Whitepaper:

Deploying Machine Learning at Scale with Serverless Microservices

TABLE OF CONTENTS

ABOUT ALGORITHMIA:

Algorithmia empowers every developer and company to deploy, manage, and share their AI/ML model portfolio with ease. In addition to the Algorithm Marketplace, Algorithmia uses the serverless AI Layer to power our Hosting AI/ML Models and Enterprise Services.

Our platform serves as a data connector, pulling data from any cloud or on premises server. Developers can input their algorithms in any language (Python, Java, Scala, Node, Rust, Ruby, or R), and a universal API is automatically generated.

Democratizing access to algorithmic intelligence.

Sign-Up for Free or Schedule a Demo Today

Introduction

This whitepaper surveys the challenges inherent in Machine Learning deployment, how emerging trends of Microservices and Serverless architecture can help, and why an AI Layer might be a great fit for your organization.

Many companies today are struggling to answer a simple question: how should they deploy machine learning models at scale?

> *72% of business leaders have indicated that they view AI as a business advantage (PwC).*

More and more frequently, companies are creating internal initiatives for garnering a competitive advantage using AI and ML. A number of developing trends are driving this increased usability of Machine Learning in the coming years:

- Wide availability of data storage and compute power
- Improved and robust open source tools and frameworks
- A growing appreciation for algorithmic decision making.

Designing and deploying Machine Learning at scale is challenging, no matter the size of your team. Data Scientists are simply not trained in the often overwhelmingly complex discipline of **deployment**, or turning their models into scalable applications. The intricacies of load balancing, event handling, and container management are a segment of the Machine Learning pipeline in of themselves, and there's no straightforward playbook for how to make them work together.

If you want to see meaningful ROI on your Machine Learning investments and build a competitive advantage this year, you'll first *need* to solve this last mile deployment problem.

Challenges With Deploying Machine Learning At Scale

Even once you've already hired a Data Science team who can build models, there's a real gap between *building models* and *building models for scale*. Data Scientists are experts in creating models locally, but deployment is a totally different challenge. It draws on skill sets from distributed systems, senior-level software engineering, and cloud architecture. These are rarely part of a Data Scientist's toolkit.

In addition to the hurdles faced in any large scale software deployment job, Machine Learning introduces a few new quirks into the equation:

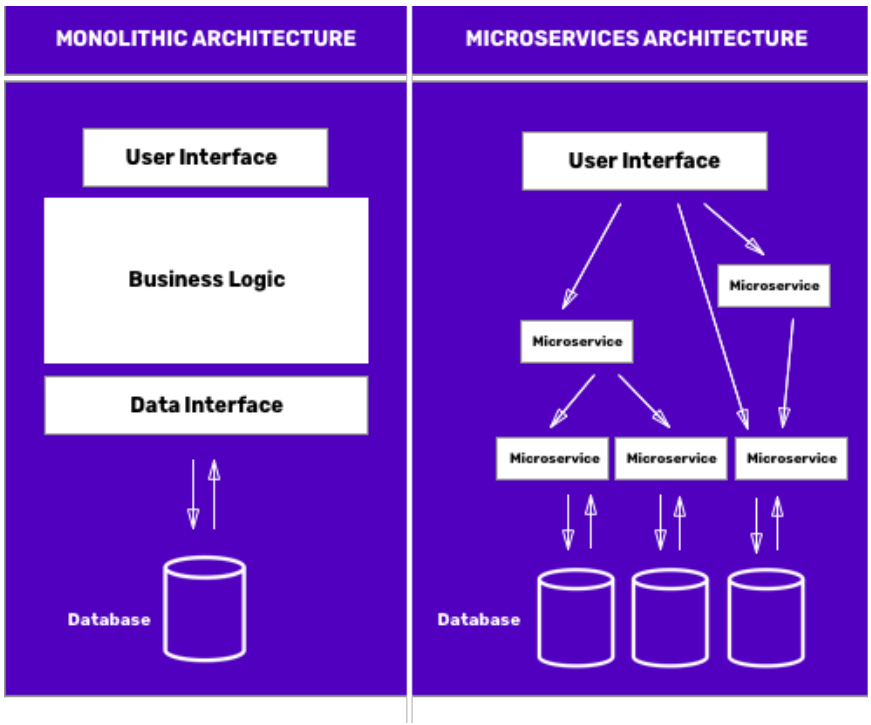
- **Multiple Languages**: (Python, R, Scala) are routinely used, even for different parts of the same model.
- **Parallel GPU Usage**: Deep Learning often requires a lot parallelization, like what's offered by GPUs. These are rarely part of your existing core infrastructure.
- **Unpredictable Costs**: Usage of Machine Learning models for inference can follow a spiked and unpredictable pattern, making cost-efficient scaling difficult.

To deal with this complexity, organizations have turned to creating teams of DevOps engineers exclusively focused on deploying and scaling Machine Learning models. These engineers focus on building cloud infrastructure, implementing APIs, and designing automation that enables Machine Learning engineers to get their models into production.

Unfortunately, these engineers are far and few in between as this developing field becomes more grounded. The experience required for these job postings is extensive and spans disciplines, some of which have only been around for the past few years (like Docker and Kubernetes). Even once hired, having Data Scientists rely on other teams is suboptimal. These modern DevOps teams can struggle to create a reliable deployment system that satisfies the involved stakeholders and their competing needs.

DevOps teams focused on machine learning still struggle to create reliable deployment platforms that suits the needs of the organization and allow them to scale efficiently and at cost.

Microservices And Machine Learning Models



Microservices is an emerging software architecture that's quickly picking up steam among both large enterprises and smaller early stage companies.

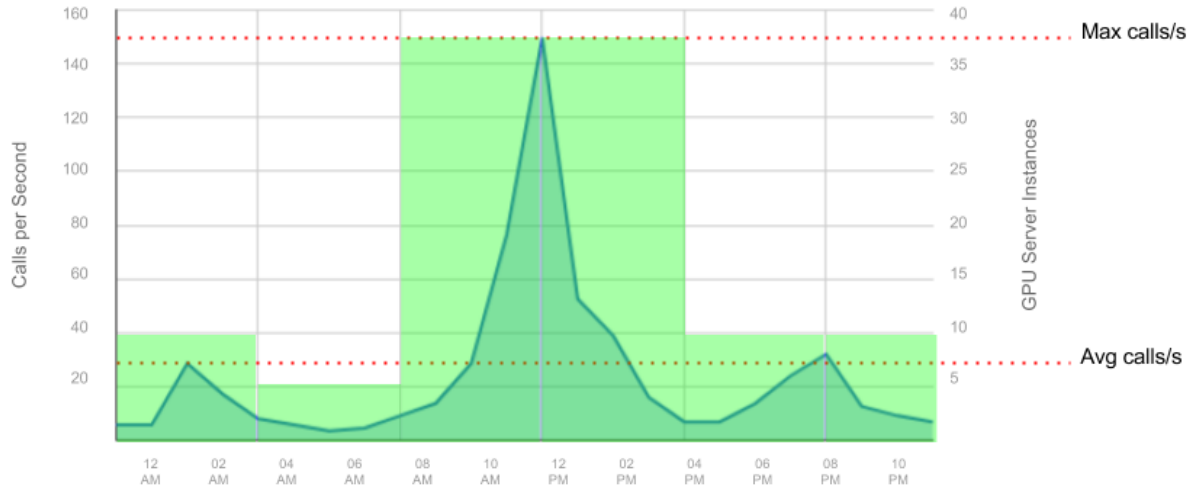
> *91% of respondents to a LightStep survey are using or are planning to use Microservices in their organization.*

Instead of developing most or all of your application code in one place – or as a *monolith* – Microservices-style development packages each application component in an individual piece, usually with a RESTful API endpoint for access. A production application developed as Microservices has communicating components, all developed and maintained separately.

> *"Microservices are a natural paradigm for developing a machine-learning pipeline." – Nitin Bandugula* (MapR)

Machine Learning is an ideal application for Microservices. Models are inherently separate: they need to interact with each other but also maintain independence. For models and parts of a pipeline, API endpoints are an operational way to integrate and create an easier development process.

The Serverless Movement And Elastic Scale



Compute time (and associated costs) of an auto-scaling non-serverless architecture

Along with Microservices, the *Serverless Framework* is redefining how some organizations deploy their software.

Two challenges that a serverless architecture solves are:

1. DevOps doesn't need to manually deploy and constantly manage VMs
2. Data Scientists can pipeline models together with ease

Instead of managing and provisioning your own servers (even on a public cloud provider), a Serverless service will take care of any infrastructure decisions including elastic scaling, while all you need to do is upload your code. It's often referred to as Functions as a Service, since you run your code without worrying about anything else.

> "*While an initial IaaS [Infrastructure as a Service] deployment can take hours or days, a typical serverless application can be deployed to an established account for the first time within minutes.*" (Deloitte)

Compute time (and associated costs) of a serverless architecture: you only pay for what you use

A Serverless architecture is an excellent fit for Machine Learning workloads.

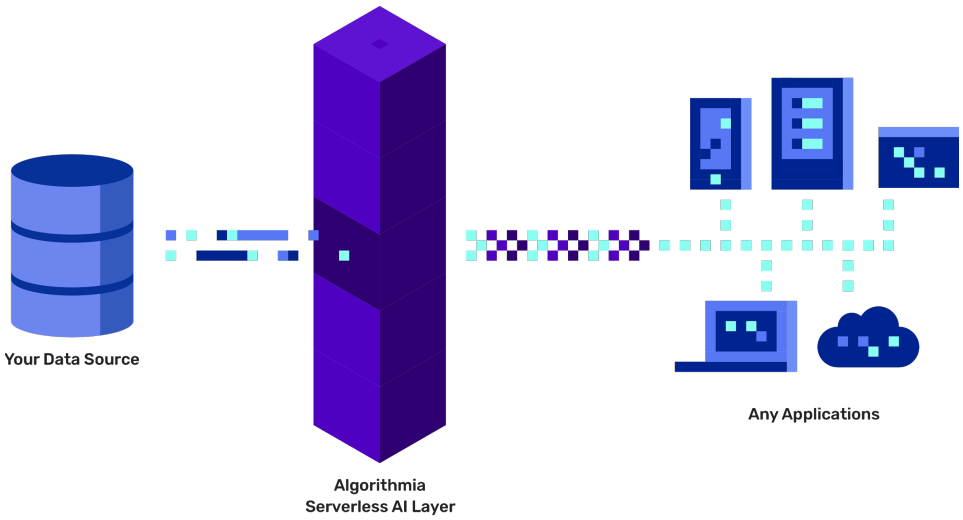
- On the training side, most Data Scientists don't have the expertise to **distribute complex training loads** across both GPUs and CPUs. A serverless architecture automates that process.
- On the inference side, calls to Machine Learning applications can be unpredictable and clustered. **Scaling up and down to meet demand** while avoiding high costs is crucial.
- Both training and inference often require use of **specialized GPU hardware**, which shouldn't be part of your core infrastructure.

The Serverless paradigm is also a strong fit for a Microservices-based architecture. With parts of your application packaged as individual functions, deploying them on a Serverless platform is as easy as a code upload.

The AI Layer: Deploy Your Machine Learning As Serverless Microservices

With the challenges of deploying Machine Learning and the benefits of Serverless and Microservices in mind, the ideal deployment platform needs to satisfy a number of competing needs. A Serverless architecture that allows you to automate the complex infrastructure decisions along with the interoperability offered by Microservices is key.

But Data Science and DevOps teams don't exist in a vacuum. Algorithms and models need to be versioned, managed, and searchable. Teams working together are always stronger, and a varied set of business-focused features are integral to a useful Machine Learning platform.



> *"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

This is exactly what we've built at Algorithmia. Our platform answers the key questions that drive your final push to ROI on Machine Learning projects.

1. **How Do I Free Up My Data Scientists To Focus On Their Core Competencies?**

The Algorithmia platform takes deployment time down to minutes. Data Scientists can just Git push or manually upload the models they've already built to our web-IDE, and we take care of the rest. Automating the DevOps process means Data Science gets to focus on Data Science.

2. **How Do I Make Responsible Infrastructure Decisions That Fit The Needs Of My Machine Learning Workloads?**

Our platform deploys serverlessly across both CPUs and GPUs. You don't need to focus on any major infrastructure decisions, and can avoid the vendor lock-in from relying exclusively on one cloud provider. Algorithmia is also the only serverless platform with GPU support.

3. **How Do I Manage Multiple Languages And Create Interoperability Between My Models?**

Algorithmia deploys your Machine Learning models as Microservices with API endpoints, along with clients available in all major programming languages. You can write your models in whatever language you want, call them in whatever language you want, and be sure that they'll all deploy together seamlessly.

4. **How Do I Version, Catalog, And Manage The Algorithms My Team Creates?**

Through our experience supporting Fortune 50 clients to government agencies, we've developed a robust enterprise ready feature set. We automatically version your models, offer searchability for your algorithms, and boast enterprise-grade security features at any level.

The Algorithmia platform offers benefits across the team structure of any Machine Learning project:

For Management	For Data Scientists	For DevOps and Engineers
Final step on the road to ROI for Machine Learning	Use the language(s) you want	Massively parallel computing
Feel safe with robust security and management features	Save time by pipelining	Only pay for what you use
Keep Data Science focused on their core competencies	No more DevOps required	CPUs and GPUs optimized for Machine Learning
Pay as you go model for elastic scaling	White-glove support	Serverless architecture that scales to meet your needs

> *"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

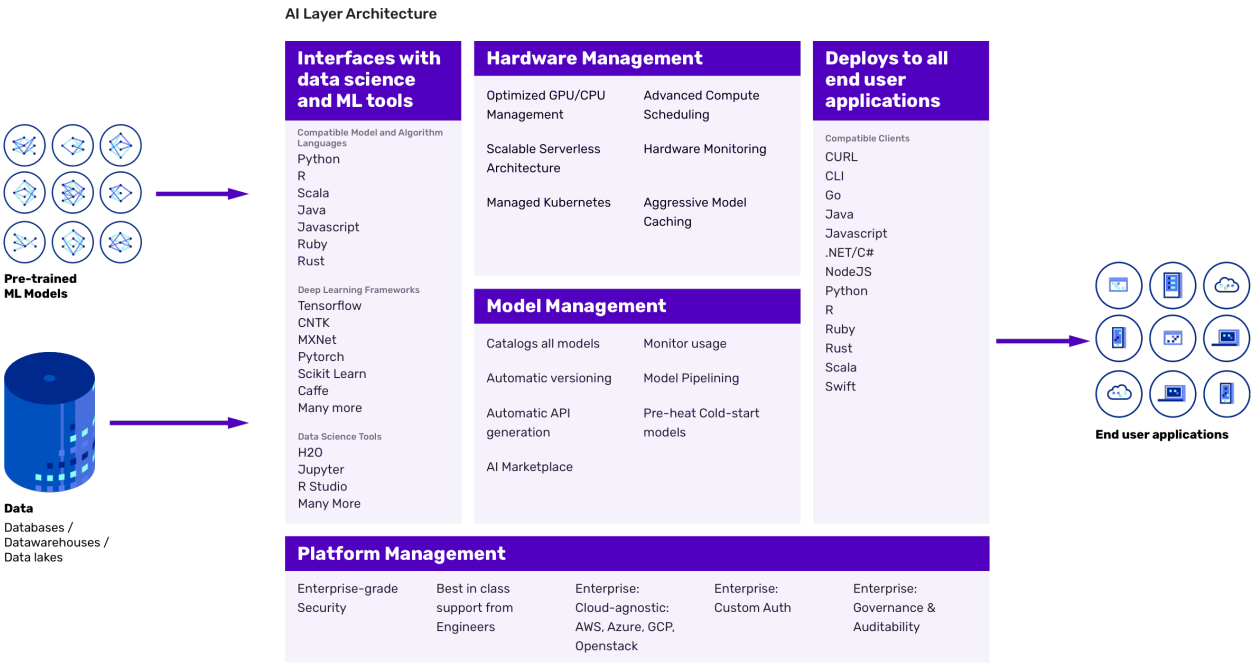
The Future of Machine Learning At Your Organization

The future of Machine Learning is a world where you don't need to worry about the technical challenges of scaling, and can focus on combining acute domain expertise with modeling to create meaningful outcomes.

We believe that every application will have an AI Layer – a platform that stands between the data backend (data and models) and the end user application. The AI Layer helps cross the river and makes it easier to create feedback loops between data and use case.

As managers allocate budget and revisit their investments in data and Machine Learning, the last mile will prove to be a key bottleneck in reaching meaningful ROI.

Algorithmia's platform is an excellent option to bridge that gap.



ALGORITHMIA

Deploy Machine Learning at scale with the Serverless AI Layer.

Sign-Up for Free or Schedule a Demo Today