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

Hadoop analytics provisioning based on a virtual infrastructure

Thursday, 25 September 2014 11:00 (25 minutes)

More than a year ago CESGA started to provide a Big Data service based on Hadoop. The initial request for this type of service came from the LIA2 research group working in Gaia, an ambitious mission from the European Space Agency. The initial implementation of the service was based on physical resources that were allocated through advanced reservations in one of our supercomputer clusters.

Unfortunately this lead to delays in the provisioning of the resources and the need for reconfigurations each time the Hadoop cluster was created. At that time, a devoted cluster was not a viable alternative because the demand for such service did not warranty a continuous usage resources.

After evaluating different alternatives, the service was moved to our private cloud infrastructure were the provisioning could be done in a much more flexible way. The performance loss was very small for Gaia's jobs so their executions where moved to this platform where they used up to 100 nodes to run their jobs.

The new platform allowed to offer a Hadoop on demand service to all our users where they could get familiar with the Hadoop ecosystem and develop their our algorithms.

Different alternatives to extend the virtual infrastructure have been evaluated, including an extensive study of the suitability of FedCloud to run federated Hadoop clusters and a comparison with Amazon EC2. The results of such study are promising, showing a small degradation of performance for small to medium jobs, making the FedCloud platform suitable for development and testing purposes.

Primary author: Dr LOPEZ CACHEIRO, Javier (CESGA)

Co-authors: FERNANDEZ, Carlos (FCTSG); ALVAREZ, Ivan (FCTSG)

Presenter: ALVAREZ, Ivan (FCTSG)

Session Classification: EGI-GEANT Symposium: Platforms