

## Development of Big Data framework for Cherenkov Telescope Array

In the astrophysical community very large observatories are being developed, such as SKA and CTA. These observatories will produce large amounts of data (several PB/y) that will need to be archived and analyzed. Furthermore, real-time processing is required to study transients and to respond to external science alerts. These data can be defined Big Data because they have the three Big Data key concepts: volume, variety and velocity. In this paper we will describe the technology that we are developing to manage the data produced by the Cherenkov Telescope Array. CTA will produce GB/s of data and kHz of triggers from more than 100 telescopes located in two different sites. These data will be analyzed in real-time and then archived for further off-line analysis. We faced this Big Data problem using Open Source technologies: Apache Spark and Apache Kafka. The first one is a framework for distributed cluster-computing which allows to develop streaming applications able to analyze a huge amount of data. The second one provides an high-throughput and low-latency framework for real-time data feeds. Using these two key technologies we are able to manage the high data rate of CTA and also to reuse in our streaming application the code developed for the off-line analysis.

**Author:** PARMIGGIANI, Nicolo (Istituto Nazionale di Astrofisica (INAF))

**Presenter:** PARMIGGIANI, Nicolo (Istituto Nazionale di Astrofisica (INAF))

**Session Classification:** Session 5a: Challenges in science data management: science gateways - Chair: C. Knapic