Big Data Analysis for Predictive Maintenance at the INFN-CNAF Data Center using Machine Learning Approaches

Leticia Decker de Sousa*†✉, Luca Giommi*†, Simone Rossi Tisbeni†, Fabio Viola‡, Barbara Martelli‡ and Daniele Bonacorsi*†
*University of Bologna, Bologna, Italy
{leticia.deckerde, luca.giommi3, simone.rossi22}@unibo.it
†INFN Bologna (Italian National Institute of Nuclear Physics), Bologna, Italy
‡INFN-CNAF, Bologna, Italy
{fabio.viola, barbara.martelli}@cnaf.infn.it

***Abstract*—Predictive maintenance is a hot topic in research. It is widely applicable to the field of supporting and monitoring computing systems with the twofold objectives of increasing the operational efficiency and reducing costs through the prediction of faults. The ultimate goal of this project is to build a complete predictive maintenance solution for data centers, based on the extraction of content from the log files of services running onsite. Currently, the major Italian Worldwide Large Hadron Collider (LHC) Computing Grid data center is the INFN-CNAF, placed in Bologna, that mainly relies on reactive maintenance. In order to improve the data center quality of services (QoS), this work uses the log data as feedstock. Typically, log files are unstructured or semi-structured files, with heterogeneous data, revealing information about the system status that could be useful for its maintenance. In general, because of its characteristics, this kind of data is hard to process with standard Machine Learning (ML) supervised solutions without a deeply time and resource-consuming solution. In addition, this information can be complemented with collected environmental data to further refine event predictions or system diagnostics.**

I. Introduction

Predictive maintenance [1] is a hot topic in research, involving several domains - both scientific and industrial - aimed at increasing the operational efficiency and reducing costs by predicting and preventing faults. There are different approaches to maintenance that differ on the time of intervention in relation to a system failure. Starting from the simplest one, the approaches are reactive (1), preventive (2), and predictive (3). In (1), the human intervention occurs after a failure is detected to restore a good operating status. On the contrary, (2) is used to schedule periodic maintenance interventions to preserve a good operating status and preventing faults. With predictive maintenance instead, it is possible to intervene before a fault occurs [2]. With respect to (1), this approach allows reducing the costs and inconveniences deriving from system downtime and repairs. Compared to (2) instead, (3) allows cutting the costs imputable to unnecessary hardware replacement.

Predictive maintenance is already being applied in multiple domains, as in railways [3] or wind turbines [4], just to name a few. In this paper we discuss the application of a predictive maintenance solution to data centers. In fact, the ultimate goal of this project is to build a complete predictive maintenance solution for a data center, based on the extraction of content from log files of services running onsite.

The use-case on which this work focuses is represented by the major Italian data center in the Worldwide LHC Computing Grid [5] (WLCG), the INFN-CNAF in Bologna. It hosts a Tier-1 center with approximately 40,000 CPU cores, 40 PB of disk storage, 90 PB of tape storage, and it is connected to the Italian (GARR) and European (GEANT) research network infrastructure with more than 200 Gbps [6], [7]. As of today, in order to keep a high QoS, the center collects log data from 1197 machines. Since May 2019, CNAF opened to the local research community a NFS-accessible repository of log files, representing the ideal use case for this project.

The rest of the paper is organized as follows: in Section II, the problem faced in this project is described in detail. Section III presents an overview of the past research activities carried out at INFN-CNAF upon which the current project relies. The project's work packages are presented in Section IV. Section V proposes an overview of the current state of the art in literature. Eventually, conclusions are drawn.

II. The Problem

Data centers, as the INFN-CNAF, may substantially benefit from the adoption of predictive maintenance techniques. In fact, switching from reactive maintenance to this approach would reduce the system downtime after a failure as well as the costs for disaster recovery. However, despite preliminary initiatives and projects in progress (better detailed in Section III), currently, maintenance in the computing center selected as our use case is mainly still based on a reactive strategy and a substantial part of the work is still done by hand. The ability to predict future occurrences of problems is a crucial asset in terms of operational efficiency and automation. Moreover, a real-time monitoring of the data center would positively affects LHC experiments, with an uninterrupted availability and acceptable reaction time in debugging and troubleshooting.

In order to improve the data center QoS, we proposes a predictive maintenance pipeline using log data as feedstock.

Typically, log files are unstructured or semi-structured heterogeneous data used to extract helpful information for maintaining a computing system. In general, because of its characteristics, this kind of data is hard to work with the standard ML supervised solutions without a deeply time and resource-consuming solution. In addition, this information can be complemented with environment collected data aimed at refining the event predictions or the system diagnostics.

Fig. 1. The overview of the predictive maintenance problem

As we can see in Fig. 1, the solution, at a glance, is provided by a complete automated software that receives a generic raw log data and provides diagnosis and predictions related to the system behaviour. In this context, the diagnosis and predictions are given by a discrete probability distribution of the catalogued system events, and they differ in the time line position in which they refer: the diagnosis looks to the past and the prediction to the future of the computing system behaviour. Since computing systems are dynamics, information of both the epochs is important because, in this scenario, the system predicting based on learned events can not be perfect.

III. Previous INFN related activities

Since 2018, the INFN-CNAF Log Group is working in Predictive Maintenance in the following projects:

- a system based on the Elastic Stack Suite that collects, parses, catalogues the log data allowing queries by content, and categorises anomalies using an embedded unsupervised ML tool [8];
- a supervised ML approach that predicts anomalies, considering normal and anomalous periods of computing system behaviour, using the StoRM service log files after a proper filtering step with regular expressions [9];
- an unsupervised clusterization method based on measurements of Levenshtein distance to classify log entries [2];
- a prototype to identify a normal and an anomalous time windows considering the data generation rate and an One-class Support Vector Machine approach to classify the system behaviour as normal or anomalous (*Anomaly Identification*) [10];
- an unsupervised method based on a dictionary centroid to identify log template (*Log Data Pre-processing*) and extract parameters(*Parameter Extraction*) using a long-short string search strategy comparing the log messages two-by-two (in progress 2019);
- a Log Files Generation Statistics Report that characterises the workload of the log data, producing the probability distribution of each generated log file (in progress 2019);
- a Big Data Infrastructure that supports and applies the developed approaches in production (in progress 2019);
- a Log Simulator that generates artificial log files used to validate the developed work, providing a Log Benchmark (in progress 2019).

The project topics are complementary and aim to provide a final plug-and-play solution that can be constantly improved by a local optimisation strategy in a modular way.

IV. Project Work Packages

The project is a modular solution composed by work packages addressing each of the steps needed to generate predictions and diagnosis from a input log. The log data (or log file) that feeds the Workflow is made by log messages that are composed by a fixed part, the template, and a variable part, the parameters. Each line of the log file is a log message in this generic format, and this is the main assumption made in whole project.

This project is divided in modules as it is shown in Fig. 2. The data processing begins in the *Log Data Pre-processing* (1) that identifies the log message templates from the raw and unstructured log data, generating a classified and structured output file. In parallel, the log data is given to an *Anomaly Identification* (2) algorithm [10] to identify the most promising pieces of log data concerning finding events deeply related to the actual (or future) system status. This is an important optimising step since we have very large amount of log data and limited computing resources.

Then, a supervised *Template Classification* (3) is done using the output of the pre-processing stage (1), and feeding the *Log Sequence Pattern Processing* (4) with real-time outputs, if necessary. This step explores the log data into the time-windows that identify an anomalous system behaviour, looking for log message patterns and errors. The aim is to identify the problem causes or to predict the system status. These patterns give us the correlation among a set of log messages, identifying a “signature” of an event.

An anomaly (2) is not necessarily a problem to be solved, therefore we separate the two cases using a *Log Sentiment Analysis* (5) approach. At this point, the details from the filtered and processed log data can be extracted to generate predictions and diagnosis according to the necessity, using the *Parameter Extraction* (6). To refine the results, a *Cross-file Anomaly Tracking* (7) can be made looking for correlated information in other log files.

Each software solution module in Fig. 2, except (7), filters the initial amount of log data, in order to identify the piece of data that holds the really useful information. This is crucial to manage the computing system, focusing on decreasing costs in terms of time and resource consumption, and manpower.

Fig. 2 also gives us the actual project status. The concluded modules are the *Log Data Pre-processing*, *Anomaly Identification* [10], and *Parameter Extraction* (boxes with dashed edge), while the next step is *Template Classification* (dotted edge). Dark grey and black boxes respectively represent input and output data of our approach.

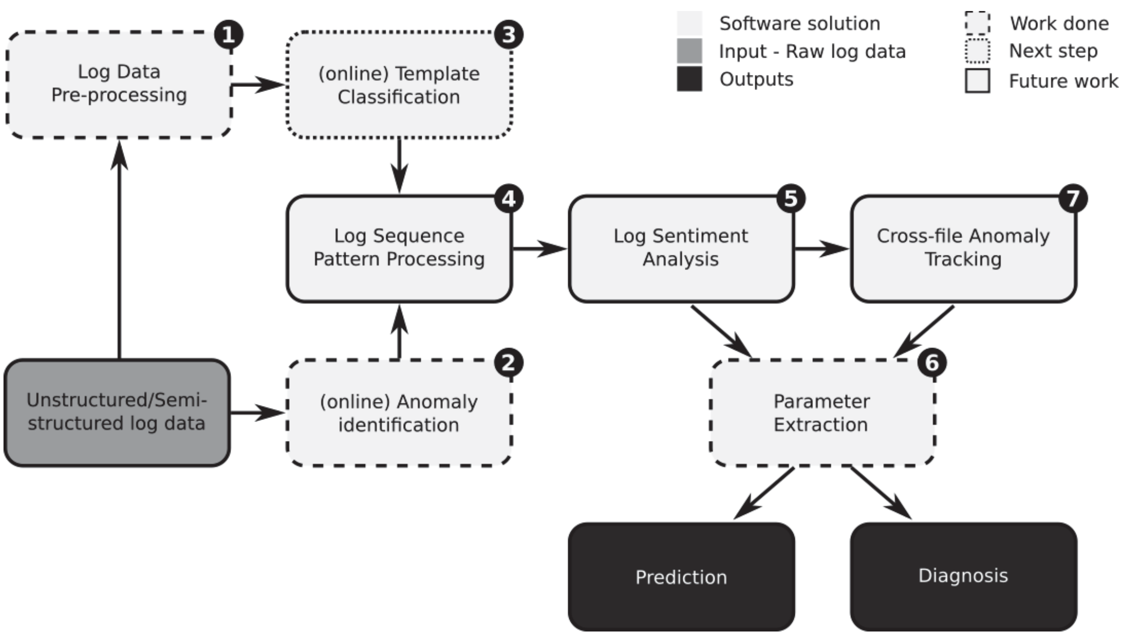

Fig. 2. The modular solution: the project work packages

V. Related Work

The growth of multiple emerging technologies contributes to the massive increase of data throughput on the Internet as well as in local networks [11]. Therefore, monitoring systems are nowadays focused on big data analysis, i.e. processing a very large streaming of unstructured data in order to reveal hidden patterns [12], [13]. In general, the intention is to identify faults [14], [15] or threats [11], [16] through batch or real-time anomaly detection. In particular, the data center is generally a rich environment of connection technologies, linking hardware and software components even in different physical machines and locations, resulting in a complex puzzle. In this scenario, it is important to concentrate efforts on diagnostic data analysis, in order to keep the services level of availability, accessibility and reliability according to the QoS agreements [12], [15], [17].

Overall, one of the main data sources to monitor the system status is the analysis of the logs [13] generated by multiple services at the center. Considering an event as an unusual occurrence, its identification is a high-dimensional anomaly detection problem. Due to exponential searching space, data snooping-bias and irrelevant features, it is a challenging topic. To overcome it, a previous work [18] combines an unsupervised Deep Belief Networks approach with an anomaly detection technique based on one-class Support Vector Machine (SVM). Anomaly detection can be explored by the classical Machine Learning approaches [16]. In particular, it can be addressed by unsupervised techniques, including methods as K-mean, and Expectation-Maximisation Clustering; passing through supervised ones as Classification-Tree, Fuzzy Logic, Neural Networks, and SVM; statistical methods as Bayes Networks; and finally, hybrid possibilities as Cascading Supervised techniques and Combining Unsupervised/Supervised Approaches. One important variation of this problem is the real-time anomaly detection (RTAD). An approach is t treats the problem of real-time anomaly detection considering streaming data with an online sequence memory algorithm called Hierarchical Temporal Memory. Another project [11] reviews the RTAD applied to network security, concluding that it is an open problem once the current approaches are not efficient enough.

In monitoring systems, predictive maintenance is the forward step from anomaly detection. A related paper [12] applies predictive maintenance in a cloud manufacturing case, providing a set of machine tools from an exhaustive combination of Machine Learning algorithms, including Principal and Independent Components Analyses, and two types of features selections.

Still in a cloud manufacturing issue [19], a research analyses the quality of a real-time predictive management cloud-based system, distributed by a cloud-storage provider, using the time-to-failure metric on hard disk drives (HDD). In another work [13], also in a HDD monitoring in a cloud scenario, it is presented a real-time predictive maintenance system based on Apache Spark to identify a forthcoming HDD failures in data centers. Still in the prevent equipment downtime [20], a work approaches a multiple-instance learning method by mining equipment logs.

VI. Conclusion

This work presents a complete plug-and-play software strategy to approach a predictive system maintenance problem using a generic log data in order to provide diagnosis and predictions outputs. The input is a generic unstructured or semi-structured log data used to generate a discrete probability distribution (DPD) of the catalogued events. This DPD can represent problem causes or future events, being a diagnosis or a prediction, respectively.

This is a 2-year project in progress. The next scheduled step is the implementation of *Online Template Classification*, following by the above-mentioned *Log Sequence Pattern Processing*, *Log Sentiment Analysis*, and *Cross-File Anomaly Tracking*.

ACKNOWLEDGEMENT

The authors would like to acknowledge INFN Bologna and INFN-CNAF for their financial support to this activity.

REFERENCES

[1] H. M. Hashemian et al., "State-of-the-art predictive maintenance techniques," *IEEE Transactions on Instrumentation and measurement*, vol. 60, no. 1, pp. 226–236, 2010.

[2] S. Rossi Tisbeni, "Big data analytics towards predictive maintenance at the INFN-CNAF computing centre.," Master's thesis, University of Bologna, 2019.

[3] P. C. L. Gerum et al.., "Data-driven predictive maintenance scheduling policies for railways," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 137–154, 2019.

[4] M. Canizo et al., "Real-time predictive maintenance for wind turbines using big data frameworks," in *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 70–77, IEEE, 2017.

[5] *Welcome to the Worldwide LHC Computing Grid*. http:/wlcg.web.cern.ch/.

[6] *INFN-CNAF*. https://www.cnaf.infn.it/wlcg-tier-1-data-center/.

[7] L. Dell'Agnello et al., ed., *CNAF 2017 – Annual Report*. 2017.

[8] T. Diotalevi et al., "Collection and harmonization of system logs and prototypal analytics services with the elastic (ELK) suite at the INFN-CNAF computing centre," in *International Symposium on Grids & Clouds (ISGC). Taipei, Taiwan: Proceedings of Science*, 2019.

[9] L. Giommi et al., "Towards predictive maintenance with machine learning at the INFN-CNAF computing centre," in *International Symposium on Grids & Clouds (ISGC). Taipei, Taiwan: Proceedings of Science*, 2019.

[10] F. Minarini, "Anomaly detection prototype for log-based predictive maintenance at INFN-CNAF," Master's thesis, University of Bologna, 2019.

[11] R. A. A. Habeeb et al., "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019.

[12] B. Schmidt et al., "Predictive maintenance of machine tool linear axes: A case from manufacturing industry," *Procedia Manufacturing*, vol. 17, pp. 118–125, 2018.

[13] C.-J. Su et al., "Real-time big data analytics for hard disk drive predictive maintenance," *Computers & Electrical Engineering*, vol. 71, pp. 93–101, 2018.

[14] A. Ambre et al., "Insider threat detection using log analysis and event correlation," *Procedia Computer Science*, vol. 45, pp. 436–445, 2015.

[15] C. Ciccotelli, *Practical Fault Detection and Diagnosis in Data Centers*. PhD thesis, 2016.

[16] S. Agrawal et al., "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[17] N. Pereira et al., "Building a microscope for the data center," in *International Conference on Wireless Algorithms, Systems, and Applications*, pp. 619–630, Springer, 2012.

[18] S. M. Erfani et al., "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, 2016.

[19] A. R. Mashhadi et al., "Moving towards real-time data-driven quality monitoring: A case study of hard disk drives," *Procedia Manufacturing*, vol. 26, pp. 1107–1115, 2018.

[20] R. Sipos et al., "Log-based predictive maintenance," in *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1867–1876, ACM, 2014.