

SAM

Adam Lyon

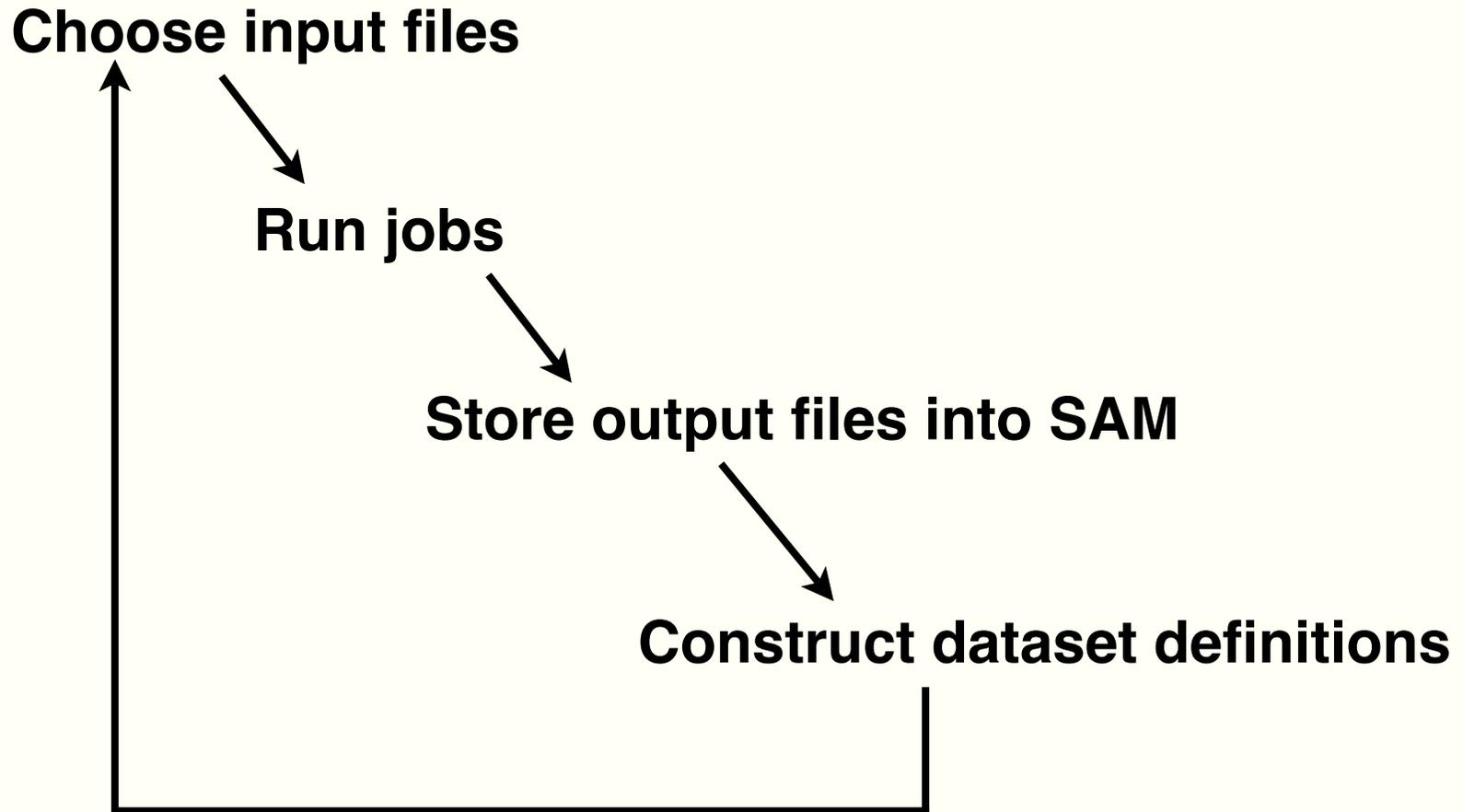
SAMGrid Project Leader – CD/REX/SUP Leader – DØ

Learn how to use SAM for retrieving and storing files

The outline of this talk

Very brief introduction to SAM for those who don't know or who haven't wanted to know. Why is using SAM important?

How to...



SAM is DØ's data handling system

Sequential data Access via Meta-data ...

You ask for and get your files one at a time

You select the files based on a meta-data query

In practice, SAM...

Is your only access to the tape system (Enstore)

Catalogs DØ's files

Caches and delivers files to jobs (or to your desktop)

Keeps track of file usage and job status

What can SAM do for you?

Get files from and put files to tape

Efficiently deliver files to your job

(here efficiently means for the whole of DØ)

Keeps track of your job status so you can recover lost output

Backup and cache your important files

How to use SAM

Command line:

```
setup sam  
sam # see list of all commands  
sam command without arguments # see options
```

How to get SAM Information (DØ at work)



<p>Collaboration</p> <p>Organization, Office Maps, People & Institutions, DØ Physicist Photo Gallery, Flags and Institutions Collaboration Banner, Flags and Map Collaboration Banner, Flags and Map DØ+CDF, Country Homepages, Author List & Masthead, Institutions & Contacts, Information for new arrivals (new), (old), Official DØ Photographs, Collaboration Meeting Archives</p>	<p>General</p> <p>DØ Calendar, DØ Agenda Server, Agenda Server Overview, Daily Meetings, Meeting Rooms, DØ News, DØ Notes, DØ Wiki, All DØ meeting/ old, Speakers Bureau, Institutional Board, Advisory Council, All Experimenters' Meeting, DØ Requisitions, Internal Docs, Video Conferencing, University of DØ</p>
<p>Detector</p> <p>Run II Operations, Online Shift Calendar, Online Logbook, Shifter Tutorials, Live Events, Run IIb Upgrade, Accelerator Status, Run II Luminosity, Luminosity Monitoring, Data Quality</p>	<p>Physics</p> <p>Run II Results, Publications, DØ Theses, Algorithm Groups, Physics Groups, Run II Editorial Boards</p>
<p>Computing & Core Software</p> <p>Online, Computer Accounts, Infrastructure, Framework, Monte Carlo, MC Production, Tools, SAM, DØ Computing Systems, Network Monitoring, DØmino0x, Remote Computing, DØ Grid, (Re)Processing, ClueDØ, DØ PC Support, Computing Planning Board</p>	<p>Software Algorithms</p> <p>Tracking, Calorimetry, Muons, Alignment/Calibration, Level 3, RECO, Simulation, Trigger Simulation, Event Display, Trigger Study Group</p>
<p>Trigger</p>	<p>Useful links</p>

Useful SAM information shown

Fermi National Accelerator Laboratory

The DØ Experiment

DØ at Work Computing Getting Started Documentation Systems
Infrastructure Production Organization Accounts

SAM Data Handling at DØ

User Links <ul style="list-style-type: none">• What is SAM?• Quick Start Guide to SAM• Form for registering as a DØ SAM user• SAM tutorials: DØ software tutorial (includes SAM use)• SAM Tips and Tricks• SAM commands• Browse the SAM Meta-data (old, but works) ←• DØ database browser (newer)• Dataset Definition Editor (super-new) ←• Grid job submission (in testing)	Monitoring SAM <ul style="list-style-type: none">• Data Handling At A Glance – an overall picture of DH• Station monitoring for FNAL SAM stations (an improved SamTV)• Sam At A Glance – all active DØ SAM stations• SAMGrid monitoring – SAM stations w/ Grid-enabled job submission• ENSTORE Tape Status – lists no-access tapes• ENSTORE System Status – the robotic tape systems for DØ• FCP Monitor – monitor file transfer queues• SAM Transfer Plots – transfer rates for cabsrv1/2 stations• DØ SAMGrid Usage Metrics
What to do in case of SAM problems <ul style="list-style-type: none">• Email D0sam-admin@fnal.gov• SAM Issue Tracker• Archives of sam-users mailing list• The operations model: what is the coverage for SAM problem resolution?	Offline Shifter Links <p>Offline Shift Coordinator: Kin Yip</p> <ul style="list-style-type: none">• Offline Shifter SAM page (~ old Backdoor page - some good docs here)• Sam_admin commands• SAM Shifter FAQ• Today's offline shifters• Roster of offline shifters
SAM Station Admin Links <ul style="list-style-type: none">• SAM Installation• Installation 'diaries': recent experience with station installation• List of SAM station admins• Triage for station problems• JIM Installation	The SAMGrid Project <p>SAMGrid project leader: Adam Lyon, CD</p> <ul style="list-style-type: none">• Project Web Page• SAMGrid phone pages• Recent talks on SAM status or performance: Bird Review of Run II Computing, Sept 03 DØ Grid Strategy Workshop, Oct 2003 Link to CD Project Briefings on SAMGrid

Why use SAM instead of doing your own data handling?

All data in SAM are backed up to tape!

(Tapes are not cheap, but if you are sane you can't break the bank)

Data delivery is managed with a big cache, so you can...

Save room on the project disks

Project disks will be accessed less (*faster for everyone*)

No need to clean/update skims on project disks

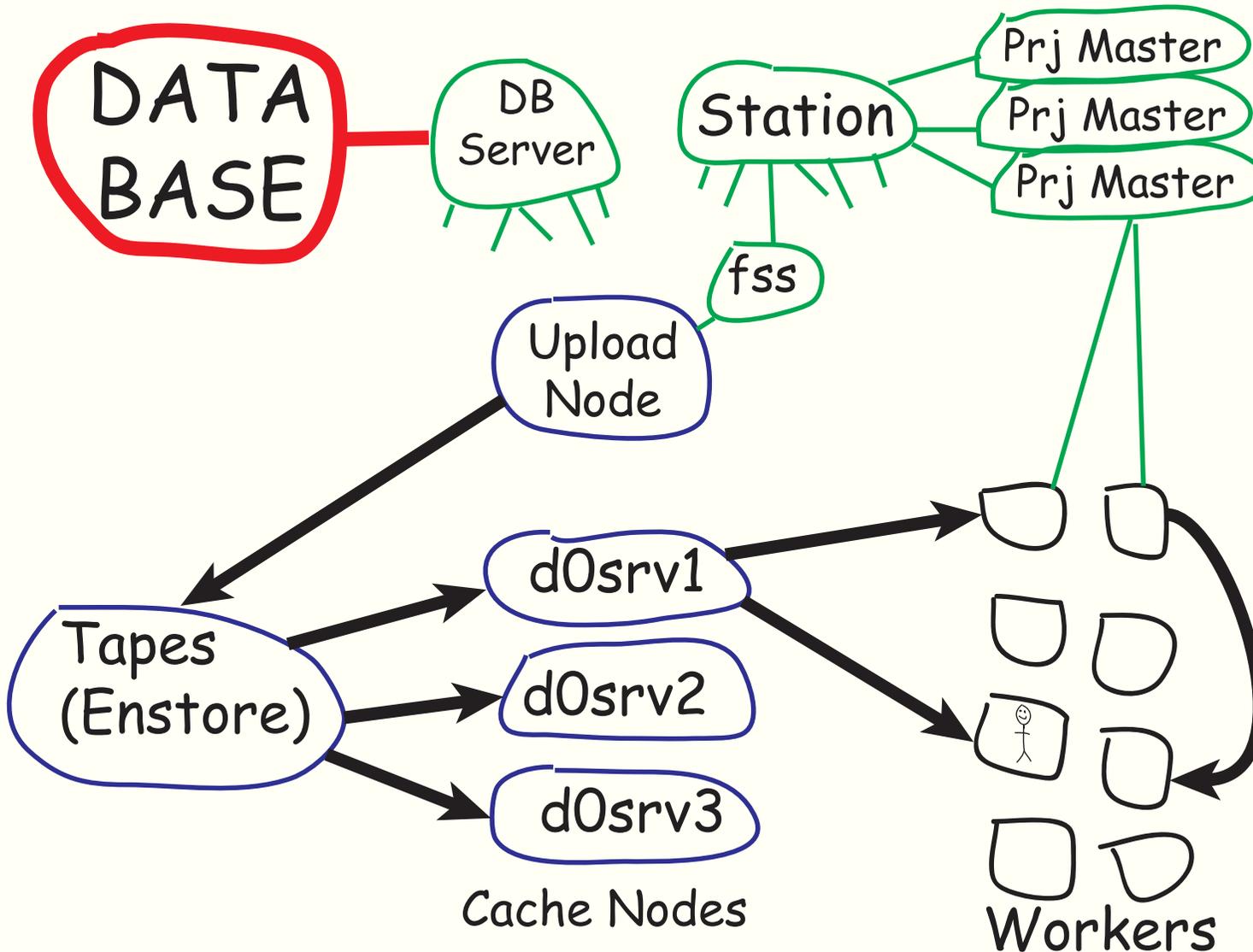
(*Popular files stay in the cache, unused files go away*)

SAM can handle lots of simultaneous access

SAM keeps track of failed jobs - makes recovery easier

But SAM is intimidating - will hopefully show it is not as difficult as it looks

There's lots going on behind the scenes



Total ~5 PB on tape; 640 TB big cache; 240 TB worker cache

How to run jobs with SAM

- 1. Write/build/copy/steal analysis code**
- 2. Choose or create a dataset definition* to run over**
- 3. Config cafe to run with SAM (if necessary)**
- 4. run_cafe (or run your group's scripts)**

***dataset definition: A meta-data query that resolves to a list of files (hopefully you can use a predefined dataset)
You typically set the dataset name in your job config.**

***snapshot: The real list of files the query returned when your job started (typically hidden from you)**

Common Samples Group has lots of good info...

Fermi National Accelerator Laboratory

The DØ Experiment

For the Public | DØ Results | DØ Collaboration | DØ at Work

2010 Winter Results [Dmitri to continue as Spokes for another two years! Congratulations!](#)

Collaboration Organization , Office Maps , People & Institutions , DØ Physicist Photo Gallery , Flags and Institutions Collaboration Banner , Flags and Map Collaboration Banner , Flags and Map DØ+CDF Country Homepages , Author List & Masthead , Institutions & Contacts , Information for new arrivals (new) , (old), Official DØ Photographs , Collaboration Meeting Archives	General DØ Calendar , DØ Agenda Server , Agenda Server Overview , Daily Meetings , Meeting Rooms , DØ News , DØ Notes , DØ Wiki , All DØ meeting/ old , Speakers Bureau , Institutional Board , Advisory Council , All Experimenters' Meeting , DØ Requisitions , Internal Docs , Video Conferencing , University of DØ
Detector Run II Operations , Online Shift Calendar , Online Logbook , Shifter Tutorials , Live Events , Run II Upgrade , Accelerator Status , Run II Luminosity , Luminosity Monitor , Data Quality	Physics Run II Results , Publications , DØ Theses , Algorithm Groups , Physics Groups , Run II Editorial Boards
Computing & Core Software Online , Computer Accounts , Infrastructure , Framework , Monte Carlo , MC Production , Tools , SAM , DØ Computing Systems , Network Monitoring , DØminox , Remote Computing , DØ Grid , (Re)Processing , ClueDØ , DØ PC Support , Computing Planning Board	Software Algorithms Tracking , Calor RECO , Simulation , Trigger Study G
Trigger TriggerMeister , Trigger Board , L1 , L2 , L3/DAQ , Current DAQ Rates	Useful links HEP Experiment useful links , Fermilab: Proce , Various: useful , DØ Logos , Fermilab Trainin , DØ Search , Fermilab Phone , Fermilab Intern

Physics Groups / Common Analysis

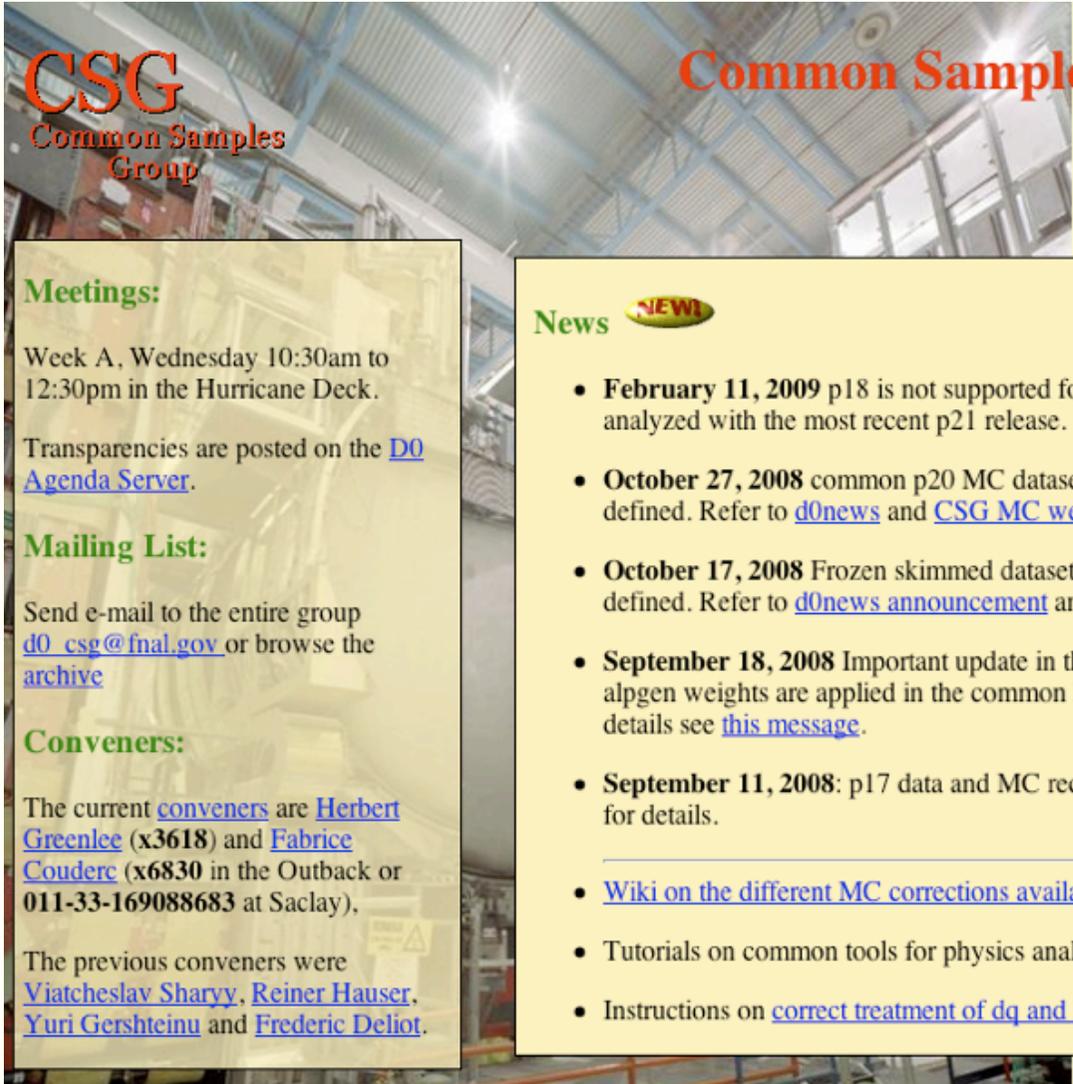


Physics Coordinators: [Gregorio Bernardi](#) and [Marco Verzocchi](#)

Physics conveners [mailing list](#) and [archive](#); algorithm and physics conveners [mailing list](#) and [archive](#)

Physics Groups (Meeting time)	Conveners
Electroweak (Week B: 9:00-11:00, Mon)	Jan Stark , Junjie Zhu
B Physics (Week B: 9:00-11:00, Thurs)	Mari Corcoran , Penny Kasper
Higgs (Week A: 9:00-11:00, Thurs)	Aurelio Juste , Wade Fisher , Krisztian Peters
New Phenomena (Week A: 11:00-13:00, Thurs)	Gustaaf Broojimans , Arnaud Duperrin
QCD (Week B: 11:00-12:30, Wed)	Dmitri Bandurin , Sabine Lammers , Don Lincoln
Top Quark (Week A: 11:00-13:00, Fri)	Frederic Deliot , Aran Garcia-Bellido , Christian Schwanenberger
Working Groups	
Jet Energy Scale (Week B: 9:00-10:30, Wed)	Shabnam Jabeen , Mike Wang , Markus Wobisch
Common Analysis (Week A+B: 9:00-11:00, Tue)	Fabrice Couderc , Herb Greenlee
Luminosity (Week A: 10:00-11:30pm, Fri)	Marjorie Corcoran , Greg Snow
V+Jets (Week A+B: 9:00-11:00, Tue)	Slava Shary , Lidija Zivkovic
Trigger (Week A+B: 11:00-12:30pm, Mon)	Marc Buehler , Rick Jesik
MC Requests Coordination	Tibor Kurca
Data Quality Group	
Data Format Working Group	

Instructions for running jobs



CSG

Common Samples Group

Common Samples Group



Meetings:
Week A, Wednesday 10:30am to 12:30pm in the Hurricane Deck.
Transparencies are posted on the [DØ Agenda Server](#).

Mailing List:
Send e-mail to the entire group d0_csg@fnal.gov or browse the [archive](#)

Conveners:
The current [conveners](#) are [Herbert Greenlee](#) (x3618) and [Fabrice Couderc](#) (x6830 in the Outback or 011-33-169088683 at Saclay),
The previous conveners were [Viatcheslav Sharyy](#), [Reiner Hauser](#), [Yuri Gershteinu](#) and [Frederic Deliot](#).

News 

- **February 11, 2009** p18 is not supported for common tools. p18 caf-trees must be analyzed with the most recent p21 release.
- **October 27, 2008** common p20 MC datasets for Winter 2009 physics analyses are defined. Refer to [d0news](#) and [CSG MC web pages](#).
- **October 17, 2008** Frozen skimmed datasets for Winter 2009 physics analyses are defined. Refer to [d0news announcement](#) and [pass 4 data page](#).
- **September 18, 2008** Important update in the heavy flavor skimming and the way how alphen weights are applied in the common MC tool package (caf_mc_util). For more details see [this message](#).
- **September 11, 2008:** p17 data and MC recafings are finished. See [ADM presentation](#) for details.

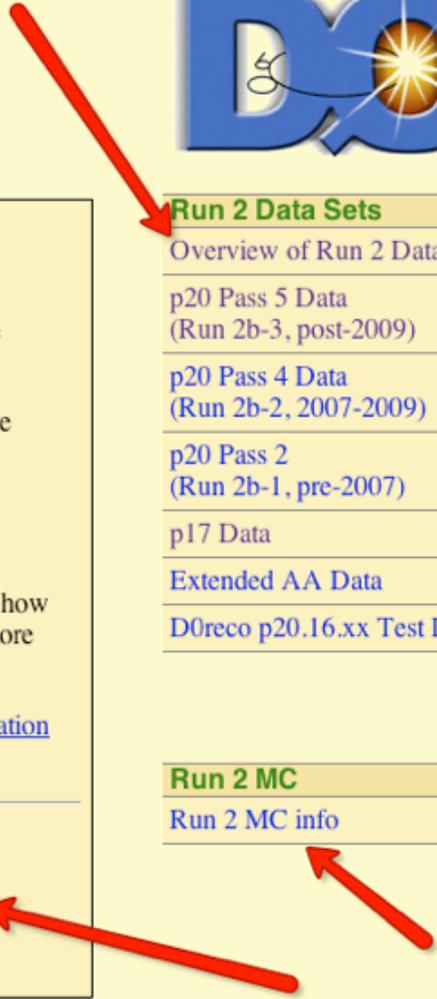
- [Wiki on the different MC corrections available within CAFe](#).
- Tutorials on common tools for physics analyses are available in the [CAF section](#).
- Instructions on [correct treatment of dq and luminosity determination](#) are available

Run 2 Data Sets

- Overview of Run 2 Data
- p20 Pass 5 Data (Run 2b-3, post-2009)
- p20 Pass 4 Data (Run 2b-2, 2007-2009)
- p20 Pass 2 (Run 2b-1, pre-2007)
- p17 Data
- Extended AA Data
- DØreco p20.16.xx Test Data

Run 2 MC

- Run 2 MC info



Selecting datasets

CSG_CAF_XXXXX_PASS2_p21.10.00

where XXXXX is the skim name.

When using the `genLBNtables` script for the generation of the parentage tables use the option `-pass p21pass2fixed`, even when using the unified datasets (the `genLBNtables` script will pick up the correct parentage also for the unfixed files). If you are analyzing the regenerated CAF trees, please use the option `-pass p21pass2regaf`.

Datasets (Unified Data)

Skim	Thumbnail Dataset Definition	CAF Tree Dataset Definition	Regenerated CAF Tree Dataset Definition
EMinclusive	CSskim-EMinclusive-PASS2-p21.03.00-allfix2007	CSG_CAF_EMinclusive_PASS2_p21.05.00_all_fixed2007	CSG_CAF_EMinclusive_PASS2_p21.10.00
MUinclusive	CSskim-MUinclusive-PASS2-p21.03.00-allfix2007	CSG_CAF_MUinclusive_PASS2_p21.05.00_all_fixed2007	CSG_CAF_MUinclusive_PASS2_p21.10.00
ZBMB	CSskim-ZBMB-PASS2-p21.03.00-allfix2007	CSG_CAF_ZBMB_PASS2_p21.05.00_all_fixed2007	CSG_CAF_ZBMB_PASS2_p21.10.00
NP	CSskim-NP-PASS2-p21.03.00-allfix2007	CSG_CAF_NP_PASS2_p21.05.00_all_fixed2007	CSG_CAF_NP_PASS2_p21.10.00
TAUTRIG	CSskim-TAUTRIG-PASS2-p21.03.00-allfix2007	CSG_CAF_TAUTRIG_PASS2_p21.05.00_all_fixed2007	CSG_CAF_TAUTRIG_PASS2_p21.10.00
2EMhighpt	CSskim-2EMhighpt-PASS2-p21.03.00-allfix2007	CSG_CAF_2EMhighpt_PASS2_p21.05.00_all_fixed2007	CSG_CAF_2EMhighpt_PASS2_p21.10.00
2MUhighpt	CSskim-2MUhighpt-PASS2-p21.03.00-allfix2007	CSG_CAF_2MUhighpt_PASS2_p21.05.00_all_fixed2007	CSG_CAF_2MUhighpt_PASS2_p21.10.00
JPSI	CSskim-JPSI-PASS2-p21.03.00-allfix2007	CSG_CAF_JPSI_PASS2_p21.05.00_all_fixed2007_1	CSG_CAF_JPSI_PASS2_p21.10.00
3JET	CSskim-3JET-PASS2-p21.03.00-allfix2007	CSG_CAF_3JET_PASS2_p21.05.00_all_fixed2007	CSG_CAF_3JET_PASS2_p21.10.00
QCD	CSskim-QCD-PASS2-p21.03.00-allfix2007	CSG_CAF_QCD_PASS2_p21.05.00_all_fixed2007	CSG_CAF_QCD_PASS2_p21.10.00
Higgs	CSskim-Higgs-PASS2-p21.03.00-allfix2007	CSG_CAF_Higgs_PASS2_p21.05.00_all_fixed2007	CSG_CAF_Higgs_PASS2_p21.10.00
TOPJETTRIG	CSskim-TOPJETTRIG-PASS2-p21.03.00-allfix2007	CSG_CAF_TOPJETTRIG_PASS2_p21.05.00_all_fixed2007	CSG_CAF_TOPJETTRIG_PASS2_p21.10.00
EMMU	CSskim-EMMU-PASS2-p21.03.00-allfix2007	CSG_CAF_EMMU_PASS2_p21.05.00_all_fixed2007	CSG_CAF_EMMU_PASS2_p21.10.00

Best to use predefined dataset definitions than trying to roll your own, if possible.

If your group doesn't make dataset defs, ask them to (they can ask us for help).

Running cafe

CSG
Common Samples
Group

Common Analysis Format, *CAFe* Framework and CAF Common Tools

Quick Links

[Complete Starter](#)
[FAQs](#)
[Package Versions](#)
[CAF Classes](#)

Tutorials

[Common Tools Overview](#)
[cafe Tutorial 2005](#)
[caf_util Tutorial 2005](#)
[caf_util Updates 2007](#)
[Manchester Tutorial 2006](#)
[MSU 2007 Tutorial](#)

CAF communication

- A mailing list has been setup for discussion of CAF: d0-caf-users@fnal.gov
 - How to [subscribe](#).
 - [View archive](#).
- [D0 Wiki](#) CAF page.

Tutorials and documentation

- [2007 overview](#) of available common tools.
- To start your analysis in CAF: [the 2005 CAF common tools tutorial](#) and [2007 updates](#).
- To understand CAF and *CAFe* and start developing using *CAFe* framework: [the Vancouver tutorial](#).
- A complete analysis example can be found in Gavin's [Manchester 2006 Tutorial](#) and Mike's [MSU 2007 Tutorial](#).
- A [hypertext version](#) of the current CAF tree classes is available.
- [CAFe Frequently Asked Questions Wiki page](#).
- The CAF format was developed by the [Data Format Working Group](#). You can find there the [initial report](#) of the first working group and the [content document](#) of the implementation group.

Installation

Follow [this instruction](#). The recent packages versions list available [here](#).

Packages

[tmb_tree \(p18-br\)](#)
[cafe \(p18-br\)](#)
[caf_util](#)
[caf_mc_util](#)
[caf_dq, dq_util, dq_defs](#)
[B-tagging](#)
[caf_eff_utils](#)
[caf_trigger](#)
[muo_cert](#)

Follow instructions on <http://www-d0.fnal.gov/Run2Physics/cs/caf/>

There are lots of tutorials.

There's also vjets_cafe...

<https://plone4.fnal.gov/P1/D0Wiki/physics/VplusJets/CAFtools/>

How to store files into SAM

- 1. Run your job to make output files AND construct meta-data**
- 2. Declare meta-data to SAM database**
- 3. Store files to tape**

Metadata is the most mysterious part (but can be easy)

#2 & #3 can be done simultaneously

**The tape system prefers "Large" files. 1 GB files are optimal
Enstore will not accept files < 100 MB (a tape holds 800 GB)**

Have your jobs write as much meta-data as possible

You can construct meta-data after the job is done, but it is **MUCH** easier if the job does it...

Cafe will write the metadata for you ONLY IF YOU SET ALL OF:

```
SAM.Family: xx  
SAM.Application: yy  
SAM.Version: zz
```

in your config file...

Application name, family, version specify the job you ran.
You can make these things up, but they must be in the database.

```
family=analysis, application=cafe-bid, version=top-2006-v6  
family=higgs2ww_lvlv, application=hwwmumu_skim, version=v3_mar2010
```

Send mail to d0sam-admin@fnal.gov to put yours in the DB
(you need that done **before you run your jobs)**

More about running jobs

Be careful about using someone else's Application data. You may inadvertently add files to their datasets!

Example:

```
caf_tools/bin/runcafe -cabsrv1
-def=CSG_CAF_EMinclusive_PASS2_p21.10.00
-filecut=30 -jobs=15 -outfiles="*root *log *.e* *.o* *html *py"
-outdir=/work/chimera-clued0/lyon/dzero/working-p21.18.00/results/
-- skim_data_EMinclusive.config
```



```
# In config file ...
##### SAVE SKIMMED TMBTree #####
skim.File:                skim_%f
skim.Disable:             EMcnn
skim.FilesPerOutput:     5
SAM.Application:         adamTest
SAM.Version:             1
SAM.Family:              adamTest
##### THE END #####
```

Now, obsess over the job progress...

You can track your project easily with Station Monitoring

fnal-cabsrv1 station

Generated at 2010-03-25 16:15:23

There are 377 running projects

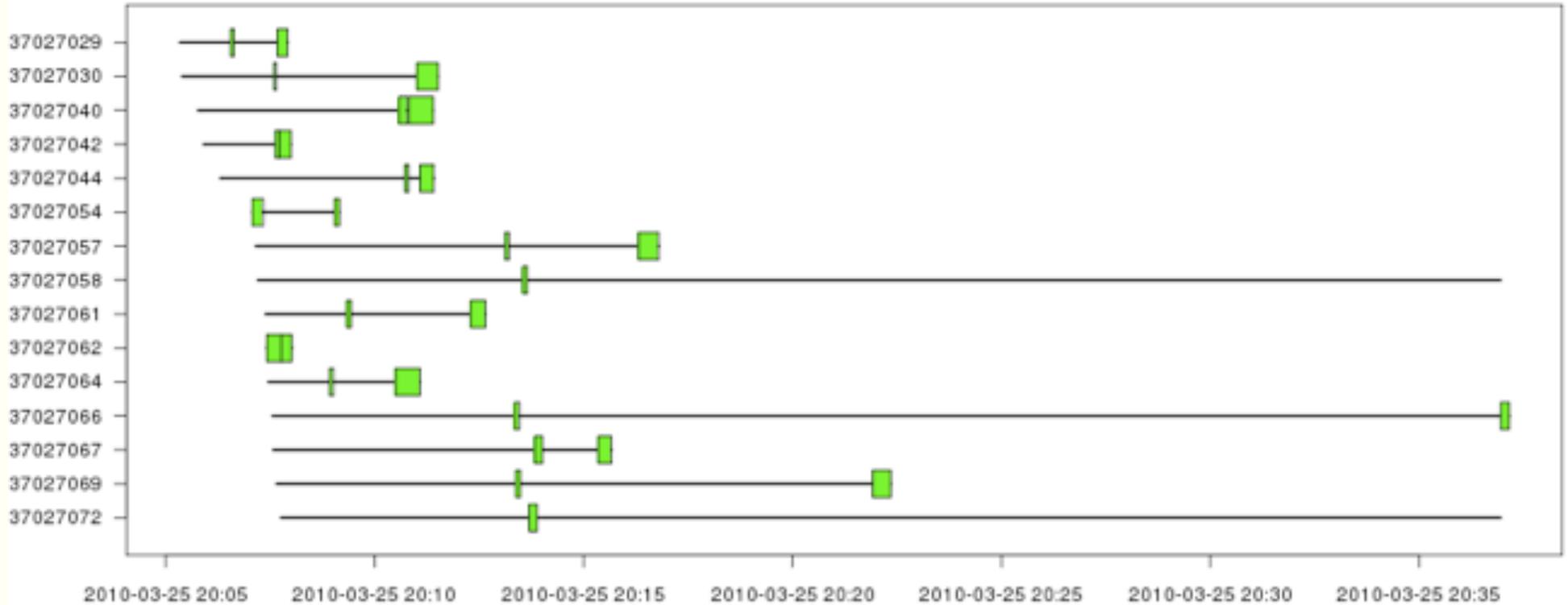
[Show recent completed projects](#) | [List all processes](#)

Project name ▲ ▼	Project Id ▲ ▼	Owner ▲ ▼	Status ▲ ▼	Files in snapshot ▲ ▼	Files seen ▲ ▼	Processes ▲ ▼	Idle/Waiting processes ▲ ▼	Last activity ▲ ▼	Longest waiting process ▲ ▼	Mean wait time (per file) ▲ ▼	Mean busy time (per file) ▲ ▼
CSG Skim p21.18.00 p20.16.07-r-25032010-151607	4206311	csgprod	running	68	18	17	-	file delivered at 2010-03-25 16:01:01	-	15min 47s	27min 57s
CSG alpgenpythia gamz+2c tautau+2c 250 1960 p211100 v3-25032010-155240	4206496	assantos	running	53	25	20	14	file delivered at 2010-03-25 16:03:40	4 minutes (process started)	47s	1min 9s
CSG alpgenpythia gamz tautau 75 130 p181301 v2-24032010-112220	4201945	shary	running	677	293	1	-	file delivered at 2010-03-25 15:57:32	-	1min 31s	4min 19s
CSG alpgenpythia gamz tautau 75 130 p181301 v2-24032010-112227	4201944	shary	running	677	360	1	-	file delivered at 2010-03-25 16:00:27	-	2min 35s	2min 9s
CSG alpgenpythia t+t 2b+4lpc m172 p211100 v3-25032010-150457	4206257	assantos	running	178	175	10	3	file delivered at 2010-03-25 16:03:00	7 minutes (consumed)	1min 19s	1min 0s
CSG alpgenpythia t+t 2l+2nu+2b m172 p211100 v3-25032010-150528	4206258	assantos	running	156	103	10	4	file delivered at 2010-03-25 16:03:31	8 minutes (consumed)	56s	3min 17s
CSG alpgenpythia t+t lnu+2b+2lpc m172 p211100 v3-25032010-150432	4206232	assantos	running	180	179	10	1	process ended at 2010-03-25 15:58:45	7 minutes (consumed)	1min 1s	1min 27s

Your project is the collection of jobs running over the files that satisfy the dataset definition.

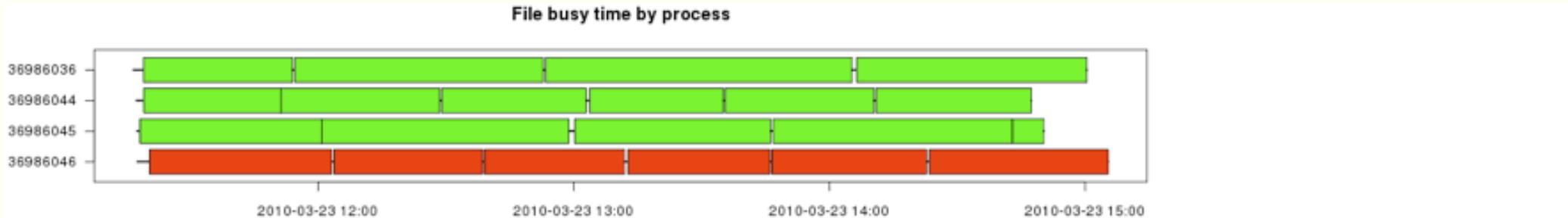
Here's the job I ran...

File busy time by process



My job had very fast execution time. Doesn't give SAM time to keep up. Not very efficient for me - but no big deal

Another example...



Processes

Process Id	Node name	Status	Description	Files seen	Last change
36986036	d0cs0909.fnal.gov	completed	3692521.d0cabsrv1.fnal.gov	4	2010-03-23 15:00:26 (process ended - completed)
36986044	d0cs2039.fnal.gov	completed	3692519.d0cabsrv1.fnal.gov	6	2010-03-23 14:47:24 (process ended - completed)
36986045	d0cs0911.fnal.gov	completed	3692522.d0cabsrv1.fnal.gov	5	2010-03-23 14:50:21 (process ended - completed)
36986046	d0cs0851.fnal.gov	error	3692520.d0cabsrv1.fnal.gov	6	2010-03-23 15:05:25 (process ended - error)

See this in text with

```
sam get project summary --project=...
```

Note the failed job. All of the output from that job is lost. Recovery is much easier now...

Recovering from job failures

```
> sam generate strict recovery project \  
--project=naosman_28163_20100323111244 --print-query
```

```
(snapshot_id 1927304 minus (consumer_id 4170379 and  
consumed_status consumed))  
or (consumer_id 4170379 and cf_pid 36986046)
```

To make a dataset...

```
> sam generate strict recovery project \  
--project=naosman_28163_20100323111244 \  
--recoDefName=myRecoveryDefinitionName
```

Now run the job again with the recovery dataset definition

Don't forget the strict keyword!!

Back to the metadata...

Cafe produced filename.metadata.py for me! Way easy!

```
from SamFile.SamDataFile import SamDataFile
from SamFile.SamDataFile import SamTime, SamSize, CRC, ApplicationFamily, DataType
from SamFile.SamDataFile import Params, ParamValue
from SamFile.SamDataFile import LumBlockRange, LumBlockRangeList
from SamFile.SamDataFile import RunDescriptor, RunDescriptorList
from SamFile.SamDataFile import NameOrId, NameOrIdList, SamLongList
from SamFile.SamDataFile import CaseInsensitiveDictionary
from SamFile.SamDataFile import DerivedDetectorFile
TheFile = DerivedDetectorFile({
    'fileName' : 'skim_CAF-CSGv3-CSskim-EMinclusive-20060224-081101-2111644_p17.09.03_p18.05.00.root',
    'fileId' : 0L,
    'fileType' : 'derivedDetector',
    'fileFormat' : 'unknown',
    'fileSize' : SamSize('3.7379MB'),
    'crc' : CRC('unknown crc value', 'unknown crc type'),
    'fileContentStatus' : 'good',
    'eventCount' : 100,
    'firstEvent' : 26188441,
    'lastEvent' : 26207420,
    'dataTier' : 'root-tree-bygroup',
    'applicationFamily' : ApplicationFamily('adamTest', 'adamTest', '1'),
    'group' : 'dzero',
    'processId' : 0L,
    'parents' : NameOrIdList([
        'CAF-CSGv3-CSskim-EMinclusive-20060224-081101-2111644_p17.09.03_p18.05.00.root']),
    'runDescriptorList' : [
        RunDescriptor('physics data taking', 188500)],
    'datastream' : 'notstreamed',
})
```

Metadata notes

Getting file type correct is important:

```
> sam get registered file types
```

```
nonPhysicsGeneric    # Good for log files
physicsGeneric       # Good for histogram/tree files
importedDetector     # Raw data
derivedDetector      # Descendants of raw data
importedSimulated    # Simulated data
derivedSimulated     # Descendants of raw data
unknown
```

You **NEVER** use importedDetector or importedSimulated

The minimum required metadata depends on the file type
If storing CAF trees w/o metadata from job, can use physicsGeneric

Required Metadata

```
<chimera-clued0> sam describe metadata requirements --fileType=nonPhysicsGeneric
```

```
Metadata requirements for fileType = 'nonPhysicsGeneric':
```

Required attributes	Data Type	Default Value
crc	CRC	CRC('unknown crc value', 'unknown crc type')
fileContentStatus	string	good
fileFormat	string	unknown
fileId	long	0
fileName	string	_UNKNOWN_FILE_NAME_
fileSize	SamSize	0.00B
fileType	string	unknown
group	NameOrId	None

```
<chimera-clued0> sam describe metadata requirements --fileType=physicsGeneric
```

```
Metadata requirements for fileType = 'physicsGeneric':
```

Required attributes	Data Type	Default Value
crc	CRC	CRC('unknown crc value', 'unknown crc type')
dataTier	string	None
fileContentStatus	string	good
fileFormat	string	unknown
fileId	long	0
fileName	string	_UNKNOWN_FILE_NAME_
fileSize	SamSize	0.00B
fileType	string	unknown
group	NameOrId	None

Required Metadata

```
<chimera-clued0> sam get registered file formats
compressed evpack
dspack
ethereal
evpack
gzipped-tar
root
run-1-sta
tar
unknown
aadst
```

For data-tier, you would most likely choose...

root-tree-bygroup # Your CAF trees (you don't have to use derivedDetector)

root-bygroup # Other Root files

You must ONLY use a -bygroup data-tier, to indicate unofficial production

Discover metadata

```
<chimera-clued0> sam get metadata --fileName=CAF-CSGv3-CSskim-EMinclusive-20060224-081101-2111644_p17.09.03
DerivedDetectorFile({
  'fileName' : 'CAF-CSGv3-CSskim-EMinclusive-20060224-081101-2111644_p17.09.03_p18.05.00.root',
  'fileId' : 10558950L,
  'fileType' : 'derivedDetector',
  'fileFormat' : 'unknown',
  'fileSize' : SamSize('1.31GB'),
  'crc' : CRC('1199986640L', 'adler 32 crc type'),
  'fileContentStatus' : 'good',
  'eventCount' : 38895L,
  'dataTier' : 'root-tree-bygroup',
  'firstEvent' : 26188593L,
  'lastEvent' : 37006515L,
  'startTime' : SamTime(1145320110.0),
  'endTime' : SamTime(1145329583.0),
  'processId' : 9707815L,
  'applicationFamily' : ApplicationFamily(appFamily='treemaker', appName='tmb_analyze', appVersion='csgca
  'group' : 'dzero',
  'parents' : NameOrIdList(['CSskim-EMinclusive-20060224-081101-2111644_p17.09.03', 'CSskim-EMI
581_p17.09.03']),
  'sourceSplit' : 0L,
  'destMerge' : 0L,
  'datastream' : 'all',
  'lumBlockRangeList' : LumBlockRangeList([LumBlockRange(2443955L, 3038309L)]),
  'runDescriptorList' : RunDescriptorList([RunDescriptor(runType='physics data taking', runNumber=188500)
})
```

Example metadata file for a tar file

```
from SamFile.SamDataFile import *
TheFile = NonPhysicsGenericFile({
    'fileName': 'WGamma-20100212-out.tgz',
    'fileType': 'nonPhysicsGeneric',
    'fileFormat': 'gzipped-tar',
    'fileSize': SamSize('123.3MB'), # stat -L -c %s <filename>
    'fileContentStatus': 'good',
    'group': 'dzero',
    'params': Params({'global': CaseInsensitiveDictionary(
        {'description': '"My Stuff"'})}), #optional
})
```

Check your metadata with...

```
sam verify metadata --descriptionFile=...
```

Note that only verifies the python, not the metadata content nor does it check for missing items

How to store a FEW files into SAM...

Create the ...metadata.py files

On clued0, move files to /work/jetsam-clued0/<you> or /work/lagan-clued0/<you> and log into machine

On d0mino, move files to a /prj_root/3nnn or /prj_root/5nnn disk (Bluearc NFS - faster)

[Be sure to clean up files when done!]

For each file, issue...

```
sam store --station=<station> --sourceFile=/path/fileToStore  
         --descriptionFile=<metadataFile.py>
```

For synchronous, add --waitForCompletion

<station> = fnal-cabsrv2 if on d0mino, otherwise clued0

Notes on file stores

Storing files may take many minutes. Instead of blocking, issue store command and periodically poll for status...

```
sam get file transfer request status --station=<station>  
      --transferIdentifier=<fileNameToStore>
```

If something goes wrong, can store again. But if metadata was accepted, then you cannot specify the metadata again; instead use

```
sam store --resubmit --station=... --sourceFile=...
```

To check that your file made it to tape,

```
sam locate <fileName>
```

A location starting with /pnfs is tape

How to store a LOT of files

We don't have a generic solution, as each situation has different requirements (but maybe we can help make something usable by most)

Some existing good examples:

Amnon/Adam scripts: I archived to,

```
wget http://d0dbweb.fnal.gov/rexipedia/lyon/samStoreExampleScript/storeData.py
```

```
wget http://d0dbweb.fnal.gov/rexipedia/lyon/samStoreExampleScript/countEvents.C
```

**Python script is specialized and cannot be used as-is, but shows:
Running asynchronous stores in parallel, forming metadata,
checking the store results**

More examples

vjets_cafe/skimming/store_files.sh [CVS]

Shell scripts for parallel stores with lots of error handling

csg_skimming/python/store_skims.py [CVS]

Simple script, but little error checking

These scripts assume that metadata already exist.

Forming metadata from scratch is doable, but may need external tools or tricks to fill in all the items (e.g. countEvents.C, looking at parent file metadata)

Once the files are stored, how do you retrieve them?

Best way is to make a dataset definition using the application name and version that you set in your job. (Try not to use file names directly).

```
sam create definition --defName="my_vjets" --group=dzero  
  --dim="APPL_NAME vjets_skim and VERSION v00.04.00  
    and FILE_NAME vjets-recaffed-ejets-fall2008%_p17.09%"
```

Or use the dataset definition editor

Now you can share that definition with your group.

How to copy files in SAM to your local disk

Sometimes you want files locally to experiment with (try your code)

The easiest way to get files to your clued0 disk is with the `sam get dataset` command.

```
setup sam ; setup fcp # Don't forget fcp
cd somewhereWithSpace
# Get specific files
sam get dataset --fileList=file1,file2,file3

# Get all files in a dataset (BE CAREFUL!)
sam get dataset --defname=datasetName
```

It will get files from CAB if they are cached there, otherwise will copy from tape.

Notes for sam get dataset

Will retrieve an entire dataset. To get just a few files, do

```
sam translate constraints --dim="dataset_def_name <dataset>"  
(or --dim="__set__ <dataset>" if above returns nothing),  
pick out the files you want, and retrieve by file names.
```

The command can take minutes, especially if retrieving from tape.

You can check if a file is cached on CAB with, e.g.

```
> sam locate \  
CAF-CSGv1-CSskim-EMinclusive-20070209-210519-4517025_p21.03.00_112414_p21.10.00.root  
  
['cchpssd0.in2p3.fr:/hpss/in2p3.fr/group/d0/upload'  
'd0srv054.fnal.gov:/sam/cache7//boo'  
'/pnfs/sam/dzero/db5/derivedDetector/csg-p21.10.00-p20.07.01-p20.08.xx/dzero/root-  
tree-bygroup/eminclusive/0003,299@psf262']
```

/pnfs/... are tape

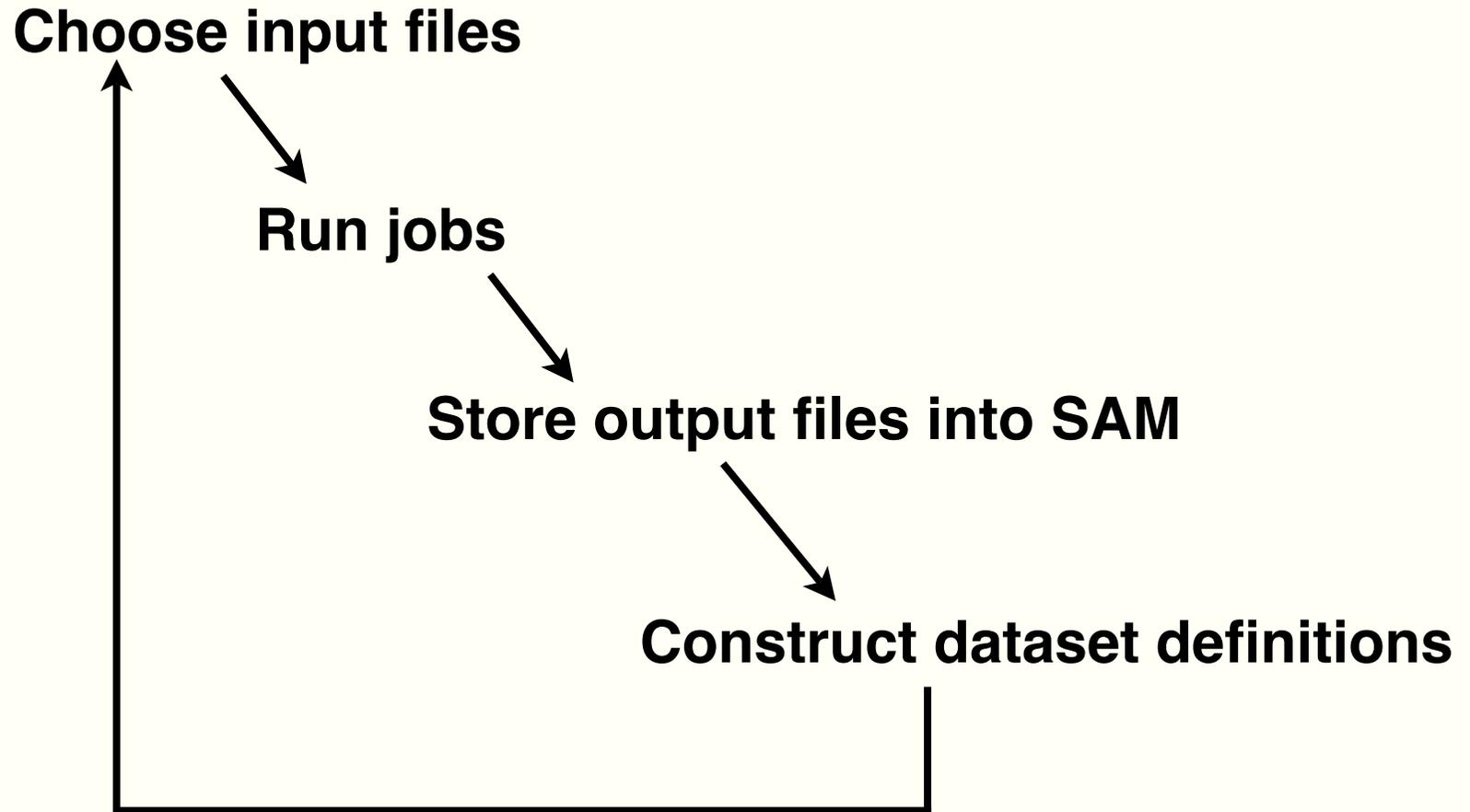
cchpssd0... is the tape system at Lyon, France

d0srv... is CAB cache!!!

What you've learned

Using SAM is to your, and your experiment's, advantage

How to...



Need SAM help? Send mail to d0sam-admin@fnal.gov