EPJ Web of Conferences **245**, 01028 (2020) https://doi.org/10.1051/epjconf/202024501028
CHEP 2019

DAQExpert - the service to increase CMS data-taking efficiency

Gilbert Badaro[8], *Ulf* Behrens[6], *James* Branson[2], *Philipp* Brummer[1,9], *Sergio* Cittolin[2], *Diego* Da Silva-Gomes[3,1], *Georgiana-Lavinia* Darlea[4], *Christian* Deldicque[1], *Marc* Dobson[1], *Nicolas* Doualot[3,1], *Jonathan Richard* Fulcher[1], *Dominique* Gigi[1], *Maciej* Gladki[1,*], *Frank* Glege[1], *Dejan* Golubovic[1], *Guillelmo* Gomez-Ceballos[4], *Jeroen* Hegeman[1], *Thomas Owen* James[1], *Wei* Li[6], *Audrius* Mecionis[3,7], *Frans* Meijers[1], *Emilio* Meschi[1], *Remigius K.* Mommsen[3], *Keyshav* Mor[1], *Srecko* Morovic[2], *Luciano* Orsini[1], *Ioannis* Papakrivopoulos[5,1], *Christoph* Paus[4], *Andrea* Petrucci[2], *Marco* Pieri[2], *Dinyar* Rabady[1], *Kolyo* Raychino[1], *Attila* Racz[1], *Alvaro* Rodriguez-Garcia[1], *Hannes* Sakulin[1], *Christoph* Schwick[1], *Dainius* Simelevicius[7,1], *Panagiotis* Soursos[1], *Andre* Stahl[6], *Mantas* Stankevicius[3,7], *Uthayanath* Suthakar[1], *Cristina* Vazquez-Velez[1], *Awais* Zahid[1], and *Petr* Zejdl[3,1]

[1]CERN, Geneva, Switzerland
[2]University of California, San Diego, San Diego, California, USA
[3]FNAL, Chicago, Illinois, USA
[4]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[5]Technical University of Athens, Athens, Greece
[6]Rice University, Houston, Texas, USA
[7]Vilnius University, Vilnius, Lithuania
[8]American University of Beirut, Beirut, Lebanon
[9]Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract. The Data Acquisition (DAQ) system of the Compact Muon Solenoid (CMS) experiment at the LHC is a complex system responsible for the data readout, event building and recording of accepted events. Its proper functioning plays a critical role in the data-taking efficiency of the CMS experiment. In order to ensure high availability and recover promptly in the event of hardware or software failure of the subsystems, an expert system, the DAQ Expert, has been developed. It aims at improving the data taking efficiency, reducing the human error in the operations and minimising the on-call expert demand. Introduced in the beginning of 2017, it assists the shift crew and the system experts in recovering from operational faults, streamlining the post mortem analysis and, at the end of Run 2, triggering fully automatic recovery without human intervention. DAQ Expert analyses the real-time monitoring data originating from the DAQ components and the high-level trigger updated every few seconds. It pinpoints data flow problems, and recovers them automatically or after given operator approval. We analyse the CMS downtime in the 2018 run focusing on what was improved with the introduction of automated recovery; present challenges and design of encoding the expert knowledge into automated recovery jobs. Furthermore, we demonstrate the web-based, ReactJS interfaces that ensure an effective cooperation between the human operators in the control room and the automated recovery system. We report on the operational experience with automated recovery.

*Corresponding author: e-mail: maciej.gladki@cern.ch

1 Introduction

The Compact Muon Solenoid (CMS)[1] Data Acquisition (DAQ) system is responsible for reading out the data from one of the two general purpose experiments at the Large Hadron Collider (LHC). The accelerator complex provides proton-proton bunch crossings at a rate of 40 MHz, and the average size of each collision event is 1-2 MB. A two level trigger is in place in order to select only the most interesting data for storage and further analysis. At the first level, a hardware trigger selects the events at a rate of 100 kHz. Full events are read out and built from all detector electronics yielding a throughput of 200 GB/s. At the second level, the High Level Trigger farm of 35 000 cores reduces the event rate to O(1 kHz).

Due to the complexity of the system, issues related to hardware, software and networking cannot be excluded. The proper functioning of all components of the system is required for reliable data taking, otherwise the dataflow may be stuck or degraded. In order to minimize the downtime of the system, various recovery procedures have been prepared by the system experts. The operator crew, rotating in the control room 24/7, supervises the data taking and follows recovery procedures if needed. Support from on-call experts is available around the clock.

There is a human factor involved in this scheme of operations. We have observed that operators may make mistakes under time pressure and will add latency to the total intervention time. During LHC Run-1 and LHC Run-2 various automation mechanisms were introduced into the system [2, 3].

2 DAQExpert

DAQExpert [4] is a service to identify and mitigate dataflow issues in order to improve the efficiency of the experiment. It provides guidance to operators and enables system experts to define the steps required to resolve operational issues in a timely manner. Additionally, it provides them with tools to perform post-mortem analysis. It constantly analyses the monitoring data from the Run Control system [5–8], and based on procedures defined by experts, finds the optimal way to recover when dataflow is stuck.

2.1 Scope

The datataking efficiency of CMS was 95.87% uptime in 2018. This is measured as a percentage of system uptime during total time of Stable Beams delivered by the LHC. There were 2184 hours of Stable Beams in total in 2018. Due to various issues with power supplies, other infrastructure, the LHC and the DAQ system 90 hours were not recorded. Forty-six hours of this downtime was assigned to DAQ related issues (93% subdetector problems, 7% central DAQ). DAQExpert aims to reduce this DAQ downtime, while the remaining 44 hours of downtime are outside of its influence.

2.2 Impact

The intervention time (a.k.a. mean time to repair, MTTR) is a key metric to measure the reliability of services, including the DAQ system. The intervention time consists of the reaction time of the operator and the recovery time itself, which is subject to constant change. There

Table 1. Distribution of reaction time over the years of operation and a measure of overal reduction calculated between the earliest (2016) and the latest available measurements (2018)

Percentile	Reaction time in 2016 [s]	Reaction time in 2017 [s]	Reaction time in 2018 [s]	Overall reduction [%]
95th	322	177	132	59
75th	100	78	41	59
50th	85	49	29	66
25th	46	23	21	54

are many factors influencing the recovery time, namely: subsystem development, improvements in the run control system and special running conditions. The MTTR is therefore an inappropriate metric to measure the impact of DAQExpert. The reaction time, on the other hand, depends only on the individual abilities and alertness of operators which is expected to remain stable over the period of investigation. Moreover, the reaction time constitutes a considerable part of the total intervention time. Therefore it is an adequate metric to measure the impact of DAQExpert.

DAQExpert aims to reduce the operator's reaction time and avoid human errors that are likely under time pressure. It was introduced gradually during Run-2. The service was first made available in the beginning of 2017, and has since been enhanced with user experience improvements and extended expert input. We have observed mean reaction time reducing from 101 seconds in 2016, to 65 seconds in 2017, and 47 seconds in 2018. This improvement can be attributed to the DAQExpert guidance.

2.3 The Human Factor

The human factor is introduced whenever the operator is allowed to make a final decision on the recovery procedure. The breakdown of the reaction time in Table 1 reveals a stable trend over the investigated period. Overall operator reaction time has been reduced between 2016 and 2018 by 59% for the fastest 95% and 75% of reactions, 66% for the fastest 50% of reactions and 54% for the fastest 25% of reactions. There is little improvement in 2018 in the fastest 25% of operator reactions, improving from 23 to 21 seconds. Figure 1 shows that vast majority of reaction times were in the range of 20-25 seconds with a significant positive tail. Additionally, it is clear that further reduction below 10 seconds is not feasible with humans operators involved in the process.

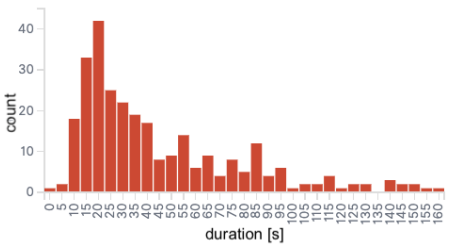

Figure 1. Reaction time histogram, based on 2016-2018 data

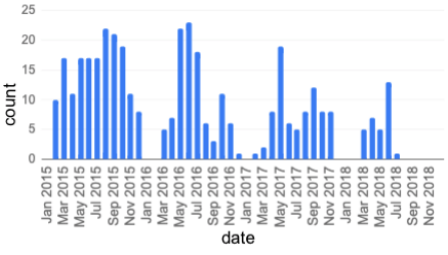

Figure 2. External help demand represented by the number of night-time calls to the on-call expert per month

The human reaction time accumulates to 4-6 hours of downtime per year, based on 2017 and 2018 data. Moreover, wrong decision overhead and improper usage of tools accumulates to another two hours of downtime per year, based on a detailed case-by-case analysis of operations data from August 2018.

The most impactful way to improve the DAQ operations is to bypass the operator wherever possible.

3 Automatic Recovery

Automatic recovery is a functionality of the DAQExpert system. It has been introduced in order to avoid the latency of human actions, and the risk of wrong decisions. It was introduced in the end of Run-2, and the first recoveries were successfully carried out without operator involvement.

3.1 Architecture

The DAQExpert adopts a microservices architecture and consists of multiple services (see Figure 3), to provide guidance, enable post mortem analysis and conduct automatic recoveries. The snapshot service collects all relevant data from the monitoring data sources and persists them for further use. The reasoning service consumes the monitoring data in real time, in order to identify potential data taking problems. It allows DAQ system experts to encode their domain knowledge in the form of Logic Modules [4, 9]. The notification service dispatches notifications to the system experts and operators. The latest addition, the controller, is in charge of performing the recovery procedure that so far was a responsibility of the operators. It sends the commands to the CMS Run Control system and keeps the operator crew updated with its actions. Web clients provide an interface for operators, with timely guidance (see Figure 4). Other user interfaces are dedicated to system experts and enable

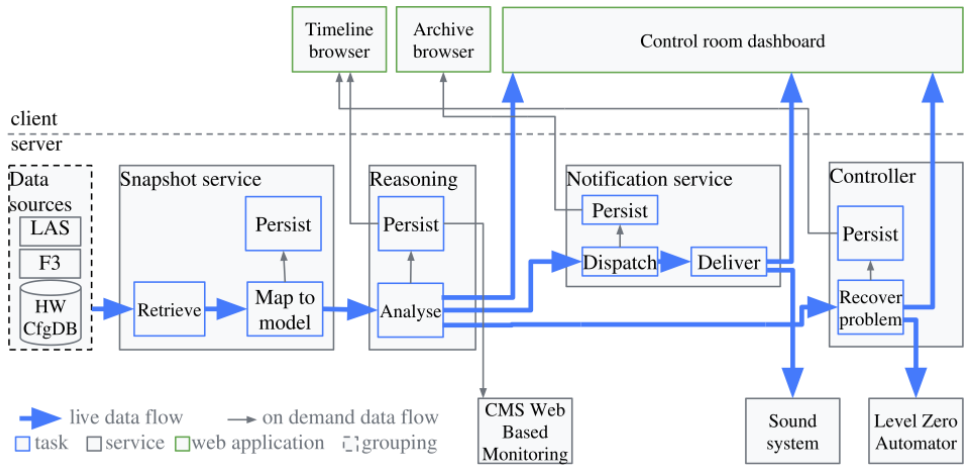

Figure 3. Architecture of the system. LAS (Live Access Service) monitoring component of the DAQ system, F3 (File-based Filter Farm) component of the DAQ system responsible for online event selection, HWCfgDB - DAQ physical and logical infrastructure database, Level Zero Automator - top level component to control the RCMS system.

post mortem analysis. The project adopts Web technologies: ReactJS [10] and Bootstrap [11] for building the reactive user interfaces; Java [12], Spring framework [13], websocket [14] and servlet [15] for building the backend services with RESTful APIs delivering live and on-demand data. Oracle database [16], Hibernate [17], JDBC [18] are used for persistence; and Apache Tomcat [19] for serving the services.

3.2 First Recovery

First recoveries driven entirely by DAQExpert without any operator involvement were tested at the end of Run-2. The detailed report of the automatic procedure, which follows the expert recommendations is shown in Figure 5. Although there is not enough data to determine the impact of this improvement, the possibility of eliminating human latency and errors have been demonstrated. The reaction time has been vastly reduced, now consisting of only monitoring delay and necessary sanity checks. Based on operational data since 2016 we estimate that this feature would reduce the downtime of CMS by around 8 hours per year. This corresponds to 17% of total DAQ related downtime and 9% of total CMS downtime.

4 Summary

The expert tool is evolving together with the DAQ system, and it is improving as we learn more about the operational challenges. Its coverage and features are being improved and we continue to observe a positive impact on data taking efficiency measured by the reduction in reaction time of operators during intervention, and the demand for expert help in the form of

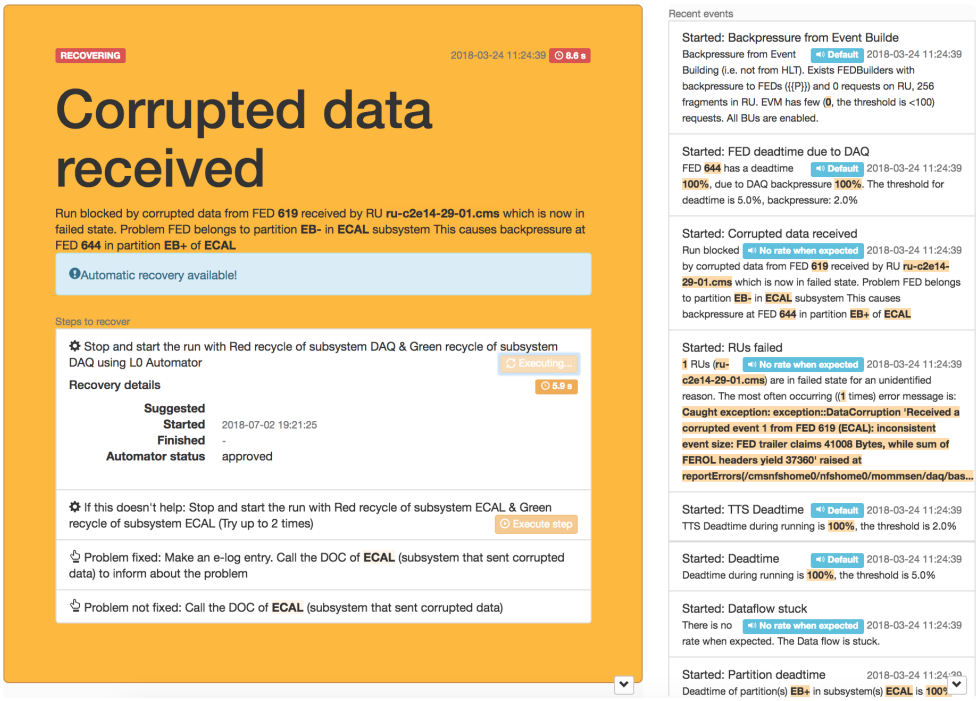

Figure 4. Dashboard for operators

night time calls to DAQ experts as shown in Figure 2. A considerable part of intervention time, the reaction time, is being reduced each year of operation. Beginning with 101 seconds on average in 2016 the DAQExpert helped to reduce it to 65 seconds in 2017 and 47 seconds in 2018.

Identifying that the human factor was limiting further improvements, efforts have been taken to bypass the operator in order to enable even quicker recoveries from data taking issues. The first automatic recoveries were observed at the end of Run-2 eliminating the overhead of human reaction latency, bringing the total reaction time down to the delay from the monitoring system, and the necessary sanity checks.

In 2018 the operator reaction time and overhead of wrong decisions accumulated to 8 hours of downtime and corresponded to 9% of the total CMS downtime. By introducing automatic recoveries we anticipate to significantly reduce this downtime in coming years.

The relevant metrics show that DAQExpert operations were successful in Run-2 and based on operational data from the last several years, a significant improvement in the CMS data taking efficiency is expected for Run-3.

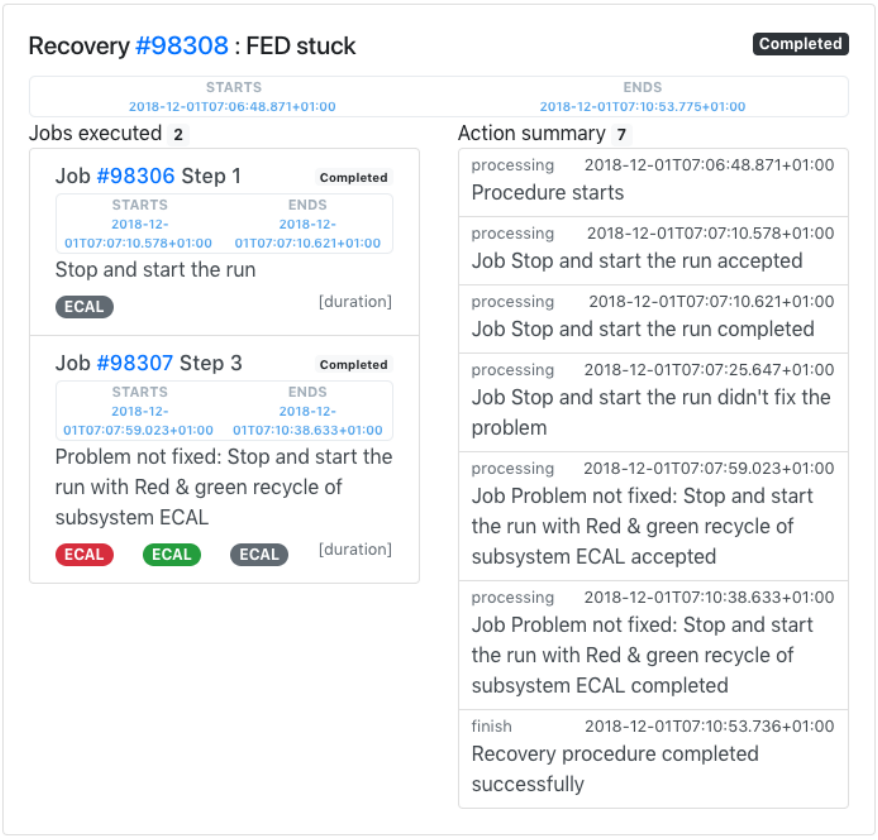

Figure 5. Automatic recovery summary of a datataking problem "FED stuck". FED (Front-End Driver) component of the DAQ system, Red/Green recycle - one of possible recovery actions, ECAL - the CMS electromagnetic calorimeter.

References

[1] CMS Collaboration, JINST 3 S08004 (2008)
[2] H. Sakulin et al., J. Phys.: Conf. Ser. **898**, 032028 (2017)
[3] H. Sakulin et al., J. Phys.: Conf. Ser. **513**, 012031 (2014)
[4] M. Gladki et al., J. Phys.: Conf. Ser. **1085**, 032021 (2018)
[5] M. Gulmini et al., eConf **C0303241**, THGT002 (2003)
[6] A. Oh et al., J. Phys. Conf. Ser. **119**, 022010 (2008)
[7] A. Petrucci et al., PoS ACAT **026**, (2007)
[8] H. Sakulin et al., IEEE Trans. Nucl. Sci. **59**, 1597-1604 (2012)
[9] H. Sakulin et al., EPJ Web Conf. **214**, 01015 (2019)
[10] ReactJS [software] https://reactjs.org/ [Accessed 2020-02-10]
[11] Bootstrap [software] https://getbootstrap.com/ [Accessed 2020-02-10]
[12] Java [software] https://www.java.com/ [Accessed 2020-02-10]
[13] Spring Framework [software] https://spring.io/ [Accessed 2020-02-10]
[14] Websocket [software] https://www.websocket.org/ [Accessed 2020-02-10]
[15] Servlet [software] https://www.oracle.com/technetwork/java/javaee/servlet/index.html [Accessed 2020-02-10]
[16] Oracle database [software] https://www.oracle.com/database/ [Accessed 2020-02-10]
[17] Hibernate [software] https://hibernate.org/ [Accessed 2020-02-10]
[18] JDBC [software] https://docs.oracle.com/javase/8/docs/technotes/guides/jdbc/ [Accessed 2020-02-10]
[19] Apache Tomcat [software] http://tomcat.apache.org/ [Accessed 2020-02-10]