Optimizing Data Migration Using Machine Learning

Aijun An

Department of Electrical Engineering and Computer Science
York University
aan@yorku.ca

1 Data Migration

Data migration refers to the process of transferring data from one storage system or computing environment to another, commonly when enterprises move from on-premises servers to cloud platforms or between clouds. It is a critical operation for organizations seeking to upgrade infrastructure, consolidate data sources, or adopt cloud-based solutions.

Industrial demand for efficient data migration is growing, as illustrated by two common scenarios. First, enterprises frequently migrate production systems and datasets across locations via the internet. In such cases, minimizing business disruption is crucial to prevent financial loss, making maximizing migration throughput a top priority. Second, companies with hybrid infrastructures often perform frequent data transfers to leverage cloud-based AI and analytics services. Here, the goal shifts to minimizing network costs, particularly bandwidth expenses incurred by regular dataset updates for maintaining freshness.

The key objectives in data migration are thus twofold: maximizing throughput and minimizing cost. However, existing methods typically compress data directly into smaller files for transfer, without exploiting underlying data distribution patterns to enhance compression efficiency.

2 Talk Abstract

In this talk, I will present an IBM CAS project on optimizing data migration using machine learning, conducted in collaboration with the IBM Cloud DB2 Data Migration team. I will describe a data migration framework that integrates online classification and tuple reordering to enhance migration throughput and reduce costs. Online classification dynamically groups similar records during migration, enabling more effective compression and faster data transfer. Tuple reordering, guided by functional dependencies, further improves performance by optimizing the similarity among adjacent tuples within each chunk.

In addition, the efficiency of data migration depends on selecting an appropriate compression level, which varies with data characteristics and network conditions. Since there is an inherent trade-off between compression ratio and processing time, I will present a predictive method for determining the optimal compression level based on data properties and available bandwidth.

These advancements demonstrate significant potential for improving data migration efficiency, particularly for large-scale datasets and bandwidth-constrained environments.

3 Acknowledgments

The project is supported by an IBM CAS grant and an NSERC Alliance grant. I would like to acknowledge my co-PI, Dr. Xiaohui Yu from York University, our IBM collaborator Dariusz Jania, and the dedicated team of graduate and undergraduate students and a postdoctoral fellow at York University: Qinxin Du, Zhongxin Hu, Kaiyu Li, Xingjian Mao, Jingpeng Pan, and Yunfei Peng, for their valuable contributions to this project. The contents presented in this talk have been or will be published in [3, 1, 2]

References

[1] Zhongxin Hu, Kaiyu Li, Xingjian Mao, Jingfeng Pan, Yunfei Peng, Aijun An, Xiaohui Yu, and Dariusz Jania. Damocro: A data migration framework using online classification and reordering. In *CIKM*, pages 4546–4553, 2024.

[2] Xingjian Mao, Xiaohui Yu, Aijun An, and Dariusz Jania. Optimizing the data migration process using machine learning techniques. In *CASCON*, 2025.

[3] Jingfeng Pan, Yunfei Peng, Kaiyu Li, Aijun An, Xiaohui Yu, and Dariusz Jania. Optimizing data migration using online clustering. In *CASCON*, pages 173–178, 2023.