The EventIndex DB setup in the Oracle RDBMS (EIO)

EventIndex workshop, Milan (Italy), June 2018

Elizabeth Gallas, Gancho Dimitrov, Petya Vasileva

EIO features
(reminder)

EIO Can perform efficiently all tasks it is designed for

- Event-picking in a fraction of a second
- Find duplication within each dataset
- Compute dataset overlaps
- Compute number of unique files per dataset
- Compute stats per lumiblock
- Cross-checks with AMI
- Associate AMI container to datasets
- Overwrite datasets in case of necessity
- Fast and efficient removal of datasets content (new)

EIO DB setup deals well with the large amount of entries

- As of 1st June 2018, the whole EIO catalog hosts **138 billion event records in 61371 datasets. Space usage: 2.8TB table segments, 2.5TB index segments.**
- From 1st Jan until 1st June 2018: **stored 13.4B event records within 4218 datasets.**

PROJECT	DATATYPE	DATASETS	BILLION_ROWS
data15_13TeV	AOD	13	0.02
data15_13TeV	DAOD	1445	1.35
data16_13TeV	AOD	18	0.75
data16_13TeV	DAOD	613	0.58
data17_13TeV	AOD	360	4.87
data17_13TeV	DAOD	1057	2.09
data17_5TeV	AOD	42	1.33
data18_13TeV	AOD	440	1.85
data18_13TeV	DAOD	228	0.57

EIO handling of special cases

- **Overwrite dataset content (automatic):**
  1st Jan - 1st June 2018: **1173 overwritten datasets having 2.3B event records.**

ACTION	PROJECT	DATATYPE	DATASETS	BILLION_ROWS
Overwrite dataset	data15_13TeV	AOD	13	0.02
Overwrite dataset	data15_13TeV	DAOD	1114	1.29
Overwrite dataset	data16_13TeV	AOD	3	0.11
Overwrite dataset	data16_13TeV	DAOD	13	0.26
Overwrite dataset	data17_13TeV	AOD	6	0.19
Overwrite dataset	data17_13TeV	DAOD	13	0.02
Overwrite dataset	data17_5TeV	AOD	11	0.44

- Introduced **improvement in the data loading PLSQL procedure** to recover automatically from deadlock situations. Could happen when the data loading procedure and the one responsible to overwrite existing EI datasets get into race condition on table's DDL action.

EIO handling of special cases (cont.)

- **Delete dataset content** (on request):
  1st Jan - 1st June 2018: **deleted 24 datasets having 81.4M event records.**

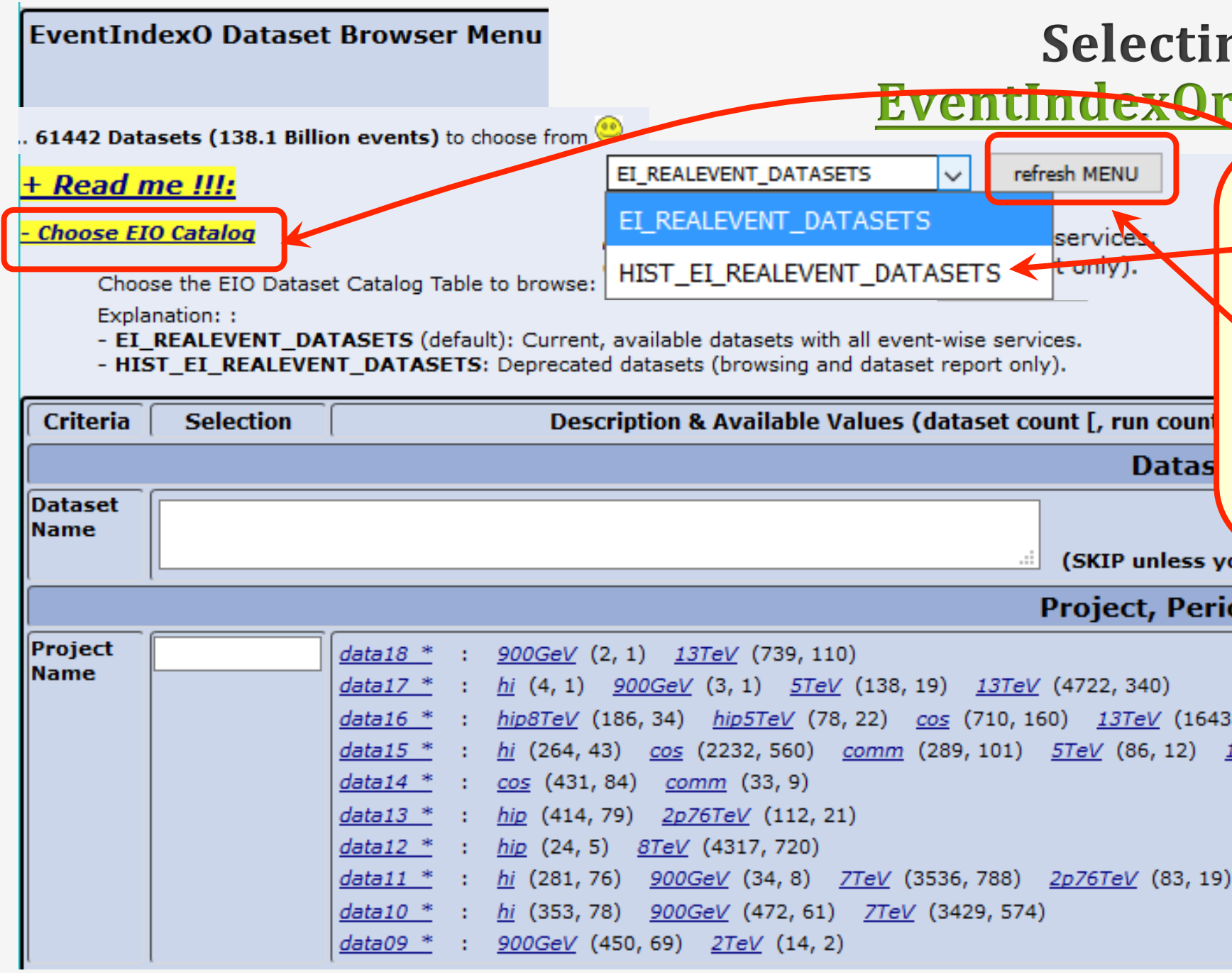
ACTION	PROJECT	DATATYPE	DATASETS	MILLION_ROWS
Delete dataset	data15_13TeV	DAOD	12	0.11
Delete dataset	data16_13TeV	AOD	3	79.93
Delete dataset	data16_13TeV	DAOD	7	0.04
Delete dataset	data16_valid	DAOD	1	0.14
Delete dataset	data17_13TeV	AOD	1	1.14

- **Fast and efficient operation via dedicated PLSQL procedure**: about 20 seconds execution time for 24 datasets.
  Important: leaves trace information by storing aside datasets definition.

Removal of EIO datasets

- Recent use cases prompting selective removal of datasets (as examples).
  Why ? Leaving them in the system had no benefit (and created some confusion)
  - March 2018: 3 Datasets with invalid Stream Names
  - April 2018: 21 Datasets with wrong Project Name for the Run number
    (they all had been TRASHED in AMI and deleted from DDM) → EIO Rank = 97
- We decided when we remove datasets (as above, or because they are 'replaced'):
  - We remove its events, duplicated events, overlaps, counts by LB, etc.
  - We keep dataset-level information (at the time of removal) separately
    - These are visible via the EIO Dataset Browser (next slide)
- Future Removal of datasets ? No pressing need at the moment …
  - Candidates ? Such as TRASHED or not found in AMI (have EIO Rank >= 94)
  - But we want to avoid removal of special (DAOD) or unique (AOD) datasets

Selecting Removed Datasets: EventIndexOracle Dataset Browser



Open "**Choose EIO Catalog**" section
In pull-down menu, choose:
HIST_EI_REALEVENT_DATASETS,
Click on "**refresh MENU**" button

The Browser will then switch to show the datasets which have been removed (**next slide**).

Removed Datasets: EventIndexOracle Dataset Browser

Form input set Dataset Table to 'HIST_EI_REALEVENT_DATASETS'

EventIndexO Dataset Browser Menu

3136 Datasets (12.7 Billion events) to choose from

Criteria	Selection	Description & Available Values (dataset count [, run count])
Dataset Name (or ID)		
Dataset Name		(SKIP unless you KNOW it) (DSID)
Project, Period, Run criteria		
Project Name		data17 * : 5TeV (11, 10) 13TeV (28, 23) data16 * : valid (1, 1) cos (2, 2) 13TeV (929, 236) data15 * : hi (2, 1) cos (21, 14) comm (3, 3) 5TeV (5, 5) 13TeV (2134, 146)
Period Name		Enter an ATLAS Data Period name (i.e. B6, or B or AllYear).
Run(s)		Enter one/more Real Data Run Numbers (current selection criteria: **441** runs) Example runs for DAOD overlap studies: data16_13TeV 300655 (June), 304008 (July), 311321 (Oct);
Other dataset name criteria		
Stream		subset * : HLT1 (1) physics * : ZeroBias (113) MinBias_lbcustom (1) MinBias (21) Main_lbcustom (1) Main (2500) Late (96) L1Calo (4) EnhancedBias (2) CosmicMuons (1) CosmicCalo (195) BphysDelayed (22) express * : express (109) debugrec * : hlt (68) debug * : all (2)

- Service Options:

refresh MENU | Dataset Report

- You are now 'browsing' the datasets in HIST_EI_REALEVENT_DATASETS Count of Datasets and Events – as shown
- The menu gives an overview of removed datasets:
  - All are data15, data16, data17
    - dataset and run counts per project name
  - Similarly, dataset count by Stream Name
- The "Service Options" are limited to
  - Refresh MENU
  - Start again (clear the form)
  - Dataset Report
- The Dataset Report will show (for datasets matching your input criteria)
  - The saved dataset-level metadata
  - Timestamp and reason for deletion

EIO Event Lookup: Improved multiple project warning

- A few cases recently where EventPicking failed because users input run/event lists with runs in more than one project
  - Panda doesn't like that – should warn users !
  - Here are the run/events from a recent case

EventIndexO EventLookup

+ Instructions

Run/Event List: The EventLookup service needs a list of the events you want in each Run or Dataset. Provide the event list in the textarea box below (one line per run event pair). A local file from your computer with your event list can be uploaded using this button: Upload Run/Event File or you can copy/paste your run/event list into the text box.
After a file upload, you can edit your event list in the textbox.
Example input Run/Event text file: *RealRun279984_4evnts*. Other examples in *Run/Eve Lists*

```
281411  2093114832
283429  3943405491
300571  1169486033
300800   865636609
```

Choose your Stream: physics_Main

Choose the input DataType Format: AOD

Choose the GUID type(s) to lookup: ☐ StreamAOD, ☐ StreamESD, ☑ StreamRAW,
+ more GUID types (127):

After inserting your run/event list, Click here: **LookupEvents**

EventIndexO Error LookupEvents

Data Type Format (dtype) : AOD
Stream/PhysicsShort (stream) : physics_Main

Input Summary:

- Input stream: **physics_Main**
- Input GUIDType: **StreamRAW**
- Input Run/Event text file has 4 lines of input.
- Max event # sought: 3,943,405,491
- ***+ Target Datasets (10):***

Error determining a common Project Name:

More than one project name found (2) !
Your input run numbers have different Project Names:
data15_13TeV, data16_13TeV.
All Runs in the input list should be in the SAME Project
--> it is a Panda requirement for Event Picking.

Check your project names for your requested Runs: *COMA Runs*
then split your requests by project name.

COMA Online Runs Report

Run Number (runs) : 281411, 283429, 300571, 300800

Project Run	StartTime *Events*	*Period*
data16_13TeV 300800	16-Jun-03 17:41 121448353	AllYear, B, B3
data16_13TeV 300571	16-May-31 04:09 137335031	AllYear, B, B2
data15_13TeV 283429	15-Oct-26 14:58 292037135	AllYear, J, J4
data15_13TeV 281411	15-Oct-11 14:00 176189901	AllYear, H, H3

Summary

- EventIndex Oracle is performing well: for the developers and the users
  - For the developers:
    - Efficient automated processes manage the data content with minimal maintenance
    - New: dataset removal procedure is available
      - The decision of what to remove remains with experts
    - New: interfaces enabled to show the datasets which have been removed and why
  - For the users:
    - Many event-wise services (as mentioned previously) are available with
      - Quick response time
      - Helpful messages and warnings
        - New: improved warnings in EventLookup about requests for events in multiple projects.
  - For all:
    - The system is well provisioned to handle more data.

THANK YOU!