Contribution ID: **52** Type: **Oral Presentation**

Extracting and Structuring Real Log Files for Efficient Monitoring and Automated Computing Operations at a WLCG Tier

Wednesday, 26 August 2020 16:30 (30 minutes)

The efforts to ensure smooth operations of WLCG computing centers towards the HL-LHC challenges are crucial to support all the research communities in the next decades. In this context, the addition of automation and intelligence to the computing operations are a key factor to achieve a real-time response to problems and minimize the person-power needed to do so in the daily operations. In this paper, we focus on the case of the Italian Tier-1, as a generic data center that collects a multi-terabyte amount of log data yearly from more than 1000 machines and services. We present a strategy to ingest unstructured log data from various sources, rearrange them in a structured and classified manner, enabling the automated extraction of useful information. This work is based on the premise that: 1) previously, there is no knowledge about the log file structure or content; 2) each log file is a set of log messages of several types; and 3) each of the log message's type is generated through a template. The template represents the fixed part of the log message and informs about the occurred event's type. While in turn, the parameters constitute its variable part, giving some details of a specific event. Both give us useful information about the system status. To extract templates and parameters from log messages, an approach based on Decker Clusterization is hereby proposed. This strategy relies on a dictionary of word frequencies as centroid for each generated cluster, in which a new log message is assigned to the cluster with the biggest calculated similarity. Related to the similarity, the word frequency is used as weight in a normalized weighted sum that takes into account the log message words present in the cluster's dictionary. This dictionary saves a list of words, their corresponding identifiers, and their frequencies in the relative cluster. Once the clusterization is done, the log messages of each cluster are mapped to a numerical version, where each word is replaced by the corresponding identifier given by the dictionary. The parameters' extraction is achieved through a long-equal-sequence search strategy. The developed work furnishes a fully and directly applicable approach capable of extract both template and parameters of the log messages in a generic log file.

Primary author: DECKER DE SOUSA, Leticia (Data Science and Computation, Università di Bologna (UNIBO))

Co-author: Prof. BONACORSI, Daniele (University of Bologna)

Presenter: DECKER DE SOUSA, Leticia (Data Science and Computation, Università di Bologna (UNIBO))

Session Classification: Network, Security, Infrastructure & Operations Session

Track Classification: Network, Security, Infrastructure & Operations