ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-TUPHA035

DATA ANALYTICS REPORTING TOOL FOR CERN SCADA SYSTEMS*

P. J. Seweryn, M. Gonzalez-Berges, J. B. Schofield, F. M. Tilaro, CERN, Geneva, Switzerland

Abstract

This paper describes the concept of a generic data analytics reporting tool for SCADA (Supervisory Control and Data Acquisition) systems at CERN. The tool is a response to a growing demand for smart solutions in the supervision and analysis of control systems data. Large scale data analytics is a rapidly advancing field, but simply performing the analysis is not enough; the results must be made available to the appropriate users (for example operators and process engineers). The tool can report data analytics for objects such as valves and PID controllers directly into the SCADA systems used for operations. More complex analyses involving process interconnections (such as correlation analysis based on machine learning) can also be displayed. A pilot project is being developed for the WinCC Open Architecture (WinCC OA) SCADA system using Hadoop for storage. The reporting tool obtains the metadata and analysis results from Hadoop using Impala, but can easily be switched to any database system that supports SQL standards.

INTRODUCTION

In addition to the vast amount of physics data created at CERN, the control systems serving the accelerator complex and the supporting technical infrastructure generate very large amounts of data themselves. The analysis of this data is an important method for identifying and solving problems, which are not triggered by the alarms, in the underlying processes with the aim of maximizing the operational time of the CERN installations. A growing interest in smart solutions for the supervision and analysis of such data is resulting in the rapid development of many tools (e.g. Spark – a part of the Hadoop ecosystem) and advancements in machine learning. Often however, these analysis tools and techniques are applied by data analysis experts, and the results are not always easily available to process engineers and operators who might have use of this information in the daily running of the plants.

This article describes a tool that aims to 'close the loop' on the data analytics, allowing operators and process engineers easy access to the analytical data directly from the SCADA systems used at CERN. By making the results of data analytics available, operators and process engineers could gain useful and insightful information on how to maintain production systems in a scope of efficient and reliable work. The work presented in this paper is a pilot project and a proof of concept that was launched at the end of 2016, within the scope of the openlab collaboration at CERN.

* The project developed and supported through CERN openlab collaboration with Siemens, http://openlab.cern/

DATA ANALYTICS AND CERN CONTROL SYSTEMS

The Large Hadron Collider (LHC), its injector complex and its technical infrastructure represent one of the largest and most complex industrial control systems ever built by mankind. The volume of control data generated by these industrial facilities is growing year after year. The analysis of this data is crucial for evaluating and improving the performance, efficiency and predictability of the entire control system. Multiple innovative data-driven strategies have been designed and developed to deal with the ingestion, processing and storing of the huge amount of control data within a reasonable period of time.

Benefits of Analysing CERN Control Systems

The majority of raw control data does not provide a lot of value in its initial unprocessed state. Therefore multiple types of analysis have been developed aiming at distinctive, sometimes even divergent, control system aspects. As it has been already pointed out by other studies [1, 2] the detection of anomalies / disturbances in an industrial process represents a key factor in the quality of the overall control system. Several descriptive analyses have been deployed to condense long lists of control alarms into shorter, more useful system status summaries. One of the most recurrent problems was the classification of faults and their related root cause analysis. More sophisticated algorithms have been developed to detect anomalies or malfunctions in various CERN control systems [3]. These predictive analyses exploit statistical, data mining and machine learning techniques to extract a model from the historical data, which is then used for anomaly detection against the online streams of control data.

The analytics use cases have been divided into three different categories: online monitoring, fault diagnosis and engineering design. All of them focus on the control data to offer analytical services as added value on top of the traditional industrial control services.

Currently, the monitoring and operation of CERN industrial systems is mostly based on specific WinCC OA applications. Therefore, the online monitoring analytics studies aim at enhancing the services provided by the SCADA systems. For example the alarms analysis system, based on Complex Event Processing (CEP [4]) engines, has been designed to continuously collect events generated by each control device, and discover anomalous patterns [5, 6]. Specifically, the stream of historical events is analysed to calculate the number of events generated by individual devices under nominal conditions; then this information is used as a dynamic threshold to detect anomalies [7], with the assumption that a fault will generate a higher number of events. A similar approach has been adopted by the

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-TUPHA035

model learning algorithms for the detection of faulty sensors [8]. Signal oscillation detection analysis represents another example of online monitoring activity [9, 10]. This has been applied to the CERN cryogenics system for online detection of those valves that were opening and closing with abnormal oscillatory movements.

The fault diagnosis activities include those analyses that are performed offline after a fault occurrence. The root cause analysis for the Gas system is an example of such analysis. It consists of a fault isolation method based on event pattern matching. The engineering design is the third family of analysis aiming at optimizing specific control applications' aspects. An example of this is the automatic evaluation of PID performance, which supports engineers to improve the tuning of control loops [11]. All of these analyses, as well as their results, are computed and stored in a Hadoop cluster.

SCADA SYSTEM

SIMATIC WinCC Open Architecture, developed by Siemens ETM [12], is the most widely used SCADA system at CERN. It is possible to create standalone as well as distributed systems with the number of machines ranging from 2 to 2048 and linked via a network [13].

Two frameworks built on top of WinCC OA, JCOP [14] and UNICOS [15], have been developed at CERN for almost twenty years and form the basis of the majority of CERN SCADA systems. These frameworks are used in around 600 controls applications related to all aspects of the accelerators, experiments and technical infrastructure.

Figure 1 presents an example of the operator's panel in one of the cryogenic production systems. The usage of the frameworks simplifies development of applications and guarantees a unified user experience.

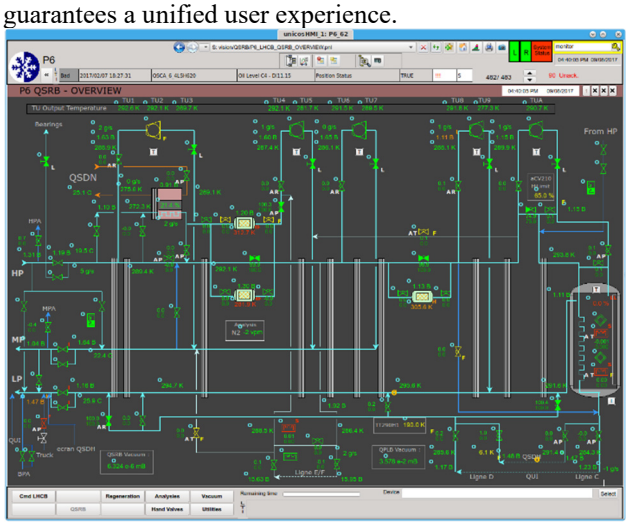

Figure 1: Operator's panel in one of cryogenic applications built with the JCOP and UNICOS frameworks.

DATA ANALYTICS REPORTING

The Data Analytics Reporting Tool connects the results of the analytics performed within the Hadoop cluster with the SCADA synoptic views of the related process objects. The analytics results are stored in the Hadoop cluster and could be accessed with SQL queries from the panels and libraries of the tool.

The *Data Analytics Notifier* informs the user that anomalies have been detected using the standard UNICOS method (i.e. blinking letters 'DA' next to the object's graphical widget – Fig. 2). This allows viewing and tracking the global situation in a particular HMI.

Based on the number of detected anomalies for a given process object, appropriate colours are applied to labels as presented in Fig. 2. In case of crossing predefined thresholds, the 'DA' label blinks red, below that value it blinks yellow and when the count is zero it is invisible. The user may preview the analyses count in the tooltip of the 'DA' label – simply by hovering the cursor over it.

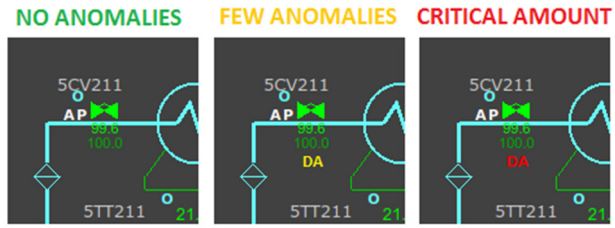

Figure 2: Different states of the UNICOS valve widget, based on the amount of detected anomalies.

This feature can be managed and configured by a dedicated panel, presented in Fig. 3. This panel allows the user to see a global overview of the situation with regard to the data analytics, collected in one place.

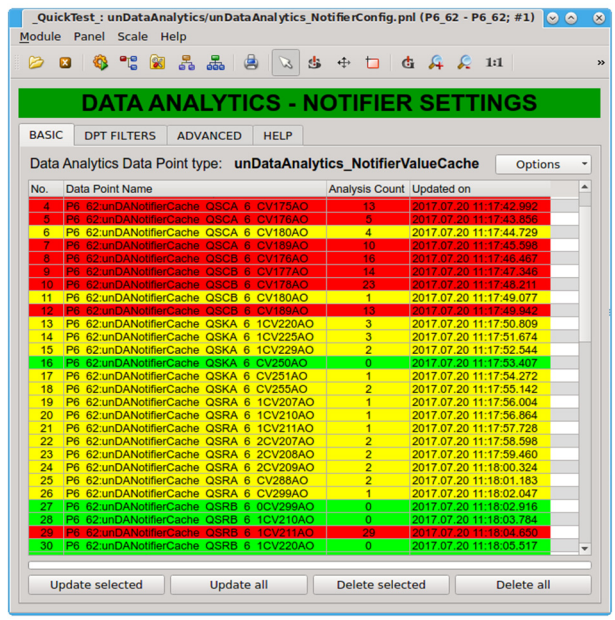

Figure 3: The *Data Analytics Notifier* configuration window.

Viewing Analyses

To view a particular device's analyses the entry point is always a dedicated graphical interface called a faceplate. The 'Analytics' tab in this faceplate, presented in Fig. 4, will be shown depending on the existence of any analysis for that object. The user can adjust the time range with several predefined options or by manual selection of the

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-TUPHA035

timestamps. Moreover, there is a dynamically populated list of all available analysis types in the database that allows filtering the results by this criterion.

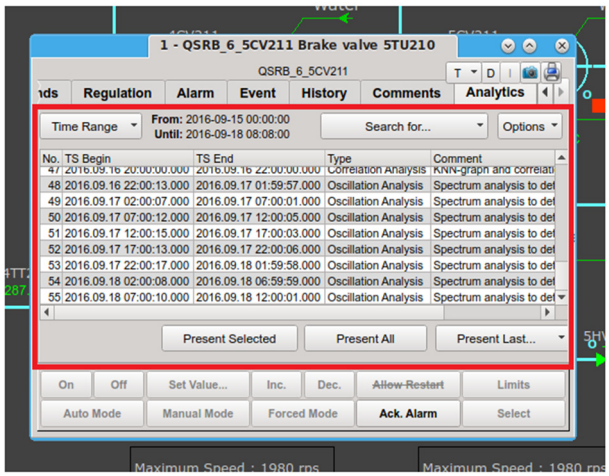

Figure 4: The Analytics tab embedded in the object's faceplate.

Finally, the user can visualize the results of the selected analyses in the view, presented in Fig. 5. This view allows storing, managing and switching between different results sets. The idea behind this panel is to collect all the opened analyses results panels in one place, rather than displaying countless independent panels. This assures clarity and order while working with the control system, preventing the situation in which the user is flooded with newly opened panels. Analyses results from different faceplates are opened within the same view which is an independent module.

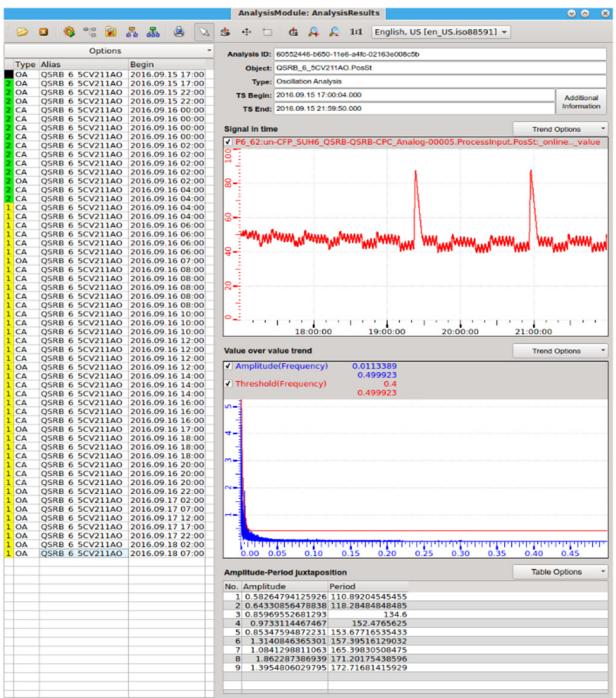

Figure 5: The view for displaying and managing different analyses results sets.

IMPLEMENTATION

Scope and Architecture

Figure 6 presents, in a simplified way, how the data from the production processes is currently being handled at CERN. Data is acquired from the field devices by the SCADA system (WinCC OA) and is then filtered and stored in an Oracle database. From there, it is being transferred on a daily basis to the Hadoop [16] cluster where large scale data analyses can be conducted using, for example, Spark [17]. The final link, marked with the red colour, is the Reporting System that reads the analytics results from Hadoop and presents them to the operator. For the moment, the reporting tool also accesses the original signals directly from the Oracle archive as it is much simpler in WinCC OA. The archiver architecture is being reviewed at the moment with the aim of simplifying it [18].

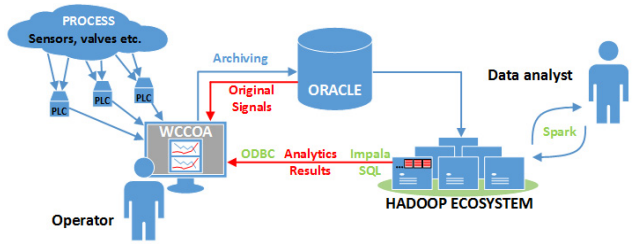

Figure 6: Simplified schema of the data acquisition, storage and extraction for the analytics reporting tool.

Technologies

Siemens' WinCC OA is the primary SCADA system at CERN and many additions are being constantly developed and upgraded, among others the C++ extension called *ctrlRBDAccess*. This extension allows communicating with different databases using a set of functions and APIs (Application Programmable Interface), one of which is ODBC (Open Database Connectivity), within the CTRL scripts (CTRL is the built-in WinCC OA scripting language). Using ODBC mechanisms it is possible to communicate with any database that incorporates the SQL standards. Hence, it is possible to easily switch from Impala (part of the Hadoop ecosystem) [19] to any other database by simply modifying the connection parameters. This assures flexibility and clarity with regards to the selection of storage for the data analytics results and its availability for any other system that one may want to use.

As mentioned above, the analytics database resides within the Hadoop cluster and it is accessible among others using Impala and SQL. Data is stored in Parquet files using Snappy compression.

Analysis Results Database Structure

The database, whose schema is shown in Fig. 7, consists of several tables. The main table (*ANALYSIS_ALL*) is an entry point that stores all metadata of all analyses like signal name, timestamps, additional comments etc. Other tables hold the results of specific analyses types which are accessible using the id of a particular analysis. This means that each type of a data set (e.g. data for plotting amplitude

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-TUPHA035

and threshold against frequency) is stored in a dedicated table. This guarantees clarity of the schema and easiness in adding further types of data sets for different new analyses that will be performed in the future.

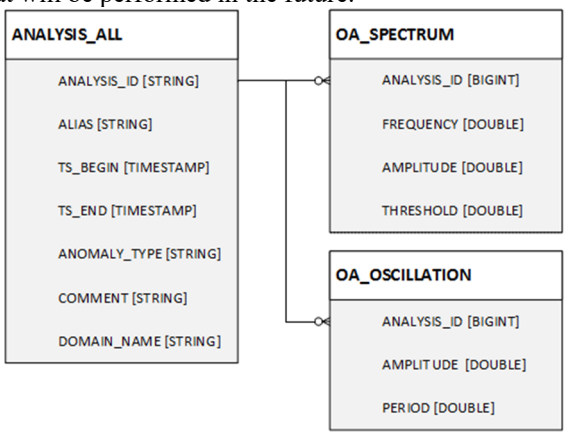

Figure 7: Part of the analytics database schema.

It is worth mentioning that the database is still evolving and expanding. More metadata will be stored to assure the integrity and versatility of usage for different systems. For example, control systems have direct access to the archived original signals (Oracle Archiver in WinCC OA), but custom applications or web services which would like to refer to it, need to know their location. Another question is to determine how to store the data: whether it should be a single database or each control application or groups of controls applications should have their own databases.

Analysis Results Presentation

Due to the existence of multiple analyses types, which includes analyses for single objects (e.g. oscillation analysis, PID performance), multiple objects (e.g. correlation analysis) and even entire subsystems and systems (e.g. electron cloud analysis), one of the project's goals was to provide a generic and versatile user interface and experience.

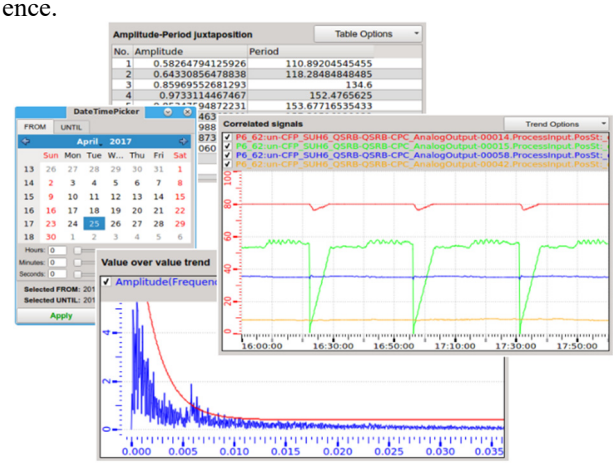

Figure 8: Examples of templates.

Since all analyses present similar data in a similar way (e.g. trends, tables, plain text) a set of templates was created. These templates, presented in Fig. 8, are a set of parameterised reference objects and associated functions which assure similar user interface and experience when opening a panel for any analysis type. Usage of the templates reduces the amount of code that needs to be written by assuming certain similarities between ways of presenting data. Some of the templates are static e.g. displaying some information/metadata about an analysis, others are dynamic and parametrized, like tables and trends.

In order for the templates to work properly, to be as simple as possible and easy to maintain, some libraries were created. In order to be compliant with the two frameworks widely used at CERN, i.e. JCOP and UNICOS, the tool was divided into a set of framework components:

- *fwDACore* is a core component, that includes all generic elements like templates, simple objects libraries, but also the library used to communicate and extract the data using the C++ extension *ctrlRDBAccess*. It allows configuring the connection, preparing and installing the ODBC configuration file and adjusting some settings.
- *unDataAnalytics* is a component used to present the data analytics to the operator in a way compatible with the UNICOS framework. It consists of faceplate tabs, views for presenting multiple analyses results and a feature called the *Data Analytics Notifier* which was described earlier. This component requires the *fwDACore* and could be treated as a basic/sample implementation with further possibility to expand it.
- Analyses results panels are a group of components and they were created as examples of how templates, mentioned before, could be used. These components are oriented mainly on the application level and what user needs. Hence, it is possible to install only selected panels and prepare custom versions of the others.

Figure 9 presents an example of a panel for an electron cloud analysis. It is a bit more complicated than others like oscillation and correlation analyses panels as it is a very specific type of analysis, describing the behaviour of an entire LHC sector. Because of this, it is being displayed independently i.e. from the toolbox menu, with the possibility to change the time range and the mathematical model of the analysis.

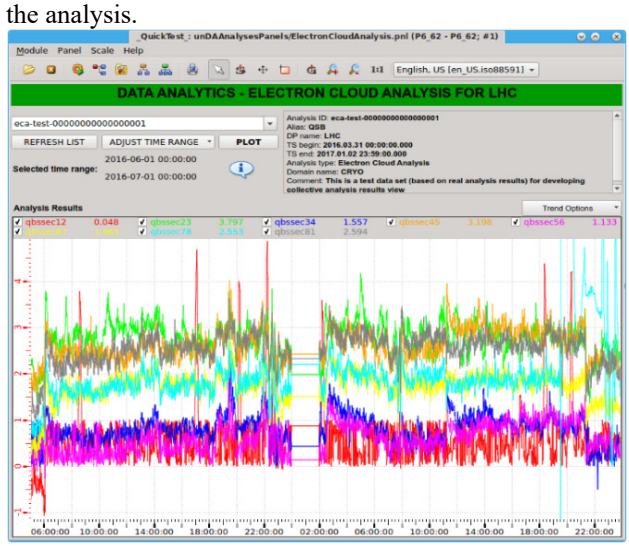

Figure 9: Panel for an electron cloud analysis.

ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-TUPHA035

Limitations of the Current Approach

During the development process certain obstacles emerged. The primary problem was that the user experience was impacted by high CPU and memory usage.

Plotting multiple huge data sets is a time consuming and resource intensive task. Moreover, presenting multiple panels at once can freeze the UI (User Interface) of the control system which is unacceptable in the case of production systems. The proposed solution was the view, described earlier, that collects and manages all opened analyses results. Using asynchronous data extraction and the synchronized mechanism of WinCC OA it was possible to manage opening a huge amount of data sets while avoiding a system crash. The efficiency of the presented solution was acceptable but another issue appeared. The typical result set usually consists of several plots and/or curves which can stress the memory of the machine that is responsible for running the controls application, especially when multiple analyses panels are opened. Unfortunately, removing the data set and reinitialising it (i.e. extracting the data and plotting it) each time would not be performant. An idea that could reduce the impact of this problem on the user experience and also improve the graphical design of the user interface is described in the following section.

Ideas for the Future

There is an idea to replace the current solution of displaying analyses results panels in the view. Instead of creating analysis panels and using templates or regular UI controls a web service could take over that task. A web display control (*WebView*) could be embedded into the view and the navigation table would be used to send appropriate requests. There are several advantages of this solution. First of all it would enhance the graphical user interface with modern web based trends and graphs. Secondly, the memory load from the local machine that runs the application would be transferred to the remote service. Moreover, this would reduce the source code, especially the parts for switching the analysis panels and templates themselves. This means less code to maintain and upgrade on the side of the controls systems. Finally, a web service could also be accessible from other applications and normal web browsers. This would grant an easier and wider range of access for a larger amount of people involved in maintaining and controlling production systems.

CONCLUSIONS

The tool presented in this paper has been created from a need to 'close the loop' between analysis of archived data, and the daily operation of the processes. This need has come about due to the rapid recent developments in methods and tools for data analysis (as seen for example in the growth of the Hadoop ecosystem). While these tools are very powerful in the hands of data scientists, they are typically not used directly by process engineers and operators, and the results of the data analysis may therefore not be leveraged as well as they could be. The tool described in this paper allows direct reporting of analysis results within the SCADA system, and provides a number of levels of information, from simple warning symbols to indicate that anomalies may be present in particular process objects, to the display of trends and frequency domain plots for an individual analysis.

REFERENCES

[1] S.J. Qin, "Control performance monitoring – a review and assessment", Comput. Chem. Eng. 23.
[2] L. Desborough, R. Miller, "Increasing customer value of industrial control performance monitoring -Honeywell's Experience", AIChE Symposium Series. 98.
[3] A. Voitier, M. Gonzalez-Berges, M. Roshchin, F. Tilaro, "Formalizing Expert Knowledge in Order to Analyse CERN's Control Systems", ICALEPCS 2015, Melbourne, Australia.
[4] E. Wu, U. Berkeley, "High-performance complex event processing over streams", SIGMOD'06 Proceedings of the 2006 ACM SIGMOD international conference on Management of data, Pages 407-418.
[5] V.J. Hodge, J. Austin, "A Survey of Outlier Detection Methodologies", J. Artif Intell Rev (2004) 22: 85. https://doi.org/10.1007/s10462-004-4304-y
[6] S. Ramaswamy, R. Rastogi, K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets", SIGMOD'00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data Pages 427-438.
[7] E. Knorr, "Algorithms for Mining Distance-Based Outliers in Large Datasets", VLDB'98 Proceedings of the 24rd International Conference on Very Large Data Bases Pages 392-403.
[8] F. Tilaro, B. Bradu, M. Gonzalez-Berges, F. Varela, M. Roshchin, "Model learning algorithms for faulty sensors detection in CERN control systems".
[9] T. Miao, D.E. Seborg, "Automatic detection of excessively oscillatory feedback control loops", Control Applications, 1999. Proceedings of the 1999 IEEE International Conference on Control Applications (Cat. No.99CH36328)
[10] F. Tilaro, B. Bradu, M. Gonzalez-Berges, M. Roshchin, "An expert knowledge base methodology for online detection of signal oscillation", 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), June 2017.
[11] B. Bradu, E. Blanco Vinuela, F. Tilaro, "Automatic PID performance monitoring applied to LHC cryogenics".
[12] ETM, http://www.etm.at
[13] Siemens, https://www.siemens.com
[14] JCOP, http://jcop.web.cern.ch
[15] UNICOS, http://unicos.web.cern.ch
[16] Apache Hadoop, https://hadoop.apache.org
[17] Apache Spark, https://spark.apache.org
[18] P. Golonka, M. Gonzalez-Berges, J. Guzik, R. Kulaga, "Future Archiver for CERN SCADA Systems".
[19] Cloudera Impala, https://www.cloudera.com