Distributing and storing data efficiently by means of special datasets in the ATLAS collaboration

Karsten Köneke (for the ATLAS collaboration)
CERN, CH-1211 Genve 23, Switzerland
E-mail: karsten.koeneke@cern.ch

Abstract. With the start of the LHC physics program, the ATLAS experiment started to record vast amounts of data. This data has to be distributed and stored on the world-wide computing grid in a smart way in order to enable an effective and efficient analysis by physicists. This article describes how the ATLAS collaboration chose to create specialized reduced datasets in order to efficiently use computing resources and facilitate physics analyses.

1. Introduction
The ATLAS experiment [1] is one of four experiments situated at an interaction point of the Large Hadron Collider (LHC) [2, 3, 4] at CERN, the European Organization for Nuclear Research located near Geneva, Switzerland. The first collisions of the two proton beams at a center-of-momentum (c.m.) energy of 7 TeV took place half a year before this conference. Later on in the long-lasting LHC program, the c.m. energy will be increased to the full design c.m. energy of 14 TeV. The physics goals of the ATLAS experiment contain, e.g., precision measurements of the W and Z boson, the top quark, and other Standard Model (SM) processes, as well as the search for the Higgs boson, for Supersymmetry, for Large Extra Dimensions, or any other so far unmeasured physics processes [1].

The volume of data that is produced by the ATLAS experiment in 2010 exceeds everything previously existing within high-energy physics. Several Petabytes have been recorded in its first year. This enormous amount of data gets analyzed by about 3000 members of the ATLAS collaboration. This grand challenge requires new ways of organizing, distributing, and analyzing data. The new concept of distributed computing over the LHC computing grid [5] is designed to cope with these challenges (see Sec. 2).

The required computing resources and time needed to process the full dataset is very large, see Sec. 3. An introduction of datasets that are specialized to certain tasks mitigate this problem because a given sub-group of the ATLAS collaboration is generally only interested in a special sub-set of the data. This path is chosen by introducing the new Derived Physics Datasets (DPD), see Sec. 4. Finally, a summary of this article is given in Sec. 5.

2. The ATLAS distributed analysis model
The ATLAS experiment records events at a rate of about 200 Hz which corresponds to a data rate of over 300 Mbyte/s. In order to process this unprecedented large amount of data and make it available to the world-wide community of ATLAS physicists, the ATLAS collaboration
adopted a tree–like structure with several layers, or computing tiers, that is the ATLAS part of the LHC computing grid [6].

The raw data gets transferred from the ATLAS experiment to the Tier 0 computing center at CERN and a first–pass event reconstruction is performed resulting in the Event Summary Data (ESD) which contains all the reconstructed objects, but not any more all the raw data. Its average per–event size is approximately 1.5 Mbytes and it is produced for all events. A subsequent processing step, also at the Tier 0, is producing the Analysis Object Data (AOD) from the ESD. The AOD is a size–reduced data set that contains less detail than the ESD. The AOD is the primary data format for physics analysis with an average per–event size of about 150 kbytes and it is also produced for all events. Both the ESD and the AOD are stored in the same format, the POOL format [7].

Due to the large overall size of the raw data, only two copies of the raw data are stored, distributed between the Tier 0 and the Tier 1s. All data files are copied to magnetic tape storage and only a very small fraction of the ESD and an even smaller fraction of the raw data will be also kept on disk drives for random access. This small fraction, together with all AOD files are then distributed to the next layer of the LHC computing grid, the Tier 2s, where they will reside on disk drives available for usage by all members of the ATLAS collaboration.

3. Challenges of the ATLAS distributed analysis model
The detailed understanding of the physics response of the ATLAS detector is one of the most important tasks necessary for ensuring excellent physics results. So is the performance evaluation and improvements of the ATLAS reconstruction software. Very detailed detector information are needed for these kind of studies, information that is not usually available at the level of the AOD, but is available in the ESD. This poses a problem since the users of the ATLAS collaboration have only full access to data stored on disk storage at the Tier 2s, i.e., to the AOD and a very small subset of the ESD. Also, a typical study requires processing the full dataset multiple times. It would require vast computing and time resources if the user is required to process the full AOD dataset multiple times.

4. New specialized derived physics datasets
In order to circumvent the possible issues mentioned in the previous section, a new class of datasets has been designed, the so–called Derived Physics Datasets (DPDs). The individual user is usually only interested in a small fraction of the total data volume since this user will only do one or two types of analysis. Most of the possible data analysis tasks that users intend to do can be grouped into a few categories. Three distinct classes of data analysis have been identified and for each of these three classes a special set of DPDs has been designed.

The first class of data analysis was focusing on the commissioning of the ATLAS detector using cosmic muon events taken while the LHC is not operational and also data events from single–beam LHC operation. The commissioning DPDs designed for these tasks are described in Sec. 4.2. Determining and understanding the performance of the ATLAS detector and reconstruction software is always a very important task. For these performance studies, the performance DPDs described in Sec. 4.3 have been designed. The third category of data analysis is comprised by the actual final physics analyses. The physics DPDs tailored for these types of analyses are described in Sec. 4.4.

4.1. Tools for the DPD production
A few DPDs of each of the three categories mentioned above are produced with the specific analysis goals in mind from each of the several ATLAS trigger streams. To be able to produce several DPDs from a given ATLAS trigger stream, a new framework within the ATLAS software has been utilized that allows to produce several output streams from a single input stream. Each
of the configured output streams can schedule its own computation and selection algorithms and tools. This design permits the production of very different output DPDs in a single processing job.

To fit all DPDs within grid storage capacity constraints, the total DPD size is restricted to be about the same size as the total AOD size. This is achieved by employing several size reduction techniques.

- **Skimming.** This is a basic event selection. Only events that are interesting for the specific types of analysis that are envisioned to be done with the DPD at hand are kept. For example only events that contain at least two well reconstructed and identified electrons above a certain transverse energy threshold are kept. This DPD can then be used to study, e.g., $Z \rightarrow e^+e^-$ events or high-mass resonance searches.

- **Trimming.** Here, all objects of a certain type are removed. One can, e.g., remove all calorimeter cells from every event. One example is a DPD designed for track reconstruction performance studies where the calorimeter cells are not needed any more. This reduces the size of each individual event.

- **Thinning.** Only selected objects of a given type are removed. One example is given by a DPD that is designed to evaluate and optimize the performance of electron and photon reconstruction. Only calorimeter cells in the vicinity of interesting electrons or photons are kept while all other calorimeter cells far away are removed. This again reduces the size of each individual event.

- **Slimming.** Some parts of an individual object are removed. For example one can remove redundant error matrices from a reconstructed track. This is used when some very detailed information about an object is not needed any more in the final analysis. This also reduces the size of each event.

Since most DPDs do not contain all events any more, the information of how many events were actually processed from the input data needs to be stored. An event counter is run during the DPD production job and the total number of events and the number of selected events is stored in each output DPD. If this information is already found in the input file, as in case of subsequent processing steps, the total event count is simply added from all processed input files and the updated numbers are written into the output DPD. Also, the luminosity information is transferred correctly to each DPD, even if no event from the input file was selected.

In addition to this, the user needs to know why a given event was actually selected. For most DPDs, there are several distinct event filtering algorithms defined and an event can be accepted by either of these. The information which of the possible several event filtering algorithms actually accepted the event at hand is stored inside the DPD for each individual event.

### 4.2. The commissioning DPDs

Commissioning of the ATLAS detector using cosmic muon events and LHC single-beam data was very important before the LHC delivered large amounts of colliding-beam data. The studies performed with these cosmic muon events are, e.g., detector alignment studies using the muon tracks, measurements of the energy deposits of a cosmic muon inside the ATLAS calorimeters, cosmic muon trigger studies, and related tasks. For these types of studies, the event content of the ESD is the most suitable data set.

Several hundred million cosmic muon events have been recorded, comprising a total size of the order of 1 Pbyte of ESD files. For most analysis tasks, only a small subset of these events is needed, usually about 500k events. These are events where, e.g., a sufficiently localized energy deposition in the electromagnetic calorimeter was found, or events where a track close to the nominal interaction point was found that left a sufficient number of hits in the ATLAS silicon pixel detector. Eleven different categories have been identified and thus, eleven distinct
commissioning DPDs have been implemented. All of these have different event selection criteria but keep the full ESD event content. They are also known as commissioning Derived ESDs (DESDs). Each of the resulting commissioning DPDs has a total size of only 1 Tbyte or less. Therefore, they are easily stored on disk storage systems at the Tier 2s and are thus readily available to the whole ATLAS collaboration.

4.3. The performance DPDs

Very detailed information is needed in order to allow studying the detector performance and the performance of the software algorithms, information that is available in the ESD, but not necessarily in the AOD. Thus, the performance DPDs are produced from the ESDs and are also described as performance DESDs. At the time of the first data processing, the performance DPDs are produced at the Tier 0, after the reconstruction jobs produced the ESDs, thus minimizing possible time delays.

The performance DPDs are then distributed to disk storage at the Tier 2s, making them easily available to the whole ATLAS collaboration for analysis. This gives users access to ESD–level information, something which is not possible using the too large ESDs directly since they are saved on tape storage systems at the Tier 1s. The overall size restriction for all performance DPDs combined amounts to the same overall size as all AODs combined.

Nine distinct performance DPDs have been defined and implemented. Each of these nine performance DPDs is produced from the ESDs of only one trigger stream. This scheme is chosen because a specific performance analysis uses in most cases events from only one trigger stream.

Four of these nine performance DPDs keep the full ESD event content. In order to reduce the overall size to the required limit, only very specific events are selected. One example is a DPD designed for studying jet reconstruction algorithms and calibrating jets using a recoiling jet against either a photon or a $Z$ boson decaying into two electrons. Here, the event is selected if either a very well reconstructed and identified photon with large transverse energy is found or a loosely reconstructed $Z \rightarrow e^+e^-$ event is found.

Another five performance DPDs employ a somewhat looser event selection and reduce in addition the event size. Two of these reduce the per–event size by dropping all objects of a given type, i.e., the events are trimmed (see Sec. 4.1). In one case, the DPD is tailored towards studying calorimeter seeded jets. Here, the detailed tracking information is removed and only basic tracks are retained, in addition to the full calorimeter information. In the other case, the DPD is designed to study the tracking performance and thus, the calorimeter cells can be safely removed. The other two DPDs with a loose event selection utilize the possibility to remove individual objects of a given type, i.e., a given object category is thinned (see Sec. 4.1). One example is a DPD produced from the muon trigger stream where events are kept that contain either at least one muon with large transverse momentum, at least one $J/\psi \rightarrow \mu^+\mu^-$ candidate, or at least one $Z^{(0)} \rightarrow \mu^+\mu^-$ candidate. In addition to these event selection criteria, a simple pre-scale is sampling the events as well to obtain an unbiased subset of events. When an event is selected, only the calorimeter cells and the detailed information of the ATLAS inner tracking detector are kept which are close to a reconstructed muon.

In addition to the performance DPDs produced from ESDs described above, there are also a few performance DPDs produced directly from the raw data. Those performance DPDs select raw events containing well identified leptonically decaying $Z$ or $W$ bosons. The event selection is performed based on reconstructed quantities. But the resulting output DPDs have the same format as the raw data, just fewer events, and are thus known as Derived RAW (DRAW). In order to avoid loading and processing the raw data twice, these DPDs are produced in the same main reconstruction job as the ESD.
4.4. The physics DPDs

Physics analyses are performed in general on very specific final states and can be done using information that is stored in the AOD, without the need to also use more detailed information from the ESD. Thus, physics DPDs, which are produced from AODs and are thus also known as Derived AODs (DAODs), can have a rather stringent event selection applied, also since each of them should not exceed a total size of about 1% of the total AOD size.

Physics DPDs usually retain for each kept event the whole AOD content. In order to meet the size requirement a rather stringent event selection is used. The event selection is based on loosely reconstructed and identified objects. For example events that contain two good electrons with a certain minimum transverse energy are kept for one DPD. For another DPD, events that contain both a loosely identified electron and muon are kept. The former example will be only produced for the electron/photon trigger stream whereas the later example will be produced for both the electron/photon trigger stream and the muon trigger stream.

Physics analyses can benefit greatly from these physics DPDs. This is due to the large reduction in the DPD size which increases the physics analysis processing speed by a factor of about 100. The physics DPDs are produced at all the Tier 1s and then copied to disk storage at the Tier 2s for physics analyses.

4.5. Integration of DPDs into the ATLAS data production

All DPDs mentioned in the previous sections are so–called primary DPDs. These primary DPDs are produced centrally by the ATLAS data production team. Right after recording the raw data and the reconstruction of the events at the Tier 0 (see Sec. 2), the primary performance DPDs (see Sec. 4.3) are produced from the ESD, also at the Tier 0. They are then distributed to disk storage systems at the Tier 2s such that users can access them at any time. The primary physics DPDs (see Sec. 4.4) are produced from the AODs at the Tier 1s, also shortly after the AODs are available there. This processing scheme is illustrated in Fig. 1.

4.6. Different levels of DPDs

There is the option that users or user groups produce secondary and even tertiary DPDs. The general concept is that several secondary DPDs are produced from one primary DPD and that one or several tertiary DPDs are produced from one secondary DPD, thus building a tree–like data hierarchy. However, this hierarchy is not enforced. Users or user groups can also decide to produce a secondary or tertiary DPD directly from the AOD. But the concept is that each further layer of datasets is more specialized to a certain analysis. Also, the size reduction techniques described in Sec. 4.1 have to be applied more and more stringently the more a DPD is specialized for a small number of users. The secondary DPDs are restricted to be of the same data format as the ESDs, AODs, and primary DPDs. They can however store additional information which the user can compute during the production, e.g., Z boson candidates can be stored for each event. The tertiary DPDs can be completely defined by the user, but the files should be saved in the ROOT format [8] and produced with officially maintained tools.

5. Summary

Before the introduction of DPDs, there were potential problems for users to access the needed information. One issue was that detailed detector information is not distributed to disk storage systems at the Tier 2s due to size limitations. Another benefit of DPDs is the faster processing time and lower consumption of computing resources when reading these DPDs, as compared to reading ESDs or AODs. The concept of the DPDs is beneficial due to the creation of several distinct and more specialized datasets that are useful for only a certain class of analysis and also the possibility to specifically select only certain quantities from an input dataset. Several tools
Figure 1. The ATLAS data processing scheme. Shown is the path the data takes, starting at the ATLAS experiment at the LHC interaction point 1. The trigger system defines several trigger streams for the raw data which are then transferred to the Tier 0 computing center at CERN. Here, the event reconstruction is performed, resulting in the ESDs. Still at the Tier 0, the AOD and performance DPD production is done. The physics DPDs are then produced at the Tier 1s. All AODs and DPDs are subsequently transferred to disk storage at the Tier 2s.

that allow the creation of these specialized DPDs have been developed and integrated into the ATLAS software framework, including automatic bookkeeping tools.

Three different classes of primary DPDs have been identified. The first class is used only for cosmic muon and single–beam events. The second class is designed for evaluating the performance of the ATLAS detector and the reconstruction software. The primary physics DPDs comprise the third class of DPDs. These DPDs are used for the physics analyses.

The introduction of the DPDs has thus mitigated some possible issues and also improved the analysis speed due to the specialized content and event selection. The DPDs are expected to considerably improve the user access to the required data and advance the analysis speed.

References
[1] Aad G et al. (ATLAS) 2008 JINST 3 S08003
[2] Bruning O S et al. CERN-2004-003-V-1
[3] Bunin O et al. CERN-2004-003-V-2
[4] Benedikt M et al. CERN-2004-003-V-3
[5] Bird I et al. CERN-LHCC-2005-024
[6] Duckeck G et al. (ATLAS) CERN-LHCC-2005-022
[7] The POOL Project http://pool.cern.ch/
[8] The ROOT Project http://root.cern.ch/