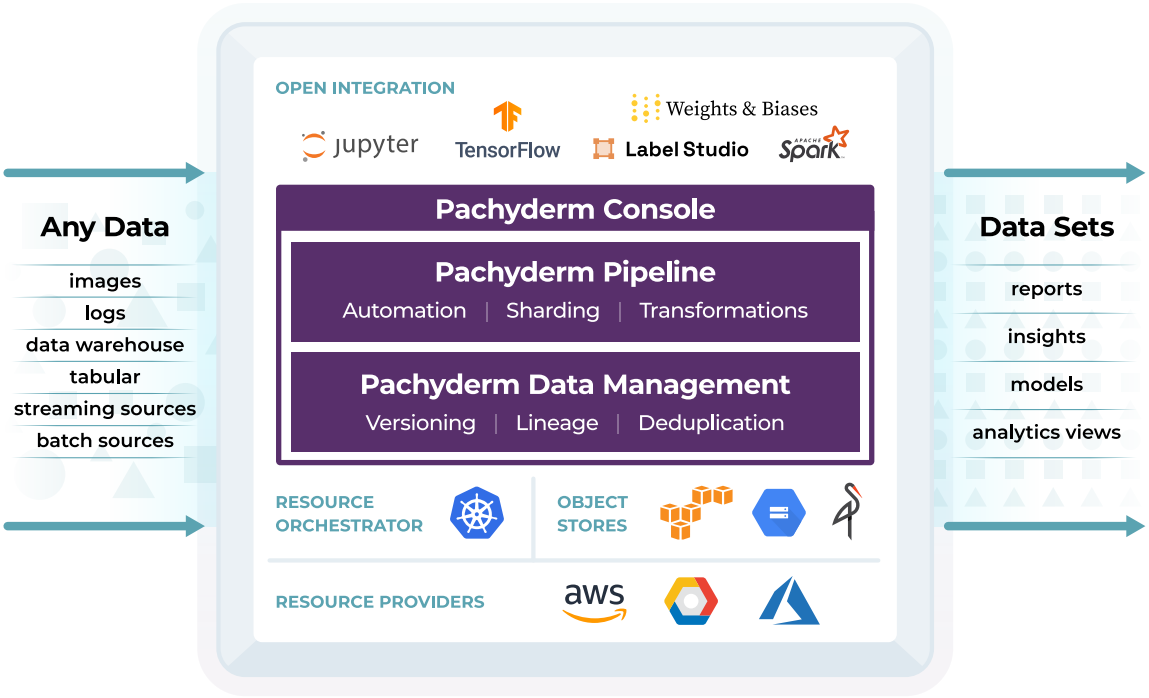
Pachyderm: Data-Driven Pipelines

Automate Data Transformations with Data Versioning and Lineage.



What Pachyderm Provides:

- **Data-driven pipelines** are automatically triggered based on detecting changes.
- **Immutable data lineage** with data versioning of any data type.
- **Autoscaling and parallel processing** built on Kubernetes for resource orchestration.
- Uses **standard object stores** for data storage with automatic deduplication.
- Container-native provides complete autonomy to use **any programming language**.
- Runs across **all major cloud providers** and on-premises installations.

> “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.”

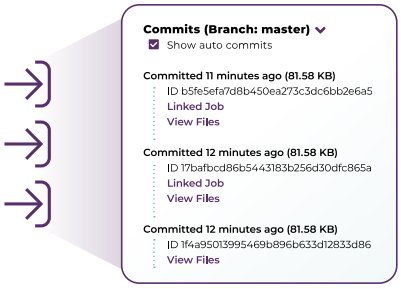
OLIVER WALTER
RESEARCH ENGINEER ASR
FRAUNHOFER-GESELLSCHAFT

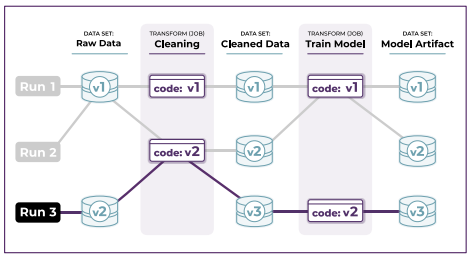
Fraunhofer

Key Benefits of Pachyderm include:

Cost-Effective Scalability

- Deliver reliable results optimizing resource utilization and maximizing developer efficiency.
- Optimize resource utilization with completely automated data-driven pipelines.
- Deduplication of data and code saves infrastructure costs.



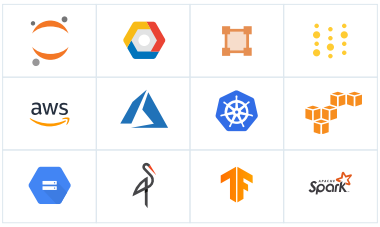


Reproducibility

- Ensure reproducibility and compliance
- Immutable data lineage and data versioning.
- Familiar git-like structure of commits, branches, & repos.

Flexibility

- Leverage your infrastructure investments.
- Run on your existing cloud or on-premises infrastructure.
- Run again any type, size, or scale of data in both batch or real-time pipelines.



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

NAGRAJ KASHYAP
CORPORATE VICE PRESIDENT,
MICROSOFT

Microsoft

Built for Data Engineers

Pachyderm is container-native, running with standard containerized tooling and allows data engineers complete autonomy to use whatever languages or libraries are best for the job.

Pachyderm is data-agnostic, supporting both unstructured data such as videos and images as well as tabular data from data warehouses.

Pipelines are intelligently triggered by detecting changes to data, which is all automatically version controlled by the platform.

Chosen by Leaders

Reduce costs and time to results with automatic intelligent “diff-based” data processing, data deduplication, and dynamic scalability.

Ensure reproducibility and compliance via immutable data lineage and data versioning of all data types and logic – input data, data processing logic, output results, metadata, and models.

Increase team efficiency and collaboration via git-like structure of commits, branches, and rollbacks.

Loved by the Organization

We understand that you support Data Scientists, MLOps, and other infrastructure teams. They will love Pachyderm too!

Data Science Support: Let Pachyderm be the single source of truth for your data. Use familiar Jupyter notebooks to experiment and iterate with your data collaboratively, while always remaining in sync.

MLOps Support: We work with the standard Kubernetes tools, integrate into existing systems, and run across all cloud and on-premises providers.

Pachyderm Products

Pachyderm is available in two major ediitons - Community and Enterprise. Community Edition is available on GitHub and licensed under a developer-friendly open license. The Enterprise Edition is available directly from the company under a commercial license. Both editions enable the building of a robust data-pipelining solution with data versioning and lineage.

Community Edition

Pachyderm Community Edition includes the core platform with data-driven pipelines, versioning, and immutable data lineage. The Community Edition is designed for small teams who prefer to build and support their own software. There are limits on the number of pipelines and parallel workers supported.

Enterprise Edition

Pachyderm Enterprise Edition is designed for organizations and teams that require more advanced administrative features and no limits with world-class support. The Enterprise Edition expands on the features of Community Edition, with pluggable authentication and role-based access controls.

Both editions run across all major cloud providers and on-premises installations.

FEATURES	Community	Enterprise
Console	✓	✓
Notebook Support	✓	✓
Immutable Data Lineage	✓	✓
Data Version Control	✓	✓
Deduplication	✓	✓
Data-Driven Pipelines	16	Unlimited
Parallel Processing (Parallel Workers)	8	Unlimited
Role Based Access Controls (RBAC)		✓
Pluggable Auth – Login with your IdP		✓
Enterprise Support		✓

Key Features

CONSOLE

Console is a complete web UI for visualizing running pipelines and exploring your data.

- Map out the overall structure and flow of all pipelines.
- View repositories, commit histories, and preview data directly in your browser.
- Follow job statuses, pipeline processes, and execution history.

NOTEBOOK

JupyterLab Mount Extension that selectively maps the contents of data repositories right into your Jupyter environment.

- Ideal for Data Scientists to explore and analyze data.
- Run and test pipeline code against versioned data.
- Create reliable, shareable development environments.

About Pachyderm

Pachyderm empowers data engineering teams to automate complex pipelines. Our unique architecture is cost-effective at scale and enables sophisticated data transformations across any types of data. We provide auto-scaling and parallelized processing of multi-stage, language-agnostic pipelines with data versioning and data lineage tracking.

Our data-driven approach means new data is handled incrementally without unnecessary reprocessing. Pachyderm delivers the ultimate CI/CD engine for data.

Our products solve a variety of machine learning and large-scale data transformation use cases, such as – NLP, image/video processing, genomics analysis, IoT stream processing, and risk analysis. We are backed by Benchmark, Y Combinator, and Microsoft's Venture Fund.

> "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

Contact Pachyderm

To learn more about the Pachyderm data-driven pipelines, contact us:

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