AMI Tutorial Tbilisi October 2010

Solveig Albrand
Jerome Fulachier
Fabian Lambert

LPSC Grenoble
Laboratoire de Physique Subatomique et de Cosmologie

Plan

- AMI (ATLAS Metadata Interface) is a tool for DATASET DISCOVERY
  - Introduction to ATLAS datasets
  - Introduction to dataset nomenclature
- How AMI works – and what it aims to do
- Practical exercises on Thursday afternoon

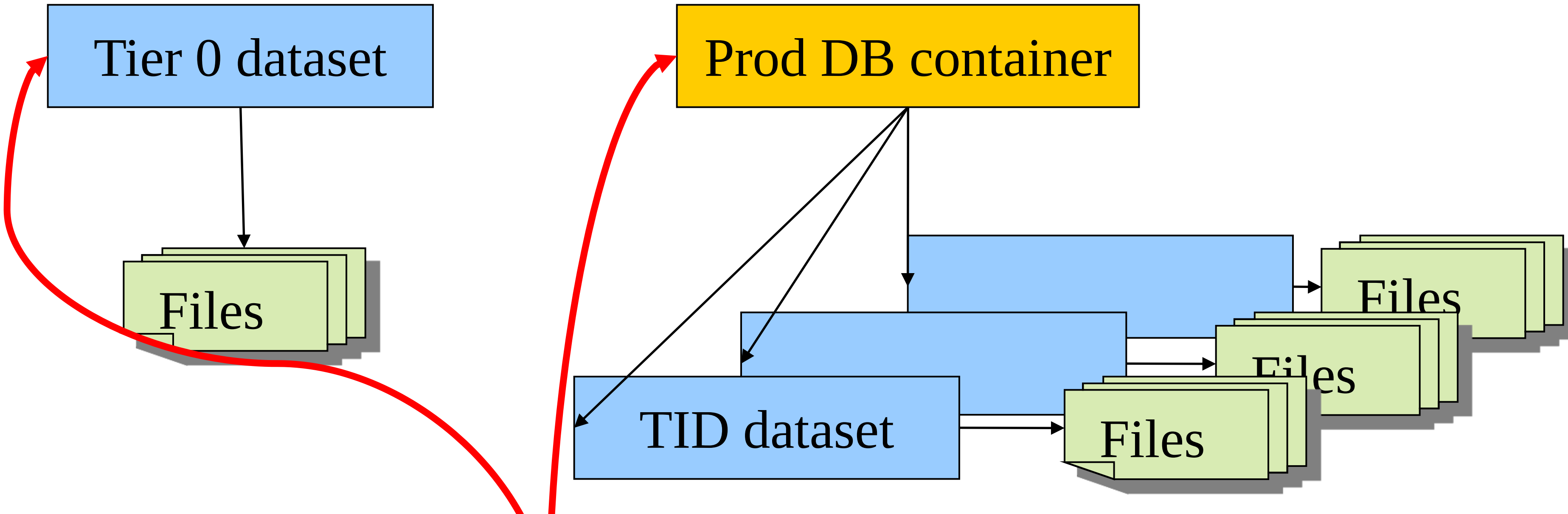
An introduction to datasets

- An ATLAS dataset is either
  - A number of files, output of ATLAS DAQ, and managed by Tier 0. (RAW data and fast reconstruction)
  - A number of files produced by a production system tasks. Each task makes a set of datasets. (MC and reprocessing of real data)
    - The primary datasets are identified by their TASK NUMBER. ("TID")

CONTAINER datasets

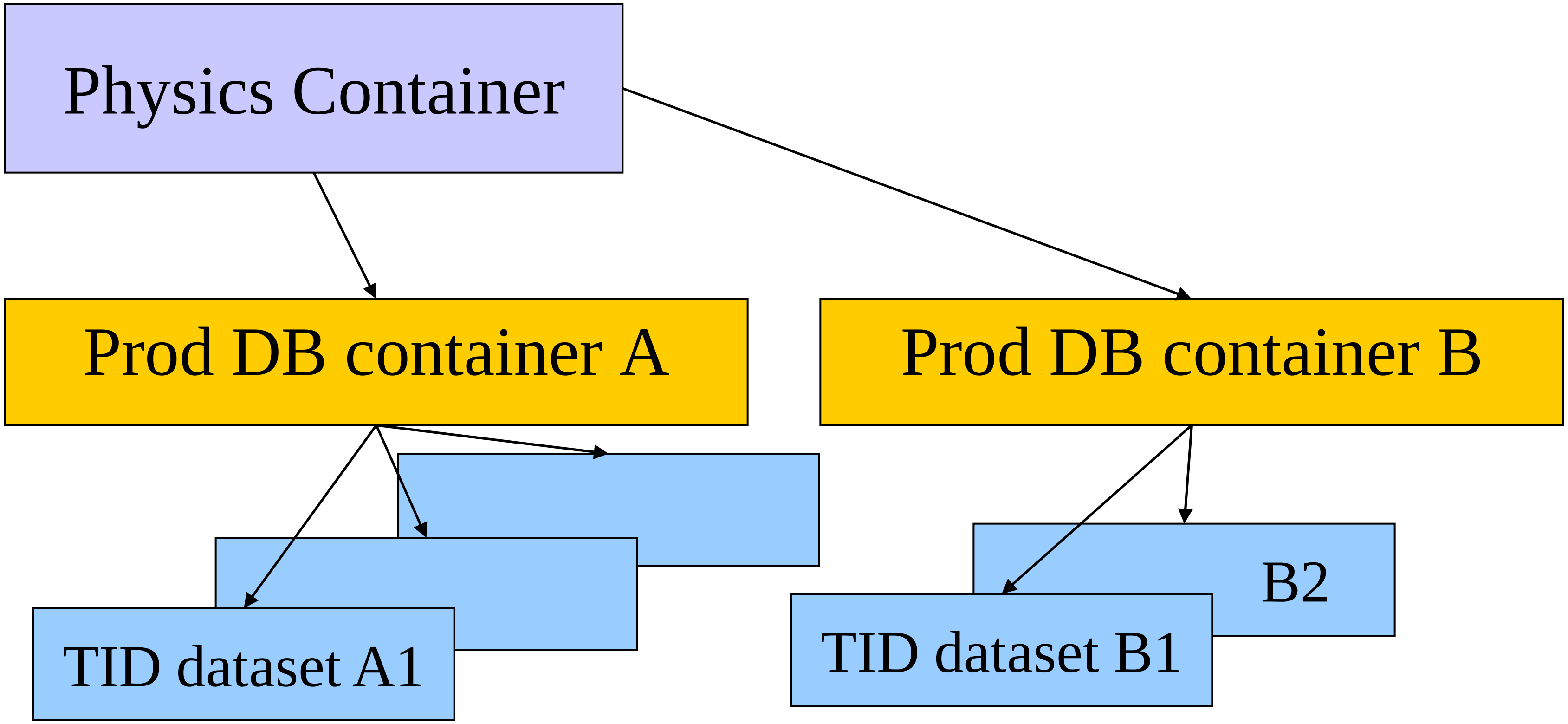
- One level of HIERARCHY (CONTAINER datasets) has been introduced to facilitate data management so that basic datasets do not become too large (>10000 files).
- There are basically TWO sorts of containers
  - A container of TID datasets. Containers group together datasets made by several production tasks with the same configuration, but different numbers of events. (Typically used for Monte Carlo simulation)
  - A Physics Container. A selection of primary containers prepared by Data preparation coordination, based on "good run lists". (An EMULATION of a second level of hierarchy, typically used to group data from several related runs described as "periods".)
- N.B. Tier 0 does not use CONTAINER datasets

Datasets

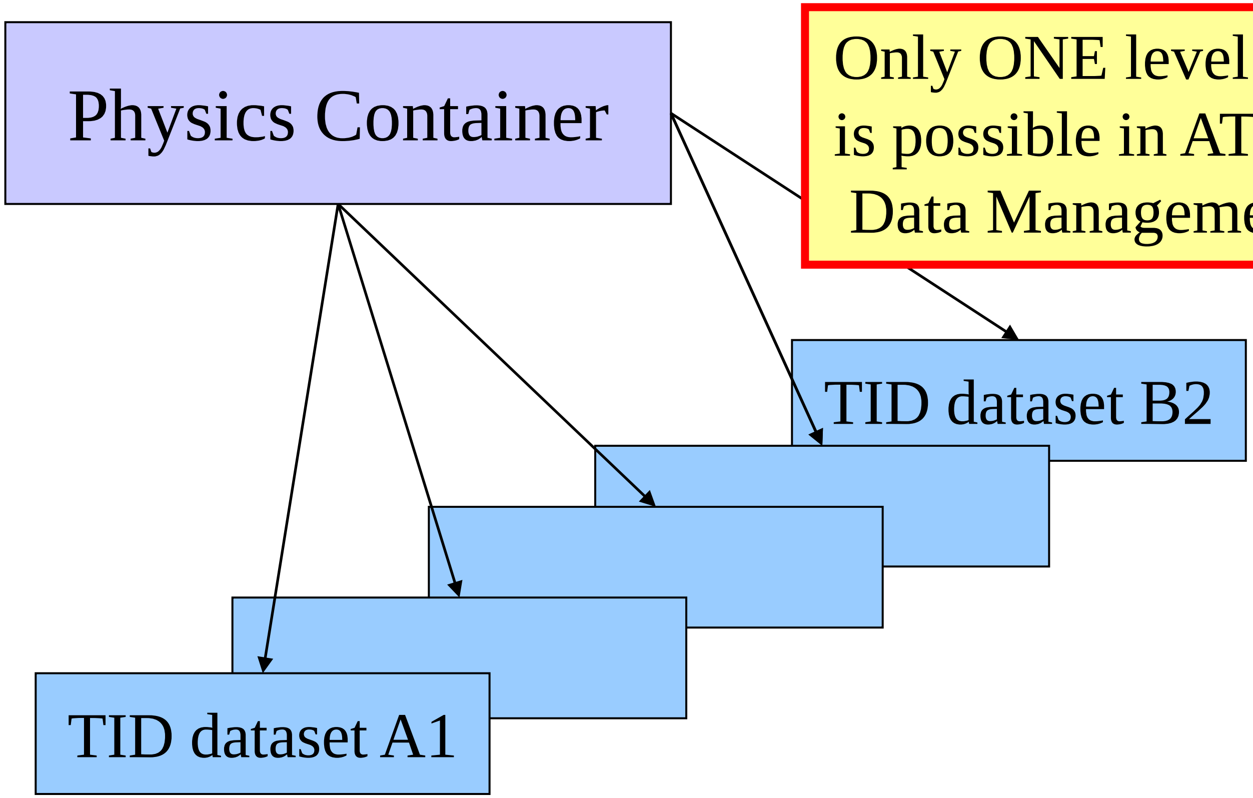


AMI catalogues these datasets ("logicalDatasetName") at the same level.

Physics Containers- logical view for reprocessed data



Physics Containers- physical view



Only ONE level of hierarchy is possible in ATLAS Distributed Data Management (DQ2.

N.B. Physics Containers of Tier 0 data are just containers of Tier0 primary datasets

ATLAS dataset nomenclature

- Datasets are grouped by TYPE.
  Each TYPE has a nomenclature template
  - REAL DATA :

    Project.runNumber.streamType.productionStep.dataType.version[/]
  - SIMULATED DATA

    Project.datasetNumber.physicsShort.productionStep.dataType.version[/]

    Where

    Project : dataNN_* or mcNN_*

    Production step : [daq|recon|simul|merge|……..]

    dataType : [RAW|AOD|ESD|HITS|…………]

    Version : A concatenation of configuration tags (AMI tags)

➔*nomenclature specification doc link*

mc09.106099.PhojetCdiff.evgen.EVNT.e456

A more recent example

mc10_7TeV.113218.pythia_minbias_truthJet60.evgen.EVNT.e608/

Task 170959
4 times 10000 jobs

mc10_7TeV.113218.pythia_minbias_truthJet60.evgen.EVNT.e608**_tid170959_00**

40000 files
1 output file/job

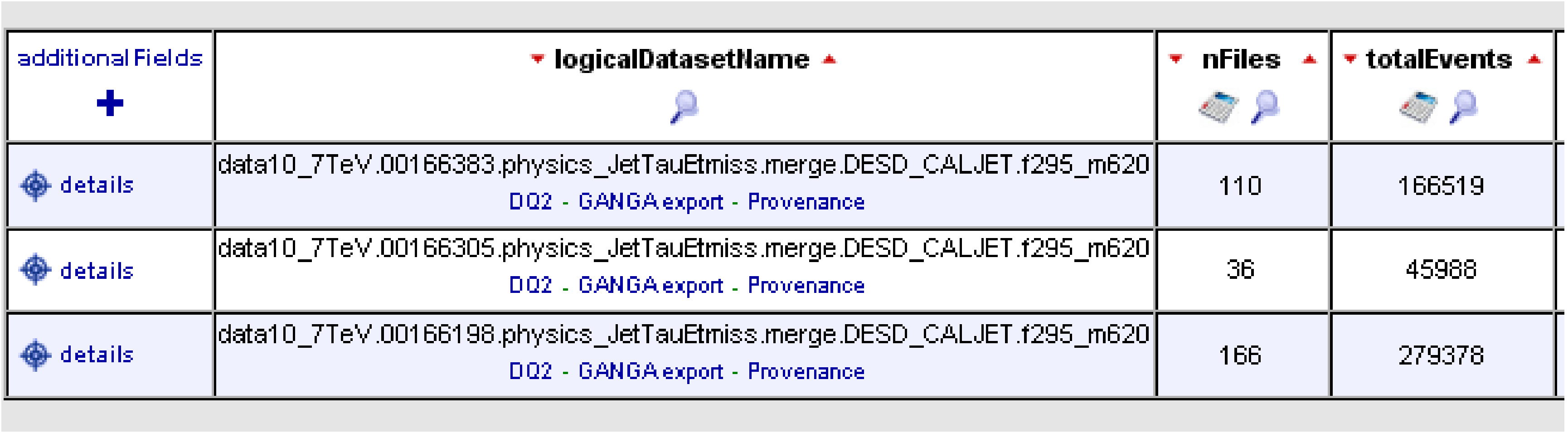
EVNT.170959._000001.pool.root.1

20 million events
500 /file

Physics Container example

data10_7TeV.periodG6.physics_JetTauEtmiss.PhysCont.DESD_CALJET.t0pro04_v01[/]

contains 3 TIER 0 datasets



additional Fields +	logicalDatasetName	nFiles	totalEvents
details	data10_7TeV.00166383.physics_JetTauEtmiss.merge.DESD_CALJET.f295_m620 DQ2 - GANGA export - Provenance	110	166519
details	data10_7TeV.00166305.physics_JetTauEtmiss.merge.DESD_CALJET.f295_m620 DQ2 - GANGA export - Provenance	36	45988
details	data10_7TeV.00166198.physics_JetTauEtmiss.merge.DESD_CALJET.f295_m620 DQ2 - GANGA export - Provenance	166	279378

Points to remember

- The ATLAS distributed data management (DDM) introduced containers because if one simply groups together all files with the same physics (but made by different production tasks) together the number of files to transfer from site to site becomes too large to handle.
- Tier 0 does not use "containers".
- Physicists doing analysis on official datasets should normally let the infrastructure take care of the container/not container problem.
- DQ2 (the DDM tool) requires the slash. AMI tries to keep this transparent for users.

About AMI

- Dataset Discovery - Means finding the names of valid datasets to use in your analysis.
- ATLAS Metadata Interface.
  - A generic cataloging framework – used in ATLAS for dataset discovery (and also Tag Collector + one or two other things)
  - Portal page. http://ami.in2p3.fr/ (Full Tutorial of AMI takes ~ 90 minutes)
  - Deployed at CCIN2P3 (French Tier 1).
  - Dedicated ORACLE cluster.
  - Two tomcat servers, apache front end with load balancing.

Things people say.

- *“[AMI] has an impressively complete information content within a somewhat complex user interface, which is by part due to the quantity of the available information.”*
- *“Thanks for AMI, it is a really great tool to help find data sets. I noticed one thing just now that should be fixed though.“*
- *“I browsed AMI, but I couldn't find anything of use there …”*
- *“AMI is one of the more user-friendly ATLAS products! “*
- *“..it's been extremely frustrating ……. why can’t it be more like GOOGLE?”*

Please retain that there are several ways of getting to the same information, that we cannot invent information, and that we are (almost!) always pleased to receive comments, complaints and suggestions.

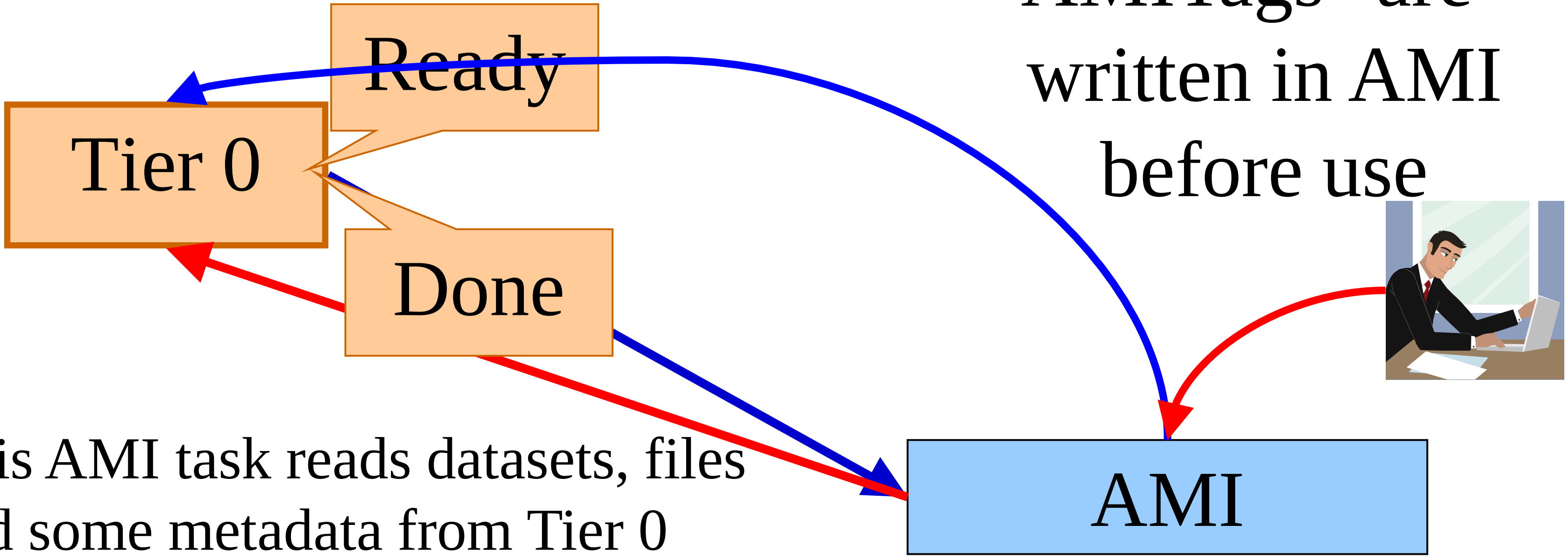
AMI is a MEDIATOR interface

- *"a software module that exploits encoded knowledge about some sets or subsets of data to create information for a higher layer of applications"*. (Gio Wiederhold, Mediators in the architecture of future information systems, *IEEE Computer Magazine*, Vol 25, No3, p38-49 March 1993)
- To put it another way, a mediator is an application which puts some domain expertise between the user and a group of data sources, so that information coming from these different sources can be aggregated.

Where does AMI get its data?

- **Real data :** From the Tier 0 : DAQ data, and first reconstruction is registered in AMI < 5 minutes after it is declared “ready” (both datasets and files).
- **Monte Carlo and reprocessing.**
  - From the Task Request DB : Tasks, dataset names, MC and reprocessing configuration tags ("AMI tags")
  - From the production DB : Finished tasks – files and metadata.
- **From physicists with AMI writer role.**
  - M.C. GEN datasets
  - MC Dataset number info, physics group owner,…
  - Corrected cross sections and comments. (coordinated by Borut Kersevan and Claire Gwenlan.)
  - Tier0 configuration tags (AMI tags for real data.)
  - Group datasets not made in the production system
  - ….

Tier 0



Ready

Tier 0

Done

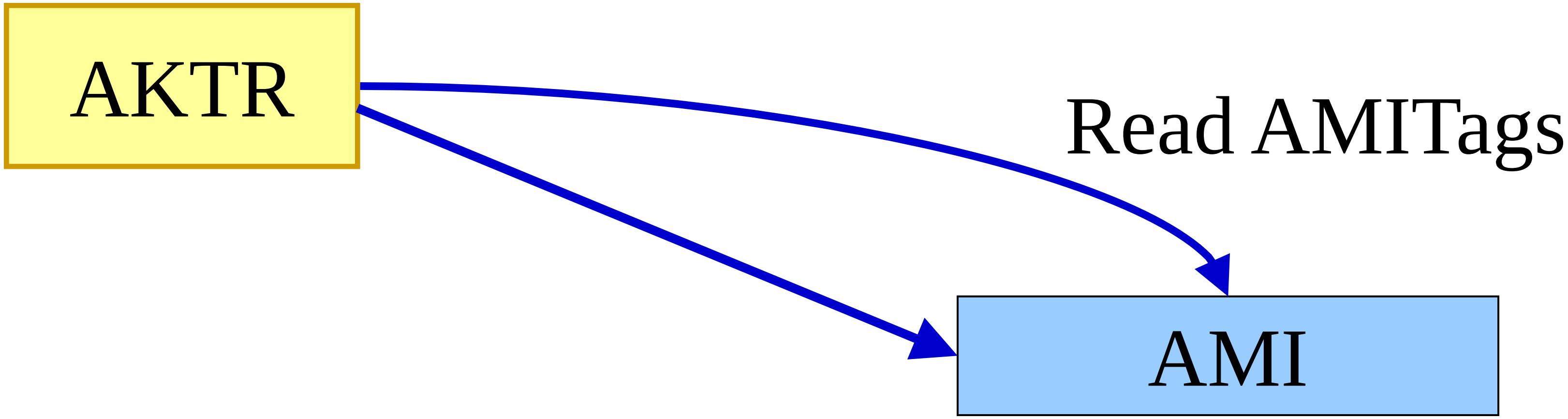
AMI

"AMITags" are written in AMI before use

This AMI task reads datasets, files and some metadata from Tier 0 using a semaphore mechanism.

Simple and efficient – AMI only get datasets when they are completed.

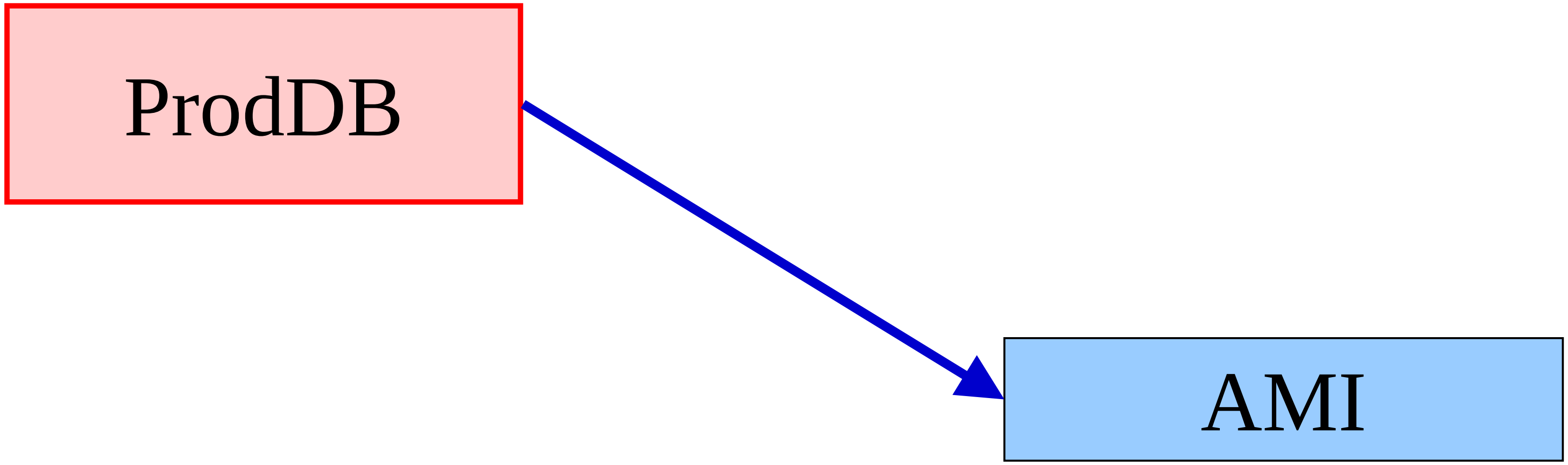
AKTR (Task Request)



1. Reads new tasks, once they are no longer pending, if they are not aborted before submission.
2. Reads provenance,
3. Reads updates to a bad status.
4. Reads MC and reprocessing AMITags.

N.B. Does NOT read finished/done status here.

ProdDB



1. Follows RUNNING status of tasks.
2. When FINISHED reads metadata.xml for all jobs of the task.

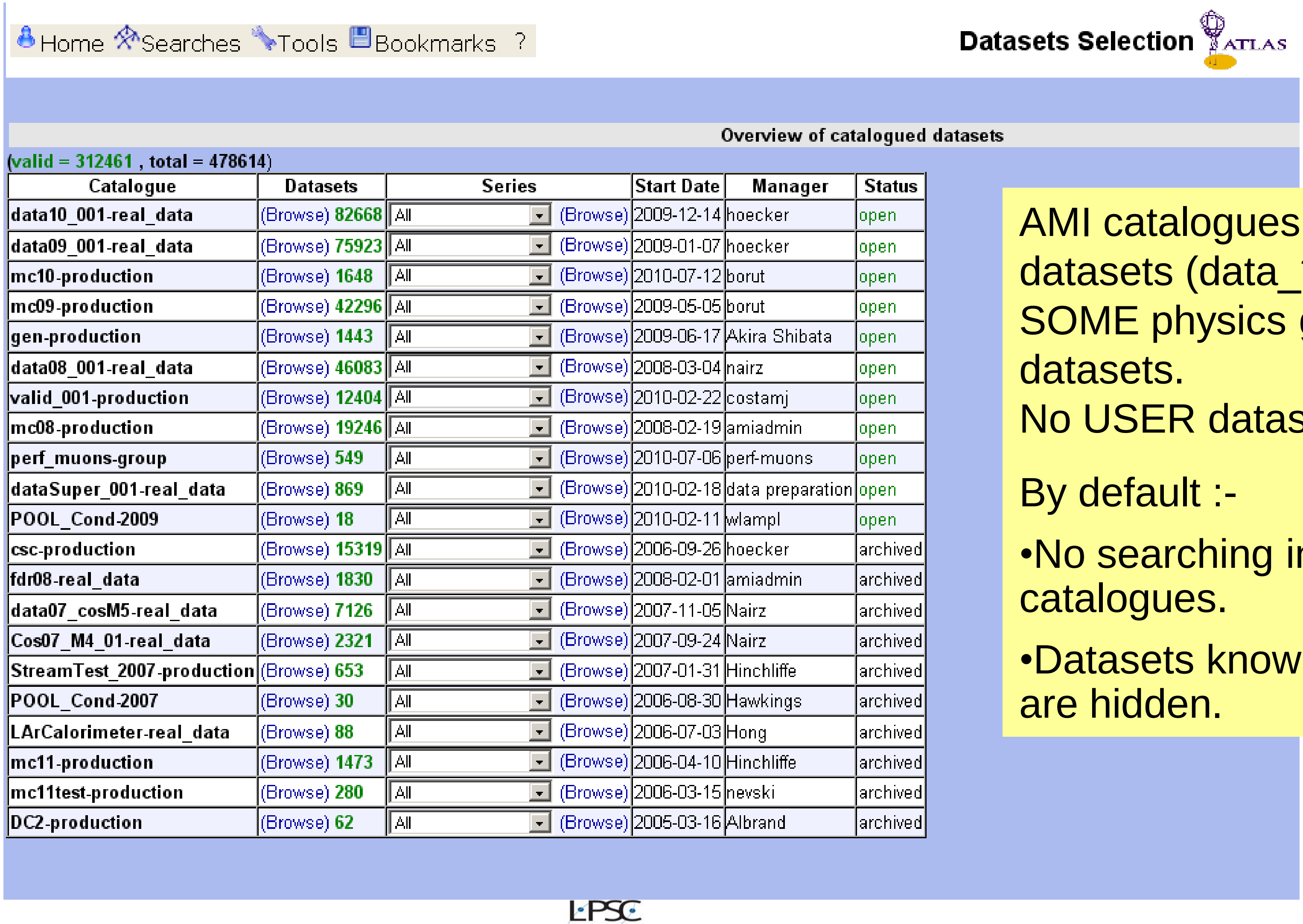
Overview page :

- Notice that not all datasets are VALID.
  - Invalid because all tasks failed, declared bad, or deleted.
  - Invalid datasets are hidden by default.
  - More info about dataset states in AMI
- Project tags are controlled, for official datasets; they follow the nomenclature specification and are declared by data preparation coordination. (data09_7TeV for example)
  - If you know the project tag this is perhaps the simplest way to get to the list.
- Note: AMI searches over all non-archived catalogues in parallel. Archived catalogues can be searched on request. Each catalogue can have a different schema.

***Tutorial** link

http://ami.in2p3.fr/opencms/opencms/AMI/www/Tutorial/Simple_search_interface.html

OVERVIEW PAGE



Datasets Selection

Overview of catalogued datasets

(valid = 312461 , total = 478614)

Catalogue	Datasets	Series		Start Date	Manager	Status
data10_001-real_data	(Browse) 82668	All	(Browse)	2009-12-14	hoecker	open
data09_001-real_data	(Browse) 75923	All	(Browse)	2009-01-07	hoecker	open
mc10-production	(Browse) 1648	All	(Browse)	2010-07-12	borut	open
mc09-production	(Browse) 42296	All	(Browse)	2009-05-05	borut	open
gen-production	(Browse) 1443	All	(Browse)	2009-06-17	Akira Shibata	open
data08_001-real_data	(Browse) 46083	All	(Browse)	2008-03-04	nairz	open
valid_001-production	(Browse) 12404	All	(Browse)	2010-02-22	costamj	open
mc08-production	(Browse) 19246	All	(Browse)	2008-02-19	amiadmin	open
perf_muons-group	(Browse) 549	All	(Browse)	2010-07-06	perf-muons	open
dataSuper_001-real_data	(Browse) 869	All	(Browse)	2010-02-18	data preparation	open
POOL_Cond-2009	(Browse) 18	All	(Browse)	2010-02-11	wlampl	open
csc-production	(Browse) 15319	All	(Browse)	2006-09-26	hoecker	archived
fdr08-real_data	(Browse) 1830	All	(Browse)	2008-02-01	amiadmin	archived
data07_cosM5-real_data	(Browse) 7126	All	(Browse)	2007-11-05	Nairz	archived
Cos07_M4_01-real_data	(Browse) 2321	All	(Browse)	2007-09-24	Nairz	archived
StreamTest_2007-production	(Browse) 653	All	(Browse)	2007-01-31	Hinchliffe	archived
POOL_Cond-2007	(Browse) 30	All	(Browse)	2006-08-30	Hawkings	archived
LArCalorimeter-real_data	(Browse) 88	All	(Browse)	2006-07-03	Hong	archived
mc11-production	(Browse) 1473	All	(Browse)	2006-04-10	Hinchliffe	archived
mc11test-production	(Browse) 280	All	(Browse)	2006-03-15	nevski	archived
DC2-production	(Browse) 62	All	(Browse)	2005-03-16	Albrand	archived

AMI catalogues OFFICIAL datasets (data_*, mc_*, and SOME physics group datasets.

No USER datasets as yet.

By default :-

- No searching in "archived" catalogues.
- Datasets known to be bad are hidden.

Configuration Tags (also known as AMI tags)

- A concatenation of configurations for successive processes.
- Example: **e466_s667_s668_d258_r1026_r1051** (last field of dataset name)

  e466 → event generation parameters

  s667, s668 → simulation parameters (simul.HITS, merge.HITS)

  d258 → digitization

  r1026,r1051 → reconstruction/ reprocessing parameters

- Interpretation of Config tags http://ami.in2p3.fr/opencms/opencms/AMI/www/ReferenceTables/
- Searching starting from the Config Tag. http://ami.in2p3.fr/opencms/opencms/AMI/www/Tutorial/ConfigTags
- Comparing tags (try r1026 and r1051) from the simple search page.

AMI Accounts

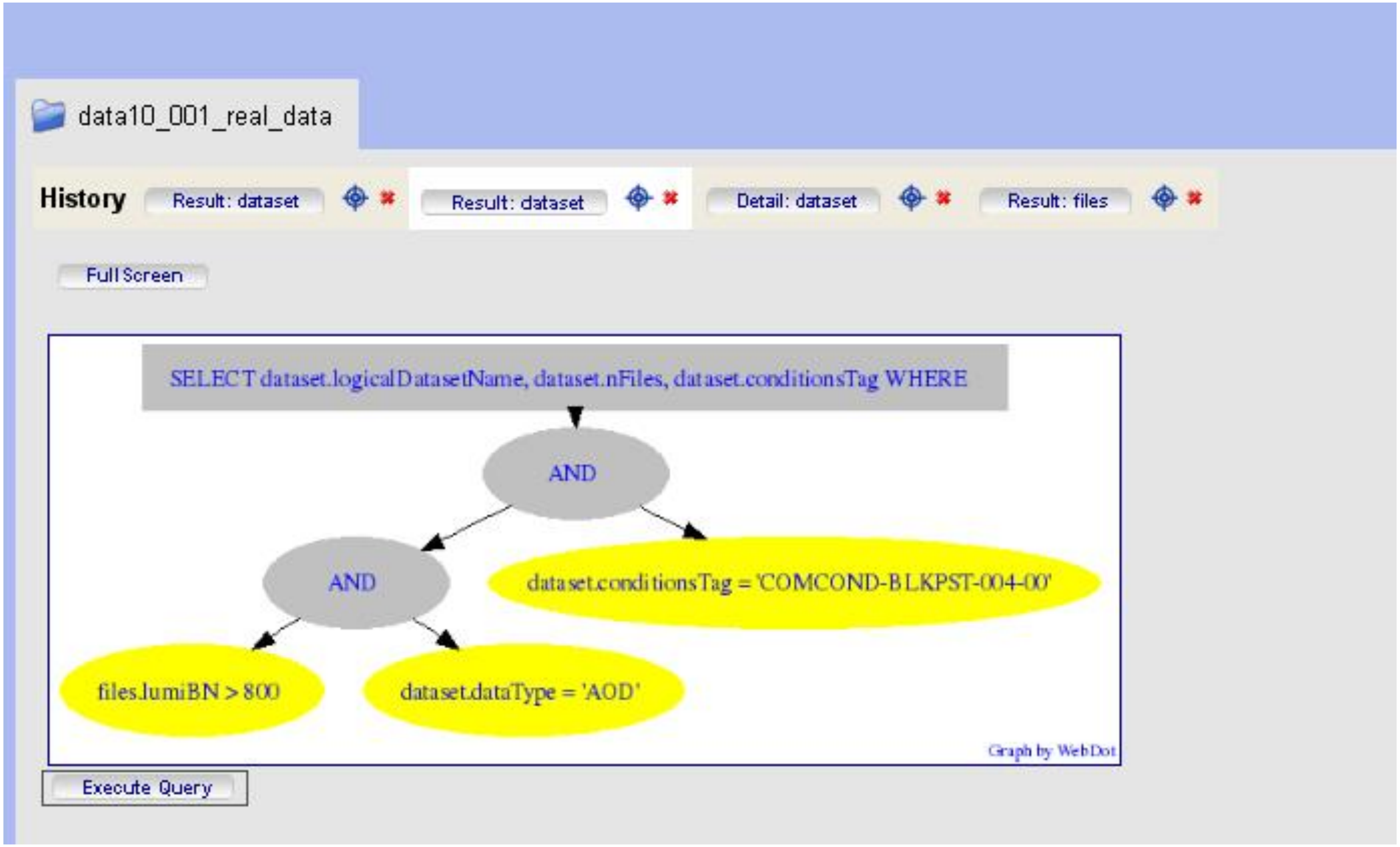
- Logging on to AMI.
  - In general you do not need to log on to read (at the moment)
  - You can make an AMI account to access a personal page.
  - You must log on for any writing operation.
  - Once you log on to AMI you can make bookmarks.
  - Tutorial link : http://ami.in2p3.fr/opencms/opencms/AMI/www/Tutorial/Other_AMI_basic_functionalities.html

Other stuff

- pyAMI. Everything in AMI can be obtained from the python client.
  http://ami.in2p3.fr/opencms/opencms/AMI/www/Client/pyAMISecure_and_cmt

  http://ami.in2p3.fr/opencms/opencms/AMI/www/Client/pyAMIUserGuide.pdf

- Ad Hoc queries. (For really advanced users!)
  http://ami.in2p3.fr/opencms/opencms/AMI/www/Tutorial/Refine_the_search.html
- Example: Which AOD datasets have more than 800 lumi blocks and used conditions Tag COMCOND-BLKPST-004-00 ?



SELECT dataset.logicalDatasetName, dataset.nFiles, dataset.conditionsTag WHERE

AND

AND

dataset.conditionsTag = 'COMCOND-BLKPST-004-00'

files.lumiBN > 800

dataset.dataType = 'AOD'

On Thursday

- "Complex data searches"
- Exercises in using AMI
- http://ami.in2p3.fr/opencms/opencms/AMI/www/Tutorial/