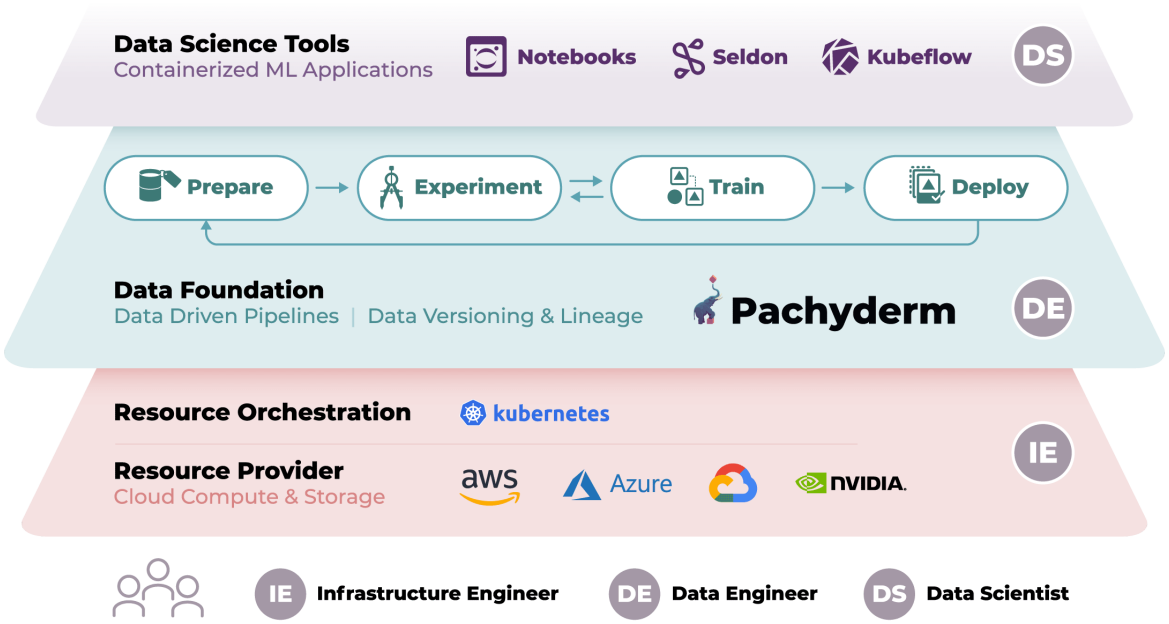
Pachyderm

SOLUTION BRIEF

The Leader in Data Versioning and Pipelines for MLOps

Pachyderm provides the data foundation that allows data science teams to automate and scale their machine learning lifecycle while guaranteeing reproducibility



To get the full value out of machine learning (ML) investments, data science teams need MLOps tools that allow them to iterate rapidly to get improved models out to customers quickly. Pachyderm provides a robust data layer that allows teams to productionize their ML life cycle resulting in quick delivery while lowering cloud costs and meeting data governance requirements. Key benefits of Pachyderm include:

Data-Driven Automation

Automate your MLOps tool chain with Data-Driven Pipelines and Data Versioning for increased productivity and reduced risk

- Automatically trigger pipelines when new data arrives
- Ability to process only new or changed data through incremental processing
- Code agnostic – supports any library or language

Petabyte Scalability

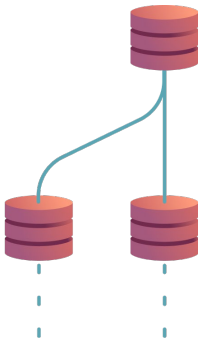
Rapidly process the largest unstructured and structured data sets with automatic parallel and incremental processing that requires no code changes

- Parallel processing that requires no code changes
- Powerful content-based deduplication that lowers storage and compute costs
- Kubernetes and container native

End-to-End Reproducibility

Iterate quickly while still meeting audit and data governance requirements through end-to-end reproducibility and immutable data lineage

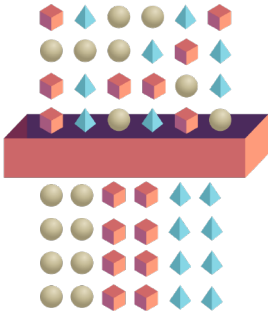
- Faster data debugging
- Ideal for meeting data governance requirements
- Ease compliance and audit tasks



Automated Data Versioning

Pachyderm's Data Versioning gives teams an automated and performant way to keep track of all data changes:

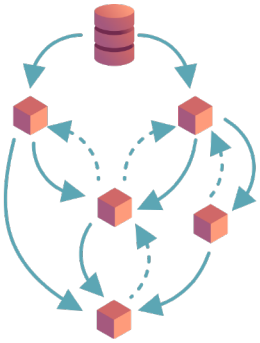
- Utilizes a Git-like structure that enables effective team collaboration through commits, branches and rollbacks
- Powerful content-based deduplication reduces the cost of storing and accessing large data sets
- File-based versioning provides a complete audit trail for all data and artifacts across pipeline stages including intermediate results
- Stored as native objects (not metadata pointers) so that versioning is automated and guaranteed



Data-Driven Pipelines

Pachyderm's Containerized Pipelines speed data processing while lowering compute costs:

- Kubernetes native approach supports any library or language
- Autoscale with parallel processing of data without writing additional code
- Automated pipelines execute whenever new data is committed
- Incremental processing saves compute by only processing differences and automatically skipping duplicate data
- Pipeline steps have JSON/YAML defined inputs and outputs that ease debuggingImmutable



Immutable Data Lineage

Pachyderm's Data Lineage provides an immutable record for all activities and assets in the ML lifecycle:

- Track every version of your code, models, and data
- Maintain reproducibility of data and code for compliance
- Manage relationships between historical data states

Pachyderm's Global IDs make it easy for teams to track any result all the way back to its raw input, including all analysis, parameters, code, and intermediate results.

> "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."

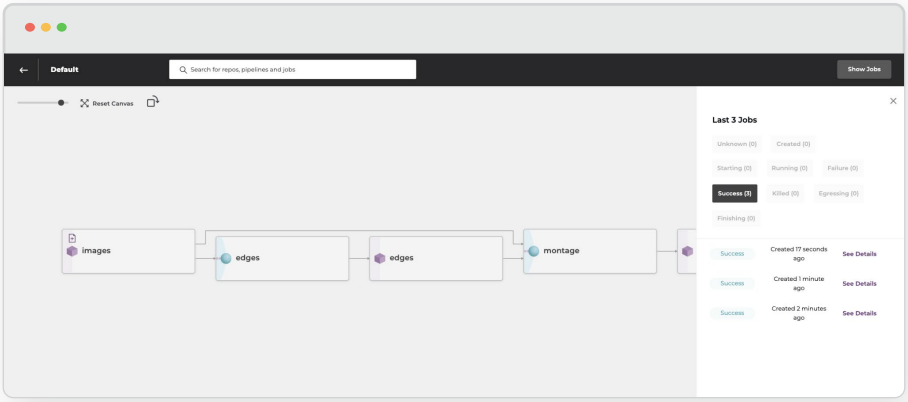
GEORGE BONEV, PHD
MACHINE LEARNING ENGINEER,
LIVEPERSON

LIVEPERSON

> "Pachyderm is on its way to becoming the next big data infrastructure company."

NAGRAJ KASHYAP
CORPORATE VICE PRESIDENT,
MICROSOFT

Microsoft



Console

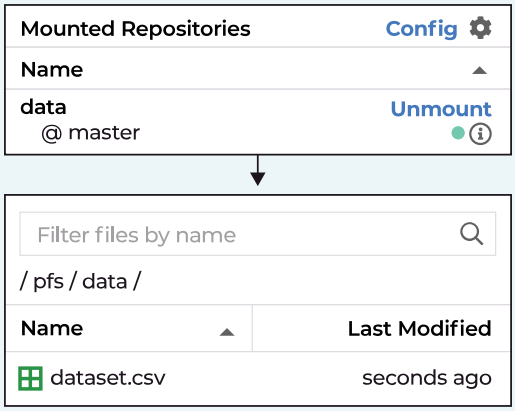
The Pachyderm Console provides an intuitive visualization of your DAG (directed acyclic graph), and aids with debugging and reproducibility with Global IDs.

- See the overall structure and flow of all your pipelines
- Ease pipeline and workflow design
- Facilitate collaboration across teams on shared DAGs
- Drill into pipelines and job details for easy debugging

> "Today, our workload runs in under a day due to incremental processing, thanks to Pachyderm. We were able to push out more models by training and serving them in parallel."

OLIVER WALTER
RESEARCH ENGINEER ASR
FRAUNHOFER-GESELLSCHAFT

Fraunhofer



Notebooks

Pachyderm's JupyterLab Mount Extension provides a point-and-click interface to Pachyderm versioned data:

- Accelerate experimentation with easy and intuitive access to versioned data
- Mount any Pachyderm data repository locally for convenient access
- Work with versioned data like it's on your own file system. No Pachyderm knowledge required
- Explore data with a built in file browser
- Collaborate across teams with a single source of truth for your data

Enterprise Administration

Pachyderm provides robust tools for deploying and administering Pachyderm at scale across different teams in your organization

- Helm 3 provides robust and standards-based deployment on any public or private cloud
- Enterprise Server provides easy centralized licensing and administration of all Pachyderm clusters / workspaces
- Use any identity provider with Pachyderm's pluggable authentication
- Role Based Access Control (RBAC), allows for fine grained control over access to clusters and data

MLOps Tool Chain Integration

Pachyderm is Kubernetes native and can be deployed across a wide range of public and private clouds. Pachyderm provides a range of integration options so that it can work smoothly with MLOps solutions across the entire ML lifecycle including labeling, experimentation, training, serving, and monitoring.

AI Infrastructure Alliance

Pachyderm is proud to be a leader in the AI Infrastructure Alliance (AIIA, pronounced ay-ya) – a community of the world's leaders in MLOps. AIIA is building unified best practices for best-in-class MLOps workflows in a collaborative and open forum.

> "Pachyderm reduced our data processing time from 7 weeks, to just 7 hours."

EYAL HELDENBERG
VOICE AI PRODUCT MANAGER,
LOGMEIN

LogMeIn

> "One of the powerful features of Pachyderm is that data is treated as a first-class citizen."

VINCENT KOOPS
SENIOR DATA SCIENTIST,
RTL NEDERLANDS

rtl

Pachyderm Products

Pachyderm offers commercial and open source data management products to help you build a robust MLOps stack that will stand the test of time.

Enterprise Edition

Pachyderm Enterprise Edition is designed for the largest projects in highly secure environments. Along with world-class support, your team also gets access to our full range of premium features including Pachyderm Console, authentication and access controls (RBAC), no scaling limits, JupyterHub integration, and centralized multiple cluster management.

Community Edition

Pachyderm Community Edition is our open source version of Pachyderm. With Pachyderm Community Edition, you get the core Data Versioning and Pipeline features of Pachyderm that you can deploy locally or in the cloud of your choosing. If you need help, there's an entire community of experts ready to offer their assistance.

Features	Enterprise Edition	Community Edition
Automated Data Versioning	◆	◆
Immutable Data Lineage	◆	◆
Data-Driven Pipelines	Unlimited	16 pipelines
GPU Support	◆	◆
Parallel Processing and Auto-Scaling	Unlimited	8 parallel workers
Global Identifiers for Easy Reproducibility	◆	◆
Incremental Processing	◆	◆
Spouts Streaming Data Architecture	◆	◆
S3 & FUSE Client Support	◆	◆
Prometheus Metrics	◆	◆
Helm 3	◆	◆
JupyterLab Mount Extension	◆	◆
Pachyderm Console (Pachyderm UI)	◆	◆
Role Based Access Controls (RBAC)	◆	Trial
Centralized Multiple Cluster Management	◆	Trial
Pluggable Auth - Login with your IdP	◆	Trial
Enterprise-Grade Support	◆	

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.

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