PROCEEDINGS OF SCIENCE

Early experience with the Run 2 ATLAS analysis model

Jack Cranshaw on behalf of the ATLAS collaboration[*] [†]
Argonne National Laboratory
E-mail: cranshaw@anl.gov

During the long shutdown of the LHC, the ATLAS collaboration redesigned its analysis model based on experience gained during Run 1. The main components are a new analysis format and Event Data Model which can be read directly by ROOT, as well as a "Derivation Framework" that takes the petabyte-scale output from ATLAS reconstruction and produces smaller samples targeted at specific analyses, using the central production system. We will discuss the technical and operational aspects of this new system and review its performance during the first year of 13 TeV data taking.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

[*] Speaker.

[†] Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract DE-AC02-06CH11357.

PoS(ICHEP2016)170

1. Introduction

The ATLAS[1] Run 1 analysis model had several features which were not expected to scale to Run 2. During Run 1 the analysis and reconstruction data products were incompatible. This led to the duplication of data in order to accommodate both formats and also meant that CPU time was wasted converting the formats[2]. On top of the resource inefficiencies, different environments between reconstruction and analysis and between different analysis groups made cross checks of physics analyses more difficult.

Consequently there was a clear need for improvement in Run 2. The improvement followed two main thrusts: a common data format for reconstruction and analysis and a data reduction framework to prepare prefiltered samples for physics groups in a production environment using tools that the physics groups could reuse locally.

ATLAS uses multiple data formats, but the data products that connect reconstruction to analysis are written in ROOT[3]. During Run 1, the reconstruction format was designed for fast retrieval of groups of events and optimized for space. Physics analyses, on the other hand, wanted fast retrieval of individual variables and to not have to rebuild objects. For Run 2 we are using a new format which is a compromise for both environments, but is able to provide the event-wise access needed by reconstruction or the column-wise access needed by analysis by changing the ROOT settings used for writing. The tool we use for this is called an auxiliary store and is described in detail in previous proceedings[4].

2. Data reduction framework

A feature common to many physics analyses is the use of intermediate-sized data products at an early stage of the analysis procedure. Typically these formats are made directly from the retained output of the reconstruction (known in ATLAS [1] as Analysis Object Data or AOD) and often have the following features:

1. They are centrally produced for both data and simulation, and their size is usually between one hundredth and one thousandth of the input data size.

2. They are typically aimed at one analysis or perhaps a group of related analyses (e.g. sharing the same final state).

3. They apply calibrations and object selections (often shared with other physics groups) as they are made.

4. They usually contain all of the information necessary to perform smearing, scaling, selection, calibration and other operations on reconstructed objects (collectively known in ATLAS as combined performance operations), and the systematic uncertainties related to these operations.

5. They are typically reproduced 10-12 times per year but are often read several times per week by the analysis teams.

The process of data reduction can be broken down into several distinct categories. ATLAS uses the following terminology for its data reduction operations:

- *Skimming* is the removal of whole events based on some criteria related to the features of the event.
- *Thinning* is the removal of individual objects within an event based on some criteria related to the features of the object, e.g. a kinematic requirement.
- *Slimming* is the removal of variables within a given object type, uniformly across all objects of that type for all events. The same variables are removed for every event and object.
- *Augmentation* is the addition of information needed for analysis during the data reduction operation.

3. Implementation

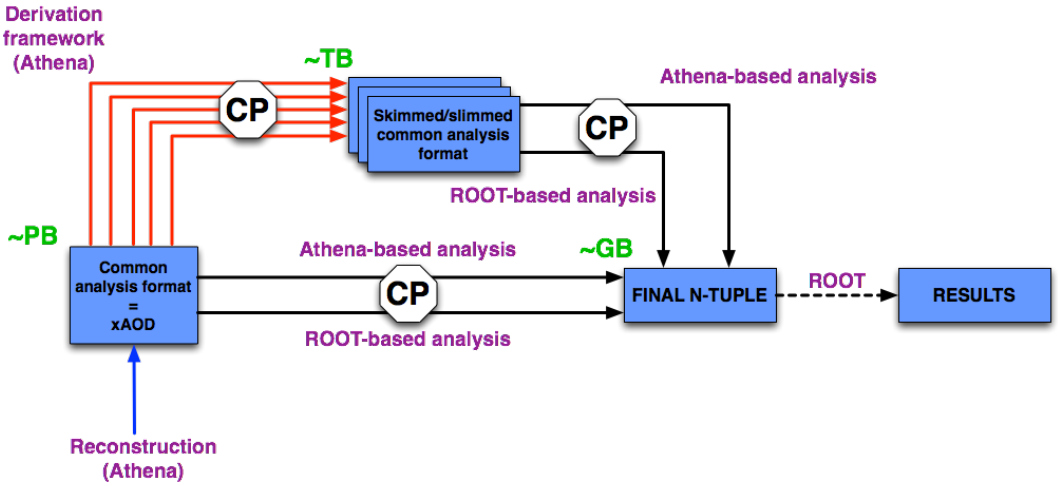

Figure 1: The ATLAS Run 2 analysis model showing how the reconstruction output (AOD) is transformed by derivation framework into multiple streams of DAOD. The original AOD and the DAOD have compatible data models which implies that analysis software can use either as input.

Figure 1 shows the data flow for the ATLAS Run 2 analysis model where the data reduction is done by the derivation framework. The derivation framework is used to create the intermediate data products from the reconstruction output (AOD) by removing (and adding) information while maintaining the structure and EDM used in the original AOD. The third and final component of the model is the analysis framework, which is used by physicists to read the derived data products, apply various combined performance tools and produce the final small n-tuples, from which plots are produced and upon which statistical analyses are based.

The role of the derivation framework should therefore be seen in terms of both software and computing: it provides physicists with tools to define the intermediate formats, and it runs their jobs on the central production system. Analysers are thereby freed from the trouble of designing

their own intermediate formats, and the considerable labour involved with running over the entire data sample themselves. Users still have access to the full xAOD as is implied by the lower route in Figure 1, but gradually the majority are moving to the centrally produced derived products.

The derivation framework uses the ATLAS framework, Athena, but derivation developers (physicists) don't interact directly with Athena but rather with the interface that the derivation framework provides. Athena was chosen because it allows derivations to leverage the well developed and tested I/O infrastructure for streaming and access to the reconstruction tools and algorithms.

The interface/toolkit includes tools for doing the four data reduction processes described above: skimming, thinning, slimming, and augmentation. Augmentation takes two forms, adding new containers of objects such as modified jets and decorating existing objects with extra variables, e.g. 'good muon'. Another innovation which has greatly simplified implementation was the development of code to do expression evaluation so that operations could be defined in text selections such as `count(Muons.pt > 25.0*GeV && abs(Muons.eta) < 2.5) >= 4` or `InDetTrackParticles.pt > 5.0*GeV` rather than hidden in bits of code.

Each derivation is defined in a single configuration file called a carriage and these carriages are grouped into trains. Various criteria are used to define the trains. Carriages that are part of the same physics group are often in the same train for ease of management. Carriages that use similar physics corrections or calibrations may be grouped together to save cpu. Finally, carriages with similar output sizes are grouped together to avoid or improve merging into final outputs. If necessary, special trains can be made if a problem is found with just a few carriages and it would waste grid resources to rerun production trains. Production trains are tested daily against the latest software changes.

4. Performance

As shown in Table 1, thirteen physics groups have defined around one hundred derivation samples. This number has changed during the run as carriages are added or removed based on physics needs. The derivation framework reads reconstructed output (AOD) and writes the various derivations (DAOD).

Table 1: Number and variety of derivations by ATLAS physics and combined performance groups. The derivations are run in groups called trains and each derivation is a carriage.

	B Physics	EGamma	Flav. Tag	Inner Det.	Muon	Tau	Exotics
carriages	2	8	4	1	5	2	18
	Higgs	Jet	SM	SUSY	Tile	Top	**Total**
carriages	20	11	5	13	1	4	**94**

The recommendations to the physics groups are that each derivation size should be no more than 1% of the input AOD size and each physics group is allowed a cumulative size up to 4% of the total AOD size for data (Monte Carlo size limits are handled separately). Figure 2 shows the fractions for the various derivations for data and Monte Carlo. This shows that most of the

derivations cluster around the recommendations but can exceed them where necessary. One of the successes of the framework has been that these decisions can be discussed in a physics context in ATLAS physics coordination and the decisions quickly translated into action on the software side.

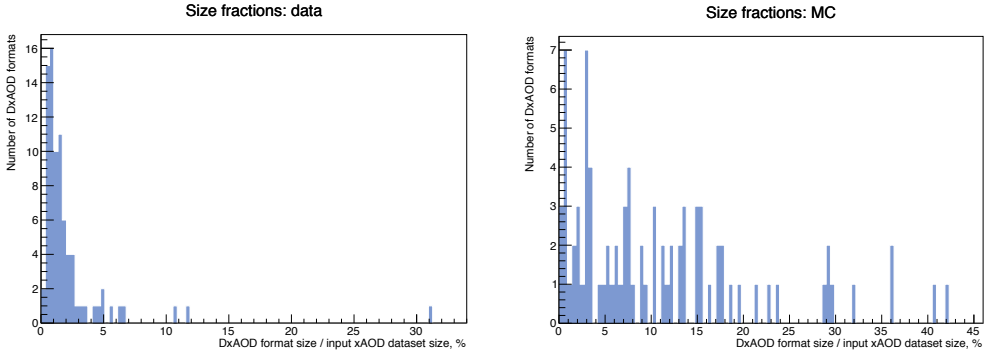

Figure 2: Size fractions for the various derivations for data and Monte Carlo.

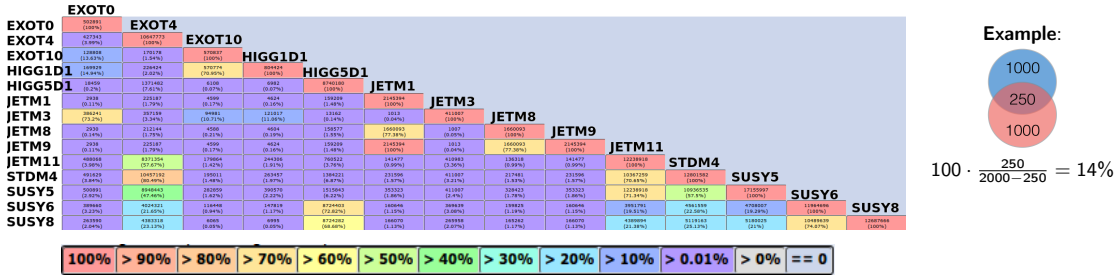

Figure 3: Monitoring of event-wise overlap fractions for the 14 derivations (out of 94) with >70% overlap.

One of the resource problems in Run 1 was the duplication of data across physics groups. Event overlaps among the derivations are monitored using the ATLAS event level metadata system, the Event Index[5]. The majority of the 94 derivations have event-wise overlaps with other derivations of less than 10%. Figure 3 shows the overlaps for the 14 derivations which had an event-wise overlap with at least one other derivation of at least 70%. We define the pairwise overlap as the intersection divided by the union. When a large overlap is detected it triggers an investigation, but not a change. If there is not a large overlap in content as well as events, or if there is an operational reason to have two derivations, then they may be kept. Monitoring overlaps has also informed physics groups that there are areas where collaboration might be useful.

5. Summary

The ATLAS Run 2 analysis model has been a success and has used resources much more economically than the Run 1 version. We are able to run of order 100 derivations in of order 10 trains. The model is also flexible enough that when necessary we have rearranged trains by moving carriages around. This is possible due to efficient nightly testing. Weekly coordination

meetings provide regular feedback from the physics groups. The system has worked both for data and Monte Carlo derivations. The interface allows physics groups the ability to manage their derivation sizes and they have used these successfully to fit within resource constraints. Over time physicists have moved away from using the primary AOD to using the prefiltered samples as they have what they need and are managed by the physics groups themselves. Also the format and production of derivations outputs have placed no serious constraints on the development of physics analysis frameworks[6]. It is foreseen that this system will work successfully for the rest of Run 2.

References

[1] ATLAS Collaboration 2008, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST 3* **S08003** [doi:10.1088/1748-0221/3/08/S08003]

[2] P. Laycock, et. al., *Derived Physics Data Production in ATLAS: Experience with Run 1 and Looking Ahead*, *J. Phys.: Conf. Ser.* **513** (2014) 032052]doi:10.1088/1742-6596/513/3/032052].

[3] R. Brun et. al., *ROOT: An object oriented data analysis framework*, *Nucl.Instrum.Meth.* **A389** (1997) 81-86

[4] A. Buckley et. al., *Implementation of the ATLAS Run 2 event data model*, *J. Phys.: Conf. Ser.* **664** (2015) no.7, 072045 [doi:10.1088/1742-6596/664/7/072045].

[5] D. Barberis, et. al., *The ATLAS EventIndex: architecture, design choices, deployment and first operations experience*, *J. Phys.: Conf. Ser.* **664** (2015) no.4, 042003 [doi:10.1088/1742-6596/664/4/042003].

[6] D. Adams, et. al., *Dual-use tools and systematics-aware analysis workflows in the ATLAS Run-2 analysis model*, *J.Phys.Conf.Ser.* **664** (2015) no.3, 032007 [doi:10.1088/1742-6596/664/3/032007]