DataCI: A Platform for Data-Centric AI on Streaming Data

Huaizheng Zhang [* 1] **Yizheng Huang** [* 1] **Yuanming Li** [* 1]

Abstract

We introduce DataCI, a comprehensive open-source platform designed specifically for data-centric AI in dynamic streaming data settings. DataCI provides 1) an infrastructure with rich APIs for seamless streaming dataset management, data-centric pipeline development and evaluation on streaming scenarios, 2) an carefully designed versioning control function to track the pipeline lineage, and 3) an intuitive graphical interface for a better interactive user experience. Preliminary studies and demonstrations attest to the easy-to-use and effectiveness of DataCI, highlighting its potential to revolutionize the practice of data-centric AI in streaming data contexts.

1. Introduction

Data-centric AI has shown promising results in many domains (Kirillov et al., 2023; Xu et al., 2021; Liu et al., 2020; Bartolo et al., 2021). Compared to model-centric AI, it pays more attention to data processing like data cleaning and labeling instead of model structure design, avoiding *garbage in, garbage out*. Both industry and academics have responded to the growing interest in data-centric AI and made significant progress (Melgar et al., 2021; Zha et al., 2023; Mazumder et al., 2022; Parrish et al., 2023).

However, current data-centric frameworks mainly focusing on static batch settings, are inadequate to address the unique challenges that unbounded and continuous streaming data present. First, adopting these frameworks to a streaming data environment requires much more tedious efforts, including streaming data pre-processing, data-centric function version control, etc. Second, several issues unique to streaming data, such as determining the optimal frequency for running data-centric pipelines, are underexplored (Shankar et al., 2022; Böther et al., 2023).

---

[*]Equal contribution [1]Independent Researcher. Correspondence to: Huaizheng Zhang <zhanghuaizheng0806@gmail.com>.
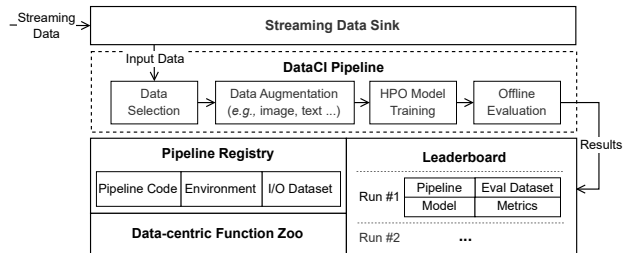
Figure 1. DataCI architecture.

In light of these challenges, we introduce an early-stage work, that demonstrates the potential of data-centric AI in streaming data settings, providing both practical showcases and quantitative analysis. Our contributions are two-fold:

First, We introduce DataCI (data continuous integration), an open-source platform that serves as a comprehensive development tool and testbed. DataCI offers a range of modular features, including a Streaming Data Sink, a Pipeline Registry, a Data-centric Function Zoo, a Pipeline Orchestration Module, and a Leaderboard. Furthermore, a graphical user interface allows users to easily manage their pipeline and visualize results, facilitating benchmark studies and the exploration of new data-centric methods for streaming data. **Second**, we provide demonstrations of DataCI and preliminary studies to highlight its user-friendly design and overall necessity.

This vision paper is an invitation to further study and enhance the application of data-centric AI to streaming data. We look forward to feedback and collaboration.

2. System Design

DataCI offers an end-to-end data-centric ML experience on dynamic streaming data. As shown in Figure 1, it introduces five primary components.

Streaming Data Sink. DataCI uses the sink to periodically collect the data from the connected streaming system and save it into different partitions. These data partitions will be tagged according to their arrival time and used for later training and evaluation.

Pipeline Registry. This module enables users to build a data-centric pipeline with ease by introducing flexible APIs. These APIs cover dataset I/O, pipeline staging, and running
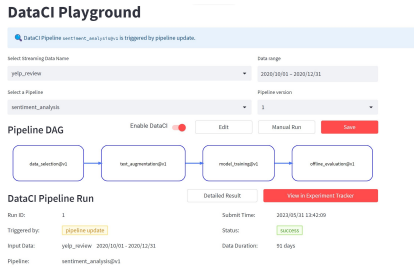
Figure 2. An interactive web interface to build, launch, compare, and visualize DataCI pipelines.

environment management, speeding up the development of a DataCI pipeline. Furthermore, a **versioning control** function is introduced to track the pipeline lineage.

Data-centric Function Zoo. Different from the model hubs from HuggingFace (Wolf et al., 2020) and PyTorch (Paszke et al., 2017), we provide this module to store data processing methods such as data selection, data augmentation, and tricks applied in a specific scenario (e.g., prompting). Users can share and reuse the functions for pipeline building.

Pipeline Orchestration. To run a data-centric pipeline, we leverage the idea from pipeline orchestration systems like Airflow. The data flow shown in Figure 3 will pass through every pre-defined stage in the pipeline and then trigger the evaluation to test the whole pipeline and know the benefit of the new data-centric function. If the pipeline further passes the A/B test, the new function or the new pipeline can be deployed to the product.

Leaderboard. Once the evaluation is finished, the results will be sent to the leaderboard. We use the {run #No.} as the index and store the pipeline name with version, the evaluation dataset, the model name with training hype parameters, and the metric for easy reproduction and comparison.

3. Demonstration

We demonstrate our system from two views, namely user experience investigation and quantity analysis.

User experience. Users' experience is our top priority. To satisfy them, we equip DataCI infrastructure with a **playground** as shown in Figure 2. This playground enables users can try our system in an interactive manner. The playground can be grouped into three sections. First, users select the data from Streaming Data Sink and pre-defined pipeline with a specific version in Pipeline Registry. Second, users can manually launch the pipeline and our playground will show a DAG (directed acyclic graph) for better visualization. Users can replace one function in the DAG and generate a new pipeline version for a quick control experiment. Third, the experiment running details will be presented in the playground for reference.
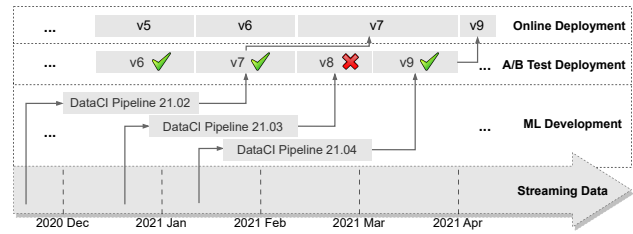

Figure 3. We use Yelp data to simulate a streaming data scenario for the DataCI evaluation.

Quantitative analysis. We simulate a real-world case by using Yelp dataset and sending them into our system in a streaming mode as shown in Figure 3. Assume that the online pipeline has upgraded to v5, which is our starting point. We build a pipeline v6, which passed the A/B test and was deployed to the online production. We keep using the latest data from Streaming Data Sink to develop new data-centric pipelines and test them. Only v8 fails as it can not outperform v7 as shown in Figure 4. Also, from Figure 4, we find if we keep using v6 without a quick pipeline update, the online performance will drop dramatically.

This is a very preliminary study to show that a system for quick building and evaluating a data-centric pipeline on streaming data is necessary, as in the real world, data distributions change very frequently. However, this study also poses many questions that need further exploration. For example, how can we decide the upgrade frequency? Is there any better metric to measure the pipeline's performance in this streaming scenario? etc.

4. Conclusion

Data-centric AI exploration has highlighted the shortcomings of existing tools in streaming data environments. To combat this, we introduced DataCI, an open-source platform that bridges these gaps. With its modular features and intuitive interface, DataCI streamlines streaming data management and method deployment. Preliminary studies affirm DataCI's potential to revolutionize data-centric AI in dynamic contexts.
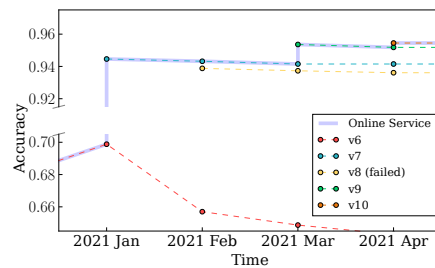

Figure 4. Performance comparison between multiple data-centric pipeline versions.

References

Bartolo, M., Thrush, T., Riedel, S., Stenetorp, P., Jia, R., and Kiela, D. Models in the loop: Aiding crowdworkers with generative annotation assistants. *arXiv preprint arXiv:2112.09062*, 2021.

Böther, M., Strati, F., Gsteiger, V., and Klimovic, A. Towards a platform and benchmark suite for model training on dynamic datasets. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pp. 8–17, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Liu, Z., Ding, G., Bukkittu, A., Gupta, M., Gao, P., Ahmed, A., Zhang, S., Gao, X., Singhavi, S., Li, L., Wei, W., Hu, Z., Shi, H., Liang, X., Mitamura, T., Xing, E., and Hu, Z. A data-centric framework for composable NLP workflows. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 197–204, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.26. URL https://aclanthology.org/2020.emnlp-demos.26.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., Diamos, G., He, L., Kiela, D., Jurado, D., et al. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*, 2022.

Melgar, L. A., Dao, D., Gan, S., Gürel, N. M., Hollenstein, N., Jiang, J., Karlas, B., Lemmin, T., Li, T., Li, Y., Rao, S. X., Rausch, J., Renggli, C., Rimanic, L., Weber, M., Zhang, S., Zhao, Z., Schawinski, K., Wu, W., and Zhang, C. Ease.ml: A lifecycle management system for machine learning. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings*. www.cidrdb.org, 2021. URL http://cidrdb.org/cidr2021/papers/cidr2021_paper26.pdf.

Parrish, A., Kirk, H. R., Quaye, J., Rastogi, C., Bartolo, M., Inel, O., Ciro, J., Mosquera, R., Howard, A., Cukierski, W., et al. Adversarial nibbler: A data-centric challenge for improving the safety of text-to-image models. *arXiv preprint arXiv:2305.14384*, 2023.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Shankar, S., Herman, B., and Parameswaran, A. G. Rethinking streaming machine learning evaluation. *arXiv preprint arXiv:2205.11473*, 2022.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xu, L., Liu, J., Pan, X., Lu, X., and Hou, X. Dataclue: A benchmark suite for data-centric nlp. *arXiv preprint arXiv:2111.08647*, 2021.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.