The NanoAOD event data format in CMS

Marco Peruzzi[1], Giovanni Petrucciani[1] and Andrea Rizzi[2] for the CMS Collaboration

[1]CERN EP Department, CH-1211 Geneva 23, Switzerland
[2]University and INFN of Pisa, Italy

E-mail: Marco.Peruzzi@cern.ch

Abstract. NanoAOD is an event data format that has recently been commissioned by the CMS Collaboration. It only includes high level physics object information and is about 20 times more compact than the MiniAOD format. NanoAOD can be easily customised for development activities and supports automated data analysis workflows. The current status and perspectives of NanoAOD design and implementation are reviewed.

1. Introduction

The CMS Collaboration [1] has recently commissioned a new compact event data format, named NanoAOD, with the aim of serving the needs of a substantial fraction of its physics analyses with a per-event payload of about 1-2 kB. The MiniAOD format [2] - the most compact one, so far - is about 20 times larger, and currently supports a very large majority of the CMS analysis efforts. NanoAOD achieves such a strong data reduction by retaining only high level information on physics objects, such as jets and leptons, dropping their individual constituents, and reducing the precision of stored variables.

The design rationale of NanoAOD is based on previous experience from user ntuples and considerations on which minimal set of object information can support a large set of analysis efforts. Its initial prototyping was reported in ref. [3]. We present here a summary of NanoAOD features (section 2) and describe the latest developments of the project and the impact of its introduction on analysis workflows (sections 3-5). An outlook on future perspectives is finally given in section 6.

2. Design and technical implementation

The content of a NanoAOD file takes the form of a flat ROOT [4] TTree, where each entry corresponds to one event. The branch names are used to identify variables that belong to the same collection of high level objects, as exemplified in table 1. Links between two objects in different collections, as in the case of geometrically overlapping objects, are stored using the index of the second object in its collection as a data member attached to the first object. A detailed description of the TTree structure and the reasons that have led us to these design choices can be found in ref. [3].

The fraction of payload taken by each physics object collection is summarized in table 2. For each object, only a limited set of variables are retained and their numerical precision is individually tuned in light of the intrinsic resolution. In particular, NanoAOD does not

Table 1. Format of variables in the NanoAOD format.

Name	Type	Content
nMuon	integer	Number of muons in the event
Muon_pt[nMuon]	array(float)	Muon transverse momentum
Muon_jetIdx[nMuon]	array(integer)	Index of jet matched to muon
MET_pt	float	Missing transverse energy

Table 2. Fraction of payload used by each object collection in the NanoAOD format, for tt simulated events with the LHC Run 2 pileup scenario.

Collection	Size
Jets	23%
Generator particles	19%
Electrons	7%
Trigger objects	6%
Generator-level jets	6%
Tau leptons	5%
Secondary vertices	5%
High level trigger	4%
Photons	4%
Muons	4%
Other	17%

include all variables needed to calculate identification decisions; it rather contains directly the decision bits, or the score of top level discriminant variables, such as those obtained by multivariate analysis methods. Systematic variations of object observables are also dropped whenever possible, in favor of their on-the-fly recalculation in later steps of the analysis. For instance, this is the case for uncertainties in the jet energy calibration (order of 50 variations), which can all be expressed as a function of calibration meta-data and only a few jet observables.

The per-event information contained in the TTree described above is complemented by coarse grain, auxiliary information in additional TTrees. This feature is used, for instance, to store sums of generator weights and allow for normalization of simulated processes, or for run-dependent data taking conditions.

A key feature of the NanoAOD format is the inclusion of documentation within the ROOT file. The branch title field is used to store a minimal description of each variable, and can be dynamically customized by different configurations of the NanoAOD producer software. This is of great value for tracking the content of each NanoAOD production, and allows for a quicker learning curve for collaborators adopting this data format in their workflows.

NanoAOD can be produced from MiniAOD files at a rate of about 10 events per second on a single CPU core. This is comparable to typical user ntuples used for analysis, and much faster than all other central production workflows (with the exception of computationally trivial ones, such as event skims based on trigger decisions or object multiplicity). This sets the typical timescale to run a full NanoAOD production on data collected in one year, about 10 billion events, to one week. This value can vary depending on the load due to other workflows running

at the same time on CMS computing resources.

3. Recent additions to the event content

The NanoAOD event content undergoes a continuous process of tuning, based on the experience we are accumulating with analyses adopting it in place of the MiniAOD + user ntuples processing model.

The jet information has been recently extended to include jet substructure taggers based on machine learning techniques, such as deep neural networks. The inclusion of this information poses a significant challenge in terms of payload size, as their output is typically highly dimensional to tackle multi-classification problems. Moreover, the evaluation of complex multivariate discriminants takes an increasing fraction of data processing time, with a larger relative impact on high-rate workflows such as NanoAOD.

Recent experience has also shown that the possibility to recalibrate jets and missing transverse energy is highly valued by analyzers. We have extended the NanoAOD content to all reconstructed jets with a transverse momentum of at least 10 GeV, slimming significantly the content for softer jets that only enter the missing transverse energy calculation, but are not used as standalone objects.

4. Impact on analysis model

The impact of NanoAOD on the typical analysis model in CMS has still to be observed in its entirety, due to the relatively recent introduction of this data format. Its adoption by analysis groups is quickly growing, and we are optimistic that this trend will continue throughout the current LHC shutdown period and Run 3.

We expect that a key role will be played by post-processing tools, that are used to perform the last analysis-specific steps of the data reduction chain. These steps cannot be made fully general, as they depend on the chosen definition of identified physics objects. Nevertheless, they follow a common logic for a large fraction of physics analyses. This allowed us to define a set of generic tools, categorized according to their function:

- calculation of systematic variations of object observables, data to simulation corrections and scale factors for identification efficiency;
- slimming and merging collections whose objects are used together in later steps (e.g. create a generic collection of light leptons from identified electrons and muons);
- calculation of more complex event variables based on several physics objects (e.g. for tagging hadronic top decays from resolved jets, calculation of invariant masses, transverse mass variables based on leptons and missing transverse energy).

These items were privately implemented in each analysis framework so far, on the basis of sets of instructions provided by object performance groups. A large part of them has now been included in NanoAOD production; for the remaining part, we envisage to centrally maintain a set of tools in a modular post-processing framework, under a tight review by object performance groups. At the same time, we expect that innovative methodologies developed by analysis groups will be implemented in the same modular framework, favoring their more efficient sharing within the collaboration.

NanoAOD also has the potential to facilitate analysis preservation. The key feature, in this perspective, is the possibility to read NanoAOD files without ROOT dictionaries or any other experiment specific software. We expect this will make it easier to the keep the data accessible in the future, and convert them to different formats if desired. Moreover, dependencies of analysis specific code on the rest of the CMS reconstruction software will be reduced. At the same time, the flat TTree structure also allows to profit more easily from recent developments in ROOT,

such as the RDataFrame interface [4, 5], as well as from other standard tools for data analysis and machine learning.

5. Impact on calibration workflows

Groups in charge of the definition and maintenance of high level physics objects perform regular studies on the observables relevant to them. For instance, jet and electron energies are corrected for detector effects in order to stabilize their response and optimize their resolution. In a similar way, the efficiency for physics objects to pass identification requirements is measured in data and simulation, in order to derive per-object corrections which are then used in all CMS physics analyses.

These workflows are typically based on custom intermediate data formats, not always optimized in terms of event content and maintained separately by each group. They require significant manual intervention each time they have to be run, even if the analysis methodology is quite mature and seldom evolves. In several cases, they have reached a level of complexity that is comparable to a full physics analysis in terms of number of input datasets and sources of systematic uncertainty. Moreover, these corrections have to be derived for each data taking period, as detector conditions evolve with time, and depend on each other - for instance, scale factors for the identification of b jets depend on jet energy calibration.

An improved efficiency in the execution of these tasks would be of great benefit, and we believe that the NanoAOD data format can play an important role in this respect. In the first place, the central production of these datasets either removes the need for private group ntuples, or at least it standardizes their basic structure. At the same time, a strong effort is being put into the automation of later steps in calibration workflows. Moreover, NanoAOD favors a tighter integration between calibration and other analysis activities. For instance, it makes it easier to use the most robust methods for measuring the performance of analysis-specific object identification discriminants.

6. Outlook

The introduction of the NanoAOD data format has a dramatic impact on the estimates for computing resources needed by the CMS experiment during the High-Luminosity operation phase of the LHC. Assuming a design target of 50% analysis coverage with NanoAOD is met by then, the projected needs [6] for disk storage in 2027 are decreased by a factor of about 2, corresponding to a reduction of more than 2 EB. The needed CPU processing power is also decreased by about 15%, as user ntuple production is partially replaced by the central NanoAOD workflow.

A large set of analyses based on the full dataset collected during the LHC Run 2 will be completed in the next months. The data format has been fully commissioned in time for this major round of data analysis, and is now part of all central workflows for dataset production. We believe that the experience and feedback we are accumulating in this moment will prove invaluable to steer the future development of the project, and meet the analysis coverage goal.

As the NanoAOD code base stabilizes, we anticipate that our focus will shift on exercising repeated productions of NanoAOD datasets with refined content and varying conditions for analysis (e.g. object calibration). Another direction of improvement is the extension of the included generator information. Alternate choices for generator scales and parton distribution functions are currently expressed in terms of multiple event weights, that cannot fit within the typical NanoAOD event size. We envisage to explore ways to reduce this information and encourage a standardization of all workflows consuming it, in a tight collaboration with generator experts and analysis groups.

References

[1] Chatrchyan S *et al* (CMS Collaboration) 2008 The CMS Experiment at the CERN LHC *JINST* **3** S08004 (doi:10.1088/1748-0221/3/08/S08004)

[2] Petrucciani G, Rizzi A and Vuosalo C for the CMS Collaboration 2015 Mini-AOD: A New Analysis Data Format for CMS *J. Phys.: Conf. Series* **664** 072052 (doi:10.1088/1742-6596/664/7/072052)

[3] Rizzi A, Petrucciani G and Peruzzi M for the CMS Collaboration 2018 A further reduction in CMS event data for analysis: the NANOAOD format *Preprint* CMS CR-2018/396 (to appear in the proceedings of the CHEP 2018 conference)

[4] Brun R and Rademakers F 1997 ROOT - An Object Oriented Data Analysis Framework, *Proc. AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res.* A **389** 81-86 (doi:10.1016/S0168-9002(97)00048-X). See also "ROOT" [software], Release v6.16/00, 05/02/2019, doi:10.5281/zenodo.2557526

[5] Guiraud E, Naumann A and Piparo D 2017 TDataFrame: functional chains for ROOT data analyses (Version v1.0) *Zenodo* doi:10.5281/zenodo.260230

[6] Boccali T and Klute M 2018 CMS report at WLCG LHCC referees meeting *Slides* https://indico.cern.ch/event/685793