A Probabilistic Analysis of Data Popularity in ATLAS Data Caching

Computing in High Energy and Nuclear Physics (CHEP) 2012, 21-25 May 2012, New York City, NY, USA

Mikhail Titov[1], Gergely Záruba[1], Alexei Klimentov[2], Kaushik De[1] for the ATLAS Collaboration

[1] The University of Texas at Arlington, [2] Brookhaven National Laboratory

ABSTRACT

Efficient distribution of physics data over ATLAS grid sites is one of the most important tasks for user data processing. The initial static data distribution model (Planned Data Placement) has been shown to create a bottleneck in data processing, where the availability of popular datasets had a profound effect on the utilization of the specific computational resources and resulted in uneven loads; this was caused by over-replication of unpopular data that have filled up disk storage space while under-utilizing of some processing resources due to low numbers of replicas of popular (desired) data. Thus, a new data distribution mechanism was implemented within the production and distributed analysis system PanDA; PD2P (PanDA Dynamic Data Placement) dynamically reacts to user data needs [1], basing dataset distribution principally on user demand. Data deletion is also demand driven (by the Replica Reduction Agent), reducing the numbers of replicas for unpopular data [2]. This dynamic model has led to substantial improvements in efficient utilization of storage and processing resources [3].

Based on the above experience, in this work we seek to further improve the data placement policy by investigating in detail how data popularity can be obtained and its impact on prediction of data popularity in the future. It is necessary to precisely define what data popularity means, what types of data popularity exist, how it can be measured, and most importantly, how the history of the data can help to predict the popularity of derived data. We introduce *locality of the popularity:* a dataset may be only of local interest to the specific user or may have a wide (global) interest. We also extend the idea of the *"data temperature scale"* model and a popularity measure. The history data of PanDA jobs including both production and analysis is used for our analysis of the behavior of data usage.

INTRODUCTION

One of the possible ways to predict dataset popularity is by using classical probability theory. This approach is defined in the first stage of the "Two-stage replica replacement algorithm" [4]. More precisely, the prediction of the replica value is *"the prediction of the number of times that a replica corresponding to the identifier f would be accessed in the next n requests, based on the previous m requests (in a fixed future time window):"*

$$V(f, m, n) = \sum_{i=1}^{n} Pi(f)$$

where *Pi(f)* is the probability of receiving a request for file *f* at request *i.*

However, this approach is not applicable when looking at job statistics (all jobs including production and analysis) obtained from PanDA for the year 2011. Figure 1 shows the decreasing trend (in popularity) in the number of successful jobs for two types of inputs, where inputs are: i.) datasets before reprocessing; ii.) datasets after reprocessing. These datasets were produced from the RAW dataset "data11_7TeV.00184130.physics_JetTauEtmiss.merge.RAW" that was the most popular among RAW datasets based on information from *DQ2 Popularity* [2] on September 2011. The obtained data will help to track the behavior of derived datasets usage.

The summary of number of jobs (characteristics of plots from the figure 1):

- ESD datasets - # jobs (with non-reprocessed datasets) = 22 731, # jobs (with reprocessed datasets, configuration tag is "r2603") = 18 756 (82.5%)
- AOD datasets - # jobs (with non-reprocessed datasets) = 21 726, # jobs (with reprocessed datasets, configuration tags are "r2603_p659") = 15 492 (71.3%)
- NTUP datasets - # jobs (with non-reprocessed datasets) = 18 376, # jobs (with datasets that are derived from reprocessed datasets) = 8 360 (45.5%)

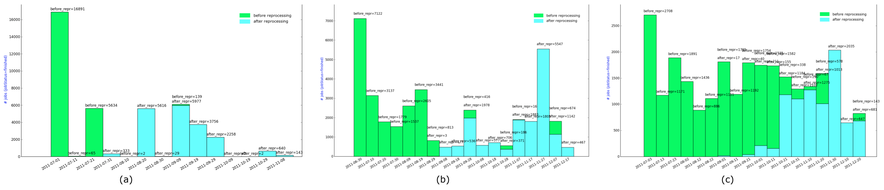

Figure 1. Number of successful jobs with dataset inputs before and after reprocessing (data11_7TeV.00184130.physics_JetTauEtmiss): (a) ESD datasets; (b) AOD datasets; (c) NTUP datasets

Table 1. Ratio between jobs with reprocessed datasets and non-reprocessed datasets (all campaigns during 2011)

Data type	Job final state	Ratio
ESD	finished	0.23 (23%)
	failed / cancelled	0.27 (27%)
AOD	finished	0.69 (69%)
	failed / cancelled	0.43 (43%)
NTUP	finished	0.38 (38%)
	failed / cancelled	0.35 (35%)

RESEARCH QUESTIONS & METHODS

There are three main sets of entities that have to be considered for data replication (all other parameters are defined as attributes for objects from these sets): i.) set of datasets $\{D_i\}$ (the term "dataset" is used for the job input object; in terms of PanDA job - an input can be either dataset that is a unit for data replication in ATLAS Data Distribution Management system, container that consists of datasets, dataset pattern, container pattern, or group of datasets / containers / patterns that are coma-separated), ii.) users $\{U_j\}$ (grid identifiers of PanDA job owners), iii.) sites $\{S_k\}$ (the term "site" is used for any storage or computing element). The main idea is to precisely identify all interactions between only dataset objects and between all the above three sets of objects (see figure 2). More precisely, dataset objects can interact with dataset objects; users can be both on the receiving and initiating end of interactions with dataset objects; similarly, users interact both ways with sites; however, site objects should not directly interact with dataset objects. This is what we call the *user-oriented approach*.

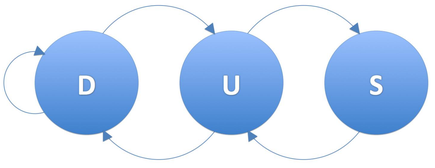

Figure 2. The interaction model between basic objects: D – Datasets; U – Users; S – Sites

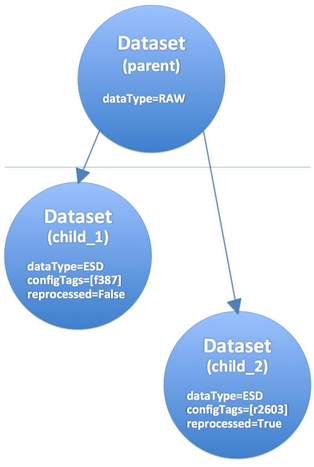

Figure 3. The Model of Datasets Interactions (MDI)

The definition of popularity is based on the popularity of certain datasets among users but not among sites; we propose a user-oriented approach for determining popularity. Future location of data requested by a user must be user- but not site-dependent. The site is chosen based on its cost and its popularity for a particular user (as the main parameters for the site definition for data caching).

Local Popularity of dataset D for user U: $D_{LP} = \frac{J_{DU}}{J_D}$ where J_{DU} is the number of jobs with the input dataset D that are submitted by the user U, J_D is the total number of jobs with input dataset D

Global Popularity of dataset D: $D_{GP} = \frac{U_D}{U} \times \frac{J_D}{J}$ where U_D is the number of users who submitted jobs with the dataset D, U is the total number of users, J_D is the total number of jobs with the input dataset D, J is the total number of jobs

For each user, weights are assigned for all sites based on two components: cost and popularity. These components can be represented as follows:

$$S_{cost} = \frac{FJ_{US}}{SJ_{US} + FJ_{US}} = \frac{FJ_{US}}{J_{US}}$$

$$S_{popularity} = \frac{J_{US}}{J_S}$$

where SJ_{US} is the number of successful jobs that are submitted by the user U at the site S, FJ_{US} is the number of failure jobs that are submitted by the user U at the site S, J_{US} is the total number of jobs that are submitted by the user U at the site S, J_S is the total number of jobs at the site S

Popularity measure is one of the basic parameters that is used for the definition of the *Dataset Object* in the *Model of Datasets Interactions (MDI).* The *MDI* is a directed acyclic graph (see figure 3) that shows interactions between dataset objects: how a particular dataset object relates to other dataset objects.

The *Dataset Object* is a vertex in the *MDI* with the following parameters: i.) the reference to the parent dataset object; ii.) the list of ancestors; iii.) the global popularity and list of users with corresponding local popularity values; iv.) a list of values containing a similarity metric between the new dataset object and previous dataset objects with the same parent dataset object (shows the percentage of how many jobs were reproduced with the new dataset object as input data).

The *Object Link* is an edge in the *MDI* with the following parameters: i.) references to the ancestor and the descendant dataset objects; ii.) the list of transformation attributes that leads from parents to children; iii.) the similarity metric between parent and child dataset objects (shows the percentage of how many jobs were reproduced with the child object as input data), i.e., the weight of the link.

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest [5]. All previously defined parameters are employed as corresponding conditions for a Bayesian network based model, which will help in calculating corresponding probabilities for data popularity.

ATLAS data can be classified into one of the following categories based on their importance and popularity (i.e., the *Data Temperature Scale*): {hot, warm, cold, frozen, obsolete}. Each state has its own set of conditions; this categorization is utilized in our data popularity model. Probabilities for each possible data state can be estimated based on Bayesian network model and used in ATLAS Data Caching as a control parameter for the replication of new datasets.

CONCLUSION

In this study we have investigated the temporal behavior of user access to datasets. Data in our analysis have been restructured and stored in a NoSQL MongoDB database that provided an advantage in retrieving and further processing of statistic information. Experiments for analyzing the behavior of the data popularity have shown that classical probabilistic models are unfit to handle the interactions among our variables. We believe that the methodology of Bayesian networks is a key for the definition of the probability of data popularity, and further research work will be focused on modeling data placement for PanDA jobs using Bayesian networks.

REFERENCES

1. The ATLAS Collaboration, Tadashi Maeno, Kaushik De, Sergey Panitkin, "PD2P : PanDA Dynamic Data Placement for ATLAS", CHEP, New York City, NY, USA, 21-25 May 2012
2. Angelos Molfetas, Fernando Barreiro Megino, Andrii Tykhonov, Vincent Garonne, Simone Campana, Mario Lassnig, Martin Barisits, Gancho Dimitrov, Florbela Tique Aires Viegas, "Popularity framework to process dataset tracers and its application on dynamic replica reduction in the ATLAS experiment", CHEP, Taipei, Taiwan, 18-22 October 2010
3. Alexei Klimentov, "ATLAS Distributed Computing Challenges and Plans for the Future", CERN Document Server, 18 September 2011
4. Tian Tian, Junzhou Luo, "A Prediction-based Two-Stage Replica Replacement Algorithm", the 11th International Conference on CSCWD, Melbourne, Australia, 26-28 April 2007
5. David Heckerman, "Bayesian Networks for Data Mining", Data Mining and Knowledge Discovery Journal, 1997

ACKNOWLEDGMENT

We would like to thank the PanDA team for providing the data for our analysis and for their continued support. We would like to express our appreciations to both CERN and UTA for the facilities provided to conduct this research work. Finally, we would like to thank I. Ueda for all the time he has spent to make this poster better.