Contribution ID: **1** Type: **not specified**

# Optimising Data Access Analytics: Integrating dCache BillingDB with PIC's Scalable Big Data Platform

*Tuesday 5 November 2024 14:00 (30 minutes)*

PIC has developed CosmoHub, a scientific platform built on top of Hadoop and Apache Hive, which facilitates scalable reading, writing and managing huge astronomical datasets. This platform supports a global community of scientists, eliminating the need for users to be familiar with Structured Query Language (SQL). CosmoHub officially serves data from major international collaborations, including the Legacy Survey of Space and Time (LSST), the Euclid space mission, the Dark Energy Survey (DES), the Physics of the Accelerating Universe Survey (PAUS), the Gaia ESA Archive, and the Marenostrum Institut de Ciències de l'Espai (MICE) simulations.

This platform is highly scalable and adaptable for various data analytics applications. The recent integration of PIC's dCache billing database records has enabled the exploration of extensive data access logs at PIC, covering roughly eight years. We will share insights from the analysis of CMS data access at PIC, which involved processing approximately 350 million entries using PIC's Hadoop infrastructure. The current system operates with around 1,000 cores, 10 TiB of RAM, 50 TB of NVMe (for caching), and 2 PiB of usable storage. Data is accessed through HiveQL and Jupyter notebooks, with advanced Python scripts enabling efficient interaction.

This framework significantly accelerated data processing, reducing execution times for plot generation to under a minute - a task that previously took several hours using PIC's PostgreSQL databases. This enhanced performance opens up new possibilities for integrating additional data sources, such as job submissions from the local HTCondor batch system, enabling advanced analytics on large datasets.

## Desired slot length

20

## Speaker release

Yes

**Authors:** FLIX MOLINA, Jose (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES)); Mr SANTAMARIA RIBA, Marc (PIC)

**Co-authors:** PLANAS, Elena (PIC); CARRETERO PALACIOS, Jorge (Port d'Informació Científica); Mr TALLADA-CRESPÍ, Pau (PIC-CIEMAT)

**Presenters:** FLIX MOLINA, Jose (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES)); Mr SANTAMARIA RIBA, Marc (PIC)

**Session Classification:** Storage & Filesystems

**Track Classification:** Storage & Filesystems