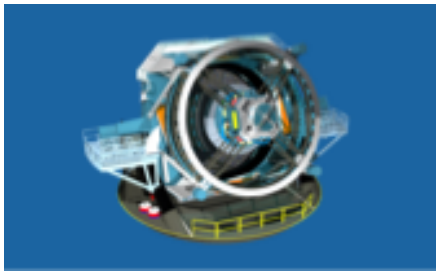

Contribution ID: **45** Type: **not specified**

Exploring Spark and MongoDb for LSST

Thursday, 15 June 2017 12:20 (30 minutes)

Spark is a very promising technology offering distributed data and computing mechanisms.

At LAL(Orsay) we have started to look at how the typical computing workflows used in LSST could use the Spark eco-system:

How to distribute algorithms in a map-reduce approach
How to format various data structures to partition them in a distributed file system
Thus, a OpenStack based cluster has been configured at LAL with Spark and its various associated components, and several models are experienced to evaluate the performance and configuration parameters (memory, CPU, ...)

In the same context, in the process in exploring various technologies related with QServ or the catalog access techniques, we are working on two promissing technologies: MongoDB and Spark DataFrames, both offering a natural data or processing distribution approach.

The method is similar for both: we exploit one limited dataset (2To) (sources and objects) and try and apply the benchmarking queries that used to be applied to QServ.

The concepts, the ingestion, and the querying methods are explored, in particular looking at possible functional or performance limitations for both systems.

Several platforms are used for this study:

The Galactica cluster at Clermont (Petasky context)
OpenStack at LAL (VirtualData context)
A test cluster CCIN2P3.

Topic:

Computing infrastructure and data management

Presenter: Mr ARNAULT, Christian (CNRS)

Session Classification: Workshop