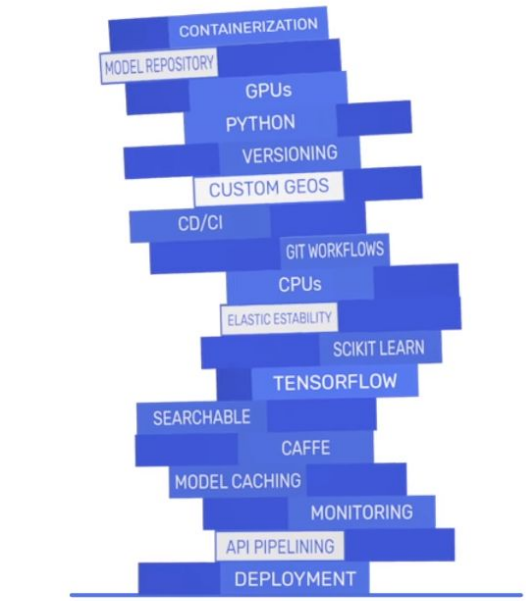
ALGORITHMIA

Algorithmia Enterprise: Realize your ROI with AI

The Problem: AI Deployment

You've collected and cleaned your data, hired data scientists, developed algorithms, trained neural networks, and deployed your machine learning models manually. When it was only a handful of models, they could fit into your normal systems and workflow, but once you began to deploy your model investments at scale, seemingly small issues began to snowball. Most organizations get stuck here. Algorithmia can help.

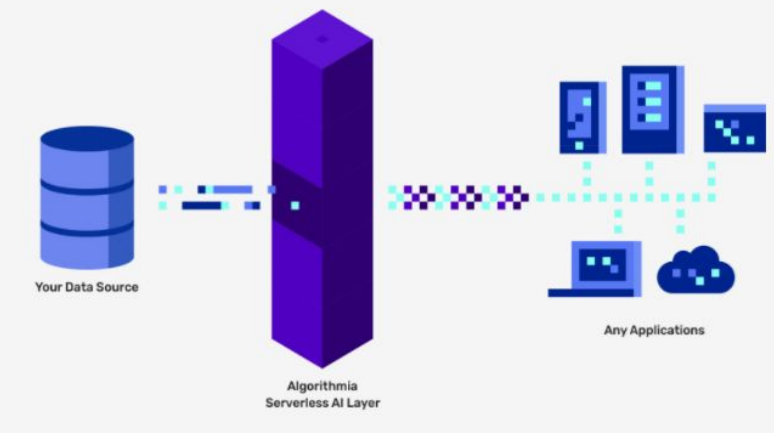


Every serious machine learning program has an AI Layer

Deploying ML across your enterprise requires a unique, scalable, flexible, and secure infrastructure to deliver potential benefits. Facebook, Uber, and Google have spent years developing private software infrastructure to power their advanced AI. Data scientists are a scarce resource and can only effectively tackle one challenge at a time. Let data scientists focus on data science and let Algorithmia's AI Layer focus on deployment and operational scalability.

The Solution: The Serverless AI Layer

Accelerate development and begin deploying your own AI at scale.



"As someone that has spent years designing and deploying machine learning systems, I'm impressed by Algorithmia's serverless microservice architecture – it's a great solution for organizations that want to deploy AI at any scale,"

~ Anna Patterson, VP of Engineering, Artificial Intelligence at Google

The AI Layer automates and optimizes deployment of your ML portfolio

Algorithmia Enterprise is the foundation layer for intelligent software. It enables organizations to centralize and productionize their models in a cohesive workflow. Companies use the discoverable registry for live algorithms from across teams and technology stacks, making any piece of code instantly available as a production-ready microservice. Algorithms range from your internal functions and machine learning models to third-party state-of-the-art contributions from university labs and independent researchers. Algorithmia Enterprise makes every version of every algorithm accessible as a unique, versioned, low-latency and highly reliable REST API endpoint that horizontally scales on demand and is ready to be integrated with any app, internet-connected device, or ETL pipeline. The software runs behind your firewall as a Virtual Private Cloud (VPC) or a dedicated on-premises cluster. It enables data scientists and application developers to experiment and deploy faster than ever, all while building up their internal algorithmic intelligence capability as a strategic asset.

Catalogs and enables algorithm searches:
Your models will be cataloged and searchable by every data scientist in your organization, eliminating redundant efforts. Every version is made accessible as a unique API endpoint that can be modified and saved to encourage collaboration and innovation. The model's authors can set various permissions to best align with internal security and compliance requirements.

Empowers users to operate in the language, framework, and cloud of choice:
The serverless AI layer is stack- and cloud-agnostic. It enables work in Java, Scala, Python, Ruby, NodeJS, Rust, or R and can combine microservices in multiple languages. Combine, pipe, and remix functions from different technology stacks to create new solutions.

Automatically, elastically scales:
Automatically scale up or down in response to fluctuating throughput needs, substantially reducing data center costs by avoiding over-provisioned clusters.

Provides stringent security:
Modify your serverless AI layer to meet any government or regulatory requirements from HIPAA to GDPR. Monitor cluster-wide health metrics and user-specific audit trails.

Enables deep learning:
Horizontally deploy and scale deep learning models over GPU clusters. Effortlessly transform complicated models to universal REST APIs.

Provides a library of algorithms:
Cherry-pick the latest algorithms from the public marketplace that are written by world-class researchers. Instantly expand the tool chain of your internal developers and scientists**.**

Has best-in-class support:
Algorithmia Enterprise is the technology behind the largest algorithm marketplace in the world and is offered with 24/7 hands-on engineering support.

> *"Today most AI/ML models are still being deployed manually, which requires a lot of time, coordination, and engineering resources. We're working with Algorithmia to help companies deploy, iterate, and scale faster on Azure with the Enterprise AI Layer,"*
> *~ Prashant Sharma, Senior Program Manager, Microsoft for Startups*
>
> *"Algorithmia empowers U.S. Government agencies to rapidly deploy new capabilities to the AI layer. The platform delivers security, scalability and discoverability so data scientists can focus on problem solving."*
> *~Katie Gray, Principal of Investments at In-Q-Tel*

Implementation

FAQs and tips on how to maximize your serverless AI layer potential

---

Installation:

Data Source
Algorithmia's serverless AI layer is cloud-agnostic and can be deployed on any private infrastructure behind strict firewalls, including hybrid and public cloud (i.e AWS, Azure, Google Cloud Platform, etc.) and on-premises. All deployments are isolated within their own private cloud or can be combined with a cloud.

DevOps
Algorithmia engineers take on the role of first-line DevOps support. Managed services will also receive at least, bi-weekly updates, and one week of internal testing.

Scaling
The AI layer can scale horizontally as worker nodes are automatically added or removed proportionate to demand. Local administrators can also manually control or edit this process, as desired.

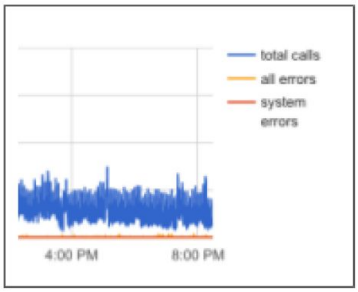
Monitoring:

Dashboard

Our bundled Grafana dashboards provides you total awareness of the state of AI deployments at your organization. Because all metrics and events are exposed via REST API endpoints you can also integrate them into your existing dashboard solutions to maintain legacy trackers. You can hone in on what matters most by designing a trigger that sends you an alert when specified thresholds have been passed.

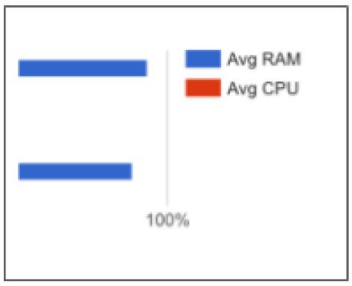
Monitor thousands of internal services in our clear and usable charts.

API Calls



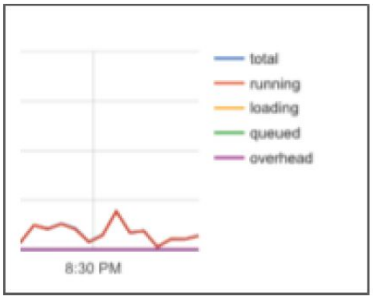
Service Health

Cluster Utilization

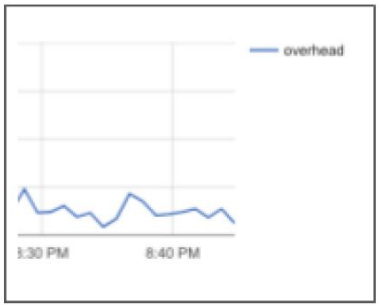


Algorithmia empowers you to monitor your APIs to prevent bottlenecks and spot trends.

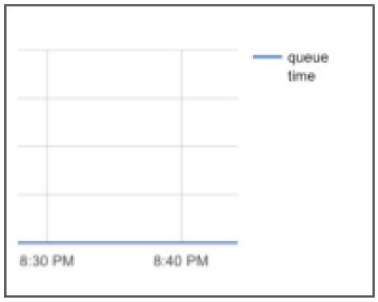
API Runtime



API Overhead Timing

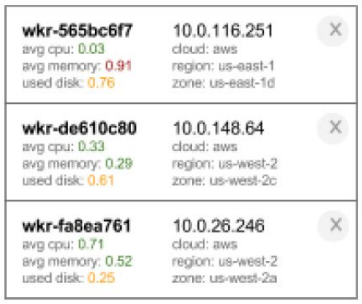


API Queue Timing

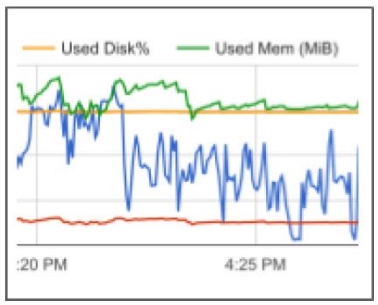


You have the ability to monitor, kill, and add individual worker nodes and inspect algorithms.

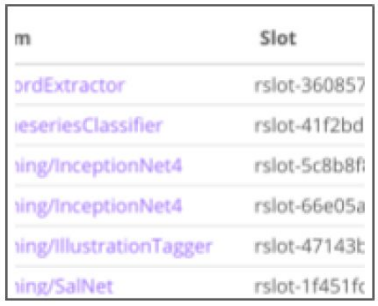
Worker Utilization



Resource Utilization

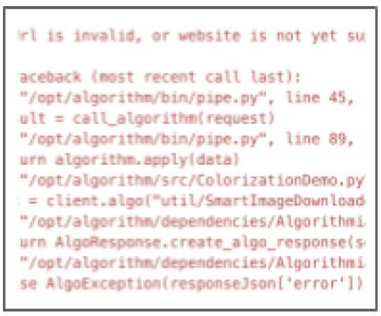


Running Algorithms

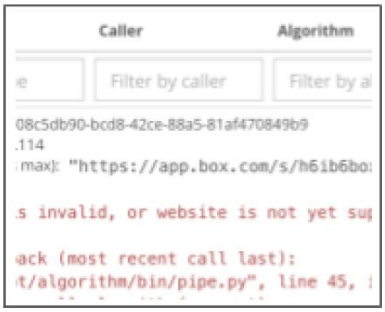


Meticulous error logs enable you to hone in on issues and improve future experience.

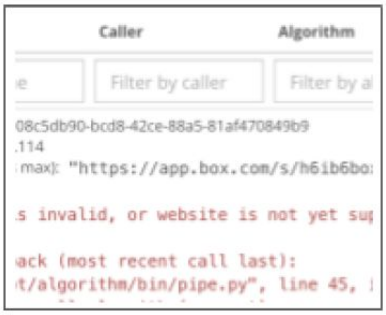
Exception Messages



Configurable Verbosity



Searchable Logs

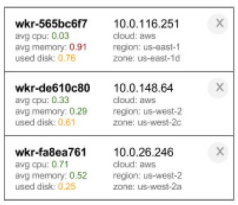


Security and Compliance

We specialize in operating in highly regulated industries

Cluster Configuration:

The Serverless AI Layer creates clusters that run code inside a private network, behind complex firewall rules. We can ensure your data never leaves specific boundaries and is compliant with government-grade security regulations.

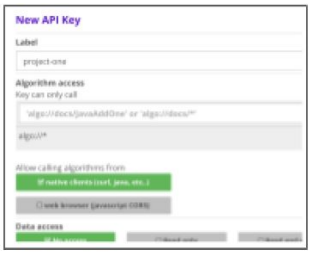


Data sovereignty

The AI layer can work as a multi-region cluster to comply with data sovereignty requirements. A typical setup allows for separate compute pools per region and a URL prefix for each region.

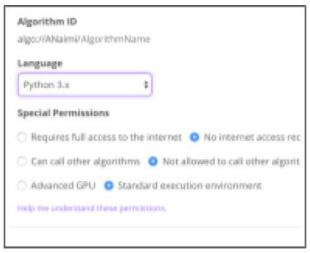
Session Isolation
Every API call operates in its own memory space and does not share or leak memory to other jobs. Each API call instantiates a dedicated Docker container, which is destroyed after execution, for perfect isolation at high performance.

Three Layer Permissions:

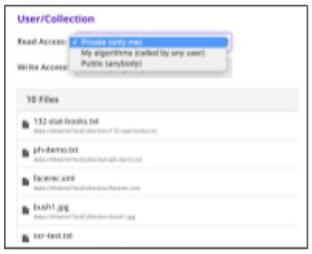


To safeguard data, the AI Layer implements permissions on three layers.

API keys
Users create a distinct personal API key for each project or experiment. Each key is specific to one algorithm and is individually auditable and revocable, and users are given explicit read/write permission over data sources.



Algorithms
Authors must specify whether their algorithm requires access to the network and if it requires access to call other algorithms. Without these permissions, algorithms will be executed in sandboxed Docker containers without those resources.



Data Sources
Data sources or collections can be configured to allow/disallow read/write access from other users. This can help control data access in multi-business line enterprises.

Organization & QA Workflows:

Your large organization can be divided into teams to mirror your own structure. Different teams have unique mandates and access rights but can also seamlessly share access for collaboration.

Teams
Users can create and join teams that share access allowing a group of people to maintain algorithms instead of one individual.

Private vs. Public
Algorithms can be marked private or public. Private algorithms will only be accessible to the owning individual or team. Public algorithms are accessible to all users on the AI Layer.

QA Publishing Workflow
You can build in quality assurance by creating a publishing workflow. To ensure a minimum level of quality while still allowing individuals to experiment, you can choose to have all algorithms approved by a compliance officer before it is accessible to the public pool.

Audit Trails:

Algorithmia Enterprise empowers you to identify suspicious activity and understand your utilization.

API Logging
The AI Layer logs every API call, showing which user used what API key to call what algorithm and version. You can configure the logging module to capture more or less data, depending on your needs.

Usage Attribution
The AI Layer measures each user's contribution to the total cluster utilization, allowing users to understand how teams and individuals are contributing to computing costs.

Error Logs
Every exception anywhere in the platform is captured including user-generated algorithms. These can be configured to capture full or partial input, error message, stack trace, and trigger an alert to take corrective action.

Your Options*

How Algorithmia Enterprise stacks up against the alternatives

Capabilities	Algorithmia	Sagemaker	AWS Lambda	Google Cloud	Azure Functions
Elastic horizontal Scaling	✓	✓	✓	✓	✓
Serverless Microservices	✓	✓	✓	✓	✓
Real time usage analytics[1]	✓	✓	✓	✓	✓
Real time cluster analytics[2]	✓	✓	✓	✓	✓
Automatic API generation	✓	✓	✓	✓	✓

Chain API Calls	✓	✓	✓	✓	✓
Org structure-based permissioning[3]	✓	✓	✓	X	X
Support for python, r, ruby java, scala, nodeJS	✓	✓	X	X	X
GPU for Deep Learning	✓	✓	X	X	X
Custom Docker Environment[4]	✓	✓	X	X	X
50 minute API call timeout [5]	✓	✓	X	X	X
Searchable model repository	✓	X	X	X	X
Algorithm marketplace	✓	X	X	X	X
Git Workflows and push to deploy	✓	X	X	X	X
Cloud Agnostic: hybrid and on-premises	✓	X	X	X	X
Data Source Agnostic	✓	X	X	X	X
Automatic Versioning with alias	✓	X	X	X	X

*This chart compares Algorithmia Enterprise with the services offered under Google Cloud Function, Azure Function, AWS Lambda, and AWS Sagemaker. Other services offered under different payment packages were not considered.

[1] Real time is defined as within 1 minute.

[2] Real time is defined as within 5 seconds.

[3] Permissions can be applied to one person and/or configured to reflect org structure by choosing to add team members from a simple dropdown. Permissions are automatically applied to everyone on the team.

[4] End user can send individual containers that can be reused on any platform.

[5] A green check indicates 50 minutes or longer, a red x indicates less than 50 minutes.