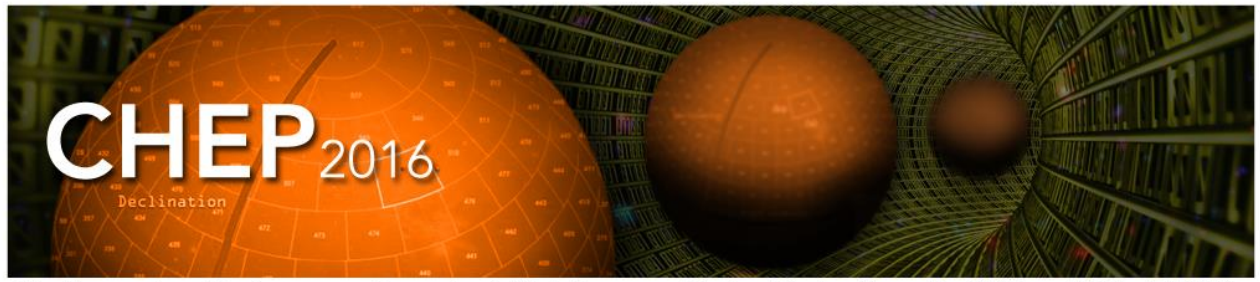


22nd International Conference on Computing in High Energy and Nuclear Physics, Hosted by SLAC and LBNL, Fall 2016

An Oracle-based Event Index for ATLAS

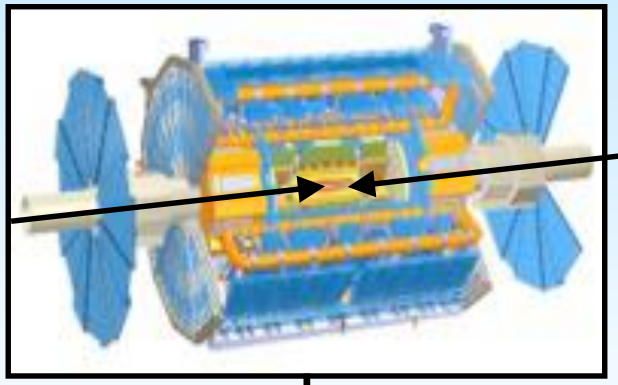
Elizabeth Gallas, Gancho Dimitrov, Petya Petrova,
Zbigniew Baranowski, Luca Canali, Andrei Dumitru,
Andrea Formica

CHEP 2016 Conference
San Francisco, CA, USA
Oct 8-14, 2016

Outline

- Introduction(s):
  - The ATLAS experiment at the LHC/CERN
  - ATLAS data processing stages
  - What is the ATLAS EventIndex (EI) ?
    - And what is EIO (Event Index Oracle) ?
- Requirements & Challenges; Features & Solutions
  - Use cases
  - Schema design
  - Data sources and Data selection
  - Storage and upload optimizations
  - User Interfaces
- Summary and Conclusions

ATLAS Data Overview



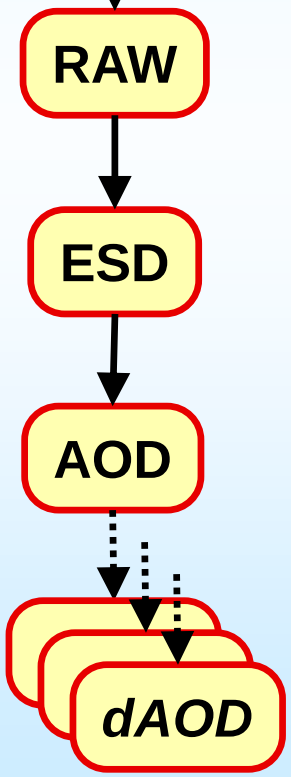
ATLAS studies particle collisions (**events**) in the heart of its multi-component particle detector

Collision rate: 600M/sec → after Trigger filtering → Recording rate: 1000/sec

→ Events per year: Billions!

Events are recorded during '**Runs**' (1-24 hours) sectioned by Luminosity Blocks '**LBN**' (minute periods of stable beam luminosity)

→ Events are assigned a unique **Event Number** in the Run



Event data is written to files for offline processing

Files: unit of data management & job-wise processing

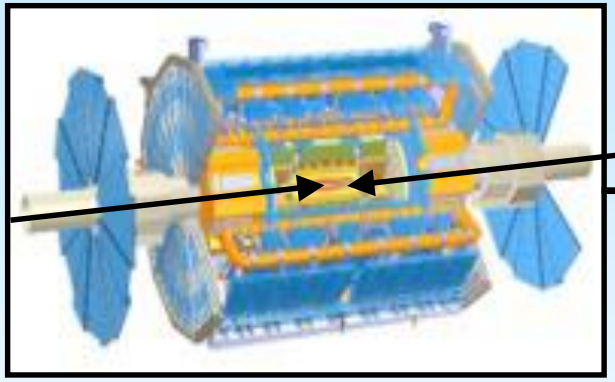
A dataset is the set of files with events from one Run

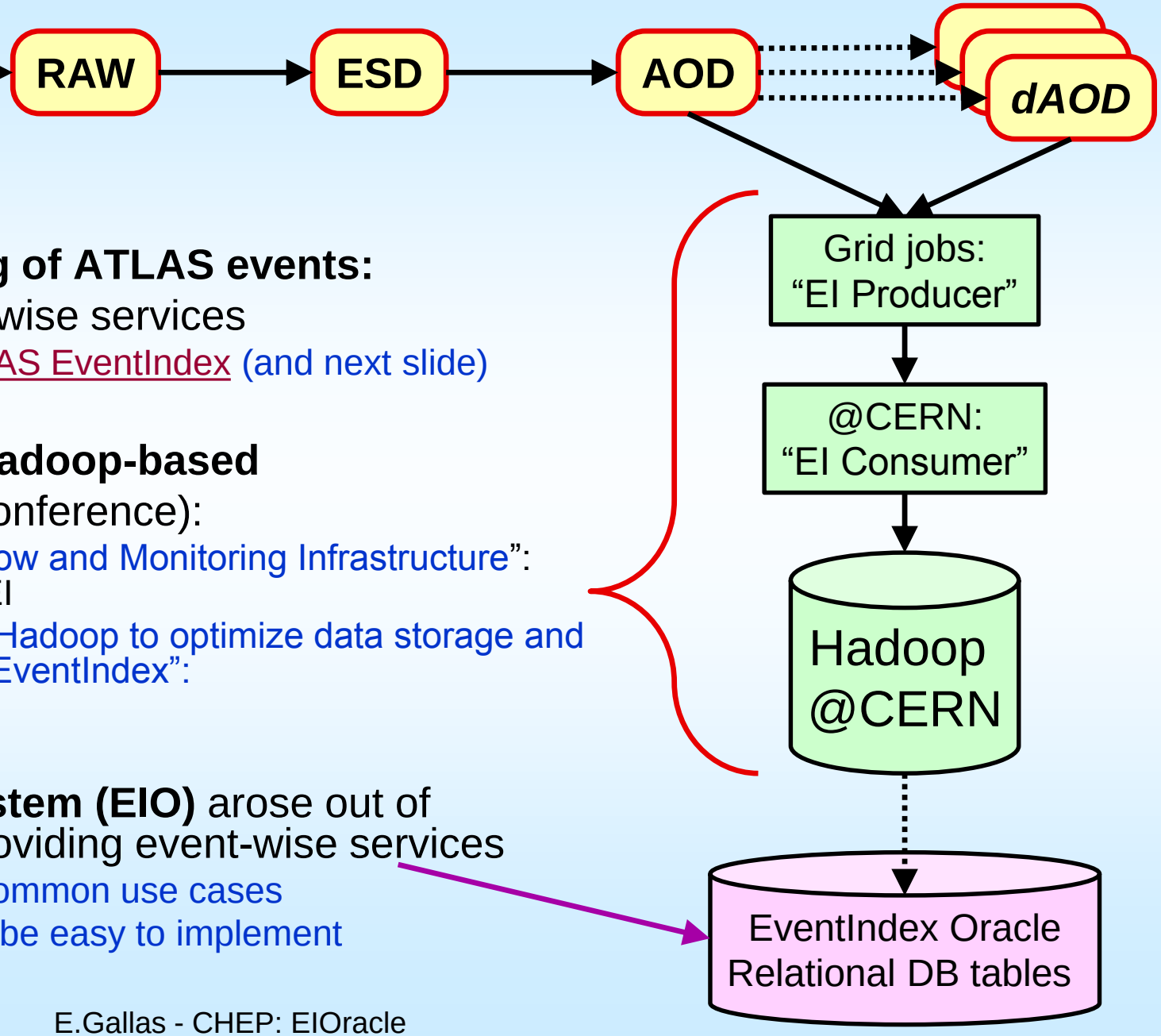
Datasets: unit of task-wise processing & physics analysis

ATLAS data file formats produced in stages of processing:

- **RAW**: raw output from detector components
- **ESD** (Event Summary Data): initial reconstruction of raw data processed with clustering, tracking, … algorithms
- **AOD** (Analysis Object Data): summary of reconstructed event suitable for most analysis (**all events**)
- **dAOD** ('derived' AOD): datasets formed from AOD customized for the needs of specific physics groups (**subset of events**)

What is ATLAS EventIndex ? and EIO ?





ATLAS EventIndex (EI) is a catalog of ATLAS events:

- Purpose: To provide various event-wise services
  - Described in CHEP 2015: The ATLAS EventIndex (and next slide)

- **Master EI storage repository is Hadoop-based**

  CHEP 2016 References (in this conference):

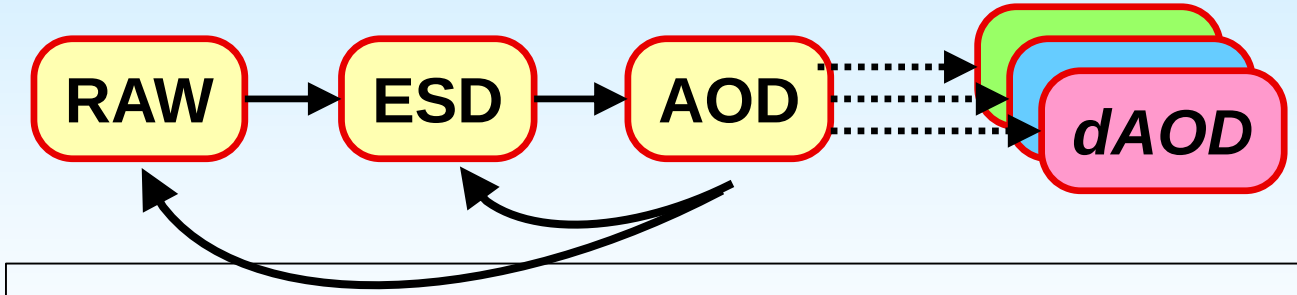
  1. "ATLAS EventIndex General Dataflow and Monitoring Infrastructure": describes the data collection for EI
  2. "A study of data representations in Hadoop to optimize data storage and search performance of the ATLAS EventIndex": exploring other data formats

- The **EventIndex Oracle-based system (EIO)** arose out of exploring various technologies for providing event-wise services
  - EIO Services include all the most common use cases and some other services proved to be easy to implement

General goals in developing EI Oracle:

Serve the main use cases with excellent performance

See what else we can do with it !



Main use case: **"Event Lookup"**

For any Event: **return references to file(s)** along the processing chain containing the event

(called the GUID == the Global Unique Identifier)

- For making event displays of special events
- For selecting a subset of events for special studies or processing

→ How does an index help ?

Datasets can have millions of events stored in thousands of files … It saves computing resources knowing which files contain events of interest

Goal: "Event Lookup" in a fraction of a second !

Other considerations and use cases:

Integrity checks:

- Identify event duplication w/in datasets and other issues in processing

  e.g. Duplication can occur at any stage: online or offline

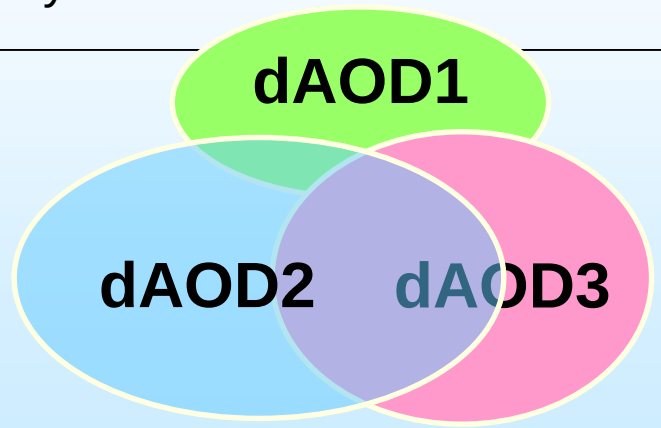
→ EIO data **integrity is important** for "Lookup" and any services as well as

to relay any problems found to ATLAS processing experts

Study "event overlap": events in common between the Datasets of a Run

- By trigger stream, processing, or filtering

→ much **easier to do with a database** than with a file system



Quantities of interest to satisfy main use cases:

Event-wise Metadata

Dataset Properties

- Project
  - Year of data taking
  - LHC beam type
- RUN_NUMBER
  - Online data taking run identifier
- Stream
  - Trigger system determines the output data stream of an event
- Production Step
  - Internal ATLAS processing step
- Data Format
  - e.g. RAW, ESD, AOD, dAOD
- AMITag
  - Version of processing of this set of events

Event Properties

- Event Number
  - Unique in the dataset
- LBN : Luminosity Block Number
  - A segmentation of the Run
- BCID : LHC bunch crossing ID
- GUID type and GUID reference 0
  - The dataset file being indexed
- GUID type and GUID reference 1
  - The file of its upstream parent dataset
- GUID type and GUID reference 2
  - The file of its further upstream parent

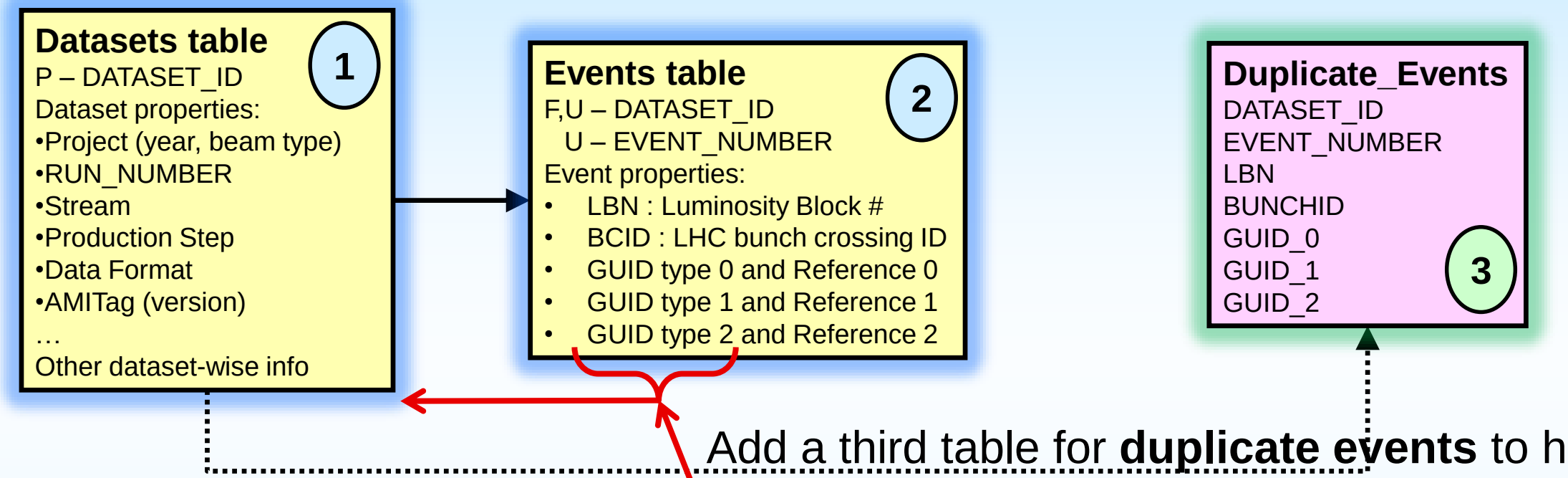
The system must catalog **Billions** of these records

Challenges:

- Sheer number of EVENTS:
  - Billions of events (currently 25.6 x 10^9 events)
  - Loading and deletion workflow (1 to O(10^6) events in such transactions)
  - Duplicate event detection and handling
- Auxiliary data for cross checks, browsing, ranking, and reports

Event Metadata (for the main use cases) is structured and the structure is very simple (2 tables)

Datasets and their Events in a Relational Database



Add a third table for **duplicate events** to help understand the nature and origin of duplication

Other immediate relational optimizations:

- Store the three GUID types at the dataset level
  - They vary by dataset, but are common to all events in a dataset
- Store only up to 3 levels of GUIDs per event
  - In principle, can be as many as 4 references per event, but in practice, 3 is sufficient for use cases and sometimes only 2 are needed

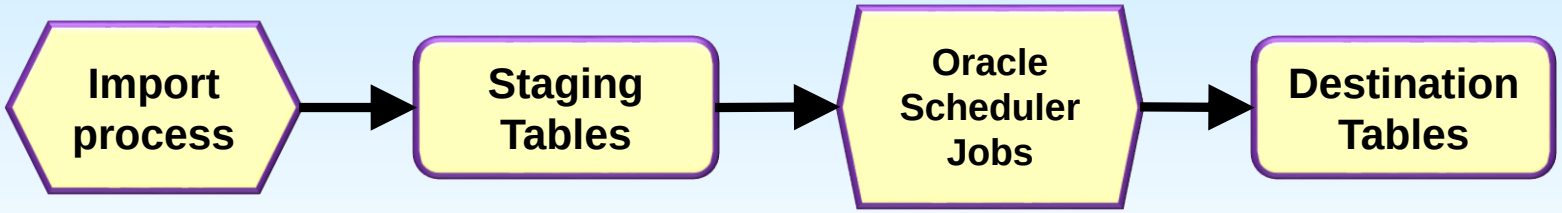
Efficient EVENT storage is achieved using:

Normalization: basic relational model techniques

and some extreme database tuning !

Event table volume-related challenges & solutions !

1. New data:



- Import into a staging table (no constraints, no indexes)
- Oracle scheduler jobs: verify, optimize, then move the events to destination tables **EVENTS** (and possibly the **DUPLICATE_EVENTS** table)

2. Further reduce Event table volume/row:

Use non-standard Oracle datatypes for GUID references:

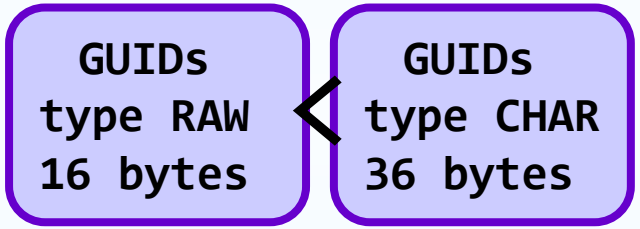
example: "21EC2020-3AEA-4069-A2DD-08002B30309D"

× 3 GUIDs per event: 108 bytes in CHAR datatype

→ Solution: use Oracle "RAW" data type

→ Oracle functions convert to CHAR as needed

→ Lookup always by Event #, not GUID itself



3. OLTP Compression:
   - Reduces data volume and improves transaction speed
   - Moderates the side effects of transactions (1 to O(10^6) events)
     - Minimizes undo and redo footprint on the database

4. Partition the Events table to handle removal of obsolete datasets
   - "LIST" partitioned (by Dataset ID): delete events → by partition removal !

more about Datasets

Datasets table
- Project (year, beam type)
- RUN_NUMBER
- Stream
- Production Step
- Data Format
- AMITag (version)

…
- GUID_TYPE0, 1, 2
- EIO_DATASET_STATUS
- AMI_DATASET_STATUS
- EIO_INSERT_DATE
- DATASET_CREATE_DATE
- DATASET_UPDATE_DATE
- EIO_COUNT_EVENTS
- AMI_COUNT_EVENTS
- RAW_COUNT_EVENTS
- UNIQ_DUPL_EVENTS
- NUM_DUPLICATES
- LATEST_RANK
- AMI_FILE_COUNT
- GUID0_COUNT
- GUID1_COUNT
- GUID2_COUNT
- INCONTAINER

In pre- and post- upload processing, we

- exploit information in other ATLAS metadata catalogs
  - AMI: Dataset-level metadata (CHEP 2016)
  - COMA: Run-level metadata (CHEP 2012)

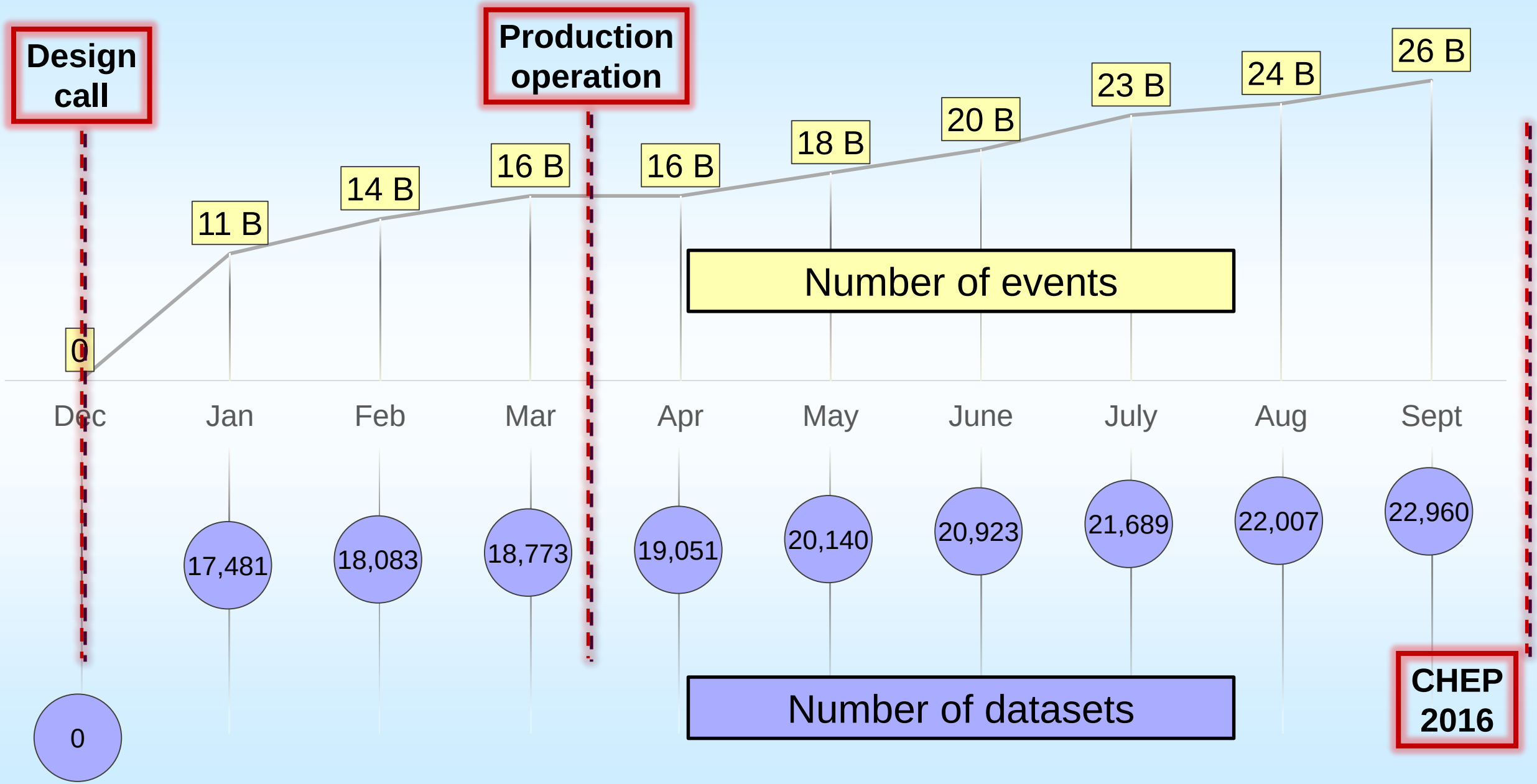
  } Also in Oracle
- aggregate information from the Events table

This data is used during processing (to make decisions) and is sometimes added to the Datasets table to provide more robust and coherent interfaces and reports.

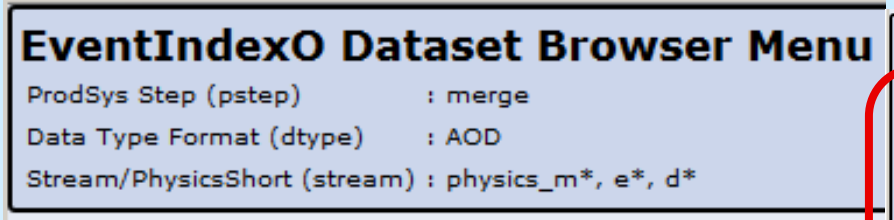
- Specifically:

Filter input datasets
  Some Runs and Streams are eliminated
Check dataset status on insert and with time
Checks of event counts:
  Dataset completeness
Show counts of duplicated events
  Number of duplicate events
  Number of unique duplicate events
Show additional counts
  Number of data files
  Number of unique GUID counts
Rank the datasets to find the ‘latest’ processing

ATLAS Event Index Oracle System Evolution



User Interfaces

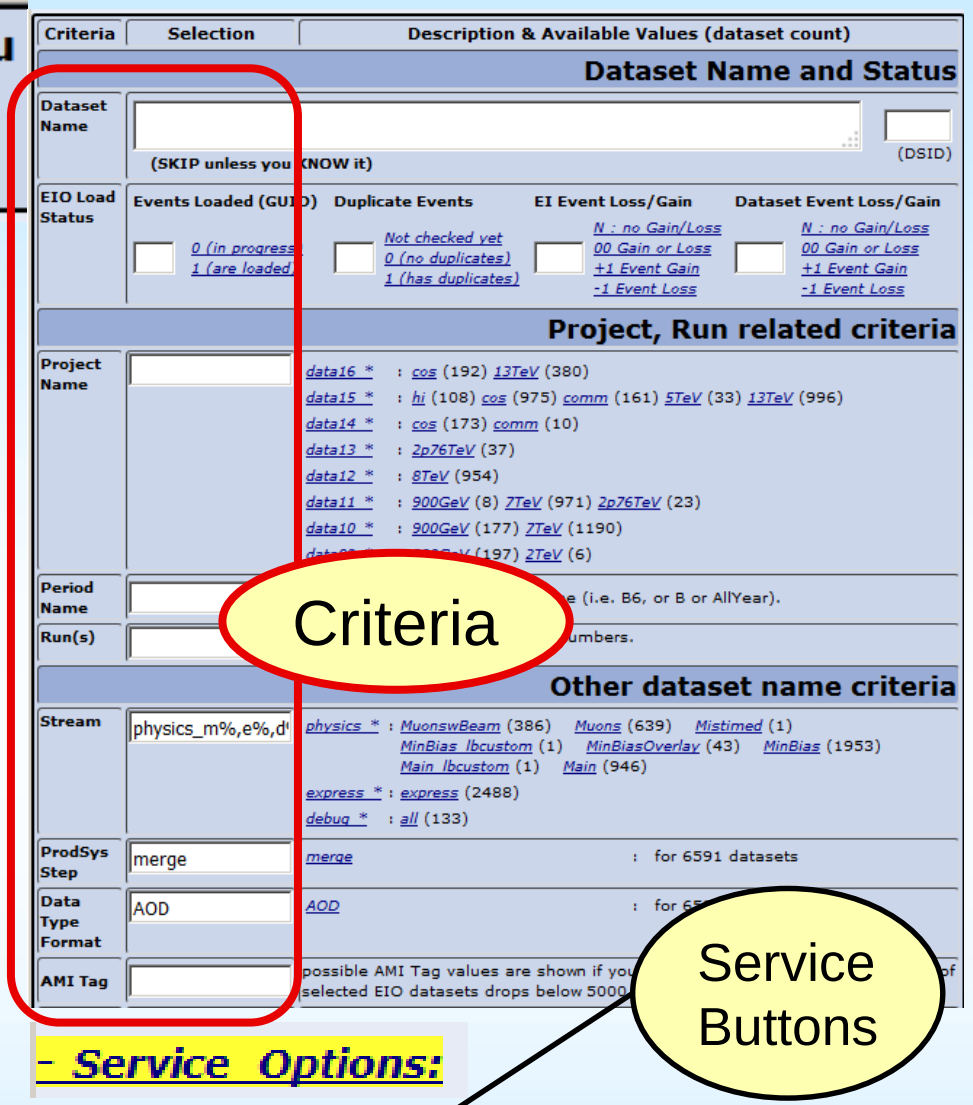


The EI Oracle Dataset Browser:

- Users easily find indexed datasets and their properties

Robust set of services are available from the browser:

- Event Lookup
  - Return GUIDs for events of interest
- Dataset Report
  - Show details about the collected dataset
  - Provide links to related reports in other systems
- Dataset Overlaps Report → example on next slide …
  - Show events in common between datasets of the same Run
- Duplicate Event Report
  - Show datasets with duplicate events and
  - Investigate the source of duplication
- Count by BCID and Count by LB Report
  - Show event counts broken down by these properties
- Missing Event Report
  - Can show which events are missing in some cases



EventIndexO Dataset Overlaps Report

AMI Tag Name (ami) : f708*
Data Type Format (dtype) : DAOD_*
Stream/PhysicsShort (stream) : physics_Main
Run Number (runs) : 300655

Dataset selection criteria

Dataset	DAOD EXOT0	DAOD EXOT4	DAOD EXOT10	DAOD HIGG1D1	DAOD HIGG5D1	DAOD JETM1	DAOD JETM3	DAOD JETM8	DAOD JETM9	DAOD JETM11	DAOD STDM4	DAOD SUSY5	DAOD SUSY6	DAOD SUSY8
DAOD_EXOT0	502891 (100%)													
DAOD_EXOT4	427343 (3.99%)	10647773 (100%)												
DAOD_EXOT10	128808 (13.63%)	170178 (1.54%)	570837 (100%)											
DAOD_HIGG1D1	169929 (14.94%)	226424 (2.02%)	570774 (70.95%)	804424 (100%)										
DAOD_HIGG5D1	18459 (0.2%)	1371482 (7.61%)	6108 (0.07%)	6982 (0.07%)	8740180 (100%)									
DAOD_JETM1	2938 (0.11%)	225187 (1.79%)	4599 (0.17%)	4624 (0.16%)	159209 (1.48%)	2145394 (100%)								
DAOD_JETM3	386241 (73.2%)	357159 (3.34%)	94981 (10.71%)	121017 (11.06%)	13162 (0.14%)	1013 (0.04%)	411007 (100%)							
DAOD_JETM8	2930 (0.14%)	212144 (1.75%)	4588 (0.21%)	4604 (0.19%)	158577 (1.55%)	1660093 (77.38%)	1007 (0.05%)	1660093 (100%)						
DAOD_JETM9	2938 (0.11%)	225187 (1.79%)	4599 (0.17%)	4624 (0.16%)	159209 (1.48%)	2145394 (100%)	1013 (0.04%)	1660093 (77.38%)	2145394 (100%)					
DAOD_JETM11	488068 (3.98%)	8371354 (57.67%)	179864 (1.42%)	244306 (1.91%)	760522 (3.76%)	141477 (0.99%)	410983 (3.36%)	136318 (0.99%)	141477 (0.99%)	12238918 (100%)				
DAOD_STDM4	491629 (3.84%)	10457192 (80.49%)	195011 (1.48%)	263457 (1.97%)	1384221 (6.87%)	231596 (1.57%)	411007 (3.21%)	217481 (1.53%)	231596 (1.57%)	10367259 (70.65%)	12801582 (100%)			
DAOD_SUSY5	500891 (2.92%)	8948443 (47.46%)	282859 (1.62%)	390570 (2.22%)	1515843 (6.22%)	353323 (1.86%)	411007 (2.4%)	328423 (1.78%)	353323 (1.86%)	12238918 (71.34%)	10936535 (57.5%)	17155997 (100%)		
DAOD_SUSY6	389660 (3.23%)	4024321 (21.65%)	116448 (0.94%)	147819 (1.17%)	8724403 (72.82%)	160646 (1.15%)	369639 (3.08%)	159825 (1.19%)	160646 (1.15%)	3951791 (19.51%)	4561559 (22.58%)	4708007 (19.29%)	11964696 (100%)	
DAOD_SUSY8	263590 (2.04%)	4383318 (23.13%)	6065 (0.05%)	6995 (0.05%)	8724282 (68.68%)	166070 (1.13%)	265958 (2.07%)	165262 (1.17%)	166070 (1.13%)	4389894 (21.38%)	5119163 (25.13%)	5180025 (21%)	10489639 (74.07%)	12687666 (100%)
	DAOD EXOT0	DAOD EXOT4	DAOD EXOT10	DAOD HIGG1D1	DAOD HIGG5D1	DAOD JETM1	DAOD JETM3	DAOD JETM8	DAOD JETM9	DAOD JETM11	DAOD STDM4	DAOD SUSY5	DAOD SUSY6	DAOD SUSY8

Color legend for percentage overlap:

100%	> 90%	> 80%	> 70%	> 60%	> 50%	> 40%	> 30%	> 20%	> 10%	> 0.01%	> 0%	== 0

Purpose: Show event counts (and %) of events in common between selected datasets of a Run

Simple steps from the EI Oracle Browser:

- User specifies the Run and datasets of interest
- Then clicks on the "Dataset Overlaps" Button

Interface Features:

- A configurable threshold: output matrix shows only overlaps above the threshold
  - The threshold for this snapshot is 70%
- Offers 2 overlap algorithms:
  - Intersection / Union x 100 % (default)
  - Intersection / (Count in Dataset 1) x 100%

Action: DatasetOverlaps — **88 Datasets** meet the input criteria.

+ Datasets (88) for Run 300655:

+ Raw Overlap counts (3916):

+ Filter selected datasets:

- Overlap Matrix (14 x 14):

Summary

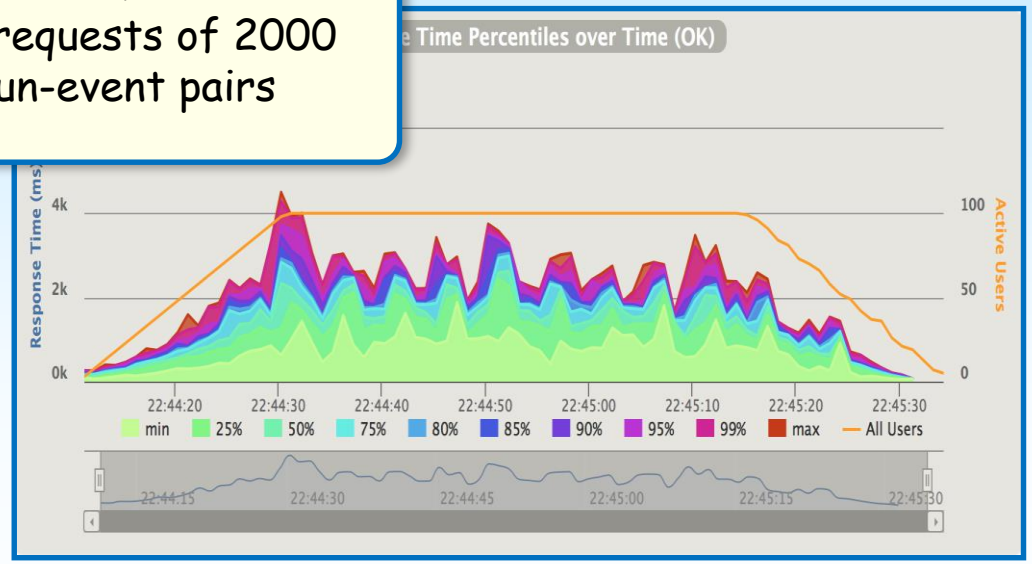
ATLAS Oracle-based Event Index EIO

- arose out of exploring various technologies for providing event-wise services using ATLAS Event Index data

Impressive minimization of resources while achieving performance goals

- Raw data volume: significantly reduced using a relational model and a number of carefully chosen techniques available in Oracle DBMS
  - Non-optimized volume: ~210 bytes/event
    - Volume to store 25E9 events: 5 TB !
  - EIO optimizations: volume ~20 bytes/event !
    - 26E9 rows currently in the system: 510 GB ! (plus index overhead: 465 GB)

Performance tests with 100 users show AVG < 2s for requests of 2000 run-event pairs



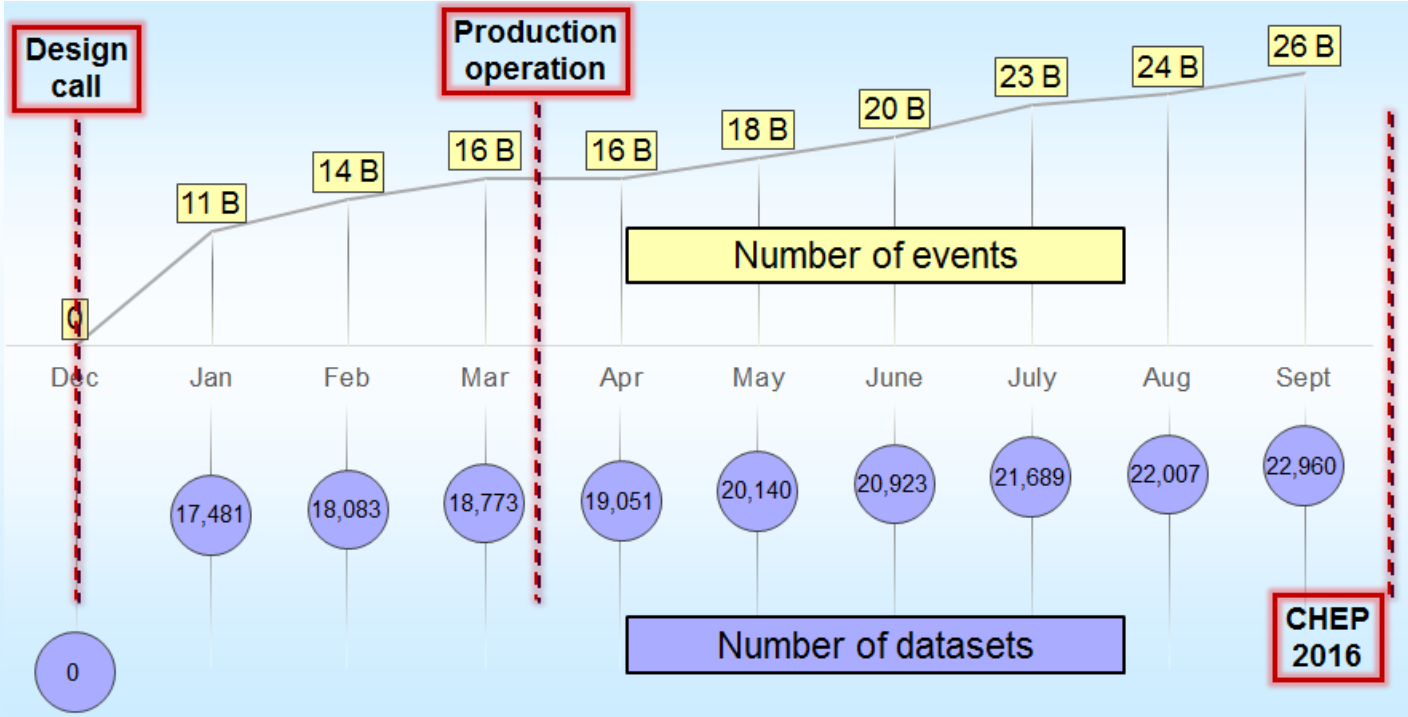
Performance is excellent for “Lookup”:

- Returns references to single events in a fraction of a second
- Simulation of 100 users: AVG < 2s for requests of 2000 run-event pairs

A number of integrity issues were found in the initial EI data

- Many checks are now integrated into the EI Collector – now caught further upstream

- Event-level metadata services for ATLAS based on an Oracle DBMS is deployed
  - Users and Experts: Like the performance and the interfaces
- Services cover the most common use cases
  - and additional services were easy to implement
- Success of the project rests on hyper-efficient modelling of the storage underlying the services
  - The system is simple, robust and reliable
- EIO reached a production state very quickly !
  - Design discussions started:
    - December 2015
  - Production DB and interfaces
    - Early March 2016
  - Smooth operations in Run 2
  - Designed to handle future rates
    - 10's of billions of events/year



Conclusions