ATLAS EventIndex & Trigger Counter

Carlos García Montoro (IFIC – CSIC / UV)

on behalf of the IFIC EventIndex Team

I Reunión Presencial VMGRID

12th April 2018

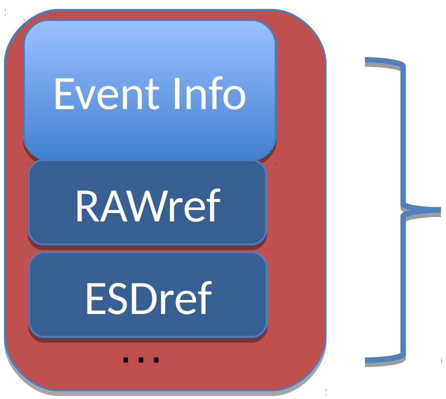
Outline

- EventIndex
  - Introduction and Use Cases.
  - Architecture.
  - Data Collections, Supervisor and Web.
  - Performance.
  - Current Issues.
- Trigger Counter
  - Introduction.
  - Web Service.
  - Internal Architecture.
  - File Formats.
  - Some Numbers.

Event Index

EventIndex: an event catalog

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



Events are 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 and HLT)
- 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

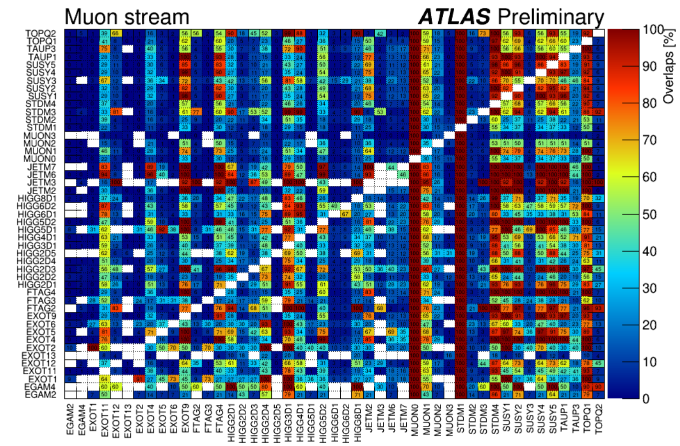
1) Event picking: users able to select single events depending on constraints. Order of hundreds of concurrent users, with requests ranging from 1 event (common case) to 30k events (occasional).

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.

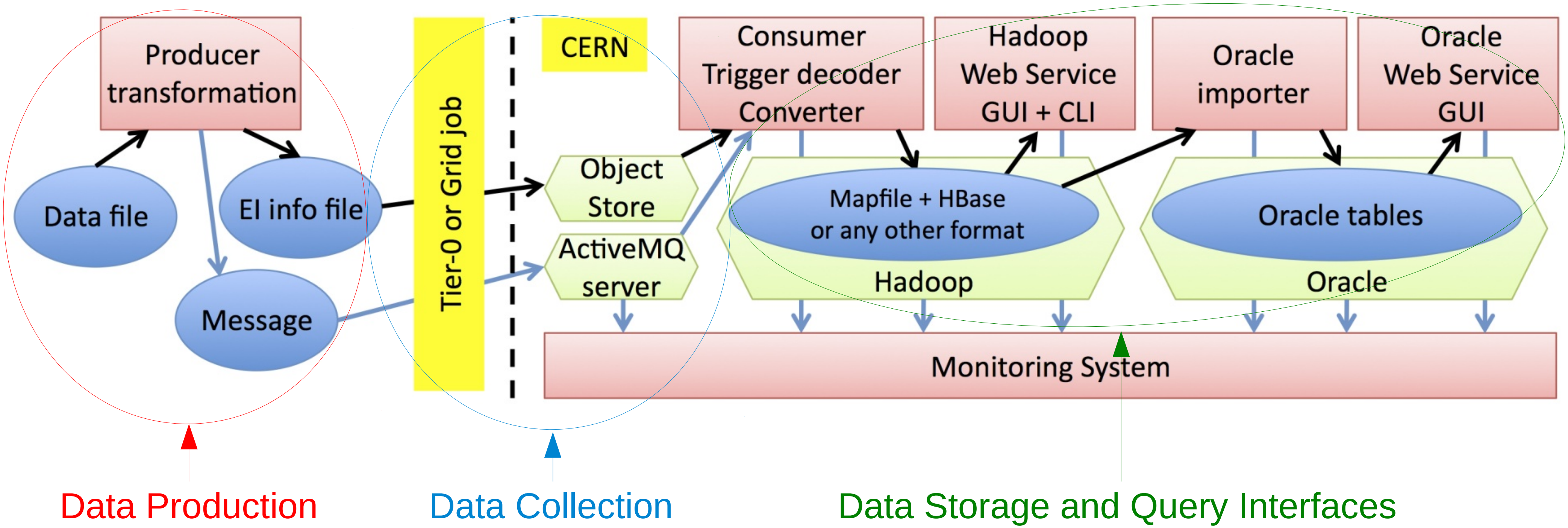
3) Trigger checks and event skimming: Count or give an event list based on trigger selection.

- **Trigger Overlap detection:** number of events in a real data Run/Stream satisfying trigger X which also satisfies trigger Y.



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

EventIndex Architecture



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

Javier, Álvaro and Carlos are responsible for all Data Collection task: a distributed producer/consumer architecture to collect all indexed data and ingest it to the Data Storage services.

Pepe has documented the use cases for the users.

Abandoned Messaging Data Collection

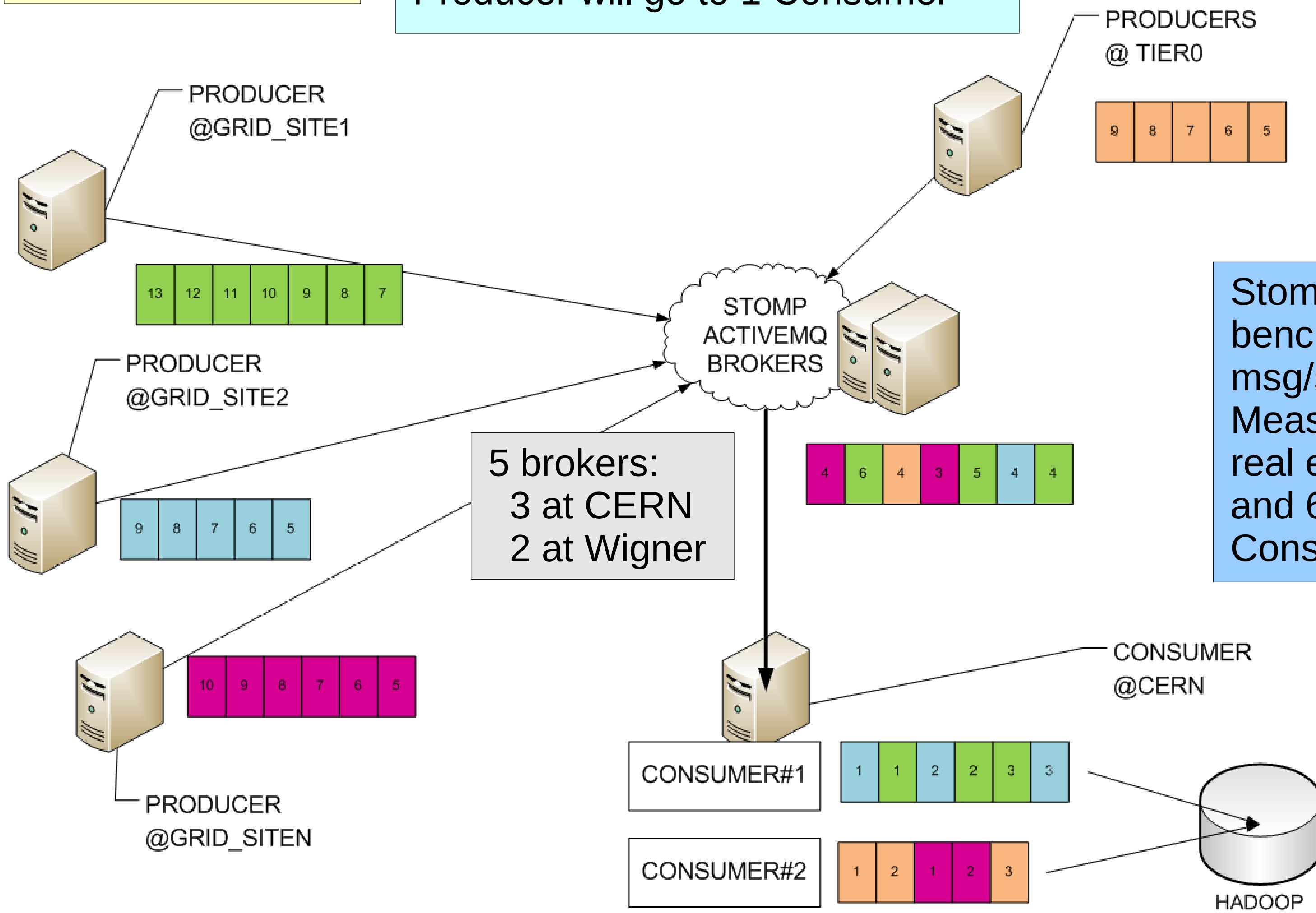
Message size is set small 1-10kB to keep broker queues agile.

Producers tag messages by group (JMSXGroupID)

Ensures that all messages from 1 Producer will go to 1 Consumer

Atomic transactions on Producers:
if connections breaks no partial processing occurs.

Status messages are sent from producers and consumers to an alternate queue.



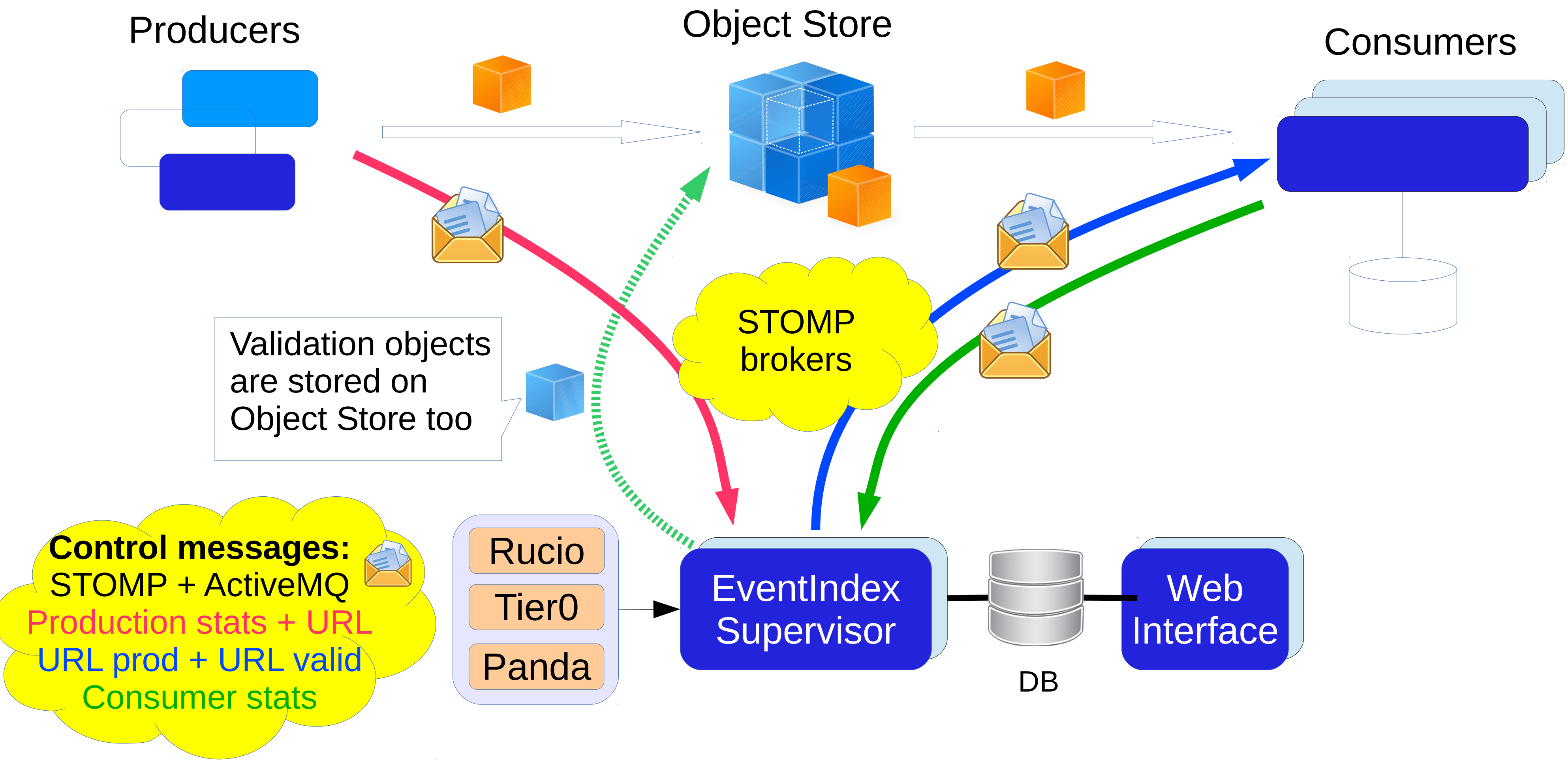
5 brokers:
3 at CERN
2 at Wigner

Stomp performance in our benchmarks reaches over 350 msg/s and 10Mb/s per Producer**.** Measured performance for sending real events reached 200K event/s and 60Mb/s (1Broker, 6 Prod/s, 4 Cons, 50K events/job)

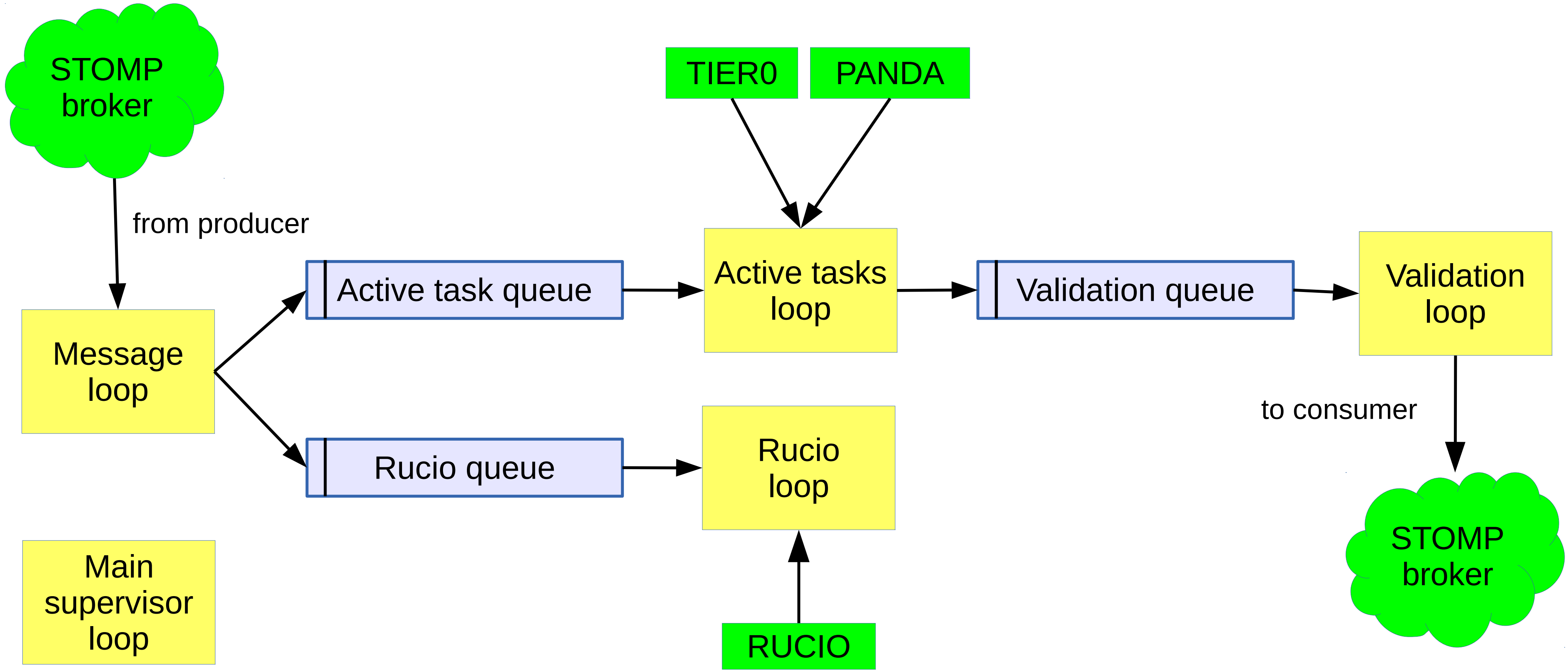
Files stored on Mapfile format usable by Hadoop Core Services

Current Object Store Data Collection

Data production is shared between producers and consumers by means of CERN's Object Store facility based on Ceph, using its Amazon S3 compatible interface



Data Collection Supervisor



Overview of the Web Interface: Tier0

Overview of the system state and load at a glance.

EventIndex Data Collection | Tier 0 ok | Panda ! | Active 0 | Rucio 0 | Validation 0 | Brokers 1 | Health ! | Hi Carlos García Montoro!

Tier 0

Filters

Show 25 entries

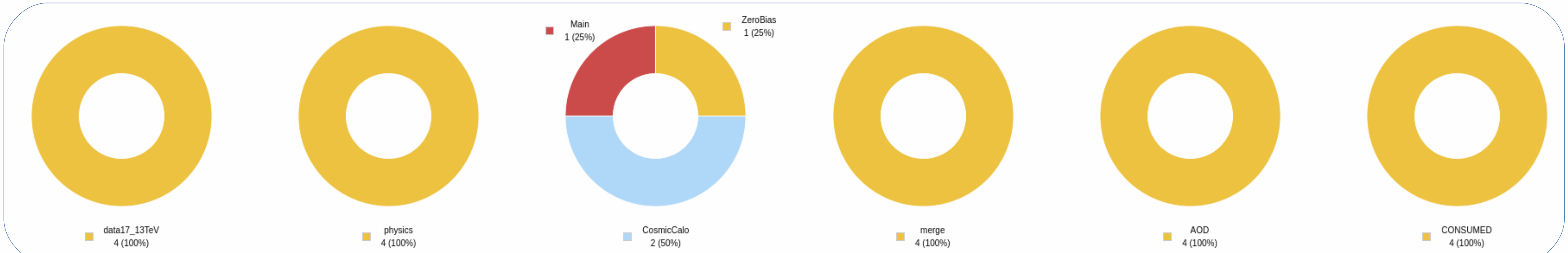
Search within results

Progress at a glance.

TaskId	Dataset	Creation	Task State	EventIndex State	Files	Events	Jobs Total	Jobs Done	Jobs Run.	Jobs Proc.	Jobs Prod.
2183747	data17_13TeV:data17_13TeV.00325140.physics_CosmicCalo.merge.AOD.f826_m1807	2017-06-07 15:17:00	FINISHED	EI_CONSUMED	1 / 1	77 / 77	1	1	0	0	1
2183736	data17_13TeV:data17_13TeV.00325042.physics_ZeroBias.merge.AOD.f826_m1807	2017-06-07 08:31:00	FINISHED	EI_CONSUMED	1 / 1	12 / 12	1	1	0	0	1
2183746	data17_13TeV:data17_13TeV.00325024.physics_Main.merge.AOD.f822_m1799	2017-06-07 15:17:00	FINISHED	EI_CONSUMED	1 / 1	10 / 10	1	1	0	0	1
2183745	data17_13TeV:data17_13TeV.00325024.physics_CosmicCalo.merge.AOD.f822_m1799	2017-06-07 15:09:00	FINISHED	EI_CONSUMED	1 / 1	1223 / 1223	1	1	0	0	1

Showing 1 to 4 of 4 entries

Previous 1 Next



Links to own pages and sources of the information.

Graphs which summarizes the data retrieved.

A Panda Task Page

Actions that privileged users can perform. States, including expected ones. All collected info.

EventIndex Data Collection | Tier 0 ok | Panda ! | Active 0 | Rucio 0 | Validation 0 | Brokers 1 | Health ! | Hi Carlos García Montoro!

Panda Task #11453518

mc15_5TeV:mc15_5TeV.420127.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee_BosonPt25Min.merge.AOD.e5883_d1424_r8670_r8052_tid11223415_00

0 EI_REGISTERED 2017-06-05 23:55:57
1 EI_PANDA 2017-06-05 23:55:57
2 EI_FAILED 2017-06-06 15:11:47

Declare Obsolete

Task

Id	EI Id	Virtual Task	Name	Status	Type	Username
11453518	1	1	user.sanchezj.spb.test.20170605_2315/	finished	anal	FRANCISCO JAVIER SANCHEZ MARTINEZ@csic.es

Task Times

Creation	Start	State Change	Modification	End
2017-06-05 21:18:48	2017-06-05 21:18:48	2017-06-06 01:28:52	2017-06-06 01:28:52	2017-06-06 01:28:52

Jobs

Total	Produced	Validated	Defined	Waiting	Pending	Assigned	Throttled	Activated	Sent	Starting	Running	Holding	Transferring	Finished	Failed	Cancelled	Merging	Closed
18	8	0	0	0	0	0	0	0	0	0	0	0	0	9	9	0	0	0

Files

Events

Validation

Id: URL:

Files	Events	Unique Events	Time

Consumer

Active Tasks Queue

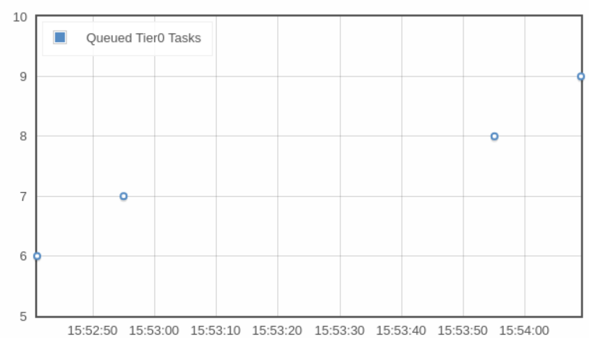
EventIndex Data Collection | Tier 0 ok | Panda n/a | Active 4 | Rucio 0 | Validation 2 | Brokers 5 | Health ok | Hi Carlos García Montoro!

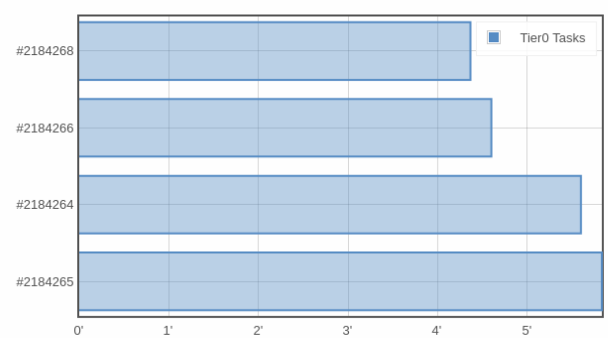
Active Tasks Queue

Show 25 entries   Search within results

Id	Task Id	Task Type	Queued Since	Modified
6	2184265	TIER0_TASK	2017-06-09 15:52:41	2017-06-09 15:52:41
7	2184264	TIER0_TASK	2017-06-09 15:52:55	2017-06-09 15:52:55
8	2184266	TIER0_TASK	2017-06-09 15:53:55	2017-06-09 15:53:55
9	2184268	TIER0_TASK	2017-06-09 15:54:09	2017-06-09 15:54:09

Showing 1 to 4 of 4 entries   Previous 1 Next





Brokers & Services Health

EventIndex Data Collection | Tier 0 ok | Panda n/a | Active 3 | Rucio 0 | Validation 0 | Brokers 5 | Health ok | Hi Carlos García Montoro!

Brokers

Show 10 entries

Search within results

									B				B		
3	mb201.cern.ch	▶	0	113	3 minutes ago	112	4 minutes ago	569	196.6 KiB	3 minutes ago	14 ms	1	203 B	an hour ago	112
4	mb204.cern.ch	▶	0	109	19 minutes ago	108	20 minutes ago	533	168.79 KiB	3 minutes ago	12 ms	1	203 B	40 minutes ago	108
6	mb119.cern.ch	▶	0	109	9 minutes ago	108	10 minutes ago	537	169.34 KiB	3 minutes ago	1 ms	2	406 B	22 minutes ago	108
7	mb120.cern.ch	▶	0	109	6 minutes ago	108	7 minutes ago	571	196.97 KiB	3 minutes ago	1 ms	2	406 B	an hour ago	108
8	mb121.cern.ch	▶	0	108	3 minutes ago	107	4 minutes ago	552	182.85 KiB	3 minutes ago	3 ms	1	203 B	an hour ago	107

Showing 1 to 5 of 5 entries

Previous 1 Next

EventIndex Data Collection | Tier 0 ok | Panda ! | Active 0 | Rucio 0 | Validation 0 | Brokers 1 | Health ! | Hi Carlos García Montoro!

Services Health

Show 25 entries

Search within results

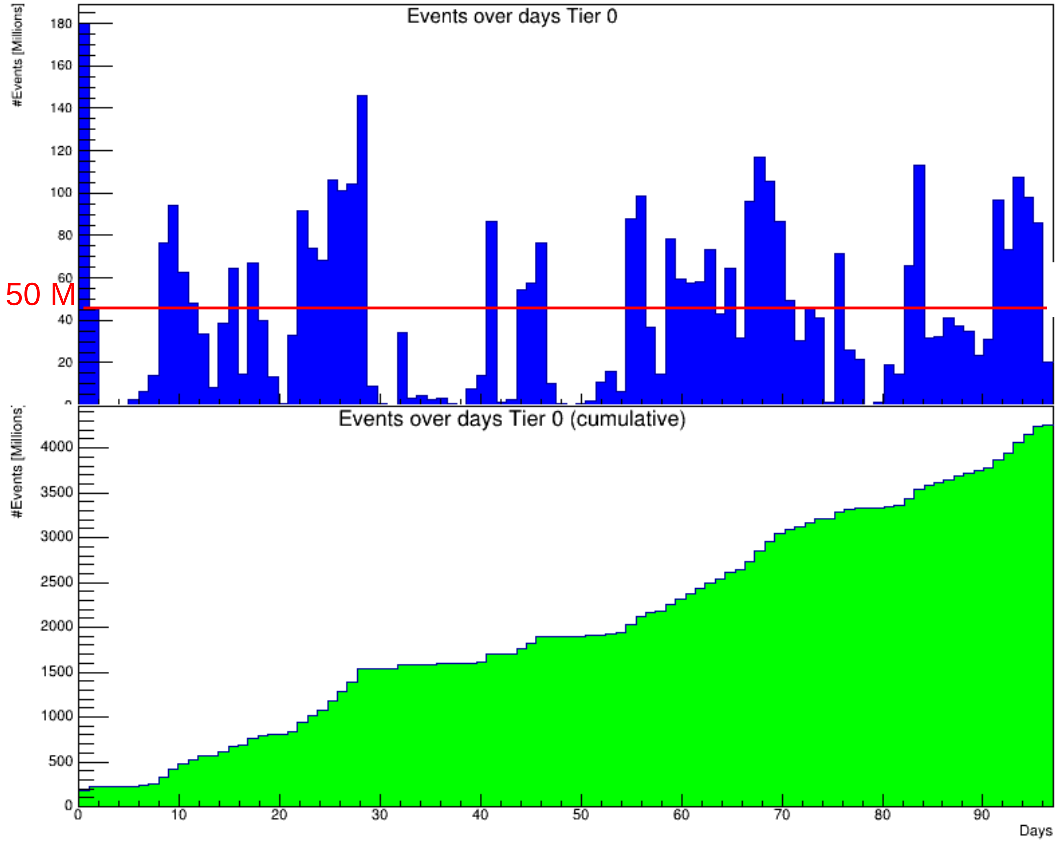
Id	Service	State	Message	Time
1	rucio	0	OK	2017-06-05 23:56:44
2	tier0	200	OK	2017-06-06 16:02:56
3	panda	-1	No JSON object could be decoded	2017-06-06 15:11:47
4	EOS	0	OK	2017-06-07 15:34:17

Showing 1 to 4 of 4 entries

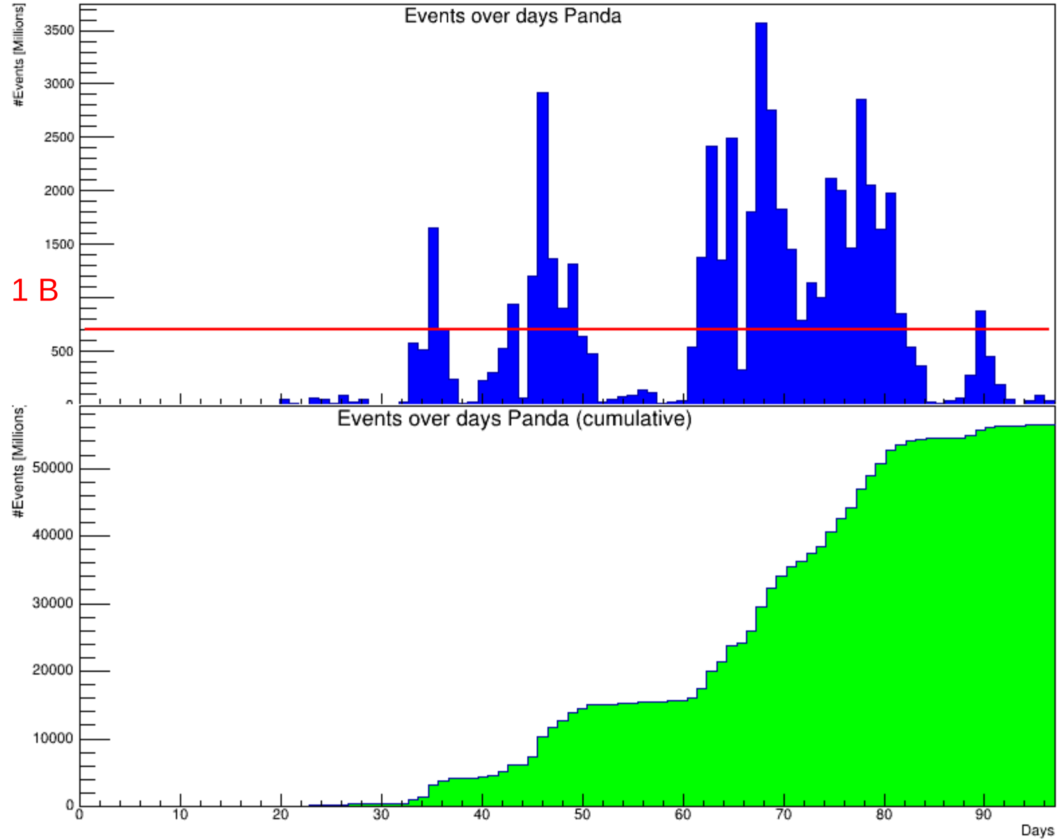
Previous 1 Next

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.

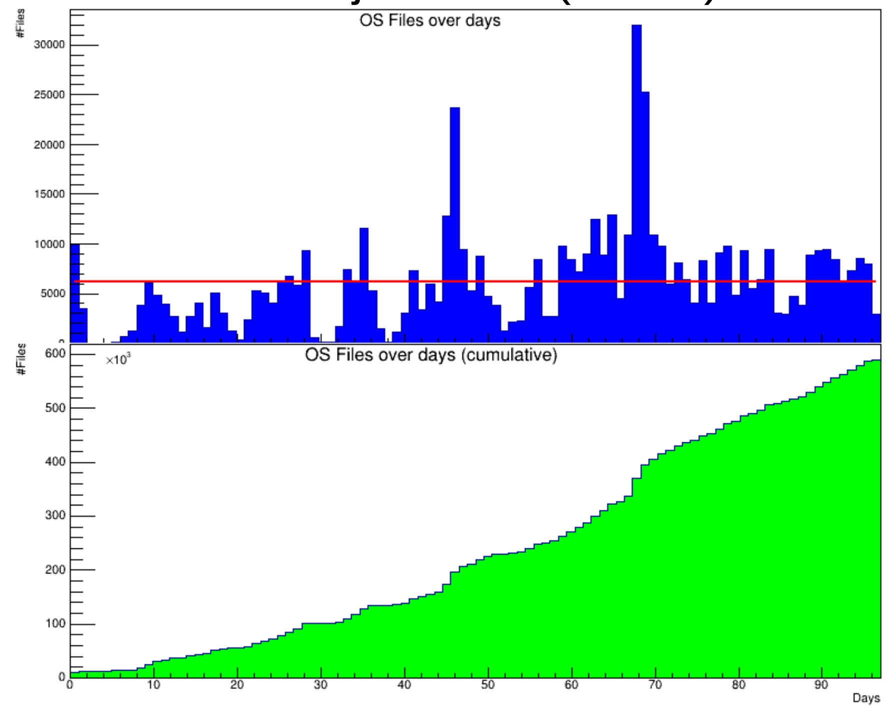
Data Ingestion to Hadoop

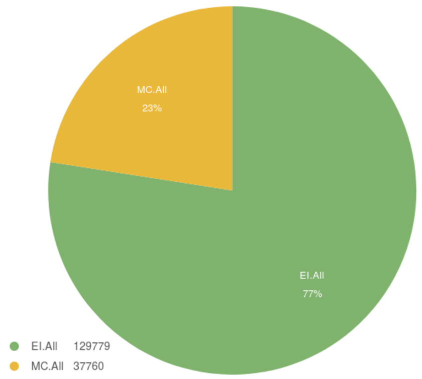
- Since mid 2017 DataCollection uses the Object Store to temporary maintain the eventindex files created by the Producers:
  - Creating ~7 K objects/day (~22 GB) that contain the ~1 B events indexed per day.
  - Data is factor 10 compressed with respect to original eventindex data.
  - Current ObjectStore data (March 2018) 1.335.000 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)





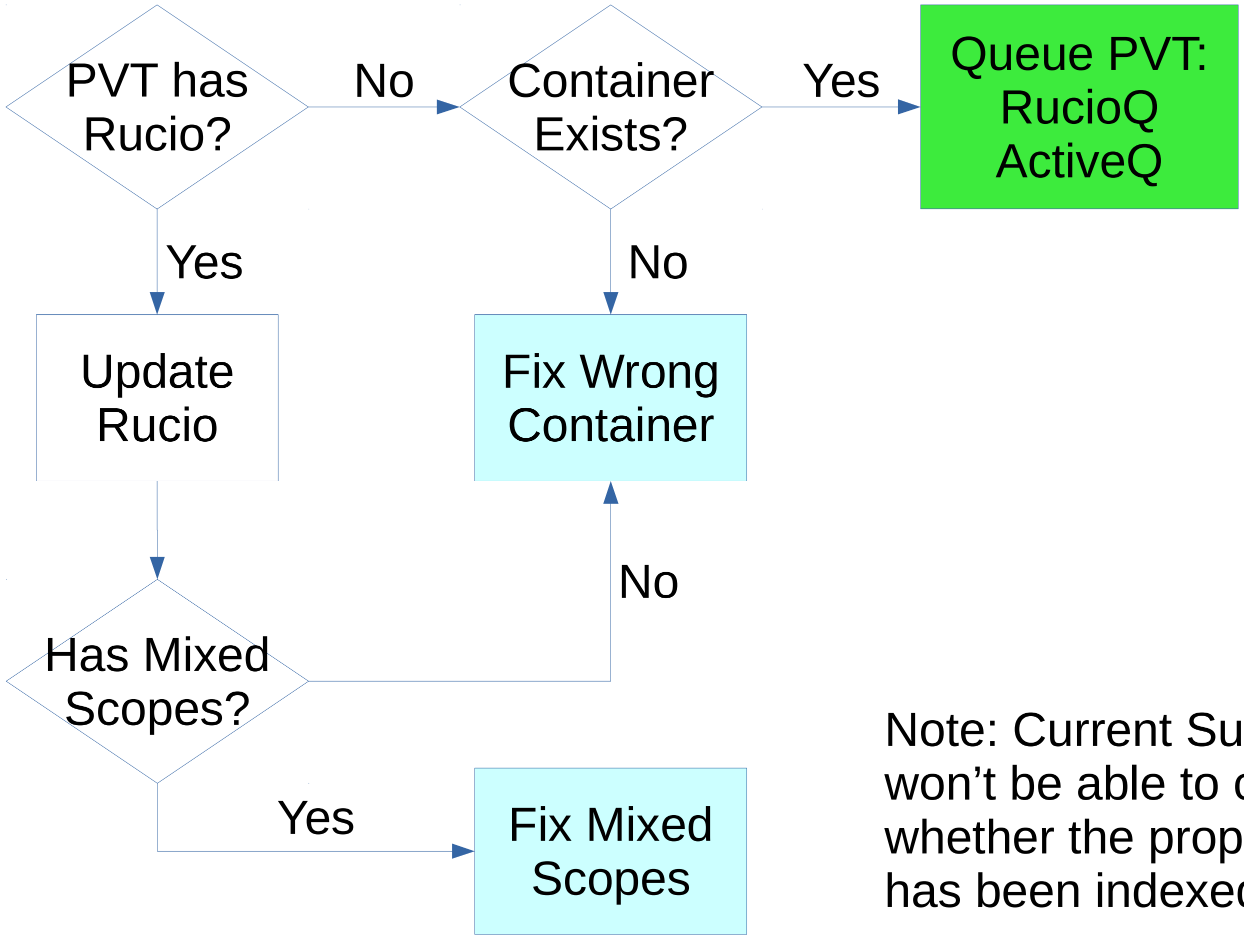
Hadoop (HDFS)

The Container Name Issue

- Current supervisor doesn’t certainly know about the intended containers to be indexed:
  - It receives messages from Panda tasks and it deduces the container name based on some assumed rules:
    - A dataset belongs to a single physics parent container.
    - The container name is the dataset name without the “_tid” part.
- Above rules have been broken:
  - Some datasets have multiple parent containers.
  - Some containers have datasets which scope mismatch.
  - Some datasets have merging steps that are not reflected into the name of its container.
- Many “Panda Virtual Tasks” (PVT), in charge of processing containers, were stuck because of this.

Short Term Action: Detect and Fix Containers with Issues

- Look for all Panda Virtual Tasks (PVTs) that have a PandaTask (PT) with jeditaskid but without Rucio information. For each PVT:



Note: Current Supervisor won't be able to certainly know whether the proper container has been indexed!

Fix Mixed Scopes

1) For each ds in Rucio:
   1) Retrieve all PTs by datasetname ordered by jeditaskid DESC.
      1) If there is none, register one and continue.
      2) Store the list of PVTs of those PTs in PVTSet, except the PVT itself.
      3) The first PT is considered the most recent. If it was in other PVT:
         1) Move it to the current PVT, if it is not being consumed. Fail if it is.
         2) Revalidate PT in the context of its new PVT if it was already consumed.
      4) The other PTs should either be moved to the PVT and declared obsolete if they had run, or should be removed if they were just registered but not ran.
2) For PVT in PVTSet, assert that it is empty and that it is not being consumed, or fail. Remove it.
3) For PT in PVT
   1) If PT still belongs to the Container, update its Rucio if it has none or has never run.
   2) If not, assert that PT no longer exist in Rucio and declare obsolete or remove if it has never run.
4) Reactivate PVT and PTs to validate.

Fix Wrong Container

1) Look for a PT of PVT that has run but has no Rucio.
   1) Get a single parent container discarding those of scope group.prj-evind.
      1) If there are more than one, discard those with the same scope of PVT.
      2) If at this point there is not exactly one parent, continue with other candidate PT.
   2) If the parent container corresponds to a know otherPVT, update Rucio in otherPVT and return fix_mixed_scopes(otherPVT)
   3) Otherwise, try to transform PVT to be the one that handles the parent container:
      1) Retrieve Rucio of the parent container.
      2) Assert that each task of PVT process a dataset of the parent container.
      3) For each ds of parent container:
         1) If a PT process ds, assert it must be in PVT or error.
         2) If no known PT process ds, register one.
      4) Revalidate consumed PTs in the context of the new PVT

Mid and Long Term Actions

- Register containers to be indexed and the supervisor itself attaches datasets to the technical container.
- Change the relationship between PT and PVT to be a many-to-many.
- Change the way consumer and data storage structures information:
  - Store events per dataset instead of per container.
  - Consider containers as aggregations of datasets.
- Migrate supervisor database from sqlite to other RDBMS which should less prone to locks due to concurrency.

Trigger Counter

What is the trigger?

- A trigger component represents a relevant condition that an event either satisfies or not:
  - L1: Signals with some threshold at the detector, depending on the detector configuration (SMK).
  - HLT: Software conditions, i.e. logical composition of L1 triggers.
- There are several arrays of trigger components for each level:
  - L1: TAV $\subseteq$ TAP $\subseteq$ TBP. Less than 512 possible trigger components per SMK.
  - HLT: PH, PT, RS. Less than 4096 possible trigger components per SMK.
- The trigger decides whether an event is permanently stored on disk/tape or just discarded.

What is AEI Trigger Counter?
What does it do?

- AEI Trigger Counter is a **web service** to provide meaningful information about the L1 and HL Trigger of datasets in a “human time” scale.
- It can **COUNT**, **LIST** and **OVERLAP** a particular trigger of a given dataset
  - Can filter by logical expression based on triggers, LB and/or BCID.
    - Expression is converted into **Reverse Polish Notation** and it is **evaluated**.
  - Can group results per LB and BCID.
- A map-reduce task runs and its results are collected, cached, and offered in up to three different formats:
  - An interactive plot is drawn when not listing.
  - The user can always download a CSV file with the results.
  - The raw data in JSON format is always offered to further process it, as it is done by the plot.

Trigger Counter Web Interface

Run Number: 334955

Stream: physics_Backgrour

Dataset: data17_13TeV.00334955.physics_Background.merge.AOD.f863_m1865

Operation: COUNT

Trigger: TAV

Expression:

Group by: Nothing

Limit:

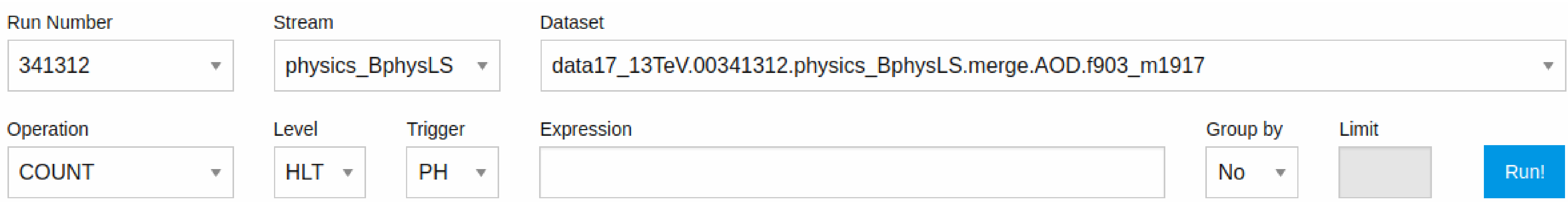
Run!

Show/Hide Triggers

0: L1_EM3	1: L1_EM7	2: L1_EM12	3: L1_EM8VH
4: L1_EM10VH	5: L1_EM15VHI_2TAU12IM_4J12	6: L1_EM15	7: L1_EM15VH
8: L1_EM18VHI	9: L1_EM20VH	10: L1_EM20VHI	11: L1_EM22VHI
12: L1_EM3_EMPTY	13: L1_EM7_EMPTY	14: L1_MU4	15: L1_MU6
16: L1_MU10	17: L1_MU21	18: L1_MU20	19: L1_MU4_EMPTY
20: L1_MU4_FIRSTEMPTY	21: L1_MU11_EMPTY	22: L1_MU4_UNPAIRED_ISO	23: L1_2EM3
24: L1_2EM7	25: L1_2EM10VH	26: L1_EM20VH_3J20	27: L1_2EM15
28: L1_2EM15VH	29: L1_EM7_2EM3	30: L1_EM12_2EM3	31: L1_EM15VH_3EM7
32: L1_2MU4	33: L1_2MU6	34: L1_2MU10	35: L1_2MU20_OVERLAY
36: L1_MU10_2MU6	37: L1_MU11_2MU6	38: L1_3MU4	39: L1_MU6_2MU4
40: L1_3MU6	41: L1_4J15.0ETA25	42: L1_EM15I_MU4	43: L1_2EM8VH_MU10
44: L1_EM15VH_MU10	45: L1_TAU12	46: L1_4MU4	47: L1_TAU12IM
48: L1_TE15.0ETA24	49: L1_TAU20	50: L1_EM12_W-MT25	51: L1_TAU20IM
52: L1_TAU100	53: L1_TAU30	54: L1_TAU40	55: L1_TAU60
56: L1_TAU8	57: L1_TAU8_EMPTY	58: L1_TAU20IM_2TAU12IM	59: L1_TAU20_2TAU12
60: L1_EM15VHI_2TAU12IM	61: L1_EM15VHI_2TAU12IM_J25_3J12	62: L1_EM15VHI_TAU40_2TAU15	63: L1_MU10_TAU12IM
64: L1_MU10_TAU12IM_J25_2J12	65: L1_EM7_MU10	66: L1_MU6_EMPTY	67: L1_MU10_TAU20IM
68: L1_XE10	69: L1_EM20VHI_TAU20IM_2TAU20_J25_3J20	70: L1_TAU20IM_2TAU12IM_J25_2J20_3J12	71: L1_J25_2J20_3J12_BOX-TAU20ITAU12I
72: L1_DR-MU10TAU12I_TAU12I-J25	73: L1_MU10_TAU12I-J25	74: L1_TAU20IM_2J20_XE45	75: L1_J15.31ETA49_UNPAIRED_ISO
76: L1_TAU60_DR-TAU20ITAU12I	77: L1_EM12_W-MT35_XS30_W-15DPHI-JXE-0_W-15DPHI-EMXE	78: L1_EM15VHI_2TAU12IM_XE35	79: L1_EM12_W-MT35_XS40_W-05DPHI-JXE-0_W-05DPHI-EMXE
80: L1_EM12_W-MT35_W-90RO2-XEHT-0_W-15DPHI-JXE-0_W-15DPHI-EMXE	81: L1_MU10_TAU12IM_XE35	82: L1_TE50	83: L1_TAU20IM_2TAU12IM_XE35

AEI TC Web Service

https://ei-atlas-dc.cern.ch/tc/



COUNT
LIST
OVERLAP

L1: TAV, TAP, TBP

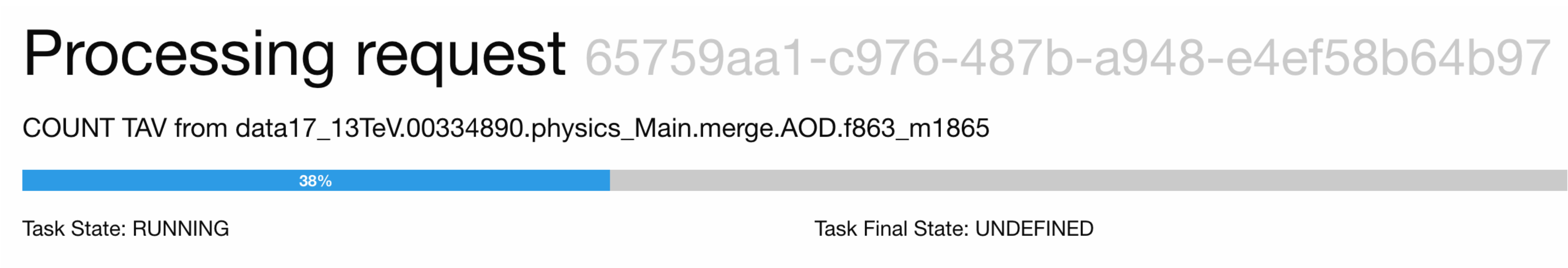
HLT: PH, PT, RS

Filtering over Triggers, LumiBlock and/or BCID

Group results by LB or BCID

Limiting Lists

Trigger Counter Web Interface: Progress of a Task

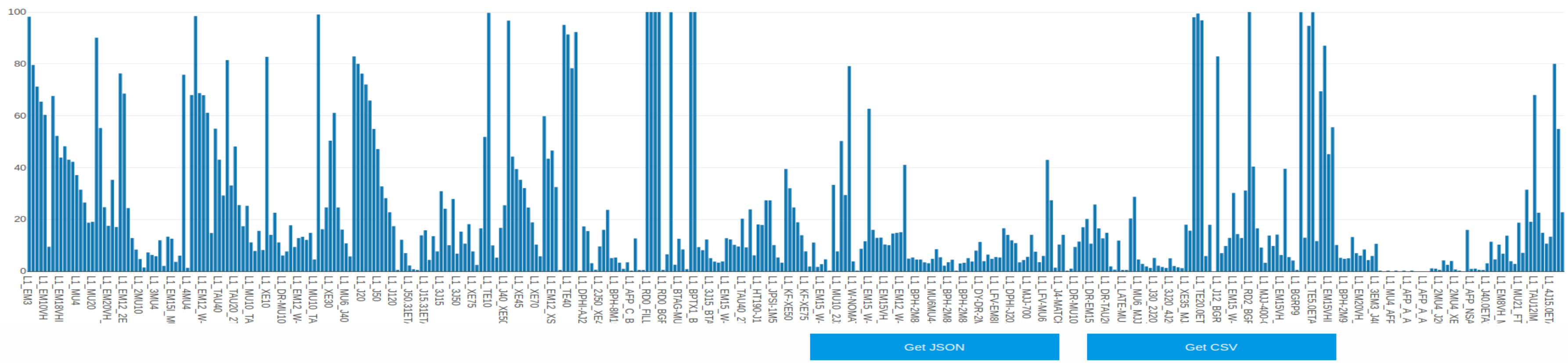


- User has output of the progress of its request, which refresh every 10 seconds.
- Map-reduce ‘state’ and ‘finalStatus’ are shown.
- If the task succeeds, the page redirects to the results.
- Trigger Counter caches the tasks, avoiding multiple runs of the same task.

COUNTing Triggers

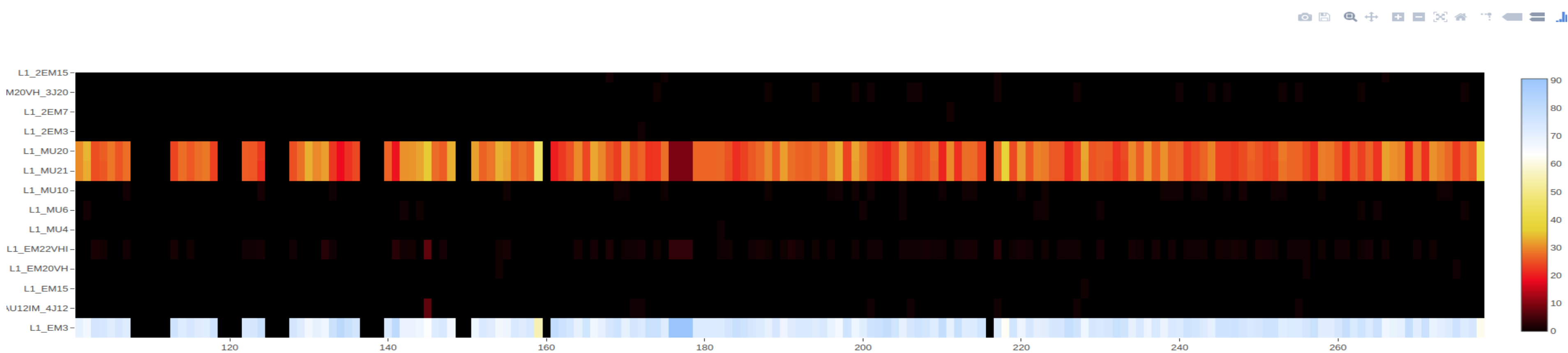
Results of 8e7f6deb-4180-4d17-884e-0ce058c574b6

COUNT TBP from data17_13TeV.00327103.physics_Main.merge.AOD.f832_m1812 where L1_EM3 or L1_MU21



Results of 84807e84-55db-469d-87af-5cc9595d9b59

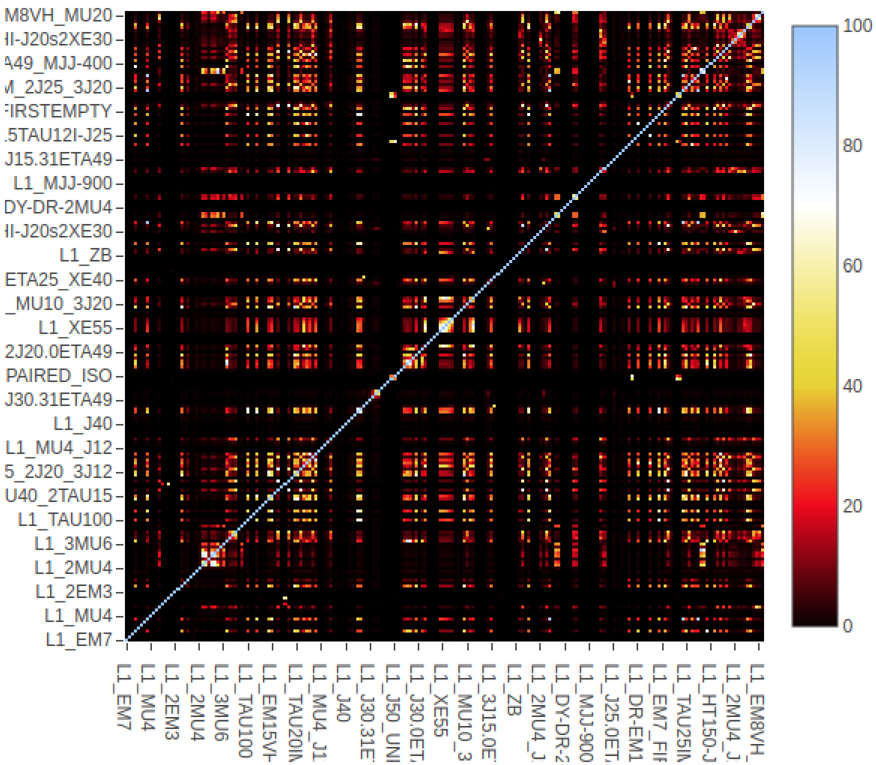
COUNT TAV from data17_13TeV.00327103.physics_Main.merge.AOD.f832_m1812 where L1_EM3 or L1_MU21 group by LumiBlock



OVERLAPing Triggers

Results of 58668579-dc70-46bd-b685-1dfd87a1b800

OVERLAP TAV from data17_13TeV.00327103.physics_Main.merge.AOD.f832_m1812 where not(L1_EM3 or L1_MU21)



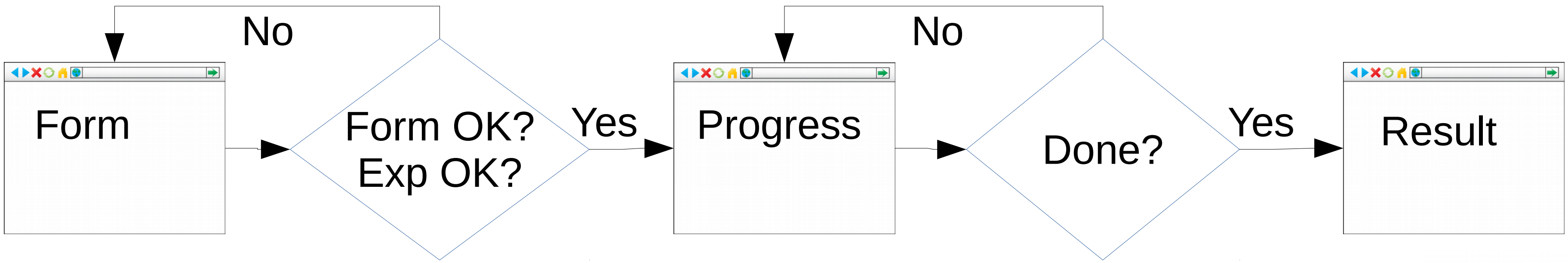
Get JSON

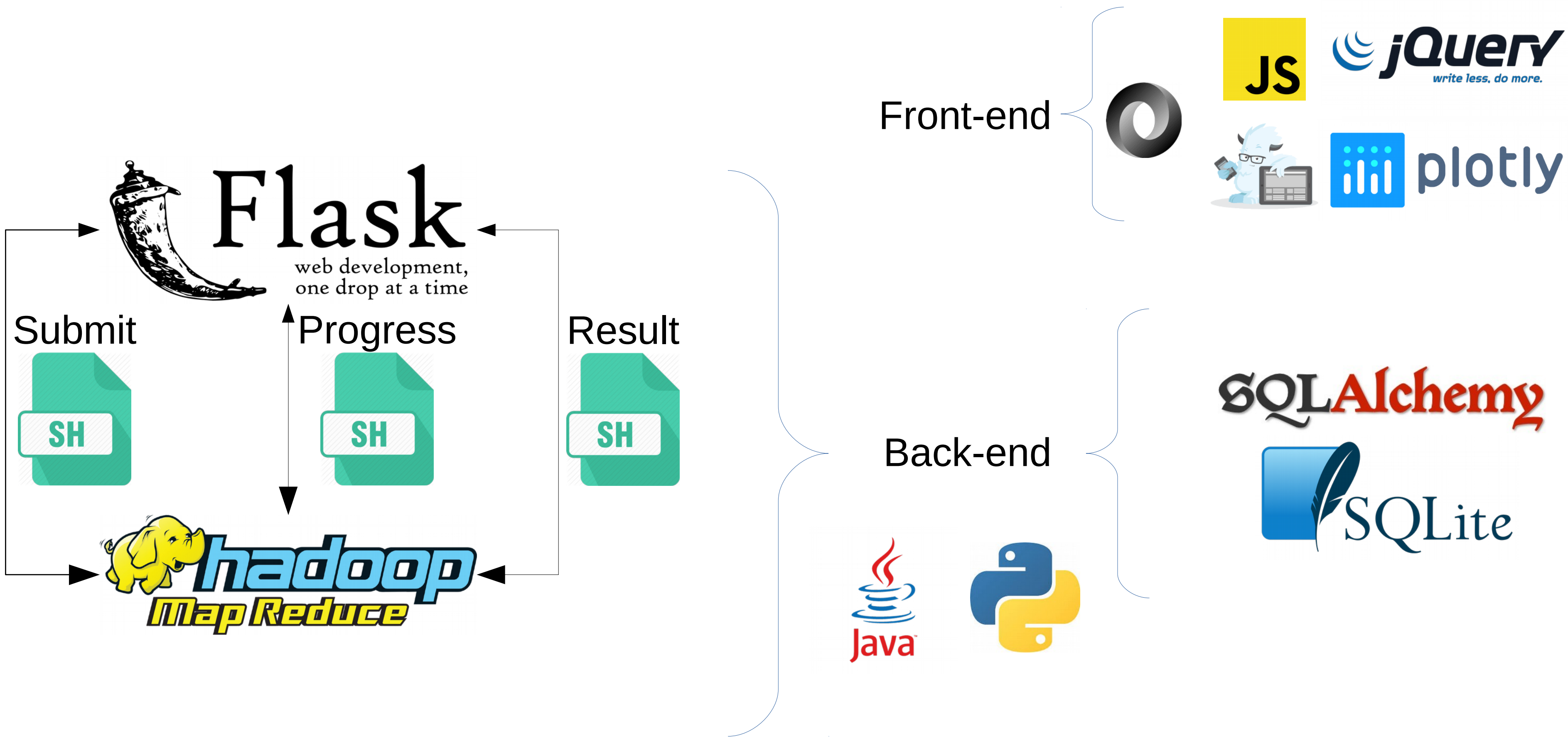
Get CSV

Behind the Scene

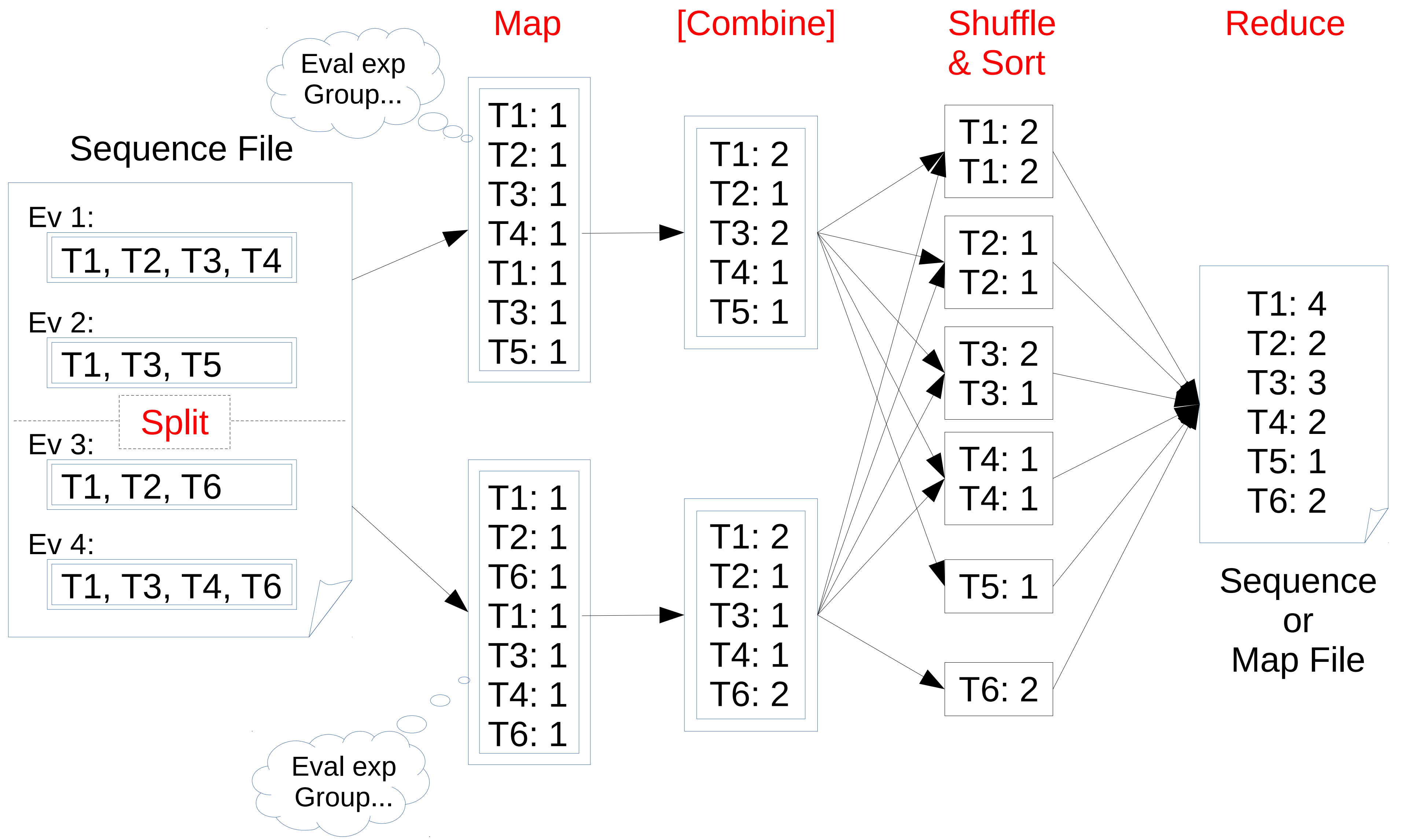
- The amount of data to store is huge and the way it is processed should not take long, but be responsive.
- We store trigger information in compressed Sequence Files using some tricks to further reduce their size.
  - **Sequence Files**: containers that contain data in key and value format.
  - The files contain minimal information of the trigger based on its properties, creating a file format which is very compact.
  - Their process can be **highly parallelizable**.
- A **MapReduce** task runs over a Sequence File.
  - Process information in parallel.
  - The proccess occurs **where the information is stored**:
    - No transfers.

Workflow & Tools





MapReduce


Map
[Combine]
Shuffle & Sort
Reduce
Eval exp Group...
Sequence File
Ev 1:
T1, T2, T3, T4
Ev 2:
T1, T3, T5
Split
Ev 3:
T1, T2, T6
Ev 4:
T1, T3, T4, T6
T1: 1 T2: 1 T3: 1 T4: 1 T1: 1 T3: 1 T5: 1
T1: 2 T2: 1 T3: 2 T4: 1 T5: 1
T1: 1 T2: 1 T6: 1 T1: 1 T3: 1 T4: 1 T6: 1
T1: 2 T2: 1 T3: 1 T4: 1 T6: 2
T1: 2 T1: 2
T2: 1 T2: 1
T3: 2 T3: 1
T4: 1 T4: 1
T5: 1
T6: 2
T1: 4 T2: 2 T3: 3 T4: 2 T5: 1 T6: 2
Sequence or Map File
Eval exp Group...

HDFS L1Trigger File Format

SEQ file with:

- Key: <LongWritable>   Value: <BytesWritable>

EventNumber		LB	BCID	L1Trigger

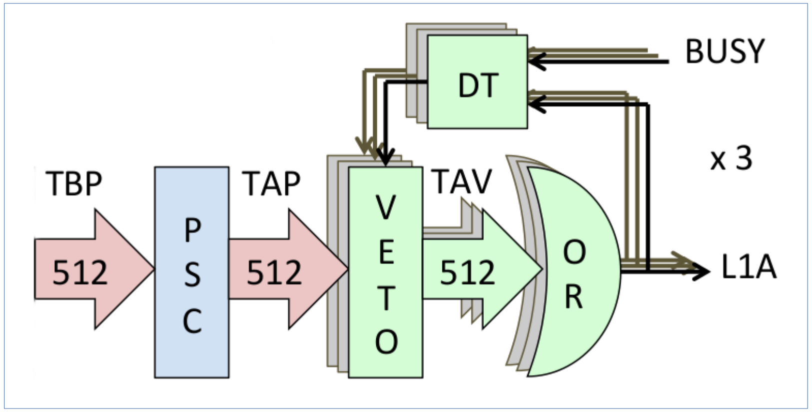
Less than 512 possible triggers, i.e. enumerating triggers, a trigger can be represented as 9 bits

- LB and BCID are uint32
- L1Trigger is a collection of packed 10 bits triggerNumbers into 32 bit words:

#TBP'	#TAP'	#TAV'		trig3	trig2	trig1	...		trigN	trig(N-1)

- Since TAV ⊆ TAP ⊆ TBP :

  TAV' = TAV
  TAP' = TAP - TAV
  TBP' = TBP - TAP



HDFS HLTrigger File Format

SEQ file with:

- Key: <LongWritable>   Value: <BytesWritable>

EventNumber	LB	BCID	HLTrigger

Less than 4096 possible triggers, i.e. enumerating triggers, a trigger can be represented as 12 bits

- LB and BCID are uint32

- HLTrigger is a collection of packed 12 bits triggerNumbers into 64 bit words:

	#PH	#PT	#RS		trig5	trig4	trig3	trig2	trig1	...					trigN	trig(N-1)

- There is no inclusion between HLT sets, therefore all triggers have to be stored

Trigger Sizes and Compression

Typical trigger counts and final Lz4 size in bytes per event:

dsname	TBP/e	TAP/e	TAV/e	TBP'/e	TAP'/e	TAV'/e	totTrig	lz4
data17_13TeV.00331951.physics_Background.merge.AOD.f848_m1844	13.1	10.6	1.1	2.4	9.6	1.1	13.1	15.6
data17_13TeV.00333994.physics_BphysLS.merge.AOD.f859_m1860	62.3	45.3	6.4	17.1	38.8	6.4	62.3	53.6
data17_13TeV.00324839.physics_CosmicCalo.merge.AOD.f823_m1799	7.2	4.6	1	2.6	3.6	1	7.2	11.2
data17_13TeV.00335302.physics_EnhancedBias.merge.AOD.f869_m1870	49.5	34	3	15.5	31	3	49.5	46.4
data17_13TeV.00326923.physics_Late.merge.AOD.f832_m1812	11.4	7.8	3.8	3.7	4	3.8	11.5	12.4
data17_13TeV.00327761.physics_Main.merge.AOD.f832_m1812	84.6	57.8	8.8	26.8	49	8.8	84.6	67.6
data17_13TeV.00334191.physics_MinBias.merge.AOD.f859_m1860	9.4	5.4	1	4	4.4	1	9.4	12
data17_13TeV.00324857.physics_ZeroBias.merge.AOD.f823_m1799	21.9	20.4	1	1.4	19.4	1	21.8	23.2

Compression ratios:

- Deflate 3.0
- Bzip2 4.5
- Lz4 1.9
- Snappy 1.8

⇨ LZ4 has been chosen because:

- It offers a good compression ratio.
- It is fast on decompression.

What do we have?

- Above 10 billion real events.
- Up to 175 million events into a single dataset.
- 1725 datasets that belong to 2017 real data
  - HLT: 367 GB
  - L1: 518 GB
- 375 Runs for several streams:
  - physics_Background
  - physics_BphysLS
  - physics_CosmicCalo
  - physics_Main
  - physics_MinBias
  - physics_ZeroBias