Data-Centric AI Infra 2.0

Nikon Rasumov [1]

Abstract

In this position paper for the DataPerf2022 workshop at ICML2022 (paper 0891), we share our considerations for an end-to-end Data-Centric AI infrastructure vision to implement Artificial Intelligence (AI). AI is trained and evaluated using datasets that undergo various changes as part of their lifecycle (privacy, drift, errors, transformations, etc). Data-Centric AI Infra helps practitioners understand and iterate on datasets for ML Models. By adopting Data-Centric AI infrastructure our customers could improve their model performance through faster, resource efficient access to AI data. We hope to connect the scientific community with the AI data problems faced in a real production environment at an exabyte scale.

1. Introduction

AI training/evaluation dataset generation and management is a critical part of the AI development lifecycle. Historically, Infrastructure in big companies has not invested in a common AI platform/infrastructure to support this important area. Instead, product/product infra teams have built various AI systems to support their own needs, many of which share common patterns and challenges (Amandalynne Paullada, 2021). AI dataset creation experience and technology is roughly split between: recommendation, content understanding and research. 1.) In recommendation, large scale datasets are created by performing streaming joins of features, datasets and labels and post-processing systems. 2.) In content understanding, joining with human labels is often required, and in rare cases feature logging/joining is required when time sensitive features are used. 3.) In research, largely image, video and text based datasets are used with their human annotations or no annotations at all (unsupervised learning).

Each domain has unique challenges having to do with the shape and processing of data that have minimal overlap with others. For example, at exabyte scale, recommendations' efficiency problems require sophistication that may not be applicable in other domains; discovery of datasets for content understanding/research is a more acute problem than recommendations (which may need discovery of features or raw signals). Yet, there are shared problems that all must solve, e.g. data governance (privacy, security, provenance, fairness, robustness), data understanding for model improvement, quick iterations on data for modeling, smooth transition from development to production, ensuring production meets all excellence requirements (e.g. freshness, training/serving skew, efficiency...), and capacity management. Due to fragmentation, innovations in one system don't apply to the other, losing economy of scale, and shared problems cannot be addressed both initially and through evolutions of systems (HazyResearch, 2022).

Ultimately, while system requirements may vary, ML practitioners need to understand and iterate on data to maximize value in data responsibly, such that: all AI datasets can be UNDERSTOOD for their privacy, fairness and model performance impact; ITERATED on in minimal amount of time with maximum model improvement; and MANAGED for governance and resource efficiency, Figure 1. As AI use cases scale up over the next few years, it is important focus on a common platform/infrastructure. To this end, we are considering building a single platform to empower ML Engineers, ML Infra Engineers and ML Researchers to quickly, efficiently and responsibly understand and iterate data to improve model performance. Concretely, this means: quickly - self-serve, fast prototyping, experimenting, deploying, debugging, backfilling, through a managed service; efficiently - large data sets, resource efficient, data maturity model, data deprecation lifecycle, capacity management; privacy aware, secure, fairness and bias analysis; model improvement through data - data augmentation, synthetic data, active learning, human annotation, weak supervision, zero-shot learning, error discovery/alerts, data cleaning, data valuation and monitoring (Data-CentricAI, 2022).

This document contains our considerations for Data-Centric AI Infra 2.0. To articulate this, we will walk through:

[1] Meta, Menlo Park, CA, USA. Correspondence to: Nikon Rasumov <rasum@fb.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

- The people that we believe matter along the way;
- The major themes and market forces;
- The Considerations for our 3 year vision;

2. Methodology

We interviewed 15 stakeholders in Data-Centric AI Infra with the following job titles: ML Engineer (5), ML Researcher (4), ML Infra Engineer(6).

3. Results

We will now walk through the jobs-to-be-done of the key stakeholders in Data-Centric AI Infra.

3.1. ML Engineers - Customers

As an ML engineer I work on problem or product teams to ship models. I combine known datasets with features and human or machine labels to create training/evaluation/testing datasets. My key objective is to improve models against top-line metrics, while at the same time, making them responsible, privacy aware, explainable and reproducible, update-able (to next versions), and deployable. Algorithms that measure how much data is required to obtain the best performance proved unreliable in practice leading to the lesson "more data is always better". However, what data to include remains an open question. Further jobs to be done include:

- Combine features and labels to form dataset
- Data Discovery and Synthesis
- Root Cause Analysis (RCA) of production issues, Monitoring
- Data Management
- Automation CI/CD

3.2. Product Infra Engineers - Builders

Sometimes also done by or in combination with Data Engineers. I create infrastructure used by Product teams. My key objective is to simplify their life, build scalable, reliable and privacy-aware infrastructure. I'm also tasked to improve the working efficiency for our ML Engineers. I hope the platforms/tools have high stability, and can truly contribute to the long-term goals for tooling consolidation and system simplification. Extensibility of platform tools means I will only build parts specific to my needs rather than the entire infra. My typical questions are: (1) How can I avoid churns from switching platforms/tools? (2) How can I increase our developers' efficiency? (3) How do I simplify debugging (e.g. either in-tools through error messages or via offline, internal service supports with feature lineages to help identify root causes)? (4) How do I establish consistent, standardized practices (e.g. consistent data type definition) for our developers? (5) How can our team maintain privacy-awareness and fairness of dataset creation? (6) How can I support troubleshooting needs? (7) How can I tune dataset performance and manage capacity budgets, deprecate low-ROI datasets and models, and configure infrastructure for new privacy policies? (8) How do I onboard product teams to a new platform? (9) How do I manage org specific training, evaluation datasets? (10) How do I provide metrics that monitor my product infrastructure? Further jobs to be done include:

- Plan and lead tool migration, consolidation, and integration
- Support training, onboarding, and troubleshooting, e.g. onboarding cookbook / detailed code examples based on organization and team's specific needs, repackage a Wiki 'landing' page.
- Establish and maintain communications with externals and internals (e.g. EMs and individual end users of tools). This can also include consolidating and passing along team members' needs and user feedback to tools' Dev teams.

3.3. ML (Applied) Researchers - Customers

As an ML Applied Researcher, I build and scale adoption of State of the Art (SOTA) approaches to solve AI challenges relevant to our products. Much of my work involves making research more easily adoptable to ML engineers within product/problem teams. To achieve this goal I build best in class models, datasets and services. I require flexibility for building a wide and robust set of models, datasets and services, but also benefit from a set of unified infrastructure components for easier customer adoption of their breakthroughs. My preferred language is python with some knowledge of SQL. I often deal with large human/machine labeled data. Examples include the engineers who build various classifiers for content integrity (hate speech, misinfo, etc). Typical questions ML Applied Researcher try to answer include: (1) What is the SOTA model/architecture for my use-case? (2) What data does it use? (3) What is the most important data and how do they contribute? (4) What other data could I use? (5) Can I use this data (privacy/health)? (6) What is my data's health? (7) How do I make sure my customers use the most up to date model as their data? (8) ROI before adding new data. Further jobs to be done include:

- Develop and apply new state-of-the-art machine learning (models, datasets, services)

- Enable Product and Problem teams to use embeddings of large scale models

Based on the universe of users above we find the following four themes.

4. Major Themes

Four major themes and metrics will be considered in Data-Centric AI Infra over the coming 3 years.

4.1. Theme A: Dataset DevX

ML dataset creation, analysis and modification is broken down into multiple incompatible languages, libraries, frameworks and tools thus taking weeks to bring a new dataset to production. Data-Centric AI Infra may include smart data selection based on the model needs, data discovery and data synthesis based on already available datasets (human and machine labeled) thus leading to a fast, self-serve experimentation, productization and superior developer experience.

Metric: The Northstar metrics are the time to productionize a dataset (dataset used to create a production model) and counter metric Pulse satisfaction survey (number of customers that are at least somewhat satisfied).

4.2. Theme B: Resource Efficiency

Large datasets contribute to a model improvement at a low computational and storage cost. It includes large, resource efficient data sets with a maturity model, and data deprecation lifecycle, using just the right amount of data when it is needed with a full self-serve capacity management.

Metric: The northstar metric is the use of a minimal amount of data to achieve top model performance.

4.3. Theme C: Model Improvement

Model improvement through data could be achieved by improving the individual steps in the dataset journey: creation, analysis, and modification of datasets. Dataset creation includes: data augmentation, synthetic data, active learning, zero-shot learning, human annotation, and weak supervision. Dataset analysis and modification includes: error discovery/RCA, subgroup bias (Responsible AI), alerts, data valuation, data cleaning, label quality, off-policy model evaluation and robustness and adversarial perturbations. It does not escape our attention that better model performance could also lead to better datasets thus creating a positive feedback loop.

Metric: The northstar metric is the model improvement produced by the dataset e.g. measured in Accuracy.

4.4. Theme D: Awareness

All usage of data in AI needs to consider external regulatory requirements and internal policies in the areas of privacy, security, fairness, robustness, transparency and governance. Data-Centric AI Infra 2.0 could help users with privacy aware and secure infrastructure and other primitives.

Metric: The northstar metric is the number of datasets that reached a certain level of awareness.

Now we will summarize the vision of how Data-Centric AI Infra could help practitioners to understand and iterate their data to improve models without decreasing their development efficiency.

5. Conclusion

In summary we are considering investing in the following Data-Centric AI Infra 3 Year Vision:

"Infrastructure to empower ML Engineers, ML Infra Engineers and ML Researchers to quickly and efficiently understand and iterate data to improve model performance."

Ultimately, while system requirements may vary the ML practitioners should understand and iterate on data to maximize value in data responsibly, such that: all AI datasets can be UNDERSTOOD for their privacy, fairness and model performance impact; ITERATED on in minimal amount of time with maximum model improvement; and MANAGED for governance and resource efficiency. As AI use cases scale up dramatically over the next few years, it is important to invest in a common platform/infrastructure.

Concretely, this means:

- Quickly: self-serve, fast prototyping, experimenting, deploying, debugging, backfilling, through a managed service
- Resource Efficiency: large data sets, resource efficient, data maturity model, data deprecation lifecycle, capacity management
- Understanding: privacy aware, secure, fairness and bias analysis
- Model Improvement through data: data augmentation, synthetic data, active learning, human annotation, weak supervision, zero-shot learning, error discovery/alerts, data cleaning, data valuation and monitoring

References

Amandalynne Paullada, Inioluwa Deborah Raji, E. M. B. E. D. A. H. Data and its (dis)contents: A survey of

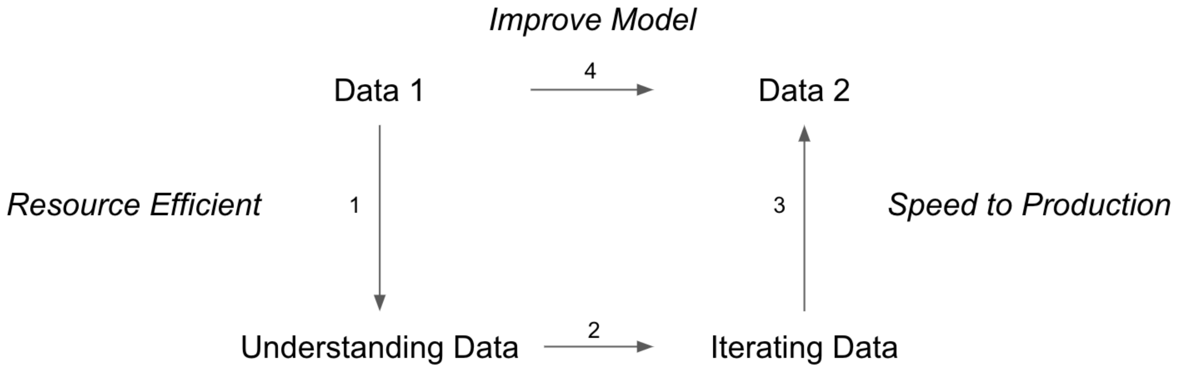

Figure 1. Data-Centric AI Infra. Dataset 1 can be understood (1) and iterated (3) to improve the model (4). in a resource efficient way (1). thus identifying the right iterations to produce a dataset 2 that improves the model performance (4) in a speedy way (3).

dataset development and use in machine learning research. *Cell REVIEW*, 2:211–229, 2021.

Data-CentricAI. https://www.datacentricai.cc/. 2022.

HazyResearch. https://github.com/hazyresearch/data-centric-ai. 2022.