EPJ Web of Conferences **214**, 06021 (2019) https://doi.org/10.1051/epjconf/201921406021
*CHEP 2018*

# A further reduction in CMS event data for analysis: the NANOAOD format

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

[1]University and INFN of Pisa
[2]CERN

**Abstract.** A new event data format has been designed and prototyped by the CMS collaboration to satisfy the needs of a large fraction of physics analyses (at least 50%) with a per event size of order 1 kB. This new format is more than a factor of 20 smaller than the MINIAOD format and contains only top level information typically used in the last steps of the analysis. The talk will review the current analysis strategy from the point of view of event format in CMS (both skims and formats such as RECO, AOD, MINIAOD, NANOAOD) and will describe the design guidelines for the new NANOAOD format.

## 1 Introduction

The typical data analysis flow in CMS [1], as well as in other LHC experiments, comprises several steps of data processing and reduction. While collisions take place at a 40 MHz rate and order of a hundred million channels are readout in the experiment, the final distributions included in published papers typically include only few observables from a very reduced set of significant events.
The data reduction starts already in the detector hardware, with zero suppression algorithms and trigger systems. While the first reduce the amount of data in each event (event content), the latter reduce the number of events to be processed. Further steps of data reduction are then applied in the reconstruction and analysis chain.
Different types of analysis are expected to need different levels of data reduction. For example, calibration processes often rely on a high level of detail in a subset of the collected events. Searches and precision measurements, on the other hand, are usually limited by the integrated luminosity hence they require a large number of events with less detail on low level detector information.

During the LHC Run 1, the following flavours of event content (data tiers) were available in CMS:

- Virgin-RAW: used only in low rate runs with heavy ions collisions (10-15 MB/ev)
- (Zero-suppressed) RAW : standard raw data event content (1 MB/ev)
- RECO: detailed information on reconstructed physics objects (3 MB/ev)
- AOD: physics objects used in analysis (400-500 kB/ev)

*e-mail: andrea.rizzi@cern.ch

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

- User defined ntuples (5 to 100 Kb/ev)

At the same time, an event selection can be applied at any step of the following chain:

- L1 Trigger
- High Level Trigger
- Primary Dataset (PD) definition: PD are sets of events with similar features at trigger level (e.g. events selected by different single muon triggers are grouped under the *SingleMuon* PD)
- Skimming: reduction of number of events based on specific trigger bits or high level reconstructed quantities (not used in most analyses)
- User selection applied at analysis level

At the end of Run 1, a new format reducing the event size by one order of magnitude with respect to AOD was designed, in order to increase the flexibility in handling the available computing resources with little impact on physics analyses. This format, named MINIAOD [2], is now the standard for CMS analyses in Run 2 and it is produced for all real and simulated events without any selection. It retains a very large flexibility as it includes all observed particles. This allows, for example, to recluster jets or to recompute b-tagging and jet subtructure observables. It also includes a high level of detail on leptons and photons, supporting their recalibration and the development of new identification techniques.

With LHC entering a regime where the collision energy will not be dramatically increased, much larger datasets will need to be processed in physics analyses to achieve their ultimate sensitivity. Therefore, a further data reduction will be soon be needed to ensure the reach of the experiment's scientific goals.
Such data reduction cannot be achieved by increasing trigger thresholds or imposing tighter skimming requirements, because the energy scale of interest will continue to be driven by the masses of the particles under study, such as the Higgs boson. A significant increase in energy thresholds applied at trigger level would significantly reduce the analysis acceptance in this phase space, canceling out the benefits from higher integrated luminosity.

For these reasons, in the same spirit of what was done for the MINIAOD format, we investigated the feasibility of a more compact data tier aiming at a decrease of more than one order of magnitude in event size.

## 2 How to achieve a further size reduction

The key to achieve size reduction is to understand how much information is strictly needed to make a given set of analyses possible. Our hypothesis has been that a large set of analyses (50% to 70%) would mostly need the same high level information (objects kinematic and identification properties), with little need of lower level details.
While a large variability in the content of user ntuples can be observed by sampling their current status, one should discriminate how much of such variability consists of slightly different implementations of a similar concept, and how much of it is due to truly different information being exploited. We actually observed that, in many cases during Run 1 and early Run 2, the user ntuples produced by a group were reused by other groups for different analysis goals.
On the basis of personal experience with Higgs and SUSY analyses, and interactions with several analysis groups in CMS, we made a list of the needed top level information (with no

focus on the exact definition of each variable) and concluded that a 1–2 kB/event goal was within reach.

The guiding principles to the design of the new format can be summarized as follows:

- No tracks or individual particle candidates
- No details linked to detector configuration
- Prefer precomputed object identification variables to inputs needed to derive them
- Allow to store complex high level analysis variables
- Limit the number of physics objects with reasonable thresholds
- Limit the information on collections with many entries (e.g. jets)
- Store each floating-point variable with a precision adequate to its physical meaning
- Do not store variations of quantities that can be recomputed from the available information

The last point is a key element for all those systematic variations (e.g. on jet energy corrections from different uncertainty sources) that can be computed on each object as a function of its properties ($f_{corr}(p_T, \eta, ...)$) in a non-event-specific way.

Finally, in order to make ntuples that are useful for several analysis groups, we needed to avoid making choices that are analysis-specific at this early stage. A typical example would be the so-called "object cross cleaning", i.e. the decision on how to remove the overlap between different physics objects reconstructed from the same input constituents (e.g. an electron or a jet, a tau jet or a quark/gluon jet). While no cleaning is performed in production, objects sharing the same origin are linked to facilitate the event interpretation after such disambiguation is performed.

## 3 The NANOAOD format

A first realization of this concept has been developed in CMS and has been named NANOAOD[1]. The present prototype is based on bare ROOT [3] TTrees with no specific dictionary needed for I/O. This mimics the typical format of user ntuples. Naming conventions on column (branches) names are used to group information belonging to the same high level object. Branches from the same object have all the same length in an event, and no further array nesting within branch elements is allowed. The branch naming convention is the following:

- The object name is the root part of all branch names (e.g. Jet, Electron, Muon) and attributes are separated from the root with an underscore (e.g. Jet_pt, Jet_eta, Muon_dxy).
- A single integer variable, for each object name, specifies the length of all branches related to that object. The name of such variable is "n" followed by the object name (e.g. nJet, nElectron, nMuon).
- Single (scalar) event-variables (e.g. GenWeight, RunNumber, etc...) are also possible without a length being specified and can optionally be grouped into single event-objects by sharing the root part of the branch name (e.g. MET_pt, MET_phi).

[1]The prefix "nano", rather than "micro", has been used to allow an intermediate format between "nano" and "mini" that would still contain all individual particles, rather than only clustered objects.

- References between objects are obtained with the Idx suffix with the following structure: *RootObject_pointedObject*Idx*OptionalQualifier* where *pointedObject* is the root name of the pointed collection. (e.g. Electron_photonIdx points to the photon associated to the same calorimetric cluster of a given electron, *GenPart_genPartIdxMother* points to the mother of a given generator-level particle). References are simply the position of the item in the collection and their content has no meaning if the pointed collection is pruned.

The title attribute of each branch is filled with minimal documentation on the content of each variable, so that the format is self-documenting. A full html reference manual with the documentation of all variables can be produced automatically from the ROOT file. This is especially useful for branches whose content is determined dynamically, and thus are different for different samples (e.g. some generator-specific information).

In addition to the TTree containing events information a few ancillary trees are stored with information that is run or luminosity section specific and that contains meta-data such as processing configuration.

## 4 The NANOAOD content

The current content of NANOAOD consists of all physics objects, including jets, electrons, photons, muons, tau leptons, trigger information, missing transverse energy, generator information, event weights and cleaning flags, primary and secondary vertices, isolated tracks, fat jets and their substructures, SoftActivity information (track jets) and more. The size used by each object collection is shown in Figure 1 and details on the individual contributions are given in Table 1.

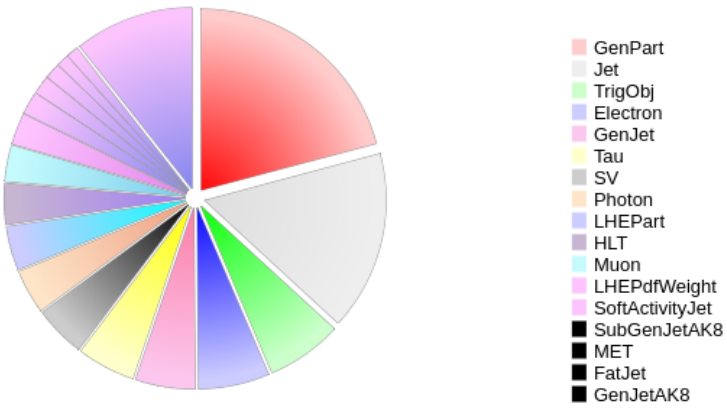


**Figure 1.** NANOAOD composition in term of relative size used by each object collection, on a simulated $t\bar{t}$ sample.

## 5 The NANOAOD tools ecosystem

While we avoid storing some sources of systematic variations, as they are not really event specific, that information eventually needs to be calculated for analysis. Moreover, even if NANOAOD is compact enough to be produced easily on large datasets without event preselection, it is often useful to reduce the number of events for a specific analysis.

| collection | variables | items/evt | kBytes/evt | bytes/item | fraction |
|---|---|---|---|---|---|
| Generator Particles | 9 | 53.38 | 0.330 | 6.3 | 20.8% |
| Jets | 33 | 8.45 | 0.253 | 30.7 | 16.0% |
| Trigger Objects | 11 | 10.32 | 0.107 | 10.6 | 6.7% |
| Electrons | 48 | 1.14 | 0.100 | 89.3 | 6.3% |
| Generator level Jets | 7 | 7.68 | 0.084 | 11.2 | 5.3% |
| Tau | 38 | 1.33 | 0.082 | 63.1 | 5.2% |
| Secondary Vertices | 13 | 2.77 | 0.072 | 26.6 | 4.5% |
| Photons | 28 | 1.50 | 0.062 | 42.3 | 3.9% |
| LHE level particles | 6 | 7.00 | 0.062 | 9.0 | 3.9% |
| HLT bits | 569 | 1.00 | 0.058 | 59.0 | 3.6% |
| Muons | 35 | 0.76 | 0.051 | 68.4 | 3.2% |
| LHE PDF Weights | 2 | 33.00 | 0.044 | 1.4 | 2.8% |
| SoftActivity Track Jets | 4 | 5.96 | 0.031 | 5.4 | 2.0% |
| AK8 Subjets at Generator level | 5 | 2.26 | 0.026 | 11.6 | 1.6% |
| Missing ET | 11 | 1.00 | 0.022 | 22.6 | 1.4% |
| AK8 Jets | 21 | 0.32 | 0.017 | 54.4 | 1.1% |
| AK8 Jets at Generator level | 7 | 1.16 | 0.016 | 14.6 | 1.0% |
| LHE Scale Weights | 2 | 9.00 | 0.016 | 1.8 | 1.0% |
| AK8 Subjets | 14 | 0.43 | 0.014 | 34.4 | 0.9% |
| LHE variables | 10 | 1.00 | 0.011 | 11.4 | 0.7% |
| Primary Vertex | 8 | 1.00 | 0.011 | 11.3 | 0.7% |
| Isolated Tracks | 13 | 0.32 | 0.011 | 34.3 | 0.7% |
| Dressed Generated Leptons | 6 | 0.62 | 0.009 | 14.6 | 0.6% |
| Parton Shower Weights | 2 | 4.00 | 0.008 | 2.2 | 0.5% |
| Generator information | 9 | 1.00 | 0.008 | 7.9 | 0.5% |
| Track Missing ET | 3 | 1.00 | 0.008 | 7.8 | 0.5% |
| Puppi Missing ET | 3 | 1.00 | 0.008 | 7.8 | 0.5% |
| b-tag weight | 2 | 1.00 | 0.006 | 6.2 | 0.4% |
| Other Primary Vertices | 2 | 2.94 | 0.006 | 2.0 | 0.4% |
| Calorimetric Missing ET | 3 | 1.00 | 0.006 | 5.7 | 0.4% |
| Raw Missing ET | 3 | 1.00 | 0.006 | 5.7 | 0.4% |
| Pileup information | 4 | 1.00 | 0.004 | 4.6 | 0.3% |
| Generated Tau | 8 | 0.21 | 0.004 | 20.4 | 0.3% |
| Generated Missing ET | 2 | 1.00 | 0.004 | 3.9 | 0.2% |
| others | 45 | | 0.03 | | 2% |

**Table 1.** Average space usage for each collection in NANOAOD for a simulated $t\bar{t}$ sample with Run 2 pileup scenario. The second column shows the number of variables associated to each item in the collection, the third column contains the average number of items per event (no maximum hard limit is enforced), the fourth column shows the size per event while the fifth has the size per item, the last column is the fraction of NANOAOD budget used by each collection.

A set of tools has therefore been created to further process NANOAOD files, maintaining the same format in input and output. We anticipate that analysis procedures of wide usage ("recipes") could be included in this framework, ensuring a quick development and consistency of analyses. The NANOAOD tools set contains, for instance, a modular event processor ("post-processor") that can be easily configured to:

- skim events,
- add systematic variations for relevant variables,
- keep or drop individual branches,
- run analysis modules from a common, centrally-maintained repository.

## 6 Data processing model

The time scale to produce a full set of NANOAOD samples from exising MINIAOD is expected to be order of 1 week for 1 year worth of data (about 10 billion events). In order to meet this goal, the processing speed to produce NANOAOD needs to be of the order of 10 Hz on a single core.
The one week target has been set keeping in mind that this is the typical turn around time of user ntuple production. This fits very well within a model where reconstruction software is only rerun about once per year, while MINIAOD files are produced (with new calibration and improved high level algorithms) only a few times per year.

## 7 Conclusions

A prototype of a data format for analysis reduced to 1–2 kB per event, NANOAOD, has been developed by the CMS collaboration. It has been recently used for physics analyses (e.g. [4]) and is being tested by a larger group of physicists in the collaboration. Further tuning of its content is expected to be needed in the next years, with the goal of having this format fully commissioned both as legacy of the Run 2 and as the event format for a large fraction of the analyses in Run 3 and beyond.

## References

[1] CMS Collaboration, The CMS Experiment at the CERN LHC, JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] G. Petrucciani, A. Rizzi, C. Vuosalo for the CMS Collaboration, Mini-AOD: A New Analysis Data Format for CMS, J.Phys.Conf.Ser. 664 (2015) no.7, 072052, doi:10.1088/1742-6596/664/7/072052, arXiv:1702.04685.

[3] R. Brun and F. Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86, doi:10.1016/S0168-9002(97)00048-X. See also http://root.cern.ch/.

[4] CMS Collaboration, Observation of Higgs boson decay to bottom quarks, Phys. Rev. Lett. 121, 121801 (2018), doi:10.1103/PhysRevLett.121.121801, arXiv:1808.08242.