CERN

Leveraging Oracle Big Data Discovery to Master CERN’s Data

Manuel Martín Márquez

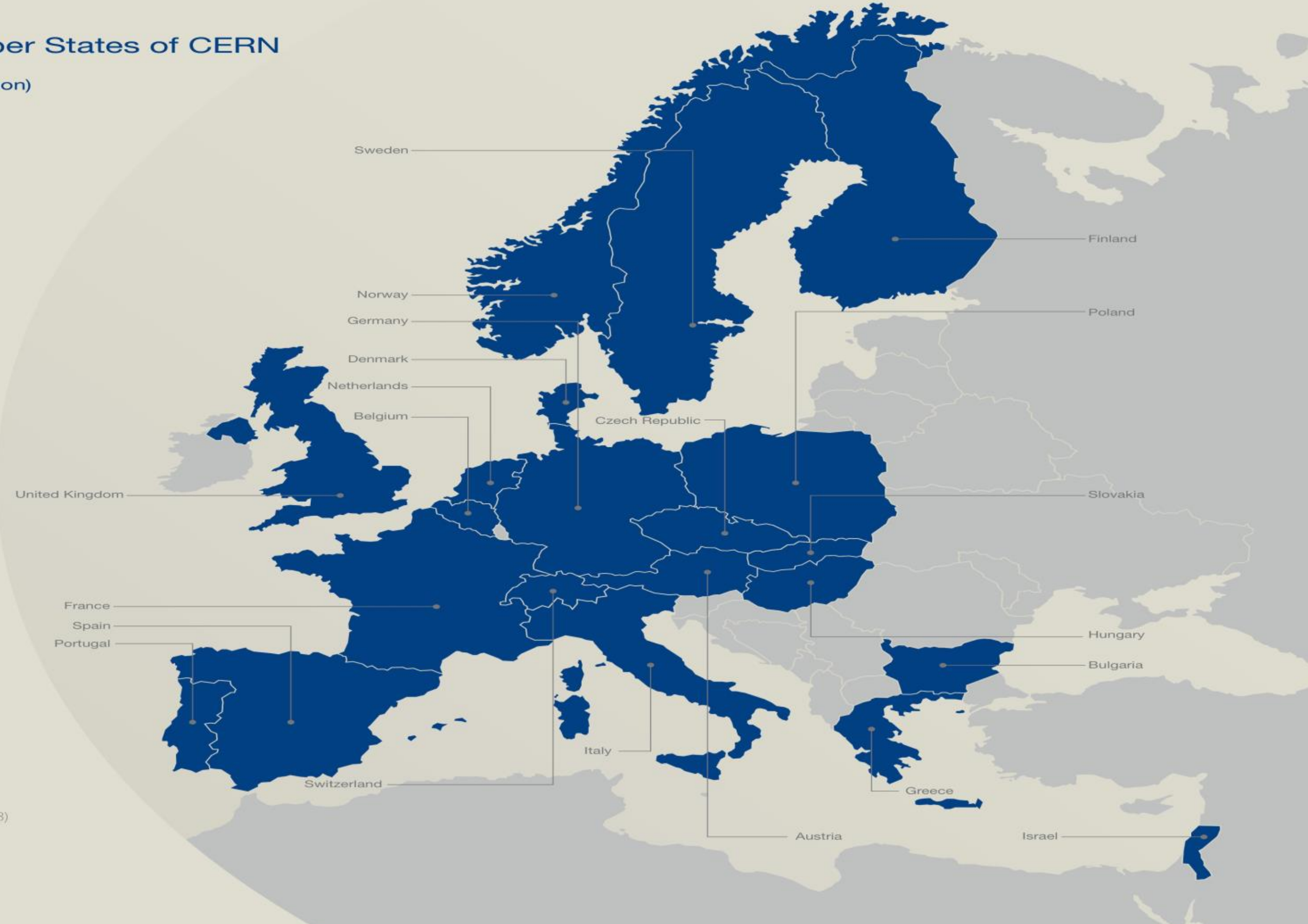
Oracle Business Analytics Innovation

12 October- Stockholm, Sweden

The twenty one Member States of CERN

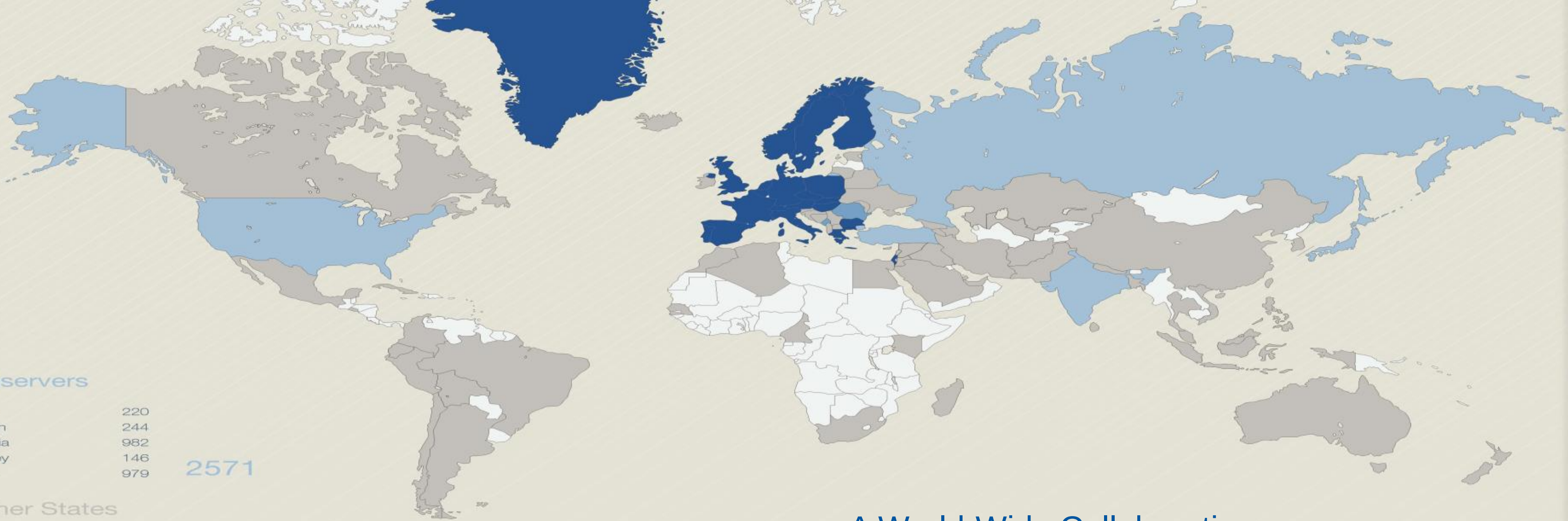
Member States (Dates of accession)

Austria (1959)
Belgium (1953)
Bulgaria (1999)
Czech Republic (1993)
Denmark (1953)
Finland (1991)
France (1953)
Germany (1953)
Greece (1953)
Hungary (1992)
Israel (2014)
Italy (1953)
Netherlands (1953)
Norway (1953)
Poland (1991)
Portugal (1986)
Slovakia (1993)
Spain (1/1961-12/1968-1/1983)
Sweden (1953)
Switzerland (1953)
United Kingdom (1953)



CERN - European Laboratory for Particle Physics

A World-Wide Collaboration



Member States

Austria	99	Greece	152	Slovakia	88
Belgium	106	Hungary	68	Spain	337
Bulgaria	75	Israel	51	Sweden	75
Czech Republic	202	Italy	1686	Switzerland	180
Denmark	53	Netherlands	153	United Kingdom	640
Finland	87	Norway	61		
France	751	Poland	229		
Germany	1150	Portugal	109		

6352

Candidate for Accession

Romania	118

Associate Members in the Pre-stage to Membership

Serbia	41

Observers

India	220
Japan	244
Russia	982
Turkey	146
USA	979

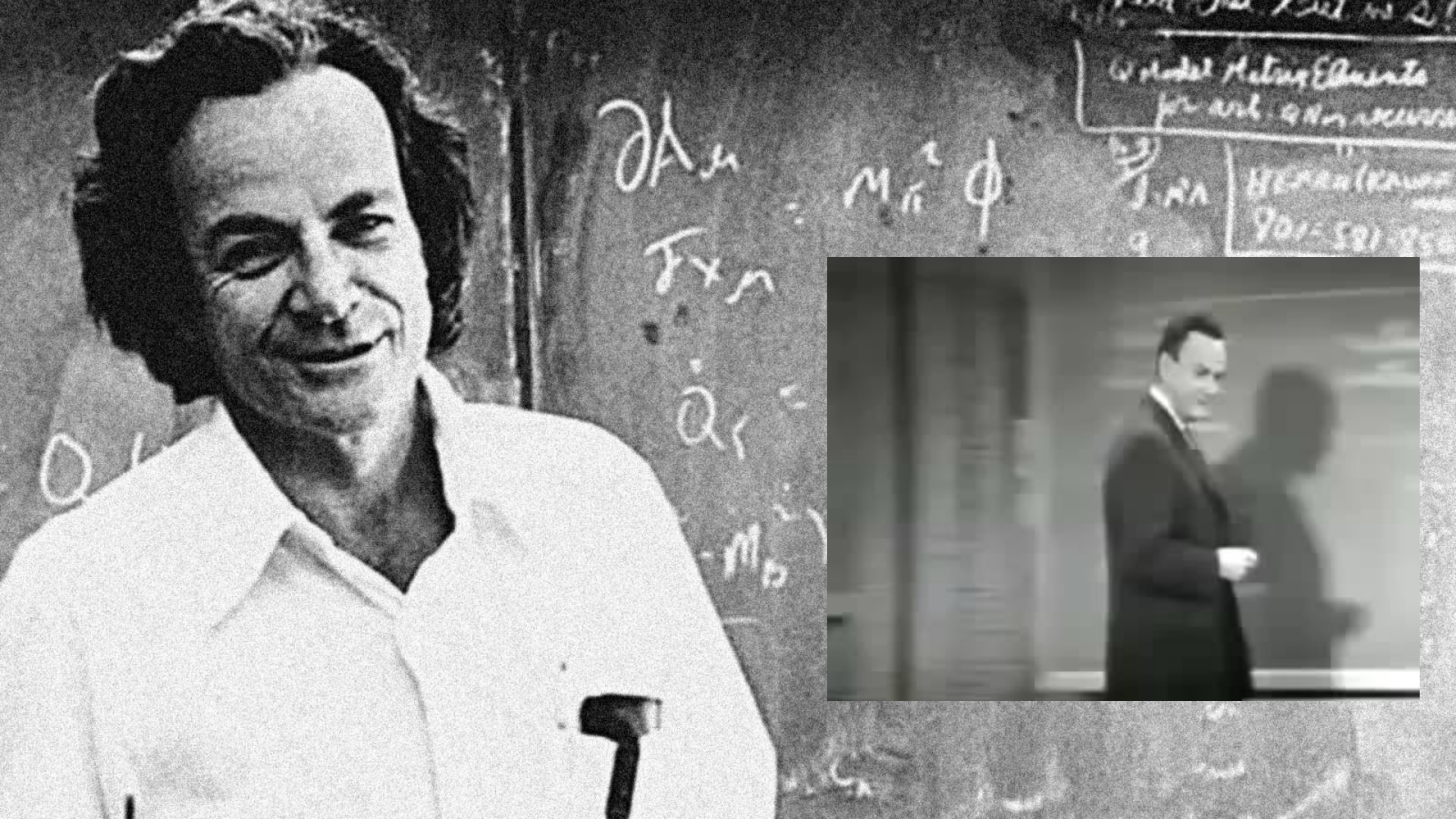
2571

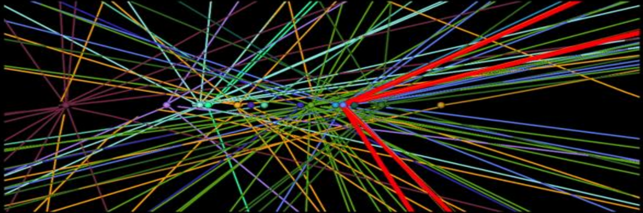
Other States

Afghanistan	1	El Salvador	1	Pakistan	41
Albania	2	Estonia	16	Palestine (O.T.).	4
Algeria	8	Georgia	36	Peru	8
Argentina	11	Gibraltar	1	Philippines	1
Armenia	25	Hong Kong	1	Saudi Arabia	3
Australia	25	Iceland	4	Senegal	1
Azerbaijan	8	Indonesia	1	Singapore	2
Bangladesh	4	Iran	28	Sint Maarten	2
Belarus	47	Ireland	22	Slovenia	27
Bolivia	3	Jordan	2	South Africa	16
Bosnia &		Kenya	1	Sri Lanka	5
Herzegovina	1	Korea, D.P.R.	1	Syria	2
Brazil	108	Korea Rep.	117	Thailand	12
Cameroon	1	Kuwait	1	T.F.Y.R.O.M.	1
Canada	134	Lebanon	12	Tunisia	6
Cape Verde	1	Lithuania	19	Ukraine	55
Chile	12	Luxembourg	4	Uzbekistan	4
China	280	Madagascar	4	Venezuela	9
China (Tapei)	45	Malaysia	15	Viet Nam	9
Colombia	30	Mauritius	1	Zimbabwe	2
Croatia	35	Mexico	64		
Cuba	7	Montenegro	3		
Cyprus	16	Morocco	12		
Ecuador	3	Nepal	5		
Egypt	19	New Zealand	7		

1415

Distribution of All CERN Users by Nationality on 14 January 2014

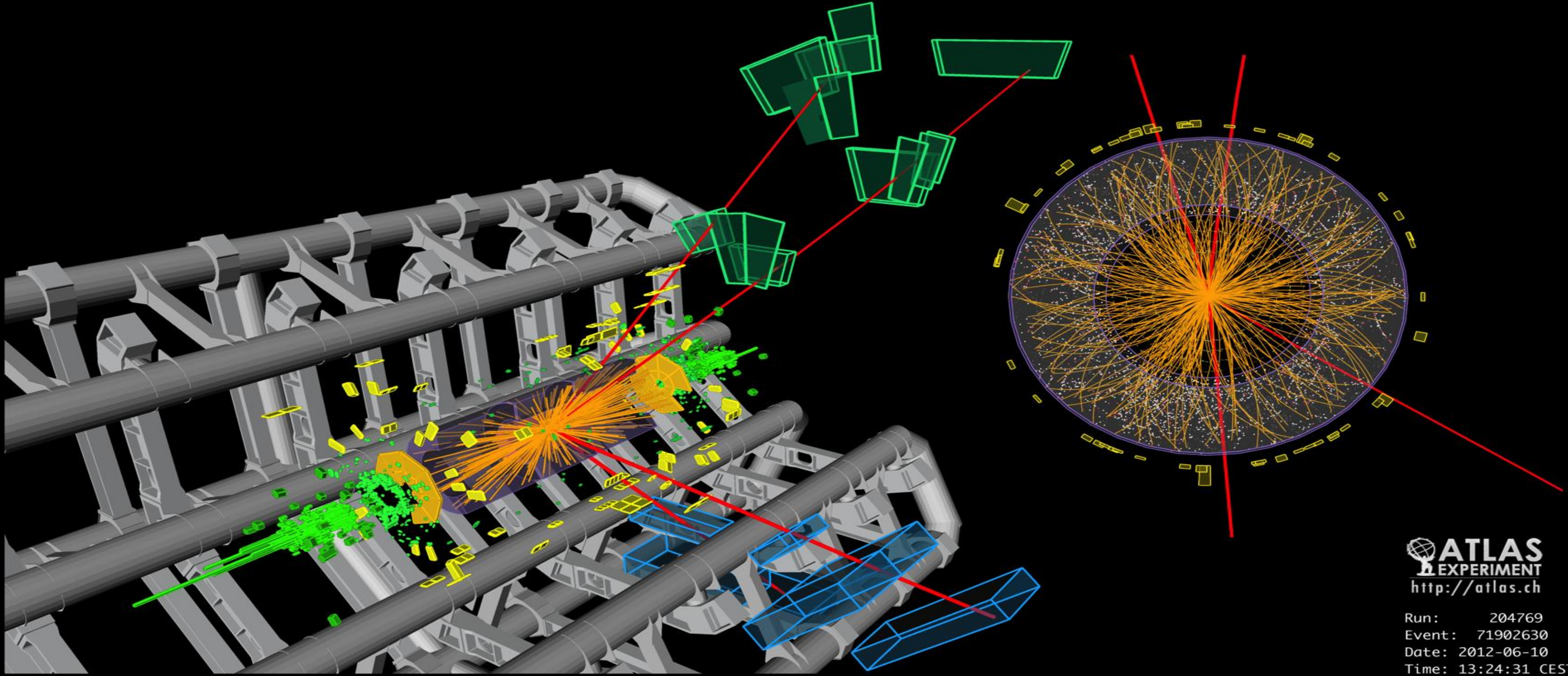

∂A_μ
M_π^2 φ
Model Matrix Elements
Q_5
-M_D



Higgs Boson Discovery

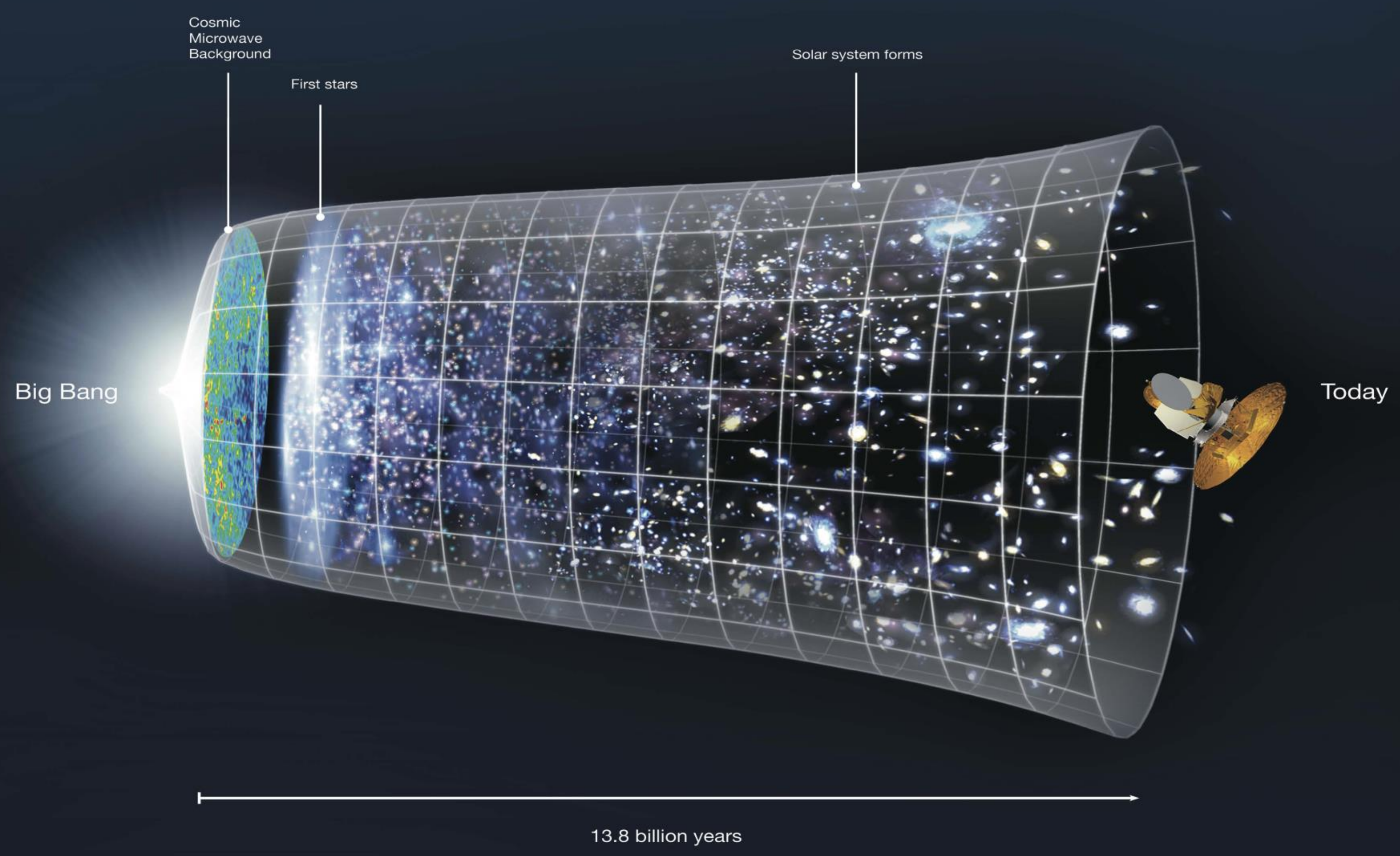
2012

Higgs to 4μ candidate event

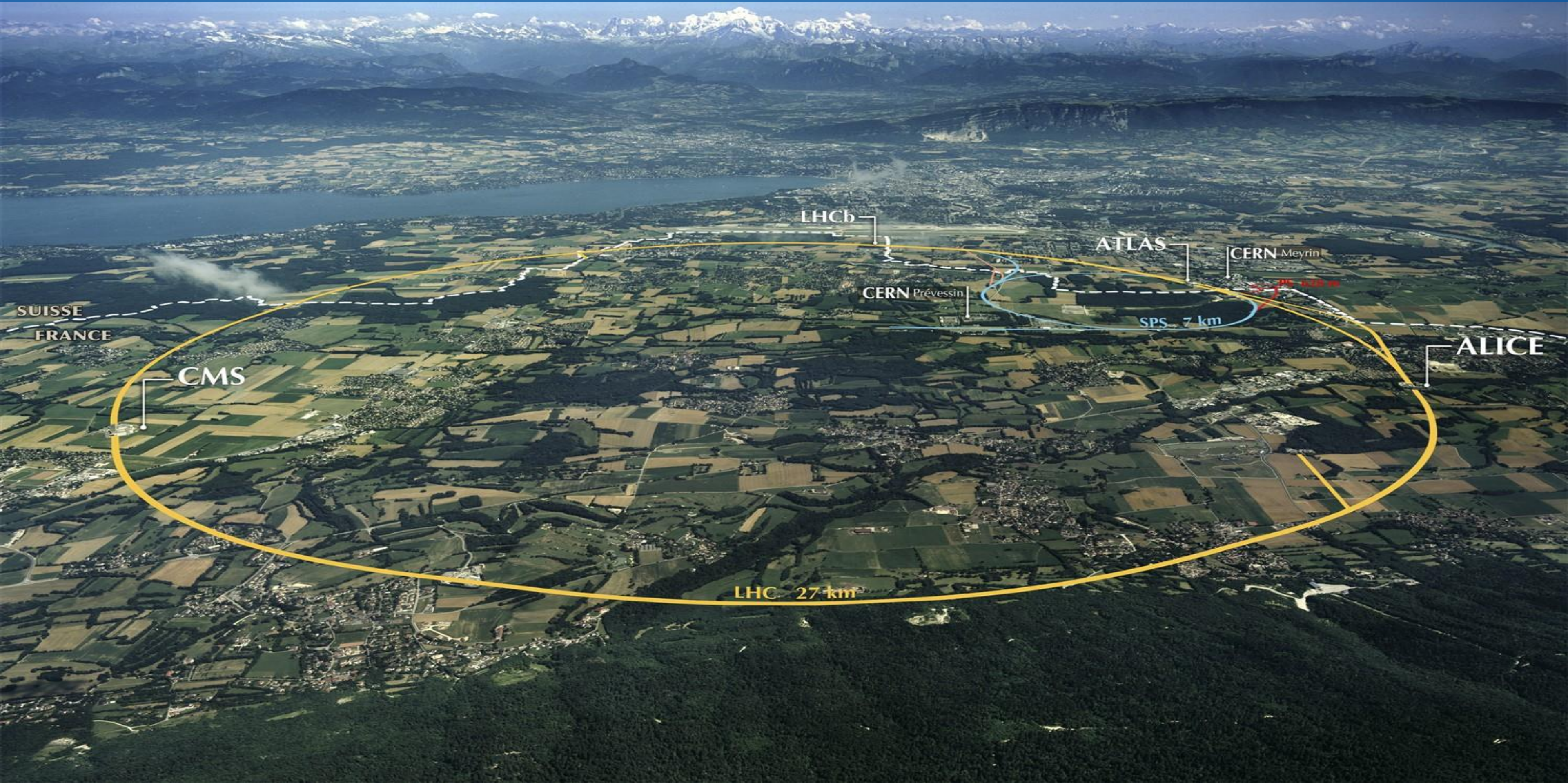


ATLAS EXPERIMENT
http://atlas.ch

Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

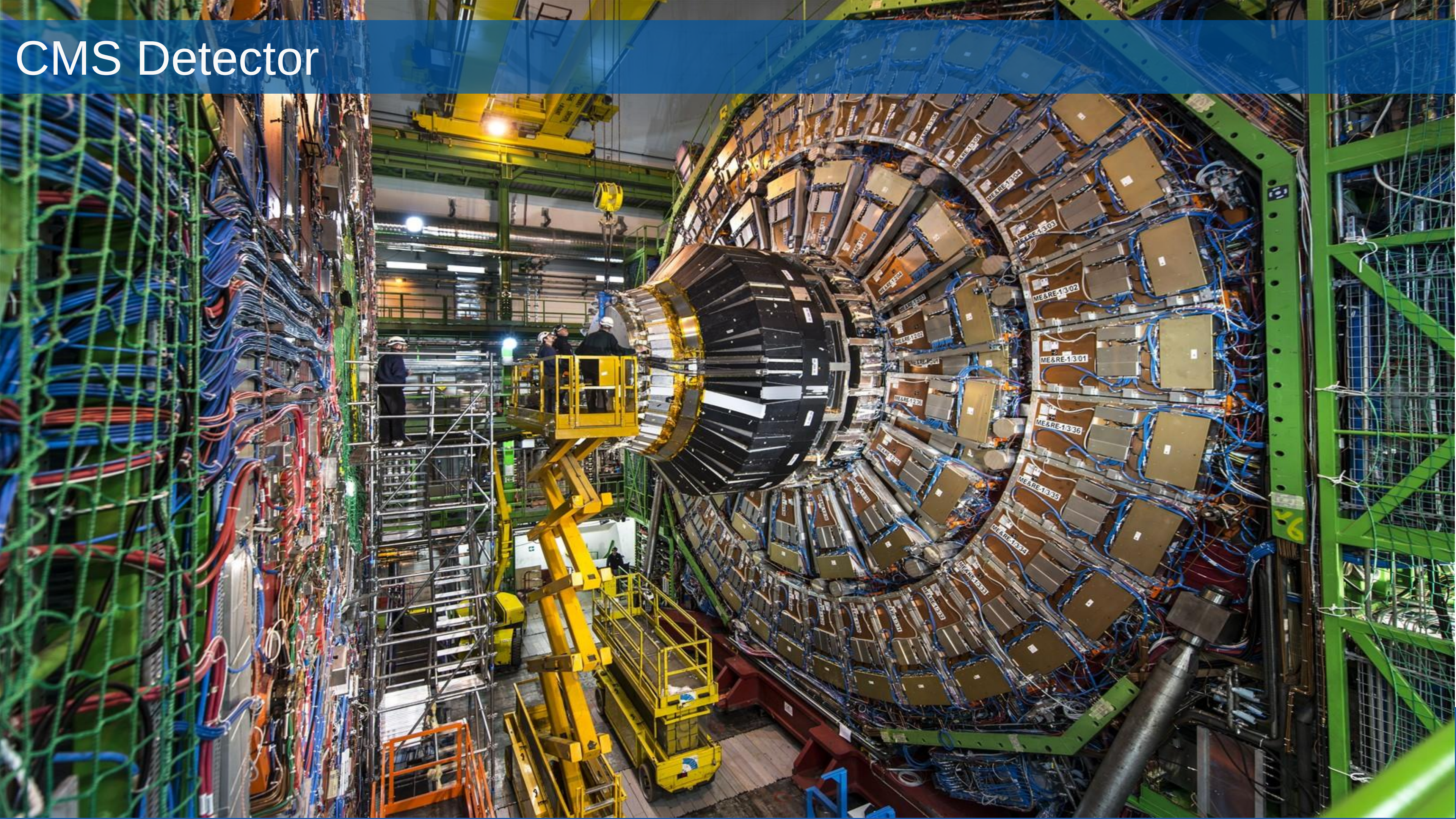

Cosmic Microwave Background
First stars
Solar system forms
Big Bang
Today
13.8 billion years

CERN Aerial View

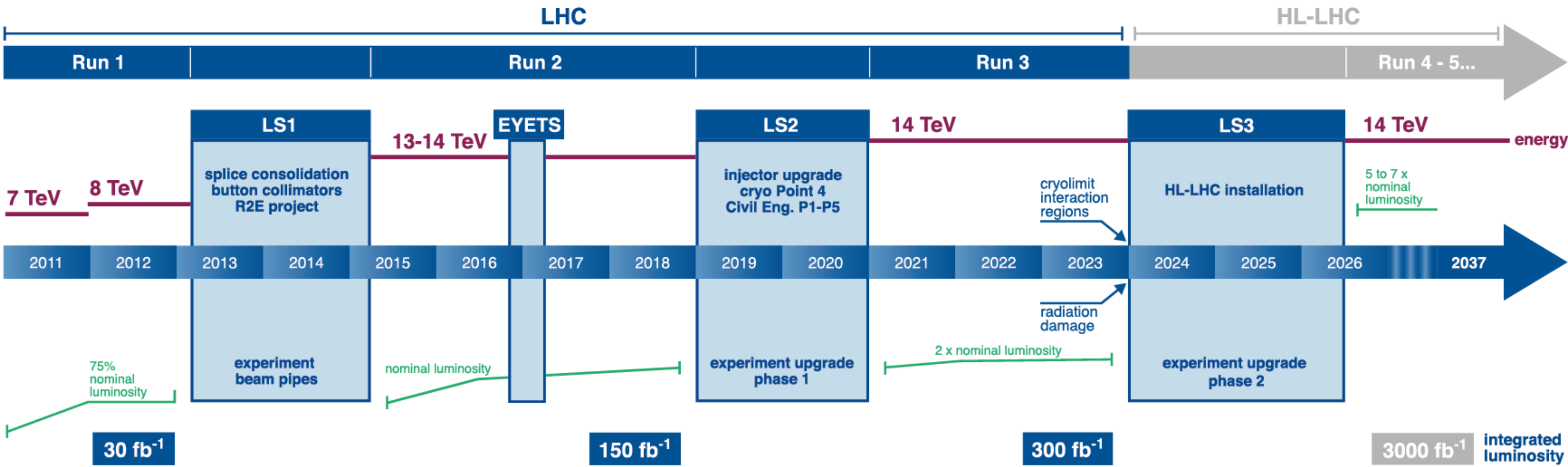


LHC Installation

CMS Detector

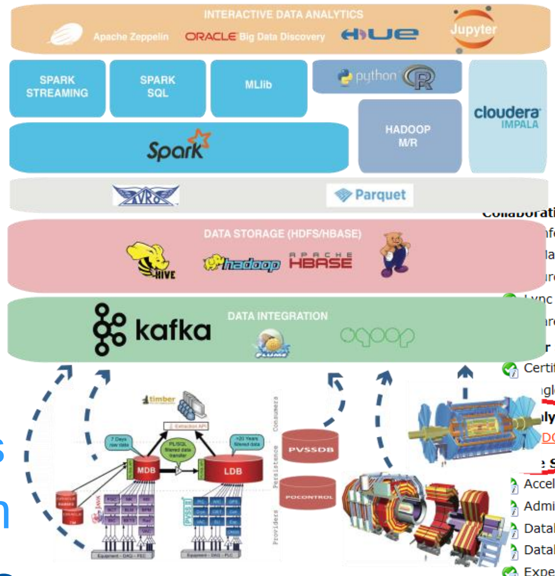


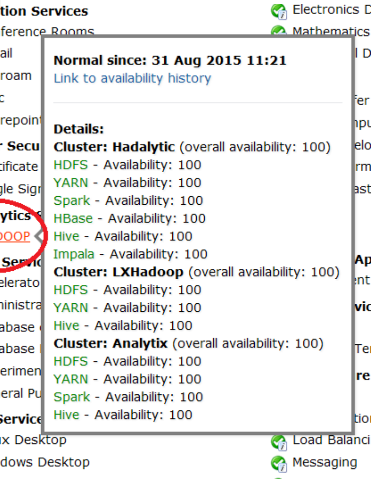
LHC / HL-LHC Plan



Hadoop and Analytics – IT-DB-SAS

- New scalable data services
  - Scalable databases
  - Hadoop ecosystem
  - Time Series databases
- Big Data Analytics
- Activities and objectives
  - Support of Hadoop Components
  - Further value of Analytics solutions
  - Define scalable platform evolution
- **Hadoop Production Service**





Collaboration Services

Conference Rooms

Mail

Eduroam

Lync

Sharepoint

Security

Certificate

Single Sign

Analytics

HADOOP

Service

Accelerato

Administra

Database

Database

Experimen

General Pu

Desktop Service

Linux Desktop

Windows Desktop

Electronics D

Mathematics

Load Balanci

Messaging

Normal since: 31 Aug 2015 11:21
Link to availability history

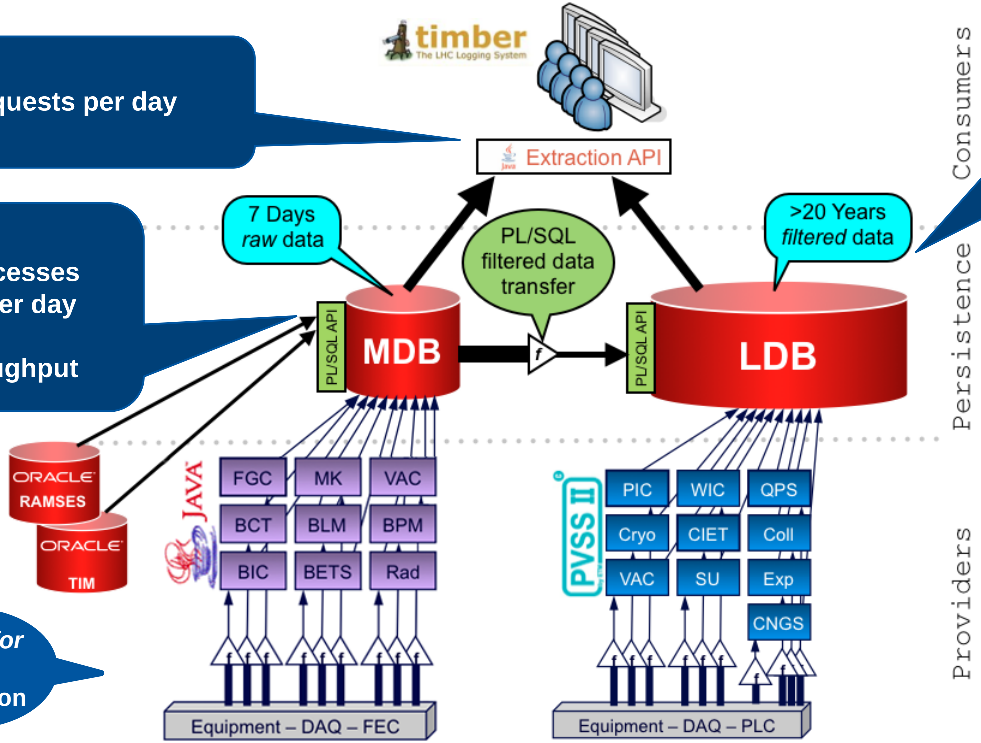
Details:
Cluster: Hadalytic (overall availability: 100)
HDFS - Availability: 100
YARN - Availability: 100
Spark - Availability: 100
HBase - Availability: 100
Hive - Availability: 100
Impala - Availability: 100
Cluster: LXHadoop (overall availability: 100)
HDFS - Availability: 100
YARN - Availability: 100
Hive - Availability: 100
Cluster: Analytix (overall availability: 100)
HDFS - Availability: 100
YARN - Availability: 100
Spark - Availability: 100
Hive - Availability: 100

CERN Accelerator Logging Service

- +800 extraction clients
- +5 million extraction requests per day
- 130 custom applications

~ 250’000 Signals
~ 50 data loading processes
~ 5.5 billion records per day
~ 275 GB / day
→ 100 TB / year throughput

~ 1 million signals
~ 300 data loading processes
~ 4 billion records per day
~ 160 GB / day
→ 52 TB / year *stored*

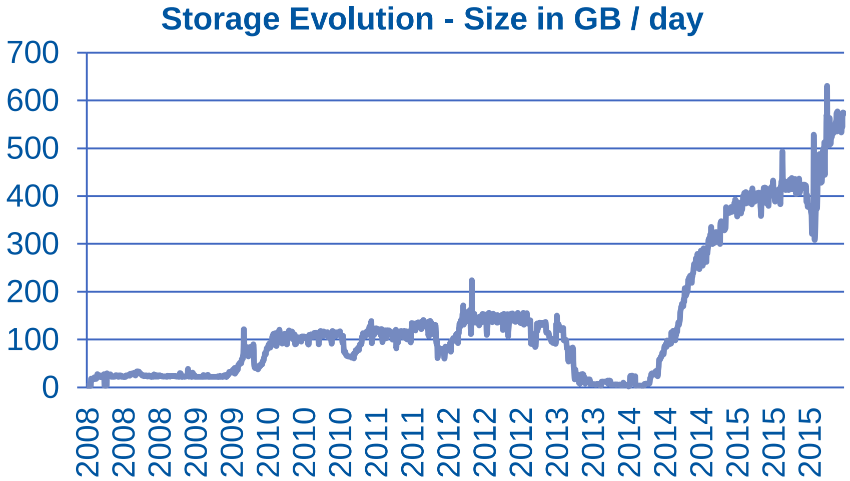


Filters for
data
Reduction

Credit: BE-CO-DS

CERN Accelerator Logging Service

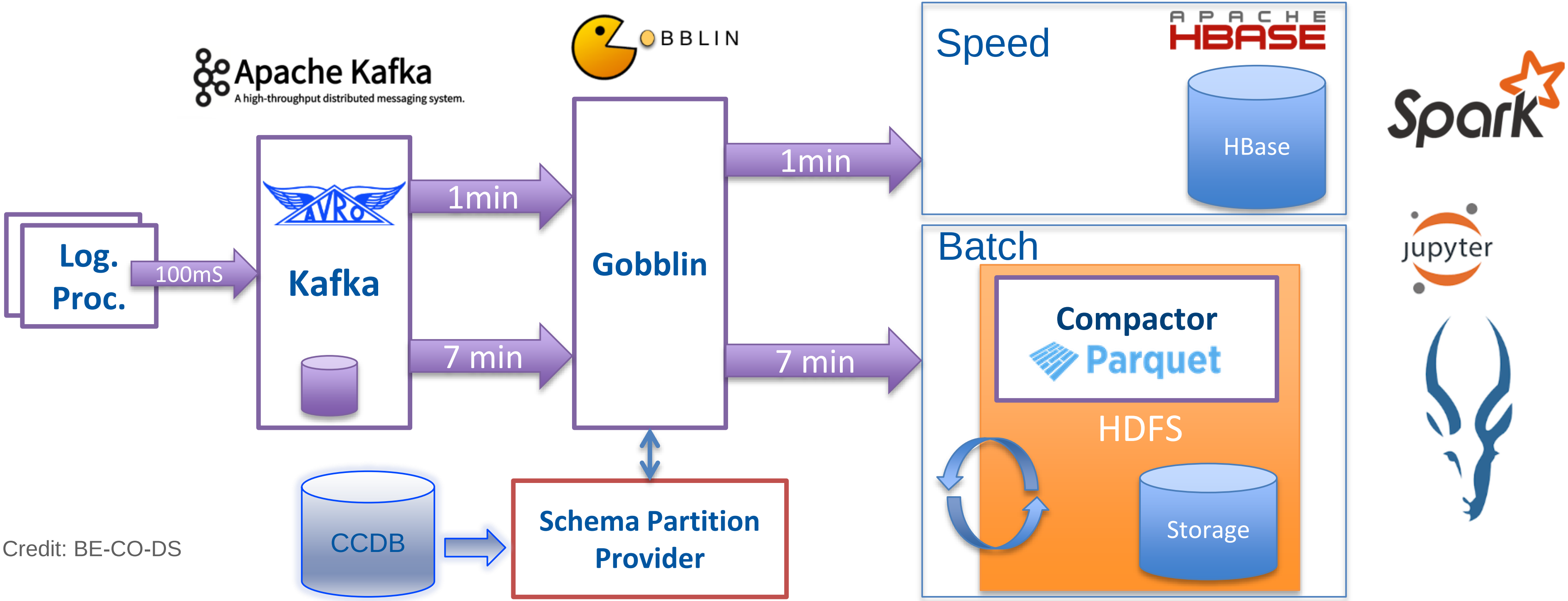
- New Landscape bring new challenges
  - Better Performance on bigger datasets
    - Big Data queries: Impala, Spark SQL
  - Leverage analytics capabilities
    - Spark Analytics: Python, ML, R
  - More heterogeneous data access models



QPS

Credit: BE-CO-DS

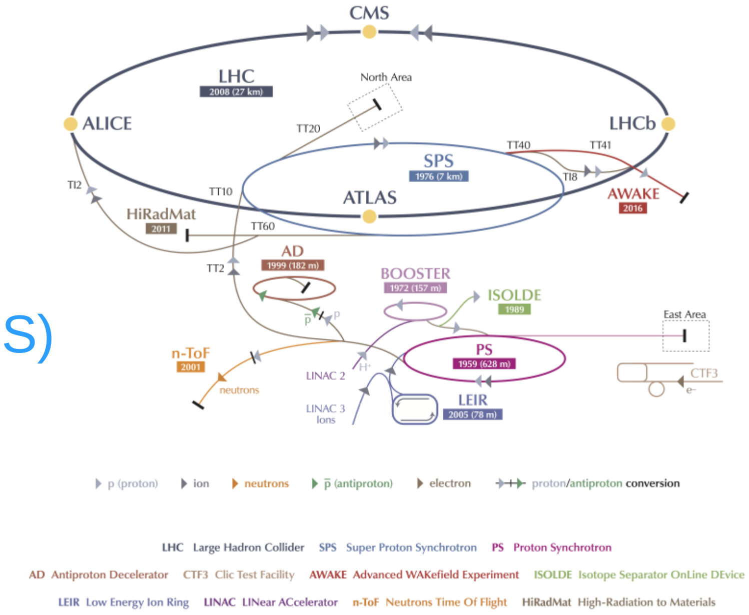
CERN Accelerator Logging Service



Credit: BE-CO-DS

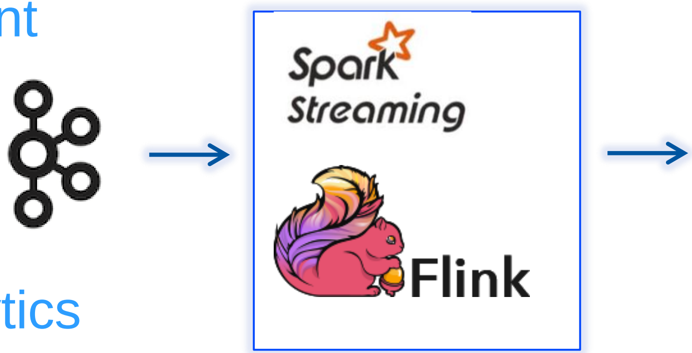
Accelerator Postmortem Analysis

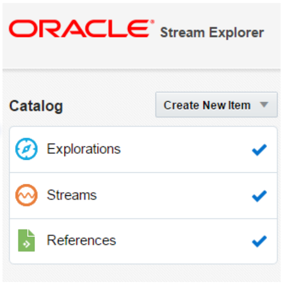
- Postmortem Analysis
  - Diagnostic on failures
    - Continue operations safely
    - Intervention Required
- Designed for CERN LHC
  - Extended to injectors complex (SPS)
  - External Post Operational Checks
  - Injection Quality Checks



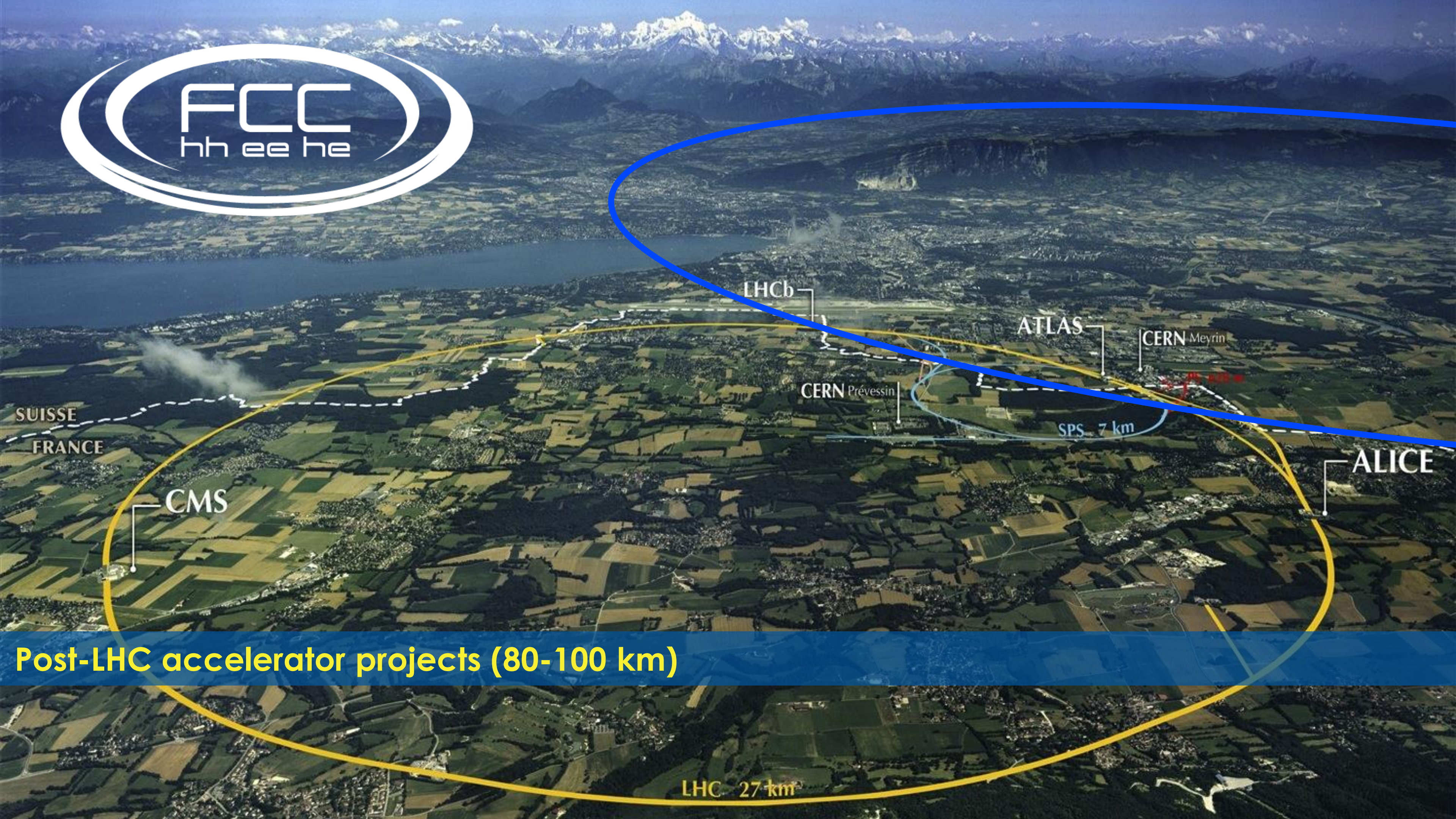
Accelerator Postmortem Analysis

- Challenges:
  - Stringent Timing Constraint
  - Better scalability
    - data storage
    - IO throughput
  - Big Data Streaming Analytics



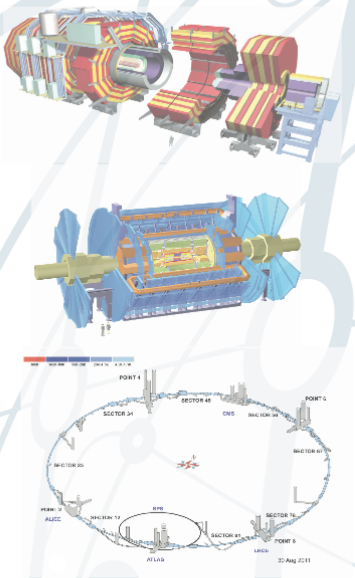


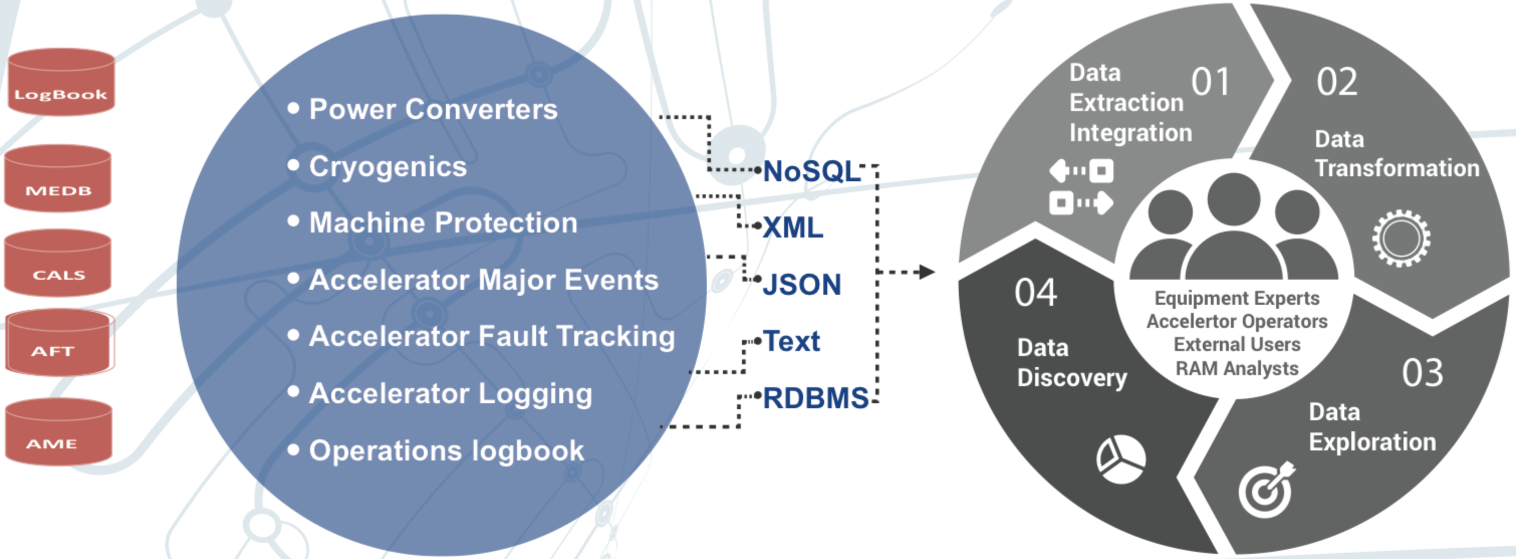
FCC
hh ee he



Post-LHC accelerator projects (80-100 km)

Scenario





LogBook

MEDB

CALS

AFT

AME

- Power Converters
- Cryogenics
- Machine Protection
- Accelerator Major Events
- Accelerator Fault Tracking
- Accelerator Logging
- Operations logbook

NoSQL

XML

JSON

Text

RDBMS

Data Extraction Integration 01

02 Data Transformation

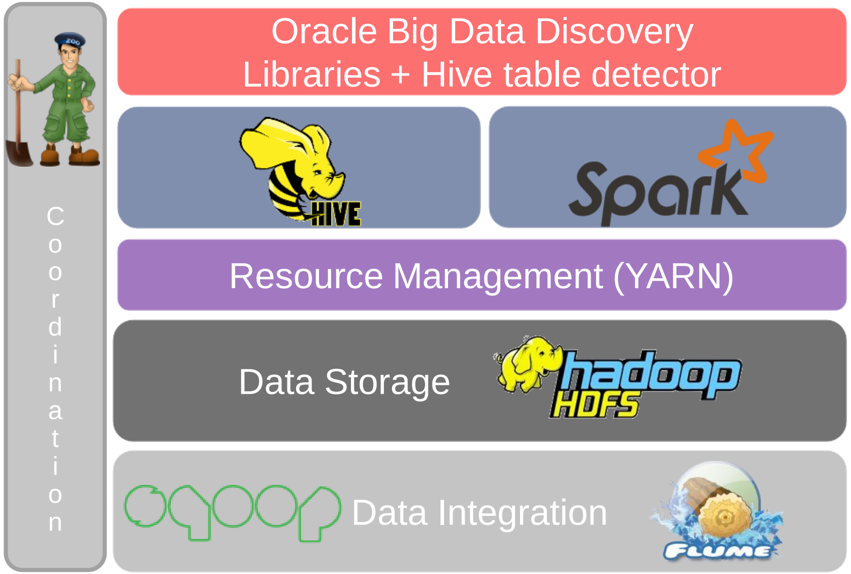
03 Data Exploration

04 Data Discovery

Equipment Experts
Accelertor Operators
External Users
RAM Analysts

Architecture overview

CDH 5.7.1
16 nodes, 24 GB ram
Intel Xeon L5520 @ 2.27GHz
165 TB HDFS



Oracle Big Data Discovery
Libraries + Hive table detector
HIVE
Spark
Resource Management (YARN)
Data Storage
hadoop HDFS
Data Integration
FLUME
Coordination

Big Data Discovery
v1.2.2
Dgraph & Studio

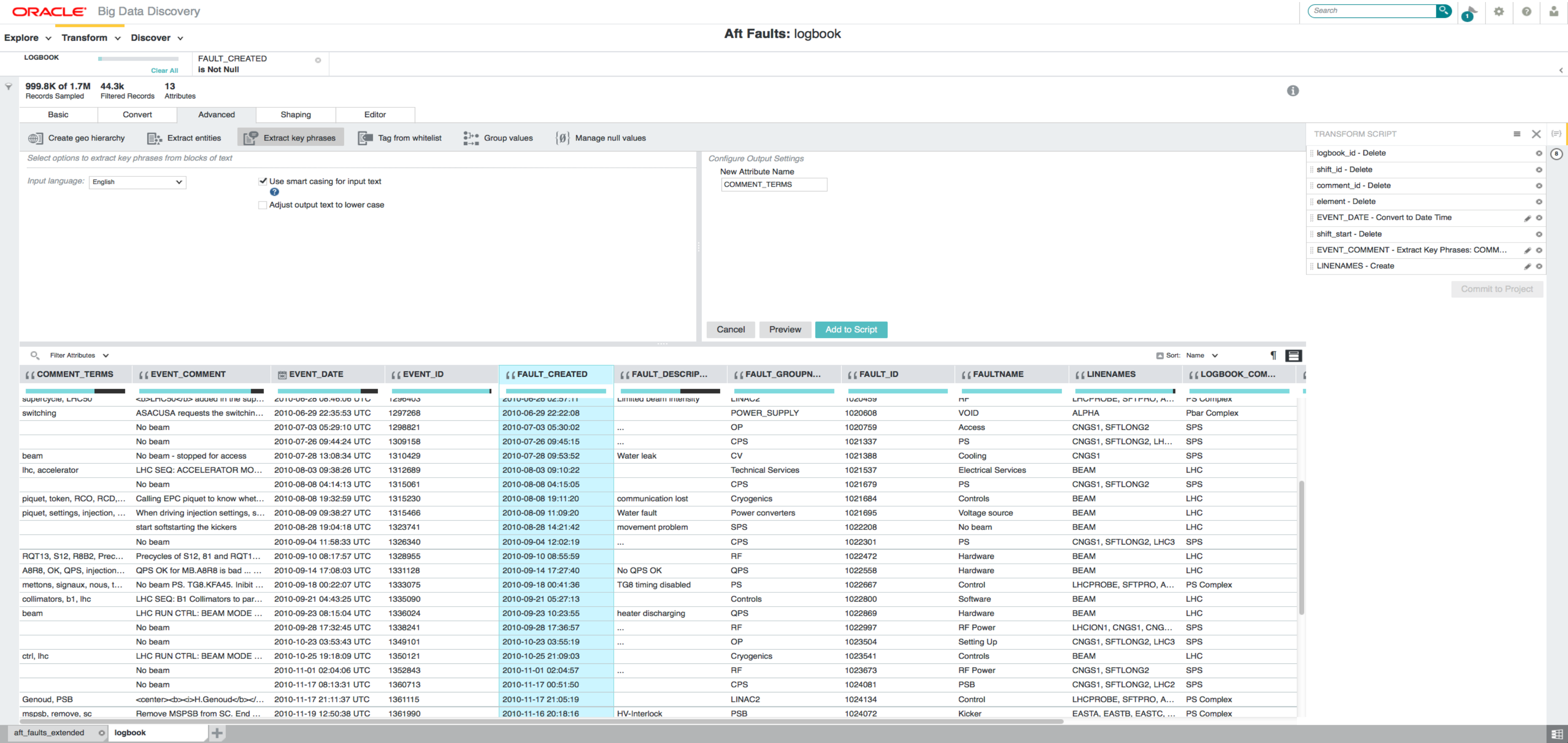
ORACLE
EXALYTICS

4x Xeon E7-8895 v2 (15 cores each)
2 TB RAM
4.8 TB Flash + 6 x 1.2 TB 10K HDD

Oracle Big Data Discovery Overview

- Data Exploration & Discovery
  - Interactive catalog of all data
  - Assess attribute statistics, data quality and outliers
  - Quick data exploration or create dashboards and applications
- Data Transformation with Spark in Hadoop
  - Apply built-in transformations or write your own scripts
  - Data Enrichment
    - Text: Entity extraction, relevant terms, sentiment, language detection
    - Geographical information: address, IP, reverse
  - Preview results, undo, commit and replay transforms
- Collaborative environment
  - Share and bookmarks
  - Create and share transformed datasets

Data Transformation UI - ETL



Aft Faults: logbook

FAULT_CREATED is Not Null

999.8K of 1.7M Records Sampled

44.3k Filtered Records

13 Attributes

Select options to extract key phrases from blocks of text

Input language: English

☑ Use smart casing for input text

□ Adjust output text to lower case

Configure Output Settings

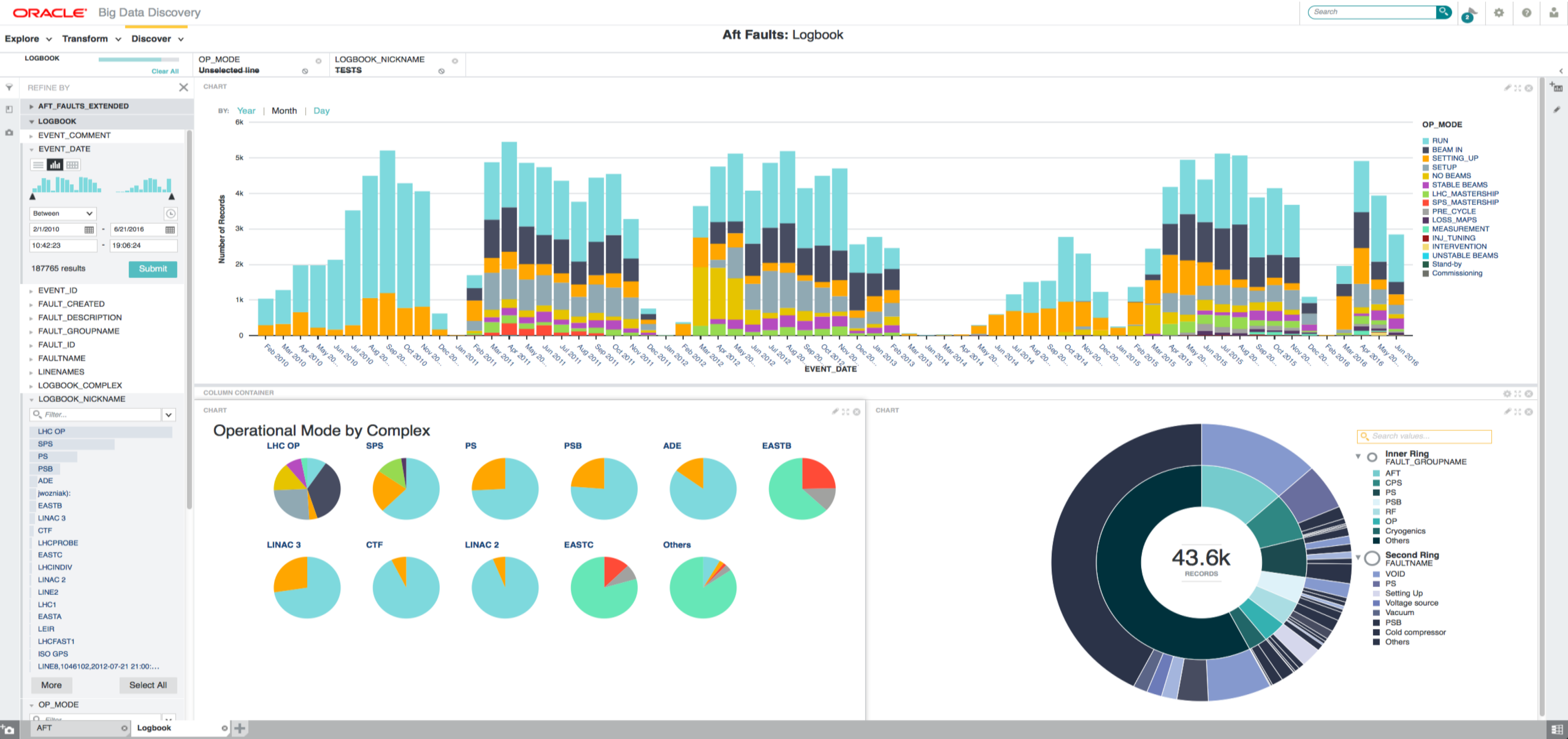
New Attribute Name: COMMENT_TERMS

TRANSFORM SCRIPT

- logbook_id - Delete
- shift_id - Delete
- comment_id - Delete
- element - Delete
- EVENT_DATE - Convert to Date Time
- shift_start - Delete
- EVENT_COMMENT - Extract Key Phrases: COMM…
- LINENAMES - Create

COMMENT_TERMS	EVENT_COMMENT	EVENT_DATE	EVENT_ID	FAULT_CREATED	FAULT_DESCRIP…	FAULT_GROUPN…	FAULT_ID	FAULTNAME	LINENAMES	LOGBOOK_COM…
supercycle, LHC50	<b>LHC50</b> added in the sup…	2010-06-28 08:46:06 UTC	1296403	2010-06-28 02:57:11	Limited beam intensity	LINAC2	1020459	RF	LHCPROBE, SFTPRO, A…	PS Complex
switching	ASACUSA requests the switchin…	2010-06-29 22:35:53 UTC	1297268	2010-06-29 22:22:08		POWER_SUPPLY	1020608	VOID	ALPHA	Pbar Complex
	No beam	2010-07-03 05:29:10 UTC	1298821	2010-07-03 05:30:02	…	OP	1020759	Access	CNGS1, SFTLONG2	SPS
	No beam	2010-07-26 09:44:24 UTC	1309158	2010-07-26 09:45:15	…	CPS	1021337	PS	CNGS1, SFTLONG2, LH…	SPS
beam	No beam - stopped for access	2010-07-28 13:08:34 UTC	1310429	2010-07-28 09:53:52	Water leak	CV	1021388	Cooling	CNGS1	SPS
lhc, accelerator	LHC SEQ: ACCELERATOR MO…	2010-08-03 09:38:26 UTC	1312689	2010-08-03 09:10:22		Technical Services	1021537	Electrical Services	BEAM	LHC
	No beam	2010-08-08 04:14:13 UTC	1315061	2010-08-08 04:15:05		CPS	1021679	PS	CNGS1, SFTLONG2	SPS
piquet, token, RCO, RCD,…	Calling EPC piquet to know whet…	2010-08-08 19:32:59 UTC	1315230	2010-08-08 19:11:20	communication lost	Cryogenics	1021684	Controls	BEAM	LHC
piquet, settings, injection, …	When driving injection settings, s…	2010-08-09 09:38:27 UTC	1315466	2010-08-09 11:09:20	Water fault	Power converters	1021695	Voltage source	BEAM	LHC
	start softstarting the kickers	2010-08-28 19:04:18 UTC	1323741	2010-08-28 14:21:42	movement problem	SPS	1022208	No beam	BEAM	LHC
	No beam	2010-09-04 11:58:33 UTC	1326340	2010-09-04 12:02:19	…	CPS	1022301	PS	CNGS1, SFTLONG2, LHC3	SPS
RQT13, S12, R8B2, Prec…	Precycles of S12, 81 and RQT1…	2010-09-10 08:17:57 UTC	1328955	2010-09-10 08:55:59		RF	1022472	Hardware	BEAM	LHC
A8R8, OK, QPS, injection…	QPS OK for MB.A8R8 is bad … …	2010-09-14 17:08:03 UTC	1331128	2010-09-14 17:27:40	No QPS OK	QPS	1022558	Hardware	BEAM	LHC
mettons, signaux, nous, t…	No beam PS. TG8.KFA45. Inibit …	2010-09-18 00:22:07 UTC	1333075	2010-09-18 00:41:36	TG8 timing disabled	PS	1022667	Control	LHCPROBE, SFTPRO, A…	PS Complex
collimators, b1, lhc	LHC SEQ: B1 Collimators to par…	2010-09-21 04:43:25 UTC	1335090	2010-09-21 05:27:13		Controls	1022800	Software	BEAM	LHC
beam	LHC RUN CTRL: BEAM MODE …	2010-09-23 08:15:04 UTC	1336024	2010-09-23 10:23:55	heater discharging	QPS	1022869	Hardware	BEAM	LHC
	No beam	2010-09-28 17:32:45 UTC	1338241	2010-09-28 17:36:57	…	RF	1022997	RF Power	LHCION1, CNGS1, CNG…	SPS
	No beam	2010-10-23 03:53:43 UTC	1349101	2010-10-23 03:55:19	…	OP	1023504	Setting Up	CNGS1, SFTLONG2, LHC3	SPS
ctrl, lhc	LHC RUN CTRL: BEAM MODE …	2010-10-25 19:18:09 UTC	1350121	2010-10-25 21:09:03		Cryogenics	1023541	Controls	BEAM	LHC
	No beam	2010-11-01 02:04:06 UTC	1352843	2010-11-01 02:04:57	…	RF	1023673	RF Power	CNGS1, SFTLONG2	SPS
	No beam	2010-11-17 08:13:31 UTC	1360713	2010-11-17 00:51:50		CPS	1024081	PSB	CNGS1, SFTLONG2, LHC2	SPS
Genoud, PSB	<center><b><i>H.Genoud</b></…	2010-11-17 21:11:37 UTC	1361115	2010-11-17 21:05:19		LINAC2	1024134	Control	LHCPROBE, SFTPRO, A…	PS Complex
mspsb, remove, sc	Remove MSPSB from SC. End …	2010-11-19 12:50:38 UTC	1361990	2010-11-16 20:18:16	HV-Interlock	PSB	1024072	Kicker	EASTA, EASTB, EASTC, …	PS Complex

Discovery Applications



Aft Faults: Logbook

OP_MODE: ~~Unselected line~~

LOGBOOK_NICKNAME: ~~TESTS~~

Between 2/1/2010 10:42:23 - 6/21/2016 19:06:24

187765 results

Advance Analytics - Notebooks

- Easy to create and share documents that contain live code
- Step by step execution reproduce the analysis, charts, etc.
- Support for multiple languages/kernels
- Multiple notebook software available
  - Jupyter/IPython
  - BDD provides notebook from version 1.2.0 (BDD Shell)
    - Can be used with Jupyter/IPython
  - HUE notebooks
  - Apache Zeppelin
  - More…

Scalable Analytics

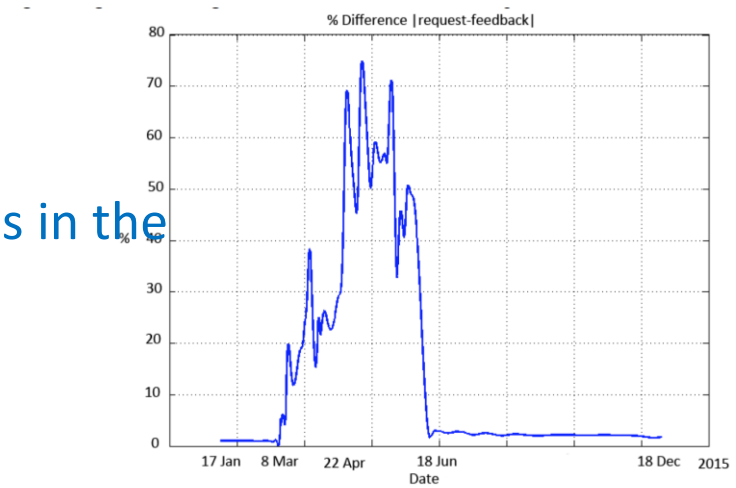
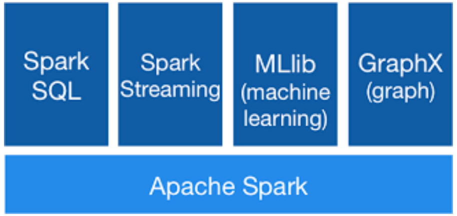


Figure 6. Absolute difference (%) between request and feedback for QSCB_6_2CV120

- Reliability of degrading components of valves in the cryogenic system of the LHC (University of Delft)
  - BDD -> Data Extraction -> Refine Calculations

  - Scalable solutions apply to all the cryogenics valves



Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Conclusions

- Hadoop is not the solution for all your problems but..
- Unlock new ways to exploit your investment on data
  - overcome technical limitations for several CERN use cases
- Allows heterogeneous data access
  - not only SQL or custom java APIs
- Once the data is in Hadoop only half of the way is done
  - Data visualization and discovery
  - Notebooks are easy to use and powerful for advanced analytics
  - Self-service tools improve productivity
    - Users should be able to do what they need without IT intervention

CERN

www.cern.ch