DOMA-FR Collaboration

Authors: DOMA-FR collaboration

The DOMA-FR collaboration, which is supported by the IN2P3 institute, is the French component of the international collaboration Data Organization, Management and Access (DOMA). This collaboration aims to define the technical, operational and functional characteristics of computerized data storage systems for scientific data around 2025. This document is the contribution of this collaboration to the GT09 (Computing, algorithms and data) working group of the national foresight process.

Context and identified issues

The DOMA international collaboration has been formally established during summer 2018 and its French part at begin of 2019. This collaboration aims to identify and characterize various elements that will participated to define what has to be a future scientific data storage system and how to exploit it.

This mission comes from a relatively trivial observation that, to be overcome, must be considered as a whole. This observation is because the need for scientific data storage services will explode in the coming years and that a purely projective approach of the current means and usages will not meet the future needs in a constrained financial envelope.

High-energy physics through the HL-LHC project has clearly identified this difficulty and to cope with it has initiated of the DOMA collaboration. However, many other experimental collaborations in other scientific fields (in particular astroparticles and astronomy) have identified the same difficulties. ESCAPE European project is complementary to DOMA and its WP2 working group aims to demonstrate the possible integration of some DOMA tools for the non-HEP communities. ESCAPE and other national and international project present a strong collaboration potential.

Research aspects to support

An overly focused approach taking one by one the different aspects of data management will not be enough. All the different aspects must be tackled in a coordinated way:

Functional aspects: In close collaboration with the experiments, some studies are to carry out to put in perspective the management models of the experimental data (DATA model, DATA access) and the functional capacities (including performances and reliability) of the future storage services. The main challenge is to be able to model the storage services characteristics according to the usages (DATA model, DATA access) in order to optimize the usage of the provided functionalities.

Operational aspects: Two elements have to be considered.

- Management of the data itself is an activity that involves both experiments / users and the service itself. It is necessary to design and define how data management, with this these new services, will be articulated by experiments and users. New, more relevant and effective, approaches are to be considered and qualified, notably based on quality of service (QoS) notions.

- Operational aspects of the storage service itself is also an important topic. The possible infrastructures (federative, datalake,) that will underlie the storage service will help to define the operational capacities of the service itself.

Technical aspects: As any service, the next generation of scientific data storage services will be based on technical solutions that will define partially the constraints applied to the services usages. The technical R&D activity to perform covers both hardware aspects (mass storage, network, disk storage / SSD) but also software aspects (transfer protocol, network transport layer, software to provide access to the service, safety), and has not to be reduced to a technical choice. From this, R&D activity must produce and/or identify the technical elements that will form the basis of the infrastructure on which the storage services will be build.

Conclusions

Given the challenges surrounding the management of scientific data beyond the 2024s, it is highlighting that the services that will store, manage and access data have to be imagined and must be invented. We have in front of us only few years in order to converge on the needs, uses and means for the experiments/users and services providers.

At the national level, there is an interest to be an active player during this phase of definition and characterization in order to have the ability to participate to the solution but also to bring via experiments (with their data model) capability to use these services. There is a potential for significant national collaboration around this theme.

At the IN2P3 institute level, the objective, although the same, is even clearer. Many experiments (notably HL-LHC but also LSST, CTA, etc.) have clearly identified the issues and are building their future on the ability of the DOMA project to propose a solution. It is the interest of the institute to support this effort, in order to be able to operate the service at our sites (CC-IN2P3 and IN2P3 sites / Engineers) but also make it "usable" by our experiments (Computing Model / Physicists - Physicists).