PROCEEDINGS OF SCIENCE

The Data Quality Monitoring and Data Certification for the CMS experiment at the LHC

Amandeep Kaur Kalsi[*] **on behalf of the CMS Collaboration**
IIHE(ULB-VUB), Université Libre de Bruxelles, ULB, Brussels, Belgium
E-mail: amandeep.kaur.kalsi@cern.ch

The Data Quality Monitoring (DQM) team has a central role in CMS providing the needed tools to monitor data in real time and to scrutinize offline data to select the baseline set to perform analyses. This paper describes in some details the current status of the DQM software and procedure used in LHC Run2, and gives an overview of the major future plans for improvement.

European Physical Society Conference on High Energy Physics - EPS-HEP2019 -
10-17 July, 2019
Ghent, Belgium

[*] Speaker.

1. Introduction

The Compact Muon Solenoid (CMS) [1] is one of the general purpose detectors at the Large Hadron Collider (LHC), operating at CERN in Geneva, designed to detect rare signals of new physics. A sophisticated Data Quality Monitoring (DQM) system is employed for the CMS detector to ensure its good operation efficiency and the reliable certification of the recorded collision data for physics measurements and searches. The DQM group is responsible for the following items [2, 3]:

- tools for the creation, filling, delivery and archival of histograms, with standardised algorithms to carry out the quality and validity tests on the distributions automatically
- real-time monitoring of the detector components, trigger, Data Acquisition (DAQ) hardware status and their visualisation
- monitoring system for the offline reconstruction, validation of calibration results, software releases and monte carlo samples
- certification of collision runs for the physics analyses by Data Certification (DC) team
- standardisation and integration of DQM components in CMS software (CMSSW) releases
- organisation and operation of the DQM shifts, tutorials and other activities

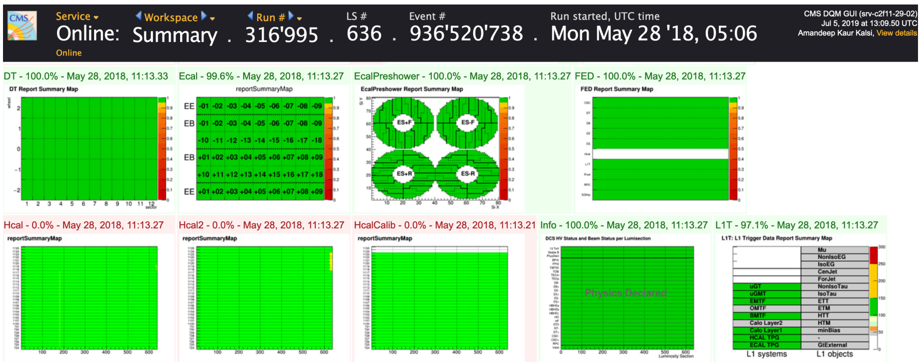
The DQM activities touch all aspects of data taking, from Online monitoring, Offline processing to the final certification of good data for analysis. It is a complex system making use of software tools and human component.

2. DQM framework

The central unit of monitoring in the DQM framework, known as Monitor Element (ME), is an object containing a ROOT histogram and the result of certain quality tests performed on demand. MEs are shared in the DQMStore, a general container class holding all the relevant plots for a single run, stored on the disk in ROOT format. Currently, run based DQM framework is employed which has two instances: the Online and the Offline DQM systems. The online DQM system is a part of High Level Trigger (HLT) [4] farm and uses the data stream produced by HLT for the production of real-time monitoring elements that provide immediate feedback about detector subsystem performance and integrity of data during data-taking. The online DQM runs over a fraction of the CMS data while offline DQM processes the full CMS event data. In general, the quality monitoring of the DQM system is a two-step process. The first step involves the creation of MEs followed by filling of information from the CMS event data. This step runs in parallel for different events in the same run. These histograms from different ROOT files of a particular run are summed over to obtain the full event statistics. These histograms are then uploaded to DQM GUI for visual inspection.

2.1 DQM GUI

DQM GUI is a web accessible software that has been developed to navigate the DQM file content and display histograms. It is a customizable application that is capable of providing histograms during live data taking for all subsystems as well as detailed information about the CMS event data. The content is organized in workspaces depending on the scope and ranges from the shift views to expert areas and high level summaries. Within a workspace, histograms can be organized into layouts to bundle related information together. A snapshot of DQM GUI is shown in Figure 1. It has three instances: Online, Offline and RelVal. The Online and Offline parts are used for the active monitoring and detailed information of CMS event data respectively, while RelVal part is used for the validation of CMSSW releases. It is flexible for the deployment of new software and usually a replica set up is maintained to test all the code before deployment to the production.


Service Online: Online
Workspace Summary
Run # 316'995
LS # 636
Event # 936'520'738
Run started, UTC time Mon May 28 '18, 05:06
CMS DQM GUI (srv-c2f11-29-02)
Jul 5, 2019 at 13:09.50 UTC
Amandeep Kaur Kalsi, View details

DT - 100.0% - May 28, 2018, 11:13.33
Ecal - 99.6% - May 28, 2018, 11:13.27
EcalPreshower - 100.0% - May 28, 2018, 11:13.27
FED - 100.0% - May 28, 2018, 11:13.27
Hcal - 0.0% - May 28, 2018, 11:13.27
Hcal2 - 0.0% - May 28, 2018, 11:13.27
HcalCalib - 0.0% - May 28, 2018, 11:13.21
Info - 100.0% - May 28, 2018, 11:13.27
L1T - 97.1% - May 28, 2018, 11:13.27

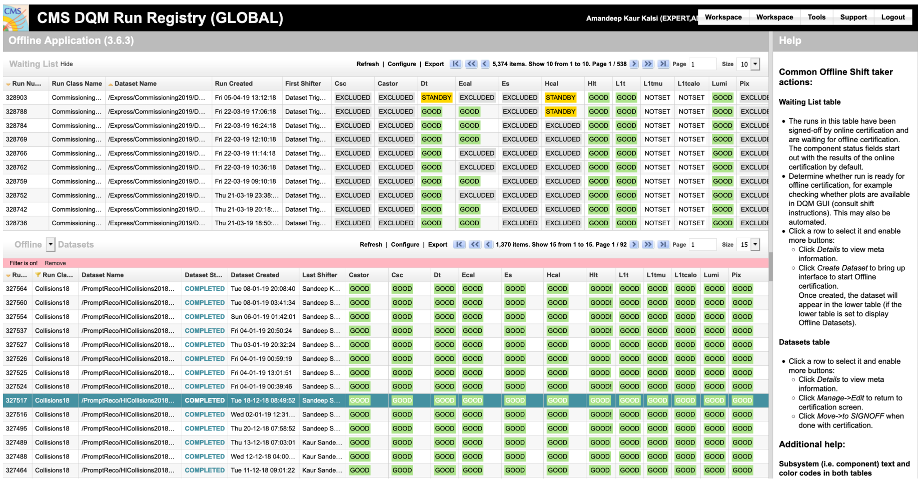
Figure 1: A snapshot of Online DQM GUI is shown. A layout of report summary maps showing the status of subsystems can be seen for a given lumisection and run. Here, each lumisection corresponds to the luminosity delivered by the LHC in ∼23 sec.

2.2 Run Registry

Run Registry (RR) is a tool used for bookkeeping of information about all CMS runs. RR consists of database-based web user interface frontend. It is a tracking tool used to certify the collected data and keeps tracks of all certification results. It provides the facilities for the manual input of data quality decisions and imports the data from other CMS web services such as CMS Web Based Monitoring (WBM), Online Monitoring System (OMS) *etc* regarding the detector and beam conditions, as well as the querying of the data and the export of selected information onto the screen and into flat output files of various formats. It has three instances: Online, Offline and User. The Online RR is used by the online shifters to add observations about the quality of run during data-taking while the offline part is used by sub-detector experts that inspects the quality of full CMS data once it is available in the offline server of DQM GUI. Several dedicated workspaces for various subsystems are provided. It provides generic API for the input/output and use of certification results through external tools. A snapshot of the offline RR is shown in Figure 2.

3. Operation, procedure and performance during Run2

The DQM system is in production since 2008 and has performed excellently during all years



CMS DQM Run Registry (GLOBAL)

Amandeep Kaur Kalsi (EXPERT,AI

Offline Application (3.6.3)

Waiting List Hide

Refresh | Configure | Export 5,374 items. Show 10 from 1 to 10. Page 1 / 538 Page 1 Size 10

Run Nu...	Run Class Name	Dataset Name	Run Created	First Shifter	Csc	Castor	Dt	Ecal	Es	Hcal	Hlt	L1t	L1tmu	L1tcalo	Lumi	Pix
328903	Commissioning...	/Express/Commissioning2019/D...	Fri 05-04-19 13:12:18	Dataset Trig...	EXCLUDED	EXCLUDED	STANDBY	EXCLUDED	EXCLUDED	STANDBY	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328788	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 17:06:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	STANDBY	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328784	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 16:24:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328769	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 12:10:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328766	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 11:14:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328762	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 10:36:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328759	Commissioning...	/Express/Commissioning2019/D...	Fri 22-03-19 09:10:18	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328752	Commissioning...	/Express/Commissioning2019/D...	Thu 21-03-19 23:38:...	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328742	Commissioning...	/Express/Commissioning2019/D...	Thu 21-03-19 20:18:...	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE
328736	Commissioning...	/Express/Commissioning2019/D...	Thu 21-03-19 18:50:...	Dataset Trig...	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED	EXCLUDED	GOOD	GOOD	NOTSET	NOTSET	GOOD	EXCLUDE

Offline Datasets

Refresh | Configure | Export 1,370 items. Show 15 from 1 to 15. Page 1 / 92 Page 1 Size 15

Filter is on! Remove

Ru...	Run Cla...	Dataset Name	Dataset St...	Dataset Created	Last Shifter	Castor	Csc	Dt	Ecal	Es	Hcal	Hlt	L1t	L1tmu	L1tcalo	Lumi	Pix
327564	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Tue 08-01-19 20:08:40	Sandeep K...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327560	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Tue 08-01-19 03:41:34	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327554	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Sun 06-01-19 01:42:01	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327537	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Fri 04-01-19 20:50:24	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327527	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Thu 03-01-19 20:32:24	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327526	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Fri 04-01-19 00:59:19	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327525	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Fri 04-01-19 13:01:51	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327524	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Fri 04-01-19 00:39:46	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327517	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Tue 18-12-18 08:49:52	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327516	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Wed 02-01-19 12:31...	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327495	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Thu 20-12-18 07:58:52	Sandeep S...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD!	GOOD	GOOD	GOOD	GOOD	GOOD
327489	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Thu 13-12-18 07:03:01	Kaur Sande...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327488	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Wed 12-12-18 04:00...	Kaur Sande...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
327464	Collisions18	/PromptReco/HICollisions2018...	COMPLETED	Tue 11-12-18 09:01:22	Kaur Sande...	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD

Help

Common Offline Shift taker actions:

Waiting List table

- The runs in this table have been signed-off by online certification and are waiting for offline certification. The component status fields start out with the results of the online certification by default.
- Determine whether run is ready for offline certification, for example checking whether plots are available in DQM GUI (consult shift instructions). This may also be automated.
- Click a row to select it and enable more buttons:
  - Click *Details* to view meta information.
  - Click *Create Dataset* to bring up interface to start Offline certification.
    Once created, the dataset will appear in the lower table (if the lower table is set to display Offline Datasets).

Datasets table

- Click a row to select it and enable more buttons:
  - Click *Details* to view meta information.
  - Click *Manage->Edit* to return to certification screen.
  - Click *Move->to SIGNOFF* when done with certification.

Additional help:

Subsystem (i.e. component) text and color codes in both tables

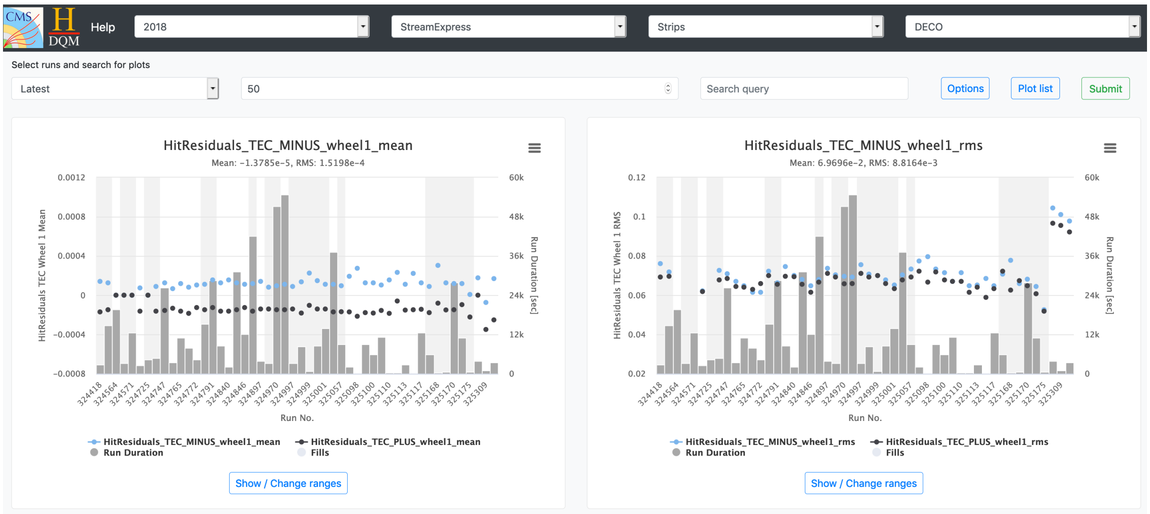
Figure 2: A snapshot of the Offline RR displaying status of various subsystem components for various runs. The green color for a particular column shows the boolean set by subsystem experts after the evaluation of CMS event data looking at the plots from the DQM GUI.

of CMS data taking. The response time of DQM GUI is less than 100 ms on an average, despite of the large number of accesses per day to both the online and the offline servers. The large number of modules running for a specific subsystem components comprise the so-called ‘DQM sequence’ and monitor both low level and high level variables at all the steps of the CMS event reconstruction. The outstanding performance of DQM system can be evaluated from the fact that during CMS Run2 data, the online DQM GUI consisted of around 22 k runs and 2×10^{9} MEs amounting to total of 650 GB of storage data while the offline DQM GUI contained about 400k runs and 4.4×10^{10} MEs accounting to $\sim$4.1 TB of storage data.

As mentioned before, the monitoring and certification of the quality of the CMS data is a multi-step procedure which spans from online data taking to the offline reprocessing of data recorded earlier. The quality assessment is based on both visual inspection of data distributions by monitoring shift persons as well as algorithmic tests of the distributions against references. The online shifter marks the runs in signoff status in online RR after its evaluation during data-taking. Offline DQM provides the so called ‘Express Stream’ for monitoring of data quality in terms of reconstruction, alignment and calibration constants with the latency of $\sim$1 hr. A team of ‘offline shifters’ and sub-detectors experts takes care of validating the data quality once the whole statistics has been processed. Sub-system experts detector uses a boolean flag in their respective workspaces for the final quality result. The central Data Certification (DC) team then uses this information from various subsystem workspaces and filters out the bad quality data to provide a file in JSON format containing certified good quality data for the physics analysis carried out in the CMS collaboration.

4. Future Plans

The DQM Framework performed well during Run2 period of LHC, yet some developments are required to adopt to the foreseen changes in the LHC conditions and CMS detector for the future LHC collisions. A new tool, known as “HISTORIC DQM (HDQM)” has been developed that provides the time evolution of the recorded data across the runs over a longer period of time. Such a technology is highly beneficial in spotting any change in detector efficiencies due to radiations and other anomalies during the course of time. A layout of HDQM webpage is shown in Figure 3. A re-designing of new Run Registry is in progress for better usability and maintainability. Machine Learning techniques are under development to support the Online data monitoring task and also to automatize the Data Certification procedure with the goal to save person power and improve efficiency. For the upcoming LHC Run3, a reorganization in the DQM configuration and core code is also ongoing.



2018

StreamExpress

Strips

DECO

Select runs and search for plots

Latest

50

Figure 3: A snapshot of HDQM page displaying the time evolution of observables showing the quality of data for a given CMS subdetector.

References

[1] CMS Collaboration, JINST **3**, S08004 (2008).

[2] L. Tuura, *et al.* [CMS Collaboration], J. Phys. Conf. Ser. **219**, 072020 (2010).

[3] F. De Guio [CMS Collaboration], J. Phys. Conf. Ser. **513**, 032024 (2014).

[4] CMS Collaboration, Int. J. Mod. Phys. Conf. Ser. **31**, 1460297 (2014)