Wallaroo

Put Your Machine Learning to Work

The fast, easy, super-charged engine for deploying, running, observing and optimizing machine learning in production

> **“Very few companies have the resources to fully leverage their data to build better products or optimize their operations. Wallaroo levels the playing field by making it easy, fast, and low cost for any enterprise to take their boldest data and ML ideas live to deliver results.”**

VID JAIN
FOUNDER & CEO, WALLAROO

How we got here

We began in 2017 as a tight-knit group of engineers dedicated to solving the increasingly common problem of analyzing large amounts of data via computational algorithms efficiently and at scale.

By applying our collective expertise in **building distributed computing systems** in industries such as high-frequency financial trading and AdTech, we built a high-performance compute engine like nothing else on the market. While our customers could now efficiently analyze their data and use it to run machine learning (ML) models at scale, they soon pointed out their next biggest challenge: bringing those models online easily, and then understanding how the models were performing to sustainably generate business value.

Like most organizations, they were doing everything by the book: data scientists would build ML models to solve a business problem, and engineers would launch them using a patchwork of open-source software and containerized model approaches. What they found, however, was that getting each model to production was like pulling teeth.

Models often had to be painstakingly re-engineered, the deployment software couldn't process data fast enough — even when running on an alarming amount of computing resources — and it was unnecessarily difficult to see how models were performing to measure their ongoing accuracy.

We knew there had to be a better way.

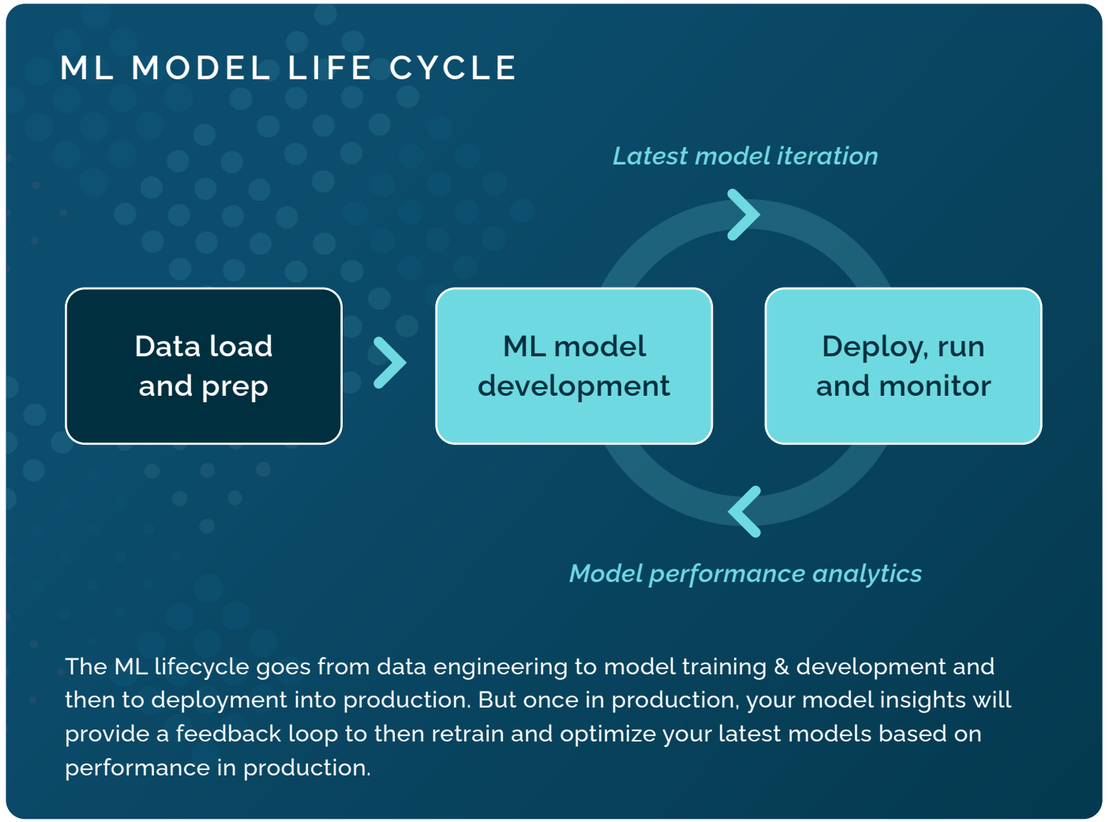
Machine learning & applied AI's "last mile problem"

Data science is a modern-day superpower, and enterprises around the world know it. Over 90% of Fortune 1000 companies are investing in Big Data, analytics, and artificial intelligence (AI), reaching over $700 billion being poured into teams of data scientists and engineers to revolutionize the way they do business.

Yet machine learning is hard, and the last mile of ML — getting the models into production to impact the bottom line — is especially hard. If businesses can't do this easily or at scale, their AI initiatives will fail, resulting in significant costs in terms of budget, manpower, and disillusionment. According to Gartner, less than half of AI prototypes make it to production, and in the end, only about 10% generate substantial ROI.

Deployment solutions — whether containerization, cobbling together various existing technologies, or customizing an analytics workhorse like Apache Spark — are cumbersome, limited in scope, expensive at scale, prone to failure, and unable to run ML models against batch and streaming data.

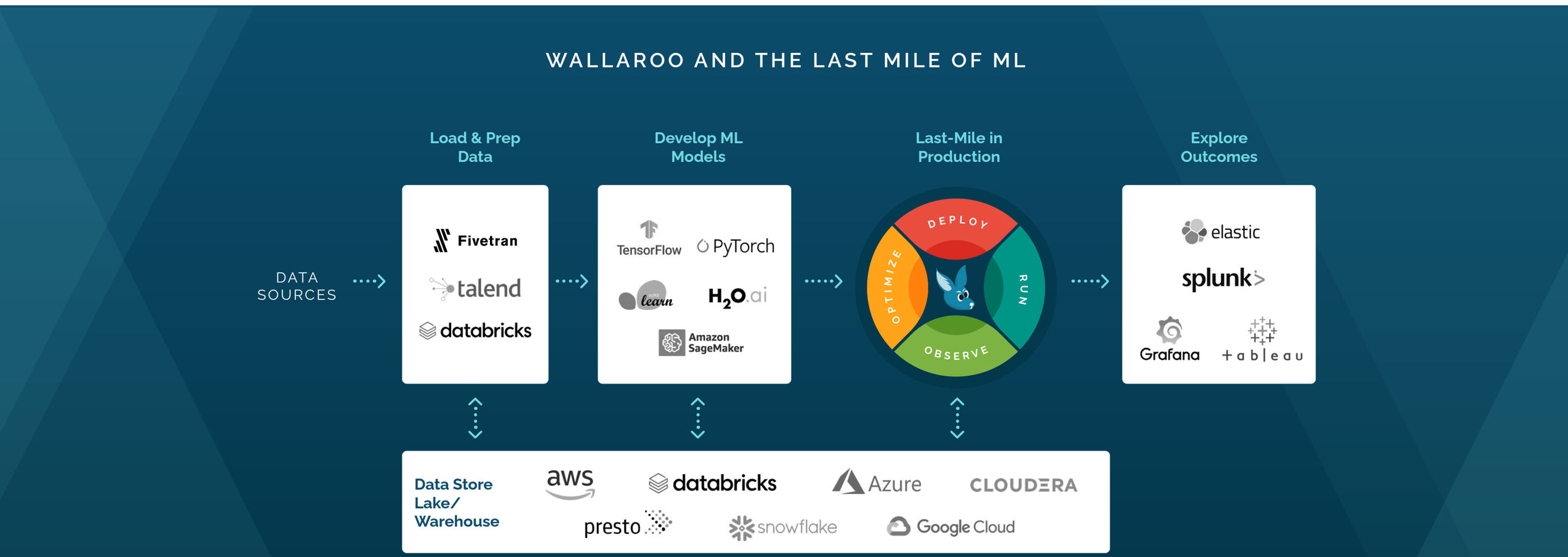
With investments in AI only trending upwards, companies hoping to achieve a return from their data will never reach their full potential as long deployment lead times and high cost to run and maintain the necessary infrastructure often outweigh the benefits.



INTRODUCING WALLAROO:

The different, better way to deploy ML

Wallaroo is a breakthrough platform for the last mile of ML, providing a simple, secure, and scalable deployment capability that fits into your end-to-end workflow.



Wallaroo gets your ML to business results faster, easier, and with a far lower investment. By streamlining the deploy/run/observe parts of an ML lifecycle and giving data scientists the freedom to use the tools they already know, Wallaroo enables your team to:

- **Deploy models in seconds**
- **Analyze data up to 12.5X faster**
- **Reduce compute costs by 80%**
- **Iterate quickly and scale easily**

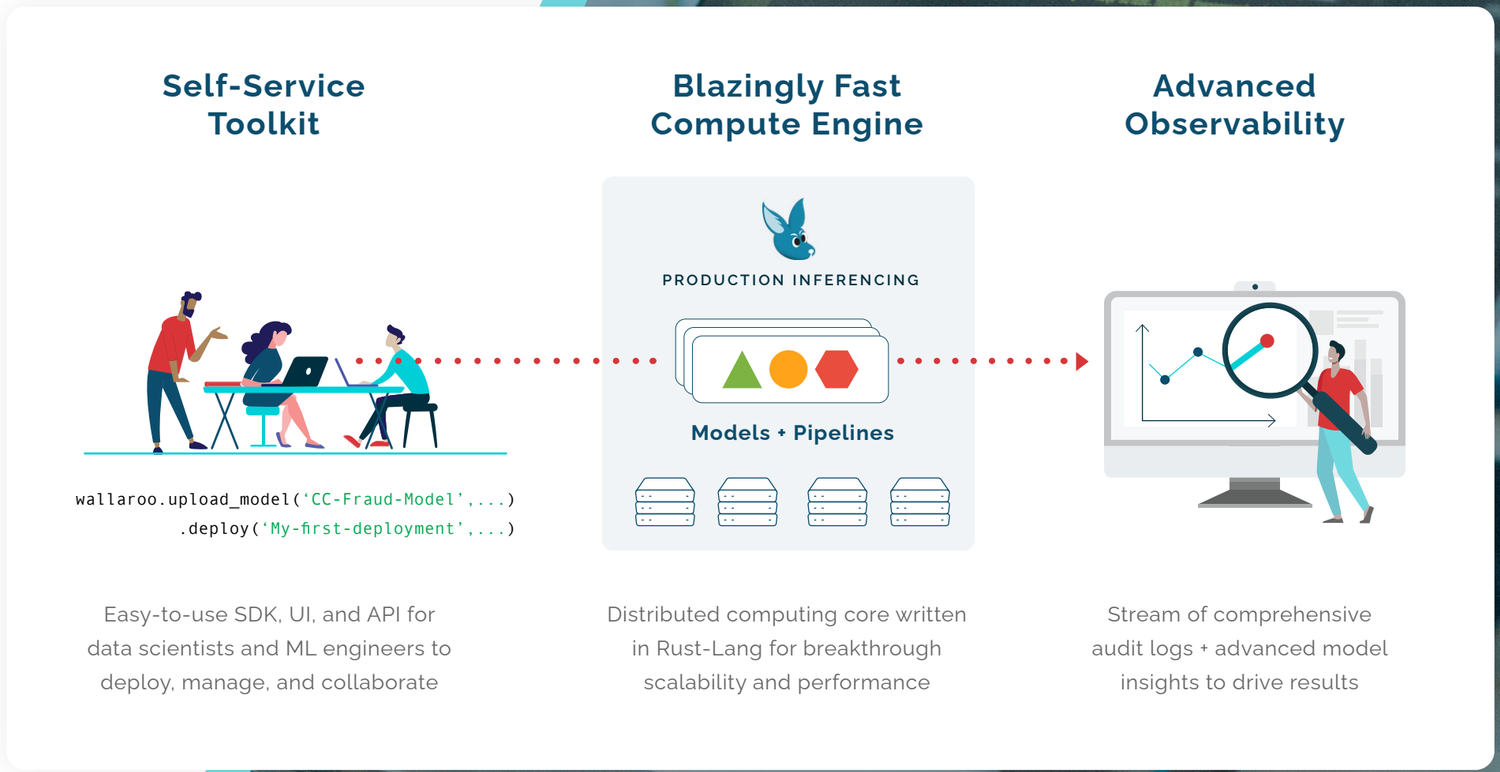
1. Takes **weeks or months** to deploy a model into production	**Deploy a model against live data in seconds**
2. **High cost** of data engineering and operational overhead	**Your existing teams can do more and new projects can launch quickly**
3. Too **expensive to scale** production environments	**Scale to process more data and more models using less infrastructure**
4. Current solution(s) **can't analyze** streaming data in real-time	**Ingest, fuse and analyze any amount of data in real-time**

Before vs. after using Wallaroo for production AI

1 easy-to-use platform.
3 key components.

This is where we normally get the question: "okay, so I see what you do and the problem you solve, but what exactly are you?" The Wallaroo platform is composed of 3 key components:

- A self-service toolkit for easy model deployment and management
- A distributed compute engine allowing you to inference faster using fewer servers
- Observability, insights, and dashboards to monitor the ongoing performance of your models in production



Wallaroo's 3 key components

Self-service toolkit: Swift and simple ML deployments

This is the component that enables data scientists to deploy their ML models against live data in two clicks of a button—whether it's to a testing, staging, or production environment. With an intuitive SDK, UI, and API along with support for common data workflows, Wallaroo takes care of the details to let data teams focus on the bigger picture.

- Easily deploy, test, and iterate ML models using the frameworks your team already know (e.g., TensorFlow, PyTorch, Scikit-learn, and XGBoost).
- Run batch jobs or streaming to capture valuable market insights as they happen.
- Refine and immediately redeploy new and improved models without complex re-engineering or operational headaches, including model management and data scientist collaboration features.

Distributed computing engine: Lightning-speed computing at lower cost

Here's where the magic happens. This highly performant, easily-scalable engine can analyze up to 100K events per second on a single server (beating the industry average of 5,000 events per second), making Wallaroo the fastest platform on the market for production ML.

- Run multiple models on a single server to drastically reduce computing costs and maintenance overhead.
- Analyze data at record-breaking speed and react to market changes in real-time for a sharper competitive edge.
- Leverage an ultrafast environment for production model scoring and pre/post-processing, with support for custom data operations.
- Scale down to run at the edge.

Typically with customers' transformer models, computer vision, complex neural networks, and NLP models we have seen 5X – 12.5X faster analysis using 80% less infrastructure compared to their previous deployments.

Observability and model insights: Real-time metrics to measure business impact

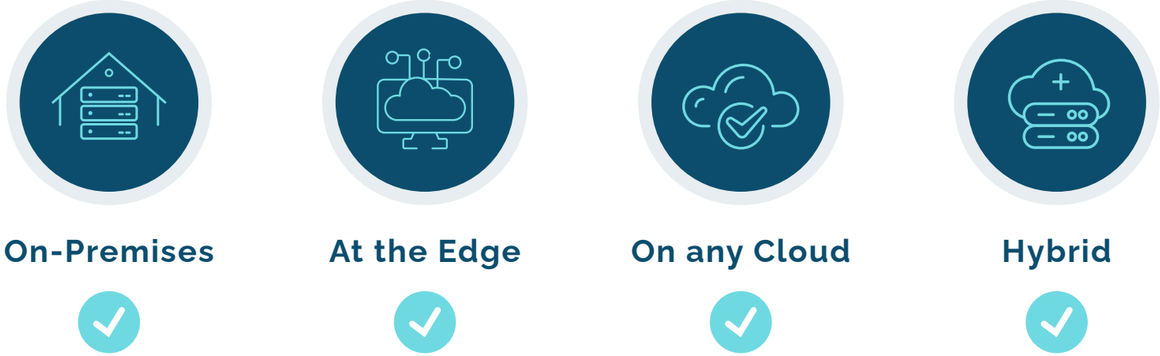
This is where it all ties together. A simple, easy-to-use interface allows anyone on your team to explore powerful metrics and detailed analytics so they can effectively track, measure, and help improve your ML's performance.

- Drill down into computing specifics like throughput, model latency, and benchmark performance for in-depth analysis.
- Validate model inputs to guard against invalid or unexpected data.
- Monitor the behavior of models and their inputs over time to understand when changes in the environment might require a model refresh.
- Use A/B testing along with shadow and staged deployments to make sure you're always using the highest-performing models.
- Simplify audits with detailed event logs and visibility into everything your compliance and risk management team needs to perform their jobs more efficiently.

Your data. Your tools. Your ecosystem.

Enterprises will often look to all-in-one MLOps platforms such as SageMaker, Databricks, or DataRobot to simplify deployment. However, these platforms force data teams to standardize on proprietary tools, processes, and formats. These tools will then lead to complexity as different business units within the same company might use different data platforms. One of our customers, for example, is all-in on a certain cloud, but because of mergers & acquisitions, its data engineering teams are supporting different deployment processes for multiple clouds.

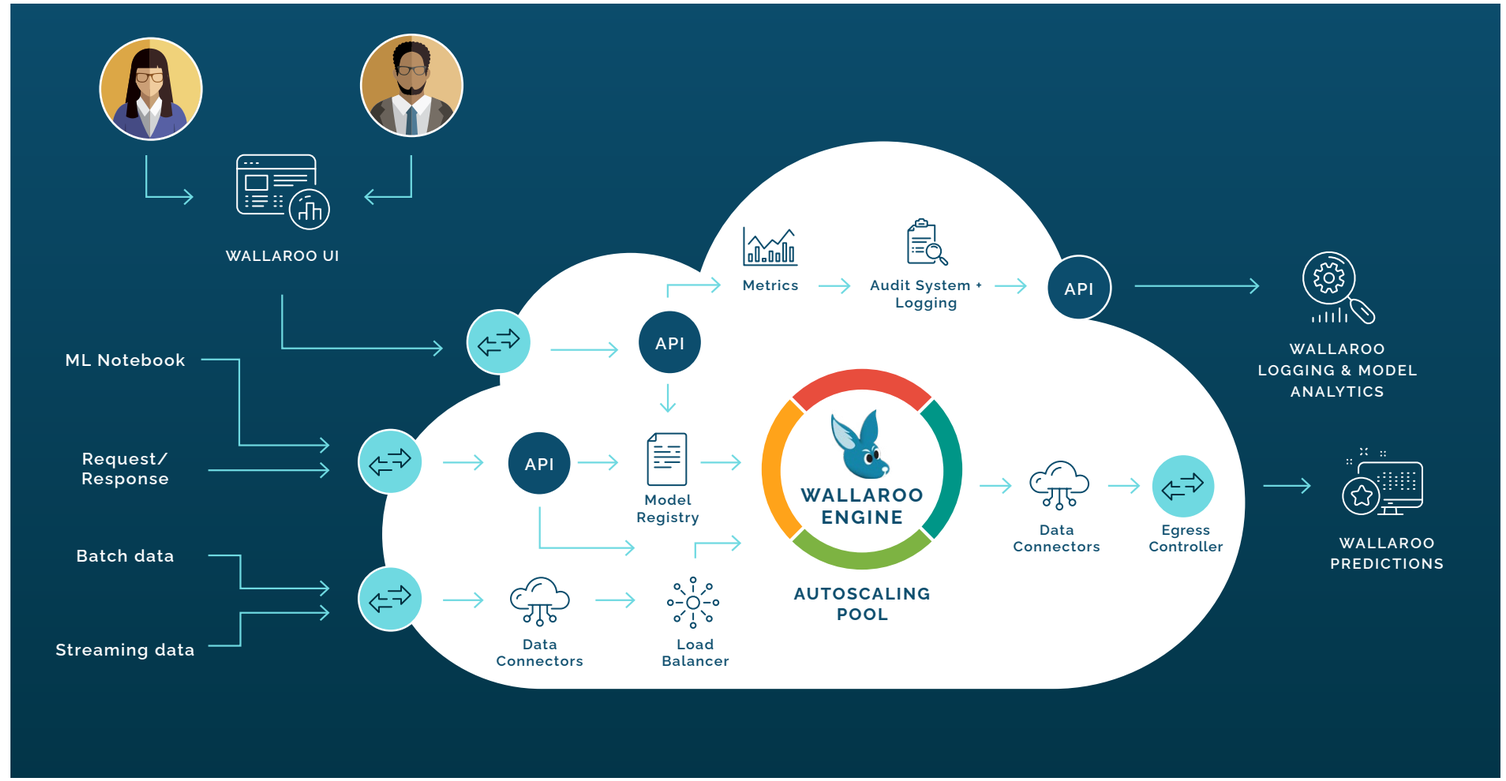
In response, companies will spend countless resources building their platform in-house, cobbling together open-source technologies such as Spark and MLflow, which might work within the current ecosystem but at the expense of performance and model observability.



Our Connector Framework neatly plugs our platform with your incoming and outgoing data points and takes care of the integration to get you up and running in no time.

- Quickly connect with popular data sources and sinks, like Apache Kafka and Amazon S3.
- Plug in custom integrations and your own in-house solutions.
- Rely on rapid support if you need to integrate something that isn't available out-of-the-box.

You can also rest assured that all your data will remain yours. Everything that goes in and out of Wallaroo is private, secure, and only visible to those with permission to see it.



Wallaroo Cluster

Built from the ground up for speed, efficiency, and ease-of-use

Wallaroo was specifically engineered for modern machine learning deployments, unlike Apache Spark, or heavy-weight containers. The core distributed computing engine is written in Rust language, not Java — so it runs at C-speeds and is Python-friendly. Our SDK was designed with data scientists in mind and has incorporated direct feedback from our customers.

Easy-to-deploy

- Designed for data scientists and ML engineers alike.
- Intuitive UI and dashboards.
- API enables flexible application & workflow integration.

Efficient Runtime

- Efficient event-by-event runtime engine for batch or streaming analysis.
- Ideal for big data or fast data.
- Supports high-throughput, latency in the microseconds.

Real-time Insights

- Experimentation, A/B testing, and champion/challenger.
- Data validation, model anomaly and data drift alerting.
- Production audit logs for deeper analysis and compliance.

Plug and Play

- Plugs into your on-prem and cloud data environment.
- Your data scientists can keep their popular ML frameworks.
- Single uniform deployment for diverse enterprise teams.

How leading companies are fueling innovation with Wallaroo

AdTech

CHALLENGE:

An ad exchange company needed to process up to 80 million events per second in real-time to optimize individual ad auction bids

OBSTACLES:

- 6 months to build an ML prototype using Apache Spark and Flink
- Required infrastructure 5x over budget

Wallaroo advantage

2 months to build and deploy ML prototypes

1 trillion events processed on just **10 servers**

↓80 % reduction in computing costs

Real Estate

CHALLENGE:

An international investment and real estate company wanted to deploy 36 pricing models with real-time customer segmentation to dynamically price thousands of storage units

OBSTACLES:

- 9 months to build an ML prototype
- 144 servers needed to run the pricing models
- Several technologies to ingest, compute, and run models

Wallaroo advantage

2 months to build and deploy ML prototypes

18 servers to run 36 pricing models

1 platform to take care of everything

↓87 % reduction in computing costs

Manufacturing

CHALLENGE:

A multinational consumer products company wanted to combine real-time demand data with supply and manufacturing data to continuously optimize their supply chain

OBSTACLES:

- 12 months to build an ML prototype
- No capacity for real-time processing
- Data engineers needed for IoT streaming

Wallaroo advantage

- Unmatched real-time data processing
- **3 months** to build and deploy ML prototypes
- Existing team easily managed the platform

Cybersecurity

CHALLENGE:

A Fortune 100 enterprise needed to deploy over 100 ML models to detect security breaches. Plus, the models had to be retrained and updated monthly

OBSTACLES:

- 2 weeks needed to retrain and deploy a new model
- 500 servers to run them
- 5 data scientists required for deployment

Wallaroo advantage

- Seconds to redeploy updated models
- **96 servers** to run over 100 ML security models
- Data scientists free to focus on innovating
- **84 %** reduction in computing costs

IoT

CHALLENGE:

The US Military needed to analyze petabytes of daily IoT data in the cloud and across millions of edge devices, including drones and ships, to swiftly detect security anomalies.

OBSTACLES:

- Common analytics solutions too bulky for edge environments
- Most data wasn't being analyzed
- Limited capacity for real-time data processing

Wallaroo advantage

- Minimal infrastructure in their cloud environment
- **50 MB** runtime for low-resource edge environment
- Real-time streaming data analysis for instant insights

Bring your boldest AI projects online with Wallaroo

Wallaroo is a platform that enables the future of AI and analytics we always wished we had: **one where cutting-edge AI and ML can be deployed in seconds, and data teams can deliver higher value at a lower cost**. We built it so your team can spend less time making your data work with your software, and more time making your data work for your business.

If you want to explore further with us, such as access to the full SDK, giving us specific feedback about functionality/semantics/integration, or discussing how we can help with your use case, email us at **deployML@wallaroo.ai**

You can find more information about Wallaroo at **wallaroo.ai/blog**

Wallaroo