Distributed Data Replication and Access Optimization for LHCb Storage System
A Position Paper

Mikhail Hushchyn[1,2,3], Philippe Charpentier[4] and Andrey Ustyuzhanin[1,2,3,5,6]

[1] *Yandex School of Data Analysis, Moscow, Russian Federation*
[2] *Yandex Data Factory, Moscow, Russian Federation*
[3] *Moscow Institute of Physics and Technology, Moscow, Russian Federation*
[4] *CERN, Geneva, Switzerland*
[5] *National Research University Higher School of Economics (HSE), Moscow, Russian Federation*
[6] *NRC Kurchatov Institute, Moscow, Russian Federation*

Keywords: Structured Data Analysis and Statistical Methods, Machine Learning, Information Extraction, Hybrid Data Storage Systems, Data Management, LHCb.

Abstract: This paper presents how machine learning algorithms and methods of statistics can be implemented to data management in hybrid data storage systems. Basicly, two different storage types are used to store data in the hybrid data storage systems. Keeping rarely used data on cheap and slow storages of type one and often used data on fast and expensive storages of type two helps to achieve optimal performance/cost ratio for the system. We use classification algorithms to estimate probability that the data will often used in future. Then, using the risks analysis we define where the data should be stored. We show how to estimate optimal number of replicas of the data using regression algorithms and Hidden Markov Model. Based on the probability, risks and the optimal nuber of data replicas our system finds optimal data distribution in the hybrid data storage system. We present the results of simulation of our method for LHCb hybrid data storage.

1 INTRODUCTION

The hybrid data storage system uses the two data storage types to store the data. The first type is relatively cheap kinds of the data storages such as magnetic tapes or HDD. Usually, the most of the data is kept on the first type of the storages. The second type is more expensive one which have high speed of data input/output in comparison with the first type. SSD is an example of the storage of the second kind. Using the second type storages helps to increase the speed of data access. However, the space of the storages of the second type is very limited, so it is highly important to estimate the data should be kept on the second type storages.

This study is useful not only for the LHCb, but for cloud providers too.

The LHCb collaboration is one of the four major experiments at the Large Hadron Collider at CERN. The detector, as well as the Monte Carlo simulations of physics events, create 15 000 PB of data every year.

The LHCb data storage system is a hybrid one. The data is kept on disk and tape storage systems. Disks are used for storing data used by physicists for analysis. They are much faster than tapes, but are way more expensive and hence disk space is limited. Therefore it is highly important to identify which datasets should be kept on disk and which ones should only be kept as archives on tape.

We use dataset access history of the LHCb data storage system for the last two years. Each time series of the access history consist of 104 points. Value of the each point is a number of accesses to a dataset for one week.

Based on the machine learning algorithms and methods of statistics we develop system for data storage management for hybrid storage systems and demonstrate the system work on the LHCb data.

Hushchyn, M., Charpentier, P. and Ustyuzhanin, A..
Distributed Data Replication and Access Optimization for LHCb Storage System - A Position Paper.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 537-540
ISBN: 978-989-758-158-8

2 RELATED WORKS

Implementation of the machine learning algorithms for data management was described in other papers. The Markov chains were used to predict the datasets popularities in a Data Management System for hybrid HDD + SSD data storage system (Lipeng, 2014). Then, the authors used the popularities and parameters of the storage system to solve data placement optimization problem.

Artificial neural networks were used to predict possible dataset accesses in near-term future in *A Popularity-Based Prediction and Data Redistribution Tool for the ATLAS Distributed Data Managemen* paper (Beermann, 2014).

The methods and the system presented in this paper is evolution of the *Disk storage management for LHCb based on Data Popularity estimator* which described in (Hushchyn, 2015). A feature of the current work is that the method and the system is suitable for any hybrid system, not only for LHCb one.

3 METHOD

The system has three separate modules. The first module predict probabilities that the datasets will be often used or to be popular in future.

The second module is used to predict number of accesses to the datasets. The datasets access history, regression algorithms and Hidden Markov Model are used for the prediction.

Based on the predicted probabilities and number of accesses the third module estimate the optimal data distribution over hybrid data storage system. The module uses the risks analysis and loss function optimization to find the optimal data distributions. The risks matrices and the loss functions represent the requirements to the data distribution.

3.1 Inputs

In this study we use only the dataset access history without any additional information about datasets. This approach allows to use our methods and system in any hybrid data storage system.

3.2 The Probabilities Prediction

Classification algorithm is used to predict dataset probabilities to be often used on a forecast horizont of N time periods (hours, weeks, months). A data manager defines datasets which are often used. For example, datasets which have more than zero number of accesses during N periods are often used for LHCb. Then, suppose that the dataset access history's time series have M time periods. Last N time periods are used to label the time series. The time series which are rarely used during the last N time periods are labeled as 0. The most popular datasets are labeled as 1. Then, $[0, M-N]$ time periods are used to train the classifier. On this study we use Gradient Tree Boosting Classifier.

The trained classifier are used to predict the probabilities to be often used or to be popular for the future N time periods. For the prediction the $[N, M]$ weeks are used. We use the area under the ROC curve and the cross-validation to measure the classification quality.

This approach demonstrate better results than regression algorithms and algorithms of time series analysis. Also, this method shows good results for time series with lack of statistics. For example, for LHCb data storage system the classification roc auc is 0.89 on train data. The figure 1 shows the classification ROC curve for LHCb data.

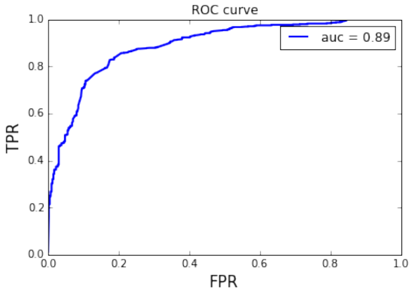

Figure 1: The classification ROC curve for LHCb data.

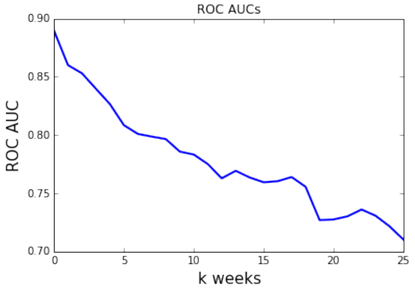

Figure 2: The classification quality for different future weeks for LHCb data.

Moreover, the classifier demonstrates that earlier time periods have smaller impact to the prediction. On example, which is shown on figure 2, $[0, 52]$ weeks were used to train the classifier. Then, the classifier was used to predict the probabilities for $[53, 78]$, ..., $[53+k, 78+k]$, ..., $[79, 104]$ weeks.

3.3 Data Distribution

The predicted popularities are used to estimate where the datasets should be kept: on type one storages or on type two. For this purpose we use the risks analysis. We use a risk matrix to calculate the total risks of the decision where to store the data. For example, the following matrix is used for two-types hybrid data storage systems:

$$M = \begin{bmatrix} C_{00} & C_{10} \\ C_{01} & C_{11} \end{bmatrix} \tag{1}$$

C_{00} - fine for the decision to keep a dataset on type 1 storage, when it should kept on the type 1 storage,

C_{01} - fine for the decision to keep a dataset on type 1 storage, when it should kept on the type 2 storage,

C_{10} - fine for the decision to keep a dataset on type 2 storage, when it should kept on the type 1 storage,

C_{11} - fine for the decision to keep a dataset on type 2 storage, when it should kept on the type 2 storage.

The predicted probabilities are probabilities that the datasets should be stored on type 1 storages (rarely used data) or on type 2 storages (often used data). Therefore, multiplying the risks matrix and the predicted probabilities for the each decision we estimate the total risks for the decisions:

$$R = \begin{bmatrix} P_0 & P_1 \end{bmatrix} \begin{bmatrix} C_{00} & C_{10} \\ C_{01} & C_{11} \end{bmatrix} = \begin{bmatrix} R_0 & R_1 \end{bmatrix} \tag{2}$$

P_0 - the predicted probability that a dataset will be rarely used,

P_1 - the predicted probability that the dataset will be often used,

R_0 - the total risk for the decision to keep the dataset on type 1 storage,

R_1 - the total risk for the decision to keep the dataset on type 2 storage.

3.4 Optimal Number of Replicas

As described in paper (Hushchyn, 2015) we use *Nadaraya-Watson kernel smoothing* algorithm and *Leave-One-Out* method for smoothing window width optimization to predict dataset future number of accesses. In some time series the Hidden Markov Model (HMM) demonstrates the better prediction results. Therefore, our system provides the HMM algorithm for the prediction of the number of accesses. Currently, we are developing the implementation of the HMM for the prediction.

Then, the predicted number of accesses are used to estimate the optimal number of replicas for the datasets:

$$Rp_{opti} = F(I) \tag{3}$$

Rp_{opti} - the optimal number of replicas for a dataset,

I - the predicted number of accesses for the dataset,

$F()$ - function for the optimal number of replicas.

Linear, quadratic or exponential functions are examples of the function $F()$.

The predicted number of access and the optimal number of replicas are helpful for optimal usage of the storage system. Moreover, the predicted number of accesses can be used to detect the datasets for which the number of replicas should be reduces to free additional space on the storage.

For LHCb the following function is used to estimate the optimal number of replicas:

$$Rp_{opti} = \sqrt{\alpha I} \tag{4}$$

α - the free parameter.

The figure 3 shows how the optimal number of replicas for a dataset depends on its predicted number of accesses and alpha value. For example, suppose the predicted number of accesses for a dataset is $I = 10$ accesses per week and $\alpha = 0.5$. Then $Rp_{optimal} = \sqrt{\alpha I} = \sqrt{0.5 * 10} = 2.24 \approx 2$ replicas.

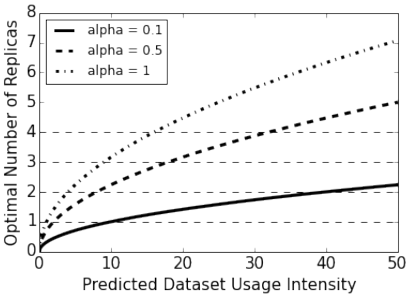

Figure 3: Dependence of optimal number of dataset replicas (Rp) from its predicted number of accesses (I) and α.

4 RESULTS

Based on the method which is described above we create the python library. This library contains tools for the probability prediction, the optimal data distribution and optimal number of replicas estimation. Moreover, the library allows to do simulation of work of our method based on the data usage history.

Also, it will be possible to use our method as webservice with docker (Docker).

Now we are developing the probability prediction method decribed above to get higher quality of the classification.

Special datasets metadata, dataset access history several features that were calculated using the access history were used as inputs in (Hushchyn, 2015).

In this study we use just accesses history. It makes possible to use the method presented here not only for LHCb but for other hybrid data storage systems. Moreover, we use risks analysis instead of the loss function optimization in (Hushchyn, 2015). This adds flexibility to data distribution requirements and allows to use our methods for hybrid data storage systems with more than two kinds of storages.

The LHCb data which was used in (Hushchyn, 2015) contains information of 7375 datasets. The method described in (Hushchyn, 2015) allows to save about 40% of disk space and makes just 8 mistakes (wrong removings of the datasets from disk). At the conference we demonstrate how much disk space can be saved using the method presented here and how much mistakes this method makes.

The method presented in (Hushchyn, 2015) and development of the study described in this paper can be viewed on (Datapop). Our study is performed by means of a Reproducible Experiment Platform (Rep) - environment for conducting data-driven research in a consistent and reproducible way.

5 CONCLUSION

In this paper we describe the key points of our method for data management in hybrid storage systems. At the conference we demonstrate how much disk space can be saved using the method presented here and how much mistakes this method makes in comparison with the method from (Hushchyn, 2015).

REFERENCES

Lipeng W, Zheng L, Qing C, Feiyi W, Sarp O, Bradley S. (2014) *30th Symposium on Mass Storage Systems and Technologies (MSST): SSD-optimized workload placement with adaptive learning and classification in HPC environments.* California. IEEE.

Beermann T., Stewart A., Maettig P. (2014) *The International Symposium on Grids and Clouds (ISGC) 2014: A Popularity-Based Prediction and Data Redistribution Tool for ATLAS Distributed Data Management.* PoS. p 4.

Hushchyn M., Charpentier P., Ustyuzhanin A. (2015) *The 21st International Conference on Computing in High Energy and Nuclear Physics: Disk storage management for LHCb based on Data Popularity estimator.* http://cds.cern.ch/record/2022203/files/LHCb-PROC-2015-019.pdf

https://www.docker.com

https://github.com/yandexdataschool/DataPopularity

https://github.com/yandex/rep