Web-based Analysis Services Report

Nov 24, 2017

Authors: Tim Bell, Luca Canali, Eric Grancher, Massimo Lamanna, Gavin McCance, Pere Mato Vila, Danilo Piparo, Jakub Moscicki, Alberto Pace, Ricardo Rocha, Tibor Simko, Tim Smith, Enric Tejedor Saavedra.

Executive Summary

Evolution of the analysis services

Web-based services (cloud services) is an important trend to innovate end-user services while optimising the service operational costs.

CERN users are constantly proposing new approaches (inspired from services existing on the web, tools used in education or other science or based on their experience in using existing computing services).

In addition, industry and open source communities have recently made available a large number of powerful and attractive tools and platforms that enable large scale data processing. "Big Data" software stacks notably provide solutions for scalable storage, distributed compute and data analysis engines, data streaming, web-based interfaces (notebooks). Some of those platforms and tools, typically available as open source products, are experiencing a very fast adoption in industry and science such that they are becoming "de facto" references in several areas of data engineering, data science and machine learning.

In parallel to users' requests, WLCG is considering to consolidate its deployment model into a relatively reduced number of resource sites (while benefitting from potentially ephemeral resources like public cloud and large installations like experiment filter farms, HPC facilities, analysis centres). In this schema, analysis facilities could be provided as web-based services located in strategic WLCG centres or spawned on relatively short-lived farms.

1. ROOT is a data analysis framework used by nearly all the HEP experiments worldwide. It provides a complete toolset for big data processing, statistical analysis, visualisation and storage. For the last couple of years, the development and the provision of web-based ROOT integrated with IT storage services is an important aspect of ROOT strategy and its collaboration with IT.
2. NXCALS is the next-generation application for accelerator logging. The accelerator logging service is critical for the operation of the accelerator complex and its legacy implementation is deployed using an Oracle database managed by IT. NXCALS is currently in active deployment and moving to a service with production-like support chain in Q1 2018. This project profits from the collaboration between BE and IT, continuing on the path of what has happened previously with the legacy system: the software and user-facing APIs are developed and supported by BE, the data platform (Hadoop and Spark in this case) is managed by IT with contribution of EP.
3. SWAN (Service for Web based ANalysis) is a web-based platform to perform interactive data analyses on the cloud using Jupyter notebooks with Python, ROOT/C++ and R kernels. It uses CERNBox for local user storage and integrates with

experiments' software and data on the cloud. SWAN permits the users to analyse the data without the need to install any software.

4. REANA (REusable ANAlyses) is a component of the CERN Analysis Preservation framework that aims at promoting controlled reproducibility and reusability of research data analyses. It provides researchers with tools to describe their analysis process in a structured manner using declarative workflow languages so that the analysis components can be preserved and the analysis can be instantiated on the containerised cloud for the revalidation, reinterpretation and reuse purposes.

Recently the availability of required technical means in common across these 3 examples (notebooks, execution containers, …) is making this approach feasible also outside our community.

The special role of CERN in setting directions for the HEP (computing) evolution and in being at the centre of collaborations across different sciences make this class of services a strategic opportunity for CERN.

About this document

The document has been compiled by authors as representatives of four IT groups (IT-CDA, IT-CM, IT-DB and IT-ST) and by EP-SFT and discussed with the IT Security team and the BE-CO (Chris and Jakub) which we thank for valuable feedback.

The first chapter is an inventory of the existing and expected use cases of Analysis Services. Currently they are developed at CERN and delivered by the EP and IT departments. Several development and integration projects exist with significant contributions from the 5 CERN groups. The second chapter describes the operation environment in the CERN computer centre. In the third chapters the authors discuss the evolution both in terms of the service itself, in its usage and in the expected evolution of the toolkit at the base of the current implementations. The document contains (links to) relevant documents to describe the current situation and possible future directions. All this reference material is presented in the Annexes.

Use cases

Intro

In this chapter we list the main use cases. These use cases cover a broad range of situations. The common point is the usage of the browser as a convenient front-end to browse and share analysis artifacts (as histogram files) to offer better usability in various areas (e.g. ease of sharing, analysis over large-latency links, …) . They can promote different ways to interact with 'back-end services' as in the case of Spark analysis or integration with different data distribution and processing units (CVMFS, EOS, HT Condor, Grid processing, …). The possibility to perform analysis in a concurrent environment effectively encourages the possibility to offer new services for non-traditional analysis activities like access to open data for outreach and education and to be the basis of knowledge preservation.

Web-based data analysis

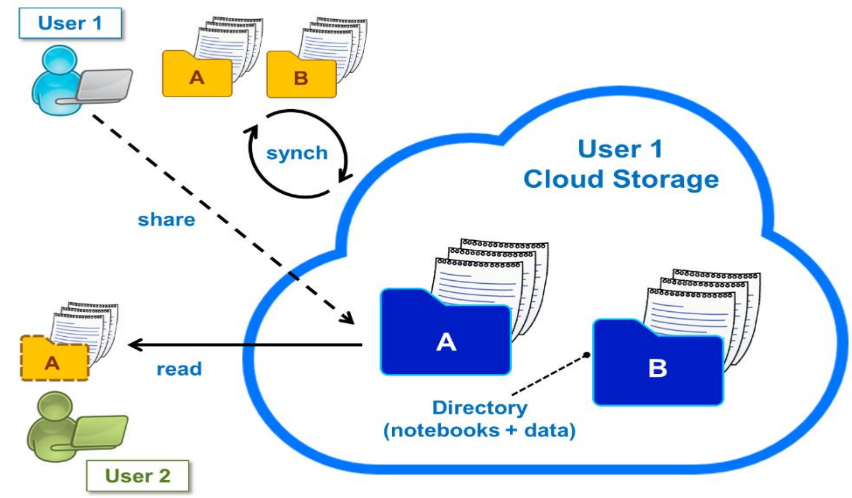
SWAN proposes to extend the HEP analysis environment with a web-based cloud platform that supports the development and execution of code, the visualisation of analysis results and the sharing of these results among scientists.

Users only need a web browser to access and work with SWAN. Once a user has started a SWAN session, she is prompted with an interface, the notebook, that allows her to combine code, formatted text and graphics in the same document. Also importantly, notebooks can run the code they contain and such execution happens entirely in the cloud. Similarly, in this cloud-based model, both the input data and the execution results reside in cloud storage.

In order to facilitate the collaboration between scientists, SWAN relies on the sharing capabilities of CERNBox, which acts as the user's home directory in SWAN and stores all the user files, including notebooks. A scientist that has produced a notebook and wants to share it with other colleagues can do so by means of CERNBox, so that these colleagues can open the shared notebook for review, execute it and even modify it and re-share it.

In the cloud-based analysis model of SWAN, the user can also synchronise the content of her CERNBox into her own machine. Thus, the results of her work in SWAN (notebooks, result files) can also be inspected locally and (possibly) offline. Moreover, the results of any local work that are stored in CERNBox will be also available in SWAN as a result of the CERNBox synchronisation capabilities.

This analysis model is also useful in the case of long-distance access where latency is big. In SWAN, the access latency is absorbed by the notebook interface: when the user is typing new content in a notebook, the code remains in her browser, and only when she chooses to execute that code a request to the server actually happens. Moreover, notebooks are saved automatically and periodically by SWAN, so that the user does not even notice this background communication with the server.



Analysis services as interface to external resources

When a user logs in to SWAN, she is assigned some computing resources to execute her notebooks. These resources, which normally correspond to a few cores and GBs of memory, are enough for lightweight, exploratory analyses.

However, for analyses that require more computational power, SWAN can be used as a bridge to access larger resources. An example of such resources are Spark computational clusters: from a SWAN notebook, a user can offload Spark jobs to external clusters, get back the results and inspect them in the notebook (see "Spark for data analysis").

More generally, effective and intuitive bridges to local batch systems (e.g. HT Condor), Grid resources and integration with the experiments' grid framework are possible and should be developed.

Spark for data analysis

BE NXCALS

NXCALS, the next generation platform for the accelerator logging system, has a critical dependency on Spark and the use of Python notebooks as the user-facing interface to data via their custom APIs. In particular, Jupyter-based notebooks (currently deployed on SWAN) allow the users to operate with the NXCALS API and analyze data of interest. A detailed explanation of the architecture of NXCALS is available for example at this talk at Spark Summit 2017: https://spark-summit.org/eu-2017/events/the-architecture-of-the-next-cern-accelerator-logging-service/

BE controls analytics

The industrial controls team has several use cases for analytics and machine learning on their controls data. Currently this is part of a pipeline that starts with copying relevant data from the controls database (currently hosted on Oracle services) into the Hadoop platform, data is subsequently processed on the Hadoop cluster using Spark. Notebooks and SWAN play an important part for the user interface and software distribution for this use case. Future developments include possible deployment of a next generation platform with controls software logging directly into a “Big Data” platform (HDFS or Kudu), therefore making analytics directly available on online data.

Physics: CMS BigData and physics analysis with Spark and ROOT

The CMS Bigdata project is currently exploring ways to perform data reduction use cases and data analysis using Spark to read and process data stored in ROOT format. Currently a small-scale demonstration for data reduction and data analysis use cases using Spark has been successfully completed (ACAT 2017 paper "CMS Analysis and Data Reduction with Apache Spark"). Further collaboration between CMS and CERN openlab is progressing to scale up the workloads to PB scale. The use case of analysis in particular fits very well with a user interface based on Jupyter notebooks and the use of Python, as in the SWAN environment. TOTEM is also experimenting on the use of Spark-ROOT.

IT Analytics working group and experiments data-management studies

Several groups in IT and experiments are using Hadoop and Spark for analysis of data originated by the operations of computing systems. These use cases typically are deployed with a "data lake" in the backend and tools for analysis at scale. Spark and notebooks interfaces are very popular tools in this context to extract value from the data for studies including understanding computing efficiency, troubleshooting, etc. ATLAS and CMS data management teams use Spark and Hadoop to perform analytics on data for their job and data file placement. See for example the ACAT 2017 paper "Exploiting Apache Spark platform for CMS computing analytics".

Similarly, analysis of "IT monitoring" data stored in the CERN Hadoop clusters can be naturally analyzed with Spark and SWAN-based notebooks.

Outside-HEP

Web-based notebooks are actually the main analysis tools outside HEP. Due to the contacts with Astronomy (e.g. SKA) and Earth Observation (e.g. EU JRC), we see that the availability of an analysis system coupled with the storage for scientific back-end (a` la SWAN) is attracting a lot of interest and be an important opportunity for extending existing collaborations. The approach of REANA (workflow engine) is equally interesting for potential future partners, for example in the medical sector.

Education & outreach

A use case where SWAN especially shines is education. Notebooks are a powerful tool to write tutorials and exercises, which students can open in SWAN, modify and execute to check their results. In order to follow the tutorial, students do not need to install any software on their machines, they just need a web browser to connect to SWAN, which eliminates any preparation burden. The same benefits apply to any outreach material produced in the form of notebooks.

A tool that is especially useful in this scenario is notebook galleries. A gallery consists of a set of notebooks, presented as screenshot images of their content, that cover a particular topic. Examples of galleries could be a course on machine learning or a tutorial on how to access, analyse and visualise open data. Thanks to galleries, users are only one click away from viewing the content they want to access: by clicking on the notebook screenshot, the user can open in SWAN that particular notebook.

CERN Analysis Preservation

The CERN Analysis Preservation framework <http://analysispreservation.cern.ch/> aims at supporting particle physics researchers with a web-based digital repository platform and a set of tools dedicated to describing, capturing, sharing and preserving the knowledge about individual particle physics analyses. The CERN Analysis Preservation project is a collaboration between the CERN IT department, the CERN Scientific Information Service, and the four LHC experiments (ALICE, ATLAS, CMS, LHCb).

The CERN Analysis Preservation framework aims at describing the current analysis practices in a structured manner that would be suitable for long-term preservation and reuse purposes. During the live analysis phase, sharing of data and analysis artefacts enables unimpeded collaboration and ultimately better results. During the publication and preservation phases, it promotes controlled reproducibility and reuse of the entire scientific process.

The ultimate goal of the CERN Analysis Preservation framework is the possibility of future reinterpretation of experimental data analyses in the light of new theories. This includes a common situation where the original analyst may have left the experimental collaboration in the meantime. The future reinterpretation of analyses requires to preserve not only to the information about the experimental datasets and the analysis software and the user code, but also the computing environment and the concrete workflow steps that were used by the original analyst to produce the original scientific results in the first place.

REANA (Reusable Analyses)

REANA (REusable ANAlyses) <http://reana.io/> is a component of the CERN Analysis Preservation framework that is dedicated to instantiating both "live" and "preserved" research data analysis on the cloud. The user specifies the input datasets and parameters, the computing environment, the analysis code and the workflow steps and REANA takes care of instantiating and running the analysis on the cloud and sharing the results back with the user.

REANA aims at supporting diverse analysis habits that exist nowadays within the various LHC collaborations and groups. It provides abstraction layers around existing workflow practices and compute environments used by the community. REANA promotes writing analysis code in a reusable manner using container technologies and offers tools and examples to facilitate the adoption of best reusable science practices by the community from the earliest stages of the analysis process.

REANA platform uses the micro-service architecture and runs on the CERN OpenStack cloud relying heavily on container technology to ensure its scalability. REANA aims at supporting several different container technologies (Docker, Singularity), several different declarative workflow engines useb in the community (Yadage, CWL, Snakemake), several different compute cloud back-ends for job execution (Kubernetes, HTCondor) and several different storage back-ends for data sharing (EOS, Ceph) used by the community.

The pilot studies carried out in close collaboration with the four LHC experiments permitted to demonstrate the feasibility of the REANA approach in the ALICE LEGO train simulation use case, the ATLAS BSM and full analysis chain example use case, or the LHCb data production and example analysis use cases.

REANA's micro-service architecture permits to profit from synergies with similar reusable science approaches existing in related scientific disciplines such as notably life sciences, bioinformatics and data science with a possibility to share components related to workflow specifications (support for CWL) or compute cloud backends (support for AWS Batch or HPC/Shifter).

Workflow-as-a-Service

The primary focus of the REANA platform is the domain of reusable science. However, the REANA platform permits to run any parallelisable workflow jobs on the shared data in the containerised compute cloud; the jobs does not have to be necessarily about particle physics analyses. The REANA platform therefore constitutes a general workflow-as-a-service solution that may be useful for running any massive IT-oriented tasks on the shared data in the containerised cloud. The use case under study includes the batch manipulation of the CERN audiovisual material by running containerised ffmpeg jobs on recorded videos.

Integration in the Computer Centre

Sustainability and harmonisation

Analysis Services as IT department services would be delivered by a combination of groups, requiring coordination during the entire service lifetime. This would mean using the tools that are commonly used for other services across the department to deploy, maintain and monitor the service. This would imply developing the appropriate Puppet configuration manifests to be able to easily instantiate new instances of SWAN with appropriate customisations and using the cadvisor container monitoring solutions widely deployed in other services such as lxplus. ServiceNow would be used for handling user requests and incidents allowing the issue to be easily passed to the different support areas such as databases, compute and storage according to the investigations. The exact mappings of components of SWAN to functional elements would be needed as part of the preparation for production.

For the compute resource provisioning, it is proposed to use OpenStack containers. With SWAN analysis load being variable according to the activities of the experiments, managing the quota allocation can be handled on a per experiment project and accounted to the pledges, growing/shrinking elastically as needed. The SWAN resource needs would need to be performed such that the procurement volumes of one of the IT bulk purchase configurations could be adjusted accordingly. This would also ensure redundancy, clear data centre capacity planning and hardware maintenance processes.

For offline analysis, a link into the HTCondor batch system would be established, potentially allowing significant workloads to be scheduled and run in non-interactive fashions while maintaining the notebook sharing. Appropriate credentials would need to be passed with the batch job to ensure access to EOS data when the job is running.

Status of the integration

SWAN and REANA are examples of services providing Analysis using the Web and building on the CERN cloud. Both rely on containers for proper workload definition, isolation and scaling, making use of existing and widely used container orchestration systems.

The OpenStack Containers project (Magnum) offers an appropriate interface for easy instantiation of these orchestration systems, with additional integration with CERN's authentication and storage systems - CVMFS and EOS in an early stage. This makes it easy to rely on common systems such as Docker Swarm or Kubernetes while benefiting from adequate integration with the remaining CERN infrastructure, including the usual service level agreements and support workflows via ServiceNow.

Additional integration is expected in areas like logging, monitoring and single sign on. The latter will likely require some development especially regarding interaction with the storage systems.

Evolution

Short-term evolution

The objective is to prepare and meet the demand from the user community for modern tools for data analytics including the demand for notebook-based frontends. In particular the SWAN solution has been adopted as a critical components for the BE-NXCALS (accelerator logging) platform and is also being used for data analysis in many different contexts including work by Experiments and IT experts. Similarly the efforts on providing data retention as initiated by the REANA project have raised interest.

The time is mature to start providing a production-quality service for notebook-based analysis managed by IT and backed by all the relevant parties, in IT and also notably in EP-SFT, as the providers of the underlying software distribution for SWAN.

<REANA>

Long-term evolution

Big-Data tools for distributed computing, data analysis and the use of notebook-based interfaces have recently disrupted the status quo for data processing in industry and have provided opportunities for the academic and research communities to profit from innovation and solutions provided as open-source products by several key-players and large communities.

We expect that current solutions for data analysis will show rapid evolution in the coming years, furthering maturing in some cases and also most likely seeing the introduction of new tools and frameworks evolving from the current offerings. These data analysis solutions are expected to grow and integrate in a the larger picture of data engineering and data science, including solutions for machine learning, data streaming and storage at scale.

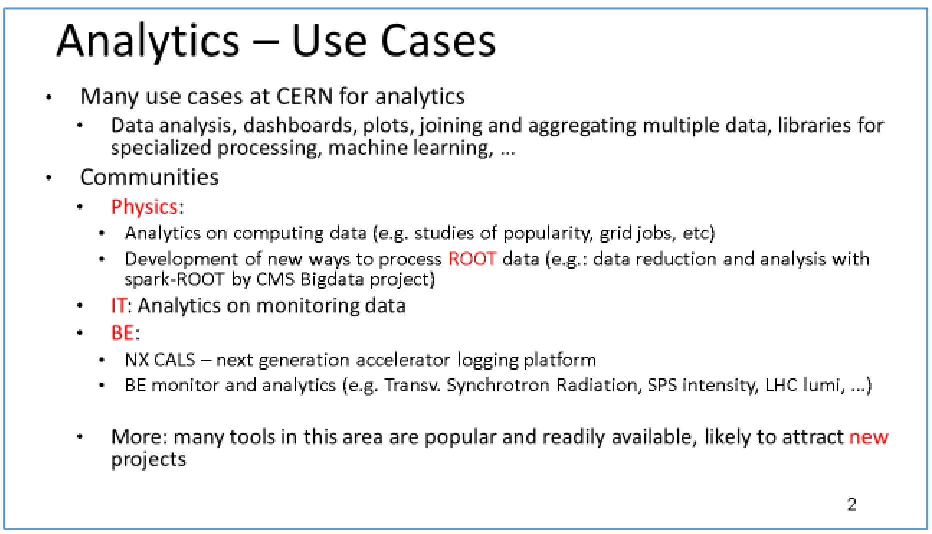
We also expect that the evolution will be driven by a few key communities who currently deploy considerable resources in research and product development. Big cloud providers and other market leaders will continue to drive the evolution of the open-source projects which we use as building blocks of our services.

Annexes

This section is a collection of reference material. It contains a number of interesting documents and link spanning from the technologies in use, to the potential technological direction, existing usage of proto services, user presentations, …

IT Management Meeting (ITMM) presentation (October 23, 2017)

Slides presented at the ITMM meeting (https://indico.cern.ch/event/662870/)



Tools and Platforms - Examples

- Many tools available – from HEP + other sciences + open source
  - Challenges and opportunities: build your solution + expect natural evolution
  - Components are modular, users adoption and traction drives many efforts
- Front ends
  - jupyter, ROOT Data Analysis Framework
- Engines and Systems
  - Spark, hadoop YARN, openstack, ROOT Data Analysis Framework, HTCondor High Throughput Computing
- Storage and data sources
  - EOS, CERNBox, hadoop HDFS, CVMFS, ceph, Databases

3

Platforms and Services

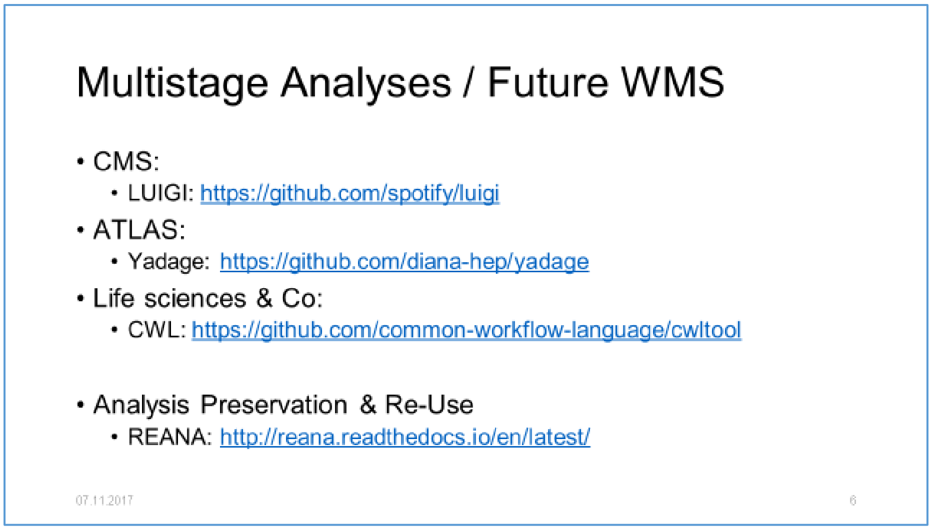
- What we think can fulfil users need
  - Analysis as a service for remote analysis -> analytics platform with notebook front-ends (e.g. SWAN)
  - Work on data preservation (e.g. CAP/REANA)
  - Solutions for data pipelines and processing (data extraction load and transformation, streaming, visualization, etc)
  - Solutions for machine learning: offer optimized platforms with software (Tensorflow, Spark, ..) and maybe hardware?? (GPUs, Intel's new chips, ..)
  - There is room for drafting a strategy doc + finding synergies with HSF?
- SWAN
  - Integrates Jupyter notebooks + relevant software (ROOT and many packages in the Python ecosystem) + integration with Spark and Hadoop and also other engines and data sources
  - Demand from users (EP, IT, BE), currently developed by EP/SFT and IT

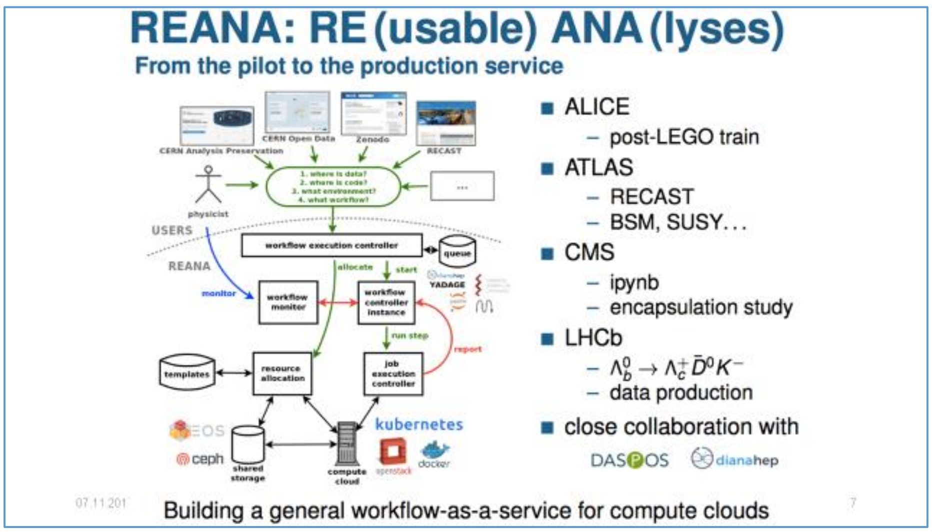
4

Whitepapers on Analysis Services at CERN

- Contributions from EP-SFT and IT (CDA, CM, DB and ST).
  - 4 pages + annexes; Massimo could be the editor
  - Deadline (draft): IT POW (~Nov 10)
- Table of Content:
  - Introduction/Executive summary
    - Architectural view, available toolchain (Jupyter, Spark, ...)
  - Use cases
    - Areas of activity and perspectives (full catalogue in the annex)
  - Integration in the CERN IT Computer Centre
    - Sustainability and harmonisation (Puppet, monitoring, ...)
    - Status of the integration (CERNBox, Condor, CVMFS, Spark, ...)
  - Evolution
    - Collaboration: 5 CERN groups to deliver analysis services to the whole community (notably BE)
  - Annexes
    - e.g. BE presentation (https://indico.cern.ch/event/668621/)

5



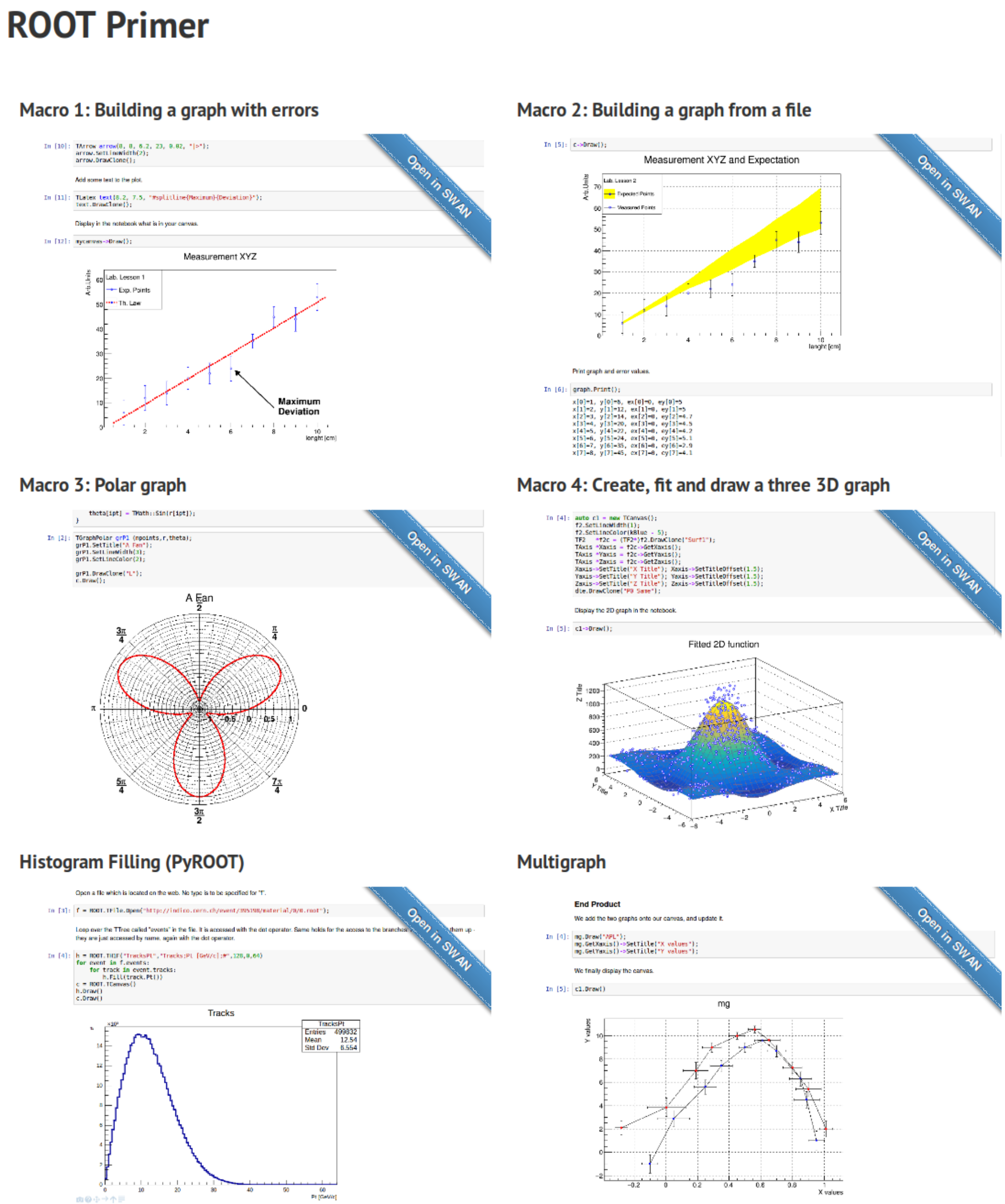


IT – BE Meeting Sept 2017

<SLIDES FROM BE?>

ROOT

The ROOT primer (as other parts of the documentations) are made available in notebook format and published on the ROOT/SWAN web site: http://swan.web.cern.ch/content/root-primer and examples/howtos .



Python notebooks for Physics analysis

In the section of the gallery http://swan.web.cern.ch/content/basic-examples many examples can be found. Notably:

- Physics: e.g. LHCb example (Anderlini) and Jet Simulation (FastJet)
- Open Data: CMS di-muon analysis
- Usage of popular analysis packages (e.g. Pandas, iPywidgets)

<TO BE ADDED>

Service for Web-based Analysis (SWAN)

SWAN provides interactive Jupyter notebooks with Python,ROOT/C++ and R kernels and it is integrated with the entire LCG software library (provided the official software environment and configuration for HEP analysis) and with EOS (data access).

The service is integrated with CERNBox home directories (EOSUSER) providing online and offline (synchronised) storage access. User files and notebooks stored in SWAN are automatically available on user's laptop (via sync clients) or central services (e.g. LXPlus via EOS fuse mounts). Sharing of notebooks is fully integrated with standard CERNBox sharing mechanisms.

SWAN is in operation since 2016. At present on average 100 notebooks are used per day (Figure 1). During specific periods (Academic Training, CSC, ...) we exceeded 200 concurrent active users.

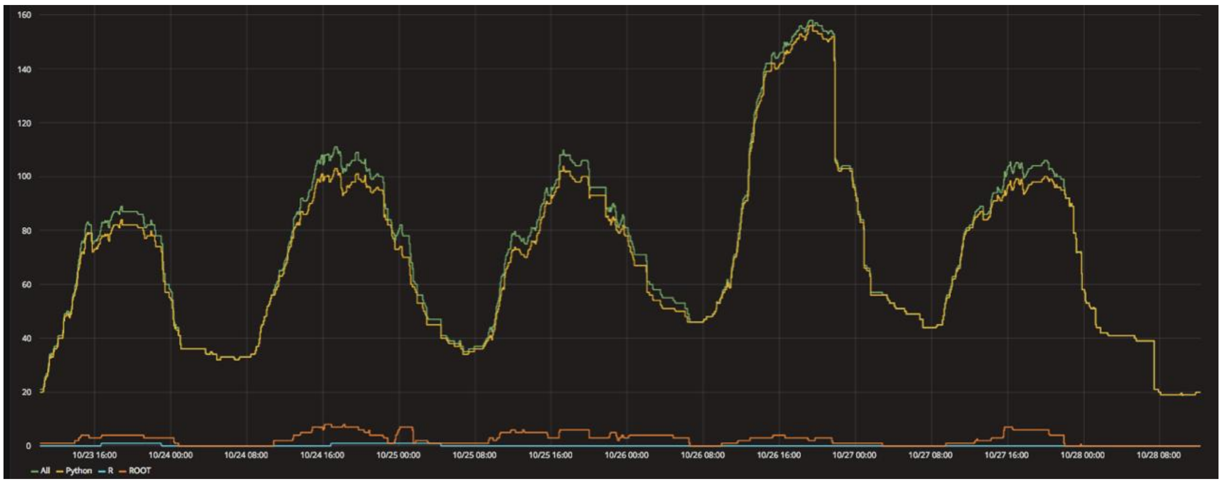


Figure 1: Weekly number of SWAN containers (user sessions) during a week (October 2017). Most of the activity is with python notebooks (PyROOT and other libraries), with a 10% contribution of C++ ROOT notebooks.

SWAN is fully integrated with IT Agile infrastructure services: the configuration is done via Puppet and the host monitoring via central monitoring service (Figure 2).

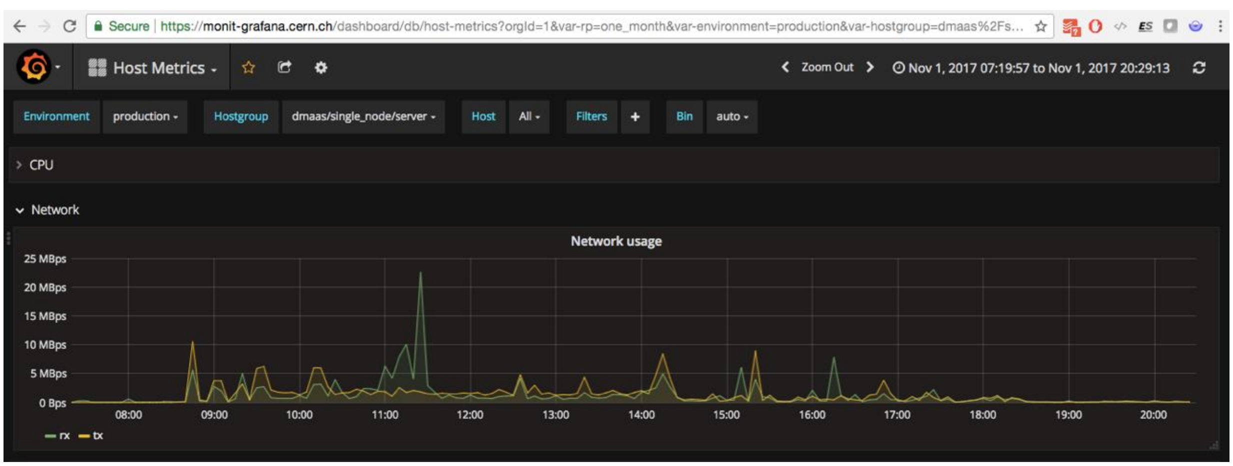


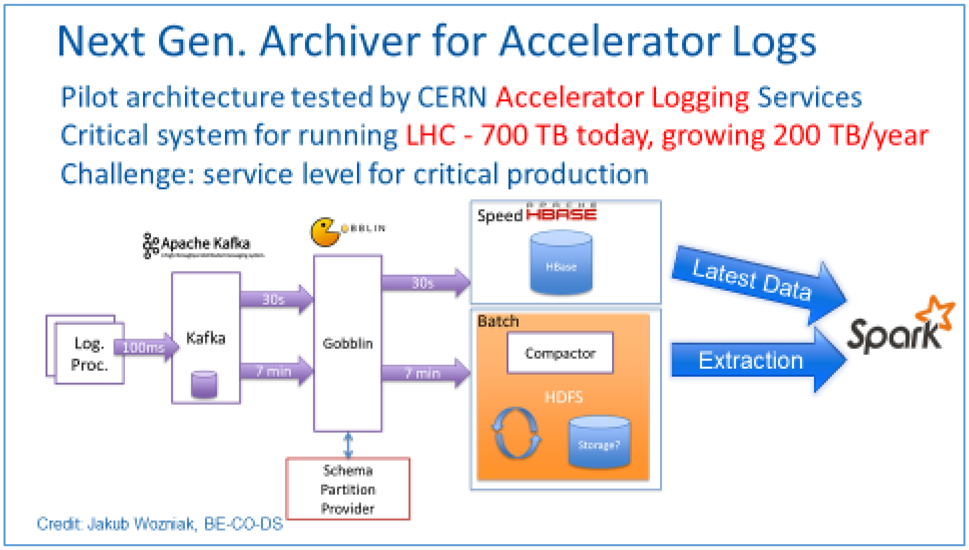
Figure 2: Front-end SWAN machines as seen by the IT monitoring.

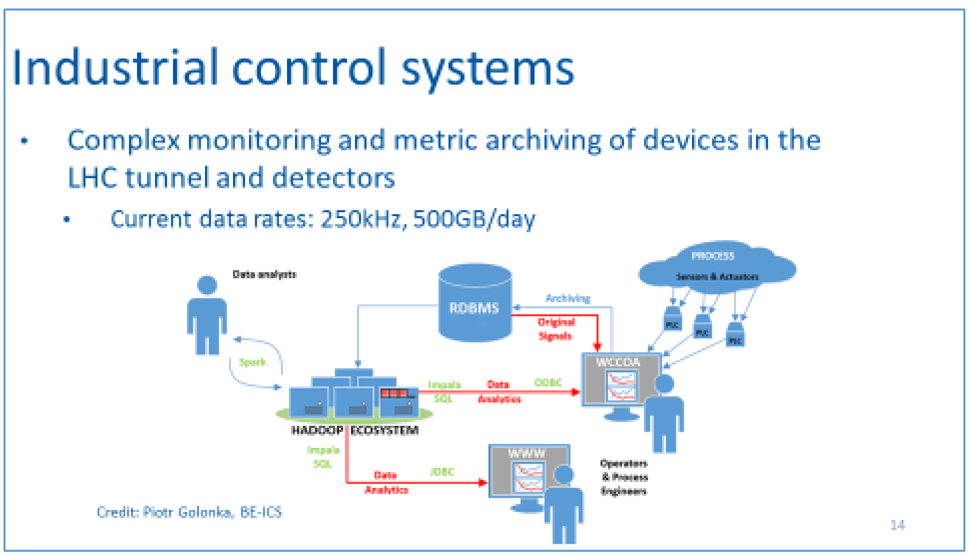
The service setup was presented to IT Security team. The service frontend is fully integrated with SSO and backend uses Kerberos 5 tickets to access underlying storage services. The SSO-KRB5 translation is done by constrained delegation with Active Directory KS4U mechanism and it is completely transparent to end users. This mechanism establishes indirect trust between storage services and SWAN host servers. AD acts as an intermediary for this trust relationship.

At present the service is deployed on 4 physical boxes (<CPU/cores, RAM statistics to be added>) for production use and a few VMs for testing and QA environment.

SPARK

Some examples can be found under http://swan.web.cern.ch/content/apache-spark: LHCb Open Data and tutorials. We list here a choice of contribution of (potential) user communities.

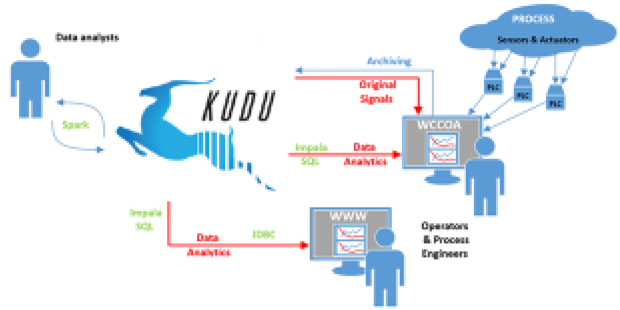




Possible Evolution - SCADA controls

What can we gain with Kudu:

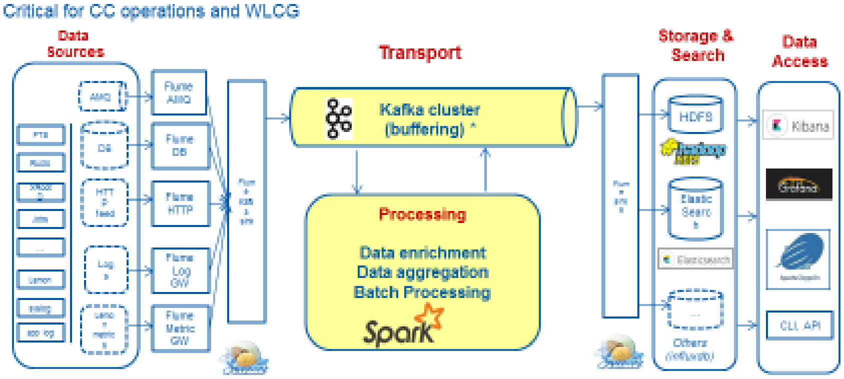
- reduce ingestion latency for analytics
- speed up live data queries and reporting
- simplification of the svstem



15

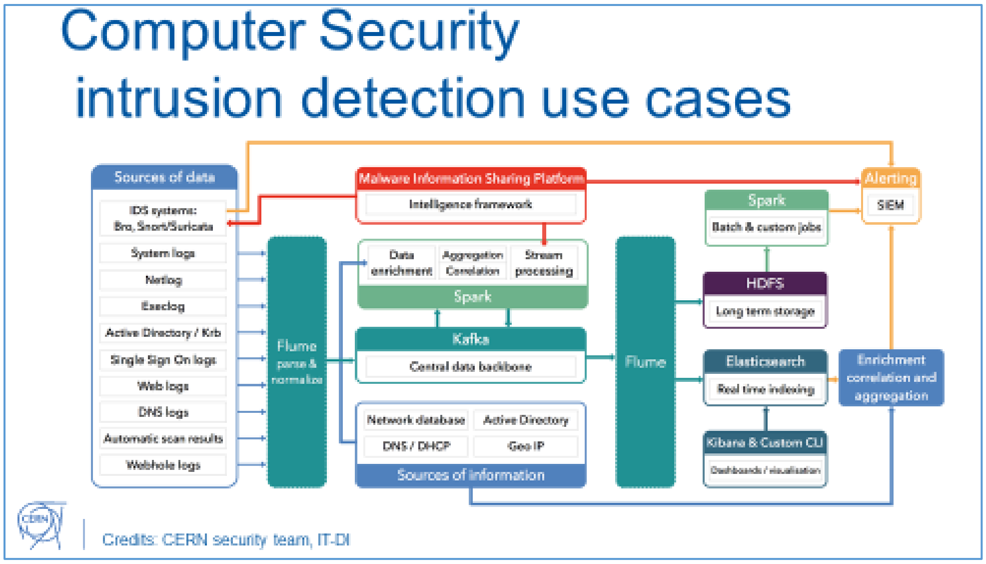
New IT Monitoring

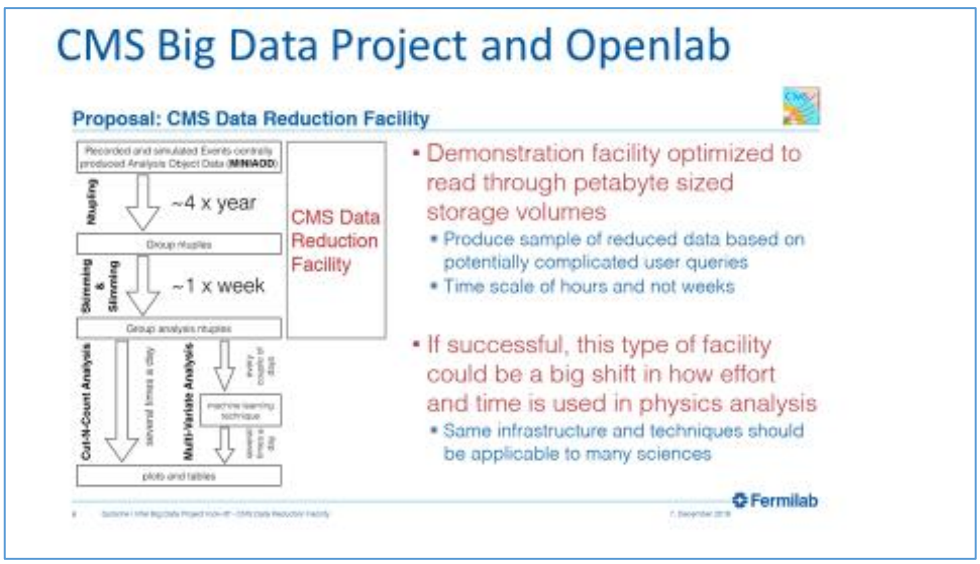
Critical for CC operations and WLCG

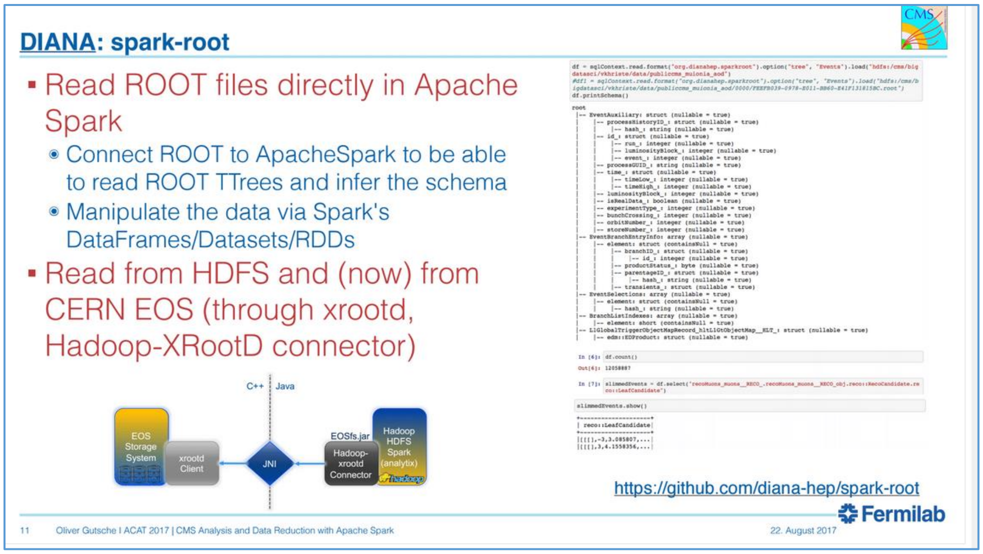


- Data now 200 GB/day, 200M events/day
- At scale 500 GB/day
- Proved effective in several occasions

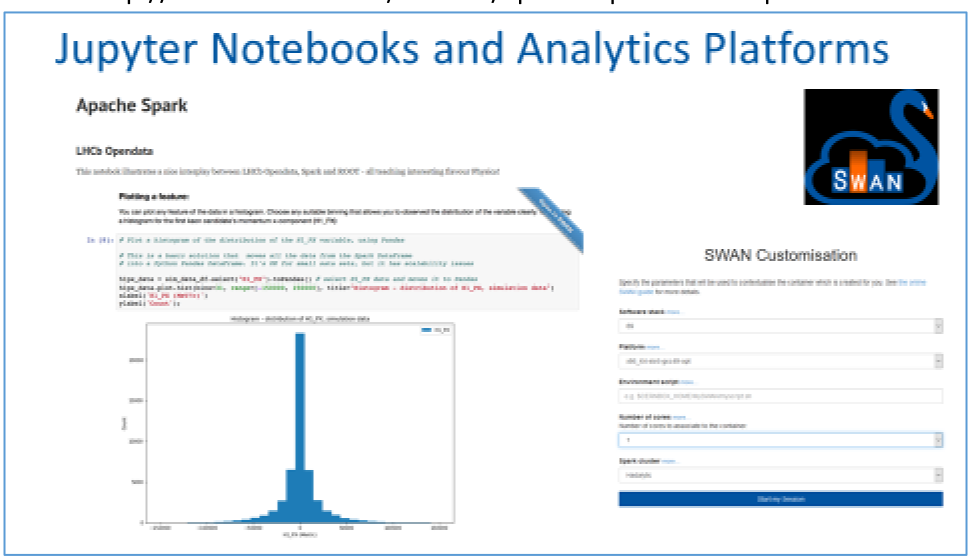
Credits: Alberto Aimar, IT-CM-MM


Computer Security
intrusion detection use cases
Sources of data
IDS systems: Bro, Snort/Suricata
System logs
Netlog
Execlog
Active Directory / Krb
Single Sign On logs
Web logs
DNS logs
Automatic scan results
Webhole logs
Flume
parse & normalize
Malware Information Sharing Platform
Intelligence framework
Data enrichment
Aggregation Correlation
Stream processing
Spark
Kafka
Central data backbone
Network database
Active Directory
DNS / DHCP
Geo IP
Sources of information
Flume
Spark
Batch & custom jobs
HDFS
Long term storage
Elasticsearch
Real time indexing
Kibana & Custom CLI
Dashboards / visualisation
Alerting
SIEM
Enrichment correlation and aggregation
Credits: CERN security team, IT-DI


CMS Big Data Project and Openlab
Proposal: CMS Data Reduction Facility
~4 x year
~1 x week
CMS Data Reduction Facility
• Demonstration facility optimized to read through petabyte sized storage volumes
• Produce sample of reduced data based on potentially complicated user queries
• Time scale of hours and not weeks
• If successful, this type of facility could be a big shift in how effort and time is used in physics analysis
• Same infrastructure and techniques should be applicable to many sciences
Fermilab



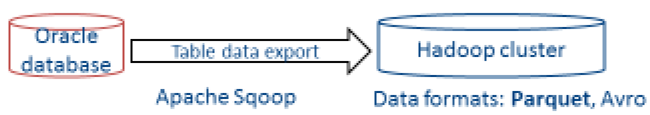
Example of using Spark for simple analysis (outreach use case). Analytics platform using SWAN and Spark. From http://swan.web.cern.ch/content/apache-spark: LHCb Open Data and tutorials



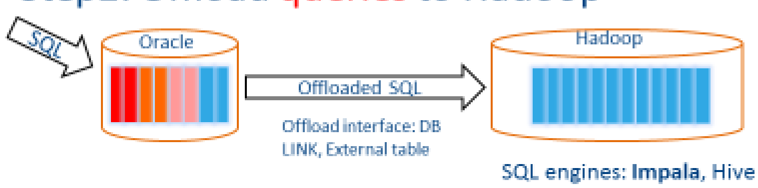
Jupyter Notebooks and Analytics Platforms

Offloading from Oracle to Hadoop

- Step1: Offload data to Hadoop



- Step2: Offload queries to Hadoop



25

Hadoop Clusters at CERN IT

- 3 current production clusters (+ 1 for QA)
- A new system for BE NXCALs (accelerator logging) platform
  - Coming in Q4 2017

Cluster Name	Configuration	Primary Usage
lxhadoop	18 nodes (cores – 576,Mem – 1.15TB,Storage – 1.17 PB)	Experiment activities
analytix	36 nodes (cores – 780,Mem – 2.62TB,Storage – 3.6 PB)	General Purpose
hadalytic	12 nodes (cores – 384,Mem – 768GB,Storage – 2.15 PB)	SQL oriented installation
NxCALS	24 nodes (cores – 1152,Mem – 12TB,Storage – 4.6 PB, SSD - 92 TB)	Accelerator Logging Service

Python notebooks for the CERN Accelerator Complex

CERN BE is one of the most active user communities of SWAN. Members of BE use SWAN regularly to access and analyse accelerator logging data that is stored in the database of the CERN Accelerator Logging Service. In order to query such database, the pytimber Python package, developed by BE, is used. Pytimber is distributed in as part of the LCG releases since July 2016, which makes it available in SWAN through the CVMFS file system.

After obtaining the data from pytimber, the usual actions include reordering/manipulating the data, doing some simple analysis, making a fit and then plotting the results. The BE colleagues use numpy to manipulate the data extracted from the database and matplotlib to display plots in the notebook. Both packages are also available in SWAN via CVMFS.

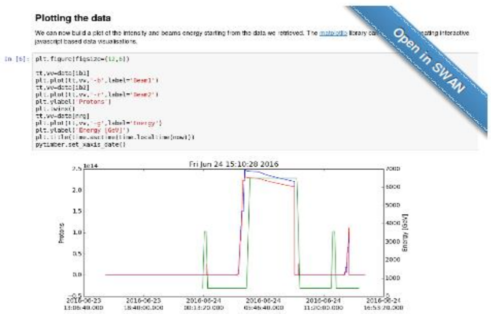
The SWAN web page includes a gallery of notebooks prepared by CERN BE (e.g. where pyTimber is used to access accelerator data as in "LHC Page 1"). See: http://swan.web.cern.ch/content/accelerator-complex.

Accelerator Complex

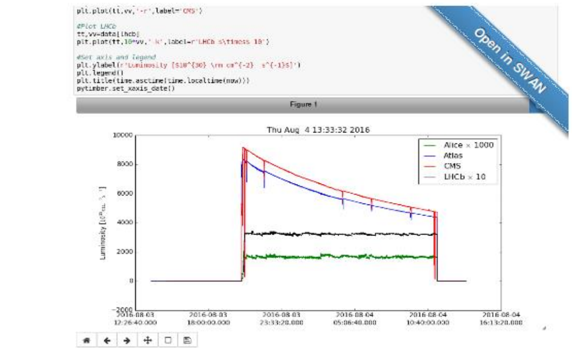
This gallery shows examples of machine studies relative to the CERN accelerators' complex.

Open in SWAN

LHC Page1

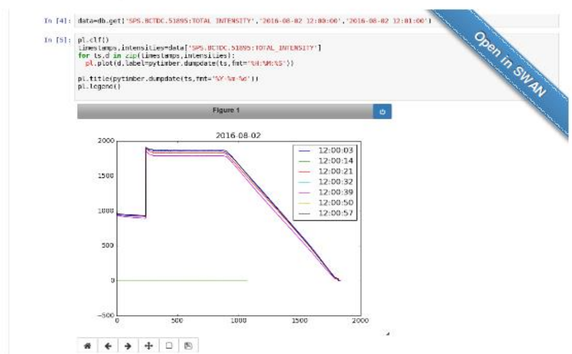


Experiments' Luminosities

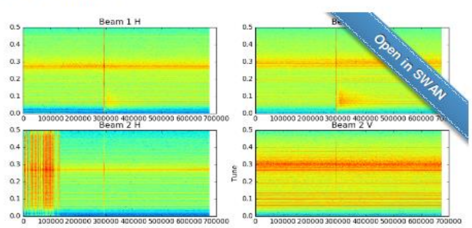


PyTimber Tutorial

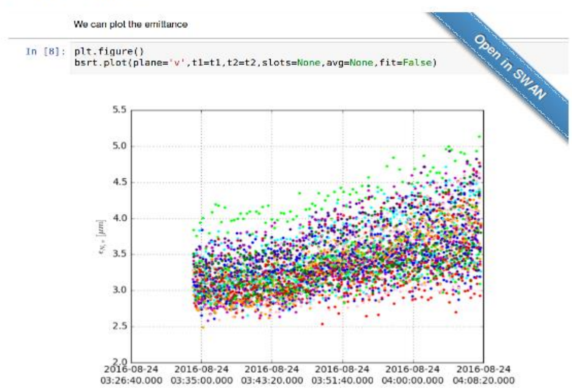
SPS Intensity



LHC BBQ Example



BSRT Example



AWAKE Experiment

SWAN is being used in production by scientists and engineers of the AWAKE experiment to analyse data coming from their acquisition system.

TOTEM Experiment

Members of TOTEM are exploring the use of SWAN combined with Spark for doing data analysis. They are currently processing files that contain ROOT tuples using batch jobs, followed by a manual reduce phase, and they want to adapt this workflow to a map-reduce Spark job.

Machine learning

Material from the IML Machine Learning Workshop (Aug 2016: https://indico.cern.ch/event/548789/). See also: http://swan.web.cern.ch/content/machine-learning

Training in HEP

Examples of training events using SWAN. In these events, the emphasis is on statistics and SWAN is used as a mean to deliver the lecture content:

- CSC: Visualization tutorial by E. Maguire: https://github.com/eamonnmag/vis-course-practical/blob/master/Visualize.ipynb; Multivariate-analysis by T. Keck: https://github.com/thomaskeck/MultivariateClassificationLecture/tree/master/exercises
- CSC: Hands-on exercises on Software Design for the Many-Core Era: https://github.com/dpiparo/swanExamples/blob/master/notebooks/CSC_Spark_Exercise.ipynb
- ROOT Summer Student Workshop: http://swan.web.cern.ch/content/root-primer
- Academic Training (e.g. Statistics course Nov 2016: CERN + Imperial College: https://indico.cern.ch/event/545212/)
- ESI Archamp training (IBD4Health School - July 2016)
- "Introduction and overview to Hadoop ecosystem and Spark“, April 2017. Slides and recordings at: https://indico.cern.ch/event/590439/

Outreach

Outreach and educational material using SWAN is listed in the SWAN gallery: http://swan.web.cern.ch/content/outreach

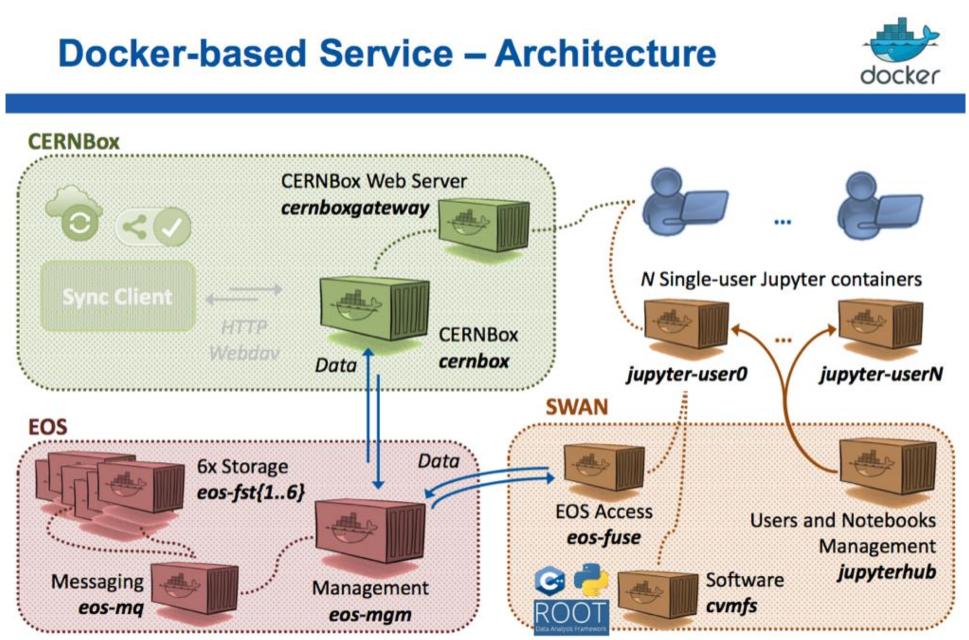
EU projects

EU project UP2U (3-year project started January 20917: up2university.eu)

Material from UP2U:

- CERN Cloud Technology for Education: https://app.box.com/embed/s/25pic76j7sngtkz8o372ikglgr3zusk3/file/141033506487
- Jupyter Notebooks: https://app.box.com/embed/s/25pic76j7sngtkz8o372ikglgr3zusk3/file/141033785754
- Up2U MicroExperiment@CERN (with Geneva University): https://app.box.com/embed/s/kvsyiqlmhwvhxsiexh441cyeojw3thhc/file/170634946867
- ITTF: Cloud-based educational services from CERN/IT in the context of the Up2U EU Project (Nov 24, 2017: https://indico.cern.ch/event/664386/) in collaboration with Geneva University

Originally developed within the UP2U project, we have prepared a 'boxed' distribution for the stack EOS-CERNBox-SWAN. This has been deployed on AWS for validation and on HelixNebula. The idea is to provide a simple yet functional complete system for evaluation, testing and for small installations (e.g. highschools). This approach has an important outreach and technology-transfer value since it drastically simplify the demonstration of our technologies. As shown in the following graph, the architecture uses Docker to organise the different components. In its simplest case, all components can coexist on a single node. This installation can also be used as an initial seed of a larger deployment (moving out the different component along with the growth of the system.



Publications

Papers

SWAN: a Service for Interactive Analysis in the Cloud: to be published in "Future Generation Computer Systems" (available as https://doi.org/10.1016/j.future.2016.11.035)

International Conferences and Workshops

Python and ROOT: Effective and Interactive Analysis of Big Data, D. Piparo at the Second Developers@CERN Forum, June 2016 (http://indico.cern.ch/event/487416/contributions/2174912)

SWAN: Data Analysis, In the Cloud, D. Piparo at the HEPData Miniworkshop, April 2016 (https://indico.cern.ch/event/512652/contributions/2143494/attachments/1262865/1867702/Swan_HEPDATA_160425.pdf (slides) and (pdf) https://indico.cern.ch/event/512652/contributions/2143494/attachments/1262865/1867703/Swan_HEPDATA_160425_Video_v2.mov (video)

DMaaS: Data Mining as a Service, D. Piparo at the "Cloud Services for Synchronisation and Sharing" Conference (CS3) , January 2016 (http://cs3.ethz.ch/presentations/Novel_applications_and_innovation/01Piparo.pdf (slides) and http://cs3.ethz.ch/presentations/Novel_applications_and_innovation/01PiparoDemo.mov (video).

Data Mining as a Service, E. Tejedor at ACAT2016 , January 2016 (https://indico.cern.ch/event/397113/contributions/1837786/ (slides) and http://iopscience.iop.org/article/10.1088/1742-6596/762/1/012039 (paper)

ROOT as a Service, D. Piparo at the "ROOT Turns 20" Workshop, September 2015 (https://indico.cern.ch/event/349459/call-for-abstracts/30/)

CERN Events

Python tools for machine data analysis and equipment control, R. De Maria BE seminar (August 2016 (http://indico.cern.ch/event/560447/)

Web-based Analysis Tools for Machine Learning, E. Tejedor at the CERN Inter-Experimental LHC Machine Learning, August 2016 (https://indico.cern.ch/event/548789)

Using ROOT with SWAN, E. Tejedor (https://indico.cern.ch/event/464391/contributions/2186139/attachments/1283264/1907382/SWAN_LHCbworkshop_010616.pdf) 7th LHCb Computing Workshop

SWAN: A Service for Web based ANalysis, D. Piparo (https://indico.cern.ch/event/479146) CMS CMG-CO Tech Meeting