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

Anomaly Detection in Data Center IT Infrastructure using Natural language Processing and Time Series Solutions

Wednesday, 22 March 2023 17:00 (20 minutes)

Data centers house IT and physical infrastructures to support researchers in transmitting, processing and exchanging data and provide resources and services with a high level of reliability. Through the usage of infrastructure monitoring platforms, it is possible to access data that provide data center status, e.g. related to services that run on the machines or to the hardware itself, to predict events of interest. Detecting unexpected anomalies is of great significance to prevent service degradation, hardware failures, data losses, and complaints from users. In the context of the data center of the Italian Institute for Nuclear Physics, which serves more than 40 international scientific collaborations in multiple scientific domains, including high-energy physics experiments running at the Large Hadron Collider in Geneva, we have performed a set of studies based on service log files and machine metrics.

Starting from our initial study aimed at combining a subset of log files and monitoring data information to detect anomaly patterns [1] involving heterogeneous unstructured data, natural language processing solutions have been applied to log files to identify words and sequences of terms as anomalies. Good results have been obtained, revealing thousands of anomalies verified by exploiting log-service messages. By defining an ad hoc clustering algorithm, various types of anomalies at the service level have been identified and grouped together. Furthermore, the adoption of a multivariate time series anomaly detection technique, called JumpStarter [2], enabled us to compute anomaly scores on monitoring data to identify the timeframe where we could overlap services and monitoring data anomalies to perform predictive maintenance analysis.

In the present work, we aim at validating the above mentioned model by considering critical scenarios and extending the range and type of monitoring data. By using error reconstruction algorithms based on, but not limited to, principal component analysis, clustering techniques, and statistical anomaly detection solutions, we plan to achieve a faster, real-time, detection of anomalies taking into consideration also the collection of past events. Furthermore, the relationship between the identified anomalies and the threshold-risk values will be assessed and shown as a dynamic level of risks to be used for predictive maintenance management. The defined pipeline can be exported to other data centers because of the usage of open souce code for its implementation. It has to be considered that training and related inference may vary depending on the amount of data provided by the data center.

References

[1] Viola, L; Ronchieri, E; Cavallaro, C. Combining log files and monitoring data to detect anomaly patterns in a data center. Computers, 11(8):117, 2022. doi: https://doi.org/10.3390/computers11080117

[2] Ma, M.; Zhang, S.; Chen, J.; Xu, J.; Li, H.; Lin, Y.; Nie, X.; Zhou, B.; Wang, Y.; Pei, D. Jump-starting multivariate time series anomaly detection for online service systems. In Proceedings of the 2021 USENIX Annual Technical Conference (USENIX ATC 21), Virtual, 14–16 July 2021; pp. 413–426.

Primary authors: Dr COSTANTINI, Alessandro (INFN CNAF); Dr SALOMONI, Davide (INFN CNAF); Dr DOINA, Duma Cristina (INFN CNAF); Dr RONCHIERI, Elisabetta (INFN CNAF); Dr GIOMMI, Luca (INFN CNAF)

Presenters: Dr COSTANTINI, Alessandro (INFN CNAF); Dr SALOMONI, Davide (INFN CNAF); Dr DOINA, Duma Cristina (INFN CNAF); Dr RONCHIERI, Elisabetta (INFN CNAF); Dr GIOMMI, Luca (INFN CNAF)

Session Classification: Artificial Intelligence (AI)

Track Classification: Track 10: Artificial Intelligence (AI)