



11 December 2013

CERN openlab IT Challenges workshop

Data analytics challenge

Thanks to all for the participation, constructive ideas,
thanks to Manuel Martin Marquez for taking the notes

General

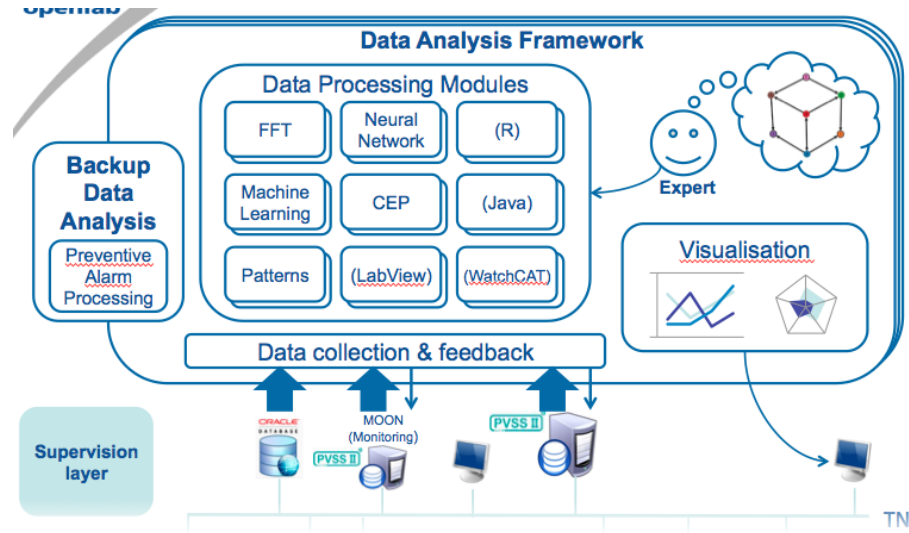
- Convince the user and establish a sponsor model, easy interface helps to explain and gain traction
- Include a dissemination aspect (training)
- Prepare 2-3 approaches and advertise them
- Start preparation in 2014

Techniques

- Show similar objects, exploratory solution
- Feedback loop from batch analysis to “real-time”
- “Real-time” is real-time at the $O(\text{second})$
- First offline analysis before real-time analysis for part (LHC logging) to first understand well the problem, example Unidentified falling objects, to convince the users
- Challenge: easy way to describe what you want, ability to translate to a lower level language

Analytics as a service

- “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)



TN

Constraints

- How to not impact production / guarantee that data can be taken(decouple production from analysis)
- Different population and needs: engineers, technicians, operators
- For part of the data, sequence is not enough, vectors is required
- ESA: offline
- Agile methodologies for data analytics, work on the process
- Yandex: plan to have DSL in DataFactory
- Different types and different format, connectors
- Communication between nodes should be transparent

TODO: 2014

- Have 4-5 use cases well described
 - Data, signification of data, goal (measurable)
- Pre-evaluate some of the technologies
- Prepare architecture and interface design for the “Analytics as a service” platform



www.cern.ch