Contribution ID: **60** Type: **not specified**

Big Data Performance in Private Clouds. Some Initial Findings on Apache Spark Clusters Deployed in OpenStack

Friday 5 November 2021 11:40 (20 minutes)

In recent years Apache Spark has become one the most important Big Data platform. In-memory processing performance and the ability to connect with any major data server/source/format have been two of the main drivers of Spark's popularity. But finding the most suitable setup for a given data processing task is challenging, depending not only on the data structure and the nature/complexity of the task, but also because of the myriad of setup parameters to be tweaked. In this paper we propose a model for assessing the processing performance of a Spark-and-Hadoop cluster, deployed on a university cloud managed with OpenStack. Randomly generated SparkSQL queries on the TPC-H benchmark schema were executed for data sets of 5GB, 10GB and 50GB, varying four data source formats and two memory settings. Predictive models built with three Machine Learning techniques (Multivariate Adaptive Regression Splines, Random Forest, and Extreme Gradient Boosting) provided encouraging results. For the given data sets, the most important predictors seem to be related with the volume of processed data and the query complexity whereas the data formats and memory settings seem less important.

Authors: Prof. FOTACHE, Marin (Al.I. Cuza University of Iasi); Mr CLUCI, Marius-Iulian (Al.I.Cuza University of Iasi)

Presenters: Prof. FOTACHE, Marin (Al.I. Cuza University of Iasi); Mr CLUCI, Marius-Iulian (Al.I.Cuza University of Iasi)

Session Classification: RO-LCG 2021 "Grid, Cloud && High Performance Computing in Science" & Cloud Computing and Network Virtualisation