



10 December 2013, Eric Grancher
CERN openlab IT Challenges workshop


Data analytics challenge

Thanks to discussions with Manuel Martin Marquez, Philippe Gayet, all participants of the workshop, discussions at Oracle HQ, etc.

Outlook

- Context: work in openlab IV
- Workshop: November 20th (<https://indico.cern.ch/conferenceDisplay.py?confId=282578>)
 - Experience
 - Challenges
- Ideas / summary

Previously in openlab IV

- Data analytics forum created 
- Work within the control competence center (Siemens partner) and the database competence center (Oracle partner)
- Successes with different use cases, interesting correlation with CASTOR, power consumption forecasting for ATLAS and CMS magnets and cryogenic systems, evaluation of ELVis and WatchCAT by Siemens, etc.



Nov 20th (Workshop) Objectives

■ Describe current Data Analytics use cases:

- How Data Analytics is already used at CERN
- Planned to be used
- What use cases are of interest within CERN
- What technologies exist today
- Find out limitations (technological and resources)

■ Data Analytics Challenges

- Data Analytics methods and technologies

Manuel Martin Marquez – CERN openlab

Pedro Andrade and Miguel Coelho dos Santos (IT)

What makes a good metric?

- **It's comparable** to another time period, group, competitor, etc.
- **It's understandable** in a way the target audience will understand
- **It's a ratio or rate**
 - Which means it's easier to act on (acquisition cost per customer)
 - It allows you to represent the tension between two things (ads shown versus bounce rate, for example)
- **It's targeted to the right audience** (Internal business, developers, marketers, investors, media)
- **It changes the way you behave**
 - "Accounting" metrics make your predictions more accurate
 - "Experimental" metrics make your future behavior more effective

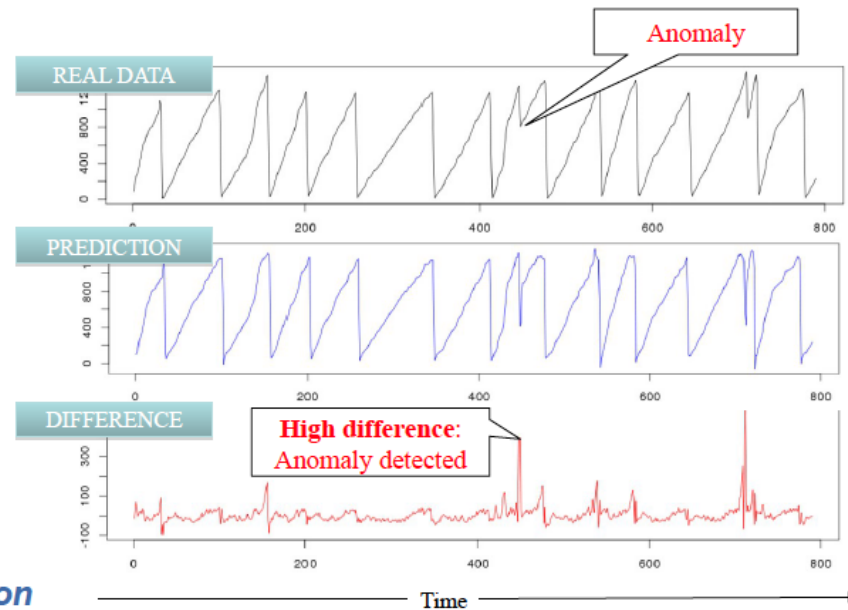
Anomaly detection

- 1) Build a **SVM (Neural Network like) model**
 - self trained
 - no supervision

- 2) Predict and compare:

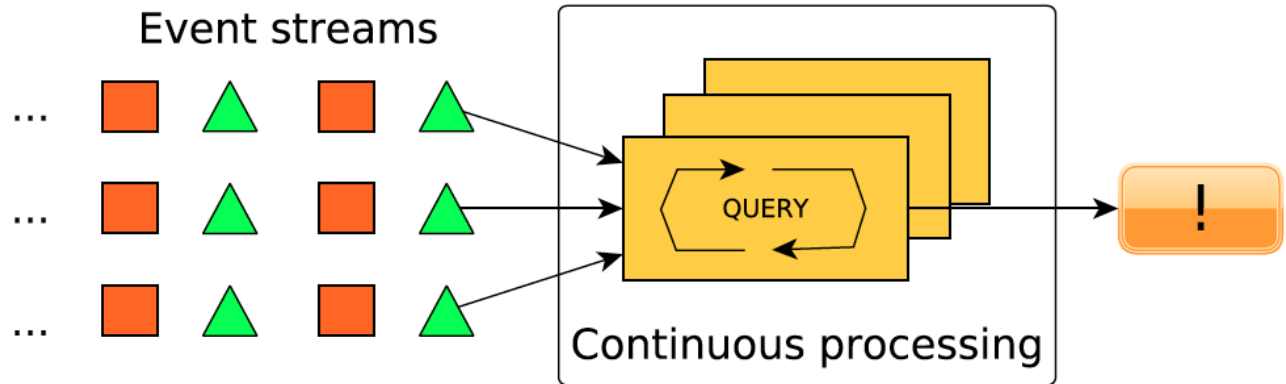
Real data Vs Prediction

- 1) **Blindly recognize anomalies**
- 2) **No other information required (i.e. thresholds)**

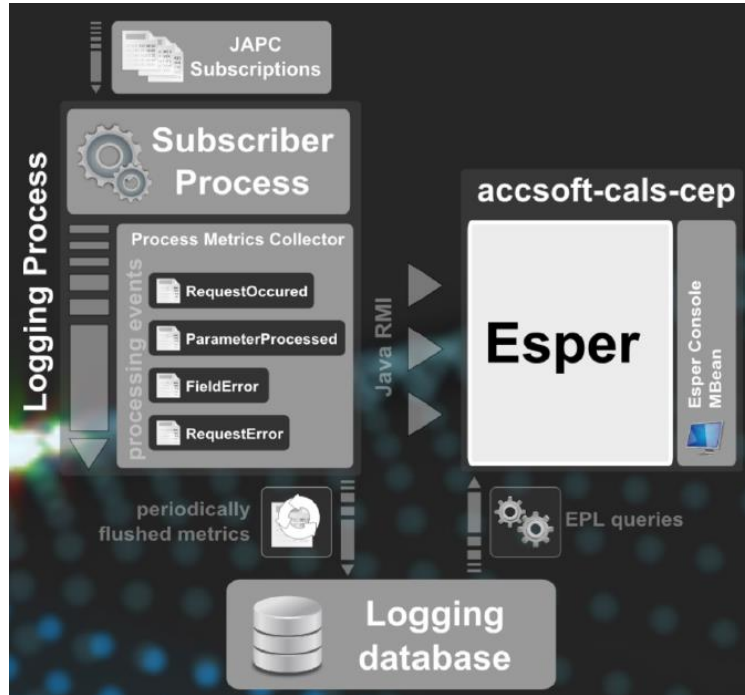


Luca Magnoni (IT-SDC)

- Continuous processing



Lukasz Burdzanowski (BE-CO)



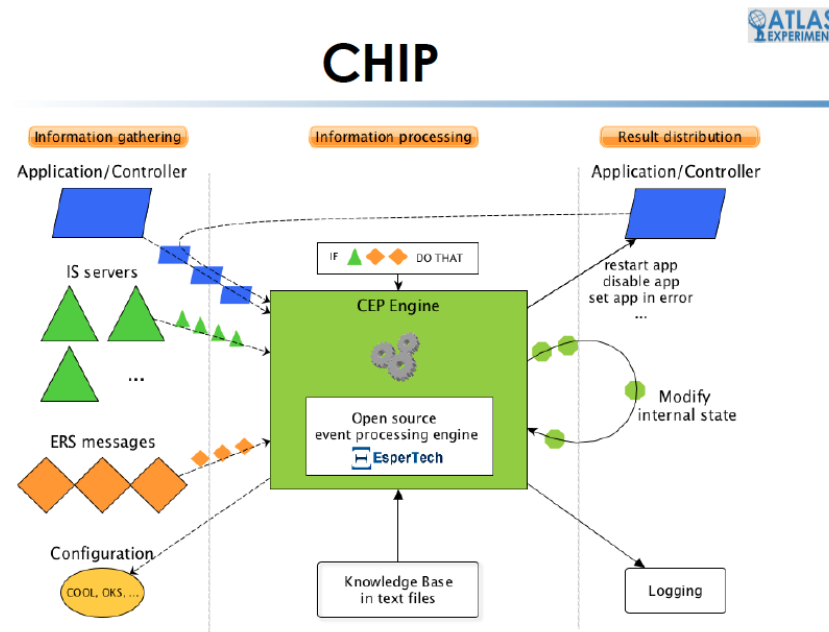
- Surveillance of CERN accelerator logging processes

4 instances logging data from:

- 5400 devices
- 11 000 parameters
- more than 200 000 distinct signals

Gabriel Anders (ATLAS)

- ATLAS shifter assistant and Central Hint and Information Processor

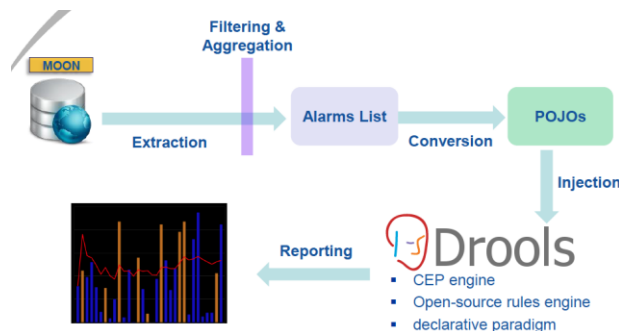


Filippo Tilaro / Axel Voitier (EN-ICE)

- Control and Monitoring system
- Alerting and reporting system
 - Manually configured
 - Based on threshold trespassing pattern
- Huge data volume

- Initial conclusions

- no single framework out of the box to analyze numerical data and not (next version of WatchCAT)
- Necessary a combination of tools for a complete data analysis (log processing, statistical analysis, pattern recognition...)
- Split this use-case into smaller ones:
 - signal analysis use-case (next version of WatchCAT will provide predictive trending capabilities)
 - automatic extraction of statistical metrics and thresholds



Chris Roderick (BE-CO)

- LHC Logging (50+ TB/year)
- Perform analysis as close to data as possible, in database analysis: built-in + ORE?

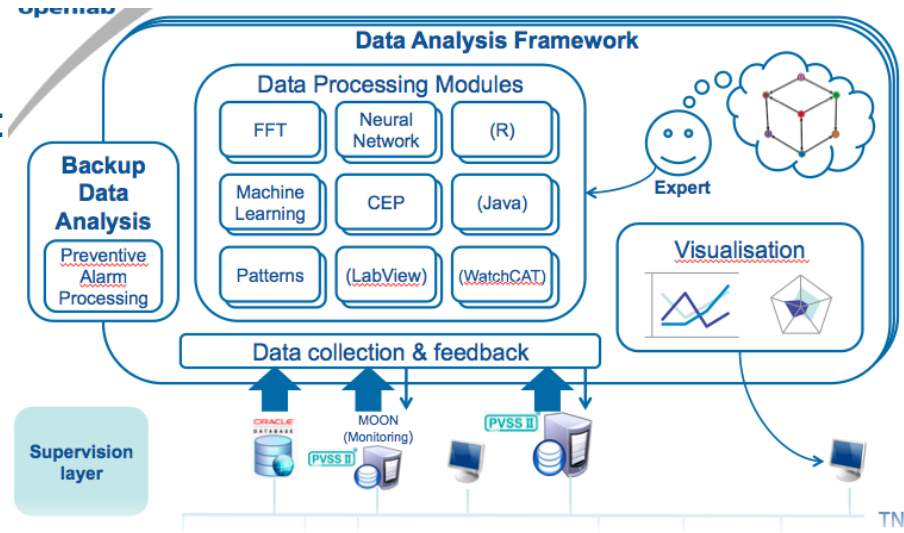
- Multi source extraction API

- Domain specific language

```
getDataForSignals(A, B, C)
alignedToTimesOfSignal(A)
filterTimesWhenValuesMatch(stddev(signalValues(D))>5)
during (
  beamModes(INJECTION, RAMP)
  of lastLHCfills(5, havingBeamModes(STABLE_BEAMS))
)
```

Axel Voitier (EN-ICE)

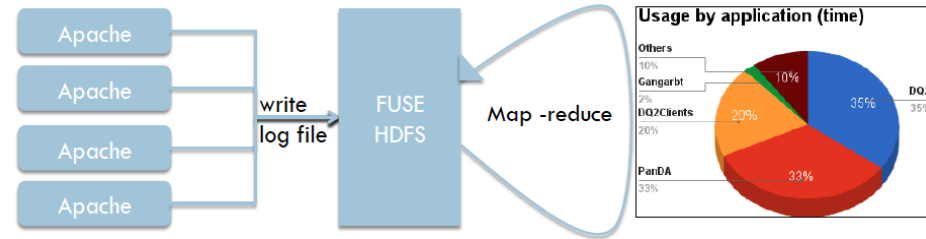
- Configurable analysis flow by user
 - + It can use custom analysis software
- High scalability of analysis processes
 - From laptop to multi-node cluster
- Stream based data processing engine: Storm
- NoSQL data storage engine
- Web-based visualisation interface
 - HTML5, Data pushed by WebSockets
 - Desktop and mobile devices



TN

Vincent Garonne (ATLAS)

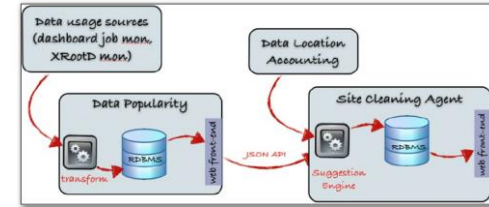
- Use cases:
 - Trace Mining (user interactions with Distributed Data Management)
 - Popularity (used for deciding which data to delete)
 - Accounting and popularity (reports on data contents/popularity)



- Log file aggregation
- ATLAS Distributed Data Management uses both SQL and NoSQL

Domenico Giordano (IT-SDC)

- Intelligent data placement models for the CMS experiment
- Need to extract further knowledge from the monitoring data in order to implement an effective data placement
 - Correlate file-access monitoring with site status
 - Readiness, queue length, storage and CPU available
 - Classify analysis activities and needed resources
 - Making recommendations
 - Learn from the past trends and patterns



Simone Campana (IT-SDC)

- Network monitoring
- Time correlation
 - During a PS throughput test, was there any known activity in the same link?
 - There is packet loss, does this appears as degraded performance somewhere at the same time
- We observe loss of performance in some network link
 - Is it a network problem and where?
 - Is it a storage problem?

Data analytics

- Real time analytics
 - Interactive
 - Alarm
- Reports
- Batch analytics, data mining, correlation, etc.



Base for discussion

- Diverse data: LHC and experiment data, computing (experiments, IT, EN, BE) as well ; at least $O(10^9/\text{day})$ rate and at least $O(10^6)$ signals. (“you make it, we break it”?)
- 1. From real time analytics to batch analysis, correlation, early prediction of upcoming failures
- 2. Lot of data, optimisation
- 3. Efficient connection and integration
- 4. Visualisation and APIs
- 5. “Analytics platform” or (Big data) “Analytics-as-a-service” (A³S ?):
 - Data fed from multiple sources (live)
 - Stored reliably
 - Data processing with multiple systems
 - Easy access, domain expert natural language (DSL)
- Strong use cases at CERN... more? linked other research organisations? (select some)

now open up!



www.cern.ch