samweb.fnal.gov:8483/sam/gm2/api/files/list/dimensions
  - https://samweb.fnal.gov:8483/sam/uboone/api/files/list/dimensions
  - https://samweb.fnal.gov:8483/sam/minerva/api/files/list/dimensions
  - https://samweb.fnal.gov:8483/sam/mu2e/api/files/list/dimensions
  - https://samweb.fnal.gov:8483/sam/<experiment>/api/files/list/dimensions
- Use the metadata to create datasets, a collection of files

Fermilab

SAM : Creating Datasets

- Before creating a dataset, use "samweb list-files" to provide a summary of files in a potential dataset

```
<dunegpvm05.fnal.gov> samweb list-files -h
Usage: samweb.py [base options] list-files [command options] <dimensions query>

List files by dimensions query

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --help-commands       list available commands

  Base options:
    -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server. If not set, defaults to
                        $SAM_EXPERIMENT if set, else $EXPERIMENT.
    --dev               use development server
    -s, --secure        always use secure (SSL) mode
    --cert=CERT         x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
    --key=KEY           x509 key for authentication (defaults to same as
                        certificate)
    --max-timeout=MAX_TIMEOUT
                        set the max timeout in seconds (After this period even
                        retriable errors will fail. This setting will not
                        interrupt an active connection; use socket-timeout for
                        that.)
    --socket-timeout=SOCKET_TIMEOUT
                        set the socket timeout in seconds (max time for data
                        to be sent or received)
    -r ROLE, --role=ROLE
                        specific role to use for authorization
    -z TIMEZONE, --timezone=TIMEZONE
                        set time zone for server responses
    -v, --verbose       Verbose mode

  list-files options:
    --fileinfo          Return additional information for each file
    --summary           Return a summary of the results instead of the full
                        list
    --expand-query      Return an expanded and normalized version of the query
    --help-dimensions   Return information on the available dimensions
    --dump-query        Return query information for these dimensions instead
                        of evaluating them
```


Fermilab

SAM : Creating Datasets

- Before creating a dataset, use “samweb list-files” to provide a summary of files in a potential dataset

```
<dunegpvm05.fnal.gov> samweb list-files -e gm2 "user gm2pro and file_format artroot and offline.requestid 5205P and run_number >= 31368 and run_number <= 32035" --summary
File count:     100850
Total size:     104491740150658
Event count:    21331213
<dunegpvm05.fnal.gov> samweb list-files -e gm2 "user gm2pro and file_format artroot and offline.requestid 5205P and run_number >= 31368 and run_number <= 32035" --dump-query
Parse tree for:

user gm2pro and file_format artroot and offline.requestid 5205P and run_number >= 31368 and run_number <= 32035

AndNode
  Dimension(user = gm2pro)
  Dimension(file_format = artroot)
  Dimension(offline.requestid = 5205P)
  Dimension(run_number >= 31368)
  Dimension(run_number <= 32035)


Normalized query:

user gm2pro and file_format artroot and offline.requestid 5205P and run_number >= 31368 and run_number <= 32035


PostgreSQL query:

SELECT DISTINCT ON (data_files.file_id) data_files.file_id, data_files.file_name
FROM data_files JOIN persons AS create_user ON data_files.create_user_id = create_user.person_id JOIN file_formats ON file_formats.file_format_id = data_files.file_format_id JOIN data_files_param_values AS dfpv_offline_requestid ON dat
a_files.file_id = dfpv_offline_requestid.file_id JOIN data_files_runs ON data_files.file_id = data_files_runs.file_id JOIN runs ON runs.run_id = data_files_runs.run_id
WHERE create_user.username = 'gm2pro' AND file_formats.file_format = 'artroot' AND dfpv_offline_requestid.param_value_id = 16410519 AND runs.run_number >= 31368 AND runs.run_number <= 32035 AND data_files.retired_date IS NULL AND data_
files.file_content_status_id = 1 AND (EXISTS (SELECT 1
FROM data_file_locations
WHERE data_file_locations.file_id = data_files.file_id))
```


Fermilab

SAM : Creating Datasets

- Creating datasets
  - samweb create-definition gm2pro_daq_preproduction_Run3E_5211P "user gm2pro and file_format artroot and offline.requestid 5211P and run_number >= 32126 and run_number <= 32319"

```
[gm2pro@gm2gpvm01:~]> samweb count-definition-files gm2pro_daq_preproduction_Run3E_5211P
32673
```

  - samweb create-definition gm2pro_daq_preproduction_Run3E_5211P_qdb "defname:gm2pro_daq_preproduction_Run3E_5211P minus isparentof:(data_tier qdb)"

- Getting the dataset description

```
[gm2pro@gm2gpvm01:~]>
[gm2pro@gm2gpvm01:~]> samweb describe-definition gm2pro_daq_preproduction_Run3E_5211P
Definition Name: gm2pro_daq_preproduction_Run3E_5211P
  Definition Id: 362145
  Creation Date: 2021-06-15T17:14:48+00:00
       Username: gm2pro
          Group: gm2
     Dimensions: user gm2pro and file_format artroot and offline.requestid 5211P and run_number >= 32126 and run_number <= 32319
[gm2pro@gm2gpvm01:~]>
```

SAM : Creating Datasets

- SAM datasets are dynamic
- Frozen datasets can be created by taking a snapshot
  - samweb take-snapshot <dataset name>
    - A snapshot id is returned

```
[gm2pro@gm2gpvm01:~]>
[gm2pro@gm2gpvm01:~]> samweb snapshot-info --snapshot-id=295365
         Snapshot Id: 295365
     Definition Name: gm2pro_daq_offline_dqc_run2C_5122A
       Definition Id: 336925
    Snapshot Version: 2
                User: gm2pro
               Group: gm2
         Create Date: 2021-05-10T17:01:02+00:00
Last Evaluation Date: 2021-06-17T16:01:40+00:00
[gm2pro@gm2gpvm01:~]> samweb list-files "snapshot_id=295365" --summary
File count:     69303
Total size:     184905216325138
Event count:    16691406
[gm2pro@gm2gpvm01:~]>
```


Fermilab

SAM : Creating Datasets

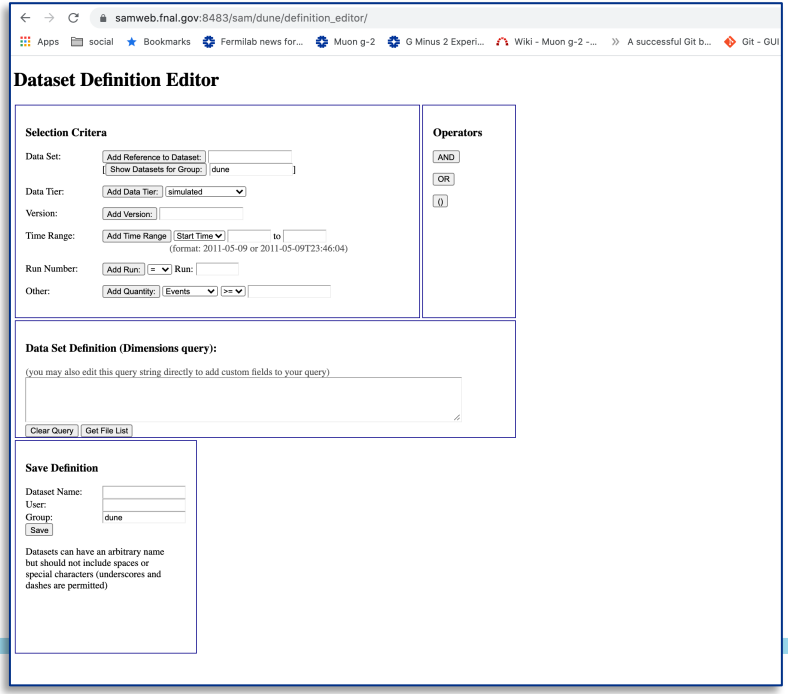
- Information on the dimension syntax
  - https://cdcvs.fnal.gov/redmine/projects/sam-web/wiki/Dimension_Syntax

- Additional query operators
  - intersect
  - isancestorof / isdescendantof

Fermilab

SAM : Creating Datasets

- Dimension editor
  - https://samweb.fnal.gov:8483/sam/dune/definition_editor/



SAM : Prestaging

- Recall that files are located on tape
- For efficient access, files should be copied from tape to dCache disk – prestaging
- If dCache is not overloaded, untouched files have a dCache lifetime for about 30 days
- Your experiment should have a script to check if the dataset is prestaged
  - Muon g-2 adopted NOvA's script

Fermilab

SAM : Prestaging

```
<gm2gpvm04.fnal.gov>
<gm2gpvm04.fnal.gov> ./samlocate.sh -p gm2pro_daq_raw_run3_PreProd_F
Authorizing ...... authorized
Fetching certificate ..... fetched
Storing certificate in /tmp/x509up_u43472
Your certificate is valid until: Fri Jun 25 01:46:37 2021
Your identity: /DC=org/DC=cilogon/C=US/O=Fermi National Accelerator Laboratory/OU=People/CN=Tammy Walton/CN=UID:twalton
Contacting  voms2.fnal.gov:15001 [/DC=org/DC=incommon/C=US/ST=Illinois/L=Batavia/O=Fermi Research Alliance/OU=Fermilab/CN=voms2.fnal.gov] "fermilab" Done
Creating proxy ............................................................ Done

Your proxy is valid until Fri Jun 18 13:46:38 2021

Fri Jun 18 01:46:38 CDT 2021

Locating files from dataset(s): gm2pro_daq_raw_run3_PreProd_F ...

Progess [============================================================] 100%: 16911/16911 done and 0 to go...


Fri Jun 18 01:55:16 CDT 2021

Processing 16911 files in total...

Progess [============================================================] 100%: 16911/16911 done and 0 to go...

Finished checking prestage status for 16911 files, 16911 (~100%) are staged.
Fri Jun 18 01:59:22 CDT 2021

<gm2gpvm04.fnal.gov>
```

SAM : Prestaging

```
<gm2gpvm04.fnal.gov>
<gm2gpvm04.fnal.gov>
<gm2gpvm04.fnal.gov> ./samlocate.sh -p gm2pro_daq_preproduction_Run3N_5200P

Fri Jun 18 02:03:13 CDT 2021

Locating files from dataset(s): gm2pro_daq_preproduction_Run3N_5200P ...

Progess [============================================================] 100%: 80618/80618 done and 0 to go...


Fri Jun 18 02:44:31 CDT 2021

Processing 80618 files in total...

Progess [============================================================] 100%: 80618/80618 done and 0 to go...

Finished checking prestage status for 80618 files, 6672 (~8%) are staged.
Fri Jun 18 03:12:13 CDT 2021

<gm2gpvm04.fnal.gov>
<gm2gpvm04.fnal.gov>
<gm2gpvm04.fnal.gov>
```


Fermilab

SAM : Prestaging

```
<gm2gpvm04.fnal.gov> samweb prestage-dataset
Exactly one of definition name or snapshot id must be provided

Usage: samweb.py [base options] prestage-dataset [command options]

Prestage a dataset

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --help-commands       list available commands

  Base options:
    -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server. If not set, defaults to
                        $SAM_EXPERIMENT if set, else $EXPERIMENT.
    --dev               use development server
    -s, --secure        always use secure (SSL) mode
    --cert=CERT         x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
    --key=KEY           x509 key for authentication (defaults to same as
                        certificate)
    --max-timeout=MAX_TIMEOUT
                        set the max timeout in seconds (After this period even
                        retriable errors will fail. This setting will not
                        interrupt an active connection; use socket-timeout for
                        that.)
    --socket-timeout=SOCKET_TIMEOUT
                        set the socket timeout in seconds (max time for data
                        to be sent or received)
    -r ROLE, --role=ROLE
                        specific role to use for authorization
    -z TIMEZONE, --timezone=TIMEZONE
                        set time zone for server responses
    -v, --verbose       Verbose mode

  prestage-dataset options:
    --defname=DEFNAME
    --snapshot-id=SNAPSHOT_ID
    --max-files=MAX_FILES
    --station=STATION
    --name=NAME         Project name
    --parallel=PARALLEL
                        Number of parallel processes to run
    --delivery-location=DELIVERY_LOCATION
                        Location to which the files should be delivered
                        (defaults to the same as the node option)
    --node=NODE         The current node name. The default is the local
                        hostname, which is appropriate for most situations
<gm2gpvm04.fnal.gov> samweb prestage-dataset --defname=gm2pro_daq_preproduction_Run3N_5200P
```

- SAM project is created
  - prestage_twalton_gm2pro_daq_preproduction_Run3N_5200P_20210617092829
- SAM projects can be monitored via webserver
- SAM projects
  - Allow you to access one, many, or all files in a dataset
  - Distribute the processing of files to many processes

Fermilab

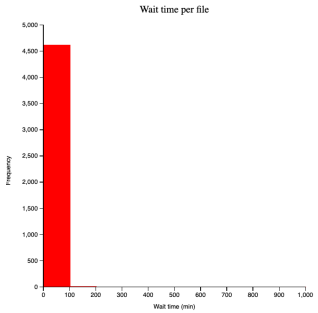
SAM : Monitoring

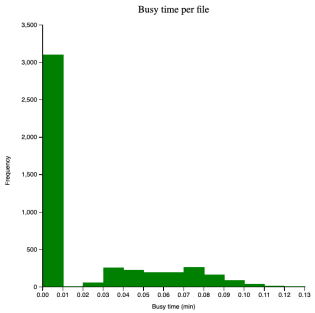
https://samweb.fnal.gov:8483/station_monitor/gm2/stations/gm2/projects/prestage_twalton_gm2pro_daq_preproduction_Run3N_5200P_20210617092829

Project prestage_twalton_gm2pro_daq_preproduction_Run3N_5200P_20210617092829

Generated at 2021-06-18 02:11:12

Project Id	638509
Status	running
Owner	twalton
Start time	2021-06-17 09:28:34
Dataset definition	gm2pro_daq_preproduction_Run3N_5200P
Snapshot Id	280553
Files in snapshot	80618
Files seen	4647
Processes	20
Busy processes	0
Finished processes	0
Waiting processes	20
Error processes	0
Cancelled processes	0
Median wait time (per file)	0s
Median transfer time (per file)	-
Median busy time (per file)	0s
Last activity	gm2preproduction_full_8390501_34338.00024.root file consumed at 2021-06-18 02:07:12





SAM : Monitoring

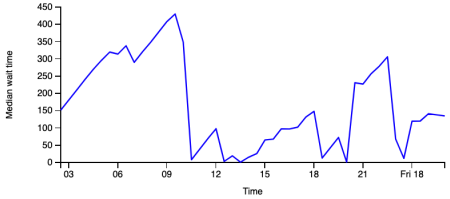
- All SAM projects can be monitored
  - https://samweb.fnal.gov:8483/station_monitor/gm2/stations/gm2/projects

gm2 station

Generated at 2021-06-18 02:13:28

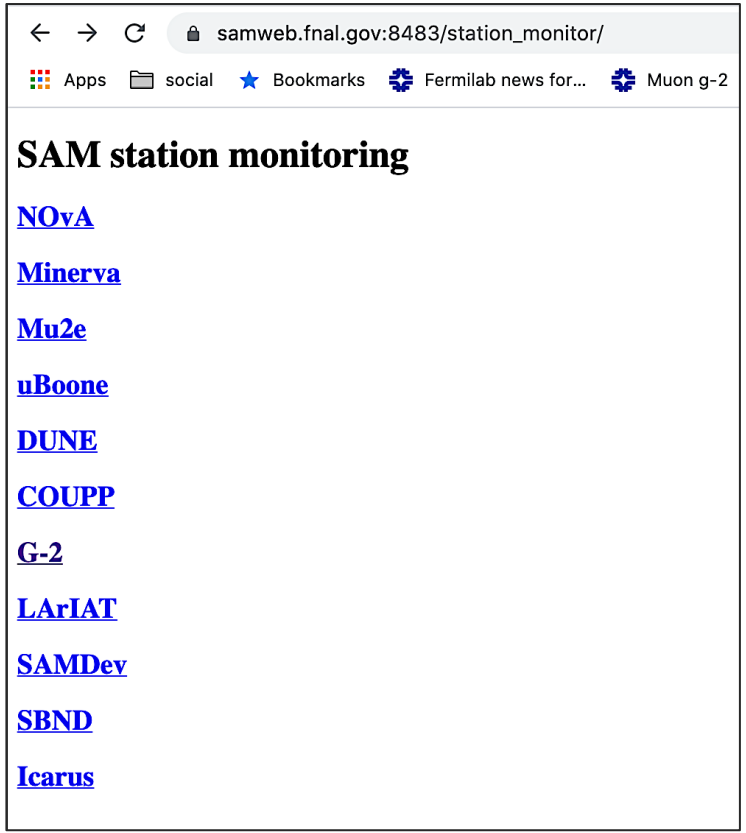
There are 51 running projects

Active processes	4254
Waiting processes	138 (3.24%)



SAM : Monitoring

- Stations
  - https://samweb.fnal.gov:8483/station_monitor/



Outline

- What is SAM?
- Using samweb
- Using SAM-for-Users

What about files not catalogue in SAM?

Fermilab

SAM For Users

- Toolkit that is used for helping analyzers to catalogue their own data
  - Handle non-production data like production data
  - Take advantage of SAM project
  - Use SAM-art (not discuss in this talk)

- Resources
  - https://indico.fnal.gov/event/13787/contributions/21757/attachments/14125/18011/dingpf_sam4users_fife_workshop_20170622.pdf
  - https://cdcvs.fnal.gov/redmine/projects/sam/wiki/SAMLite_Guide

Fermilab

SAM For Users

- List of commands

```
<gm2gpvm04.fnal.gov> sam_
sam_add_dataset              sam_clone_dataset            sam_dataset_stage_status     sam_move2archive_dataset     sam_project_caffeine         sam_revert_names
sam_archive_dataset          sam_condense_dataset         sam_extract_dataset_metadata sam_move2persistent_dataset  sam_remove_location_dataset  sam_unclone_dataset
sam_archive_directory_image  sam_copy2scratch_dataset     sam_metadata_dumper          sam_move_dataset             sam_restore_directory_image  sam_validate_dataset
sam_audit_dataset            sam_dataset_duplicate_kids   sam_modify_dataset_metadata  sam_prestage_dataset         sam_retire_dataset
<gm2gpvm04.fnal.gov> sam_
```

SAM-4-Users available commands are:

```
sam_add_dataset              (makes a new dataset)
sam_validate_dataset         (validates that all the files are present, or which aren't)
sam_audit_dataset            (check all files at destination location to see what files are and are not in named dataset)
sam_clone_dataset            (makes a replica of a dataset in a different location [i.e. copies it])
sam_unclone_dataset          (removes the replicas of a dataset in a specific location [i.e. cleans up a copy])
sam_condense_dataset         (slurp dataset through program and generate summed files)
sam_modify_dataset_metadata  (applies or modifies the metadata associated with a dataset)
sam_retire_dataset           (retires a dataset)

sam_archive_dataset          (copies a dataset to tape backed area, and removes the old copy)
sam_archive_directory_image  (copies a tarfile of a directory to a tape backed area)
sam_copy2scratch_dataset     (copies a dataset to the scratch area)
sam_move2archive_dataset     (alias for sam_archive_dataset)
sam_move2persistent_dataset  (copies a dataset to the persistent area)
sam_move_dataset             (sam_clone_dataset and clean out other locations.)
sam_prestage_dataset         (tells DCache to prestage all the files in a dataset)
sam_project_caffeine         (keeps production SAM projects awake, for when job submission is slow)
sam_remove_location_dataset  (alias for sam_unclone_dataset)
sam_restore_directory_image  (inverse of sam_archive_directory_image)
sam_revert_names             (undoes uniqifying of names from sam_add_dataset)

sam_extract_dataset_metadata (v3_2_8+ runs an extractor program and updates metadata on all dataset files)
sam_dataset_duplicate_kids   (v3_2_8+ look for multiple children of same file)
sam_dataset_stage_status     (v3_2_8+ alias for sam_validate_dataset with staging report turned on)
```

SAM For Users : Creating Datasets

```
[[gm2pro@gm2gpvm04:~]> sam_add_dataset -h
Usage: sam_add_dataset <-f file | -d directory> [options]

Add a group of files to SAM and create a dataset out of it.

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server defaults to $SAM_EXPERIMENT
                        or group name if not set
  -u USER, --user=USER  default is gm2pro
  -t TAG, --tag=TAG     the value for Dataset.Tag which will be used to
                        distinguish this new dataset default format is
                        user+date
  -n NAME, --name=NAME  the dataset name default is userdataset+user+date
  -d DIRECTORY, --directory=DIRECTORY
                        directory of files to create dataset with
  -r, --recurse         walk down all levels of directory
  -f FILE, --file=FILE  file of file paths to create dataset with
  -m METADATA, --metadata=METADATA
                        json file of metadata you would like added to all
                        files
  -s SUBPROCESS, --subprocess=SUBPROCESS
                        execute a child program in a new process to extract
                        metadata only sam_metadata_dumper currently supported
  -c CERT, --cert=CERT  x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
  --no-rename           Do not rename files to ensure uniqueness
  --replace-uuids       Replace existing uuid strings to ensure uniqueness
  --data_tier=DATA_TIER
                        default data_tier if no metadata is given; default
                        'sam-user'
  -v, --verbose         returns verbose output
[[gm2pro@gm2gpvm04:~]>
[gm2pro@gm2gpvm04:~]> sam_add_dataset --no-rename -d . -n gm2pro_qdatabase_Run3E_5211P_hadd2
```


Fermilab

SAM For Users : Creating Datasets

Made a mistake when creating the dataset

```
[[gm2pro@gm2gpvm04:~]> sam_add_dataset -h
Usage: sam_add_dataset <-f file | -d directory> [options]

Add a group of files to SAM and create a dataset out of it.

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server defaults to $SAM_EXPERIMENT
                        or group name if not set
  -u USER, --user=USER  default is gm2pro
  -t TAG, --tag=TAG     the value for Dataset.Tag which will be used to
                        distinguish this new dataset default format is
                        user+date
  -n NAME, --name=NAME  the dataset name default is userdataset+user+date
  -d DIRECTORY, --directory=DIRECTORY
                        directory of files to create dataset with
  -r, --recurse         walk down all levels of directory
  -f FILE, --file=FILE  file of file paths to create dataset with
  -m METADATA, --metadata=METADATA
                        json file of metadata you would like added to all
                        files
  -s SUBPROCESS, --subprocess=SUBPROCESS
                        execute a child program in a new process to extract
                        metadata only sam_metadata_dumper currently supported
  -c CERT, --cert=CERT  x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
  --no-rename           Do not rename files to ensure uniqueness
  --replace-uuids       Replace existing uuid strings to ensure uniqueness
  --data_tier=DATA_TIER
                        default data_tier if no metadata is given; default
                        'sam-user'
  -v, --verbose         returns verbose output
[[gm2pro@gm2gpvm04:~]>
[gm2pro@gm2gpvm04:~]> sam_add_dataset --no-rename -d . -n gm2pro_qdatabase_Run3E_5211P_hadd2
```

```
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]> sam_revert_names -h
Usage: sam_revert_names <-f file | -d directory | glob_pattern> [options]

Reverts file names that contain a uuid.

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  -d DIRECTORY, --directory=DIRECTORY
                        directory of files
  -r, --recurse         walk down all levels of directory
  -f FILE, --file=FILE  file containing file paths to revert
  -p PATTERN, --pattern=PATTERN
                        unix filename pattern like '*.root' to filter by used
                        with -d and -f and -f -
  -n, --nothing         do nothing and print what program would do
  -v, --verbose         returns verbose output
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]> sam_revert_names -d .
```


Fermilab

SAM For Users : Creating Datasets

```
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]> samweb count-definition-files gm2pro_infill_dqc_Run3E_5211P_hadd2
 327
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]> samweb describe-definition gm2pro_infill_dqc_Run3E_5211P_hadd2
 Definition Name: gm2pro_infill_dqc_Run3E_5211P_hadd2
   Definition Id: 362945
   Creation Date: 2021-06-17T19:10:37+00:00
        Username: gm2pro
           Group: gm2
      Dimensions: dataset.tag gm2pro_infill_dqc_Run3E_5211P_hadd2
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]> samweb get-metadata gm2dqc_hadd_21234540_60.root
      File Name: gm2dqc_hadd_21234540_60.root
        File Id: 91593676
    Create Date: 2021-06-17T19:10:20+00:00
           User: gm2pro
      File Size: 971138541
       Checksum: (none)
 Content Status: good
      File Type: unknown
    File Format: unknown
      Data Tier: sam-user
    Dataset.Tag: gm2pro_infill_dqc_Run3E_5211P_hadd2
[[gm2pro@gm2gpvm04:~]>
[[gm2pro@gm2gpvm04:~]>
 [gm2pro@gm2gpvm04:~]>
```

Can use samweb commands

Fermilab

SAM For Users : Creating Datasets

```
[[gm2pro@gm2gpvm04:~]> samweb get-metadata gm2dqc_hadd_21234540_60.root
     File Name: gm2dqc_hadd_21234540_60.root
       File Id: 91593676
   Create Date: 2021-06-17T19:10:20+00:00
          User: gm2pro
     File Size: 971138541
      Checksum: (none)
 Content Status: good
     File Type: unknown
   File Format: unknown
     Data Tier: sam-user
   Dataset.Tag: gm2pro_infill_dqc_Run3E_5211P_hadd2
```

To modify the metadata

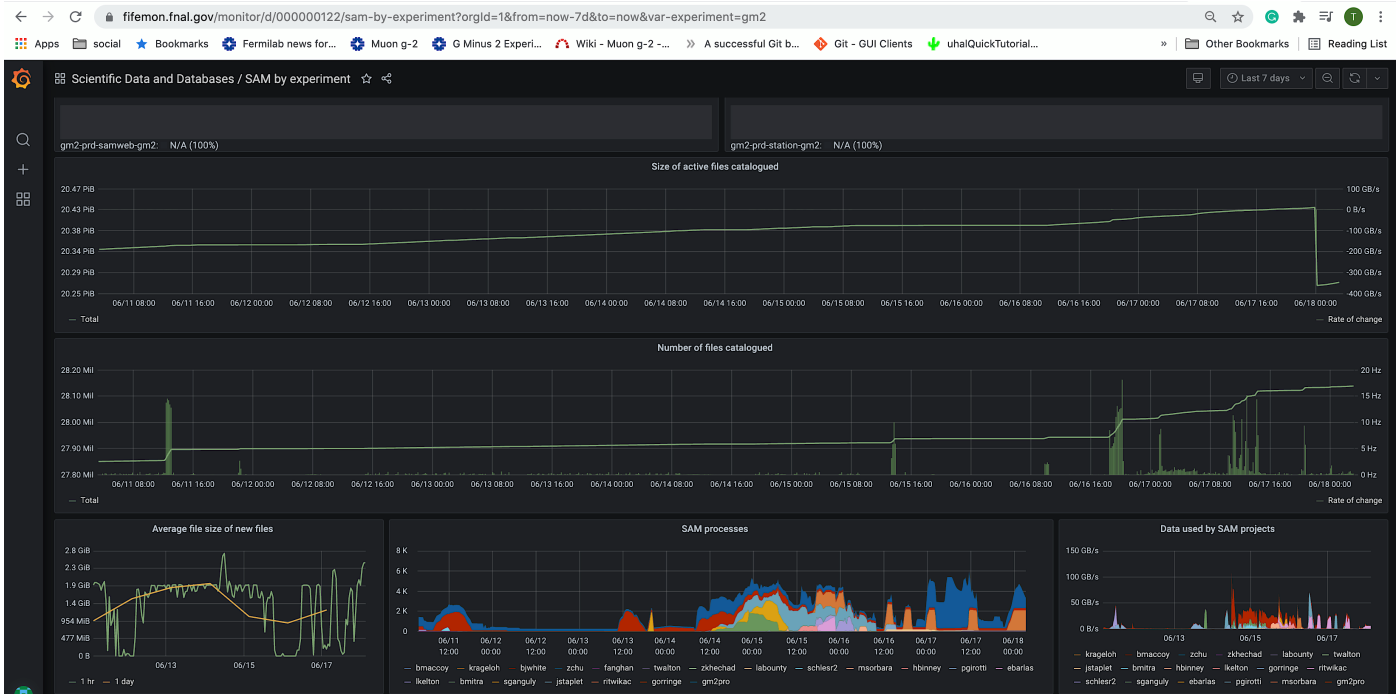
```
[gm2pro@gm2gpvm04:~]> samweb modify-metadata -h
Usage: samweb.py [base options] modify-metadata [command options] [<file name>] <name of file containing metadata parameters to modify (json format)>

Modify metadata for an existing file. If no file name is given as an argument
the metadata file must contain a list of JSON objects containing the file name
and fields to modify If the file name is given the metadata file must be a
single JSON object containing the fields to modify

Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --help-commands       list available commands

  Base options:
    -e EXPERIMENT, --experiment=EXPERIMENT
                        use this experiment server. If not set, defaults to
                        $SAM_EXPERIMENT if set, else $EXPERIMENT.
    --dev               use development server
    -s, --secure        always use secure (SSL) mode
    --cert=CERT         x509 certificate for authentication. If not specified,
                        use $X509_USER_PROXY, $X509_USER_CERT/$X509_USER_KEY
                        or standard grid proxy location
    --key=KEY           x509 key for authentication (defaults to same as
                        certificate)
    --max-timeout=MAX_TIMEOUT
                        set the max timeout in seconds (After this period even
                        retriable errors will fail. This setting will not
                        interrupt an active connection; use socket-timeout for
                        that.)
    --socket-timeout=SOCKET_TIMEOUT
                        set the socket timeout in seconds (max time for data
                        to be sent or received)
    -r ROLE, --role=ROLE
                        specific role to use for authorization
    -z TIMEZONE, --timezone=TIMEZONE
                        set time zone for server responses
    -v, --verbose       Verbose mode

  modify-metadata options:
[gm2pro@gm2gpvm04:~]>
```

Summary

- SAM is a toolkit for data handling
  - Developed by Fermilab
  - Files are managed via metadata
  - Data are stored on tape
  - Data must be prestaged for access
  - Jobs of SAM datasets are executed via SAM projects
  - Webserver for monitoring the SAM projects
  - Fermilab provides a lite SAM system
    - Needed for non-production data

- Additional information is provided throughout the talk

Fermilab

Summary

- https://fifemon.fnal.gov/monitor/d/000000122/sam-by-experiment?orgId=1&from=now-24h&to=now&var-experiment=gm2