A comparative evaluation of Big Data frameworks for log processing

Attila Boros[a], Péter Lehotay-Kéry[b], Attila Kiss[c]

[a]Eötvös Loránd University
attila9778@yahoo.com

[b]Eötvös Loránd University
lkp@caesar.elte.hu

[c]Eötvös Loránd University
kiss@inf.elte.hu

Abstract

Nowadays a huge part of collected data comes from logging systems' behaviour. Examples are complex monitored systems of different institutions where computations require powerful distributed environments to run. Our work aims the specific area of log data obtained from telecommunication operator systems with the goal to identify non-trivially detectable problems, like frequency of node restarts on a given time period or the reason of these events. In order to substitute significant new information from these system logs, it is important to use proper frameworks for analyzing them. This being a comprehensive problem, various frameworks have been proposed. In this paper we evaluate and compare Apache Spark and Elasticsearch (with Logstash) as two prominent frameworks for processing log data. Through our work we perform experiments on different problem solutions with different complexity in order to measure how non-functional features, like processing time and resource consumption vary between them. Additionally, our experimental data shows that how choosing between different frameworks can influence the performance of these computations.

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002). The project has also been supported by the Ericsson-ELTE Software Technology Lab.

Keywords: Big Data, Elasticsearch, Spark, Log analysis, Telecommunication, Network, Data analysis