Event Index Core

(Hadoop & HBase)

- Import
- Trigger Overlap
- Other News

Julius Hrivnac
Fedor Prokoshin
Grigorij Rybkin
Rainer Toebbicke
Ruijun Yuan

EI WS, 11 Dec 2017, CERN

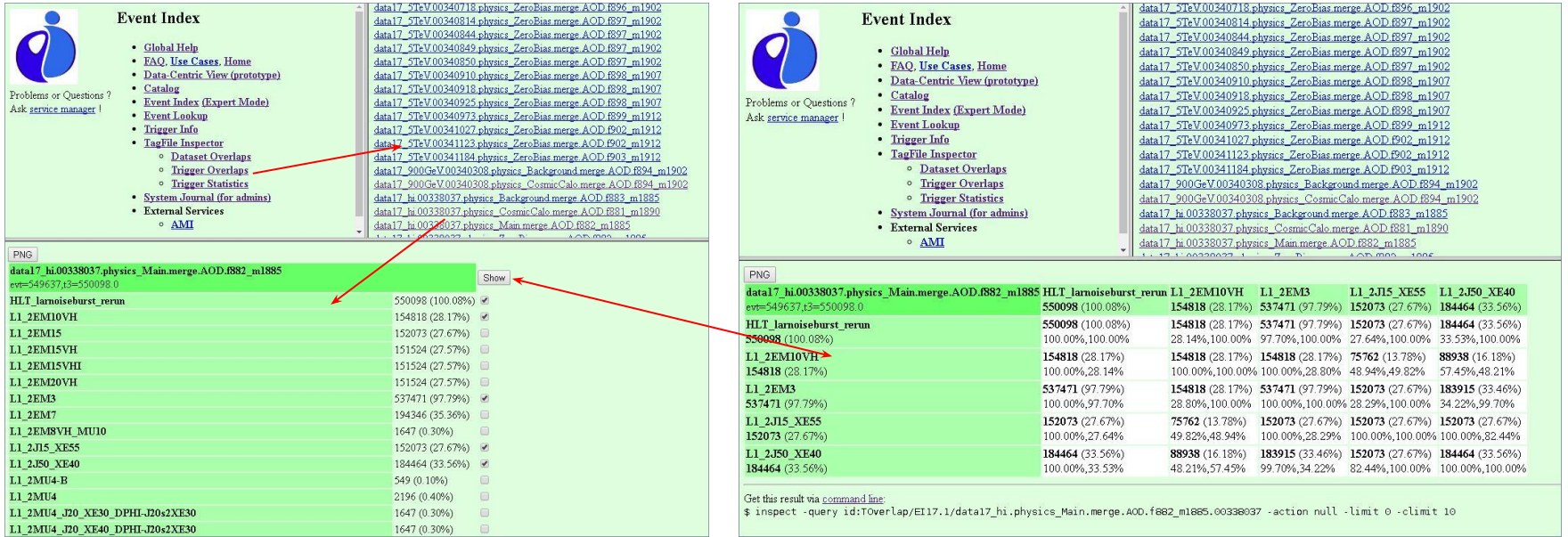
Import

- Very smooth and fast after migration to ObjectStore Consumer
  - Also because legacy importing procedure re-written
- The only (temporary) problem:
  - We haven't taken into account that new Consumer files are not sorted, which breaks 'ei -key' and 'el -api simple/rich'
  - Import has been fixed and already imported files have been sorted (by Grigorij)
- Import of 37 datasets have failed during cluster upgrade
  - Automatic re-import successful
- What about conf-directory OSDatasetsConf_2 ?
  - It contains a lot of PARTIAL conf files
- Import executed in several stages (change in one stage triggers re-execution of following stages)
  - Actual import from Consumer, sorting, indexing, adding trigger info, registering in Catalog - as datasets arrive, usually within minutes
  - Uploading to HBase for EL - 4x a day (may be slower depending on number of data)
  - Generating derived tables (DOverlaps, TOverlaps, TStats) - nightly
    - Old TagFiles processed with 100 per day, all will be done till end 2017
  - Testing (consistency, multi-events, EL,...) - nightly (EL postponed one day)
- All import steps
  - Recorded in Journal
  - Registered in Catalog (if relevant)
  - Mailed to atlevind

```
IMPORTED EVENTS
===============
EI 2009        34 444 939
EI 2010     1 196 985 327
EI 2011     1 844 680 412
EI 2012     3 313 189 845
EI 2013       331 475 303
EI 2014       436 763 032
EI 2015    38 527 094 863 (was 37 875 871 406)
EI 2016    63 203 930 561 (was 57 524 958 034)
EI 2017    17 004 632 956 (was  5 948 137 819)
MC 2012           300 000
MC 2015    28 763 092 958
MC 2016     3 825 805 416 (was  1 571 035 156)
ALL          > 158 billion events
```

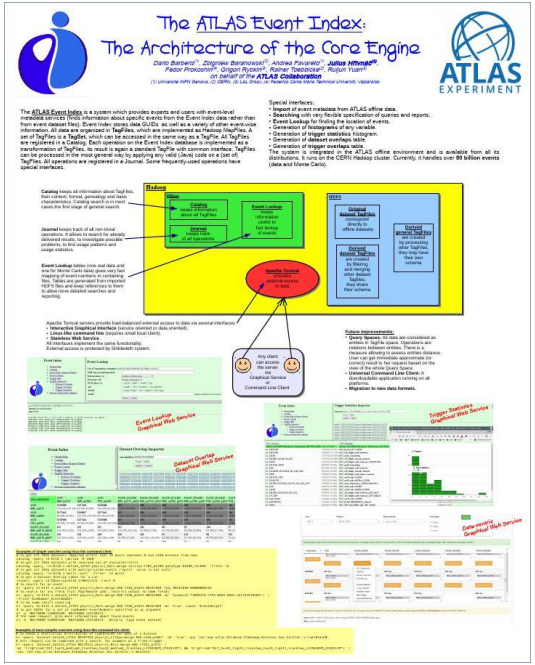
Trigger Overlap

- Generated from subset of events because full generation would take too long
  - For all newly imported datasets (and some old datasets)
  - Parallelism (M/R,...) doesn't help much because task can't be easily divided into independent subtasks (big common data)
  - Full statistics generation for subset of triggers is fast and can be done on request (via CLI or WS)
- Visualisation may be slow (as a lot of data should be parsed), esp. *the second arrow*
- Trigger Statistics is contained in Trigger Overlap (but Trigger Statistics table use all data)
- Working on it….



Other News

- All TagFiles are now using Record gzip compression
  - Gives factor 2x-3x, completely transparent to clients
  - More aggressive compression (5x-10x) would break some clients
  - Before only newly imported TagFiles were compressed, derived TagFiles were uncompressed
- New nightly tests
  - Mostly for EL: Just the same tests as PAnda, but without PAnda
- Many small fixes and improvements (Justin, Grigorij, Rainer, Fedor)
- Poster in ACAT/Seattle *(ATL-COM-SOFT-2017-073)*



Info

Web Service: https://atlas-event-index.cern.ch/EIHadoop

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

Frequently Asked Questions: https://atlas-event-index.cern.ch/doc/faq

Sources: svn+ssh://svn.cern.ch/reps/atlasoff/Database/TAGHadoop/TagConvertor

AFS: /afs/cern.ch/project/jps/reps/atlas-eies/TagConvertor

CVMFS: $ lsetup eiclient

EOS: /eos/atlas/atlascerngroupdisk/proj-evind/Results