Automated quality monitoring and validation of the CMS reconstruction software

Danilo Piparo
CERN, CH-1211 Genve 23, Switzerland
E-mail: danilo.piparo@cern.ch

Abstract. A crucial component of the CMS Software is the reconstruction, which translates the signals coming from the detector's readout electronics into concrete physics objects such as leptons, photons and jets. Given its relevance for all physics analyses, the behaviour and quality of the reconstruction code must be carefully monitored. In particular, the compatibility of its outputs between subsequent releases and the impact of the usage of new algorithms must be carefully assessed. The automated procedure adopted by CMS to accomplish this ambitious task and the innovative tools developed for that purpose are presented. The whole chain of steps is illustrated, starting from the application testing over large ensembles of datasets to emulate Tier-0, Tier-1 and Tier-2 environments, to the collection of the physical quantities in the form of several hundred thousand histograms, to the estimation of their compatibility between releases, to the final production and publication of reports characterised by an efficient representation of the information.

1. Introduction

The CMS Software (CMSSW) is a highly modular framework written in C++, which is characterised by a code base of more than two million lines and about 250 active developers constantly maintain and improve it. One of its crucial components is the reconstruction. It ensures that the electronic signals coming from all the sub-detectors are correctly transposed into energy depositions and then physics objects like photons, muons, jets, tracks and vertices. These objects are the ones then used for the data analysis. The feature sets of such a huge software project are grouped into release cycles and different cycles can coexist. The release schedule is tight, indeed:

- Pre-releases are provided once a week in order to consolidate the state of the code, test interdependencies among software components.
- Releases are cut approximately once per month. They close a development cycle, are used for central processing, Monte Carlo production and analysis.
- Amendment Releases are made available to solve specific issues and supersede the previous releases.

An automated and central validation of physics output of all releases is needed. The first step to achieve this ambitious goal is to provide for each release sufficiently large data and Monte Carlo samples for several processes, the so-called release validation datasets [1].

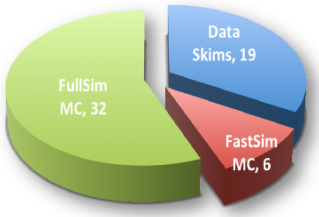

Figure 1. Composition of RelVal datasets. Several different types of events coming from real data, fully and fast simulated Monte Carlo samples are studied.

2. Production of release validation samples

As soon as any kind of CMSSW release is available, different types of samples are processed. The first group is represented by several categories of data subsets (also called "skims") obtained by selecting and grouping the events characterised by particular distinctive signatures. The second group consists in fully and fast simulated Monte Carlo events mimicking relevant processes like top quark and electroweak bosons production or important signatures like high p_T photons or leptons. This procedure guarantees in the first place that all components work together without failures during execution in a large scale environment. Moreover, this broad spectrum of signatures and event types gives the possibility not only to validate the physics objects and CMS sub-detectors but also to check the correctness of the alignment and calibration procedures.

Presently, to cover the quoted use cases, several processing workflows are submitted to a batch system (see figure 2):

- 19 different data skims.
- 32 full simulation Monte Carlo datasets.
- 6 fast simulation Monte Carlo datasets.

To deliver such an amount of data, dedicated resources are exploited at the FNAL farm (up to 1000 batch slots). This setup allows to provide all the necessary datasets within a 24 hours latency. For a given produced sample, the CMS data quality monitoring (DQM) infrastructure allows to collect the relevant quantities in a set of monitoring histograms (about 250.000) and save them in output ROOT [3] files [2]. This step represents an important data reduction with respect to the initial events n-tuples.

3. Data quality monitoring

Data quality monitoring is essential for the detector and operation efficiency. It allows to certify recorded data and produced Monte Carlo samples, allowing to spot possible problems or unwanted features. The CMS DQM provides tools for creation, transfer and archival of histograms and scalar monitor elements. In addition, visualisation of monitoring results, standardisation and integration into CMS software releases is guaranteed.

The histograms filled during the data processing are stored in ROOT files. At the end of the processing chains, these files are uploaded to the DQM servers and their content indexed in a database for performance and concurrent request handling. One interface to this database is a web GUI (see [4]) for the visualisation of the data quality monitoring. The state of the user session and application logic are held on the web server and users actions are mapped directly to http API calls.

The plots containing the monitoring histograms are created on demand with a special renderer which allows to perform quite sofisticated manipulations. For example, it can display a single

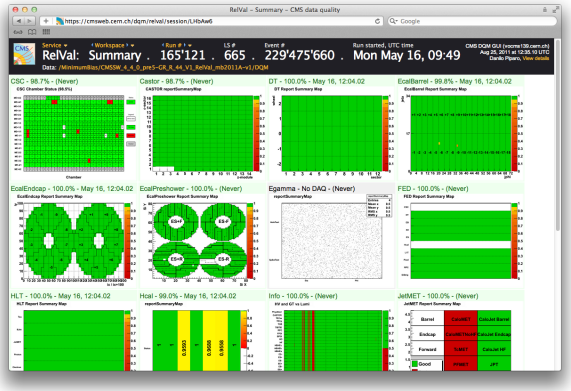

Figure 2. A snapshot of the CMS DQM web graphical user interface. The plots are created on the fly accessing the necessary information which is stored in an internal database, avoiding to rely on the ROOT I/O.

histogram or overlay multiple histograms, access basic ROOT draw options and manipulate plot axes settings.

CMS deploys four production DQM server instances. Three are used to monitor the data being recorded at CMS point 5, to check the reconstructed data coming from the Tier-0 and Tier-1s and to survey the data at the CMS CERN Analysis Facility. The fourth instance is exclusively dedicated to the release validation samples.

All CMS validation suites, for both physics objects and subdetectors, are based on the DQM technology

4. Relmon tool

RelMon is a tool to compare two sets of histograms, stored in ROOT files. Highly modular, it is written in Python and is interfaced to ROOT via the PyROOT bindings. RelMon is not dependent on any component of the CMS software.

The agreement between pairs of histograms is quantified with a statistical test. Predefined tests are available (Chi-square, Kolmogorov-Smirnov, Bin-by-Bin) and new user-defined compatibility criteria can be easily implemented. Pairs of corresponding histograms are selected by name. Moreover, the information about the outcome of the statistical tests is aggregated following the directory tree present in the ROOT files. Since the operations performed can be quite CPU intensive, the tool can exploit multiple cores to speed up the execution of the comparisons.

The outcome of the comparison of the two sets can be produced in the form of a minimalist ASCII output or an elegant browsable web report featuring effective representation of the information, intuitive diagrams such as gauges, pie charts and bar charts and images of histograms overlays to inspect the details of the comparisons (see figures 3 and 4).

RelMon provides also CMS specific features. For example an interface to fetch the histograms directly from the DQM server over an authenticated connection instead of ROOT files, a command-line tool to compare all the DQM output ROOT files coming from release validation productions or two given CMSSW releases (figure 4) and interface to DQM rendering to visualise large amounts of histograms comparisons.

RelMon allows to easily compare large sets of histograms according to a predefined criterion and produce convenient reports to spot all the differences. Moreover it is suited to be deployed in order to setup a central validation mechanism.

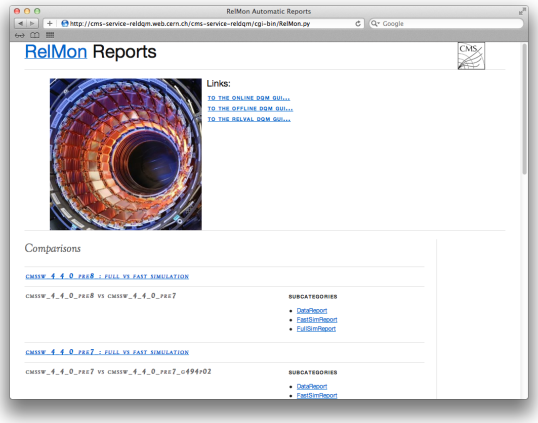

Figure 3. Central page where all the regression tests between CMSSW releases are stored.

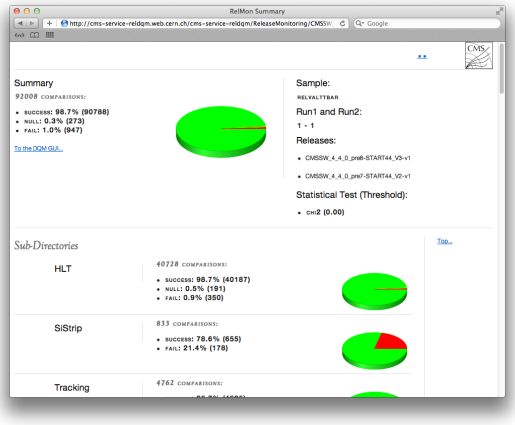

Figure 4. Piecharts and graphs are used in order to provide an overview of the global agreement between histogram sets.

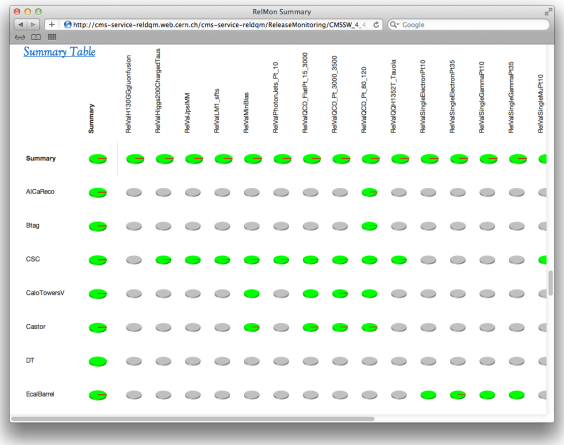

Figure 5. With the relmon reports, a complete overview about the Physics performance of a CMSSW release can be given. Large samples of histograms are compared, considering several datasets at the time.

5. Central validation

In CMS, the automatic generation of RelMon reports is triggered for every CMSSW release. Regression tests are performed across all data, full and fast simulation Monte Carlo DQM histogram samples coming from a given release and the preceding one, used as a reference. In addition, a comparison between full and fast simulation is performed. The latency between the arrival of the ROOT files containing the DQM histograms and the appearance of the RelMon report on the web is $\sim$ 30 minutes, even though more than one million histograms are compared.

This strategy exposes many potentially critical aspects of the physics output produced by the two releases: anomalies can be immediately pin-pointed. With this strategy, reconstruction coordinators gain broad overview of the overall physics performance of the release and the experts of the single physics objects and detector performance groups can immediately give their feedback about possible issues.

6. Conclusions

CMSSW is a huge software project, a fundamental part of which is represented by the reconstruction. The stability of the physics performance must be checked for each new release. In order to achieve this goal, CMS provides large groups of Data and Monte Carlo samples to perform release validation. The study of these samples, which aims to the quality monitoring and validation of the CMS reconstruction software, is performed exploiting the DQM infrastructure and RelMon. This strategy has proven to be very effective to spot anomalies and unwanted features within a very restricted timescale.

References

[1] The CMS Collaboration 2010 Validation of software releases for CMS *J. Phys. Conf. Ser.* **219** 042040

[2] The CMS Collaboration 2010 CMS data quality monitoring: systems and experiences *J. Phys. Conf. Ser.* **219** 082005

[3] R. Brun et al. 1997 ROOT - An Object Oriented Data Analysis Framework *Nucl. Inst. Meth. In Phys.* **A 389** pp 81-86

[4] The CMS Collaboration 2010 CMS data quality monitoring web service *J. Phys. Conf. Ser.* **219** 072055