Contribution ID: **29** Type: **not specified**

# Storage usage statistics through Hadoop big data technologies

*Thursday 17 October 2019 16:35 (25 minutes)*

As one the main data centres in France, the IN2P3 Computing Center (CC-IN2P3, https://cc.in2p3.fr) provides several High Energy Physics and Astroparticles Physics experiments with different storage systems that cover the different needs expressed by these experiments.

The quantity of data stored at CC-IN2P3 is growing exponentially. In 2019, about two billion files are stored. By 2030, this number of files is expected to increase by a factor of eight.

To monitor and supervise these storage systems, several applications leverage file metadata. Information such as size, number of blocks, last access time, or last update time for each storage system are used to export customized views to users, experiments, local experts, and support and management teams. However, these applications are usually monolithic and thus do not scale well. With the load expected by 2030, with a total amount of 4 TB of metadata, this could become problematic.

To improve the scaleability of these applications, CC-IN2P3 has initiated a research and development project couple months ago. The idea is to build on data analytics framework such as Hadoop/Spark/... to process the expected massive amount of storage metadata in a scalable way.
Our objective is to set up a software architecture that will act as scalable back end for all the existing and future monitoring and supervision applications. This should improve the production time of day-to-day statistics across all the storage services that are made available to the users, experiments, the support team and management of the CC-IN2P3.

A long-term objective of this project is to become able to supervise the whole life cycle of data stored on the resources of the CC-IN2P3 and thus to ensure that the Data Management Plans provided are respected by the experiments.

In this talk we will present the current status of this ongoing project, discuss the technical choices we made, present some preliminary results, and also expose the different issues we encountered along the road to success.

## Speaker release

Yes

## Desired slot length

**Author:** DUBOIS, Antoine (CNRS)

**Co-author:** AIDEL, Osman (CNRS / CC-IN2P3)

**Presenter:** DUBOIS, Antoine (CNRS)

**Session Classification:** Storage and Filesystems

**Track Classification:** Storage & Filesystems