ATLAS EventIndex y Event WhiteBoard

Sistema de indexado y catalogado de eventos para experimentos con grandes cantidades de datos.

Álvaro Fernández Casaní (IFIC – CSIC/UV)

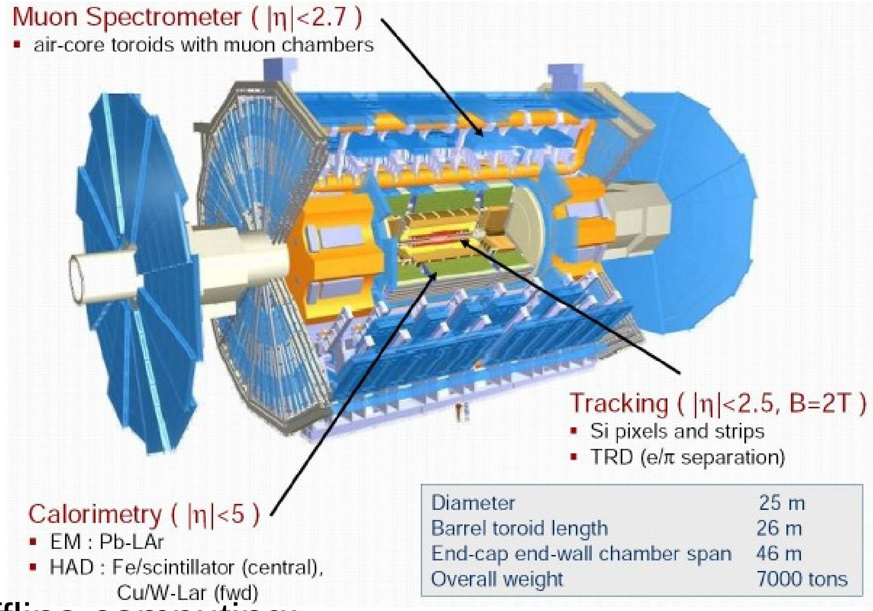
On behalf of the EventIndex team

IFIC - Jornadas Técnicas – 13 Marzo 2018

Outline

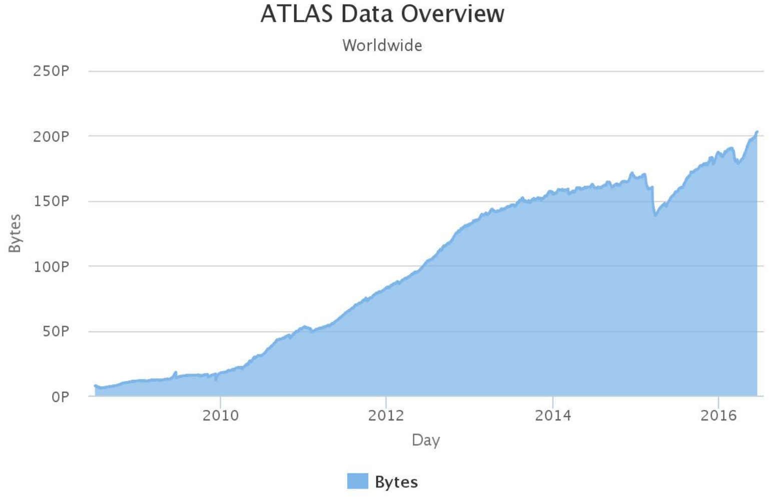
- ATLAS EntIndex project
  - Objective and use cases
- Distributed Production and Data Collection Architecture
- Performance and results
- Evolution: Event WhiteBoard
- Summary

ATLAS Computing Challenges


Muon Spectrometer (|η|<2.7)
▪ air-core toroids with muon chambers
Tracking (|η|<2.5, B=2T)
▪ Si pixels and strips
▪ TRD (e/π separation)
Calorimetry (|η|<5)
▪ EM : Pb-LAr
▪ HAD : Fe/scintillator (central), Cu/W-Lar (fwd)

Diameter	25 m
Barrel toroid length	26 m
End-cap end-wall chamber span	46 m
Overall weight	7000 tons

2016 milestone (run 2): 200 Petabytes moved



Moving >1 PB, >20 GB/s Creating: 1.5-2M files per day

- The offline computing:
  - 2018: LHC plan, 25ns BCMS, 2500b, 1.3e11 ppb,L=2.2e34
  - 1 kHz real data taking. 10^{10} events/year
  - Average event size (raw data): 0.8 MB/event

Processing:

- >150 centres with ~300-350k cores

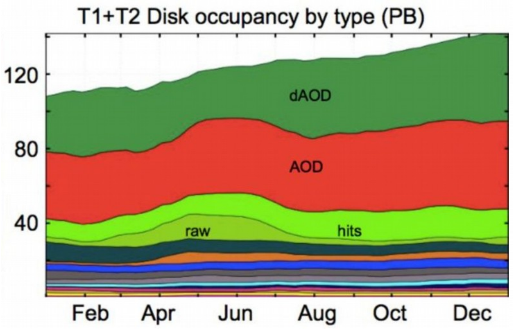
Storage:

- raw data recording rate 440 MB/sec. Reprocessings and total data moving (figures on the right)

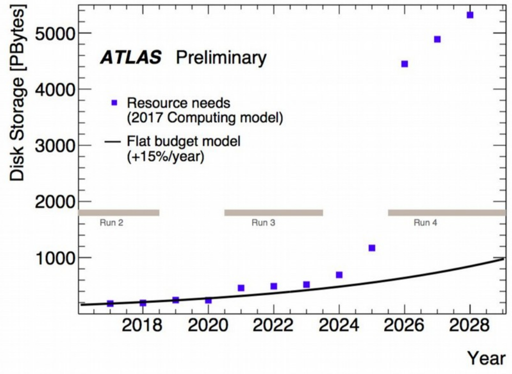
Data Access:

- Physicist spread around the world. Grid technologies to access computing and data.

2017 Storage



Forecast

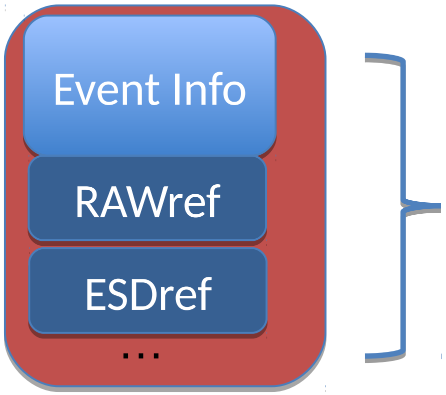


Predicted run 4 (2024): Exabyte scale

Disk storage ~6x short at HL-LHC

EventIndex: an event catalog

- **A catalog of data** (all events in all processing stages) is needed to meet multiple use cases and search criteria. A small quantity of data per event is indexed.



Events stored in files (identified by GUID)
Files are grouped into DATASETS
Wanted Event Index information ~= 300bytes to 1Kbyte per event:

- Event identifiers (run / event numbers, trigger stream, luminosity block)
- Online trigger pattern (l1, l2, ef)
- References (pointers) to the events at each processing step (RAW, ESD, AOD, DAOD) in all permanent files on storage

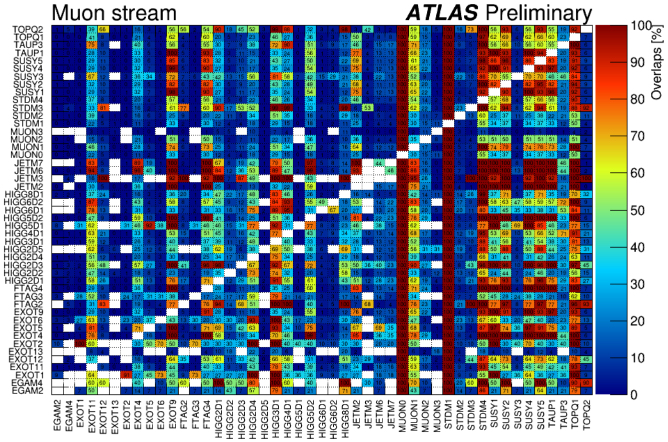
- We are indexing **Billions of Events,** stored in **Millions of files** replicated at CERN and **hundreds of grid-sites worldwide,** meaning **Petabytes of data→ A complex big data distributed system.**

Use Cases

1) Event picking: users able to select single events depending on constraints. Order of hundreds of concurrent users, with requests ranging from 1 event (common case) to 30k events (occasional).

2) Production consistency checks

- **Duplicate event checkings**: events with same Id appearing in same or different files/datasets.
- **Overlap detection** in derivation framework: construct the overlap matrix identifying common events across the different files.

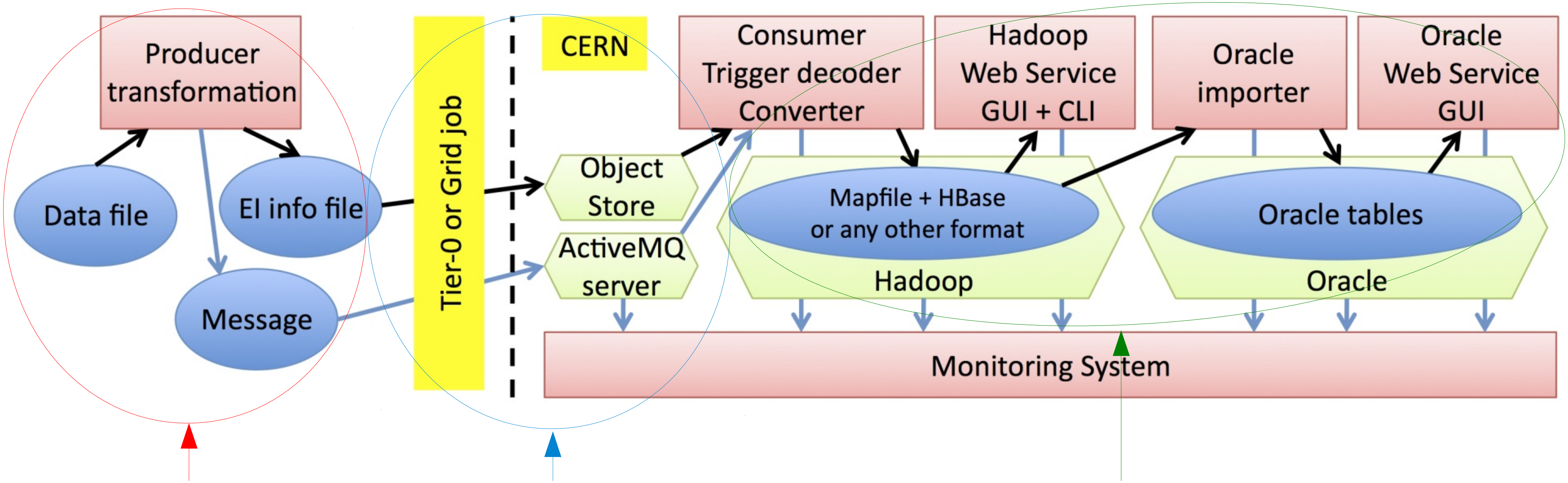


3) Trigger checks and event skimming: Count or give an event list based on trigger selection.

- **Trigger Overlap detection:** number of events in a real data Run/Stream satisfying trigger X which also satisfies trigger Y.
- (See presentation by Carlos García Montoro ‘ATLAS EventIndex Trigger Counter)

Storing and accessing thousands of files and millions of events in reasonable time is done with Hadoop Technologies.

EventIndex Architecture



Data Production

Data Collection

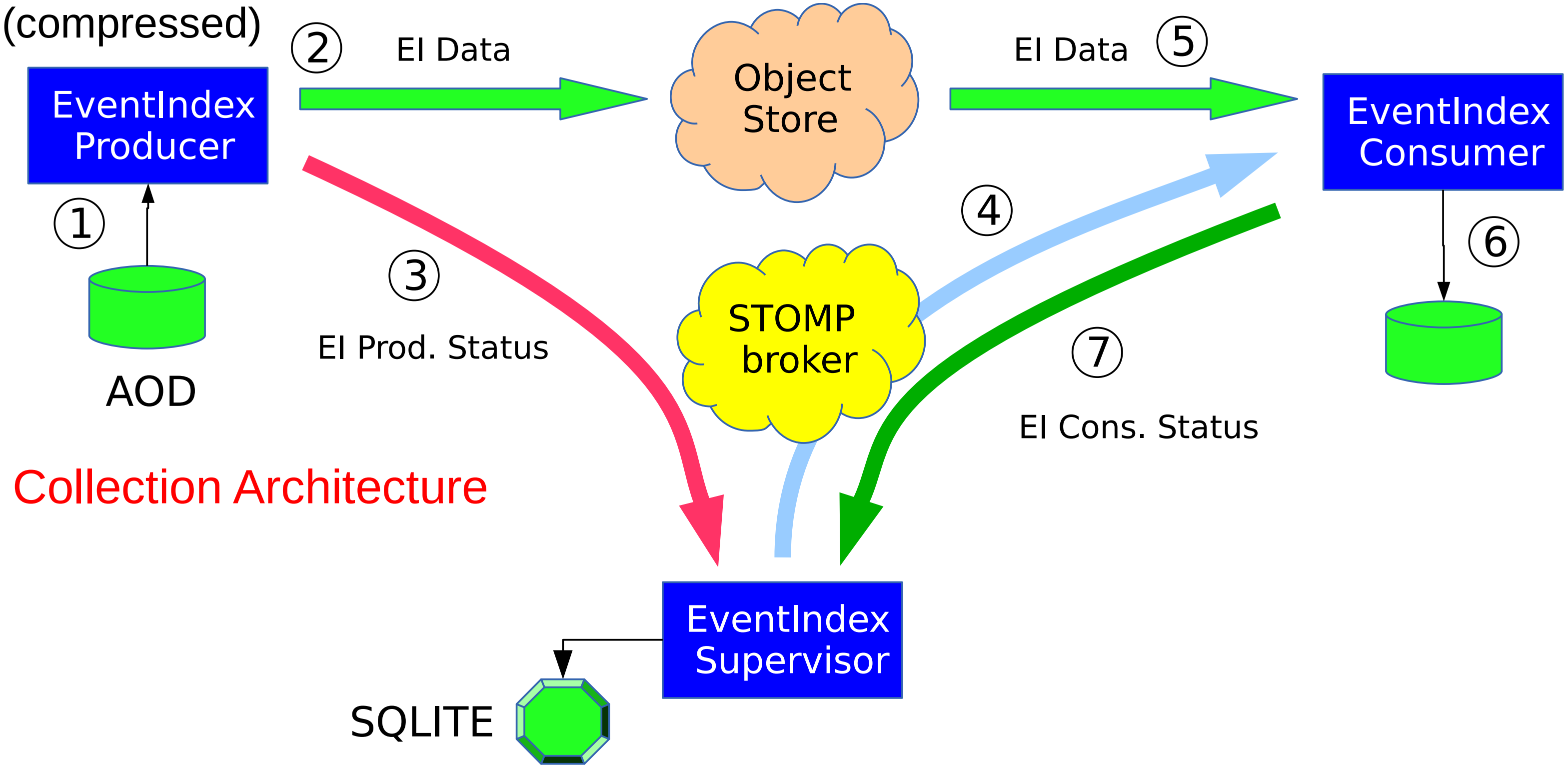
Data Storage and Query Interfaces

IFIC coordinates Data Production task to ensure all event index grid production performs correctly.

IFIC is responsible for all Data Collection task: a distributed producer/consumer architecture to collect all indexed data and ingest it to the Data Storage services.

Data collection

2015 - mid 2017: Pure Messaging Based architecture (ActiveMQ brokers / Stomp protocol). Json data encoding. (production peaks showed bottlenecks on messaging brokers)
mid 2017 onwards: ObjectStore (CEPH / S3 interface) as intermediary storage. Google protobuf data encoding (compressed)

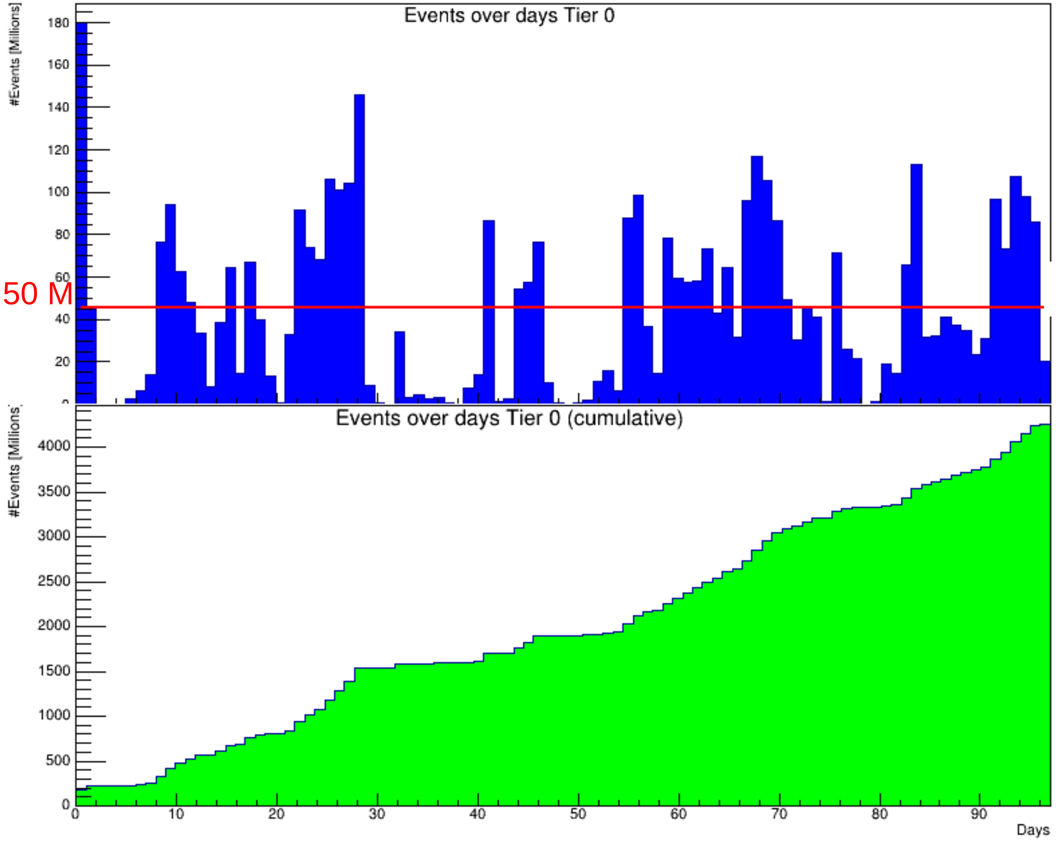


2018 Data Collection Architecture

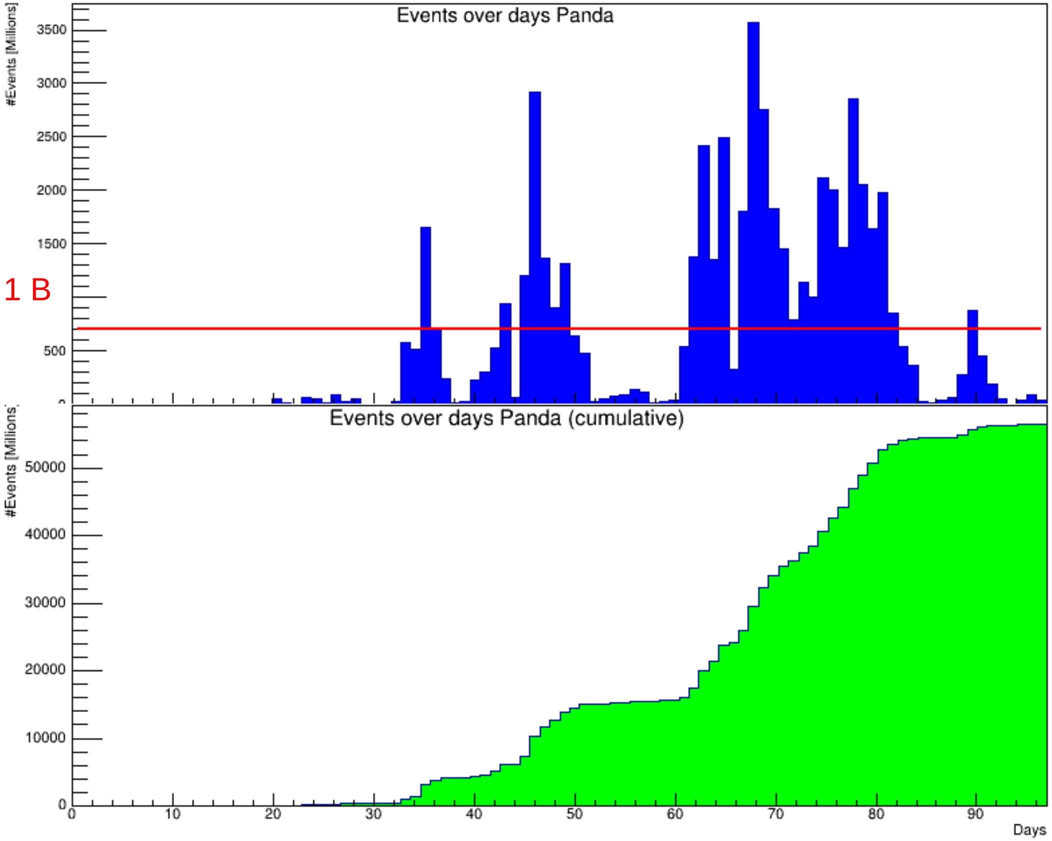
Producer: Athena Python transformation (ATLAS r21.0.32), running at Tier-0 and grid-sites. Indexes AOD data and produces an **eventindex file**, stores in **ObjectStore**
Supervisor: Controls all the process, received processing information and validates data by dataset. Signals valid unique data for ingestion to Consumers. Operated with a web interface
Consumers: Retrieves ObjectStore data, groups by dataset and ingest it into **HDFS (Hadoop distributed Filesystem)**

Event processing rates

TIER0 @ CERN



GRID



EventIndex data production:

Tier0 @ CERN : ~50 M indexed events/day (3 K jobs/day)

Grid Sites: ~1 B indexed events/day (5 K jobs/day)

- Single Consumer event throughput performance improved from 1K events/s (Messaging only), to 15K events/s (ObjectStore).
- Overcoming messaging brokers bottleneck, we can also now scale horizontally.

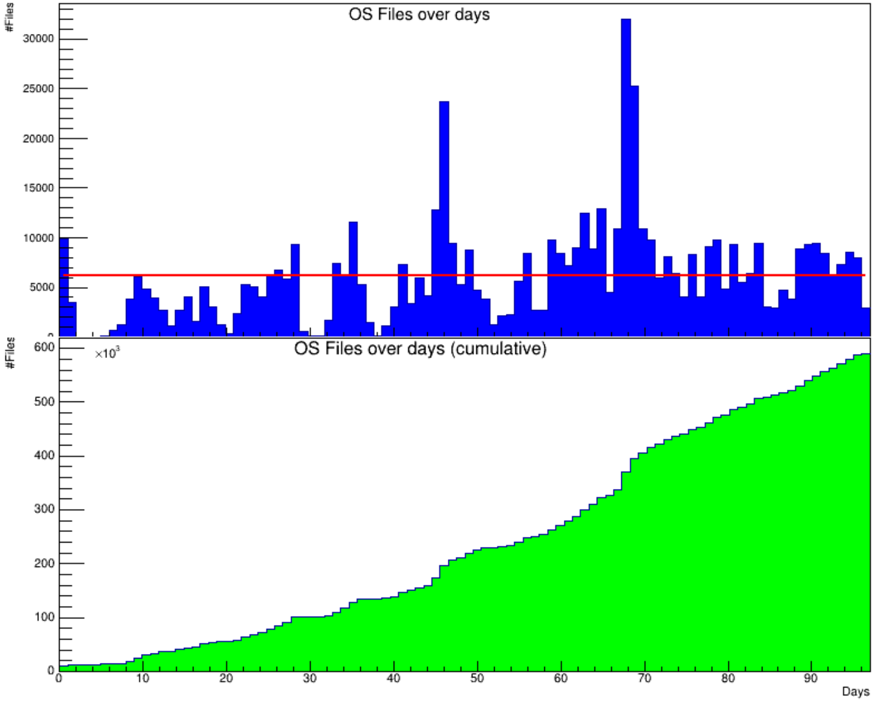
Data Ingestion to Hadoop

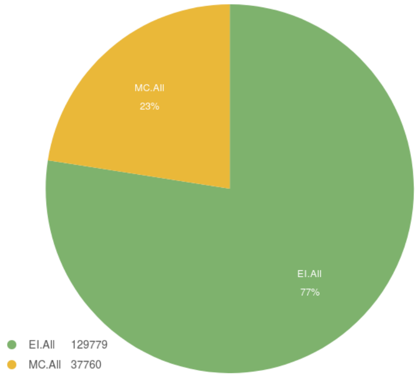
- Since mid 2017 DataCollection uses the Object Store to temporary maintain the eventindex files created by the Producers:
  - Creating ~7 K objects/day (~22 GB) that contain the ~1 B events indexed per day.
  - Data is factor 10 compressed with respect to original eventindex data.
  - Current ObjectStore data (March 2018) 1335000 objects (8TB data)

- Consumers retrieve the eventindex files from ObjectStore and write it in HDFS Hadoop

- Current EventIndex Data in Hadoop:
  - 167 TB of indexed events data (129.7 TB real data, and 37.7 TB MonteCarlo simulated data)

ObjectStore (CEPH)





Hadoop (HDFS)

Event WhiteBoard

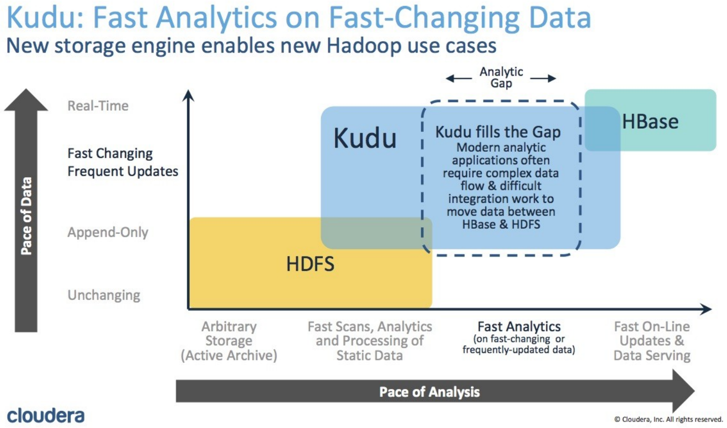
- **An evolution of the EventIndex concepts**
  - **Currently**: the same event across each processing step (RAW, ESD, AOD, DAOD, NTUP) is physically stored at different HADOOP HDFS files.
  - **Future**: One and only logical record per event (Event Identification, Inmutable information (trigger, lumiblock, …), and for each processing step:
    - Link to algorithm (processing task configuration)
    - Pointer(s) to output(s)
    - Flags for offline selections (derivations)
- **Support Virtual Datasets:**
  - A logical collection of events
    - Created either explicitly (giving a collection of Event Ids) of implicitly (selection based on some other collection or event attributes)
    - Labelling individual events by a process or a user with attributes (key:value)
- **Evolve EventIndex technologies to future demanding rates, and WhiteBoard requirements**
  - **Currently:** ALL ATLAS processes: ~30billion events/day (up to 350Hz on average) → update rate throughout the whole system (all years, real and simulated data). Read 8 M files/day and produce 3 M files
  - **Future:** due to expected trigger rates, need to scale for next ATLAS runs: at least half an order of magnitude for Run3 (2021-2023): 35 B new real events/year and 100 B new MC event/year. For run4: 100 B new real events and 300 B new MC events per year. Then sum up replicas and reprocessing

Current work on Backend storage

Optimization and unification of data storage for the EventIndex.

Apache Kudu: new columnar-based storage that allows fast insertions and retrieval.

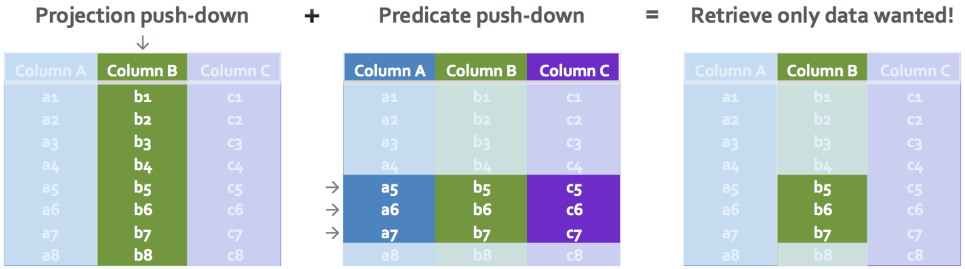
Testbed @ IFIC for current developments and tests. First test promising. Explore support for Event Whiteboard user cases.



Kudu: Fast Analytics on Fast-Changing Data
New storage engine enables new Hadoop use cases
Analytic Gap
Pace of Data
Real-Time
Fast Changing Frequent Updates
Append-Only
Unchanging
Kudu
Kudu fills the Gap
Modern analytic applications often require complex data flow & difficult integration work to move data between HBase & HDFS
HBase
HDFS
Arbitrary Storage (Active Archive)
Fast Scans, Analytics and Processing of Static Data
Fast Analytics (on fast-changing or frequently-updated data)
Fast On-Line Updates & Data Serving
Pace of Analysis
cloudera
© Cloudera, Inc. All rights reserved.

Benefit from common access patterns:

- Related data (reprocessing) sit close to each other on disc. Reduce redundancies and improve navigation.
- Recent loaded data is the most accessed data (possibility to apply in-memory data caching)



Projection push-down + Predicate push-down = Retrieve only data wanted!

Column A	Column B	Column C
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5
a6	b6	c6
a7	b7	c7
a8	b8	c8

Column A	Column B	Column C
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5
a6	b6	c6
a7	b7	c7
a8	b8	c8

Column A	Column B	Column C
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5
a6	b6	c6
a7	b7	c7
a8	b8	c8

Summary

- IFIC is contributing and leading data collection task of the ATLAS EventIndex project.
  - Operating in production indexing billions of events from thousand of grid jobs running in distributed manner worldwide.
  - Recently improved Data Collection performance, and scalability for future runs.
- Gained experience developing and operating a complex distributed system, and in the Big Data tools world:
  - Hadoop HDFS, Hbase, MapReduce framework, Apache Kudu.
- Future challenges regarding new production rates, and Event WhiteBoard use cases:
  - Current work on new Storage Technologies to support faster insertion, and low latency and analytics use cases unification.
- More information:
  - “The ATLAS EventIndex: Full chain deployment and first operation”. D. Barberis, J. Cranshaw, A. Favareto, A. Fernández Casaní, et al. Nuclear and Particle Physics Proceedings (2016), pp. 913-918. DOI information: 10.1016/j.nuclphysbps.2015.09.141
  - “ATLAS EventIndex general dataflow and monitoring infrastructure”- Á. Fernández Casaní et al 2017 J. Phys.: Conf. Ser. 898 062010

EventIndex Team

- Dario BARBERIS (Università degli Studi e INFN Genova)
- Zbigniew BARANOWSKI (CERN)
- Jack CRANSHAW (Argonne National Laboratory (US))
- Gancho DIMITROV (CERN)
- **Alvaro FERNANDEZ CASANI (IFIC - Instituto de Fisica Corpuscular)**
- Elizabeth GALLAS (University of Oxford (GB))
- **Carlos GARCIA MONTORO (IFIC - Instituto de Fisica Corpuscular)**
- Claudia GLASMAN (Universidad Autonoma de Madrid)
- **Santiago GONZÁLEZ DE LA HOZ (IFIC - Instituto de Fisica Corpuscular)**
- Julius HRIVNAC (Laboratoire de l'Accelerateur Lineaire (FR))
- David MALON (Argonne National Laboratory (US))
- Fedor PROKOSHIN (Federico Santa Maria Technical University (CL))
- Grigori RYBKIN (Université Paris-Saclay (FR)
- **Javier SANCHEZ (IFIC - Instituto de Fisica Corpuscular)**
- **Jose SALT (IFIC - Instituto de Fisica Corpuscular)**
- Rainer TOEBBICKE (CERN)
- Petya VASILEVA (CERN)
- Ruijun YUAN (Laboratoire de l'Accelerateur Lineaire (FR))
- **Recently Join to System monitoring task:** Evgeny Alexandrov (Joint Institute for Nuclear Research (RU)), Igor Alexandrov (Joint Institute for Nuclear Research (RU)), Ivan Kadochnikov (Joint Institute for Nuclear Research (RU)), Mikhail Mineev (Joint Institute for Nuclear Research (RU))
- **Functional tests task:** Miguel Villaplana (Università degli Studi e INFN Milano (IT)