TOWARDS A SECOND GENERATION DATA ANALYSIS FRAMEWORK FOR LHC TRANSIENT DATA RECORDING

S. Boychenko, C. Aguilera-Padilla, M. Dragu, M.A. Galilee, J.C. Garnier, M. Koza, K. Krol, R. Orlandi, M.C. Poeschl, T.M. Ribeiro, K. Stamos, M. Zerlauth CERN, Geneva, Switzerland
M.Z. Rela CISUC, University of Coimbra, Portugal

Abstract

During the last two years, CERNs Large Hadron Collider (LHC) and most of its equipment systems were upgraded to collide particles at an energy level twice higher compared to the first operational period between 2010 and 2013. System upgrades and the increased machine energy represent new challenges for the analysis of transient data recordings, which have to be both dependable and fast. With the LHC having operated for many years already, statistical and trend analysis across the collected data sets is a growing requirement, highlighting several constraints and limitations imposed by the current software and data storage ecosystem. Based on several analysis use-cases, this paper highlights the most important aspects and ideas towards an improved, second generation data analysis framework to serve a large variety of equipment experts and operation crews in their daily work.

INTRODUCTION

The consolidation and upgrade activities during the recent long shutdown has allowed almost doubling the collision beam energy to 13 TeV. This implies as well a much increased damage potential in case of the serious failure in comparison to previous runs. As the rest of the machine, the systems which protect the LHC have equally undergone major improvements based on the experience collected during the first operational run between 2010 and 2013. Besides the increased number of machine protection devices installed in the machine, the amount of data which they produce also increased. The software which is used for analysis of accelerator diagnostic data requires a major upgrade to keep up with the LHC expansion and the unprecedented amounts of data to be processed [1].

Based on experience of the previous hardware commissioning we have identified the main shortcomings with the current data analysis approach. Combined with the use cases collected over several operational years and the feedback provided by operators and hardware experts, we were able to define the main goals and requirements for the next generation accelerator data analysis framework. In order to solve the identified problems, an extensive study was performed, followed by a comparison between existing data storage and processing tools. Although, the integration of modern data analysis frameworks might improve the situation with respect to the currently implemented approach, we believe that specificities of collected data and environment itself pose a greater challenge. Based on the initial research, a new workload-driven approach for organizing the data is proposed, aiming to serve efficiently heterogeneous workloads. The strengths and the weaknesses of the proposed solution are being analysed, as well as we discuss the details of the integration of the proposed technique with currently deployed accelerator software stack. Finally, the goals for the future research and development of the new analysis framework are defined.

The remaining document is organized as follows: next section provides a short overview on the current analysis framework shortcomings and desired requirements for the new solution. The following section describes the novel approach which is believed to allow building the system capable of satisfying the majority of the use cases. In future work we outline the goals for future research and development. Finally in the last section a summary on the main conclusions is presented.

CURRENT ANALYSIS FRAMEWORK

During the last few years, significant efforts were made to initiate the process of LHC transient data analysis centralization. The operators and equipment experts were previously relying on multiple service-specific graphical tools to study the underlying data. The process required significant effort and concentration, especially during accelerator commissioning when thousands of tests are executed in order to validate the behaviour of control systems. Furthermore, the data correlation between different storages was either performed manually or required the development of custom modules, which would take care of the required merging process. A recent attempt on automatizing the validation process for LHC protection systems resulted in the organization and systematization of the analysis process. The results were consolidated into the AccTesting framework [2] which has proven its efficiency during the last LHC hardware commissioning phase.

Infrastructure

There are two major sources of LHC diagnostic data: The Post Mortem (PM) framework [3] and CERNs Accelerator Logging Service (CALS) [4]. Despite acquiring the similar data from a given LHC device, the two services have fundamentally different use cases and properties. The main differences between the two frameworks is the resolution of collected data. The Post Mortem system only collects data around interesting events, by retrieving all the entries from the internal buffer of the monitored devices.

Copyright © 2015 CC-BY-3.0 and by the respective authors

ISBN 978-3-95450-148-9

Internal device buffers record the data with very high frequency (up to nanosecond precision), allowing to reconstruct very precisely the accelerator and equipment system states around events such as the beam extraction. Given the amount of the devices and produced data quantities it is impossible to store all this information continuously with the same acquisition frequency during operation. On the other hand, the collection of continuous low-frequency data provides a broader overview on eventual root causes of problems. This requirement is satisfied by the CERN Accelerator Logging Service, which retrieves the data on change with a maximum frequency of a few Hz. The collected information is used not only to understand failure sources but also to analyse the long-term performance and evolution of LHC operation.

On top of the aforementioned storage systems, users can either implement external modules to perform automatic data analysis or run arbitrary data extractions. In the first case, the modules either subscribe to determined devices or poll the data continuously from the storage, in order to identify or classify complex events. In case of manual analysis, users either use the available service-specific graphical tools (which mostly do not allow for correlation of data between the sources), or provided APIs [5] which require programming skills to be used efficiently. Recently, developments towards a data querying engine have started based on specifically designed for accelerator environment Java embedded Domain Specific Language (eDSL) [6]. This eDSL provides an abstraction layer over data extraction from different storages (see Figure 1). Queries are being written in English-like syntax language, specifically designed for CERNs accelerator environment. Currently eDSL supports assertions on multiple signals, which were required to analyse the majority of the hardware tests during the last LHC commissioning phase.

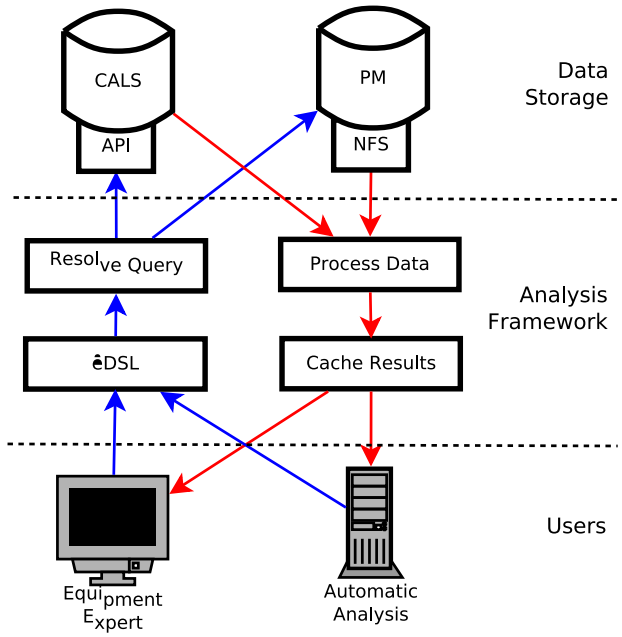

Figure 1: Java embedded Domain Specific Language data analysis framework work flow. Instead of pushing the responsibility of data analysis to users, the eDSL based solution performs analysis its servers.

Next Generation Analysis Framework Requirements

The experience of the last commissioning and operational period of the accelerator indicates the heterogeneity of the workload to be handled by the framework. When the accelerator systems are being tested the analysis of localized device types or sector based data is being performed for short time intervals (related to test execution periods). When the machine is in operation the data from the past few weeks might be retrieved to study the behaviour of specific equipment. During long shutdown phases a more extensive analysis of the overall performance of the LHC is conducted to detect the most common failure sources or to determine the efficiency of machine operation and the resulting physics output. This use case requires the extraction of data for large periods of time.

The analysis of the downsides of the currently deployed solution suggests that the new framework should be flexible enough to perform calculations close to the data source. In many use cases an expert is interested only in a small fraction of data resulting from simple aggregation operations. The limitations of the current systems impose the transfer of the whole data set in order to perform the required calculation on the users' infrastructure. The data transfer can be further optimized through efficient caching mechanisms.

A successful integration of the system into the existing accelerator environment implies the integration of the currently developed and maintained eDSL, specifically designed for accelerator diagnostic data analysis. The integration with eDSL is among the most prioritized requirements, since most of the operators are coming from areas other than software engineering. Providing a querying language specific to their respective domain would flatten significantly the learning curve needed with more generic languages and leave more time to focus on the data analysis.

Among the main constraints which prevents the system development to advance more rapidly is the performance of the retrieval of the data from underlying storages. Currently deployed storage systems provide restrictive APIs and limited access to the infrastructure, preventing the execution of optimizations which would allow the analysis framework to satisfy many of the aforementioned requirements. While the Post Mortem system is partially maintained by the Machine Protection and Electrical quality assurance group, the Accelerator Logging Service is maintained by the controls group of CERN in conjunction with the IT teams. In a recent collaboration agreement the details of possible extension of the CALS service were discussed in order to improve the access performance to the stored information.

ISBN 978-3-95450-148-9

Copyright © 2015 CC-BY-3.0 and by the respective authors

MIXED PARTITIONING SCHEME REPLICATION

After identifying the problems of the current solution and defining requirements for the proposed new analysis framework, a detailed study of the currently existing solutions for analysis of a large data sets was conducted. An important factor which was taken into account during the study was the predominant time series nature of the data. The signals with respective variables are being stored in simple structures with corresponding measurement timestamps. Once collected and stored in a database, the signal value will never be updated, thus both CALS and PM are following "write once read many" paradigm. On the other hand, the identified requirements suggest that other data dimensions should be taken into account when optimizing the storage for heterogeneous workload (for example: device type, accelerator state the measurement is related to, etc). Finally we concluded that there are multiple tools and techniques which could improve the current situation and boost data read throughput. Multiple characteristics of the data combined with the orthogonal use cases will however quickly bring forward the limitations of the current infrastructure. To overcome these predicted limitations, we propose a new strategy for data storage, named mixed partitioning scheme replication.

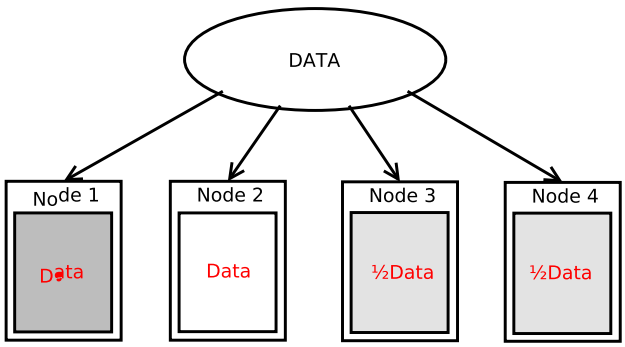

Figure 2: Basic hypothetical mixed partitioning scheme replication. The same data is partitioned differently on infrastructure nodes.

The infrastructure which is taken as an example and presented in Figure 2 consists of four nodes, with a replication factor of three (which is the recommended configuration in HDFS [7]). The number of nodes is chosen specifically to demonstrate how the data is organized when there are additional available resources. The replicas which are connected to infrastructure organize the data using different criteria. The color inside the box corresponds to different partitioning strategies. In case of Node 3 and Node 4, the data is organized using the same strategy, but is shared amongst the available resources to distribute the load evenly. Since the proposed technique is workload based, the decision to share the data among the available resources depends on the load of the participating replicas (additional resources will reduce the workload on the most busy nodes). It is important to note, that once the strategy is chosen and the data is stored, the schema does not evolve with time.

The main advantage of the proposed solution is the capability of handling efficiently highly heterogeneous workloads since different partitioning strategies can be adapted independently for particular query categories. On the other hand, the presented approach has a significant drawback related with node load balancing dependency on determined time periods. For example, during the accelerator commissioning phase, it is very unlikely that there will be some long term LHC performance analysis executed, while during the shutdown phase the probability of serving such queries is much higher in comparison to equipment test specific workloads.

To solve the load balancing problem, we propose a slightly different approach, based on the same principles of data replication using different partitioning techniques. In the Figure 3 the scenario remains the same, but in this case partitioning is done in two phases. An initial data division is being performed according to multi-dimension criteria which depends on data and workload characteristics adapted to the existing infrastructure configuration.

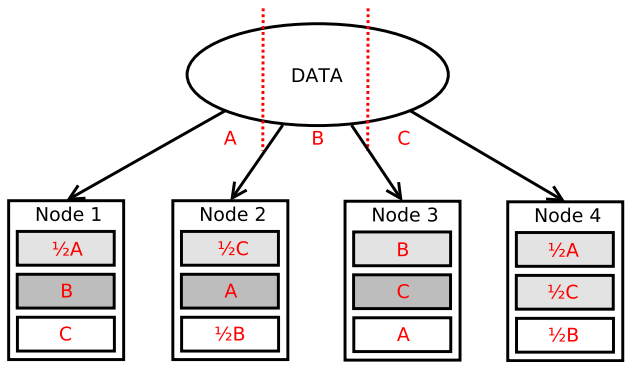

Figure 3: Complex hypothetical mixed partitioning scheme replication. The data is partitioned in two phases.

After the initial partitioning, data segments are being stored within the replicas using different partitioning strategies. Similarly to simple mixed partitioning scheme replication, the data organization is workload-driven and the busiest data segments are the first to be shared with additional resources. The main advantage of the proposed solution is the possibility of finding the ideal combination of storage strategies to optimize the load distribution throughout the whole infrastructure. On the other hand, there are several shortcomings as well with this approach. First of all, we foresee that strong data consistency mechanisms which are imposed by analysis types could impose some limitations on the system's scalability. Additionally, the complexity of the solution might raise maintainability challenges for the underlying infrastructure administrators especially in case of failure recovery mechanisms since the recovery process will require the translation of data from one scheme to another. All aforementioned factors will

Copyright © 2015 CC-BY-3.0 and by the respective authors

ISBN 978-3-95450-148-9

require detailed analysis to understand if the gained performance is enough to outweigh the downsides of the approach.

The proposed solution could be initially deployed on an intermediate storage component enhanced with the functionalities of modern engines for large-scale data processing (Spark [8] or Impala [9] for example). In addition to the possibility of working with data through eDSL, users will have the possibility to execute arbitrary requests on the data using the data processing engines native query language. The choice of data serialization and storage solution will depend on its compatibility with the chosen data processing engine and its performance with heterogeneous workloads. Initially the data will be fetched from CALS and PM. Before being available for users, the data will be pre-processed, correlated and merged, to abstract the data source where it was originally fetched from. A possible integration scenario of a mixed partitioning scheme replication with new technologies and the proposed infrastructure is depicted in the Figure 4.

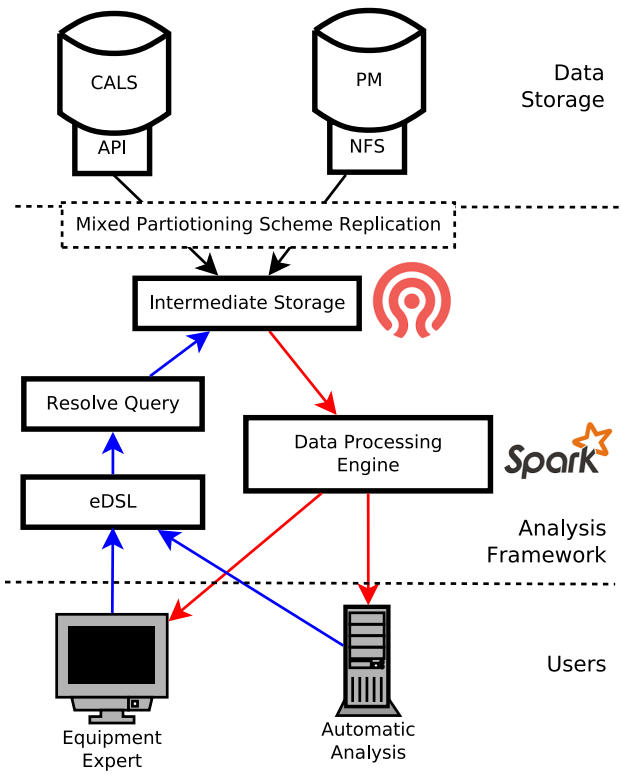

Figure 4: Mixed partitioning scheme replication integration into current infrastructure. In portrayed scenario, the data storage solution and data processing engine chosen are CEPH [10] and Spark, respectively.

FUTURE WORK

In the near future our efforts will focus on building a model approximating the real algorithm implementation which will allow to achieve early feedback about the benefits and eventual shortcomings of the proposed approach. Besides determining the efficiency of the proposed solution, the model would allow to understand how different parameters could affect the final system behaviour for different workloads. The plan is to start with a simple one node simulation, extend it to multi-node environments and finally validate the predicted performance by benchmarking the currently deployed infrastructure. In the end, the precision of the model will directly influence the request routing in the mixed partitioning scheme replication, since the algorithm should be able to determine efficiently the best resource to execute the query.

CONCLUSION

In order to address the new requirements and enhancement of the LHC for its second and subsequent operational running periods, the accelerator transient data analysis framework requires a, more performant data storage solution. Studies of the currently deployed analysis framework and literature suggest that the currently operating solutions will only be able to cover partially the required features of the data storage system for a highly performing analysis framework. A novel solution is being proposed to address both the desired characteristics as well as enhancing the storage performance.

REFERENCES

[1] K. Fuchsberger et al.,"Concept and Prototype for a Distributed Analysis Framework for LHC Machine Data", ICALEPCS'2013, San Francisco, CA, USA (2013).

[2] D. Anderson et al.,"The AccTesting Framework: An Extensible Framework for Accelerator Commissioning and Systematic Testing", ICALEPCS'2013, San Francisco, CA, USA (2013).

[3] O. Andreassen et al.,"The LHC Post Mortem Analysis Framework", ICALEPCS'2009, Kobe, Japan (2009).

[4] C. Roderick et al., "The CERN Accelerator Measurement Database: On The Road To Federation", ICALEPCS'2011, Grenoble, France (2011).

[5] C. Aguilera-Padilla et al., "Smooth migration of CERN post mortem service to a horizontally scalable service", WEPGF047, these proceedings, ICALEPCS'2015, Melbourne, Australia (2015).

[6] M. Audrain et al., "Using a Java Embedded Domain-Specific Language for LHC Test Analysis", ICALEPCS'2013, San Francisco, CA, USA (2013).

[7] https://hadoop.apache.org/

[8] http://spark.apache.org/

[9] http://impala.io/

[10] http://ceph.com/

Copyright © 2015 CC-BY-3.0 and by the respective authors

ISBN 978-3-95450-148-9