Journal of Physics: Conference Series **513** (2014) 042009 doi:10.1088/1742-6596/513/4/042009

The future of event-level information repositories, indexing, and selection in ATLAS

D Barberis[1], J Cranshaw[2], G Dimitrov[3], T Doherty[4], EJ Gallas[5], J Hrivnac[6], D Malon[2], A Nairz[3], M Nowak[7], D Quilty[4], R Sorokoletov[3], P Van Gemmeren[2], Q Zhang[2] for the ATLAS Collaboration

[1] Universita di Genova and INFN, Genova, Italy
[2] Argonne National Laboratory, Argonne, IL, USA
[3] CERN, Geneva, Switzerland
[4] Glasgow University, Glasgow, UK
[5] Oxford University, Oxford, UK
[6] LAL, Universite Paris-Sud and CNRS/IN2P3, Orsay, France
[5] Brookhaven National Laboratory, Upton, NY, USA

Corresponding author: Jack.Cranshaw@cern.ch

Abstract. ATLAS maintains a rich corpus of event-by-event information that provides a global view of the billions of events the collaboration has measured or simulated, along with sufficient auxiliary information to navigate to and retrieve data for any event at any production processing stage. This unique resource has been employed for a range of purposes, from monitoring, statistics, anomaly detection, and integrity checking, to event picking, subset selection, and sample extraction. Recent years of data-taking provide a foundation for assessment of how this resource has and has not been used in practice, of the uses for which it should be optimized, of how it should be deployed and provisioned for scalability to future data volumes, and of the areas in which enhancements to functionality would be most valuable.

This paper describes how ATLAS event-level information repositories and selection infrastructure are evolving in light of this experience, and in view of their expected roles both in wide-area event delivery services and in an evolving ATLAS analysis model in which the importance of efficient selective access to data can only grow.

1. Introduction

The ATLAS[1] experiment is a high energy physics experiment which collects data from proton beam collisions at the Large Hadron Collider (LHC) in Geneva, Switzerland. In addition to these collision data, it also generates a corpus of simulation data of comparable size. Together, these allow them to probe the basic nature of the universe. This data is composed of billions of events distributed over one hundred petabytes of raw and derived data. These data come in various formats from the RAW data written as streams of bytes to various forms of object data. The same events may exist in all or only some of these formats. From a very early stage, the ATLAS computing model[2] proposed to use an event-level metadata system to help deal with the scale of this data. This paper will look at the status and outlook for that system.

The LHC is currently undergoing an upgrade to full design center of mass energy and full luminosity. The current period of time is being referred to as the long shutdown 1 or LS1, and the period of previous data taking (2009-2013) will be referred to as Run 1 whereas Run 2 will start when LS1 ends in 2015. Many software upgrades that were not possible during data taking are currently under way, and this forms a proper gap to look at both the past and the future.

The use of event-level metadata at ATLAS has been presented multiple times at previous CHEP conferences, so this paper will refer to those for details[3] and present only a summary here.

1.1. Gathering

Metadata were gathered during the normal ATLAS data processing chain where RAW data are calibrated and reconstructed to produce event summary data (ESD) and analysis object data (AOD). The gathered information formed the TAG data and included

- event identification (including navigational information)
- trigger information and reconstruction flags
- physics quantities from the AOD

These data were produced in chunks by Grid jobs and then integrated into a central database to provide a global view of the data and cross-reference with other metadata systems at ATLAS such as AMI[4], COMA[5], and DDM[6].

1.2. Querying

Once the data were uploaded into the central database, a method was needed to track and query it. A system was developed which allowed us to load the data into multiple Oracle servers at different sites and perform queries on this system which were transparent to the user. This used a catalog of data locations and status, which was also stored in Oracle; and a set of services which parsed and executed the queries. The primary interface to this system was the Event Level Selection Service Interface, or ELSSI. ELSSI was deployed as a php-based web interface. In addition, as regular use cases were identified, web services were deployed for those specific cases.

1.3. Navigating

In addition to querying and analyzing the event-level data, the navigational capabilities could be used to retrieve data for those events. There were two primary use cases: event picking and skimming. Event picking was the case where individual event were requested for debugging, event displays, et. al. Skimming was the creation of a subset of events based on metadata in the TAG.

2. Run 1 Summary

A system which implemented the capabilities described in the introduction was deployed for Run 1 and known as the TAG data services. These services were deployed on servers at CERN and operated primarily in two environments: the Grid[7] and Oracle[8]. When running on the Grid, TAG data stored in ROOT[9] format were used rather than accessing Oracle directly. This system is operational and in use for Run 1 detector data as well as Monte Carlo. Here we'll review some of the successes and problems with the system used for Run 1.

2.1. Run 1 Accomplishments

- Data Gathering
    - A set of algorithms to generate TAG data was part of the standard ATLAS release and could be run with any of the inputs which produced a DataHeader. The metadata content was configurable in blocks maintained by the various expert groups: Data Quality, Trigger, Combined Performance, and others[10].
    - The gathered data was available in files using the standard ATLAS data handling tools as well as loaded into a central database. This allowed us to leverage the work of other groups to distribute these files on the Grid.
- Querying
    - A system for querying TAG data was integrated into the ATLAS framework and improved the framework read speed by loading only the chunks of data containing those events.

  - A system for examining the full ATLAS data store of RAW and AOD data was deployed using TAG data stored in multiple Oracle databases. This system was available through a web interface, which helped users explore the data and develop their data selection[11].
  - The file resident data in ROOT format was used by the data monitoring to provide fast feedback on data quality during 2011 and 2012 data taking.
- Navigating
  - The navigation information was able to provide addresses for files for AOD and RAW data files, and this was well integrated into the ATLAS reconstruction framework, Athena.
  - The ROOT format for TAG data could be added to other forms of ROOT data to add navigability. This allowed some of the data products derived from AOD to act as TAG files which could be used to access AOD or RAW data for events in those post-AOD data products.
  - A system for configuring Grid jobs to access AOD and RAW data was deployed on servers at CERN. These services supported both the event picking and skimming use cases[12][13].

2.2. Issues with the Run 1 system

The system for the collection of TAG data has been stable since 2011, and it is still running for current Monte Carlo production. Nevertheless, experience has indicated some possible improvements. Although storing and transferring the files as datasets allowed us to easily integrate with the standard Grid processing tools, e.g. PanDA; there were several points which required extra steps which introduced a delay between when the AOD data was available and when the TAG data was available to make selections. This delay could be a matter of weeks during reprocessing on the Grid.

Several architectural problems also surfaced over time. Some of these had to do with the manner in which the data were gathered. The TAG data were collected in a single stage during AOD processing. Consequently they did not include information about the increasing number of data products produced afterward. Even without navigational information, this could have provided valuable statistical and accounting information. The time lag between data production and TAG availability caused some users to conduct their skims/selections with other tools to provide sufficiently timely results for detector operations and physics activities. Thus the inability to append data and the timeliness of the availability of that data were obstacles.

There were also problems when using the physics metadata included in the TAG. The metadata was used extensively and successfully for data quality monitoring, but for event selections of AOD and RAW data for physics it had some issues.

- Many of the variables that physicists wanted to use for selections mapped most naturally into variable length array or vectors. Although various relational designs can be used to implement this, we did not find any that would scale for our use cases. This led to using fixed length arrays which many times provided an unacceptable truncation of the data needed for selection as well as being very difficult to query with SQL.
- The data structures within Oracle were kept simple to avoid expensive activities such as joins, but the database schema exhibited poor performance for several sets of common user queries.
- An increasing number of corrections were being applied downstream of the AOD used for TAG production which made the TAG contents imprecise.

There were also effects due to the fact that Run 1 was an extended period of running the ATLAS detector and the ATLAS computing system were all undergoing commissioning. The TAG services were part of the larger ATLAS data processing and metadata framework, and sometimes these services evolved in ways which conflicted with the way that the TAG services were designed. For

example, the ATLAS trigger system requirements evolved over time to support new data taking and Monte Carlo needs which were difficult to support with the existing TAG data structures. Examples of this included using bits outside the design range and running multiple trigger configurations in Monte Carlo. Also, physics users migrated to custom frameworks based on ROOT and custom ntuple formats. A central service based on a single upload could not support these use cases.

Finally there were some places where the design did not scale the way we expected. As the ATLAS data became large, problems developed with sending data from queries on the database to Grid jobs. This forced us to adopt a staged query model where the database effectively provided file lookup, while the actual event navigation was done using the TAG files on the Grid sites. This resulted in a co-location problem where data access required TAG files available locally.

3. Improving support for existing use cases

The TAG services used during Run 1 were undeniably useful, but over time various limitations surfaced. During the period between Run 1 and Run 2, a new project labeled the Event Index[14] has been started to address some of these limitations: data availability, comprehensiveness of the data, extensibility of the data, and better scalability. Two overarching architectural principles have driven the changes being considered

- Separate the event navigation part from the metadata gathering and querying.
- Improve the integration with the data processing to streamline both the gathering and use of event-level metadata.

During Run 1 the event-level metadata was assembled by a set of Athena algorithms, grouped into datasets, and eventually transferred to where it was needed. For Run 2, the Grid or Tier-0 should send the navigational information as messages to a service which accumulates them and puts them into a database. This could be used with any jobs where a process exists which can catalog the events in the output. This addresses both the availability and the comprehensiveness of the data.

Separating the navigation from the metadata gathering is really just the first step in adding extensibility to the data. With the navigation information coupled to an event id, there can multiple sources of metadata which reference that event id. Also, any process can use this navigation if it is capable of producing an event list. Some of the advantages are:

- The navigational data in the database has no required metadata, so it can be uploaded before the metadata is known.
- The metadata does not have to know about the files, nor is it limited to the files available when it was produced.
- Further partitioning of the metadata, for example by arrival times or mutability, is also possible.

The metadata may or may not use the same messaging system used to collect the navigational data.

As noted in the previous section, there were scalability and flexibility problems with the Oracle schemas being used. Since the conception of the TAG database in the early 2000's, new database technologies have been developed, and the Event Index project is investigating Hadoop[15] as an alternative storage technology to address the problems in Run 1. For example, a benefit of this approach is that the schema requirements are relaxed, and one can deploy almost any application to analyze the data as a Map-Reduce job.

The metadata content of the Event Index is still under discussion. Trigger information has been accepted as useful. Physics decisions, such as Higgs candidates, forward W signature, etc. (i.e. offline triggers), will also be supported. There is a discussion of whether physics variables should just be left in the hands of users who can develop processes which produce an event list however they deem best. This list could then be used with the navigational component of the Event Index.

If the metadata is separated from the navigational information, then the metadata storage could also be optimized for a different set of queries. The navigational information is naturally accessed in an event-wise manner. The metadata, on the other hand, could benefit from using the same approach that Google uses to find documents: the inverted index[16]. Rather than storing the metadata event-wise, it

would be stored and accessed by the metadata key, where each key would be associated with a list of events with that key's value. Results are then calculated using fast set operations on the returned event lists. This can even be expedited by keeping sorted lists of events within sub-run datablocks such as luminosity blocks. Luminosity blocks correspond to a number of events of order 10^5.

4. Future Prospects

When event-level metadata was first proposed for the ATLAS computing model, it was viewed as the first step in providing a full abstraction layer between the user who thinks in events and the bulk data storage which worries about files and sites. That abstraction layer was never fully realized. Two places where this could be applied to near term ATLAS objectives are the following items.

- Serving events to processes on multiple processors.
- Reducing the storage footprint by saving pointers to events rather than copying the event data multiple times, i.e. virtual streams.

It's also possible that the data storage in ATLAS could be made 'smarter', where the files are not just data blocks, but incorporate business logic and intelligence, for example,

- indexing
- metadata computations.
- reformatting.

Historically, a large fraction of HEP data has had a limited lifetime during which it is useful and maintained. Results are calculated and published, and then the physicists move on. Recently, there has been a push to preserve this data[17]. Although the most extreme of these proposals would try to preserve the entire data sample, there are categories of preservation, and one of the clearest mandates is for data used in publications. Published plots tend to be highly selective, and something like an event index could be useful for 'tagging' these events. This could have several benefits.

- Sharing the data between analyses within the experiment could be expedited and managed.
- Data for previous analyses could be replotted or reanalyzed with new reconstruction or calibrations or fitting procedures or ...
- It would be easy to export data for particular analyses for archiving or publishing.

The initial concepts discussed here are already under development, such as the event serving and indexing. Most of the others described in this section are either dependent on those initial developments or simply lacking in manpower for now. Experience during Run 2 will determine whether that changes.

5. Conclusion

Event-level metadata at ATLAS has yet to realize its full potential. For Run 2 the TAG project, which provided these capabilities during Run 1, is being replaced by an improved system called the Event Index which is both addressing some of the limitations of the TAG system as well as extending them with new features. Depending on the evolution of the systems at ATLAS and broader data preservation efforts, event-level metadata might also have a role in other activities.

References

[1] ATLAS Collaboration 2008 JINST 3 (2008) S08003.
[2] ATLAS Computing Group 2005 CERN LHCC-2005-022 ISBN 92-9083-250-9
[3] Cranshaw J *et al* 2007 *J. Phys. Conf Ser.* **119** 072012
[4] Albrand S, Doherty T, Fulachier J and Lambert F 2008 *J. Phys.Conf. Ser.* **119** 072003 (10pp)
[5] Gallas E J *et al* 2012 *J. Phys. Conf. Ser.* **396** 052033
[6] Garonne V *et al* 2012 *J. Phys. Conf. Ser.* **396** 032045
[7] Berman F, Fox G, Hey A J (Eds.). 2003 *Grid computing: making the global infrastructure a reality* Vol. 1&2. (Wiley).
[8] http://www.oracle.com

[9] Brun R and Rademakers F 1997 *Nucl. Instrum. Meth.* A **389** 81-86
[10] Assamagan K A *et al* 2006 CERN ATL-SOFT-PUB-2006-002
[11] Viegas F *et al* 2010 *J. Phys. Conf Ser.* **219** 072058
[12] Doherty T *et al* 2012 *J. Phys. Conf Ser.* **396** 052058
[13] Cranshaw J et al 2007 *J. Phys. Conf Ser.* **119** 042008
[14] Barberis D et al 2014 *The ATLAS EventIndex: an event catalogue for experiments collecting large amounts of data* (this publication)
[15] http://wiki.apache.org/hadoop/
[16] Barroso L A, Dean J and Holzle U 2003 *Micro IEEE* 10.1109/MM.2003.1196112
[17] South D M 2012 *J. Phys.: Conf. Ser.* **396** 062018