# Half-time report: Toward Efficient Data Management in Hierarchical Storage using Reinforcement Learning

Tianru Zhang *

Department of Information Technology, Uppsala university

**Abstract**

With the rapid development of big data and cloud computing, data management has become increasingly challenging. Over the years, a number of frameworks for data management have become available. Most of them are highly effective, but ultimately create data silos. It becomes difficult to move and work coherently with data as new requirements emerge. A possible solution is to use an intelligent hierarchical (multi-tier) storage system (HSS). A HSS is a meta solution that consists of different storage mediums organized as a jointly constructed storage pool. The placement of data in the hierarchy is determined by the data migration policy. It is a non-trivial task since it should be made according to the characteristics of the dataset, the tier status in a hierarchy, and access patterns. To address these challenges, we have introduced an open-source hierarchical storage framework with a dynamic migration policy based on reinforcement learning (HSM-RL) [1]. We presented a mathematical model, a software architecture, implementations based on both simulations and a live cloud-based environment, and experiments comparing the proposed RL-based strategy to three rule-based policies. Nevertheless, in many areas of data-driven science, large datasets are generated where the individual data objects can be information-sparse. A challenge is to efficiently find and work with the most interesting data as early as possible in an analysis pipeline. We dedicated to an application area of high-throughput and high-content imaging experiments, where large amounts of microscopy images are acquired and analyzed under varying experimental conditions. We associated the HSM-RL framework with the data-specific internal structure, as well as access patterns. Based on two relevant imaging datasets, Protein translocation and Phenotypic screening, we conducted experiments showing that our framework is superior and can quickly adapt to new data access requirements. It overall reduces the data processing time, and the proposed autonomous data placement strategy is superior compared to other static or semi-static data placement strategies.

## 1 Introduction

According to available statistics, the overall amount of data created worldwide had reached 79 zettabytes in 2021, and current estimates are that by 2025 the total size will be 125 zettabytes [2]. This data explosion brings numerous big data challenges for all industries, academia, and even for individuals. Large-scale data management is one of the most challenging prospects in the big data domain. Recent advancements in the field of large-scale data management range from effective interconnected storage devices to smart algorithms for efficient management at scale. These advancements have significantly enhanced storage capacity, management, and high-availability of large datasets. The continuous generation of new datasets brings new requirements that need further research and development. These new requirements not only bring conventional challenges to a much larger scale, but also add new challenges related to recently developed solutions. Conventional challenges include efficiency, scalability, and high-throughput. New emerging challenges include data management based on different storage types (volume-based or object-based solutions), pay-as-you-go models offered by service providers, and solutions that are compliant with new data privacy and security regulations. It is evident that these multi-fold challenges require effective and comprehensive solutions.

The best data management strategy depends on the type of data, access patterns, usability, and long- and short-term availability needs for the data. To fulfill these requirements, we have centralized and distributed file systems that are optimized for different hardware (e.g. Cache, SSD, and HDD) settings. There are also tape storage solutions and object stores for long-term storage. These solutions have different characteristics and are suitable for different storage requirements. However, it has been repeatedly reported that static availability of datasets

*tianru.zhang@it.uu.se

causes issues related to performance and availability. Meanwhile, data access patterns change over time, creating varying access patterns. Thus, there is no single storage solution that can address these dynamic requirements.

## 2 Hierarchical Storage and management using RL

To address the challenges mentioned above, the concept of hierarchical storage provides an overarching approach. The main idea is to connect different independent storage solutions and move the data between them according to criteria designed to meet a set of requirements. The hierarchies based on different underlying storage solutions ensure rational allocation of resources and therefore solve the problem of inefficiency due to imbalanced resource usage in traditional storage solutions. However, the actual realization of the concept is a non-trivial task. In one of our recent articles [1], we have comprehensively described the underlying challenges related to storage hierarchies and presented a solution based on reinforcement learning (RL). We have also built a simulation and a fully functional cloud-based framework for general purpose datasets. We have conducted experiments based on static and dynamic datasets to compare our RL-based solution with three realistic rule-based schemes in different scenarios. The results have proved the effectiveness, efficiency and consistency of the RL-based policy. Moreover, we have also shown the ability of the proposed hierarchical storage management framework in terms of efficient data placement with minimum resource utilization.

## 3 HSM-RL with scientific datasets

Different datasets have different characteristics that can play important roles in large-scale data management. These characteristics include total dataset size, individual object size, access patterns, and regulations. Scientific datasets are unique as the individual objects can be information-spare, such as in the case of automated microscopy. The information-sparsity in scientific datasets plays an important role in efficient data analysis. Here it is important to note that the meaning of sparsity varies with the kind of analysis under consideration. Faster access to the most relevant data would significantly improve the efficiency of analysis. In the second article [3], we focused on highlighting the unique characteristics of the scientific datasets, capitalizing on the underlying structure to efficiently build the storage hierarchies, and reduce the time and cost associated with the analysis pipelines. We have called those special characteristics the *interestingness values*, which can be a single value, a vector or a matrix that highlights the importance of an individual object in a dataset. We merged the *interestingness values* into our previous proposed HSM-RL framework by designing new RL state variables related to the *interestingness values*. With the specific design, our focus is to study the framework's capabilities to efficiently and autonomously manage scientific datasets and improve the analysis time by providing faster access to the datasets.

In order to highlight the capabilities of our framework, we performed experiments on two scientific imaging datasets: Protein translocation and Phenotypic screening. We also used other common policies such as Least Recently Used (LRU) replacement, Least Frequently Used (LFU) replacement, and Maximum feature policy as a comparison. Our results clearly illustrate that once the data is available in the system, the framework efficiently organizes the active/required subset of the data to have faster access while keeping the inactive data on the slower or less expensive tiers in the hierarchy. The framework achieves this behavior autonomously and also adapts accordingly to new emerging data access requirements. This makes it uniquely suitable for exploratory scientific analysis.

## References


[1] Tianru Zhang, Andreas Hellander, and Salman Toor. "Efficient Hierarchical Storage Management Empowered by Reinforcement Learning". In: *IEEE Transactions on Knowledge and Data Engineering* (2022), pp. 1–1. DOI: 10.1109/TKDE.2022.3176753 (cit. on pp. 1, 2).

[2] *Statista, Big data - Statistics Facts.* https://www.statista.com/topics/1464/big-data/#dossierKeyfigures (cit. on p. 1).

[3] Tianru Zhang et al. "Management of Scientific Datasets in Hierarchical Storage Using Reinforcement Learning". In: *Submitted to IEEE International Conference on Big Data 2022* (2022), pp. 1–1 (cit. on p. 2).