Methods of Data Popularity Evaluation in the ATLAS Experiment at the LHC

Olga Chuchuk, Thomas Beermann, Alessandro Di Girolamo, Maria Grigorieva, Alexei Klimentov, Mario Lassnig, Markus Schulz, Andrea Sciaba, Eugeny Tretyakov

EPJ Web of Conferences **251**, 02013 (2021) https://doi.org/10.1051/epjconf/202125102013
CHEP 2021

Methods of Data Popularity Evaluation in the ATLAS Experiment at the LHC

Thomas Beermann[5,], *Olga* Chuchuk[6,7,*], *Alessandro* Di Girolamo[7,], *Maria* Grigorieva[1,2,**], *Alexei* Klimentov[3,], *Mario* Lassnig[7,], *Markus* Schulz[7,], *Andrea* Sciaba[7,***], and *Eugeny* Tretyakov[2,4]

[1]Lomonosov Moscow State University, Russian Federation
[2]Plekhanov Russian University of Economics, Russian Federation
[3]Brookhaven National Laboratory, USA
[4]National Research Nuclear University MEPhI, Russian Federation
[5]Bergische Universitaet Wuppertal, Germany
[6]Université Côte d'Azur, France
[7]CERN, Geneva, Switzerland

Abstract. The ATLAS Experiment at the LHC generates petabytes of data that is distributed among 160 computing sites all over the world and is processed continuously by various central production and user analysis tasks. The popularity of data is typically measured as the number of accesses and plays an important role in resolving data management issues: deleting, replicating, moving between tapes, disks and caches. These data management procedures were still carried out in a semi-manual mode and now we have focused our efforts on automating it, making use of the historical knowledge about existing data management strategies. In this study we describe sources of information about data popularity and demonstrate their consistency. Based on the calculated popularity measurements, various distributions were obtained. Auxiliary information about replication and task processing allowed us to evaluate the correspondence between the number of tasks with popular data executed per site and the number of replicas per site. We also examine the popularity of user analysis data that is much less predictable than in the central production and requires more indicators than just the number of accesses.

1 Introduction

A dataset in the ATLAS experiment [1] at the LHC is the aggregation of multiple files in one logical and operational unit in a distributed computing environment. Only datasets, not single files, are transferred between sites and replicated. For this reason, in our study the popularity of data refers to the popularity of datasets.

As data replication and deletion strategies in ATLAS are typically based on manually defined static policies, sooner or later we face situations when popular datasets do not have

*e-mail: olga.chuchuk@cern.ch
**e-mail: maria.grigorieva@cern.ch
***e-mail: andrea.sciaba@cern.ch

enough replicas, or unpopular datasets, on the contrary, take up too much space. These inconsistencies may cause significant delays in tasks execution processes that could have been avoided if the policies were more automated and dynamic.

Methods for assessing data popularity in order to automate data management policies have already been investigated in earlier studies [2, 3], but none of them have been integrated in the production system of the ATLAS computing infrastructure. A possible reason that they are not used is due to insufficient integration of the following sources of information about ATLAS data: DDM (Distributed Data Management) Rucio [4], Rucio Traces [5], EOS [6] Report Logs, WMS (Workload Management System) PanDA [7]. These sources have specific sets of metrics that can be used to assess the popularity of datasets. Our studies have proved that measurements obtained from these four sources are consistent with each other. This means that data on popularity can be integrated without sacrificing the measurement accuracy, but with a gain in the number of consolidated metrics. Additionally, we demonstrate how popularity can be evaluated based on different sources of information and which auxiliary metrics can be calculated for further integration work.

2 Sources for Data Popularity Measurements

There are two general methods for the evaluation of dataset popularity depending of the source of information: how often datasets were requested on the grid, and how many tasks were executed with the datasets as an input within a given time range. Below, we describe the existing data sources that contain metrics needed for the popularity measurements.

2.1 DDM Rucio

DDM Rucio is an open source framework that provides scientific collaborations and individual users with means to organize, manage and access data at any scale. Data can be distributed across distributed data centers around the world. Rucio was originally designed to meet the requirements of the ATLAS experiment. It provides the formation of datasets from files, combining datasets into containers[1] and management of data distribution and replication to the grid.

Rucio Traces

Every data access made on the grid, either through the WMS or directly through the DDM system is tracked by the Tracer system. The trace dictionary is on a file level and includes information like the filename and scope, the corresponding dataset, the endpoint where it was accessed, the type of access, e.g. Analysis or Production upload, the user account, some file information and timings. From the Rucio server the traces are forwarded to a central ActiveMQ broker from where different consumers read this data. One consumer is the Kronos daemon in Rucio. This daemon takes the traces and updates the last access timestamp for files and datasets in the Rucio catalogue, which is used to sort replicas for the deletion. Furthermore, it also updates the *access_cnt* field for files and datasets, which simply counts the number of times when a file or any file in a dataset was accessed.

Rucio API

Rucio API provides detailed metadata about dataset parameters (project, run number, a short description of a physics process, production tags, data format) and information about their

[1]Container is a named set of datasets

replicas, including the placement. The direct popularity metric is the number of accesses (*access_cnt*).

2.2 EOS Report Logs

EOS is one of the storage systems used in the Worldwide LHC Computing Grid (WLCG). In particular, it is deployed at the CERN Data Center and provides a separate instance for each of the large LHC experiments, including ATLAS. EOS Report Logs store detailed information about the system and user file accesses that can be used to better understand the data popularity and life cycle. EOS Report Logs consist of several types of records: file accesses (generated each time a file is opened), file deletions from disks, and file deletions from the metadata space. File accesses records contain the file identifiers, the timestamps of the file opening and closing, the number of bytes read and written, the size of the file before and after each operation, and others (in total, more than 60 metrics). In the study, we are using the records outlined above to obtain the aggregated view of the file life cycle (from its creation until the deletion). We collect and process these log files using the Spark cluster facility provided by the CERN IT Department.

2.3 PanDA Database

PanDA (Production and Distributed Analysis) is a highly scalable and flexible data-driven workload management system that supports central production and user analysis data processing in ATLAS. PanDA DB is a database system serving PanDA. It registers the comprehensive historical and operating meta-information about all physics analysis tasks, jobs being executed within the distributed computing environment of the ATLAS experiment. Additionally, PanDA DB registers metadata about input and output datasets of the computing tasks. Popularity of a dataset in the PanDA DB can be measured as the number of tasks executed within a certain period of time with this dataset as an input, that refers to the number of accesses to the dataset. Besides the popularity of input datasets we can study the execution process of each task, i.e. execution time, time delays, computing sites, user name[2]. This auxiliary information allows us to evaluate the uniformity and efficiency of the resource utilization.

3 Consistency Check of Data Popularity Metrics

In this study we evaluate whether data popularity metrics calculated on various data sources are consistent. It is an important step for further research as we will integrate and join data from multiple sources.

Consistency Check of Data Popularity Metrics Between Rucio Traces and PanDA DB

For this evaluation we collected data (production tasks with input AOD[3] datasets of the Monte-Carlo type) from Rucio Traces and PanDA DB in a defined time period (01/09/2020 - 20/10/2020). Then we generated two data samples with pairs of (Dataset Name, taskID)[4] from Rucio Traces and PanDA DB. The results showed that these two samples are consistent. From PanDA DB we took tasks (with input datasets) that were started or finished within the specified time period, from Rucio Traces we selected tasks with input datasets that were transferred within the same time period.

[2]We are not going to study users statistics. For the popularity measurements we just take the number of different users who executed tasks with ATLAS datasets

[3]Analysis Object Data - these datasets are used for physics analysis tasks

[4]taskID is a task that used the dataset as an input

Consistency Check of Access Metrics between Rucio Traces and EOS

Another consistency check we performed was between the file access records from the EOS logs and the Rucio traces. For that, we selected a specific time interval, limited to October 1, 2020, from 00:00 UTC to 23:59 UTC. All Rucio traces with access times in this period and for files located at CERN were selected; similarly, all file access records from EOS logs in that period and not related to system events were selected, and the two samples were compared.

The comparison showed that 98.8% of the files accessed according to Rucio are accessed also according to EOS, and the remaining 1.2% is almost completely related to traces generated before the files were actually accessed ("direct access") and for which the access never materialized due to a job crash. On the other hand, many more files get accessed according to EOS than to Rucio, as expected due to the fact that EOS contains a significant amount of files not accessed by PanDA jobs.

4 Analysis of the ATLAS EOS instance at CERN Data Center using EOS Report Logs

We analyzed EOS Report Logs for the CERN Data Center for a period of three consecutive months (01/01/2020 - 31/03/2020), which is not a data-taking period and the tasks performed at the site mainly consisted of Monte Carlo production jobs and data analysis. As seen in the description of the EOS Report Logs, most of the storage system activities logs, except for deletions, have the same format. They include all the possible file operations (reads, writes, updates and so on). The original set of log metrics does not indicate the operation type. Nonetheless, having this information is important to understand which operations a file undergoes during its lifetime.

We use the existing metrics (size of the file on opening and closing and the number of bytes read and written) to classify operations into five categories: "Create", "Read", "Update", "Empty" and "Abnormal", where "Empty" operations have no bytes read or written and "Abnormal" are all the operations that cannot be classified as any of the previous categories. The relative impact of the operations other than "Create" and "Read" is very small (less than 1% of all operations), which confirms the assumption that most of the data are immutable.

Figure 1 gives an overview of the read/write processes happening at the EOS instances during the considered three months and shows how actively the provided disk volume was used by ATLAS and two other LHC experiments. The plot indicates that the access patterns differ from one experiment to another.

ATLAS, in comparison to the other experiments, has the biggest EOS instance volume and had the most intense workload. Its total turnover (the sum of all the bytes read and written) is over 480% of the instance volume at the time. During the examined period, all the experiments read more data than they wrote, but not by a large margin. ATLAS had 2-3 times as much read volume as the written one.

We grouped the access records by the file identifiers to obtain file-specific statistics and categorized files based on their creation and deletion times in relation to the boundaries of the considered period. Overall, we have four categories covering all the possible cases:

- Files created and deleted during the period;
- Files created during the period and deleted after or not deleted;
- Files created before and deleted during the period;
- Files created before and not deleted or deleted after the period.

The pie chart in Figure 2 shows the distribution of these categories for the ATLAS EOS instance.

The biggest fraction belongs to the files that were present on disk before and after the considered period (36.8% of total volume). For these files, the lifetime is more than three months. The ATLAS experiment produced and deleted approximately the same volume in this period, as a result, the total occupied volume did not change. A big fraction of created files were also deleted (66.9% of created volume), which indicates that there are a lot of short-lived files with a lifetime of shorter than 3 months. The fraction of files that went through the whole life cycle ("Created and Deleted") is only 31.2%, and, in the future, we plan to extend the time frame in order to increase the relative number of such files.

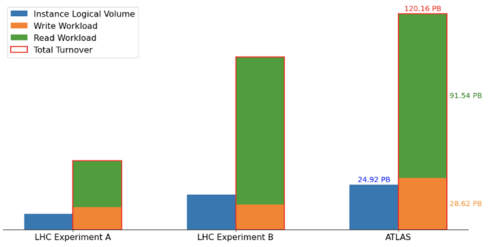


Figure 1: Total workload overview

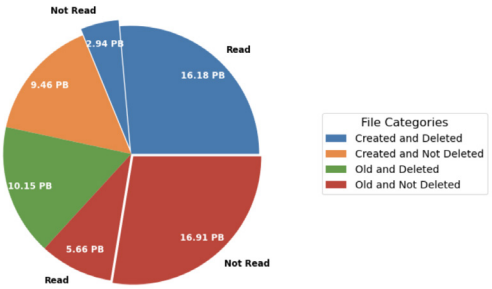


Figure 2: File categories distribution

ATLAS jobs produce a large number of log files and some of them are never read before the deletion. Overall, in the considered period, almost 3 PB of files were not read between their creation and deletion. Moreover, a big fraction of files stayed on disk without being accessed for a long time (see Figure 2). The possibility of keeping these files on a less expensive storage media should be investigated further [8].

While ~85% of the files were <1 GB, most of the occupied volume (~85%) was coming from the files >1 GB, with the average file size ~400 MB. In Figure 3 the "Read Volume" is

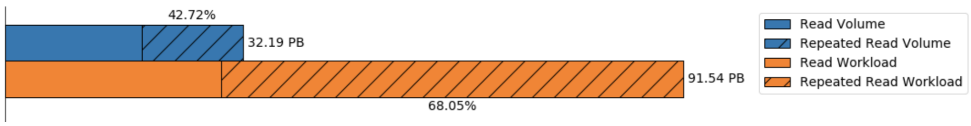
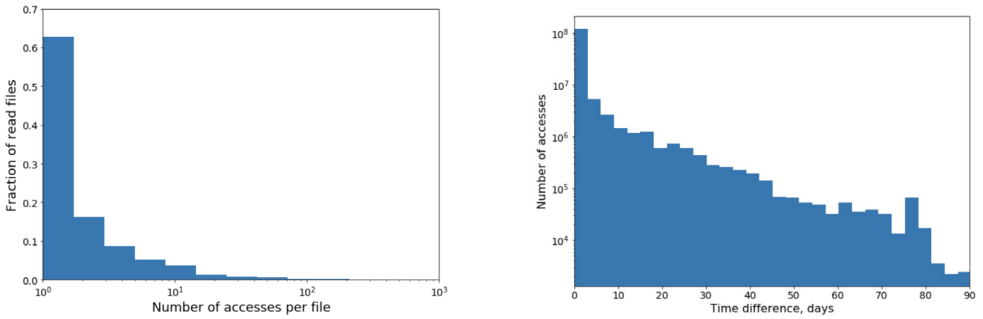


Figure 3: Read workload

the total volume of all the accessed files and the "Read Workload" is the sum of all the bytes read. The hatched parts show the fraction of the volume that was read more than once and the corresponding fraction of the workload. Some read accesses did not read the files fully, though the average read rate per file at the ATLAS instance is 95.84%. As seen in Figure 4a, most files were accessed only once (~63%), and, if a file was re-read, it was most likely to happen within a couple of hours (see Figure 4b).

5 Rucio Access Metrics for Detector and Monte-Carlo Data

As we already saw, Rucio traces provide a very convenient way to record every time that a file has been accessed as input for a job or written as output, or downloaded/uploaded by a user



(a) Number of accesses per file

(b) Time difference between consecutive accesses

Figure 4: Accesses time locality

using the Rucio client tools, and this information is stored indefinitely in a Hadoop cluster at CERN, allowing all sorts of data popularity studies. The most useful attributes in traces are the file name, its size, the dataset name, the remote site, the access time and the access event type. Another important source of information generated by Rucio are the logs, which record events like file transfers and file deletions and they can be used, for example, to study the effectiveness of data management policies.

In the vast majority of cases, ATLAS jobs process files from their local storage, possibly after they have been transferred from a remote site. In some cases, for example for sites without local storage, files will be directly copied from a remote site to the job worker node.

The first approach attempted was to examine file and dataset access patterns to a few representative ATLAS sites belonging to different categories: *Tier-0/1*, *nucleus Tier-2*[5] and *non-nucleus Tier-2*[6]. Only traces related to PanDA jobs were considered, and only files belonging to AOD, DAOD[7] and HITS[8] datasets, as they constitute the vast majority of the data accessed.

Besides the evaluation of the number of traces in this study we decided to investigate how accesses are spread over time or concentrated, that can have a big impact on data management. Additionally, we measured the fraction of weeks with accesses to a given dataset.

The distributions over datasets in Figures 5(a,b) show that DAOD dataset accesses are very few and concentrated in time during a year. Differences in distributions across different sites are small, consistent with the fact that access patterns depend primarily on the file type independently of the location at which datasets are processed. The distributions for the same quantities were derived at a global scale, looking at accesses independently of the site, and very similar distributions where obtained (Figures 5(c,d)) for AOD datasets; it is worth stressing though that only datasets that are accessed at least once contribute to the distributions.

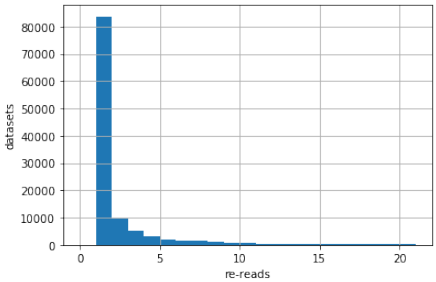
There is therefore a strong indication that ATLAS datasets, even globally, were accessed very few times during a year. Therefore, ideally they would need to be accessible from disk only for a short time. Rucio logs can be used to measure the time between a file's last access and the time it is deleted from a site. Ideally, this time should be as short as possible, but it is in fact of the order of 1-2 months, depending on the site. Another couple of metrics have been

[5]Nucleus sites are sites with larger storage size and better network connectivity that can aggregate work done at other sites

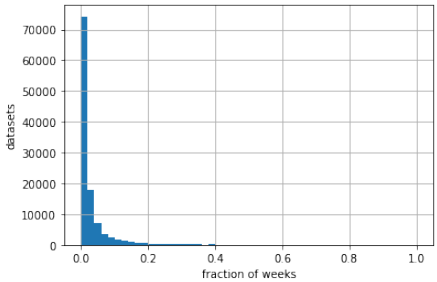
[6]Non-nucleus sites - satellite sites

[7]Derived Analysis Object Data

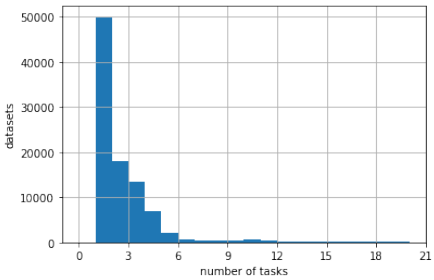
[8]Output of the simulation



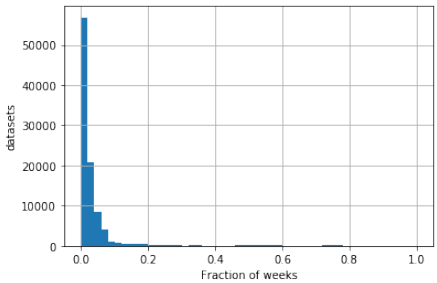
(a) No. of dataset re-reads for DAOD at BNL-ATLAS



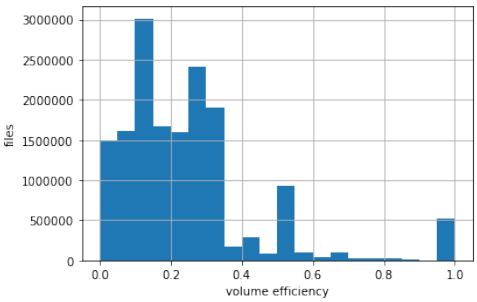
(b) Fraction of weeks with accesses to DAOD at BNL-ATLAS



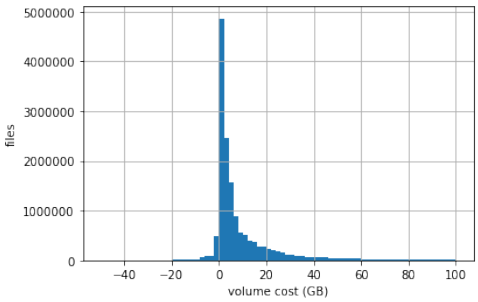
(c) No. of tasks per dataset on all input AOD



(d) Fraction of weeks with accesses to all AOD



(e) Volume Efficiency at CERN-PROD



(f) Volume Cost at CERN-PROD

Figure 5: Dataset access patterns in 2020 year

proposed to quantify how effectively disk storage is used: the volume efficiency V_{eff} and the file storage cost F_C:

$$V_{\text{eff}} = W_{\text{acc}}/W_{\text{disk}}$$
$$F_C = (W_{\text{noacc}} - W_{\text{acc}}) \times S$$

where W_{disk} is the number of weeks the file spends on disk, W_{acc} is the number of weeks when it is accessed, W_{noacc} is the number of weeks on disk and without accesses and S the file size. The distributions for CERN (Figures 5(e,f)) show that files tend to spend a large fraction of their time on disk without being accessed, consistent with the previous results; this observation suggests a reduction of the number of disk replicas as an effective way to reduce disk utilization.

6 Popularity of ATLAS User Analysis Data

Central production tasks are typically planned in advance and executed sequentially as a bunch of tasks, as can be seen at the bottom of the scatter plot in Figure 6. User analysis tasks, on the contrary, can be created and started at any time. Start points of the analysis tasks are almost randomly distributed over time. For this reason we decided to distinguish the popularity analysis of datasets utilized for user tasks. We collected data from two main data sources:

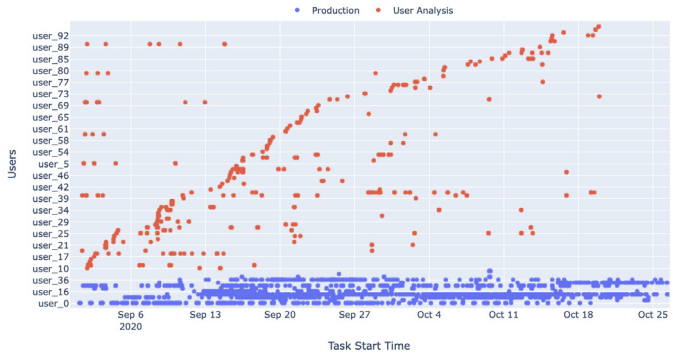


Figure 6: Start time of production and user analysis tasks

PanDA DB and Rucio API, and from two auxiliary souces: CRIC (Computing Resource Information Catalogue) [9] and CERN Phonebook Directory, and placed the integral view of this data into the ElasticSearch[9] storage. Currently this storage structure comprises 54 fields. Each data source provides access to a specific group of metadata for a particular task, dataset, computing center and user. PanDA database is used to access metadata of individual user analysis tasks. These metrics allow us to identify delays in tasks brokerage, datasets usage density, total duration of succeeded and failed tasks and many more. The WLCG infrastructure topology and experiment-specific configurations are described in CRIC. With the help of CRIC we are able to identify task execution hardware specifications and geolocations of computing centers. These metadata are useful for anticipation of task execution time [10]. The CERN Phonebook Directory is exploited to address deep study on user analysis tasks. The information provided is used to identify the user's home institution and its country with geolocation coordinates. This information is potentially important as it allows us to discover institutes and countries that carry out more physics analysis tasks of particular types of ATLAS data. The Rucio API provides us with the detailed information about datasets and replicas.

The popularity is measured with four parameters: *N tasks* - number of tasks, *N users* - number of users who carried out these tasks, *N institutes* - number of users' home institutes, *N countries* - number of countries where the home institutes are. Number of tasks refers to the number of accesses of a dataset. Other user-specific parameters allow us to measure how a datasets usage is distributed around the world, across institutes and among users [11]. Four further metrics allow us to estimate the efficiency of the execution process of datasets and correlate it with popularity metrics: *Replication* factor - average number of dataset replicas, *Computing* factor - average number of computing sites, *Duration* - task execution time from start to finish, *Delay* - task run delay time. These metrics can be applied for datasets and for groups of datasets. The Tables 1 and 2 are built on data aggregated for the

[9]Elasticsearch is a search engine based on the Lucene library

Table 1: Popularity Metrics by Projects [October 2020–January 2021]

Project	N tasks	N users	N institutes	N countries
mc16_13TeV	307 252	678	163	33
data18_13TeV	30 627	291	118	28
data17_13TeV	23 925	256	104	27
mc15_13TeV	17 859	267	104	27
mc12_8TeV	41	2	2	2

Table 2: Task Execution Parameters by Projects [October 2020–January 2021]

Project	Replication	Computing	Duration,sec	Delay,sec
mc16_13TeV	1.54	50	235 249	28 758
data18_13TeV	2.1	59	328 655	27 383
data17_13TeV	2.1	59	367 256	30 170
mc15_13TeV	4.76	40	143 088	19 518
mc12_8TeV	3.38	8	101 847	33 204

period from 01/10/2020 to 24/01/2021 (115 days), grouped by project names[10], and represent popularity and task execution process measurements. The most popular by the number of computing tasks are datasets of the “mc16_13TeV”[11] project. The projects “data18_13TeV”, “data17_13TeV”[12] and "mc15_13TeV"[13] are less poopular by the number of tasks, but still popular among users from different institutes and countries as expected. Computing factor is correlated with dataset popularity: more popular datasets are processed by far more computing sites than unpopular. Replication factor for popular datasets fluctuated from 1.5 to 4.7, but at the bottom of the table we can observe that non popular datasets have 3-5 replicas, that is clearly excessive. Delay time in most cases depends on the replication factor: it decreases with the increasing of the number of dataset replicas. For example, the diagram shows that datasets belonging to the project “mc15_13TeV”, that are not popular, have 4.76 replication factor and, consequently, not long execution time (Duration = 143088 s) and delay (19518s). The most popular project “mc16_13TeV” has a lower rate of replication (1.54), and the highest delay time (28758). We aim to accelerate the processing of the most popular datasets and to reduce delays of tasks execution. Number of dataset replicas and computing sites can affect these indicators directly or indirectly.

7 Conclusion

At each step throughout the processing and data management stack, Rucio and PanDA keep detailed logs on the operations on files, tasks and datasets. By combining this information with the access records produced by the EOS file system, CRIC and the CERN Directory, it is now possible to understand the complete life cycle of data from the global distribution down to the fraction of files that are read, covering managed and individual use. While it is highly desirable

[10]Project identifies the particular physics or computing context of a set of datasets
[11]Monte Carlo production at 13 TeV in 2016 year
[12]Real data at 13TeV in 2017-2018 years
[13]Monte Carlo production at 13 TeV in 2015 year

to expand the work to cover longer time periods, additional storage systems and more details, this level of insight will serve nevertheless as the foundation on which further analysis can be based to optimise the management and usage of storage. In addition, the concepts developed are sufficiently general that they can be adapted with relative ease for wide use beyond ATLAS. The main outcome of the studies covering AOD and DAOD datasets is that the vast majority of data is accessed only a few times, that most accesses take place within short intervals, and that a significant fraction of the data occupies disk space for extended periods without seeing active use. These findings, while still preliminary and not necessarily representative for all phases of the experiment, indicate that the usage of caches, more aggressive data deletion policies and changes between different levels of quality of service have the potential to optimise the overall cost of storage. However, before gains can be realised in practice, more work is required. Notably this would be in the continuous monitoring of data popularity, and the development and tracking of expressive metrics that can be used to guide users and site managers towards a more efficient use of resources.

Acknowledgements

We thank our ATLAS Distributed Computing colleagues, ATLAS sites, Tier-1 ATLAS centers, CERN Tier-0 operations. The work at Plekhanov University (Section 6) is funded by the Russian Science Foundation grant (project No.19-71-30008). The work at Brookhaven National Laboratory is funded in part by the U.S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing contracts.

References

[1] ATLAS collaboration, JINST **3**, S08003 (2008)
[2] A. Molfetas et al., Journal of Physics: Conference Series **331**, 062018 (2011)
[3] T. Maeno, K. De, S. Panitkin, Journal of Physics: Conference Series **396**, 32070 (2012)
[4] J. Elmsheuser, A. Di Girolamo (ATLAS), EPJ Web Conf. **214**, 03010 (2019)
[5] M.S. Barisits, Ph.D. thesis, Vienna, Tech. U. (2017)
[6] L. Mascetti et al., J. Phys. Conf. Ser. **664**, 042035 (2015)
[7] A. Klimentov, K. De, S. Jha, T. Maeno, P. Nilsson, D. Oleynik, S. Panitkin, J. Wells, T. Wenaus, J. Phys. Conf. Ser. **762**, 012021 (2016)
[8] M. Barisits, M. Borodin, A. Di Girolamo, J. Elmsheuser, D. Golubkov, A. Klimentov, M. Lassnig, T. Maeno, R. Walker, X. Zhao, EPJ Web Conf. **245**, 04035 (2020)
[9] A. Anisenkov, J. Andreeva, A. Di Girolamo, P. Paparrigopoulos, B. Vasilev, EPJ Web Conf. **245**, 03032 (2020)
[10] E. Antonov, B. Onykiy, A. Artamonov, E. Tretyakov, International Journal of Technology **11**, 1125 (2020)
[11] M. Grigorieva, E. Tretyakov, A. Klimentov, D. Golubkov, T. Korchuganova, A. Alekseev, A. Artamonov, T. Galkin, *High Energy Physics Data Popularity : ATLAS Datasets Popularity Case Study*, in *2020 Ivannikov Memorial Workshop (IVMEM)* (2020), pp. 22–28