Automated Quality Monitoring and Validation of the CMS Reconstruction Software

Danilo Piparo (CERN) - on behalf of the CMS collaboration

Introduction

The CMS Software (**CMSSW**) has **more than 2 million lines of code and about 250 active developers** constantly maintain and improve it. Its **reconstruction** component is crucial, since it ensures that the **electronic signals coming from the detector are correctly transposed into physics objects** - photons, muons, jets, tracks, vertices… - which are 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:**

- Pre-releases: consolidate the state of the code, test interdependencies among software components (1 per week)
- Releases: close a development cycle, used for central processing, Monte Carlo production and analysis (1 per month)
- Amendment Releases: solutions for specific issues, supersede the previous releases.

Automated and central validation of physics output between releases is needed

To achieve this ambitious goal, for each release, sufficiently large data and Monte Carlo samples for several processes must be provided.

Production of Release Validation Samples

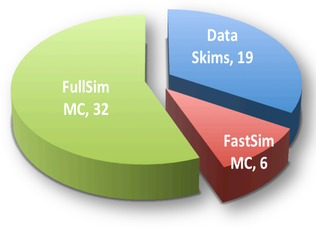
As soon as any kind of CMSSW release is available, different types of samples are processed:

- **Several categories of data skims** obtained selecting and grouping the events characterised by particular signatures
- **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:

- **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 Fermilab (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 data quality monitoring infrastructure allows to collect the relevant quantities in a set of monitoring histograms (about 250.000) and save them in output ROOT files. This step represents an important data reduction with respect to the initial events n-tuples.**

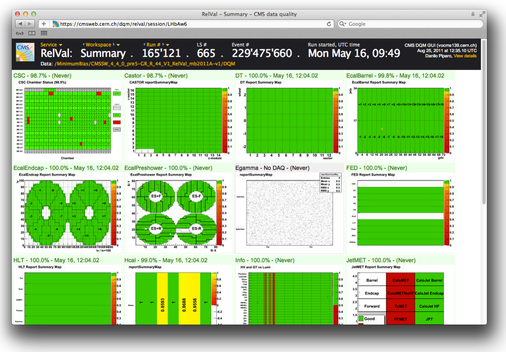
Many data and Monte Carlo samples are provided for each release for validation purposes

Data Quality Monitoring (DQM)

Data quality monitoring is essential for detector and operation efficiency. Furthermore it allows to certify recorded data and produced Monte Carlo samples.
In particular, the CMS DQM provides:

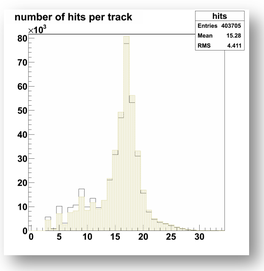
- Tools for creation, transfer and archival of histograms and scalar monitor elements
- Visualisation of monitoring results
- Standardisation and integration into CMS software releases.

The ROOT files containing histograms filled during the data processing are uploaded to the DQM server and their content indexed in an internal database for performance and concurrent requests handling.
A fundamental part of the system is a web GUI 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:

- Display single or overlay of multiple histograms
- Access basic ROOT draw options
- 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 tier-0 and tier-1s, 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

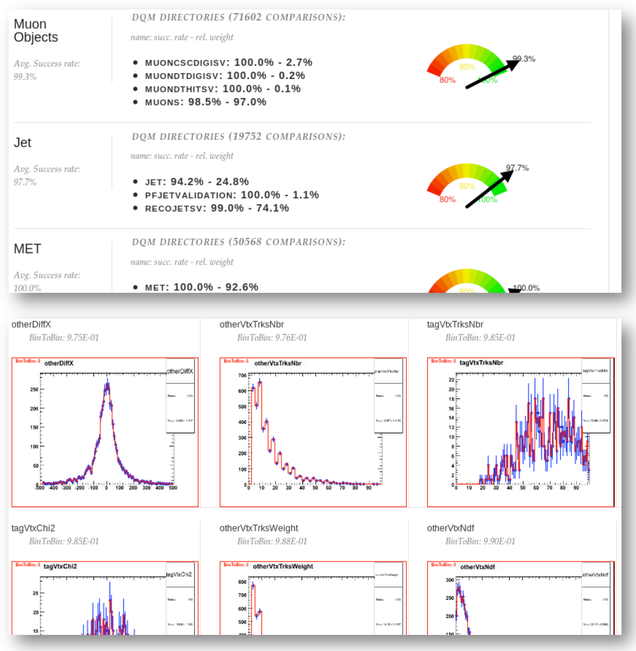
RelMon Tool

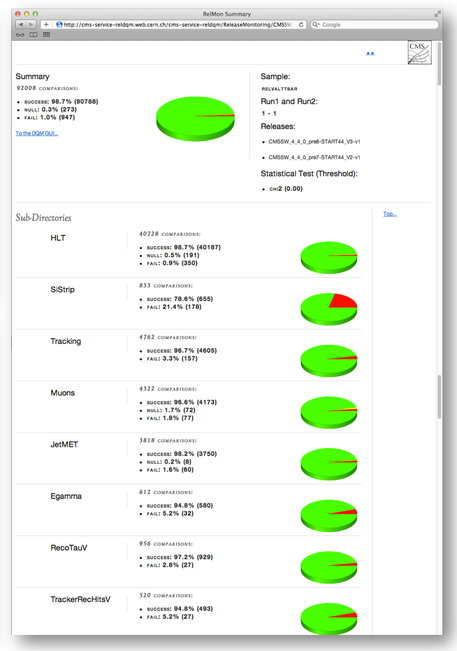
RelMon is a tool to compare two sets of histograms, stored in ROOT files. Highly modular, it is written in Python and is interfaced to ROOT via the PyROOT bindings.

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. The information about the outcome of the statistical tests is aggregated following the directory tree present in the ROOT files.
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: gauges, pie charts and bar charts
- Images of histograms overlays to inspect the details of the comparisons.





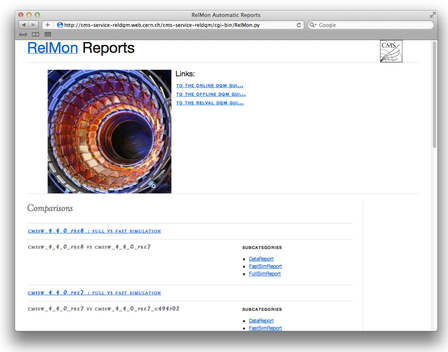
RelMon provides also CMS specific features:

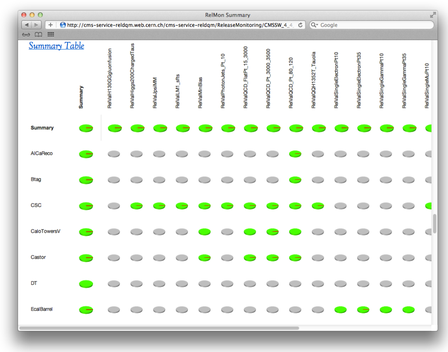
- An interface to fetch the histograms directly from the DQM server over an authenticated connection
- Command-line tool to compare all the DQM output ROOT files coming from release validation productions of two given CMSSW releases
- Interface to DQM rendering to visualise large amounts of histograms comparisons
- Creation of a summary page, displaying the status of the single CMS subsystem and an overview of the comparisons success rate of all subsystems for all RelVal samples.

RelMon allows to easily compare large histograms sets according to a predefined criterion and produce convenient reports to spot all the differences

Central Validation

The automatic generation of RelMon reports is triggered for every CMSSW release. The regression is 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 to that, 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 of about 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.
Reconstruction coordinators gain a broad overview about the overall physics performance of the release and the experts of the single Physics Objects Groups (POGs) and Detector Performance Groups (DPGs) can immediately give their feedback about possible issues.

RelMon, joint with the CMS DQM infrastructure, is a powerful tool for central automated validation of the CMSSW reconstruction code

References

- **Validation of software releases for CMS, 2010 J. Phys.: Conf. Ser. 219 042040.**
- **ROOT – An Object Oriented Data Analysis Framework, Nucl. Inst. & Meth. In Phys. Res. A 389.**
- **CMS data quality monitoring: Systems and experiences, 2010 J. Phys.: Conf. Ser. 219 082005.**
- **CMS data quality monitoring web service, 2010 J. Phys.: Conf. Ser. 219 072//055.**
- **RelMon: https://twiki.cern.ch/twiki/bin/view/CMSPublic/RelMon**

14th International Workshop on Advanced Computing and Analysis Techniques in Physics Research – Uxbridge, London UK, September 5-9, 2011