Data Store and Query System

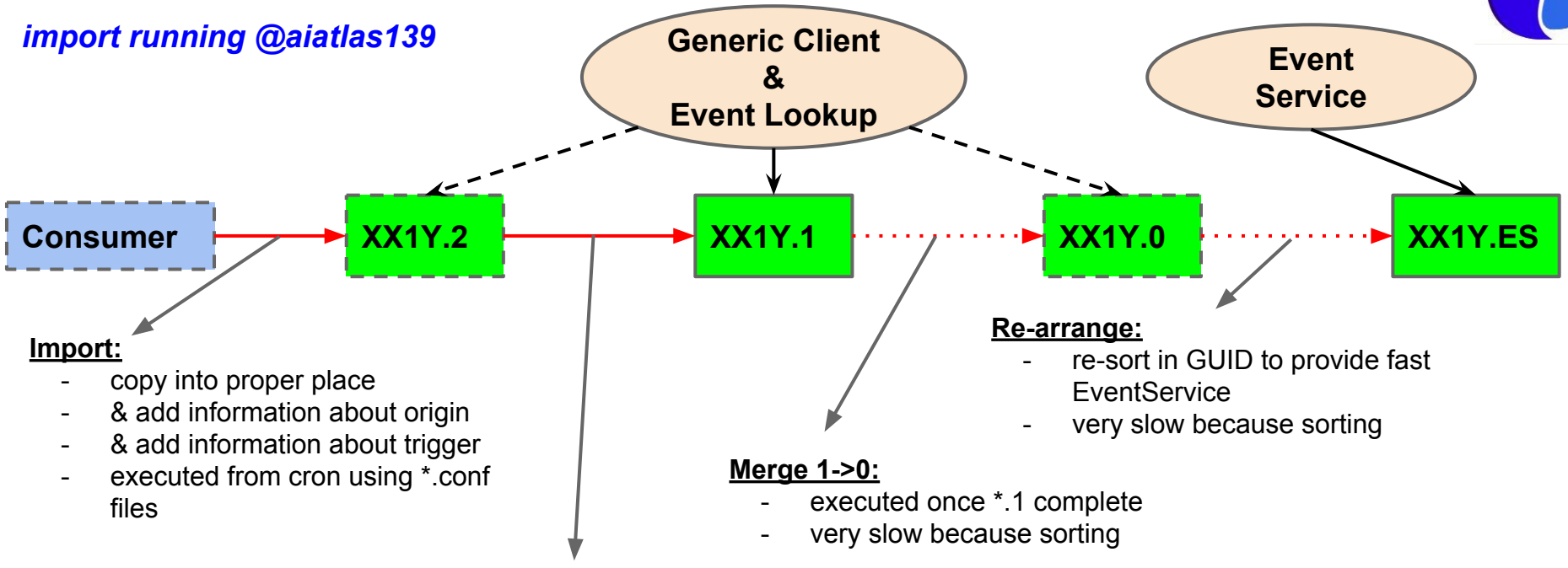
- Data Organisation & Derivation
- Data Access
- Data Verification & Correction
- Import Status
- Problems

latest version: ***1.15.0***

J.Hrivnac, EI WS, 31March15, CERN
+ Justin, Fedor, Rainer, Andrea,...

Data Organisation & Derivation (Import)

import running @aiatlas139



Import:
- copy into proper place
- & add information about origin
- & add information about trigger
- executed from cron using *.conf files

Re-arrange:
- re-sort in GUID to provide fast EventService
- very slow because sorting

Merge 1->0:
- executed once *.1 complete
- very slow because sorting

Merge 2->1:
- (now) executed immediately after Import

XX = EI | MC
Y = 0 | 1 | 2 | 3 | 4 | 5

EI15.2
EI15.1
EI15.0
EI15.ES
...

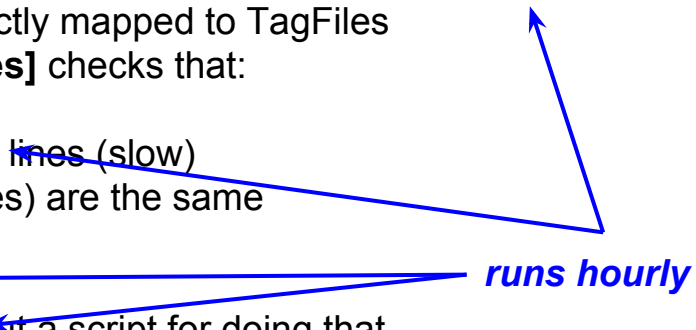
XX1Y.2/<project>.<runNumber>.<streamName>.<prodStep>.<dataType>.<version>/<guid>_<transid>_<consumerid>_<pandataskid>_<pandauserid>
XX1Y.1/<project>.<runNumber>.<streamName>.<prodStep>.<dataType>.<version>
XX1Y.0
XX1Y.ES

Data Access

- Catalog now supports several ways of searching:
  - exact, prefix, substring, regex (default is substring)
- Some attributes are available in several places:
  - path, Catalog attribute, entry field
- This allows variety of possibilities how to search:
  - **-query 'regex path:.*/EI10\.1/.*\.physics_CosmicCalo\..*\.f299_m639'**
  - **-query 'path:EI10.1 prodStep:physics_CosmicCalo version:f299_m639'**
  - **-query 'path:EI10.1'**
    -mr 'prodStep.equals("physics_CosmicCalo") && version.equals("f299_m639")'
- We will do benchmarks to choose the best way
  - the second possibility looks best, it's fast, transparent (= not regex) and position-independent
    - the "global attributes" (project, runNumber, streamName, prodStep, dataType, version) annotations are added to existing *.1 Catalog entries now, new imports will get them automatically
- Web Service GUI based on "global attributes" is under development

Data Verification & Correction

- Some problems during import:
  - Some TagFiles should be re-imported because modified/corrected
  - HDFS sometimes fails (HDFS timeout,...)
  - AFS ticket expires
  - Machine reboots
  - …
- Not fatal (except for merge to *.0, *.ES): failed data are re-imported after verification
- So data are verified after import:
  - **catalog -query <query> -verify [fix]** checks internal consistency of Catalog and dangling references
  - **importer … verif** checks that Consumer files are correctly mapped to TagFiles
  - **verificator [merge|mergelines|schema|files|attributes]** checks that:
    - merge: *.2 -> *.1 -> *.0 merged files correspond
    - mergelines: as merge, but checks also number of lines (slow)
    - schema: the schema in files collections (directories) are the same
    - files: non-referenced files
    - attributes: attributes propagation during merge
  - all commands either allow to fix the problems or write out a script for doing that
  - It would be great to have a definitive list of valid data from Consumer once a “year” is finished to perform final check



Import Status

IMPORTED

EI 2010 2	41447	MC 2010 2	0
EI 2010 1	2751	MC 2010 1	0
EI 2010 0	**1**	MC 2010 0	0
EI 2010 ES	**1**	MC 2010 ES	0
EI 2011 2	152274	MC 2011 2	0
EI 2011 1	3527	MC 2011 1	0
EI 2011 0	0	MC 2011 0	0
EI 2011 ES	0	MC 2011 ES	0
EI 2012 2	267451	MC 2012 2	40
EI 2012 1	4213	MC 2012 1	2
EI 2012 0	0	**MC 2012 0**	**1**
EI 2012 ES	0	**MC 2012 ES**	**1**
EI 2013 2	42677	MC 2013 2	0
EI 2013 1	92	MC 2013 1	0
EI 2013 0	0	MC 2013 0	0
EI 2013 ES	0	MC 2013 ES	0
EI 2014 2	6379	MC 2014 2	0
EI 2014 1	33	MC 2014 1	0
EI 2014 0	0	MC 2014 0	0
EI 2014 ES	0	MC 2014 ES	0
EI 2015 2	7968	MC 2015 2	0
EI 2015 1	167	MC 2015 1	0
EI 2015 0	**1**	MC 2015 0	0
EI 2015 ES	**1**	MC 2015 ES	0

WAITING

EI 2010	1553
EI 2011	0
EI 2012	134
EI 2013	405
EI 2014	342
EI 2015	469
MC 2010	0
MC 2011	0
MC 2012	0
MC 2013	0
MC 2014	0
MC 2015	0

UNMERGED

```
Missing TagFiles from EI10.1: 50
Missing TagFiles from EI10.2: 0
Missing TagFiles from EI11.1: 0
Missing TagFiles from EI11.2: 0
Missing TagFiles from EI12.1: 1
Missing TagFiles from EI12.2: 0
Missing TagFiles from EI13.1: 24
Missing TagFiles from EI13.2: 0
Missing TagFiles from EI14.1: 22
Missing TagFiles from EI14.2: 0
Missing TagFiles from EI15.1: 9
Missing TagFiles from EI15.2: 0
Missing TagFiles from MC10.1: 0
Missing TagFiles from MC10.2: 0
Missing TagFiles from MC11.1: 0
Missing TagFiles from MC11.2: 0
Missing TagFiles from MC12.1: 0
Missing TagFiles from MC12.2: 0
Missing TagFiles from MC13.1: 0
Missing TagFiles from MC13.2: 0
Missing TagFiles from MC14.1: 0
Missing TagFiles from MC14.2: 0
Missing TagFiles from MC15.1: 0
Missing TagFiles from MC15.2: 0
```

done

waiting

to be fixed

MISCATALOGED

```
============
1129570 INFO  (Catalog.Catalog                : 234) : Dangling entries:       1
1129570 INFO  (Catalog.Catalog                : 235) : Repeated signature:     0
1129570 INFO  (Catalog.Catalog                : 236) : Repeated path:          0
1129570 INFO  (Catalog.Catalog                : 237) : Missing  index entries: 0
```

UNCATALOGED

```
Missing from Catalog: EI10.1/data10_7TeV.00152441.physics_RNDM.merge.AOD.f239_m433
Missing from Catalog: EI10.1/data10_7TeV.00152490.express_express.merge.AOD.f241_m433
Missing from Catalog: EI10.1/data10_7TeV.00152490.physics_CosmicCalo.merge.AOD.f241_m433
...
```

from hourly diagnostics

Problems & Tasks

- Merging to *.0 and *.ES is very slow (days)
  - as it sorts billions of entries
  - growing probability of a problem during the job
  - will look at sorting algorithm to speed it up
    - via incremental merge
  - will also try combined query to *.0/*.ES with not-yet-included data from *.1
- Current structure with Hive/Pig can provide event-overlaps, someone should do it
  - Jack seems to have volunteered for that

Links

- Home:
  - http://atlas-event-index.cern.ch/doc *(atlas-event-index = aiatlas016)*
  - http://cern.ch/hrivnac/Activities/Packages/TagConvertor
- CLI:
  - **catalog, ei, inspector, importer, verificator**
- Web Service:
  - http://atlas-event-index.cern.ch/EIHadoop
  - ssh -L 8080:aiatlas016.cern.ch:80
  - java -jar EIHadoopCatalog.exe.jar *for remote Catalog CLI*
  - java -jar EIHadoopEI.exe.jar *for remote Event Index CLI*
  - java -jar EventServer.exe.jar *for remote Event Server CLI (not yet in CVMFS)*
  - java -jar EventLookup.exe.jar *for remote Event Lookup CLI (not yet CVMFS)*
  - runEventLookup.py
- SVN:
  - svn+ssh://svn.cern.ch/reps/atlasoff/Database/TAGHadoop/TagConvertor
- Ant targets:
  - ant -p *to get help on available targets*