Big data analytics for the Future Circular Collider reliability and availability studies

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 072005 doi:10.1088/1742-6596/898/7/072005

Big data analytics for the Future Circular Collider reliability and availability studies

Volodimir Begy[1], Andrea Apollonio[2], Johannes Gutleber[2], Manuel Martin-Marquez[2], Arto Niemi[3], Jussi-Pekka Penttinen[4], Elena Rogova[5], Antonio Romero-Marin[2] and Peter Sollander[2]

[1] Faculty of Computer Science, University of Vienna, Waehringerstr. 29, Vienna, Austria
[2] CERN, Geneva 23, Switzerland
[3] Tampere University of Technology, Korkeakoulunkatu 10, Tampere, Finland
[4] Ramentor Oy, Hermiankatu 8, Tampere, Finland
[5] Delft University of Technology, Mekelweg 2, Delft, Netherlands

E-mail: volodimir.begy@univie.ac.at

Abstract. Responding to the European Strategy for Particle Physics update 2013, the Future Circular Collider study explores scenarios of circular frontier colliders for the post-LHC era. One branch of the study assesses industrial approaches to model and simulate the reliability and availability of the entire particle collider complex based on the continuous monitoring of CERN's accelerator complex operation. The modelling is based on an in-depth study of the CERN injector chain and LHC, and is carried out as a cooperative effort with the HL-LHC project. The work so far has revealed that a major challenge is obtaining accelerator monitoring and operational data with sufficient quality, to automate the data quality annotation and calculation of reliability distribution functions for systems, subsystems and components where needed. A flexible data management and analytics environment that permits integrating the heterogeneous data sources, the domain-specific data quality management algorithms and the reliability modelling and simulation suite is a key enabler to complete this accelerator operation study. This paper describes the Big Data infrastructure and analytics ecosystem that has been put in operation at CERN, serving as the foundation on which reliability and availability analysis and simulations can be built. This contribution focuses on data infrastructure and data management aspects and presents case studies chosen for its validation.

1. Introduction

1.1. Future Circular Collider Study

The Future Circular Collider Study hosted by CERN is a world-wide study federating a collaboration of more than 90 organizations from 30 nations. The goal is to conceive a set of conceptual designs for frontier circular colliders for the post-LHC era [1]. The study spans physics and phenomenology, detectors and experiments, lepton and hadron colliders, their infrastructures, implementation scenarios and cost estimates. The initiative was launched in 2014 as a direct response to a specific request stated in the 2013 Update of the European Strategy for Particle Physics [2]: "CERN should undertake design studies for accelerator projects in a global context, with emphasis on proton-proton and electron-positron high-energy frontier machines. These design studies should be coupled to a vigorous accelerator R&D programme, including high-field magnets and high-gradient accelerating structures, in collaboration with

national institutes, laboratories and universities worldwide" [3]. The main deliverable of the study is a multi-volume Conceptual Design Report in time for the next strategy update. The study is partially implemented as a Horizon 2020 project (EuroCirCol GA) and federates various national R&D programs including common work with the HL-LHC project in several domains.

1.2. RAMS Modelling and Simulation

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator [4]. It has been operating since 2007 and it took 8 years to optimize its operation at an energy of 6.5 TeV, close to its design performance [5]. The integrated luminosity, as a key performance indicator for particle colliders depends directly on the availability of the accelerator complex. Ongoing research at CERN takes the LHC and its injectors as a case study to assess the potentials of industrial reliability engineering methods for the domain of scientific, industrial and medical particle accelerators. The goal of the reliability and availability study is to come to proposed conceptual designs for future circular colliders that are actually feasible from an operation point of view [6]. The reliability analysis requires good understanding of the operation, failure root causes, repair and maintenance data of the entire accelerator complex, its sub-systems, key infrastructure elements, proposed and workable schedules as well as the tight and puzzling interaction with the high-energy beam. The analysis is based on continuous modelling and simulation of subsystem cycle and operation phase-dependent failure distributions as input to provide predictive models with high fidelity as output.

1.3. Challenges

Today such reliability analysis requires significant manual effort to extract, prepare and analyze the operational and maintenance data from a large number of diverse subsystems, e.g. logging and monitoring services. Eventually, only a limited amount of the available data is suitable for the reliability and availability analysis since its quality and integrity are highly heterogeneous.

A computer science research approach should assist the integration of heterogeneous structured, semi-structured and unstructured data from different sources (industrial control systems, relational databases, file storages, ...) and semi-automated data quality annotation.

2. The Approach

2.1. Reliability and Availability Data Analytics Platform

A single uniform reliability and availability data analytics platform should address the mentioned challenges. The proposed platform encompasses characteristics of a so-called Data Lake, which is capable of hosting various types of data, organized according to various paradigms. It should permit to gradually improve the data maturity by the means of human knowledge incorporation in an automated fashion, using both open and closed loop approaches, depending on the nature of the data and the potentials for the process automation. This approach requires enriching the available raw data with metadata. The raw data should never be altered. In order to correlate different data sources, the platform enables to extract homogeneous views "on the fly". The resulting analytics ecosystem should process the monitoring data of the accelerator complex in an automated or human assisted fashion, thus allowing a continuous reliability & availability analysis of the machine complex. Achieving this goal, even only partially, would significantly contribute to the sustainability of today's and future accelerator complex already before a post-LHC machine would be constructed.

Figure 1 displays the initial architecture of the platform. The bottom part depicts an extensible layer for data ingestion from different systems. The heterogeneous data is transmitted in its original form to a horizontally scalable and configurable semantic storage tier, which includes a metadata store and interfaces for the annotation of the data quality. The analytics tier presents a query interface to the stored data based on a schema-on-read principle, provides

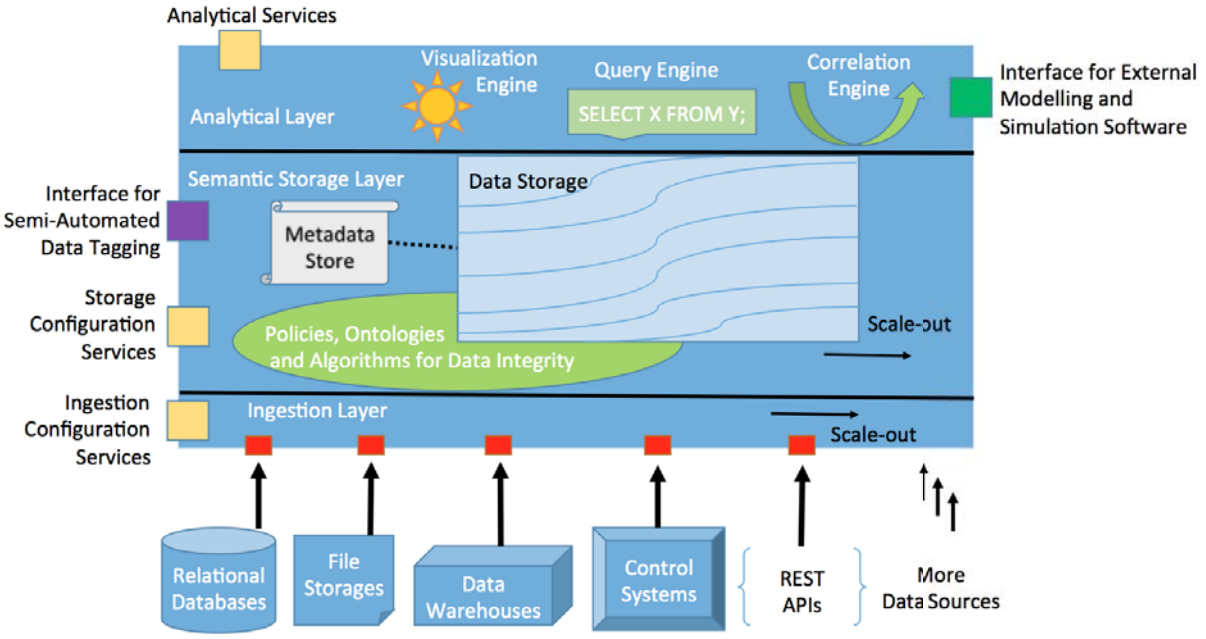

Figure 1. Initial Architecture of Reliability and Availability Data Analytics Platform

a bag of tools to visualize and correlate query results and export data to external reliability simulation and analysis software.

The depicted platform is currently being developed as a joint effort of the FCC reliability and availability study members and the CERN IT department.

2.2. Big Data Infrastructure

A so-called Big Data infrastructure based on the Apache Hadoop ecosystem deployed at CERN lies at the core of this R&D project. The Hadoop cluster is a suitable foundation for numerous reasons: the "shared nothing" architecture allows to scale the throughput of the analytics workflows, since the data are accessed and processed locally [7]. The Hadoop Distributed File System (HDFS) serves as a flexible storage layer. Furthermore, Apache has a large developer community and thus provides numerous software frameworks related to Hadoop, helping to construct the platform. This section describes the individual components of the Big Data infrastructure from hard- and software perspectives.

The cluster dedicated for the Hadoop ecosystem consists of 14 nodes, each having 64 GB of memory (total of 896 GB). Each node has 32 CPU cores (Ivy Bridge-EP 2,6 GHz) (total of 448 CPUs). There is 48 x 4 TB disk space per node. The total storage capacity of the cluster is 2,69 PB.

The deployed stack of software frameworks, which are used within analytics workflows is depicted in Figure 2. For data ingestion Apache Flume and Sqoop are used. The storage layer deploys HDFS, Apache HBase and Hive. Apache Oozie is used to schedule certain parts of the workflows. Impala is used as the SQL engine and Apache Spark is the tool of choice for large scale data processing and analytics. These frameworks provide the core functionality.

In addition, the platform still requires several custom software components. We develop them on top of the de facto standard frameworks. An example of such component is CERN HLoader. Various persistency systems for recording of the time-series data originating from control and monitoring systems have been deployed at CERN prior to the Hadoop infrastructure.

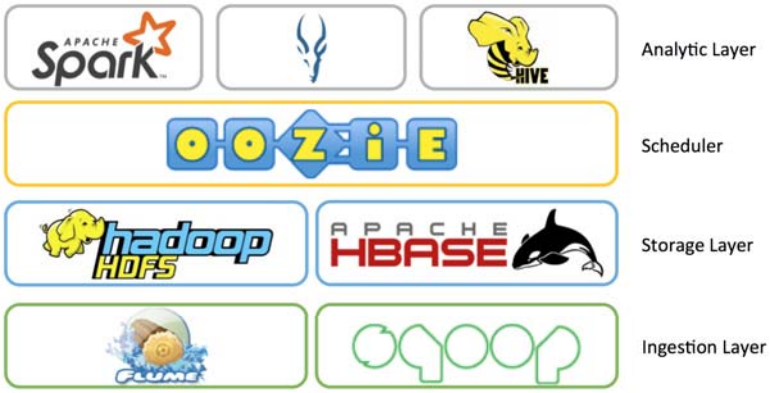

Figure 2. The Software Frameworks Stack

Thus, currently most of the raw data for the analysis needs to be ingested into HDFS first. HLoader is a tool for data ingestion from relational databases in a user friendly manner, with no deep computer science knowledge required, which is useful for instance for reliability engineers. Built around Apache Sqoop and Oozie, HLoader provides a REST API, which exposes metadata about available source and target storage systems to authorized users. The framework provides functionality to submit one-off or re-occurring Sqoop jobs using Oozie workflow or coordinator apps respectively. Apache Oozie supports event or data based conditions for triggering various actions (e.g. Spark, Hive and others). The project's source code can be found at `https://github.com/cerndb/hloader`. Figure 3 represents the architecture of the CERN HLoader framework.

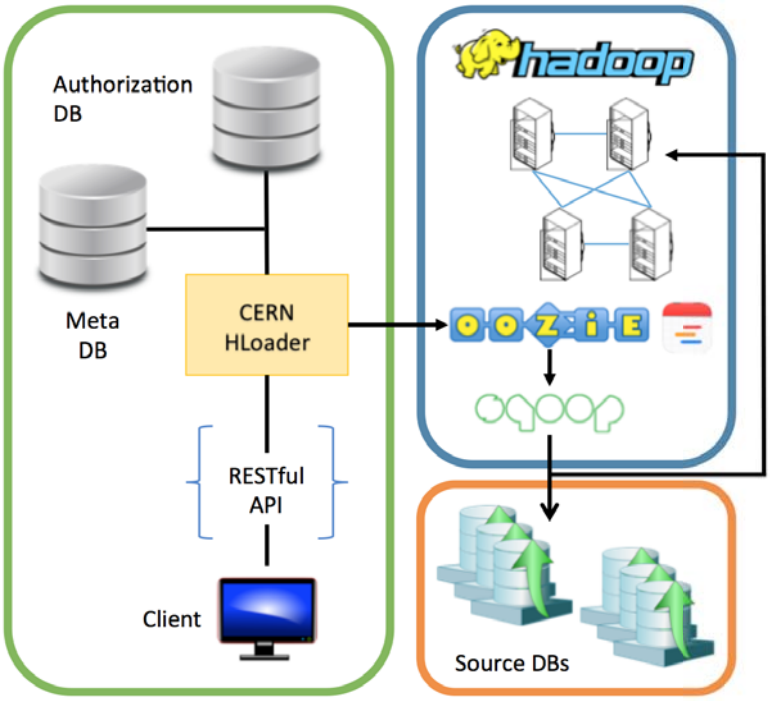

Figure 3. CERN HLoader Architecture

3. Selected Use-Cases

This section presents a non-exhaustive selection of use-cases, which are currently being implemented to validate the approach and to research methods and tools to integrate the data, to test machine learning approaches for data quality annotation and to find suitable approaches for querying diverse data in a homogeneous way.

One case is dedicated to the Super Proton Synchrotron (SPS) beam quality. SPS is the direct injector to the LHC and is therefore a determining factor of LHC availability for physics-grade beam in the collider [8]. The goal of the case study is to specify Quality of Service (QoS) parameters for SPS beams and to derive the probability of different QoS levels (e.g. good, acceptable, failed) for the LHC. Gaining a deep understanding of the SPS beam quality impact factors would eventually permit to simulate different injection scenarios and their impacts on hypothetical LHC beam-for-physics availability scenarios. For this purpose, the data from the SPS Beam Quality Monitor [9] is currently analyzed, brought in relation to actual analogue beam diagnostics and machine operation monitoring data to devise a model for defining successful, potentially successful and failing injections.

The second case study examines the dependency of the accelerator complex on the French electricity grid. This research integrates electricity grid operational data with RAMS analysis (failure distributions, fault trees, ...) of the accelerator complex components, which are critical for the beam quality. The findings of this work will assist decision making for electricity supply optimization, leading to an increased reliability & availability of the concerned machines.

Acknowledgments

The authors would like to thank Dániel Stein and Anirudha Bose for their work on CERN HLoader, Zbigniew Baranowski for his expertise on persistency systems and Joeri R. Hermans for the review of the document.

References

[1] Future Circular Collider Study Mandate (FCC-GOV-PM-001)
[2] FCC Brochure
[3] 2013 URL `http://cds.cern.ch/record/1567258`
[4] O'Luanaigh C 2014 URL `http://cds.cern.ch/record/1998498`
[5] Wenninger J LHC operation in 2015 and prospects for the future
[6] Apollonio A First results from availability studies
[7] Baranowski Z, Canali L and Grancher E 2014 *Journal of Physics: Conference Series* vol 513 (IOP Publishing) p 042001
[8] Drosdal L 2015 *LHC Injection Beam Quality During LHC Run I* Ph.D. thesis Oslko U.
[9] Papotti G, Bohl T, Wehrle U and Follin F 2011 Longitudinal beam measurements at the lhc: the lhc beam quality monitor Tech. rep.