The Compact Muon Solenoid Experiment
Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

28 November 2014 (v2, 16 December 2014)

CMS Data preparation for Run II

Francesco Fabozzi on behalf of the CMS Collaboration

Abstract

The Run II of the Large Hadron Collider will confront us with new challenges, mainly due to the higher number of interactions per bunch crossing and the reduced time distance between bunches. In order to be ready for the beginning of the run, in view of an early discovery, the CMS collaboration is currently evolving the infrastructure established during Run I to monitor the data quality, to validate the progresses on detector simulation, event reconstruction, physics objects definition, and to handle large production of simulated events. This contribution covers the development and operational aspects of data preparation at CMS for Run II and describes how the experience gained from Run I is serving the planning of the physics program for Run II.

Presented at *IEEE-NSS-MIC-2014 2014 IEEE Nuclear Science Symposium and Medical Imaging Conference*

CMS Data Preparation for Run II

Francesco Fabozzi on behalf of the CMS Collaboration

Abstract—The Run II of the Large Hadron Collider will confront us with new challenges, mainly due to the higher number of interactions per bunch crossing and the reduced time distance between bunches. In order to be ready for the beginning of the run, in view of an early discovery, the CMS collaboration is currently evolving the infrastructure established during Run I to monitor the data quality, to validate the progresses on detector simulation, event reconstruction, physics objects definition, and to handle large scale production of simulated events. This contribution covers the development and operational aspects of data preparation at CMS for Run II and describes how the experience gained from Run I is serving the planning of the physics program for Run II.

I. INTRODUCTION

THE central feature of the Compact Muon Solenoid (CMS) apparatus [1] at the CERN Large Hadron Collider (LHC) [2] is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the superconducting solenoid are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass/scintillator sampling hadron calorimeter. Muons are measured in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid. Extensive forward calorimetry complements the coverage provided by the barrel and endcap detectors. The CMS experiment uses a two-tier trigger system consisting of the first-level (L1) trigger and High Level Trigger (HLT). The L1 trigger, which is comprised of custom electronics, reduces the readout rate from the bunch crossing frequency of approximately 20 MHz to below 100 kHz. The HLT is a software-based trigger system that makes use of information from all sub-detectors, to further decrease the event rate to about 400 Hz.

In Run I the LHC provided proton-proton collisions at a center of mass energy of 7 TeV in 2010-2011, and 8 TeV in 2012, achieving a peak instantaneous luminosity record of $7.7 \cdot 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$. The CMS apparatus recorded data equivalent to an overall integrated luminosity of about 27 fb^{-1} , with an efficiency greater than 90 %. The huge amount of collected data and the top-level performances of the detector, were fundamental for developing a broad physics program, where the discovery of a new particle consistent with the Higgs boson [3] represents the main success.

In 2015 a new physics run of the LHC will start. In Run II the machine will provide proton-proton collisions at the unprecedented center of mass energy of 13 TeV. The

instantaneous luminosity is expected to reach, and possibly exceed, the design luminosity of $10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, and the time spacing between two consecutive bunch crossing will go from 50 ns to 25 ns. The record machine parameters will give the possibility to search for evidence of new physics already with the first collected data. At the same time these conditions pose new challenges to the CMS experiment: for instance the in-time pileup (overlap of events due to multiple collisions in the same bunch crossing) is expected to increase from 21 (average value during Run I) to 40, and the reduced time between bunch crossing will give rise to not negligible out-of-time pileup (overlap of events coming from different bunch crossing). Thus new or improved reconstruction algorithms and methods for pileup rejection are being developed within the collaboration. New discoveries will require optimal detector performances, together with workflows, procedures and tools to assure the highest quality of the reconstructed data.

In the following we will review the key workflows and tools for data preparation in CMS, which were fundamental for the success of Run I, and the improvements deployed in preparation for Run II.

II. DATA PREPARATION OVERVIEW

In CMS, both collected and simulated collision events are processed by complex and inter-dependent workflows (Fig. 1) which translate the raw data (low-level format that is directly related to the digital output of the readout electronics) into a higher level format, which provides physics observables (e.g., positions and momenta), that are encapsulated in physics objects for easy and intuitive access by end-users (so that collections of reconstructed muons, electrons, jets, etc. are provided for each event).

Data are centrally organized in datasets, by grouping together events sharing common topologies, to facilitate processing by analysis jobs, but at the same time fulfilling constraints dictated by an efficient use of computing and storage resources.

In order to provide physics quantities computed with the highest accuracy, the reconstruction algorithms need reliable calibration and alignment constants for all the sub-systems of the CMS detector, which are derived from workflows applied to the data itself with different latencies.

Finally, monitoring procedures are applied throughout the steps in order to promptly identify problems in the data itself, in alignment and calibration constants, in reconstruction algorithms, or in the CMS software framework, thus assuring always the highest quality of the data.

Manuscript received November 30, 2014.

F. Fabozzi is with INFN Sezione di Napoli, Complesso Universitario di Monte Sant'Angelo, Via Cintia, I-80126, Napoli, Italy, and Università della Basilicata, Viale dell'Ateneo Lucano 10, I-85100, Potenza, Italy (telephone: +39 0971 206166, e-mail: francesco.fabozzi@na.infn.it).

Fig. 1. Stream of data from collisions or simulation in the CMS experiment. Rectangular boxes represent workflows which process the data to the final format suitable for physics analysis. All the workflows and data are continuously monitored throughout the stream.

III. ALIGNMENT AND CALIBRATION

In CMS three workflows, distinguished by the time needed for their completion, are employed to derive alignment and calibration constants, which are then stored and handled by Oracle databases [4].

A quasi-online workflow provides constants to be used in input to the HLT and to 0-latency (Express) data streams, with latency of the order of few minutes from data taking. For instance, a beam-spot measurement is provided about every 2 minutes using information from tracks and pixel-only vertexing.

A prompt calibration workflow processes data from the Express stream, that is available in less than two hours from data taking, to derive constants needed in input by the prompt reconstruction workflow, which provides reconstructed data in less than 48 hours from data taking. ECAL transparency corrections are an example of constants obtained from prompt calibration.

Dedicated calibration data streams, obtained with specific event selection and reduced event content in order to optimize computing resources, are finally processed by offline calibration workflows. They deliver calibration constants aiming to optimal accuracy. For instance, alignment inter-dependencies of CMS detector sub-systems are computed at this level. Dedicated reprocessing campaigns, using in input state-of-the-art offline alignment and calibration constants, are centrally organized in order to provide recommended datasets for physics analysis.

The above workflows will stay the same also for Run II, and work has been done to consolidate the software infrastructure.

Alignment and calibration workflows have been also employed for testing the CMS computing, software and analysis infrastructure in preparation for Run II (CMS Computing, Software and Analysis Challenge 2014), by simulating different alignment and calibration scenarios (corresponding to the expected knowledge of the detector after several phases of data taking) and evaluating the corresponding impact on physics analysis performances.

IV. DATA QUALITY MONITORING

The Data Quality Monitoring (DQM) framework is a widely used tool in CMS, which is fully integrated in the general CMS software infrastructure and provides tools for booking, filling, handling, and archiving of the histograms used to monitor relevant variables [5], [6]. It is employed for real-time detector data monitoring (online DQM), for reconstructed data monitoring (offline DQM), to define the official run list for physics analysis (data certification), and for validation of the releases of the official CMS software. From the operational point of view, DQM users (e.g., online shifters, detector experts, etc.) inspect histograms on a web-based graphical user interface (GUI) with authenticated worldwide access.

In preparation for Run II the DQM framework has been upgraded in several aspects, based on Run I experience, aiming to improve performances and increase robustness of the tool. For instance, all the DQM related software packages have been migrated to be multi-thread/multi-core compliant, alongside the evolution of the entire CMS software framework. The format of DQM output has been optimized to gain a sizable reduction of memory consumption. The robustness of the online DQM has been improved by increasing automatization of procedures (for instance automatic data discovering and automatic cleanup of old files have been implemented). New functionalities have been also introduced, such as the possibility to perform comparisons of histograms from data and simulated samples, in order to monitor the level of accuracy of detector simulation.

V. DATASETS DEFINITION

Physics events recorded by CMS are centrally organized in non-exclusive datasets based on HLT results, which group together events with similar properties [7]. Selection algorithms are centrally applied to the reconstructed primary datasets in order to provide sub-samples (skims) of reduced size, useful to perform analysis tasks of common interest.

During Run I, the “data parking” technique was introduced, in order to exploit at best the trigger capabilities, beyond the limitations imposed by the limited CPU resources for offline data processing. Indeed the full HLT bandwidth was exploited by writing on tape about 450 Hz of events recorded by CMS without streaming them to the prompt reconstruction workflow, so to avoid loading the infrastructure beyond its

capacity. The events were instead reconstructed after the end of Run I, profiting of the better availability of resources. Parked datasets were extremely useful to broaden the CMS physics programs to areas not covered by triggers available in prompt reconstruction datasets.

As in Run I, a large-scale production of simulated events at 13 TeV is planned also for Run II. It is expected that of the order of ten billions of events will be produced. Centralized Monte Carlo (MC) events production is a complex workflow, which requires actions and interplay among several roles: production contacts from physics groups making simulated samples requests, generator group conveners testing and approving the requests, and production operators at CMS computing sites submitting and monitoring production jobs. The official MC production in CMS will be managed by using McM [8], a web-based platform aggregating information from all of the above users, and taking care in an automatic way of many operations related to production.

VI. PHYSICS VALIDATION

In CMS, centrally coordinated validation campaigns are routinely performed in order to monitor and validate the continuous developments in multiple areas, such as simulation packages, reconstruction algorithms, detector conditions, software framework, system architectures, and compiler versions. Each campaign requires production of the order of a hundred samples of adequate statistics, both from data and simulation, and equipped with a large set of diagnostic plots for inspection via DQM GUI.

Each campaign involves about 100 collaborators. Validation datasets and histograms are distributed promptly to validators and experts from each detectors and physics groups for inspection. Reports from validators have to be collected and reviewed in order to declare the target of the campaign validated or failing. The frequency of validation campaign is quite high: regular software release validations are performed bi-weekly, plus there are many dedicated campaigns for specific changes or updates.

Web-based tools have been developed to ease all the procedures related to validation. The RelMon tool [9] performs automated comparison of histograms in the DQM GUI from target and reference releases, based on statistical compatibility tests. Histograms with statistically significant changes are highlighted for further inspections from validators. The ValDb tool [10] aggregates validation reports for each campaign and from each group, also providing an easy bookkeeping. The platform is interfaced to an email hyper-news system to broadcast the reports to all the involved collaborators. The validation outcome from each report is highlighted with visual flags, which are useful to get a quick overview of the validation status. Detailed reports can also be inspected to get further details.

CONCLUSIONS

Data preparation in CMS was one of the crucial elements in Run I to deploy high quality physics. The same results must be

assured already from the beginning of Run II, but in more challenging conditions, in view of an early discovery. Based on the successful experience of Run I, it will be extremely important to promptly derive reliable alignment and calibration constants, to perform a continuous monitoring of data and related workflows, and to set up careful dataset definitions. The use of web-based tools will ease management of official Monte Carlo production and of validation procedures.

REFERENCES

- [1] CMS Collaboration, “The CMS experiment at the CERN LHC,” JINST 3 S08004, 2008.
- [2] L. Evans and P. Bryant (editors), “LHC machine,” JINST 3 S08001, 2008.
- [3] S. Chatrchyan *et al.* (CMS Collaboration), “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” Phys. Lett. B, vol. 716, pp. 30-61, 2012.
- [4] F. De Guio on behalf of the CMS Collaboration, “CMS alignment and calibration workflows: lesson learned and future plans,” Presented at the 37th Conference on High Energy Physics (ICHEP2014), talk available at: <http://indico.ific.uv.es/indico/getFile.py/access?contribId=239&sessionId=32&resId=0&materialId=slides&confId=2025>.
- [5] F. De Guio on behalf of the CMS Collaboration, “The CMS data quality monitoring software: experience and future prospects,” J. Phys. Conf. Ser., vol. 513, 032024, 2014.
- [6] L. Tuura for the CMS Collaboration, “CMS data quality monitoring web service,” J. Phys. Conf. Ser., vol. 219, 072055, 2010.
- [7] G. Franzoni for the CMS Collaboration, “Dataset definition for CMS operation and physics analyses,” Presented at The 37th Conference on High Energy Physics (ICHEP2014), talk available at: <http://indico.ific.uv.es/indico/getFile.py/access?contribId=240&sessionId=32&resId=0&materialId=slides&confId=2025>.
- [8] J. R. Vlimant for the CMS Collaboration, “McM: the evolution of PREP. The CMS tool for Monte Carlo request management,” Presented at The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), talk available at: <http://indico.cern.ch/contributionDisplay.py?contribId=152&confId=214784>.
- [9] D. Piparo for the CMS Collaboration, “RelMon: a general approach to QA, validation and physics analysis through comparison of large sets of histograms,” J. Phys. Conf. Ser., vol. 396, 022011, 2012.
- [10] A. Norkus for the CMS Collaboration, “ValDb: an aggregation platform to collect reports on the validation of CMS software and calibrations,” Presented at The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), talk available at: <http://indico.cern.ch/contributionDisplay.py?contribId=149&confId=214784>.