

# *Event Index Core*

## *(Hadoop & HBase)*



- Import
- Physical Containers
- Trigger Info
- Datasets Overlap
- Journal
- Important Missing Pieces

*Andrea Favareto (testing)*

*Julius Hrivnac*

*Fedor Prokoshin (trigger)*

*Rainer Toebbicke (infrastructure, HBase EL)*

*Ruijun Yuan (graphics)*



# Import

- Runs smoothly
  - 500-700 imports per day
  - EI datasets imported quickly
  - Some MC datasets imports is very slow due to their size
    - x\*100000 MapFiles per dataset
- Upload to HBase runs twice a day (and lasts several minutes)
  - Can be run more often
- Temporary files (\*.2) cleaned automatically
- Nightly checks for various inconsistencies
  - Then found problems handled by hand
    - Currently 4 datasets with different nevents in EI and AMI
- Hourly checks for duplicated events
  - Almost 100 already found, all EI15.1\*DAOD or MC15.1
  - Tagged in Catalog as multi:<multiplicity>x<number>
    - catalog -query 'id:MC15.1 multi:2' -filter 'id multi'
  - Should be fixed ? (easy)
- aiatlas016 had problems (do anyone know why ?)
  - All services transparently moved to aiatlas054 during that period

## IMPORTED

=====

EI	2009	526
EI	2010	4518
EI	2011	3944
EI	2012	4400
EI	2013	525
EI	2014	492
EI	2015	22759
EI	2016	1721
MC	2009	0
MC	2010	0
MC	2011	0
MC	2012	3
MC	2013	0
MC	2014	0
MC	2015	36692
MC	2016	0



# *Physical Containers (Event Lookup)*

- Physical containers introduced (in Catalog)
  - As \*.0 TagSets (TagSet = a set of TagFiles)
- Existing containers created
  - Most of them are incomplete (missing:<n> tag in Catalog records that)
- Unclear control flow (who/when/how records new Containers)
- Available to EL soon
  
- EL -indexer is no more equivalent to EL -simple/-rich
  - Indexer implements agreed strategies (for AMI tags, etc...)
  - Simple/Rich gives always everything (from HDFS TagFiles)
    - Useful for more detailed studies in case of problems



# Trigger Info

- Graphical output prototype
  - In 1/10000 scale
- TI is now available as Python client and direct URL (the same syntax as for EL)

## Event Index

- [Global Help](#)
- [Catalog](#)
- [Event Index \(Expert Mode\)](#)
- [Event Lookup](#)
- [Trigger Info](#)
- [Bookmarks](#)
- [System Journal \(for admins\)](#)



## Trigger Info

**-legend** Year Projets Stream Name Prod Step Data Type Version Run Number

**-query** E15.1 data15\_13TeV physics\_Main merge AOD t641\_m151 00284213

**-tlevel** ☐ L1 ☐ L2 ☒ EF or HLT ☐ all

**-tdecision** ☒ L1 AfterVeto or L2/EF/HLT Physics ☐ L1 AfterPrescale or L2/EF/HLT PassThrough ☐ L1 BeforePrescale or L2/EF/HLT Resurrected

**-toverlap** HLT\_e120\_loose,HLT\_2e17\_vloose,HLT\_2g50\_loose,HLT\_2j55\_bloose,L14J20.0ETA4 (disables -tlevel and -tdecision)

**-email** (implies asynchronous execution)

**-tpercent** ☐ Complete Result ☒ Percent of overlaps

Search Reset

Overlaps	HLT_e120	HLT_2e17	HLT_2g50	HLT_2j55	HLT_g120
HLT_e120	10000	1.0	2.0	1.0	86.0
HLT_2e17	1.0	10000	3.0	0.0	1.0
HLT_2g50	2.0	3.0	10000	0.0	14.0
HLT_2j55	1.0	0.0	0.0	10000	5.0
HLT_g120	86.0	1.0	14.0	5.0	10000



# Datasets Overlap

- Functionality available since a long time:
- *Number of overlapping events between two datasets*  
= *number of duplicated events in merged dataset*
  - There should be no duplicated events in either dataset
- Uses aux M/R job
- Use -email option as calculation can be slow for big datasets
  - => runs offline and sends results in email
- The API will be generalised to accept more datasets in one task.
- More user-friendly wrapper & GUI will be provided
- Should start systematic creation of overlap matrix for new / old datasets (could start immediately)  
?
  - Recorded (in Catalog) ?
  - Available from Journal anyway (once executed).

```
$ ei -query
id:EI15.1.data15_13TeV.physics_Main.merge.AOD.f643_m1518.00283429,EI15.1.data15_13TeV.physics_Main.merge.DAOD_EGAM3.f643_m1518_p2425.00283429
-mr true -aux net.hep.atlas.Database.EIHadoop.Accessor.Aux.Dupl
...
All 497835 results are available from TagFiles: [EICache/2016.06.14.10.30.39.511/hrivnac]

$ catalog -query
id:EI15.1.data15_13TeV.physics_Main.merge.AOD.f643_m1518.00283429,EI15.1.data15_13TeV.physics_Main.merge.DAOD_EGAM3.f643_m1518_p2425.00283429
-filter 'id nevents'
...
1 : id = EI15.1.data15_13TeV.physics_Main.merge.AOD.f643_m1518.00283429, nevents = 79569604,
2 : id = EI15.1.data15_13TeV.physics_Main.merge.DAOD_EGAM3.f643_m1518_p2425.00283429, nevents = 497835,
```

0.63%



# Journal

- Free text search possible for action, result and comment
- A lot of information already in Journal (& Catalog) - all results are there:
  - Can be exploited

## Event Index

- [Global Help](#)
- [Catalog](#)
- [Event Index \(Expert Mode\)](#)
- [Event Lookup](#)
- [Trigger Info](#)
- [Bookmarks](#)
- [System Journal \(for admins\)](#)

## Journal

start	02:34:26 14/Jun/2016	actor	<input type="radio"/> Catalog (modifying)	rc <input type="radio"/> 0 <input type="radio"/> -1 <input checked="" type="radio"/> *
stop	10:34:26 14/Jun/2016	<input type="radio"/> Event Index		
periodic refresh	<input type="checkbox"/>	<input type="radio"/> Eventy Lookup	<input type="radio"/> Trigger Info	
action		<input type="radio"/> Importer	<input checked="" type="radio"/> *	
result	e4917_s2860_r7856_r7			
comment				
<input type="button" value="Search"/> <input type="button" value="Reset"/>				

Entries of , e4917\_s2860\_r7856\_r7676 in result between 02:34:26 14/Jun/2016 and 10:34:26 14/Jun/2016:

actor ▲▼	rc ▲▼	time ▲▼	timestamp ▲▼	action ▲▼	result ▲▼	comment ▲▼	id ▲▼
EICLI	0	51	03:33:32.669 14/Jun/2016	{-outname=MC15.1.mc15_5TeV.410013.PowhegPythiaEvtG... more	Info: 50000 events found, result MC15.1.mc15_5TeV... more		<a href="#">2016.06.14.03.33.32.569/atlevind</a>
CatalogCLI	0	2	03:33:37.496 14/Jun/2016	{-modify=projectmc15_5TeV runNumber:410013 stream... more	Info: 2s spend Result: 1 : TagFile(id: MC15.1.mc1... more		<a href="#">2016.06.14.03.33.37.398/atlevind</a>



# *Important Missing Pieces*

- Commands with Shibboleth protection
  - Currently only Graphical GUI and direct URL are protected and available outside CERN
  - Prevents updates of CVMFS distribution !
- Dataflow logistics
  - Updating/removing datasets
  - Synchronising with AMI
    - AMI status changes too often
    - Maybe should be checked for each command
      - That would make it slower
- Import of BIG MC datasets
- Update TWiki (some obsolete info)
  - Up-to-date info is always available via CLI help
    - and TagConvertor Web Home
    - and Web Service GUI



# Info

**Web Service:** <https://atlas-event-index.cern.ch/ElHadoop>

**Documentation & Distribution:** <https://atlas-event-index.cern.ch/doc>

**Twiki:** <https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/EventIndex>

**Twiki Tutorial:** <https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/EventIndexTutorial>

**Sources:** `svn+ssh://svn.cern.ch/repos/atlasoff/Database/TAGHadoop/TagConvertor`

**AFS:** `/afs/cern.ch/sw/lcg/external/Java/TagConvertor`

**CVMFS:** `$ lsetup eiclient`