



The Full Machine Learning Release Guide for Startups

Who is this Presentation for:

Developers



Looking to incorporate ML into apps and services

Data scientist



Looking to be more productive

ML experts



Looking to stay at the bleeding edge



How AWS drives growth for ML startups



Activate

- Free ML training
- Free ML workshops
- Credits
- Office hours
- Dedicated team
- Reference architecture



Build

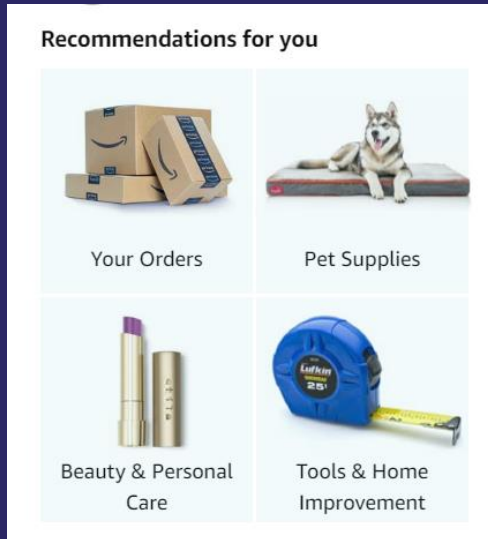
- Broadest and deepest set of products
- Technical support
- Access to beta products and roadmap
- Cost optimization reviews
- Speaking engagements
- Co-marketing



Connect

- Fundraising (VCs)
- Enterprise customers
- Partners (APN)
- AWS Marketplace
- Deep industry expertise

Amazon's machine learning innovation at scale



4,000 products
per minute sold
on Amazon.com



1.6M packages
every day



Billions of Alexa
interactions each week



First Prime Air delivery on
December 7, 2016

Amazon's machine learning innovation at scale

- Amazon went through a multi-year machine learning (ML) journey to be the ML-driven company you see today. We have been applying ML in areas such as personalization and supply chain for over 20 years. We've also improved that original personalization model significantly over time and moved it to other products, such as Amazon Prime.
- We use ML throughout our fulfillment process, including developing a forecast system that can predict the appropriate amount of demand for each product we sell worldwide to deliver on customer expectations for convenience, cost, and delivery speed.
- We've developed natural language processing technology to give customers an entirely new way to interact with technology through Alexa. We've also developed groundbreaking technology, such as autonomous flight through Prime Air drones and robotics in our fulfillment center to get packages to customers faster. This was a significant culture change, but it's something every organization can do.

Why Startups Chose AWS for ML



Why Startups Chose AWS for ML



Broadest and deepest set of AI and ML services

200+ new features and services launched in 2020 alone

Solutions for everyone

Support all three of the major machine learning frameworks



Accelerate your adoption of ML with SageMaker

Single IDE for the entire ML workflow

At least 54% lower TCO

Up to 70% cost reduction in data-labeling

Up to 90% cost reduction with managed spot training



Built on the most comprehensive cloud platform

Highly secure, reliable, fully featured data store

The strongest set of compute, storage, security, database, and analytics capabilities to build upon

Thousands of startups powered by AWS ML

coinbase



Aurora

common
edits

STUDIO71

stripe



CONVOY

Root
Insurance Co

DELHIVERY



TransferWise

affirm



DEEPMAP

cravelabs

TECTON



ABACUS.AI



AI/ML with AWS

Innovation, choice, and flexibility



100,000+

customers have used machine learning (ML) on AWS

250+

new capabilities for ML and artificial intelligence (AI) in just the last 12 months

92% of deep learning (DL) in the cloud runs on AWS

91% of cloud-based PyTorch runs on AWS

AWS ML SOLUTIONS

Reduce training time by 50%

Provide 90% scaling efficiency

Deliver 3x faster network throughput

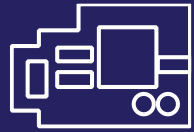
Improve price and performance by 25%

AI/ML with AWS

Innovation, choice, and flexibility

- AWS is the vendor of choice for over 100,000 customers and a leader in offering the broadest, deepest compute infrastructure for AI/ML training and inference.
- You can choose from a range of Amazon EC2 instances based on the latest CPUs, GPUs, and custom accelerators, coupled with industry-leading networking and storage to meet your budget needs and the performance requirements of your models.
- AWS is investing and innovating to deliver a range of high-performance and low-cost infrastructure options so you can choose the option that best fits your needs.
- For example, AWS is building silicon from the ground up that is optimized for ML inference performance. We used learnings from our Graviton CPU and Nitro system innovations to build the AWS Inferentia chip—a custom chip that delivers the lowest cost of inference in the cloud.
- By running ML workloads on AWS, you get on-demand access to high-performance, low-cost, easy-to-use infrastructure services for training and deploying ML and DL models.

Common startup challenges



Developers

Data silos, prep and cleaning, and overall data workflow



Skills Gap

Not enough people and not enough experience



Business Case Uses

Finding the right business use cases that could benefit from ML



ML Operations

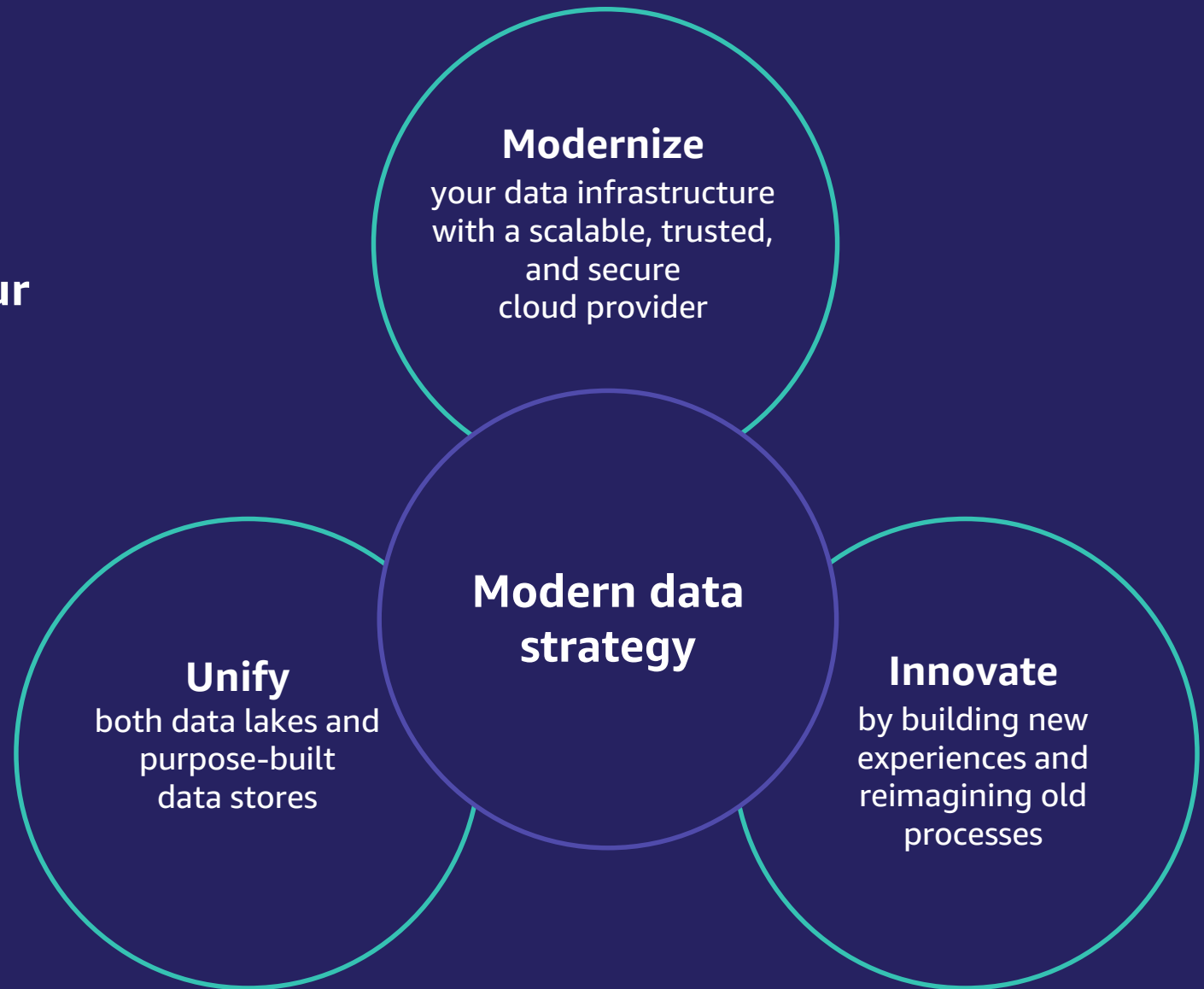
Creating and managing ML workflows is time consuming and complex

Modern Data Strategy with AWS



Modern Data Strategy with AWS

Modernize, unify, and innovate your way to a modern data strategy



Modern Data Strategy with AWS

Harnessing data is the next wave of digital transformation

1

AWS can help your startup to:

Modernize your data infrastructure with a scalable, trusted, and secure cloud provider. Organizations running legacy, on-premises data stores or self-managing in the cloud still have to perform management tasks, such as database provisioning, patching, configuration, or backups. AWS can help take care of this for you. Benefit from our unmatched experience, maturity, reliability, security, and performance for your most important applications.

Modern Data Strategy with AWS

Harnessing data is the next wave of digital transformation

2

AWS can help your startup to:

Unify: Put your data to work with secure, well-governed access to data. To make decisions quickly, you need new data stores that will scale and grow as your business needs change. You also need to connect everything together—including your data lake, data warehouse, and purpose-built data stores—into a coherent system that is secure and well governed. AWS helps you accomplish this through data lakes and purpose-built data stores.

Modern Data Strategy with AWS

Harnessing data is the next wave of digital transformation

3

AWS can help your startup to:

Innovate new experiences and reimagine old processes with AI/ML. Machine learning is one of the most disruptive technologies of the last 25 years. It can help create entirely new revenue opportunities, make better and faster decisions, and improve operational efficiencies. AWS meets you wherever you are in your journey with the broadest, most comprehensive set of ML and AI services for builders of all levels of expertise.



Starting your ML journey



Starting your ML journey



Establish a data strategy

Lay the foundation for transformation and innovation



Start with the business challenge

Find the right use case based on the needs of your business



Working together for success

Take advantage of multiple programs and training options to help you along the way

Starting your ML journey



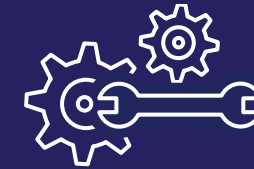
Address horizontal Use cases

Allow developers to easily add intelligence to any application



Find vertical Solutions

Discover industry-specific services that better fit your needs



Build your Own models

Make machine learning faster and easier to do with Amazon SageMaker

Getting started: Common machine learning use cases

Solve real-world problems with machine learning (ML)

Enhance the customer experience



Personalization



Contact center intelligence



Media intelligence

Delight customers while reducing operational costs

Optimize the business



Intelligent search



Intelligent document processing

Improve productivity and optimize business processes



Fraud detection



Business metrics analysis

Accelerate innovation



ML modernization

Speed up and scale up innovation with ML



Next gen DevOps

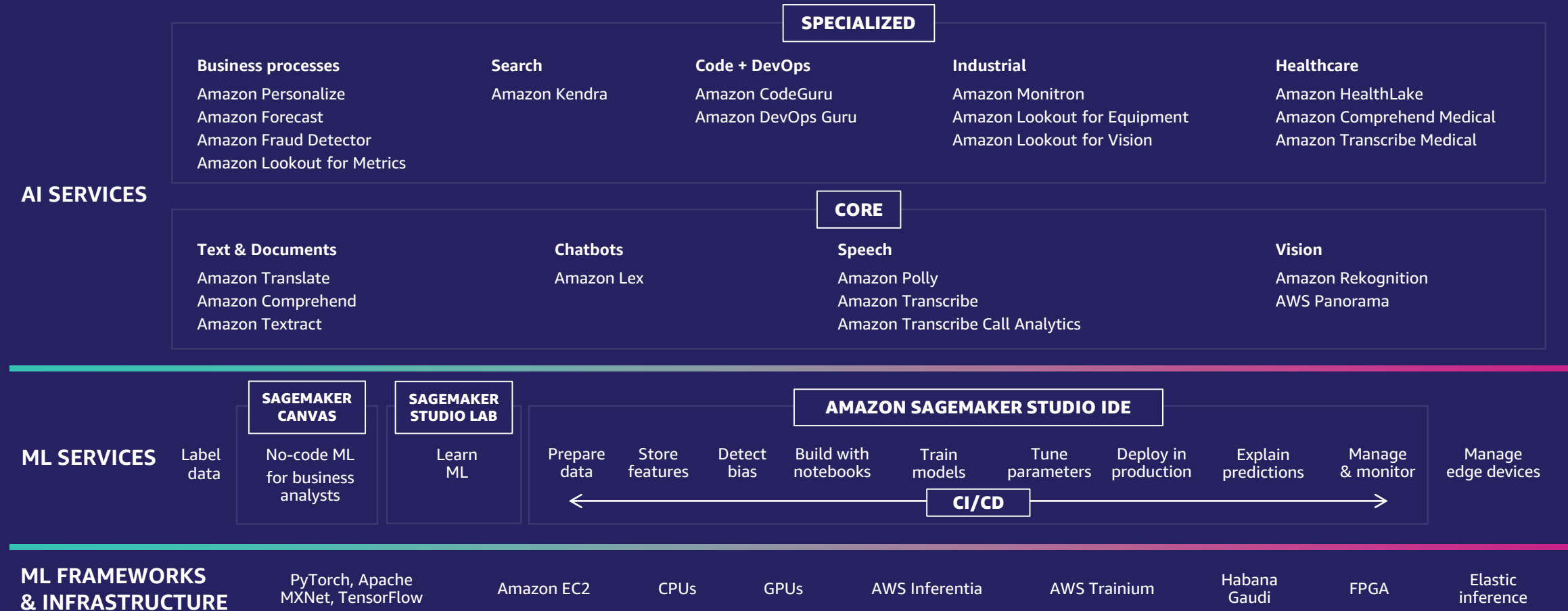
Getting started: Common machine learning use cases

Solve real-world problems with machine learning (ML)

- ML can be applied to many common challenges across various business functions and industry verticals. From enhancing customer experience to improving the productivity of technical teams, these machine learning use cases help optimize tasks that are complex, expensive, and require human overhead. AWS allows multiple easy-to-use AI Services to tackle these common use cases. No prior ML experience required. If you want to run ML workflows in the cloud and use your own machine learning models and algorithms, there is Amazon SageMaker, a fully managed service that helps data scientists and ML developers build, train, and deploy machine learning models quickly.

The AWS ML stack

Broadest and most complete set of machine learning capabilities



AI Services: Easily add intelligence to applications

No machine learning skills required



Vision



Chatbots



Business Tools



Search



Healthcare



Speech



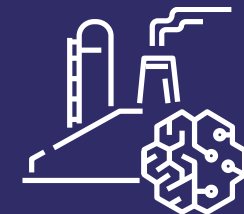
Text



Contact Centers



Code & Dev Ops



Industrial

AI Services: Easily add intelligence to applications

No machine learning skills required

- Our pre-trained AI Services provide ready-made intelligence for your applications and workflows. AI Services easily integrate with your applications to address common use cases such as personalized recommendations, modernizing your contact center, improving safety and security, and increasing customer engagement.
- Because we use the same deep learning technology that powers Amazon.com and our Amazon Machine Learning services, you get quality and accuracy from continuously learning APIs. And best of all, AWS AI Services don't require machine learning experience.

Amazon Fraud Detector

Identify fraud faster



Enhance fraud detection with ML



Any level of ML expertise can build ML fraud models



ML boost from Amazon experience and enrichments



Fewer false positives and manual reviews



Fraud staff self-service to address threats faster



Lower TCO and faster TTV

Amazon Fraud Detector

Identify fraud faster

Customers report a range of benefits from using Amazon Fraud Detector, including:

- **Enhancing fraud detection efforts with machine learning.** Use Amazon Fraud Detector to augment or replace your existing rules-based fraud detection solution.
- **Allowing users with any level of machine learning expertise to build ML models.** With Amazon Fraud Detector, fraud operations staff who are not ML experts can build and use ML-based fraud models. As a result, full-time data scientist involvement in fraud detection efforts is not needed or can be repurposed to address other organization needs.
- **Improving the results of machine learning models with Amazon's enrichments and fraud detection experience.** Most customers report that the model produced with Amazon Fraud Detector is better than the one they built themselves.
- **Handling fewer false positive alerts and overall manual reviews.** By producing models with Amazon Fraud Detector, you'll see a significant drop in the amount of work your customer service and fraud operations teams must process.
- **Allowing fraud operations teams to implement new measures to address threats.** Fraud prevention teams use Amazon Fraud Detector to instantly define and deploy new model logic that will address new attacks and bypass the need to work with IT.
- **Lowering total cost of ownership (TOC) while speeding up time to value (TTV).** Log in to the AWS console and immediately start developing and deploying fraud detection solutions using Amazon Fraud Detector, without upfront costs or commitments, or the need to install, configure, or maintain software.

Amazon Personalize

Delight customers and improve customer experience



Deliver
high-quality
recommendations



Adapt to changes
in customer intent
in real time



Train a
recommendation
model with a
few clicks



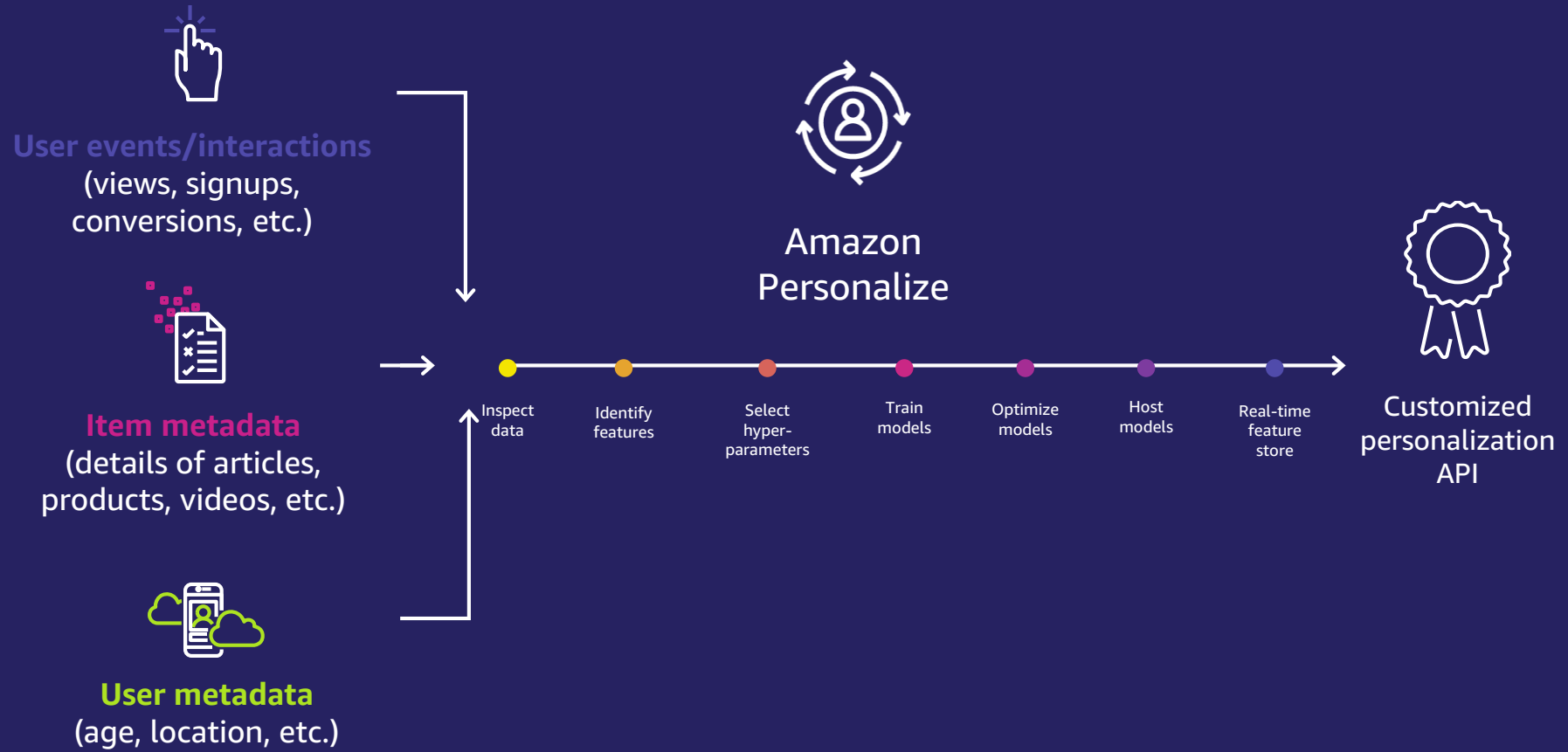
Generate
recommendations
for almost any
product or content

Amazon Personalize gives you out-of-the-box solutions to foundational use cases. You can use these capabilities to generate a broad array of personalized user experiences.

Amazon Personalize

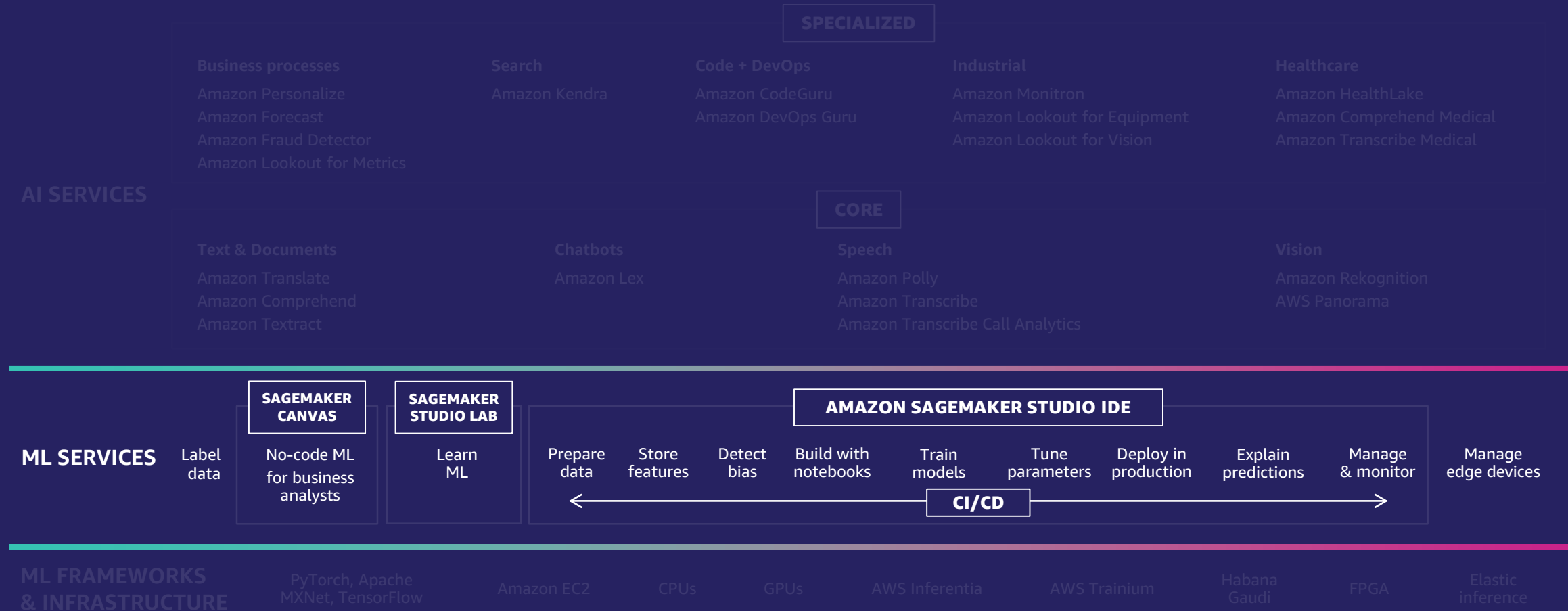
How it works

1. Processes and inspects the data
2. Identifies what is meaningful
3. Selects the right algorithms
4. Trains and optimizes a personalized models that is customized to the data



The AWS ML stack

Broadest and most complete set of machine learning capabilities



Amazon SageMaker



Integrated workbench

IDE designed specifically for ML, data preparation, experiment management, and pipelines

Managed infrastructure

Designed for ultralow latency and high throughput, automatic scaling, and distributed training

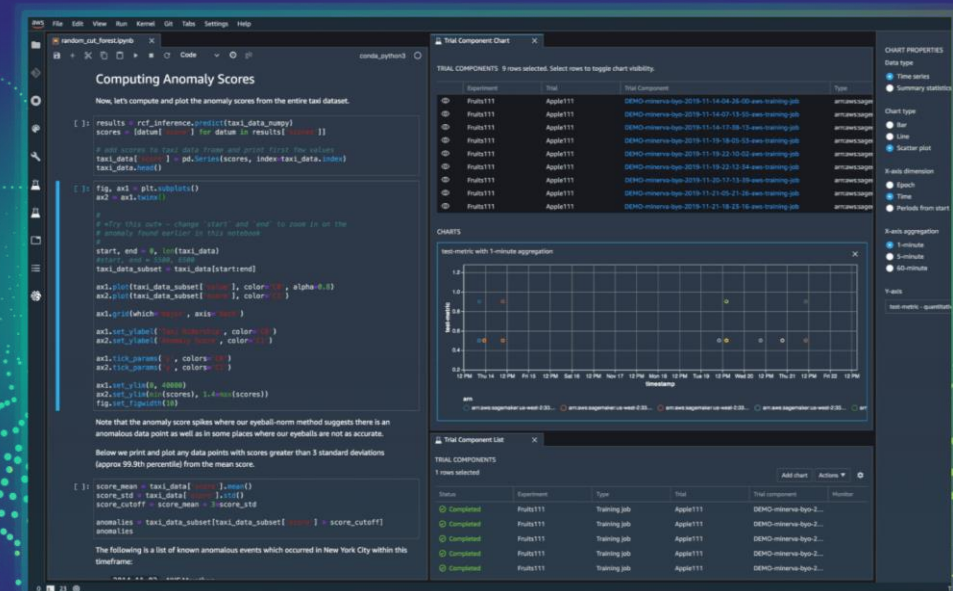
Managed tooling

Purpose-built from the ground up to work together: SageMaker Autopilot, collaboration, Jupyter notebooks, Experiments, Debugger, Model Monitor, and more

<https://aws.amazon.com/sagemaker>

Amazon SageMaker

Most complete
end-to-end ML service



Amazon SageMaker

Most complete end-to-end ML service

- Amazon SageMaker is the most complete end-to-end service for machine learning. It is a managed service for data scientists and ML operations teams that helps remove the undifferentiated heavy lifting associated with machine learning so that you have more time, resources, and energy to focus on your business. SageMaker has a lot of features and a full workbench of capabilities, there is three main pillars that describe what SageMaker is, and how it gets used.
- First off, SageMaker Studio offers an integrated workbench of tools. For example, you can launch Jupyter notebooks and JupyterLab environments instantly through SageMaker Studio. SageMaker also provides complete experiment management, data preparation, and pipeline automation and orchestration to help data scientists be more productive.

Amazon SageMaker

Most complete end-to-end ML service

- Now, if you've used a Jupyter notebook before, you know it needs to run on a computing environment. SageMaker provides fully managed servers in the cloud to make this easy for data scientists and developers. But even beyond notebooks, SageMaker provides other managed infrastructure capabilities as well. From distributed training jobs, data processing jobs, and even model hosting, SageMaker takes care of all the scaling, patching, high availability, and more associated with building, training, and hosting models.
- Finally, to help make data scientists more productive, this integrated workbench, which sits on managed infrastructure, is also enriched by a huge ecosystem of tools, all purpose-built for ML and designed from the ground up to work together.

Amazon SageMaker: Built to make ML more accessible



Amazon SageMaker:

Built to make ML more accessible

- Amazon SageMaker is the most complete end-to-end ML service, helping you through improved agility, productivity, and cost-effectiveness.
- We built SageMaker from the ground up to allow every developer and data scientist to build, train, and deploy ML models quickly and at a lower cost. To accomplish this, we provide the tools required for every step of the ML development lifecycle in one integrated, fully managed service. In fact, we launched 50+ capabilities in the past year alone, all aimed at making this process easier for our customers.
- And last year we launched Amazon SageMaker Studio so you can access all your tools in one place.

Amazon SageMaker overview

Amazon SageMaker

Prepare

SageMaker Ground Truth

Label training data for machine learning

SageMaker Data Wrangler

Aggregate and prepare data for machine learning

SageMaker Processing

Use built-in Python or bring your own (BYO) R/Spark

SageMaker Feature Store

Store, update, retrieve, and share features

Build

SageMaker Studio notebooks

Use Jupyter notebooks with elastic compute and sharing

Built-in and BYO algorithms

Use dozens of optimized algorithms or bring your own

Local mode

Test and prototype on your local machine

SageMaker Autopilot

Automatically create machine learning models with full visibility

Train and tune

One-click training

Distributed infrastructure management

SageMaker Experiments

Capture, organize, and compare every step

Automatic model tuning

Hyperparameter optimization

SageMaker Debugger

Debug training runs

Managed spot training

Reduce training cost by 90%

Deploy and manage

One-click deployment

Fully managed, ultralow latency, high throughput

Kubernetes and Kubeflow integration

Simplify Kubernetes-based machine learning

Multi-model endpoints

Reduce cost by hosting multiple models per instance

SageMaker Model Monitor

Maintain accuracy of deployed models

SageMaker Pipelines

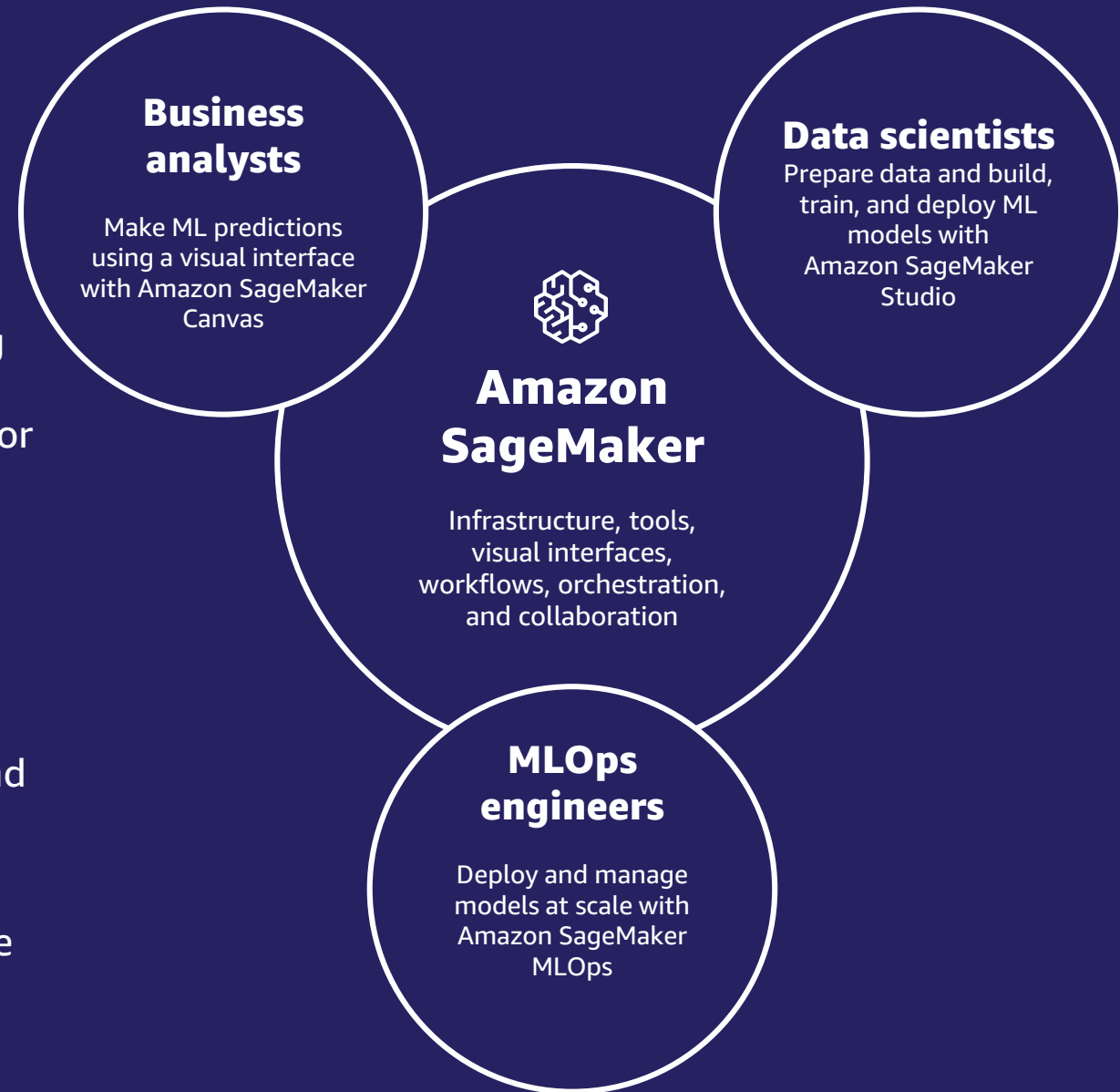
Implement workflow orchestration and automation

SageMaker Studio

Integrated development environment (IDE) for ML

Amazon SageMaker helps organizations harness ML

- Amazon SageMaker is a comprehensive machine learning service enabling business analysts, data scientists, and MLOps engineers to build, train, and deploy ML models for any use case, regardless of ML expertise.
- Business analysts can make ML predictions using a visual interface with SageMaker Canvas.
- Data scientist can easily prepare data, and build, train, and deploy ML models with SageMaker Studio.
- MLOps engineers can deploy and manage models at scale with SageMaker MLOps.



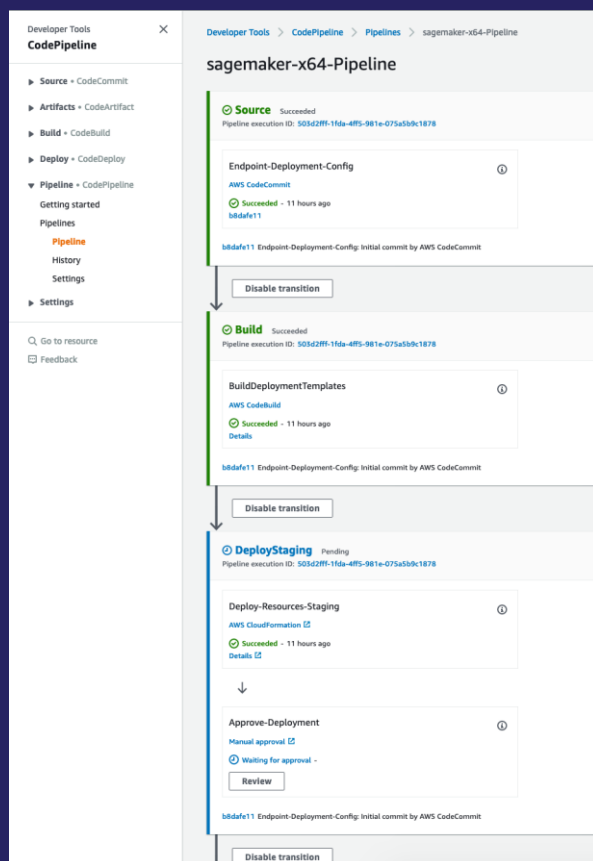
Approve models for production

The screenshot shows the Amazon SageMaker Studio interface. On the left, the 'Components and registries' sidebar is visible. The main area displays the 'Recommendations Model - Latin America' registry. A table lists model versions with columns for Name, Status, Step, Description, Status updated by, and Modified on. Version 6 is highlighted in red with a 'Rejected' status. An 'Update model version status' dialog box is open, showing the 'version 6' row selected. The dialog includes a 'Status' dropdown menu set to 'Approved' and a text area for a comment: 'The model accuracy of this model looks good. Approved.' The dialog also has 'Cancel' and 'Update status' buttons.

Name	Status	Step	Description	Status updated by	Modified on	Actions
version 6	Rejected	Staging	New model with SKLA...	Jen Cabro	10/10/20	Open model version Update model version status...
version 5	Approved	Production	Model updated on 8/1...	Jen Cabro		
version 4	Approved	Archived	Model updated on 7/15...	Jen Cabro		
version 3	Approved	Archived	Model updated on 6/15...	Jen Cabro		
version 2	Approved	Archived	Model built on 5/15/20...	Jen Cabro	10/10/20	
version 1	Approved	Archived	Model built on 4/15/20...	Jen Cabro	10/10/20	

And once those are provided to your data scientists, then your data scientists move their models through CI/CD pipelines by “approving” them within the model registry.

Deploy using fully managed CI/CD pipelines

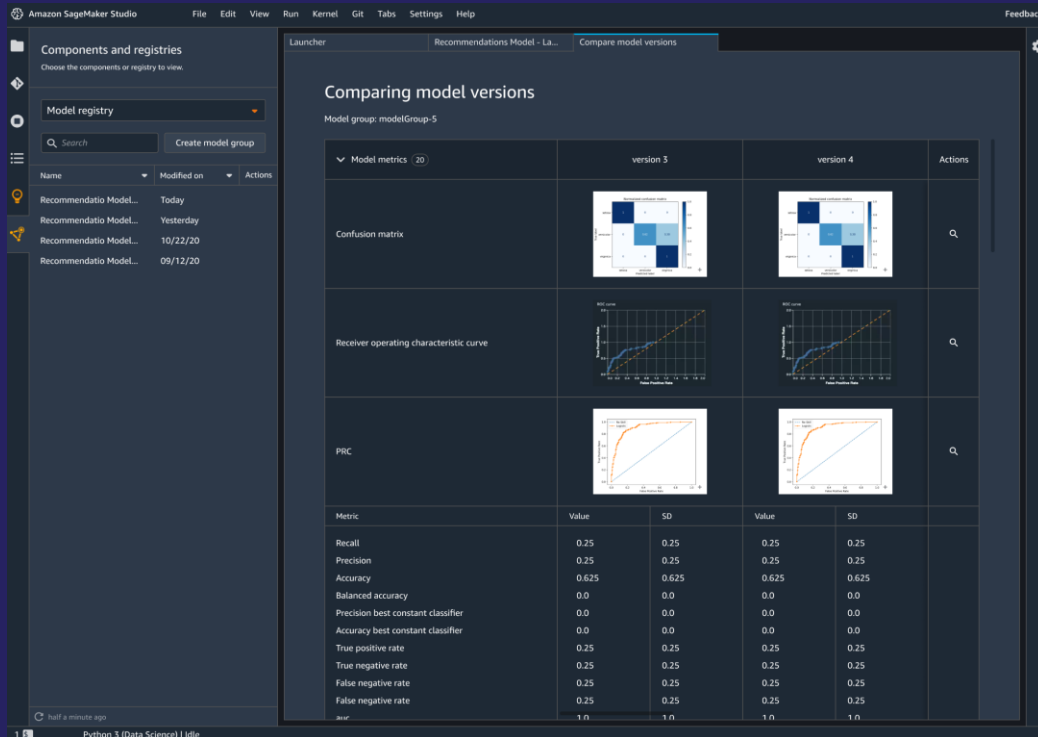


Now let's switch gears and talk about what happens after a model is trained, and after it's been added to the model registry. How do you use SageMaker Pipelines to initiate a CI/CD workflow?

Let's talk about how CI/CD workflows get created in the first place. And for that, we'll actually talk about a new persona, the DevOps engineer.

For this use case, DevOps engineers will actually continue to use AWS tooling that is familiar to them. For example, AWS CodePipelines. And you can define your CI/CD pipeline to codify how you want to move your model artifact through different environments and integrated tests.

View and compare evaluation metrics from training step



The great part about all of this integration is that you can then take advantage of the single pane of glass that SageMaker Studio provides. So after your models get deployed to production, you can see results about those models right within Studio.

Amazon SageMaker key benefits

Most complete end-to-end ML service



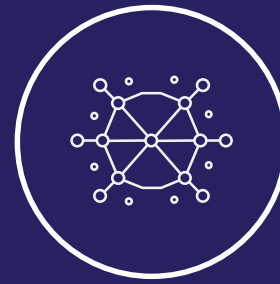
Democratize ML innovation

Empower more groups of people, including business analysts



Accelerate the ML lifecycle

Reduce training time from hours to minutes



Prepare data at scale

Access, label, and process structured and unstructured data



Streamline ML processes

Automate and standardize MLOps processes

Amazon SageMaker key benefits

Most complete end-to-end ML service

1. Enable more groups of people, including business analysts, to create ML models on their own using point-and-click, no-code visual interfaces
2. Access, label, and process massive amounts of structured data (e.g., tabular data) and unstructured data (e.g, photos, video, and audio) for ML
3. Reduce training time from hours to minutes with optimized infrastructure. Boost team productivity up to 10x with purpose-built tools
4. Automate and standardize MLOps processes across your organization to deploy and manage models in tens of thousands in production



Amazon SageMaker is DevOps ready



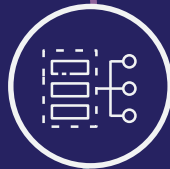
SECURITY

Security features to help you meet the strict security requirements of ML workloads



COMPLIANCE

Eligible for compliance with PCI, HIPAA, SOC 1/2/3, FedRAMP, and ISO 9001/27001/27017/27018



ML WORKFLOWS

Create automated workflows in minutes to support thousands of models



SCALABILITY

Train complex models with massive datasets

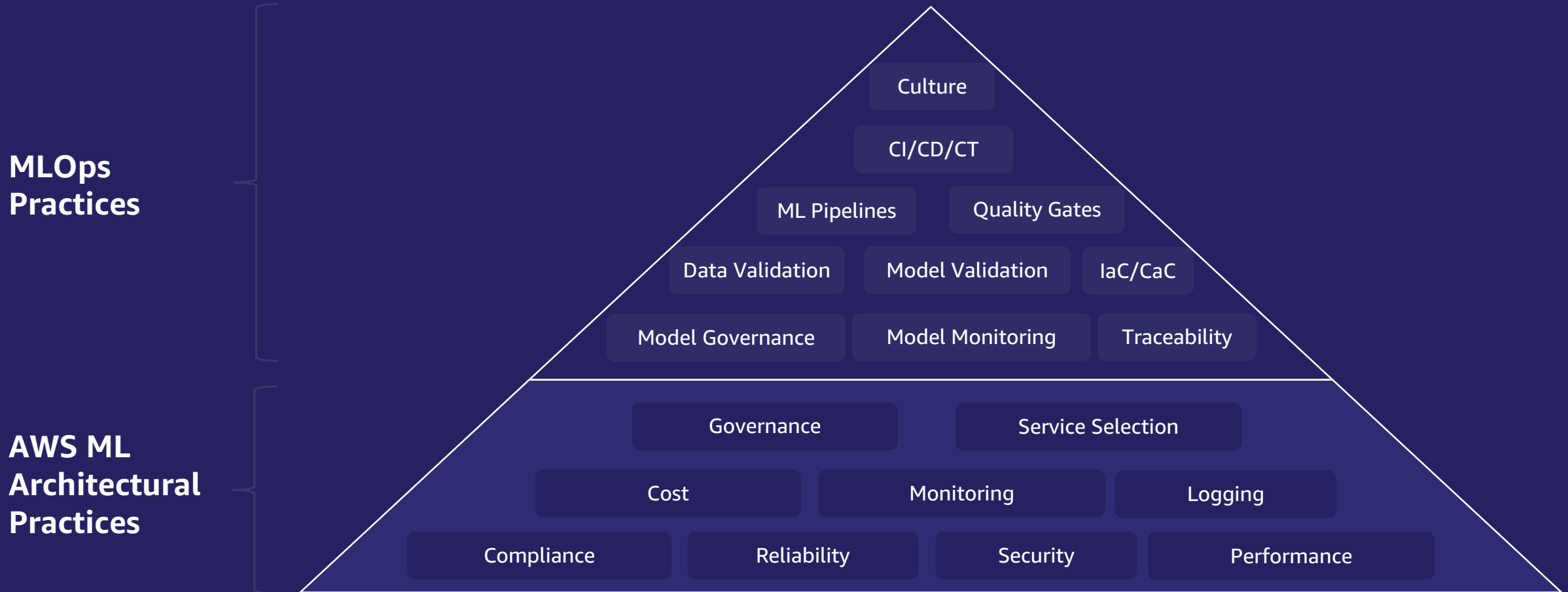


ORCHESTRATION

Automatically schedule and execute jobs with managed infrastructure

ML Operational Acceleration Framework

Foundational AWS ML Architecture & MLOps Practices



MLOps Defined

MLOps Maturity Model

	People	Data	Train	Deploy
Initial	<ul style="list-style-type: none">• Disconnected data science & IT teams• Limited cross-training	<ul style="list-style-type: none">• Ad-hoc data collection and preparation	<ul style="list-style-type: none">• Manual training & retraining• No clear path to deployment	<ul style="list-style-type: none">• Manual deployment
Repeatable	<ul style="list-style-type: none">• Improved collaboration with stakeholders• Shared project goals	<ul style="list-style-type: none">• Automated data pipelines	<ul style="list-style-type: none">• Defined path for experimentation• Automated training pipelines• Manual Model Validation	<ul style="list-style-type: none">• Automated deployment pipelines• Limited monitoring / measuring
Reliable	<ul style="list-style-type: none">• Cross-functional project teams• Some cross-training	<ul style="list-style-type: none">• Automated ML Pipelines• Data Governance	<ul style="list-style-type: none">• Experiment Management• Automated ML Pipelines• Model Governance• Automated Model Validation	<ul style="list-style-type: none">• Automated ML Pipelines• Monitoring & Logging (<i>Model, Workload, Pipeline</i>)
Scalable	<ul style="list-style-type: none">• Cross-functional project teams• Cross-training	<ul style="list-style-type: none">• CI/CD• Policy-as-Code• Configuration-as-Code• Automated Validation	<ul style="list-style-type: none">• CI/CD• Policy/Config-as-Code• Automated Model Validation• Automated Integration Validation	<ul style="list-style-type: none">• CI/CD• Policy/Infra/Config-as-Code• Model Monitoring• Dashboard & Transparency

The screenshot shows the Kubeflow Central Dashboard interface. The top navigation bar includes the Kubeflow logo and a 'Select namespace' dropdown. The main content area displays the 'Experiments' section for 'webinar-experiments', with a specific experiment 'cifar10-hpo-train-deploy' selected. The 'Graph' tab is active, showing a vertical sequence of five pipeline steps, each with a green checkmark indicating completion. The steps are: 'sagemaker-hyper...', 'update-best-mod...', 'sagemaker-traini...', 'sagemaker-creat...', and 'sagemaker-deplo...'. A sidebar on the left contains navigation options like Pipelines, Experiments, Artifacts, Executions, Archive, Documentation, Github Repo, and AI Hub Samples. At the bottom, there is a 'Build commit: 743746b' and a 'Report an Issue' link.

Amazon SageMaker capability

←→ Hyperparameter tuning job

←→ Custom function to update epochs

←→ Training job

←→ Create model

←→ Deploy inference endpoint

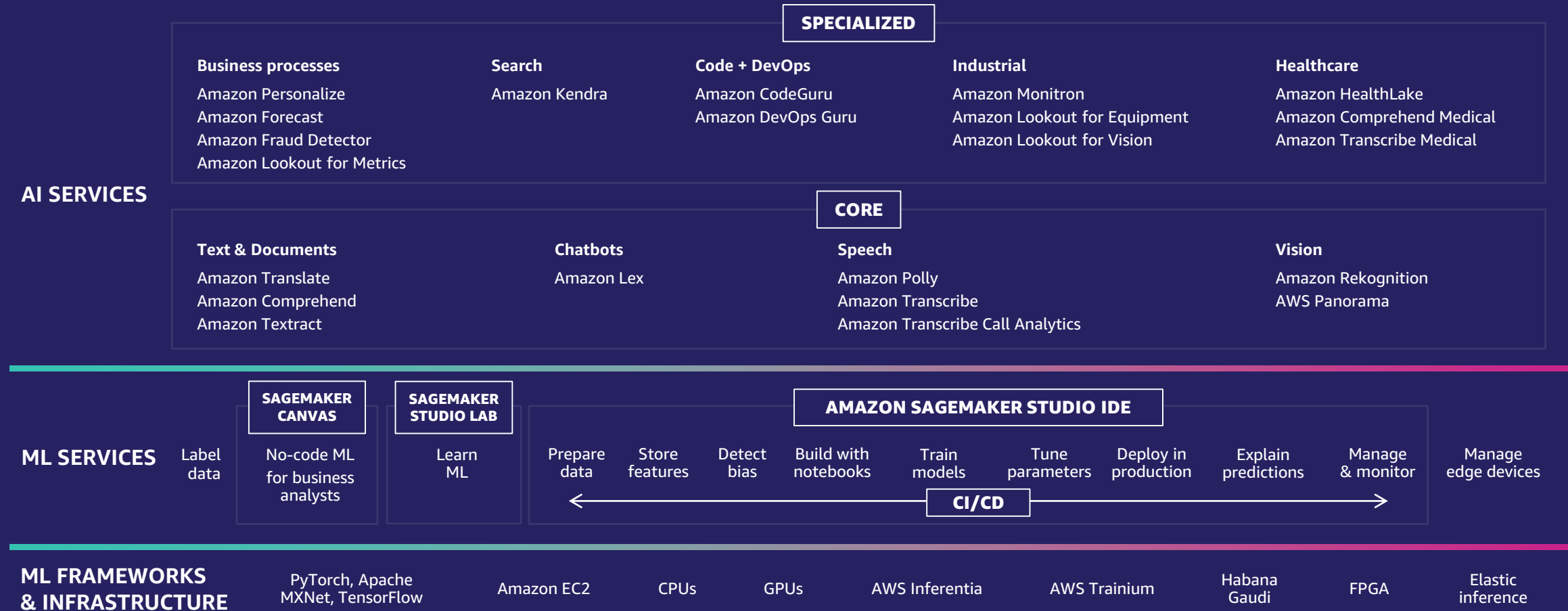
Kubeflow pipeline

ⓘ Runtime execution graph. Only steps that are currently running or have already completed are s



The AWS ML stack

Broadest and most complete set of machine learning capabilities



The AWS ML stack

Broadest and most complete set of machine learning capabilities

- At a macro level, we think about ML as having three layers of the stack
- At the bottom layer are the frameworks, interfaces, and infrastructure for expert ML practitioners.
- At the frameworks layer, three frameworks are predominantly used: TensorFlow, PyTorch, and MXNet
- Today, 92% of cloud-based TensorFlow and 91% of cloud-based PyTorch runs on AWS.
- New algorithms are being developed all the time in different frameworks. Companies want to borrow from or run those algorithms but don't want to port them to a different framework. As a result, AWS has dedicated teams for all the major frameworks so you have the right tool for the right job.

The AWS ML stack

Broadest and most complete set of machine learning capabilities

- AWS provides the broadest and deepest portfolio of ML infrastructure services with a choice of processors and accelerators to meet your unique performance and budget needs. [Amazon EC2 P4d instances](#) provide high-performance ML training in the cloud with the latest NVIDIA A100 Tensor Core GPUs coupled with first-in-the-cloud 400 Gbps instance networking. P4d instances are deployed in hyper scale clusters, called EC2 UltraClusters, that offer supercomputer-class performance for complex ML training jobs.
- For inference, [Amazon EC2 Inf1 instances](#), powered by AWS Inferentia chips, provide high performance and the lowest-cost inference in the cloud.

Broadest and deepest compute infrastructure for AI/ML

Choice of CPUs, GPUs, and accelerators for your performance and budget needs

Traditional machine learning (ML)

Deep learning (DL)

Training and inference

Inference

Training



Cascade Lake CPU, Skylake CPU
Habana Gaudi accelerators



EPYC CPU



Graviton CPU
Inferentia chip



A100, V100, T4 GPUs



Broadest and deepest compute infrastructure for AI/ML

Choice of CPUs, GPUs, and accelerators for your performance and budget needs

- AWS is a leader in offering the broadest and deepest compute infrastructure for AI/ML training and inference. We were the first to launch GPU-based instances for ML in the cloud. You can choose from a range of Amazon EC2 compute instances based on the latest CPUs, GPUs, and custom accelerators, coupled with industry-leading networking and storage to meet your budget needs and the performance requirements of your models. If you're running classical ML, you can choose our CPU-based instances to run training and inference of your small to medium-sized ML models. If you're looking for high-performance instances to run training and inference of your deep learning models, you can leverage our GPU- and accelerator-based instances.
- Our EC2 P4d instances are the highest-performance compute instances for training deep learning models in the cloud. They offer 2.5x better performance and 60% lower cost than previous-generation GPU-based P3 instances. They are designed for multi-node distributed training for large DL models. P3 instances are high-performance, cost-effective instances for training medium-to-large deep learning models and for single-node ML training. AWS is investing and innovating to deliver a range of high-performance and low-cost infrastructure options so you can choose the option that best fits your needs.

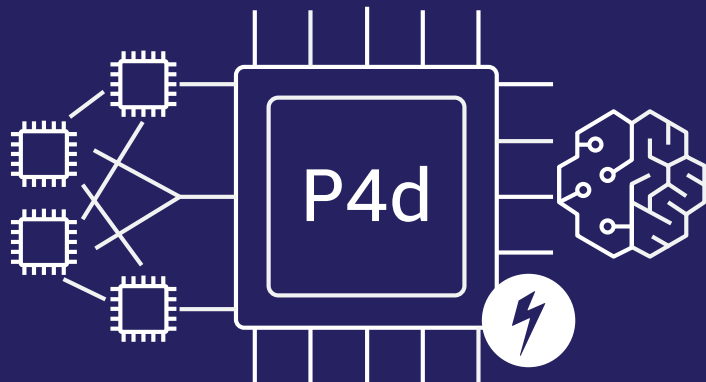
Broadest and deepest compute infrastructure for AI/ML

Choice of CPUs, GPUs, and accelerators for your performance and budget needs

- For example, EC2 DL1 instances powered by Gaudi accelerators from Habana Labs (an Intel company) deliver 40% better price performance than comparable GPU-based instances for training deep learning models. We're also building silicon from the ground up that is optimized for ML inference performance. We've leveraged learnings from our Graviton CPU and Nitro system innovations to build the AWS Inferentia chip—a custom chip that delivers the lowest cost of inference in the cloud.
- For high-performance GPU-based inference, you can use G4d instances. By running your ML workloads on AWS, you get on-demand access to high-performance, low-cost, and easy-to-use infrastructure services for training and deploying ML and DL models.

Introducing Amazon EC2 P4d instances

P4d instances



One of the most powerful GPU instances in the cloud

ML model with up to 60% lower cost to train, an average of 2.5x more deep learning performance, and 25% more GPU memory

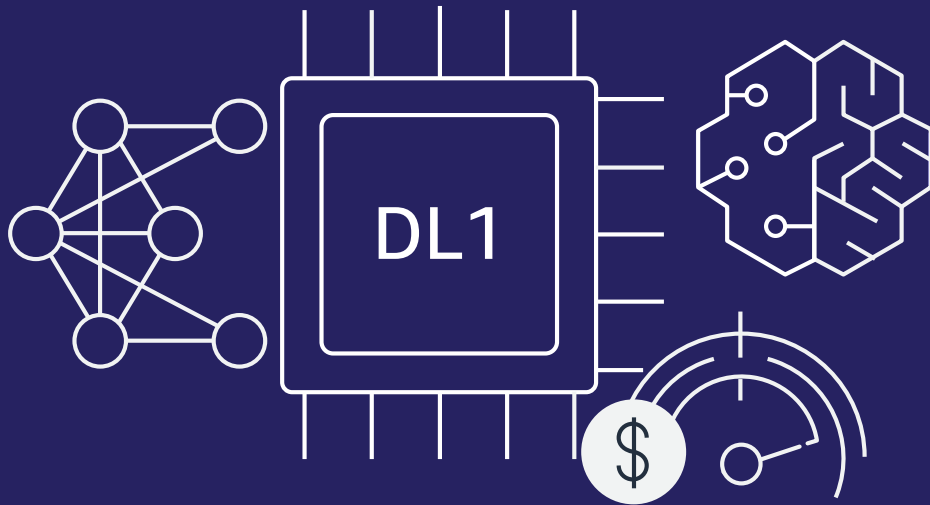
Powered by eight NVIDIA A100 GPUs and 400 Gbps of network bandwidth, and capable of 2.5 petaflops of performance

Deployed in UltraClusters consisting of thousands of tightly coupled GPUs, ideal for ML training and HPC

Introducing Amazon EC2 DL1 instances

Better price performance for training deep learning models

DL1 instances



Featuring up to eight Gaudi accelerators by Habana Labs (an Intel company)

Specifically built for training deep learning models

Up to 40% better price performance than the latest GPU instances

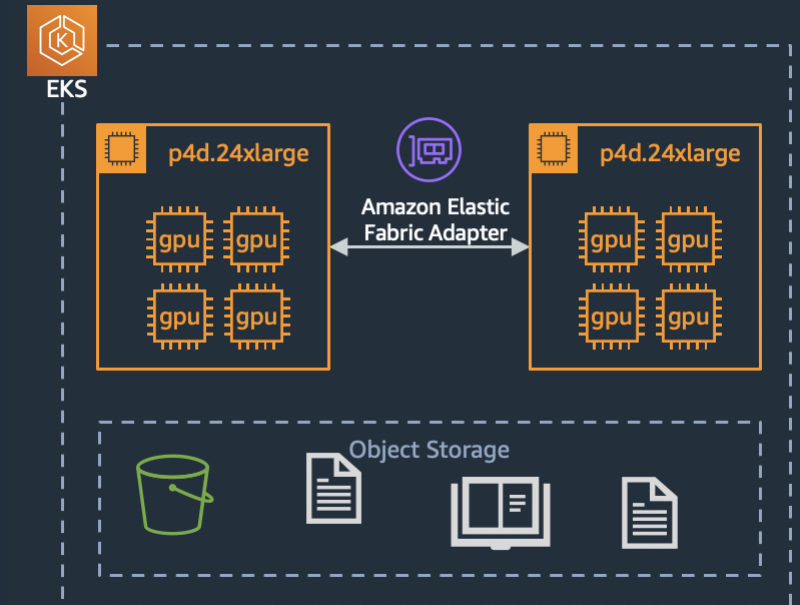
Custom software seamlessly integrated with TensorFlow and PyTorch

Get started easily using DLC, DL AMIs, or Amazon SageMaker

Launch DL1 instances via Amazon ECS and Amazon EKS for containerized ML applications

Large scale nlp startup on AWS

- CA startup aims to build the world's largest NLP model (2T+ parameters)
- Decided to run on AWS after a POC to train a 50B parameter model across 200 P4ds (800 A100 GPUs)
- Training workload is now in production across 500+ P4ds (4000 GPUs)
- Runs directly on EC2 with EKS orchestration and EFA for networking speedup
- Deployed in a single AZ, looking to extend architecture to span AZs.
- **Key infrastructural elements of the system:**
 - EKS for orchestration
 - EFA to support fast inter-node communications at scale
 - 500+ nodes of P4d and growing
 - S3 storage layer
 - Custom DLAMI built off of AWS base DLAMI
 - PyTorch + DDP + NCCL



Self managed ML

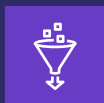
Broad choices to meet your evolving development needs

Tools

DATA PROCESSING & LABELING



Kinesis



Glue



EMR



SageMaker
Ground Truth

OSS



Kafka



APACHE
SPARK



hadoop

DL FRAMEWORKS



DLAMI



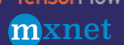
Containers

OSS

PYTORCH



TensorFlow



mxnet

DEV, TRAINING, AND TUNING (OSS)



Notebooks



Horovod



Katib



Kubeflow
Operators



AutoGluon

DASK

DEPLOY



KFServing*



AWS
Greengrass

TF Serving*, NVIDIA
Triton*

SERVERLESS



AWS
Lambda



Nuclio*

ML OPS

Code Build
Code Commit
CDK
ECR
Jenkins*

Orchestration

AWS SERVICES



EKS



ECS



Fargate



Batch



Parallel Cluster

OSS



Kubernetes

AWS/PARTNER SERVICES



Step Functions



Cloud Formation



Terraform

OSS



Kubeflow Pipelines



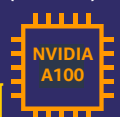
Airflow

Workflow

Infrastructure

COMPUTE

(Preview)



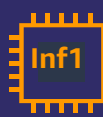
NVIDIA
A100



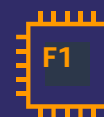
P3



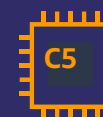
G4



Inf1



F1



C5

Accelerators

STORAGE



S3



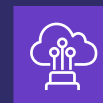
EFS



FSX

Lustre

DATA TRANSFER



Direct connect



Snowcone



Snowball



Snowball edge

NETWORKING



EFA



Getting started: Next steps



Getting started: Next steps



Collaboration

AWS as an ML expert



Discover Workshop

Identifying the use case



Training

AWS ML Embark Program

AWS deep devices

Getting started: Next steps

With the broadest and deepest set of ML services, AWS can help guide you on your ML journey. We can work with you to figure out the business challenges you want to solve by using ML and help you identify the right use case.

AWS has trainings and acceleration programs to help you get started. We offer in-person trainings for both technical and business stakeholders through our AWS ML Embark program as well as a full free online training and certification program for your developers. In addition, our portfolio of deep devices—AWS DeepRacer, AWS DeepComposer, and AWS DeepLens—were designed to provide developers with a hands-on, fun, engaging way to learn ML.





Thank you!

[Contact Us](#)

[Learn More](#)