CMS data preparation for Run II

Nancy Marinelli

On behalf of the CMS Collaboration

Abstract

The LHC Run II will bring new challenges, mainly due to the higher number of interactions per beam crossing and the reduced time spacing between the crossings. In order to be ready for the beginning of the run, the CMS Collaboration is evolving the infrastructures, developed during Run I to monitor the data quality, to validate the progress on the detector simulation, event reconstruction, physics object definition and to handle large scale production of simulated data samples. This contribution covers the development and operational aspects put in place for Run I and describes how the experience gained is guiding the planning for Run II.

Keywords: Monitoring, data, validation

1. Introduction: from Run I to Run II

During the LHC Run I, the CMS experiment [1] collected data with varying conditions in the proton-proton center-of-mass energy (7 and 8 TeV), the instantaneous luminosity and the number of collisions per beam crossing (pileup) reaching an average of 21 during the run at 8 TeV. The success achieved, which culminated with the Higgs boson discovery [2, 3], was possible also thanks to a long and careful preparation with the design and the development of techniques, tools and procedures which insured that high quality data and simulated samples were available for physics analyses. The conditions foreseen for Run II will be by far more challenging than what already encountered. The centre-of-mass energy will double, the luminosity will further increase and with it the number of pileup interactions with the addition of the possible reduction of time spacing between bunch crossings from 50 ns to 25 ns. Key items among the many steps leading from the raw data collection to the final physics object hence the final physics results.

2. Calibration and alignment

The calibration procedure [4] is designed to take place at different times during and after the data taking, with increasing degree of precision. A “quasi-online” procedure, lasting minutes, is applied for running the High Level Trigger; a dedicated express data stream with 100 Hz sampling is used. The beam spot is measured within about 2 minutes using only track information or pixel-based vertexing. The following step is the “prompt” determination of calibration constants to be used within 48 hours from the data taking, for the prompt reconstruction of physics objects. Finally there are the “offline” procedures, which rely on calibration data streams with dedicated event selections, designed to optimize for bandwidth and storage space. The offline calibration take place well after the data taking and aims at providing the best understanding of the detector. The alignment takes fully into account the inter-dependencies between the calibration of each
sub-detector. The offline calibration and alignment procedure is performed and used for data reprocessing with ultimate accuracy.

3. Data quality monitoring

The CMS data quality monitoring (DQM) [5] is divided in two parts: the online and the offline DQM systems. The former is essential for monitoring the detector data and status so to discover at early stages hardware problems and insure high efficiency in the detector operation. The offline DQM instead is principally devoted to monitor low and high level quantities in reconstructed data and ensure their quality certification. It is also used for systematic validation of simulation and reconstruction software as well as of alignment and calibration conditions. A set of simulated and real data samples are produced each time a new software release and/or new calibration conditions are issued and DQM histograms produced (Section 5). The CMS DQM framework is fully embedded in the more general CMS software (CMSSW): it supports histogram booking, filling, handling and archiving and it is provided with a standardized interface for algorithms which perform automated quality tests. The results can be visualized on a web-based graphical user interface which guarantees authenticated worldwide access. The DQM system is operated since 2008 and it has evolved since, to accommodate the increasing needs of the CMS running during Run I. In order to cope with the increased CPU time needs expected for Run II, the DQM system has been upgraded to the new multi-core multi-thread processing. Furthermore new functionalities will be deployed, such as the possibility of comparing histograms from data and simulated samples to monitor constantly the data-to-simulation level of agreement, well before reaching the actual offline analysis level where this operation normally takes place.

4. Run-dependent simulation

Data taking in 2012 took place under rapidly changing conditions: on one side the increasing luminosity delivered by the LHC, hence the increase of the pileup events and on the other the change in the noise level and in response of the ECAL. With the increasing integrated luminosity, the dark current of the silicon avalanche photodiodes (APD) used in the barrel region increased as expected [1], leading to a higher noise level. Furthermore both ECAL barrel and endcap response varied accordingly to the variation in the light-yield from the lead-tungstate crystals resulting in reduced signal-to-noise ratio. The standard simulated samples produced at the beginning of the run were found to be far from reflecting these dynamic conditions and for a specific analysis, the search of the Higgs boson decaying to two photons [6], this was not acceptable. The first experience with time dependent simulation was then successfully made; different real data samples were grouped accordingly to the specific data taking conditions; sets of conditions, representative of a given data taking period were stored in a data base together with their interval of validity. Simulated samples were then produced by choosing randomly one set of conditions which were kept stable during each production job. This special procedure was put in place during Run I specifically to produce simulated samples of H → γγ samples which were

Figure 1: The DQM Graphical User Interface is web-based and guarantees authenticated worldwide access.
used to build the Higgs boson signal model used in the analysis. It was imperative to produce simulated samples with the best accuracy for each data taking period, in order to achieve the best mass resolution.

5. Physics validation and supporting tools

A continuous software release mechanism exists in CMS according to which a new release is made each time improved reconstruction algorithms or new releases of external simulation packages (e.g. GEANT4 [7]) or simulation techniques (fast simulation) become available. In addition, new software versions are released when alignment or calibration constants need updating as well as whenever compilers, system architecture or external packages need to be kept up-to-date. Before using a given release for massive production of simulated samples or reprocessing of data, systematic validation campaigns are centrally coordinated and synchronized with each new release cycle. A set of test samples, of both simulated and real data, is produced each time; experts from each sub-detector and physics object group (about hundred CMS collaborators overall) is required to verify that all most significative quantities (e.g. detector noise, track momentum resolution, ECAL energy resolution etc. etc.) are at their optimal level and help discovering and report in a timely manner, possible problems in the reconstruction software or in any of the other areas mentioned before. On average a validation campaign takes place bi-weekly.

The activity is supported by several tools which helps visualizing the results as well as collecting and tracking the reports from the different subsystems. The RelMon tool (Fig. 3) retrieves the histograms produced by the DQM system and compare with statistical tests, with given references. A summary is also provided of the rate at which failures appear in the comparison. Upon inspection of the RelMon pages, the responsibles of each subsystem can then file a report in the ValDb (Fig. 2), which is then used by the overall coordinators of the validation campaign to draw the conclusions.

The activity is supported by several tools which helps visualizing the results as well as collecting and tracking the reports from the different subsystems. The RelMon tool (Fig. 3) retrieves the histograms produced by

Figure 2: Example of a page from the validation database: each subsystem can file a report

Figure 3: Example of the RelMon tool: for each sub-system folders containing histograms produced with the DQM system, are visualised and are statistically compared.

References


