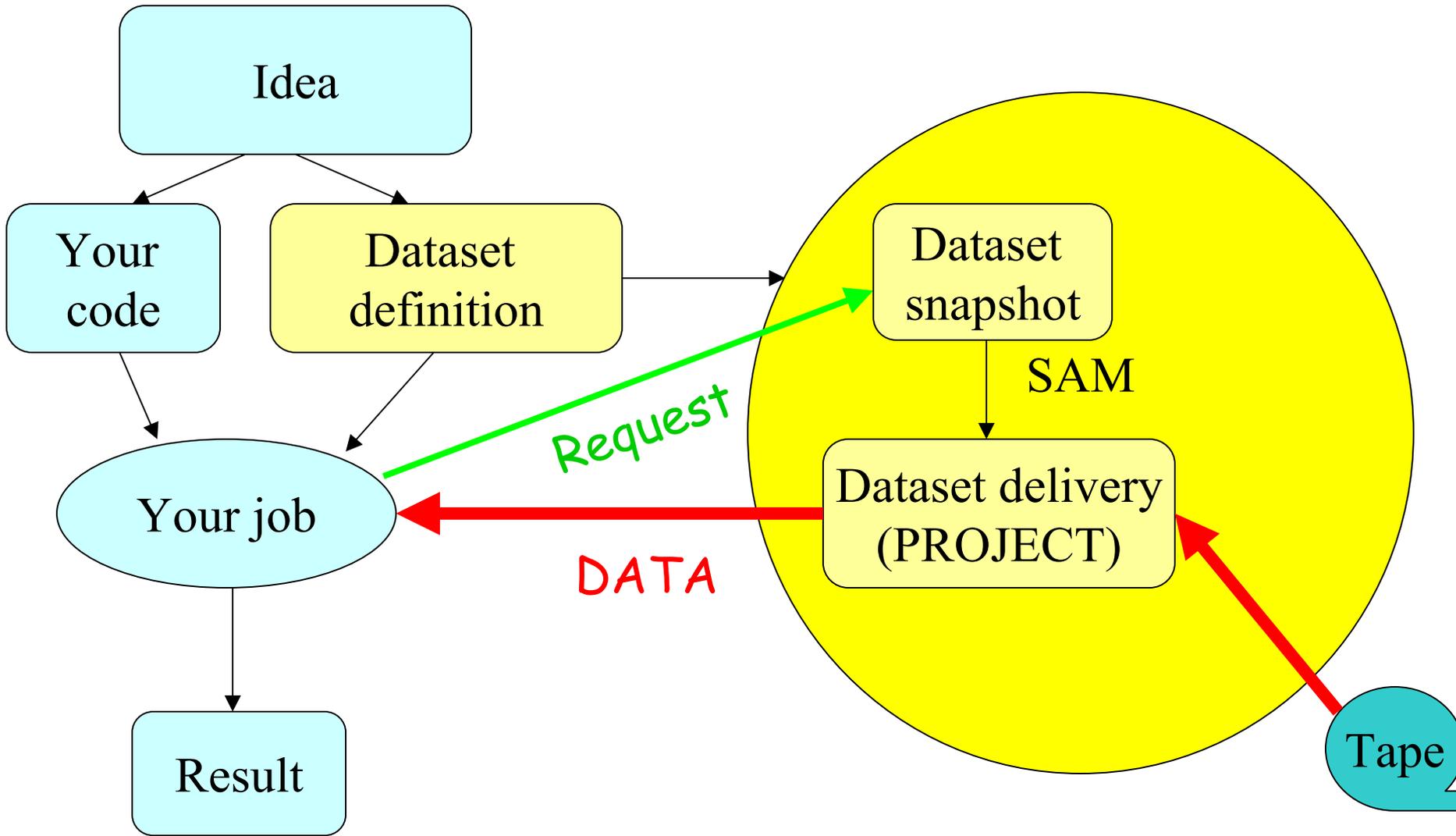


Getting data at DO

H. Schellman
October 8, 2002

Outline

- Overview of sam metadata
- What you need to select data
- Sources for more information about the data
- Relation of sam to analysis to luminosity
- Simple storing



Stored in sam

- File catalog
- Your dataset definition (an abstract description of your data)
- Each time you run, you also store:
 - The dataset snapshot
 - the list of files that satisfied the abstract definition when your job asked for them.
 - The status of the dataset consumer that ate the data.
 - Files delivered
 - Files consumed (your program said ok I got them)
- You can use all of this information to help with analysis.

More stored in sam

- Each file has lots of information
 - The usual, what's in it, format
 - Availability at various sites
 - How it was made
 - What program made it
 - Where was it made?
 - Who made it?
 - What were the input files that made it
 - What has been made from it
 - What programs were run on it
 - What files resulted...

In principle

- In principle, you have the full processing chain for files in sam.
- This allows you to keep track of what you are doing.

How to use sam in two pages

- **CTBUILD**

RegSAMManager in OBJECTS
sam_manager in LIBRARIES

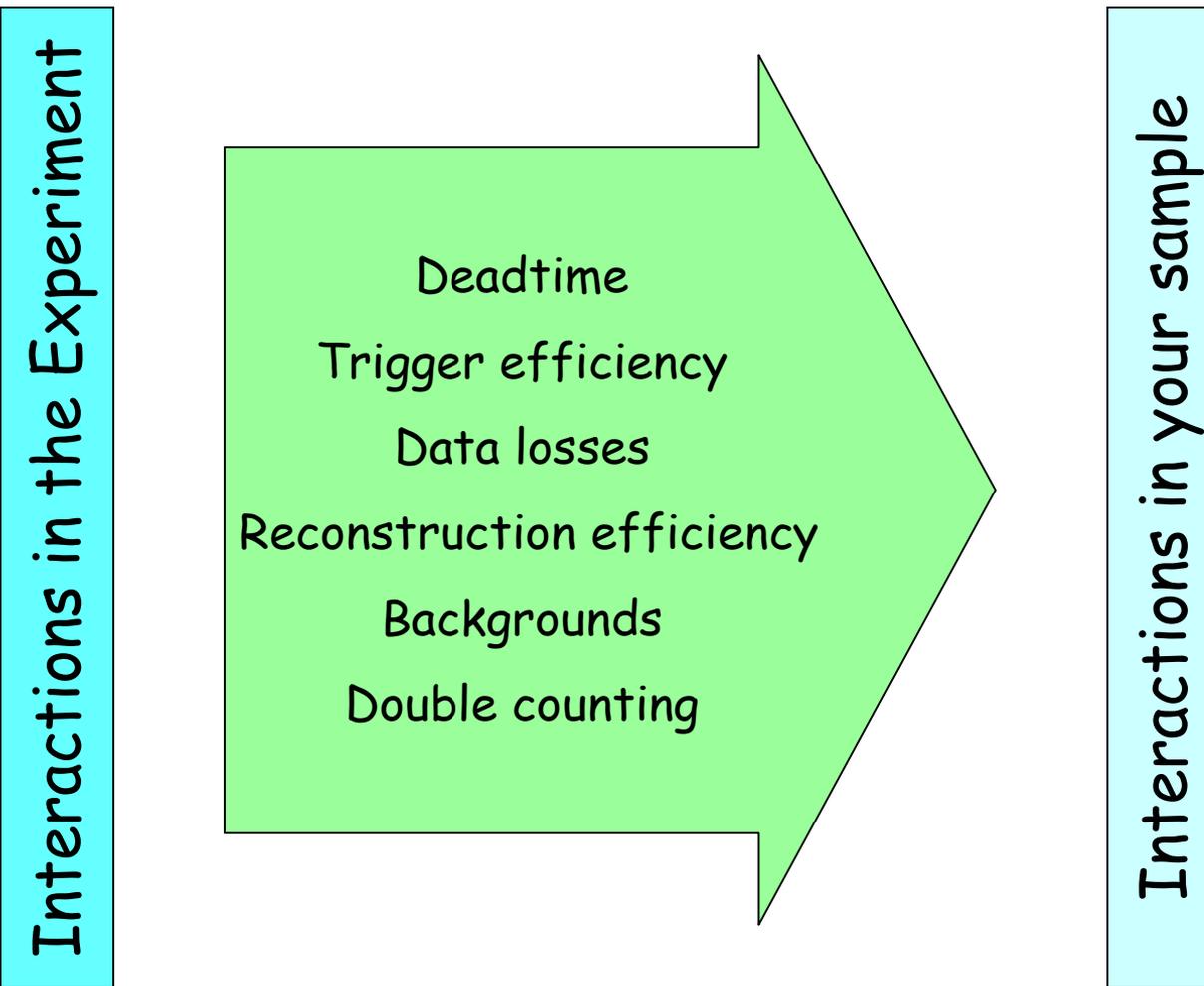
- **RCP**

packages = < ...geo sam read ... >
sam = <mypackage mySAMManager>

- **Input**

framework -input_file SAMInput:
d0tools -defname=<sam definition>

```
// -----  
// SAMManager.rcp  
// -----  
  
string PackageName = "SAMManager"  
string ApplicationFamily = "event_select"  
string ApplicationName = "event_select"  
string ApplicationVersion = "em_sel_01"  
int ProjectMasterTimeout = 240 // in minutes; default 0 (indefinitely)  
int FileRequestTimeout = 240 // in minutes; default 0 (indefinitely)  
int StoreRequestTimeout = 240 // in minutes; default 0 (indefinitely)  
// File store mode: 0 asynchronous (default), 1 synchronous  
int FileStoreMode = 0  
// Output file parentage mode: 0 opened input files (default), 1 closed  
// input files  
int FileParentageMode = 1
```



$$N_{\text{prod}} = \sigma \cdot \Phi \cdot \epsilon(t) \cdot \Delta t$$

$$N_{\text{obs}} = \epsilon(t) N_{\text{prod}}$$

Things you need to know to select data

- What type of data (raw, DST, TMB, root)
- When data taken (trigger version or run range)
- Run Data quality
- Code version
- Who made it (production, someone on CDF?)
- Where is it available
- Is it readable?
- Have I seen this data before and should I ignore?
- Is it luminosity safe?
- Near term future - streaming

SAM

- ⊕ Group
- ⊕ Person
- ⊕ Keyword
- ⊕ Dimension

- Search
- Create New
- Email help
- FAQ
- Documentation

SAM Dataset Definition Details

Definition Name: **Keyword Usages:**

 %125280%
 %125282%
 %125291%
 %130968%
 %130982%

Definition Id:

Create Date:

Work Group:

Username:

Description:

Dimension Query

*If you enter both a detailed query above and constraints below, they will be combined to make one query.
 Or, use the option to see the query before translating it into a set of files or saving it.*

Operator	Dimension	Constraint Value
and	<input type="text"/>	<input type="text"/>
and	<input type="text"/>	<input type="text"/>
and	<input type="text"/>	<input type="text"/>

Translate Constraints Results

Total File Count: 1009 Total Event Count: 11104376 Avg File Size: 3348

```

recoT_all_0000162523_mrg_003-010.raw_p11.11.00
recoT_all_0000162523_mrg_012-026.raw_p11.11.00
recoT_all_0000162587_mrg_001-001.raw_p11.11.00
recoT_all_0000162588_mrg_101-106.raw_p11.11.00
recoT_all_0000162588_mrg_108-108.raw_p11.11.00
recoT_all_0000162458_mrg_001-007.raw_p11.11.00
recoT_all_0000162593_mrg_001-004.raw_p11.11.00
recoT_all_0000162593_mrg_006-010.raw_p11.11.00
recoT_all_0000162586_mrg_001-003.raw_p11.11.00
recoT_all_0000162586_mrg_005-005.raw_p11.11.00
  
```

Dataset definitions

- DIMENSIONS - things you can select on
- OPERATORS (and, or, minus, >, <=)
- Parentheses - use these, the parser thinks it knows what you thought you meant, just like powerpoint.

Examples of definition components will be in a box throughout the talk.

The one below reuses an old definition you've already made then selects only those files which were created after 00:00 on 9/21/2002 (yep, it includes 9/21, it's an Oracle thing).

```
__set__ old_definition and create_date > 09/21/2002
```

Data-tiers

What format is the data?

- Official or 'bygroup'
- Filtered or not - have events been selected?
- Virtual or not (virtual are only for book-keeping)
- Production data_tiers
 - Raw
 - Reconstructed filtered-reco
 - Thumbnail filtered-thumbnail
 - Root-tuple filtered-root

```
data_tier thumbnail
```

Data taking periods

Trig_config_type **physics**
Trig_config_version **7.30**
Trig_config_name **%global%Cal%Muon%**

Run_number **164302-164309,165002**
Lum_min, lum_max **> 1673000**
File_partition **001**

```
(run_number 164302 and file_partition 001)  
trig_config_version 7.30
```

Data quality

- Run_quality reasonable
- Run_qual_group (run_qual good and run_quality_group muo)
- This isn't available yet for most runs...

```
(RUN_QUALITY REASONABLE and RUN_QUAL_GROUP MUO)
```

Code version

- Production versions

- p10.07.%
- p10.11%
- p10.14%
- p11.09.00,p11.10.01,p11.11.00,p11.12.01

http://www-d0.fnal.gov/computing/production/production_runs/

- If you do not specify you **WILL** get duplicate events.

```
version p11.1%
```

```
version p11.09.00,p11.10.00
```

Who made it

- Non-production stuff gets stored in data-tiers with a bygroup suffix.
- There is no guarantee that such datasets are well defined or can have luminosity calculated.

```
data_tier raw-bygroup
```

Where is it available?

- This gets you the files which are already in dOmino
- This is probably a biased sample, for example raw files on dOmino are likely to be there because the B group was picking 100,000 events.

```
full_path %dOmino
```

Have I seen it before?

Farms method - do it again unless stored to tape:

- Gets you all **raw** files which have not had **reconstructed** output stored back into sam for version **p11.11.00**

```
data_tier raw minus (file_analyzed>1 and  
version_analyzed p11.11.00 and data_tier_analyzed  
reconstructed)
```

Getting only unconsumed files - M. Verzocchi

```
sam create dataset definition --group="dzero" \  
  --defname="root1112base" \  
  --dim="version p11.12% and data_tier root-tuple"  
  
sam create dataset definition --group="dzero"  
  --defname="root1112notprocessed" \  
  --dim="__set__ root1112base minus \  
(project_name mv-proc1112-% and \  
  consumed_status consumed and consumer mverzocc )  
  
and then every time you submit a project do  
  
set TIMESTAMP=`date +%m%d%Y-%H:%M:%S`\  
setenv SAM_PROJECT "mv-proc1112-$TIMESTAMP"  
sam submit --defname="root1112notprocessed"
```

If a job fails you can recover using the timestamp...

What's out there

Browse other people's definitions in the dataset browser:

http://d0db.fna1.gov/sam_project_editor/DatasetEditor.html

Sam data Browser

http://d0db.fna1.gov/sam_data_browsing

- **General data Browser**

<http://cdfdbb.fna1.gov:8520/cdfr2/databases>

- **Runs database**

<http://d0db.fna1.gov/run/runQuery.html>

Quickie Query Lists

[Data Tiers](#)

[Logical Data Stream](#)

[Physical Data Stream](#)

[Run Types](#)

[Stations](#)

[File Locations](#)

[WorkGroups](#)

[Nodes](#)

[Parameter Categories](#)

[Parameter Types](#)

SAM Data Browsing

[Explore the SAM Data Model](#)

SAM provides dynamic data browsing capabilities available from your web browser. The following data browsing pages allow you to perform various queries against the SAM meta-data. Simply enter the query parameters that you want to use to limit your findings and click **Run** to query the SAM database. The output of your query will display directly in your web browser.

In some cases, your output will also include buttons on fields in the output. You may click on these buttons to "drill down" deeper into the database and see further relationships that may be pertinent to your original search.

These and other data browsing web pages are easy to build using the [MISWEB](#) technology.

Members	Query for all members currently registered in SAM.
Analyzed Files	Query the various files that have been analyzed per project.
Analyzed Files Summary	Query a summary count of files analyzed per period, drilling down to the specific list of consumers per file.
Data Files	Query the multitude of data files available in SAM.
Cached Files	Query the files cached on SAM stations, including pinned files.
Analysis Projects	Query the various analysis projects that are recorded in SAM.
Datasets (formerly Project Snapshots)	Query the various datasets that are recorded in SAM.
Dataset Definitions (formerly Project Definitions)	Query the various dataset definitions that are recorded in SAM.
Event Query	Query events that are recorded in SAM.
Station Queries	Query SAM stations.
D0 Farm Jobs	Query D0 Farm Batch Jobs.
Parameter Value Query	Query Parameter Values.
Request Query	Query MCRUnjob Requests.
Application Families Query	Query Application Families.

— For help contact sam-users@fnal.gov —



<http://cdfdbb.fna1.gov:8520/cdfr2/databases>

Randy Herber

CDF Run2 Data Bases Queryer and Browser

Reference Materials

All report groups condensed.

-  Submit query to report phase.
-  Clear form to default values
-  Select a report group to expanded.



Source selection

-  **The source defaults below should be adequate.**
-  Search within selected **offline** data server.



This data base specification generally is used by the **Data Catalog** group of reports and is used to obtain **Data Catalog** data regardless of report group as some reports use multiple data bases.

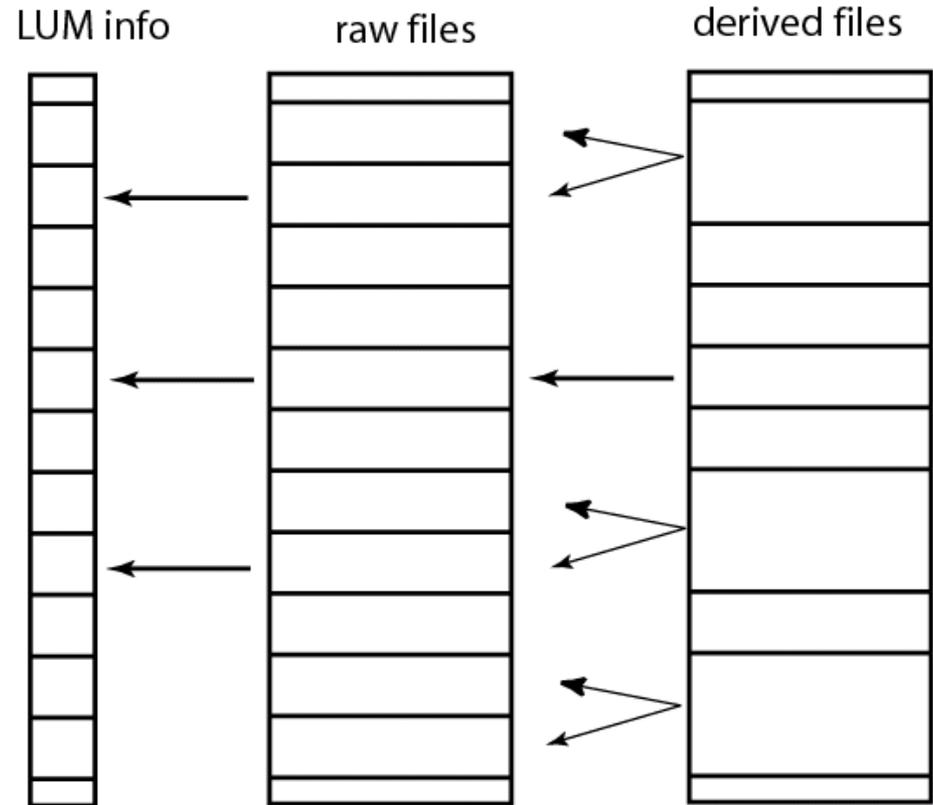
How sam relates to analysis

Sam tries to track what you do so you can normalize later

- Sam stores
 - your definition
 - what files were actually delivered
 - what processes were run on those files
 - what processes made those files
-
- In principle, you can figure out everything done to a file since it was logged.

Luminosity information is associated with files

- At D0, the luminosity for your analysis is found by associating the files in your analysis with luminosity information.
- This makes knowing which files were in your analysis vital.
- This information is stored in the file metadata



What we know about derived file

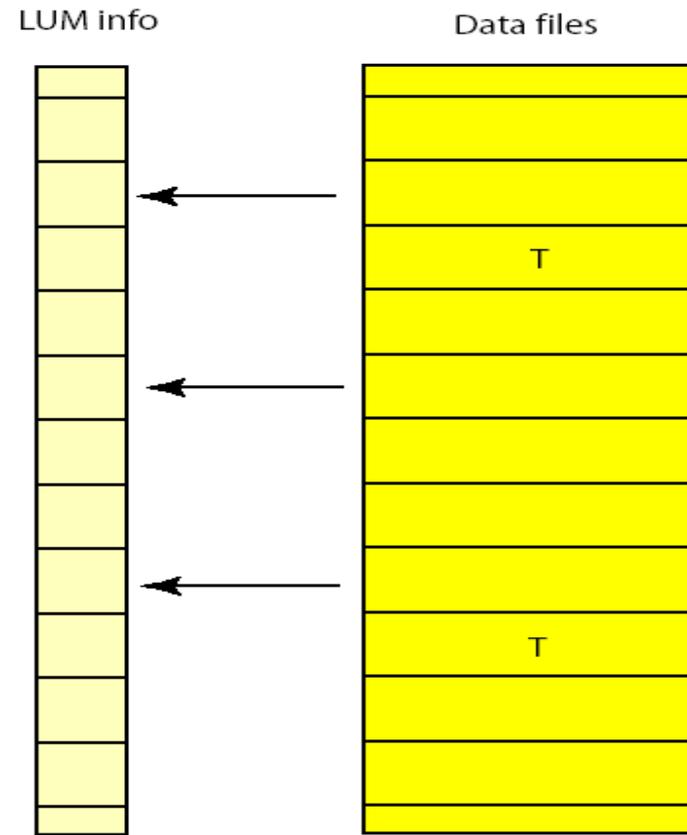
recoT_all_0000164605_mrg_210-213.raw_p11.12.01

```
runNumber: 164605
physicalDatastreamName: all,
dataTier: thumbnail,
eventCount: 8852
lumMin: 1585395, lumMax: 1585398,
version: p11.12.01, applName: recon_root,
projectName: farm.p11.12.01.18157,
projSnapId: 38056,
projectDefName: farm-dayset-2002-09-24-164605-2-
p11.12.01_20020925163504
children list: []
parent list: ['all_0000164605_210.raw',
'all_0000164605_211.raw', 'all_0000164605_212.raw',
'all_0000164605_213.raw']
```

```
sam dump file --filename=<name>
sam get metadata --filename=<name>
```

Correct example

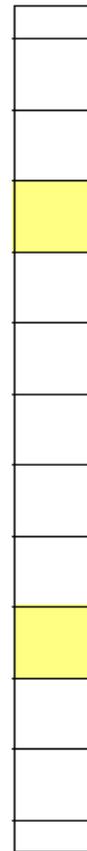
- Your data actually only has trigger T in 2 files but the list of files that trigger T could have been in is much larger.
- The luminosity corresponds to all of the time that trigger T was live, not just when it fired.
- Your correct list of input files is the full list, not just the ones with trigger T in them.



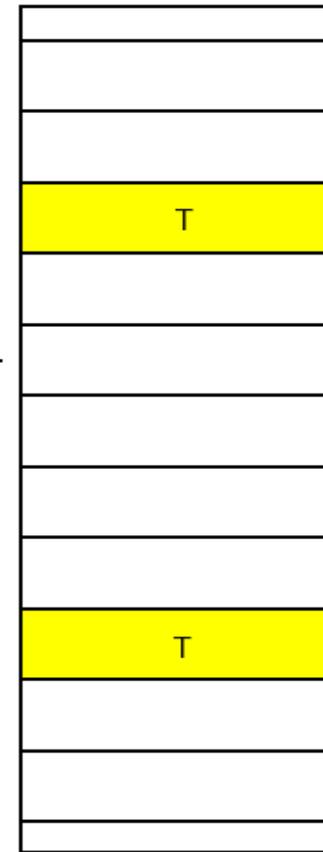
Bad example

- You logically say, why get all the files, I only need the ones with my trigger in them...
- Unless you get that set of files carefully, you will only get a small fraction of the luminosity.
- Pick events is this example!

LUM info

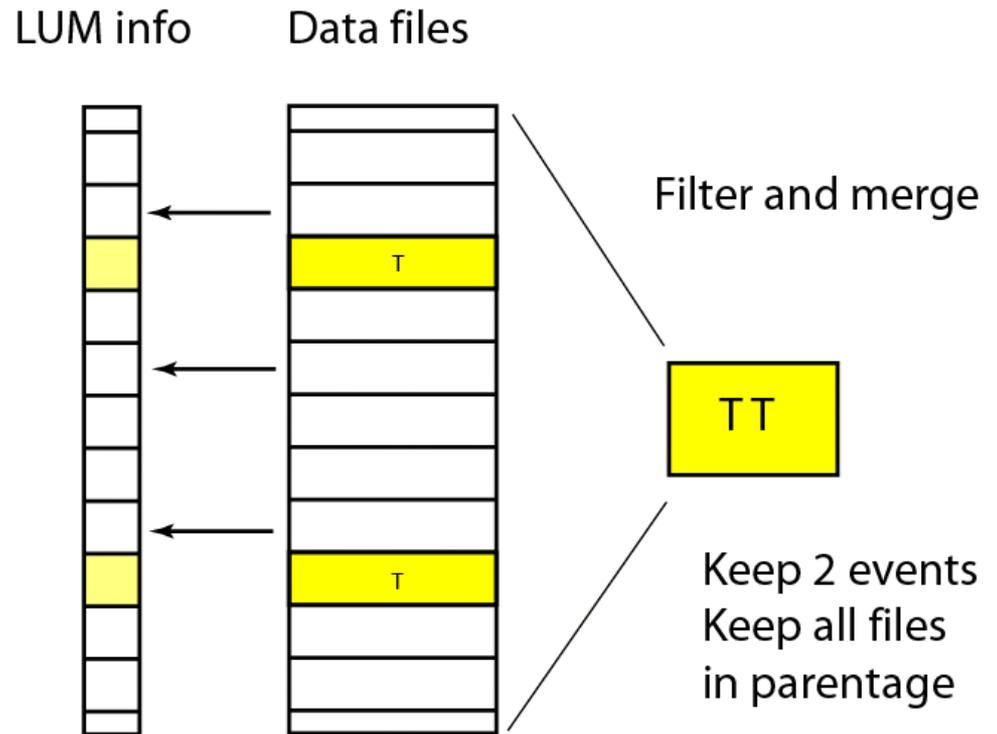


Data files



Solutions

- Production filtering merges files and includes all input files in the parentage, your big set of files becomes one smaller file which has pointers back to all of the luminosity.



Another method...

- Make 2 sam definitions, one with all data that trigger was live for, the other for those files which actually have events. Use list of files in the first one to derive your luminosity.
 - Quality cuts?
 - Reconstruction losses?

Storing files back into sam

If you use sam for input, and write output in DST or thumbnail format, you will get an output file:

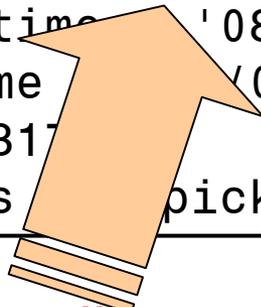
`<outputfile>.metadata.py`

as well.

With a little editing you can use this to store your output back into sam.

This is metadata produced by a copysam.py job which merged two input files which came from sam, pick_w32.dat and pick_w1.dat

```
from import_classes import *
TheFile = ProcessedFile(
    name = '2files',
    sizeK = 40104,
    events = Events(3654425, 641460, 198),
    stream = '',
    tier = 'reconstructed',
    start_time = '08/02/2002 19:56:00',
    end_time = '08/02/2002 19:56:05',
    pid = 817,
    parents = ['pick_w32.dat', 'pick_w1.dat'])
```



Currently sam calls everything reconstructed - you have to change this before you store the file! You must use a <data_tier-bygroup>

To store data you need

- A **file** to store - with a unique name - please don't use something that looks real official - others may pick up your file in a query...
- Valid **metadata** for that file, generated by the framework if you use sam input and EVPACK output. (and then edited to have the right data_tier)
- A **pnfs** location to store it to - ask your physics/id group boss.
- The WZ group has a script which does this ...

```
sam store --descrip=<meta.py> --source=$PWD  
--dest=<your pnfs location>
```

Can I normalize this?

- Unless all of the input files corresponding to the trigger list or run range for your analysis are included in the parentage, these derived files don't have the information in them to get a luminosity directly.
- In principle you use sam to get the list of parent files you should have had and use that for normalization.
- Production files do have the full information in their metadata.