Anomaly Detection in Data Center IT & Physical Infrastructure

E. Ronchieri, L. Giommi, L. B. Scarponi, A. Costantini, D. C. Duma, D. Salomoni

INFN CNAF, Bologna, Italy

1. Executive Summary

Anomaly detection (AD) in data centers is challenging due to the amount of heterogeneous data to be analyzed. They include, among others, CPU and memory consumption, network traffic, cooling, and electrical states. Defining a solution to early identify unexpected anomalies is important to prevent data losses, breakdown of systems, and any other event considered to be critical for the activity of the data center.

2. Problem Statement and AD Architecture

In the context of the INFN CNAF data center, one of the WLCG Tier-1s, a set of studies based on monitored cooling, electrical, and IT hardware and software metrics to detect anomalies have been performed.
Figure 1 shows a representation of the proposed AD architecture. Historical data, collected from remote sensors installed on the plant, are subjected to a pre-processing step, which makes them digestible to train predictive models. Finally, raw predictions require a post-processing step to generate an alarm signal that is easily understandable by and manageable for the human operator.

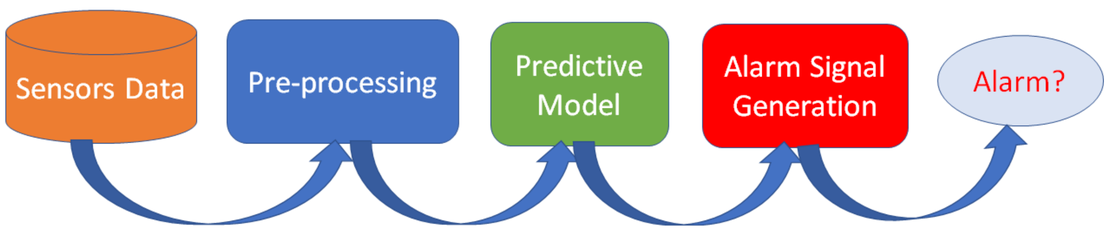


Figure 1:The proposed AD Architecture

3. Sensors Data

Figure 2 shows the types of data considered in this study.

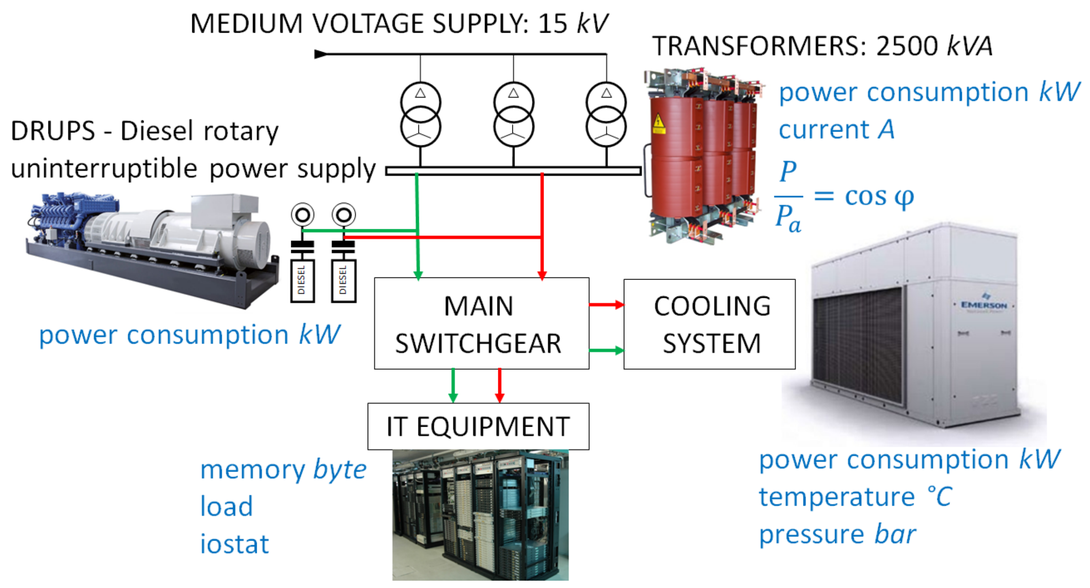
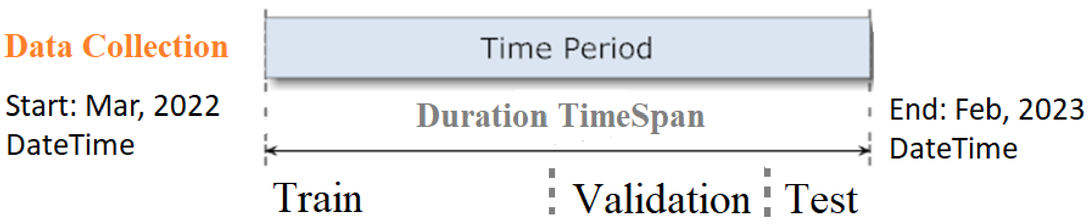


Figure 2:Data Sources - the red and green lines show the double supply.



Contact Information

- ronchieri@cnaf.infn.it - Elisabetta Ronchieri
- giommi@cnaf.infn.it - Luca Giommi
- scarponi@cnaf.infn.it - Luigi Benedetto Scarponi

4. Pre-processing

Data have been properly pre-processed before feeding predictive models. Table 1 shows a subset of assessed variables with their expected value.

Category	Variable	Unit	Expected Value
Cooling System	Tin H_2O	$°C$	$[15, 20]°C$
Cooling System	Tout H_2O	$°C$	$15°C$

Table 1:The main considered variables.

Figure 3 shows H_2O temperature values around July 28 2022 with likely anomalies, which differ from the expected values more than 3σ.

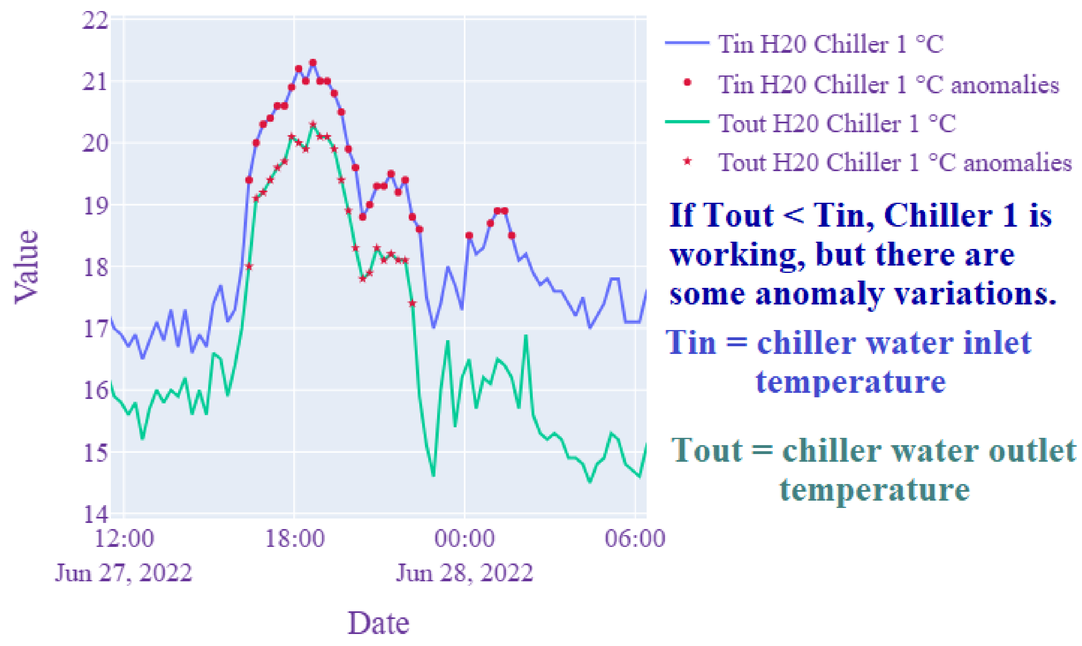


Figure 3:An example of H_2O Temperatures on Chiller 1 with related anomalies in the shape of red circles and stars.

5. Models & Discussion

The collected variables have been scaled and used to feed the proposed models, based on clustering machine learning (ML) algorithms (such as DBSCAN and K-means), to group and identify anomalous events. DBSCAN and K-means allow determining anomalies (or noisy) in the distribution. DBSCAN groups them in a cluster set to -1, as shown in Figure 4 with red dots and stars; K-means, instead, uses the last built cluster.

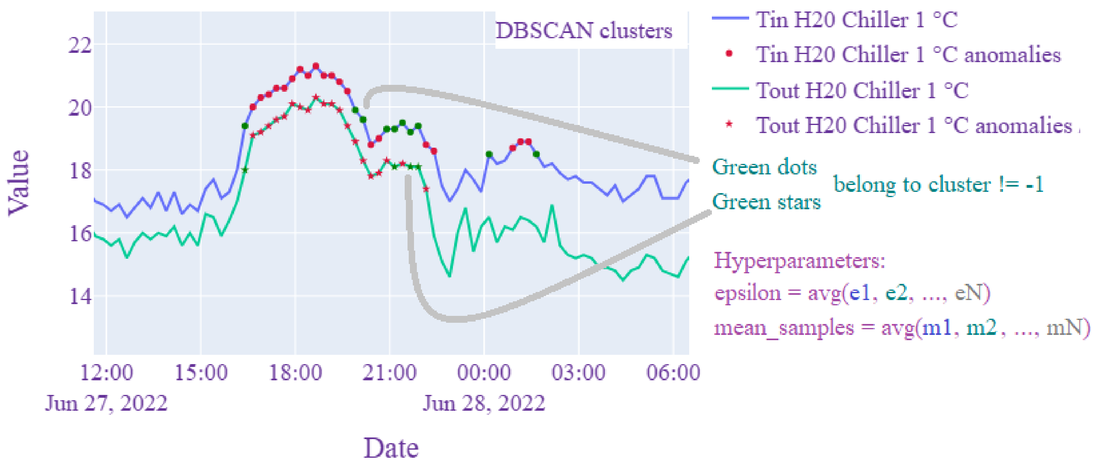


Figure 4:An example of H_2O Temperatures on Chiller 1 with related anomalies detected by using DBSCAN.

Using the presented approach, we can attribute different relevance to the likely anomalies (red and green dots and stars), contributing to turn them in proper alarm signals to trigger different actions in the data center and to refine policies aimed to address and solve arising problems.