Contribution ID: **5** Type: **Oral Presentation**

Using Big Data for Anomaly Detection

Monday, September 23, 2019 4:30 PM (15 minutes)

During the last years, Big Data technologies, and in particular Hadoop and HBase, have enabled us to expand enormously the information that we collect and store from all our servers and infrastructures. We no longer need to discard old data using round-robin-databases or restrict the number of active nagios-style checks.

Now we can take full advantage of a metric collection infrastructure that allows to perform nagios-style checks directly against the metrics database instead of directly accessing the servers.

The information currently includes tens of thousands of time-series that are stored in HBase as well as a large collection of logs stored in HDFS.

The next challenge, is to analyze this data to detect anomalous behaviour, usually this is done by the operators looking at different operational dashboards, however data anomaly set has become too large and diverse for manual interpretation.

To take advantage of these metrics, we started evaluating generic anomaly detection techniques, applying them directly to our time-series and log data. The main problem we encountered when evaluating these generic solutions is that they produce a large number of false positives that greatly reduce their usefulness. It is not practical to have a system that produces so many alerts than it is impossible for operators to investigate all of them.

So three years ago, we started developing our own custom algorithms to detect anomalies. We will show how this approach enabled us not only to have a better understanding of our systems but also to obtain accurate results for different use cases ranging from SSH attack detection to CPU malfunctioning detection.

Primary author: CACHEIRO, Javier (CESGA)

Presenter: CACHEIRO, Javier (CESGA)

Session Classification: IBERGRID Contributions

Track Classification: R&D for computing services, networking, and data-driven science.