Contribution ID: **199** Type: **Oral**

Data Mining as a Service (DMaaS)

Monday, 18 January 2016 14:00 (25 minutes)

Data Mining as a Service (DMaaS) is a software and computing infrastructure that allows interactive mining of scientific data in the cloud. It allows users to run advanced data analyses by leveraging the widely adopted Jupyter notebook interface. Furthermore, the system makes it easier to share results and scientific code, access scientific software, produce tutorials and demonstrations as well as preserve the analyses of scientists.

In order to use DMaaS, the user connects to the service with a web browser. Once authenticated, the user is presented with an interface based on the Jupyter notebooks, where she can write and execute data analyses, see their results inlined (text, graphics) and combine those with explanations about what she is doing, everything in the same document. When finished, notebooks can be saved and shared with other colleagues who can review, modify and re-run those notebooks.

The DMaaS service is entirely hosted in the cloud. All the user analyses are executed on a virtualised infrastructure, devoting to each user a private container that is isolated from the rest. Similarly, the input and output data of the analyses, as well as the notebook documents themselves, reside in cloud storage. The access and usage of this infrastructure is protected by the necessary security components, which ensure that every user is granted resources according to her credentials and permissions.

This presentation describes how a first pilot of the DMaaS service is being deployed at CERN, starting from the notebook interface that has been fully integrated with the ROOT analysis framework, in order to provide all the tools for scientists to run their analyses. Additionally, we characterise the service backend, which combines a set of IT services such as user authentication, virtual computing infrastructure, mass storage, file synchronisation, conference management tools, development portals or batch systems. The added value acquired by the combination of the aforementioned categories of services is discussed. To conclude, the experience earned during the implementation of DMaaS within the portfolio of production services of CERN is reviewed, focussing on the opportunities offered by the CERNBox synchronisation service and its massive storage backend, EOS.

Primary author: TEJEDOR SAAVEDRA, Enric (CERN)

Co-authors: PIPARO, Danilo (CERN); MOSCICKI, Jakub (CERN); MASCETTI, Luca (CERN); LAMANNA, Massimo (CERN); MATO VILA, Pere (CERN)

Presenter: TEJEDOR SAAVEDRA, Enric (CERN)

Session Classification: Track 2

Track Classification: Data Analysis - Algorithms and Tools