Three Core Principles for Accelerating AI Value

There was a time when data alone held tremendous power. With the rise of machine learning, the true power comes from what we learn from that data. It's not enough to simply have the information. To compete in today's market, companies need to unlock the hidden insights and invisible patterns in that sea of data.

Production-ready ML & DataOps Stacks Are Built with Three Principles in Mind.

The journey to mature, business driven machine learning (ML) is fraught with pitfalls: Teams can easily waste untold amounts of time with buggy and brittle data pipelines, inconsistent, poorly curated and poorly labeled datasets, not to mention disconnected tool chains that don't work together. It's not long before they find themselves outpaced by more efficient competitors built for big scale.

Teams and technology vary from project to project, and company to company. By aligning to the core capabilities defined in this report, AI & ML leaders not only increase their chance of succes, they can drive real value from machine learning much faster.

Democratization

There's more to managing your data than governance alone. While data governance gives teams policies and procedures for handling data, data democratization looks to unlock silos and share permissions across an organization securely, getting it into the hands of the data science teams that can unlock those hidden insights.

> "You need to go fast. You need to work with your existing tools, your existing languages, your existing dependencies. You want to invest as little as possible in learning, right? You just need stuff to be processed. And since Pachyderm utilizes really flexible tools like Docker and Kubernetes, it's very democratizing."

EYAL HELDENBERG,
VOICE AI PRODUCT MANAGER AT THE LOGMEIN AI CENTER OF EXCELLENCE

Why Democratize?

Democratization delivers the right balance between security and usability: It gives teams access to data so they can extract fresh value while allowing the people who are directly responsible for the storage and control of data to maintain strong controls over who and what can access the information.

Democratizing data access gives organizations an analytical edge. If your teams can't even get access to the rich datasets in your organization and they have to jump through countless unnecessary hoops then you'll find your data science teams stuck in neutral, unable to give you the insights you need to make better business decisions.

Two Sides of Democratizing Data

ACCESSIBILITY

Access to data and data interpretation tools is provisioned for teams who can make use of product development, analytics, and business intelligence.

USABILITY

There is more to democratization than simple access. Without tools agile enough for collaborative and flexible workflows between teams, data is accessible without being democratized.

Pachyderm takes a unique approach to data operations. It gives teams a centralized, source-agnostic and version controlled source of truth for data, with robust role based access control (RBAC) that allows your team to centralize permissions while opening those datasets up to fresh insights and productive collaboration. Pachyderm gives you a containerized, language-agnostic data pipeline that can quickly and easily extract data from any source, and pull it into your pipelines to deliver powerful machine learning models to production.

USE CASE:

Rapid Model Deployment for Pandemic Response

Companies and governments rely on RiskThinking.AI to model uncertain futures, predicting everything from climate change to COVID. Their key challenge: how to improve the data scientists' workflow while providing data engineers with a robust toolset for ML operations.

The less data scientists have to deal with data wrangling, versioning, and troubleshooting pipelines, the more they can focus on the future: trying out different ideas and looking at data in different ways.

Pachyderm let Riskthinking.AI's data scientists focus on the complexity of models rather than the complexity of figuring out which model was trained on which version of the dataset. It gave them the foundation to work with data and deploy any ML tool they wanted, automatically pushing the best performing models to production.

But data democratization didn't stop there: thanks to Pachyderm, data scientists could also share how models were improving over time with less technical stakeholders. Because when you can demonstrate what you're doing to your team, everyone knows the impact they're making on the bottom line, and the real world.

When your ML platform levels the playing field for data access, teams can make connections and build products that would have been impossible otherwise.

By 2025, 80% of data and analytics governance initiatives focused on business outcomes, rather than data standards, will be considered essential business capabilities.

GARTNER

Speed of Iteration

In order for ML to really transform your business and bottom line, your team needs to move fast enough to affect significant change in your process and operations. This is where self-built solutions can start to break down: if self-built tools aren't equipped for scale or support, it can lead to models that never get to production, or models that break down with no clear indication as to why and brittle datasets that deliver inconsistent results every time you roll out a new version of the model.

> "The difference was an order of magnitude faster... If it took 10 hours on the old system then it would only take an hour with Pachyderm."

GEORGE BONEV, PH.D
MACHINE LEARNING ENGINEER, LIVEPERSON

The faster you can process data, test models, and iterate on your results the more scalable your machine learning operations will be.

What usually gets in the way of ML success?

- Lack of robust machine learning operations and tools that can't handle scale
- No way to scale data processing
- Relying on error-prone manual processes
- A complete inability to quickly and easily investigate unexpected results.

This is where the right data management approach can make or break your project's success. Pachyderm shrinks data processing time for machine learning operations by delivering a data-centric pipeline that:

- Automatically parallelizes data processing without you having to write scaling specific code or cut and paste boilerplate templates
- Incrementally processes new data, which cuts cloud compute and storage load by only processing differences and automatically skipping duplicate data

USE CASE:

Improving Customer Service Outcomes

LivePerson connects people and brands through their AI-powered Conversational Cloud. Their platform lets brands provide conversational experiences on messaging channels like SMS, WhatsApp, and more, so people can stop wasting time on hold and crawling through websites. On top of that, LivePerson's conversational AI lets brands scale their ability to hold conversations with as many of their customers as possible, as quickly as possible.

Facing an influx of new customers, LivePerson faced an all-too-common processing bottleneck: their machine learning infrastructure was built with multiple different Python scripts running across many independent nodes. It all ran on a cluster of powerful machines but each machine had its own discrete, unshared storage with no standardized repository of data.

If a scientist wanted to do a grid search for a certain model they were optimizing, they'd have to track jobs across these different, loosely connected servers. Performance was monitored manually, meaning any issues would require accessing an individual server to troubleshoot it.

Pachyderm removed friction from every stage of the process to accelerate experimentation and iteration for LivePerson's team, and their customers:

- By automatically scaling and scheduling jobs, Pachyderm efficiently transformed data for model processing.
- Jobs could be tracked across the entire cluster - far less time spent finding the issue, more time available for solving the problem.
- Parallelized, incremental processing meant data scientists could process more experiments than ever, removing bottlenecks that delayed customer results.

LivePerson's ability to experiment and iterate was unleashed with Pachyderm's performance and processing improvements. With automated and centralized monitoring, critical issues can be quickly resolved with the lowest possible interruption - and the best possible customer satisfaction.

When teams have the capacity to make fast, agile iterations, they are more likely to turn their experiments into products and tools that have real business impact.

Data Auditing and Accountability

With terabytes of sensitive data getting flowing into company databases every day, it's critical that organizations can rapidly respond to everything from data breaches, to GDPR right to be forgotten requests, to policy violations. Teams not only have to worry about evaluating the ethics of their models to ensure they're transparent, fair and equitable, they also have to understand whether their models violated company policies or data protection laws around the world.

Strong auditing and accountability comes from robust data operations. If you don't have strict controls on where data came from and how it's changed over time, it's hard to deal with challenges in real time.

The key is reproducibility. Reproducibility allows data science teams to look at the output of a machine learning model, and trace its entire story back to its original input. It lets that data science team look at different versions of the data, as well as how that data changed and how those changes lead to a different outcome.

When processing sensitive or personally identifiable data with Pachyderm, data science teams produce results that can be audited and tracked to their data source, simplifying ML audit and compliance workloads. They can look backwards and forwards in time and track the total history of personal and sensitive data all the way back to its raw input, including all analysis, parameters, code, and intermediate results generated by a model.

USE CASE:

Improving Patient Outcomes with Reproducibility

A top managed healthcare provider has a dedicated team leveraging cutting edge AI to harvest long term insights and make more detailed health predictions from claims and electronic health record data. The team required data lineage, access control, and - critically - accelerated processing for massive quantities of healthcare data.

Pachyderm delivered immutable lineage for patient data, plus the parallelism and incrementality required to efficiently scale the AI team's ML processing.

With millions of patient records, they could not afford the time or cost of re-processing irrelevant records when only a small subset were relevant at any given time. Pachyderm not only processed these records in parallel, it also automatically processed only those containing new information, increasing both scale and speed while reducing costs.

"One of my first observations with Pachyderm was what matters most in the pipeline is that the right data shows up at the right place and time. That level of abstraction allows our teams to do 90% of development without even deploying anything to Pachyderm at all. When we get to production, all we do is change the pipeline path and everything runs just fine. This philosophy of creating a container at the last moment really protects the data scientists from having to deal directly with the infrastructure." - Head Engineer, Top Healthcare Provider

As a plus, Pachyderm's approach to incrementality is also key to its data lineage, giving them the tools to investigate errors and reproduce past outcomes with a simple, unbreakable change history.

With time and energy saved on tracing data lineage, AI teams using Pachyderm have more bandwidth to focus on building ethical AI, reducing model bias, and reviewing the quality of their outputs.

Conclusion: The Foundation of Successful Data Operations

Teams everywhere face the challenges of democratization, speed, and data auditing. It's not enough to deal with these problems as they happen. You need tools purpose-built for helping you scale these machine learning mountains. You need a plan to address problems in real time. The best AI teams do that by adopting and well thought out data policy, making that data available to the people that need it, and giving them the data-centric tools they need to process data smarter and with greater flexibility.

Launching ML without effective data operations tools and policies in place can lead to bottom line disasters, legal challenges, auditing failures and public relations nightmares. Pachyderm delivers the tools you need to automate data transformations at every stage of your machine learning process. It reduces duplication and corruption, provides version control and lineage of that data, and the workflow for retrieval, manipulation, and analysis of data.

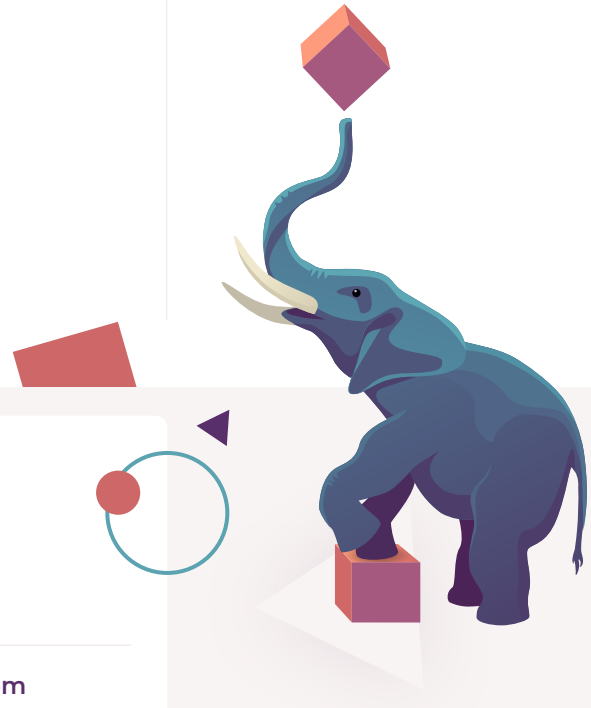
Pachyderm's team of AI experts and data scientists can help your firm unlock the true value of your data. Contact us to learn more about how Pachyderm can help get your ML and AI projects to market faster, lower data processing and storage costs, and meet strict data governance requirements.

About Pachyderm

Pachyderm is the leader in data versioning and pipelines for MLOps. We provide the data foundation that allows data science teams to automate and scale their machine learning lifecycle while guaranteeing reproducibility.

With over $40 million in three rounds of funding from leading investors like Benchmark, Microsoft M12, Y Combinator, and others, Pachyderm, Inc. offers a commercial Pachyderm Enterprise Edition and an open source Pachyderm Community Edition.

Pachyderm helps customers get their ML and AI projects to market faster, lower data processing and storage costs, and supports strict data governance requirements.



Contact Pachyderm

To learn more about Pachyderm's machine learning solutions, contact us:

info@pachyderm.com • 888-338-9597 • www.pachyderm.com