Big Data Analytics as a Service Infrastructure: Challenges, Desired Properties and Solutions

Manuel Martín Márquez
CERN- European Centre for Nuclear Research

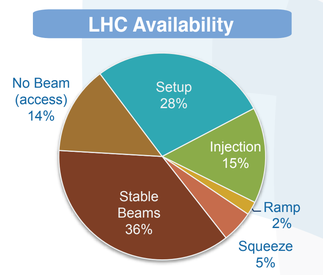
Manifesto

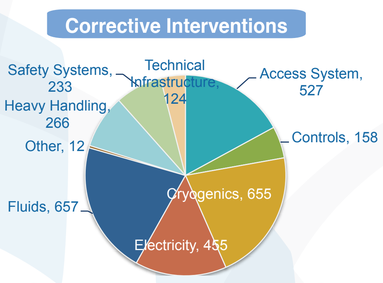
CERN's accelerator complex and detectors are an extreme data generator, every second an important amount of comprehensively heterogeneous data coming from control equipment and monitoring agents is persisted and needs to be analysed. Over the decades, CERN has applied different approaches, techniques and technologies. This has minimized the necessary collaboration to deliver cross data analytics over different domains. Essential to unlock hidden insights and correlations between the underlying processes, which enable better and more efficient daily-based accelerators operations and more informed decisions.

The proposed Big Data Analytics as a Service Infrastructure aims to: (1) Integrate the existing developments. (2) Centralize and standardize the complex data analytics needs for the CERN's research and engineering community. (3) Deliver real time and batch data analytics capabilities and (4) provide transparent data access and extraction-transformation-load, ETL, mechanisms to the different and mission-critical existing data repositories.

Data Analytics Objectives

Optimize CERN's Controls:





Maximize operation efficiency and availability

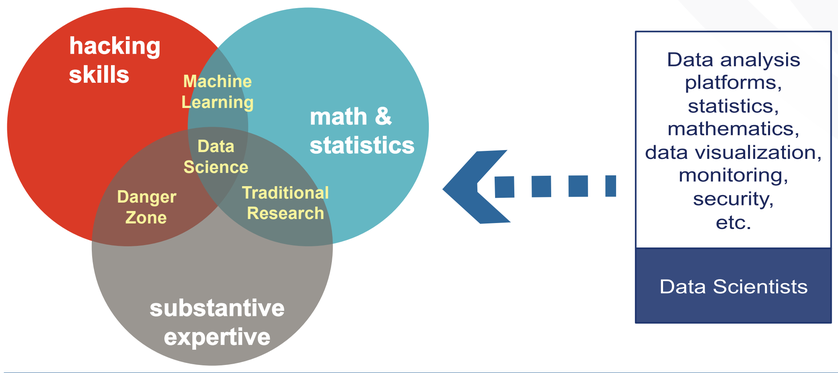
Reduce and predict faults and corrective interventions

Evolve CERN's Control and Monitoring Systems

Control and Monitoring Systems

DATA ANALYTICS

Intelligent, Predictive and Proactive Systems



Data analysis platforms, statistics, mathematics, data visualization, monitoring, security, etc.

Data Scientists

Technical Challenges

Data Access and Repositories Integration

Persist large amount of heterogenous data
- -Cryogenics, vacuumm, power converters...

Millions of control devices (time series data)
- -Sensors, actuators, monitoring agents

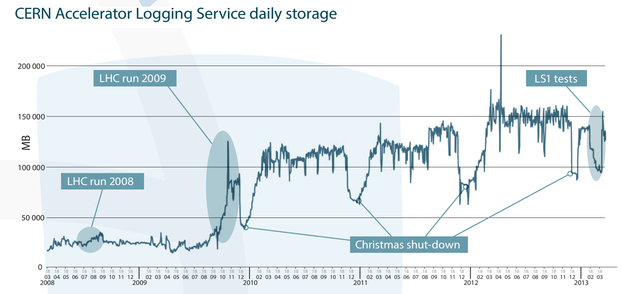
Integrate exitisting control data repositories
Provide transparent and flexible data access

Near-Real-Time processing

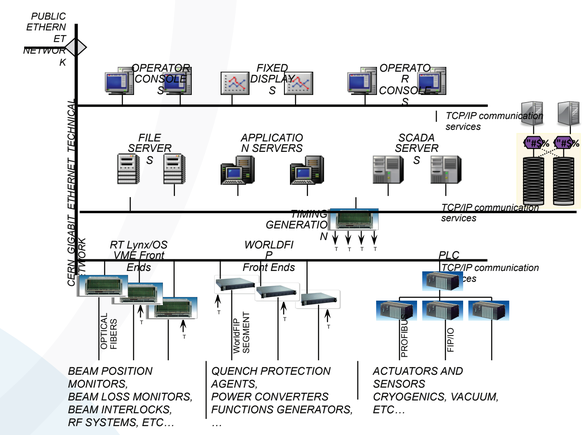
Order of GBs per second - Low latency
Integrate pre-existing knowledge and inferred
Scalable and fault-torelance

Batch and micro-batch analysis

Integrate different tools and frameworks



The logging service stores data using Oracle RAC databases, of close to one million pre-defined signals coming from heterogeneous sources, and it provides access to logged data for more than 700 registered individuals, more than 100 registered custom applications from around CERN, and even offsite access for purposes such as the CNGS experiments in Gran Sasso Italy.



Educational Aspects

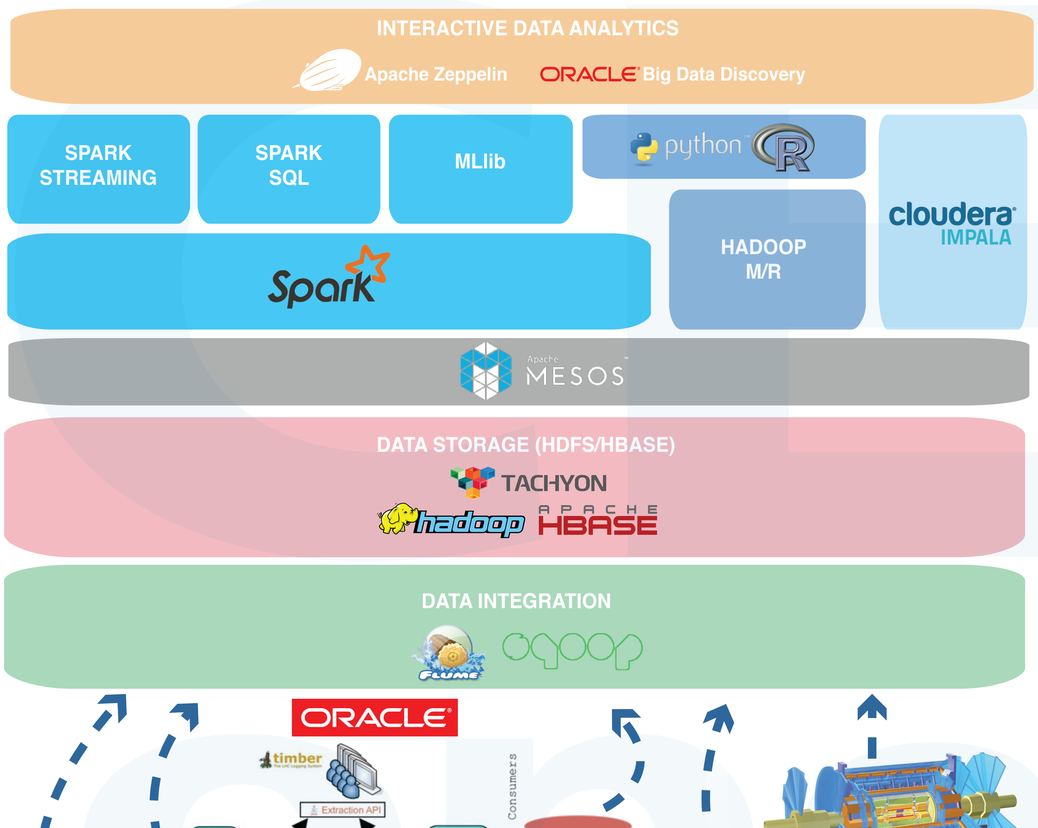
Data Scientist - General

New Professional Profile
Many domains of expertise involved

Data Scientist - CERN

Need to train engineering and contol teams

Data Analytics as a Service



Some Use Cases

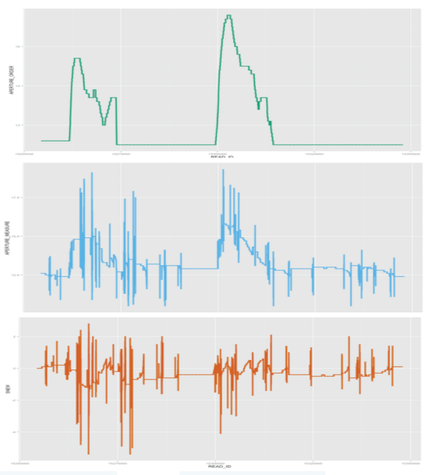
Faulty cryogenics valves detection

Signals used:
- S = aperture order - aperture measured

Features extractions based on S
- -Variance
- -Percentile 99.9
- -Rope distance – R(S)
- -Noise Band – B(S)

$$R(S) = \frac{1}{N}\sum_{i=2}^{N}|S(i) - S(i-1)| \qquad B(S) = \frac{\left|\sum_{k=1}^{N_{fft}/2} Pxx(k)\right|^2}{\sum_{k=1}^{N_{fft}/2} Pxx(k)}$$

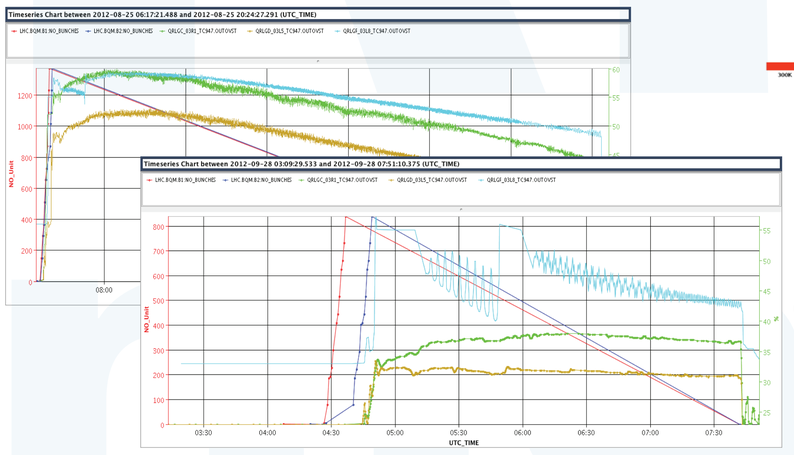
Pxx be the power spectrum of the signal S, from 0 to 0.5Hz, where S has been previously mean-centred



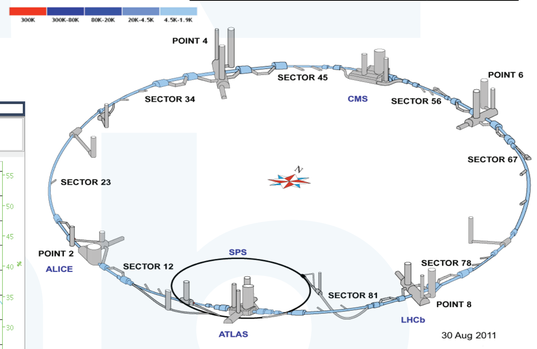
Automatic faulty valves detection system
- SVM - Support Vector Machine

Anomaly detection on beam screen cryogenics control

PID output (time series) segmentation
Segments characterization
Features extraction
Classification based on features



Instrument/Actuators	Total
Temperature [1.6 – 300 K]	10361
Pressure [0 – 20 bar]	2300
Level	923
Flow	72
Flow	2633
Control valves	3692
On/Off valves	1835
Manual valves	1916
Virtual flow meters	325
Controllers (PID)	4833



CHEP2015
OKINAWA, japan

CERN

CERN openlab