

ATLAS EventIndex: A Catalogue of all ATLAS Events.

Álvaro Fernández Casani
Carlos García Montoro
Santiago González de la Hoz

José Salt Cairols
Javier Sánchez
Miguel Villaplana Pérez





The ATLAS EventIndex

ATLAS produces **billions of events per year**.

- Events are stored in files.
- Files are grouped into datasets.
- Rucio is the distributed data management system that keeps track of files, datasets, their metadata and their replicas.
 - But Rucio does not handle event-level information.
- **EventIndex is the catalogue of all the ATLAS events.**

EventIndex Contents

	Event ID
	Trigger Decision
	Self Reference 
	Ref. to RAW 
	Ref. to AOD 
	...

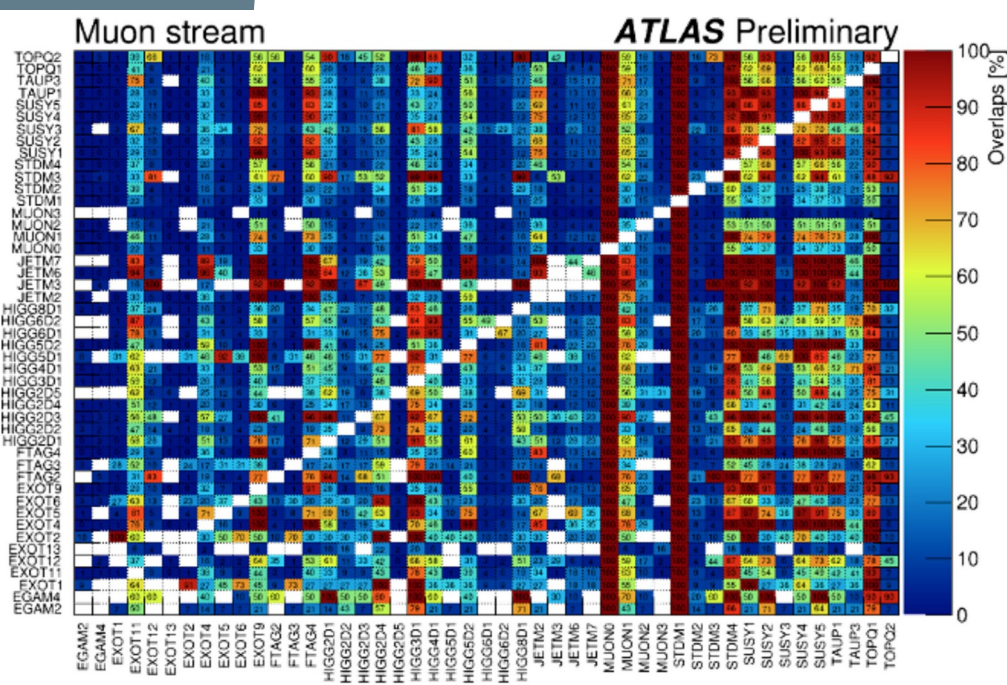
Event Records of immutable event information:

- Run and event number.
- Its own location
 - GUID, oid1, oid2.
- Its provenance.
- Luminosity Block.
- Bunch Crossing Id (BCID).
- Trigger information*.
- MC information*.

References to the events at each processing stage in all permanent files generated by central productions.

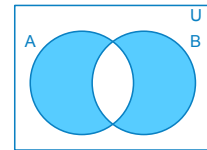
Use Cases

- **Event Picking:**
 - Give me this event in a specific format and processing version.
- **Counts or selections.**
 - Based on trigger decisions...
- **Overlaps:**
 - Of triggers in a dataset.
 - Of events between derivations.
- **Production checks:**
 - Completeness.
 - Duplicates detection.





SQL XOR NoSQL?



Before EventIndex (Run 1):

Tag Database

- Somewhat different use cases.
- Huge Relational DB.
 - **ORACLE**[®]
- Included mutable information.
- Production using DDM.
- Performance **below expectations**.

SQL

End of 2012

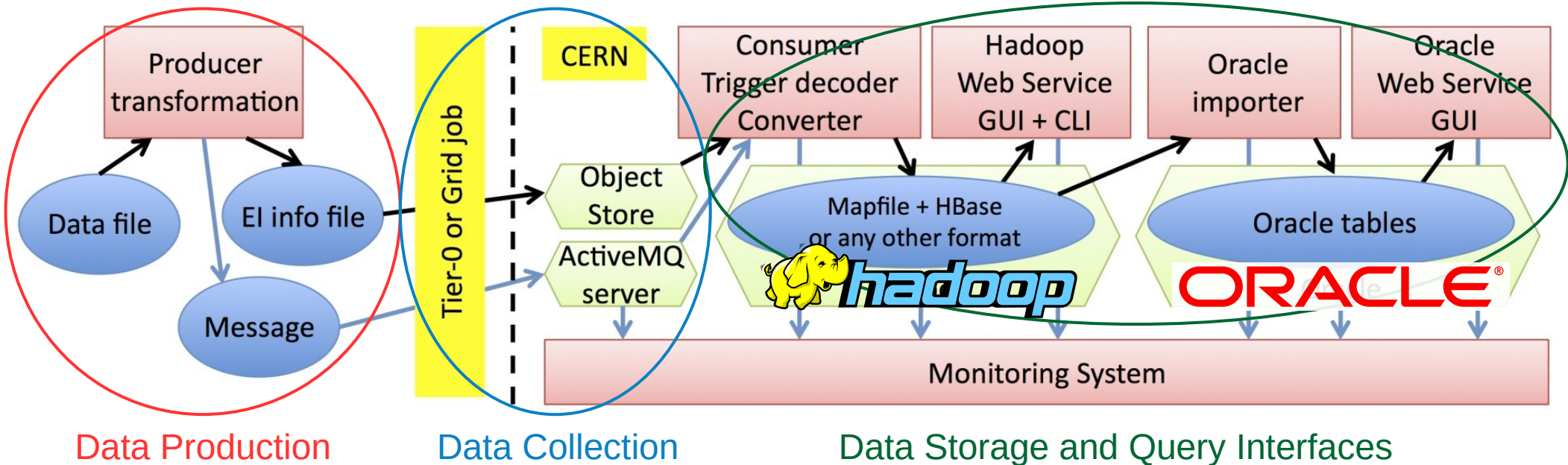
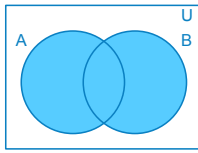
EventIndex

- Learn from previous mistakes.
 - Custom data production and data collection.
 - Only immutable info.
- Explore New BigData Technologies:
 -  **hadoop**
 - **Scalability** wrt. **data volumes**.
 - Apache: free software.

NoSQL



Architecture



Data Production

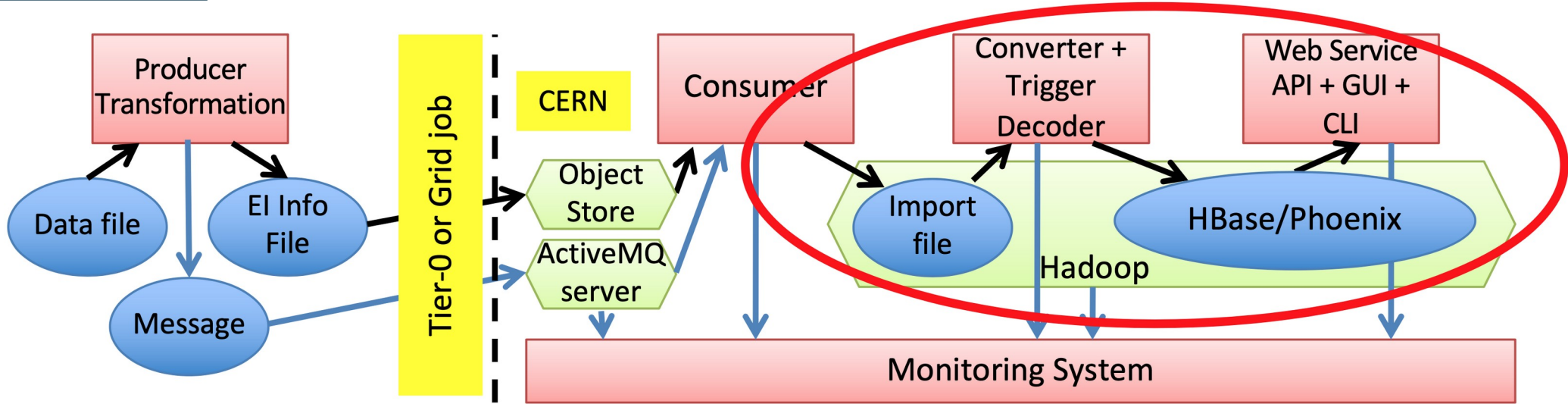
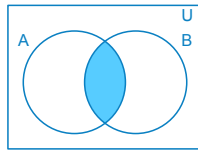
- Extract event metadata by means of Tier 0 and Grid jobs.
- Store production independently of ATLAS DDM.
- Coordinated by **IFIC**.

Data Collection

- Collect, validate and ingest production.
- Designed, developed and maintained by **IFIC**.

Storage & Query

- Permanent storage
 - Full info in Hadoop
 - HDFS MapFiles + HBase.
 - MapReduce.
 - Subset in Oracle.



2012 architecture mostly based on **HDFS**.

- Maturity of Hadoop.
- Flexibility.
- Catalogue in HBase.

New architecture based on



- Designed and developed by IFIC.
- Relational database over non relational distributed database.

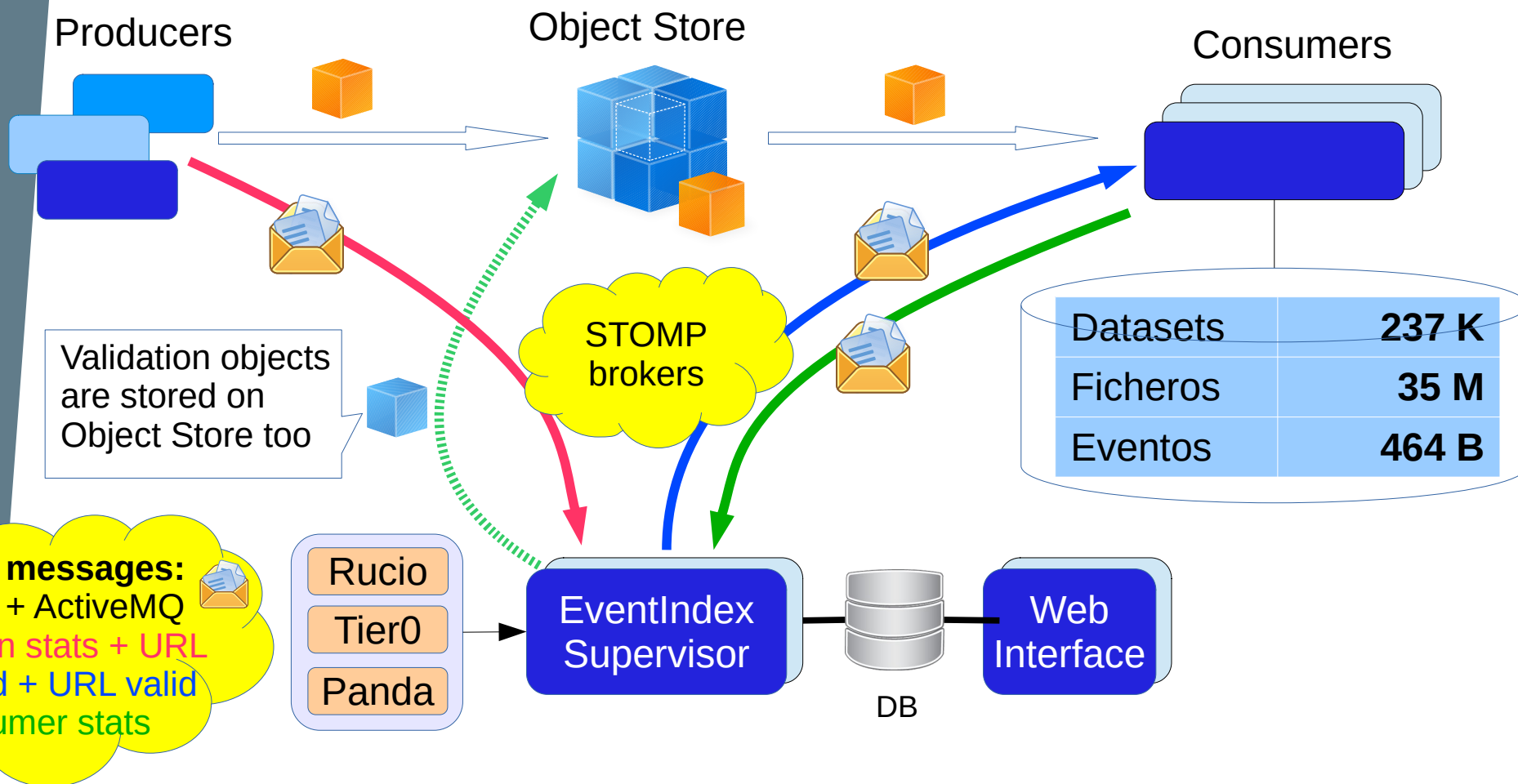
SQL AND NoSQL

Why?

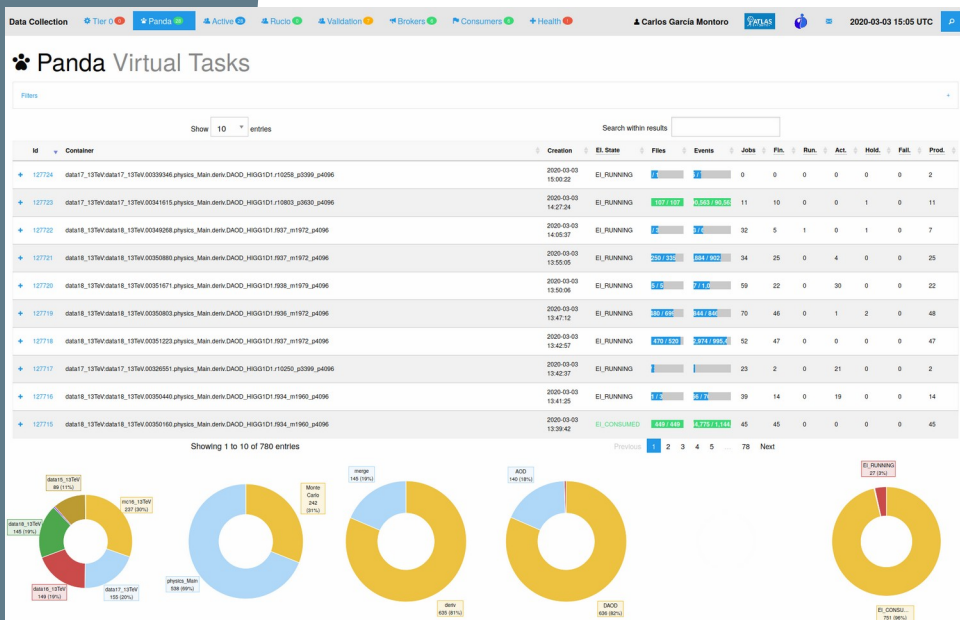
- **Data taking rates will increase:**
 - x3 during Run 3.
 - x10 for HL-LHC.
- **Target:**
 - **100B new real events/year.**
 - **300B new simulated events/year.**
- **HBase is mature:**
 - Distributed, non-relational database.
 - It provides fast lookups on top of HDFS.
- **Phoenix** provides SQL on top of Hbase.
- HBase organizes information. We decide:
 - **Families** carefully created to **reduce retrieval times.**
 - **Primary Key** wisely chosen to keep it **small** and performant, and to **maximize the locality of related information.**

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



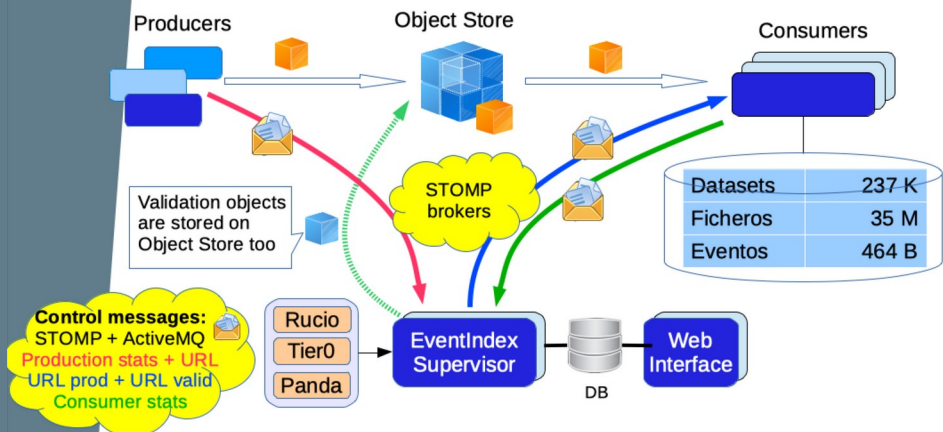
Data Collection Supervisor



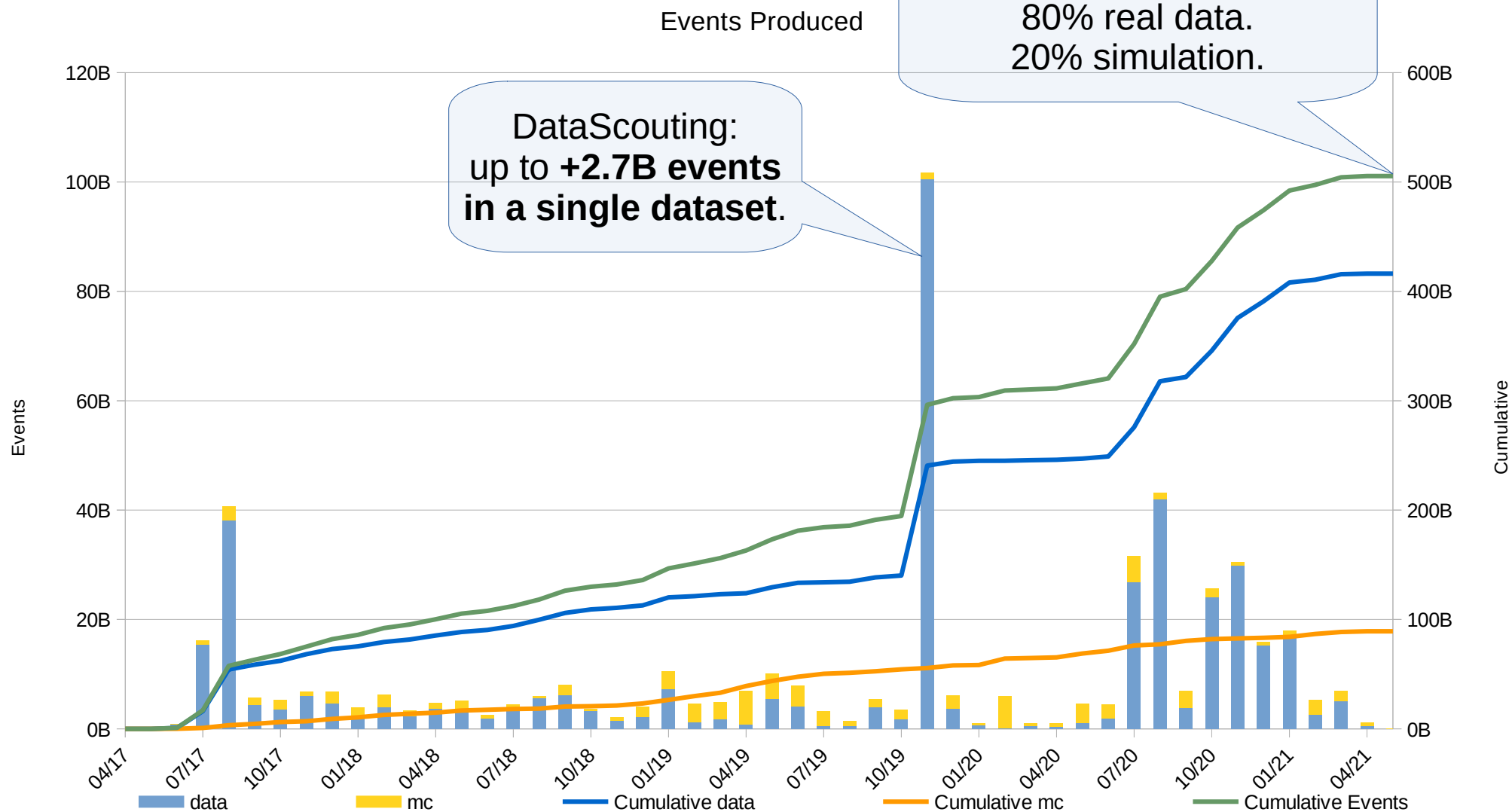
Key Component:

- It **orchestrates** the indexation procedure of all ATLAS events.
- It receives summarized information from the jobs that run at the WLCG.
- It **validates** the production, checking its correctness and completeness against Rucio.
- It **decides and commands** what should be stored in Hadoop.
- Designed, developed and maintained by **IFIC**.

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

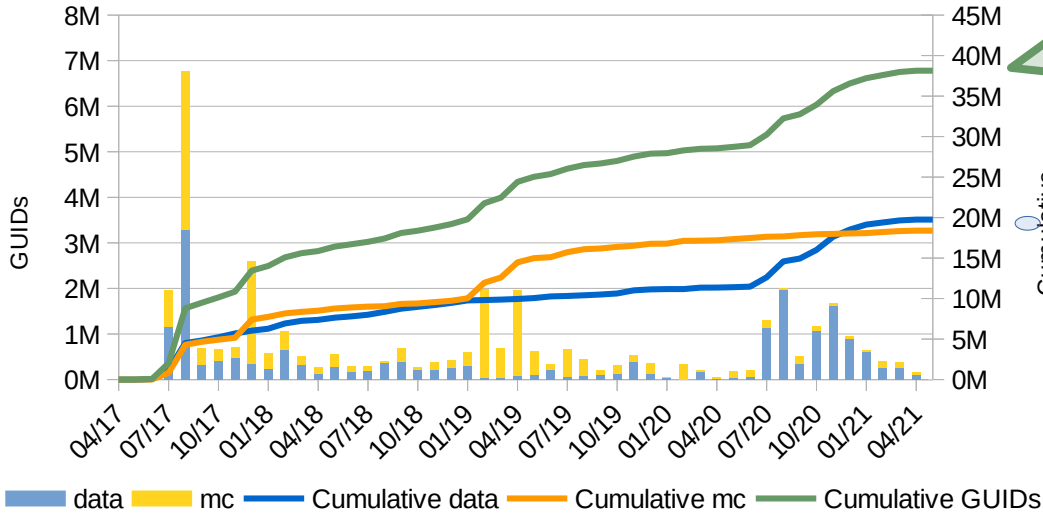


Event Records Processed



GUIDs and Jobs

Processed GUIDs

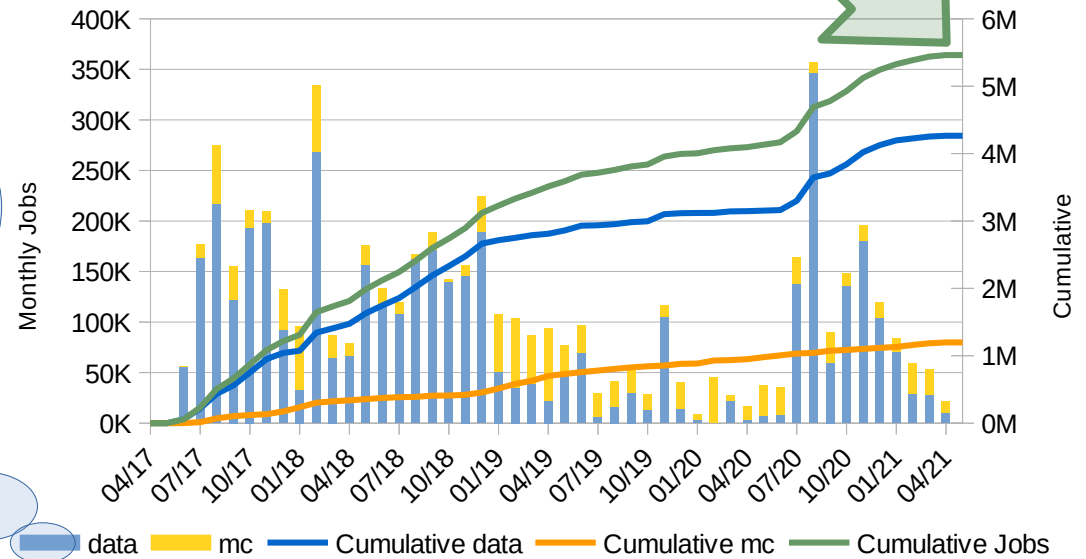


~40M GUIDs

Data ≈ mc
~20M GUIDs

~5,5M Jobs

Jobs



Far more data jobs.

- data jobs process less GUIDs than mc jobs.
 - Trigger.
 - Bigger GUIDs.

Trigger Counter

Run Number:

Stream:

Dataset:

Operation:

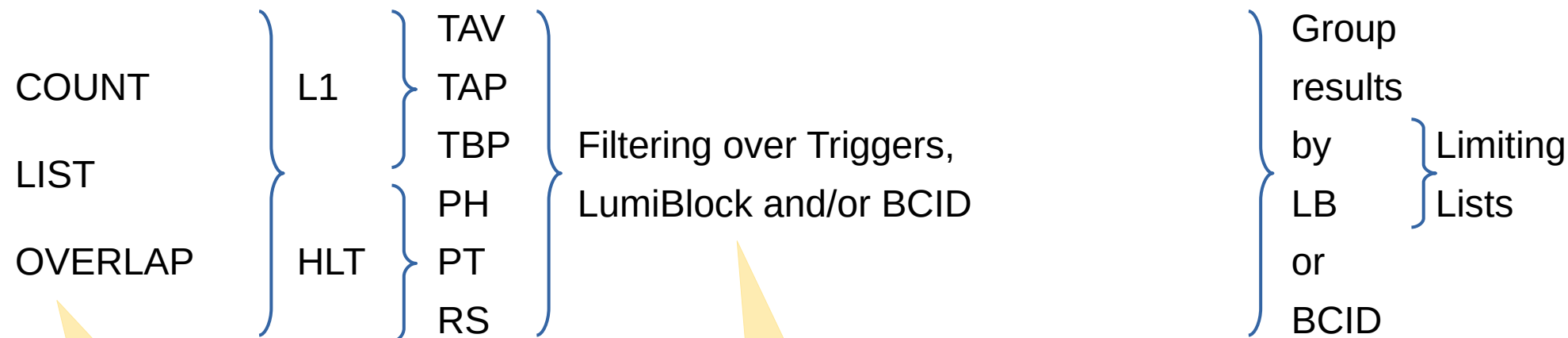
Level:

Trigger:

Expression:

Group by:

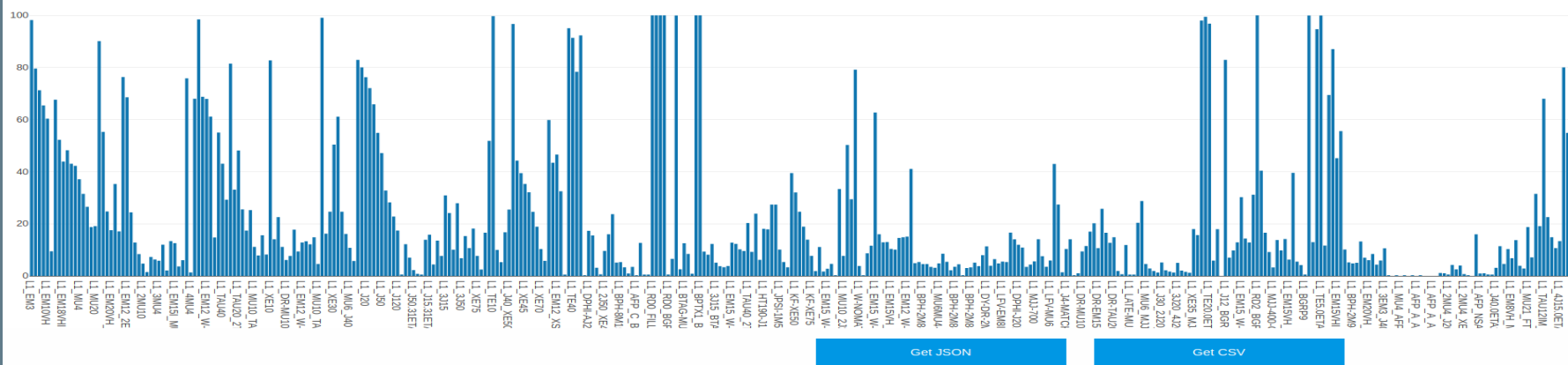
Limit:



COUNTing Triggers

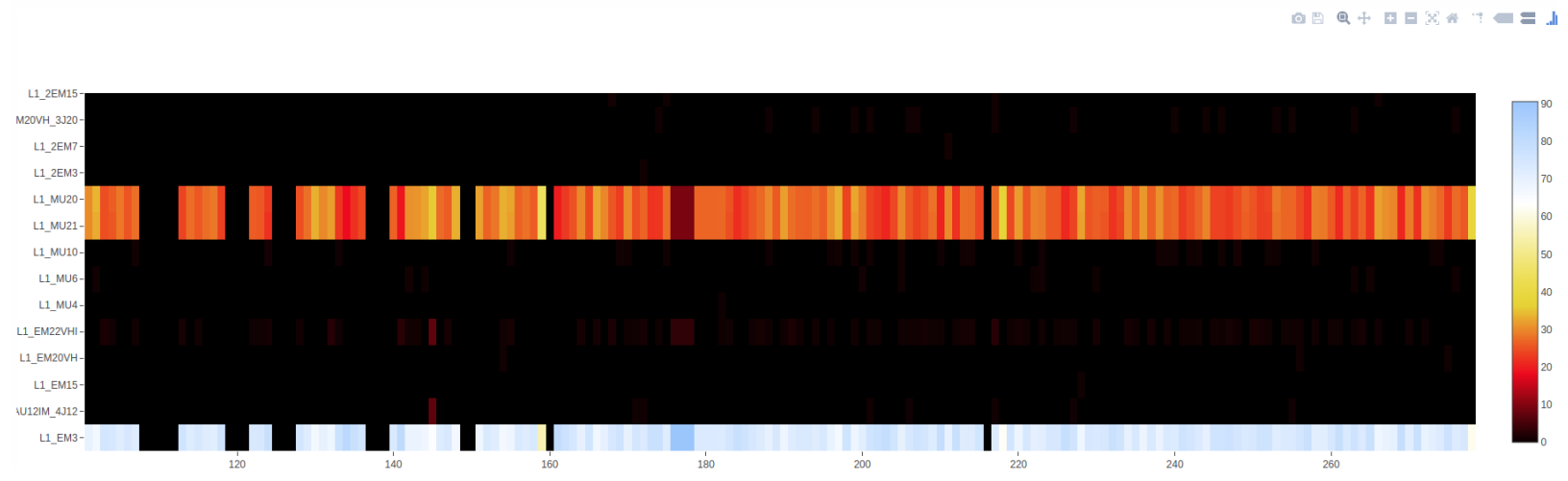
Results of 8e7f6deb-4180-4d17-884e-0ce058c574b6

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

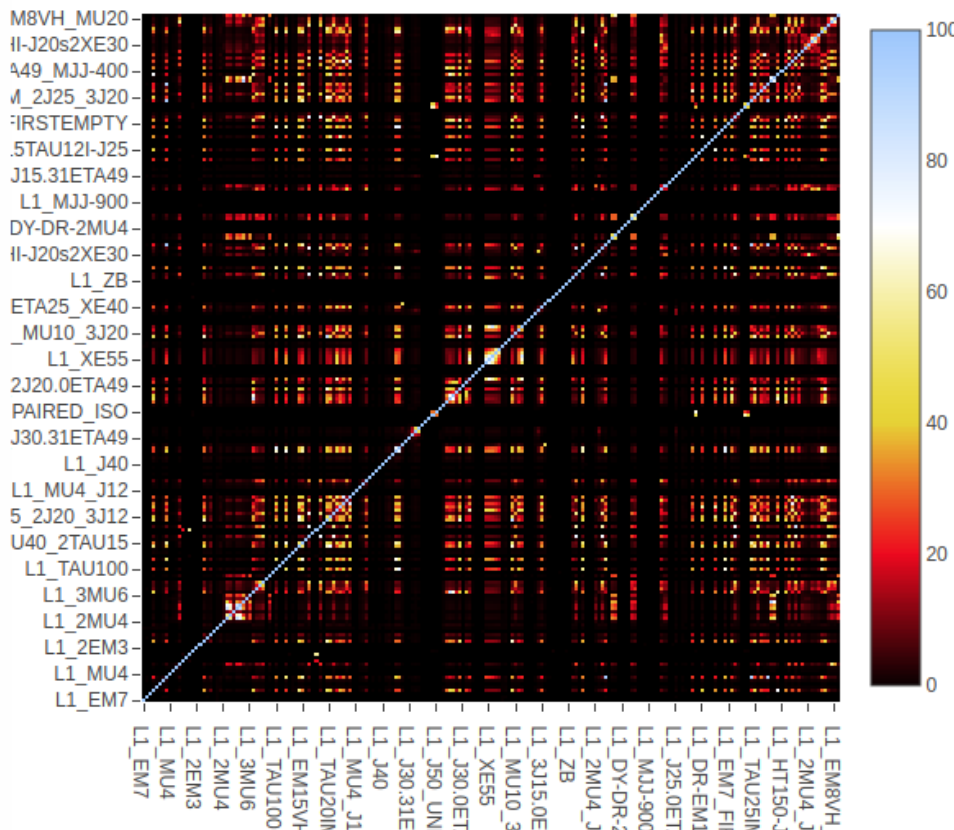
COUNT TAV from data17_13TeV.00327103.physics_Main.merge.AOD.f832_m1812 where L1_EM3 or L1_MU21 group by LumiBlock



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)



- Results available in a couple of minutes for all the operations.
- Callable from tools thanks to parametrized URLs.
- Result in different formats:
 - As interactive plots
 - Exportable as JSON.
 - Exportable as CSV files.

Get JSON

Get CSV



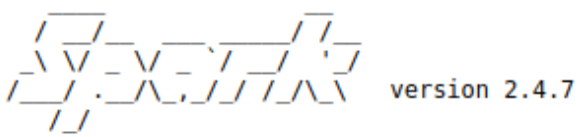
Analytics engine for large-scale data processing.



Strong statically typed general-purpose programming language:

- OOP +
- Functional Programming

Welcome to



Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_282)
 Type in expressions to have them evaluated.
 Type :help for more information.

```
scala> import eventindex.analytics.spark.Overlaps._
import eventindex.analytics.spark.Overlaps._

scala> val overlapsDF = calculateOverlaps(dspid=42328576)
overlapsDF: org.apache.spark.sql.DataFrame = [stream1: string, stream2: string ... 4 more fields]
```

scala> overlapsDF.show()

stream1	stream2	events_stream1_only	events_stream2_only	events_bothstreams	ratio
HIGG5D3	SUSY12	9699304	1064267	442599	0.0394960097874653
HIGG5D3	SUSY3	9535585	3980929	606318	0.0429317575964934
EX0T9	SUSY3	1259379	4326670	260577	0.04456878206336441
SUSY10	SUSY3	4079720	3996309	590938	0.06818279105020245
EX0T3	HIGG5D3	253665	9502989	638914	0.061460229974927776
EX0T3	EX0T9	806155	1433532	86424	0.03715385895170093
EX0T9	SUSY12	1416337	1403247	103619	0.03544707637478478
SUSY12	SUSY3	1437342	4517723	69524	0.011540040324742484
EX0T3	SUSY3	869074	4563742	23505	0.004307847723768...
EX0T3	SUSY10	420459	4198538	472120	0.09273406994967902
EX0T3	SUSY12	683780	1298067	208799	0.09531389370989196
EX0T12	EX0T3	300588	888856	3723	0.003120267322177...
HIGG5D3	SUSY10	8998119	3526874	1143784	0.0836785909961074
SUSY10	SUSY12	3550036	386244	1120622	0.22160247519133255
EX0T9	SUSY10	1155000	4305702	364956	0.06264631394427891
EX0T12	SUSY12	267322	1469877	36989	0.020848410653211497
EX0T9	HIGG5D3	964152	9586099	555804	0.05004513303778885
EX0T12	SUSY10	227381	4593728	76930	0.015706285719652293
EX0T12	HIGG5D3	177721	10015313	126590	0.012266919802504432
EX0T12	EX0T9	214777	1430422	89534	0.05161255363217279

only showing top 20 rows

Conclusions

- **EventIndex is the Catalogue of All the ATLAS Events.**
- It uses **BigData technologies** to handle **hundreds of billions of event records.**
- It **continues evolving** using state of the art technologies.
- **IFIC plays a key role** in most of its components.

Run: 359058
Event: 2965933740
2018-08-25 02:51:44 CEST



IFIC

INSTITUT DE FÍSICA
CORPUSCULAR

Thank you!

Questions?

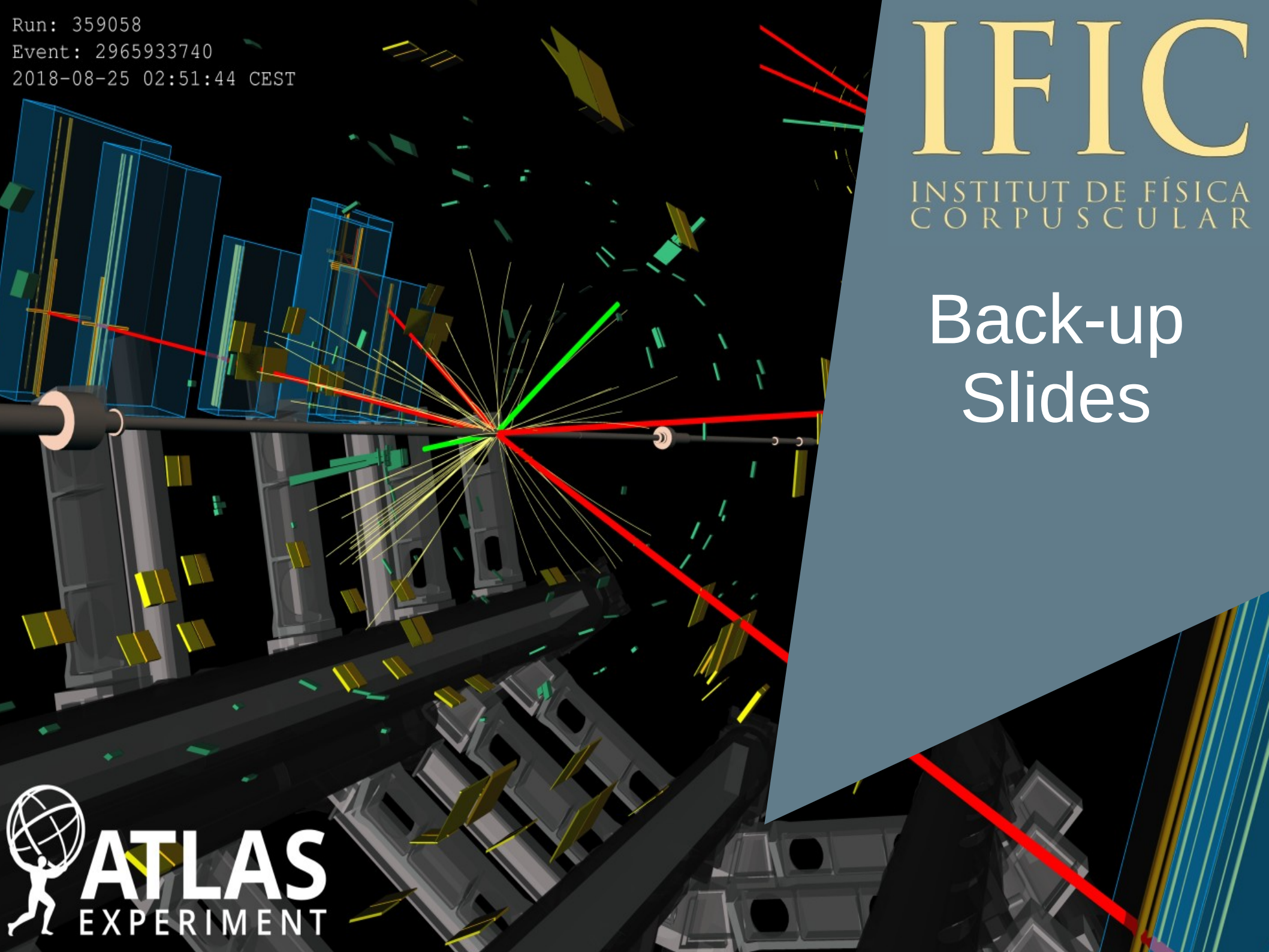


Run: 359058
Event: 2965933740
2018-08-25 02:51:44 CEST

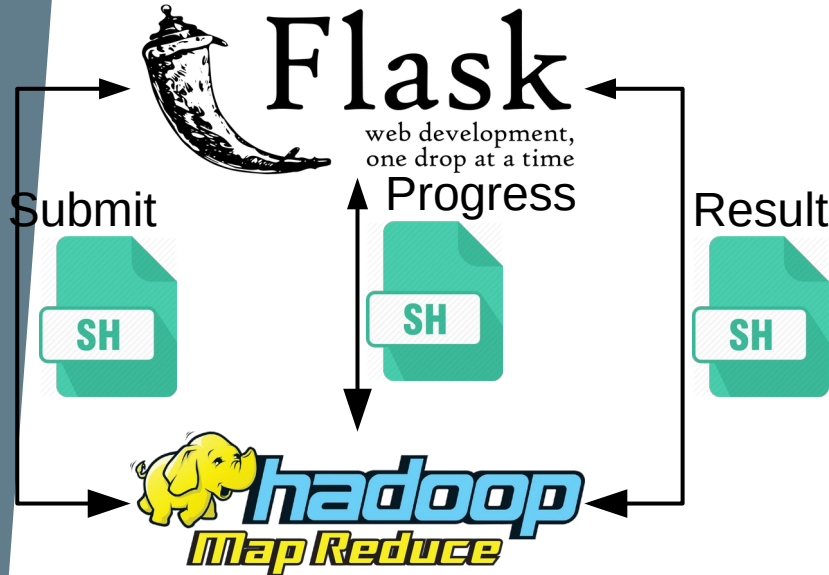
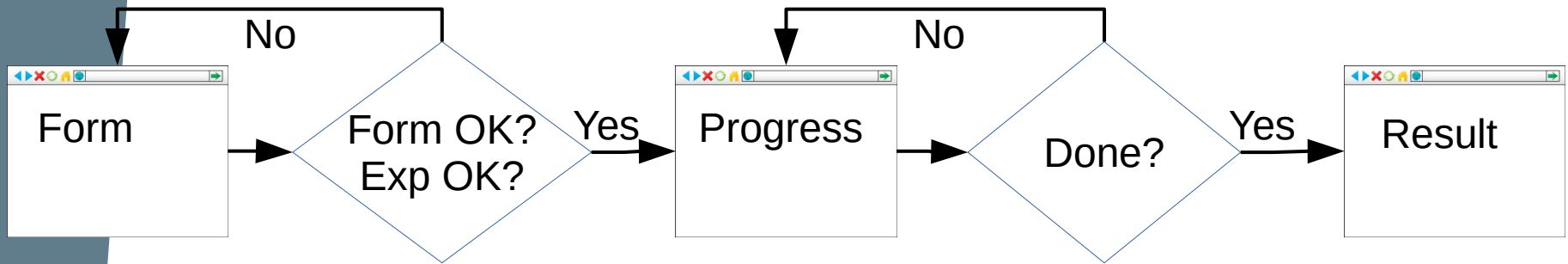
IFIC

INSTITUT DE FÍSICA
CORPUSCULAR

Back-up Slides



Inside Trigger Counter



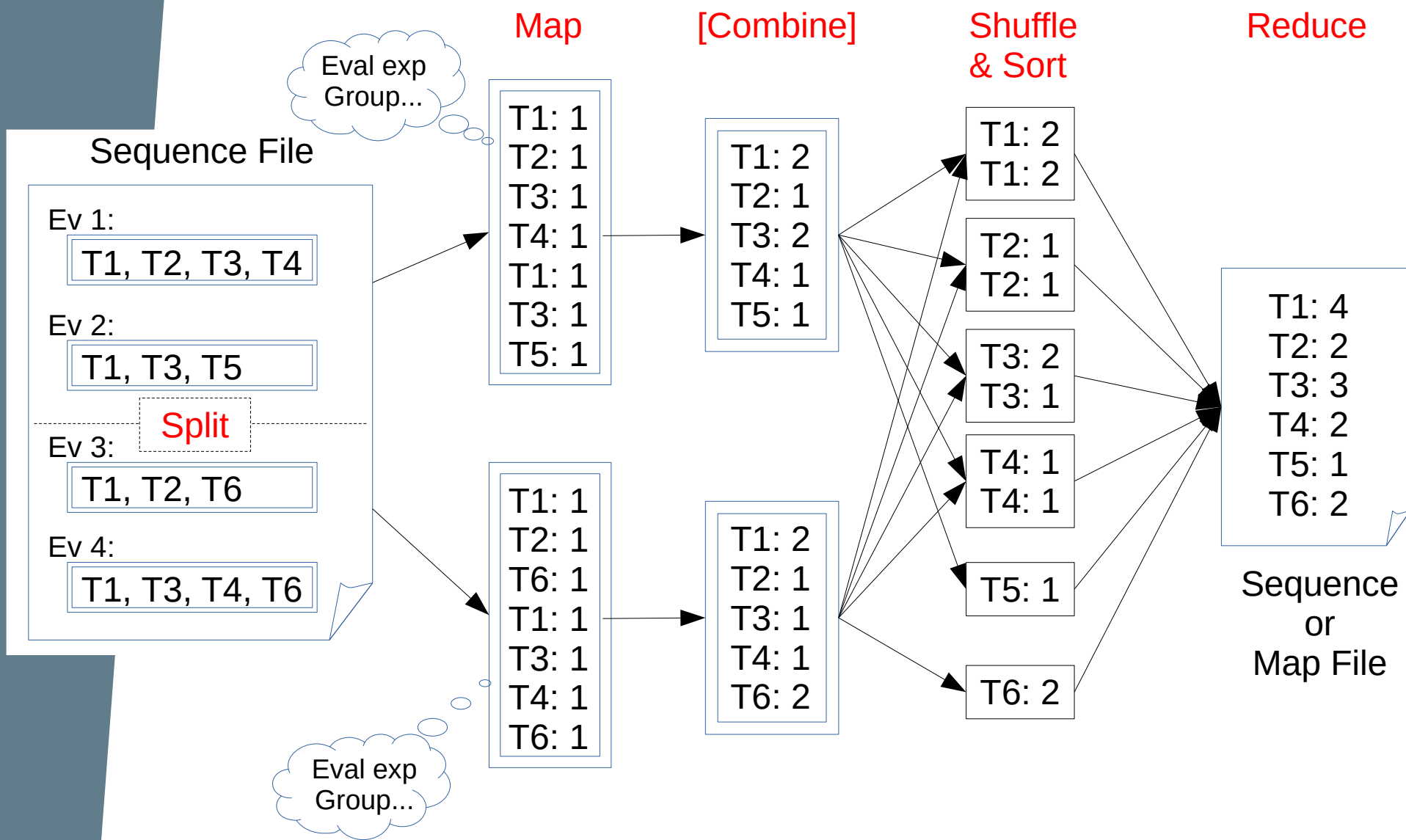
Front-end



Back-end



MapReduce



Example: Sum Values per Suit

