AWS Architecture Monthly

July 2019

Machine Learning

Notices

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions, or assurances from AWS, its affiliates, suppliers, or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Editor's Note

In last month’s (June) magazine, we offered several pieces of content related to Internet of Things (IoT). This month we’re talking about artificial intelligence (AI), namely machine learning.

Alan Turing, the British mathematician whose life and work was documented in the movie *The Imitation Game*, was a pioneer of theoretical computer science and AI. He was the first to put forth the idea that machines can think.

Jump ahead 80 years to this month when researchers asked four-time World Poker Tour title holder Darren Elias to play Texas Hold’em with Pluribus, a poker-playing bot (actually, five of these bots were at the table). Pluribus learns by playing against itself over and over and remembering which strategies worked best. The bot became world-class-level poker player in a matter of days. Read about it in the journal *Science*: https://amzn.to/science-mag-poker-bot.

If AI is making a machine more human, AI’s subset, machine learning, involves the techniques that allow these machines to make sense of the data we feed them. Machine learning is mimicking how humans learn, and Pluribus is actually learning from itself.

From self-driving cars, medical diagnostics, and facial recognition to our helpful (and sometimes nosy) pals Siri, Alexa, and Cortana, all these smart machines are constantly improving from the moment we unbox them. We humans are teaching the machines to think like us.

We hope you find this edition of Architecture Monthly useful, and we’d like your feedback. Please give us a star rating and your comments on Amazon. You can also reach out to aws-architecture-monthly@amazon.com anytime.

For July’s magazine, we assembled architectural best practices about machine learning from all over AWS, and we’ve made sure that a broad audience can appreciate it.

- **Interview**: Mahendra Bairagi, Solutions Architect, Artificial Intelligence
- **Training**: Getting in the Voice Mindset
- **Quick Start**: Predictive Data Science with Amazon SageMaker and a Data Lake on AWS
- **Blog post**: Amazon SageMaker Neo Helps Detect Objects and Classify Images on Edge Devices
- **Solution**: Fraud Detection Using Machine Learning

- **Video**: Viz.ai Uses Deep Learning to Analyze CT Scans and Save Lives
- **Whitepaper**: Power Machine Learning at Scale

Annik Stahl, Editor

Interview:

Mahendra Bairagi, Solutions Architect

I recently met with Solutions Architect and artificial intelligence expert Mahendra Bairagi to ask him some question about artificial intelligence (AI) and machine learning.

Annik: Mahendra, we often use “machine learning” and “artificial intelligence” interchangeably. What's the difference and why does it matter?

Mahendra: AI seeks to create machines that seem to — or have — human intelligence, so if artificial intelligence is making the machines intelligent, machine learning is how we accomplish that. Putting it into business terms, AI is considered more of the business use case and machine learning provides the way to solve that use case.

Annik: What are some high-level benefits of machine learning?

Mahendra: I’ve met with more than 60 customers, and I’ve noticed that the key benefits of machine learning include automation, process improvements, customer retention, personalization, new customer acquisitions, lowering of operational cost through fraud detection, forecasting, better compliance management, and creating new business lines.

AI helps our finance or banking customers with fraud detection, forecasting, and process improvement through automation like check deposit apps. Prediction, anomaly detection, and forecasting also help industrial customers.

I’ve seen customers rebuild their business around AI and machine learning. Machine learning can give you a lot more personalized services. For example, when you can log in to Amazon or Netflix, you can see that it knows what you're looking for based on your history, and it can give you a really good idea of what you should buy or watch next.

Annik: Are there other some particular benefits that regular people on the street may not recognize, meaning they don't understand that they’ve been helped by machine learning?

Mahendra: I would say process improvement is a key one. If you look at Amazon, our product delivery changed big time — we now have one-day and same-day shipping —

and that’s all because of automation and machine learning behind the scenes. It's all about process improvement, automation, and machine learning.

But behind the scenes, many of the newer improvements that we and AWS customers put together (and not just in delivery or warehouse management, but also in other businesses such as online streaming) are based on prediction, too. Let's say there’s a popular new movie being released that’s expected to be in high demand. For example, when there is demand for romantic movies during Valentine’s Day, AI notices that trend and makes sure that that particular content is cached on the Edge servers, meaning faster access to those particular movies. You can actually look at the trends and patterns and improve your content streaming. That’s also something that our customers don't see, these very subtle benefits.

Then you could look at some improvements in auto industry. Detection of pedestrians, road signs, blind spots, automatic braking — these are all base on AI. Also, in health care, claims processing is automated with the help of AI and machine learning.

Annik: How do you think machine learning helps business leaders make better decisions? Can it help identify inefficiencies or help customer engagements?

Mahendra: Machine learning is going to impact every state of technology decision-making, and we’re already seeing customers reinventing their businesses with machine learning. There are literally tens of thousands of customers today using machine learning in virtually every industry and size of company.

For example, finance and banking companies can deal with fraud prevention and fraud detection using AI and machine learning. In the past, fraud prevention used to be very reactive. But now looking at the patterns and traits, you can predict when and where the fraud is going to happen.

Also, there are customers in the sports business using machine learning to transform the customer experience. When you log in to watch a game, you get to see different angles and all the plays, and one of our customers, the NFL, uses such machine learning to improve the TV experience for the fans. But also, the teams themselves are also using machine learning to improve the athletes’ performance.

Recently I came across a few use cases in the oil and gas industry. In one, some of the coastal oil rigs, many of which are far from civilization in places like northern Alaska, have to deal with encroaching wildlife, such as polar bears. The oil rig operators can get an alert when these animals come close to the rig. This is because it’s both a safety issue as well as a case for environmental protection, especially with endangered species, about whom you’re supposed to alert the appropriate authorities.

Another interesting oil and gas case has to do with flares from the oil rigs. On a rig that’s far from civilization, you can add a smart camera that’s programmed to look for flares and then alert oil and gas companies. There are two benefits to tracking flames: 1) the companies need to comply with regulatory authorities; and 2) flares mean there’s a waste of precious natural gas. With machine learning you can detect flares and figure out the pattern.

Annik: I read this quote in Forbes: “Machine learning is our new literacy that will define and shape businesses and industries at large, perhaps in even more profound ways than the Internet did.” What do you think about that?

Mahendra: Totally. Yeah, totally. But I also believe that the Internet enabled a lot of machine learning.

Traditionally, machine learning was mostly statistical algorithms; it didn't need lot of data to train. But then computers got smarter. The Internet arrived and now we have a lot of data that’s generated by machines and people. And because of the Internet, especially cloud computing, we can use this data and train the machine learning models, especially in the case of supervised learning. It’s going to bring a lot of changes in the industry.

I see so much demand for machine learning, and it’s coming from three different areas.

Annik: Is there anything else you want to add? We want our customers to know.

Mahendra: Yeah, I will say machine learning is not a mystery anymore. You can get started very easily. Our goal is to put machine learning in hands of every developer. If you have any use cases, then working backwards from the use case to the solution will give you good results. Machine learning is not difficult at all. There is a lot of hype about it but if you can tell us your requirements, if you need any help, AWS is always here.

Training:
Getting into the Voice Mindset

Available online at: http://amzn.to/AWS-ML-training-voice

Description

This is an introductory course geared towards anyone who wants a quick introduction to voice-enabled experiences. You will learn why voice-enabled experiences are the ubiquitous expectation these days and how voice-enabled experiences add value to a business. You will learn whether a voice-enabled experience is the right choice for your business and how to choose projects for voice. The course will also give you information about how Alexa works, the different machine learning technologies that power Alexa and how Alexa integrates with Amazon Web Services to build an end-to-end enterprise voice solution. Finally, you will learn about the different Alexa tools available to begin your journey into voice development. Specifically, you will learn about the Alexa Skills Kit (ASK), Alexa Voice Service (AVS), and Alexa for business.

Intended Audience

This course is intended for:

- This course is appropriate for both those who are new to voice-enabled experiences and those who want to determine whether voice is the right choice for their business.
- Anyone interested in understanding what tools are available to begin voice development with Alexa.

Course Objectives

In this course, you will learn:

- How voice-enabled experiences are making a difference in businesses
- Whether voice-enabled experiences are the right choice for your business
- Which tools you can use to start developing skills with Alexa

Real-World Example

Artificial Intelligence and Machine Learning Opportunities for Enterprises
http://amzn.to/AWS-ML-webinar

This AWS webinar is an introduction to Artificial Intelligence and Machine Learning and some practical ways to think about how to use it in the Enterprise. It is ideal for Data Scientists and Business Decision makers working in large Enterprise businesses looking to uncover opportunities to begin an AI/ML project.

Quick Start:

Predictive Data Science with Amazon SageMaker & a Data Lake on AWS

Available online at: http://amzn.to/AWS-ML-QS-predictive

This Quick Start builds a data lake environment for building, training, and deploying machine learning (ML) models with Amazon SageMaker on the Amazon Web Services (AWS) Cloud. The deployment, which takes about 10-15 minutes, uses AWS services such as Amazon Simple Storage Service (Amazon S3), Amazon API Gateway, AWS Lambda, Amazon Kinesis Data Streams, and Amazon Kinesis Data Firehose.

Amazon SageMaker is a managed platform that enables developers and data scientists to build, train, and deploy ML models quickly and easily.

This Quick Start is for users who want to unleash the power of their data to make predictive and prescriptive models for business value, without needing to configure complex ML hardware clusters. It enables end-to-end data science, starting with raw data and ending with a prediction REST API in a production system.

The Quick Start also provides a demo scenario developed by Pariveda Solutions. The demo shows how to store raw data in Amazon S3, transform the data for consumption in Amazon SageMaker, use Amazon SageMaker to build an ML model, and host the model in a prediction API for Amazon Elastic Compute Cloud (Amazon EC2) Spot pricing.

What You'll Build

This Quick Start architecture builds the following:

- A structured data lake in Amazon S3 to hold the raw, modeled, enhanced, and transformed data
- A staging bucket for the feature engineered and transformed data that will be ingested into Amazon SageMaker
- Data transformation code hosted on AWS Lambda to prepare the raw data for consumption and ML model training, and to transform data input and output
- Amazon SageMaker automation through Lambda functions to build, manage, and create REST endpoints for new models, based on a schedule or triggered by data changes in the data lake
- Amazon API Gateway endpoints to host public APIs for enabling developers to get historical data or predictions for their applications

- Amazon Kinesis Data Streams to enable real-time processing of new data across the Ingest, Model, Enhance, and Transform stages
- Amazon Kinesis Data Firehose to deliver the results of the Model and Enhance phases to Amazon S3 for durable storage
- An Amazon CloudWatch dashboard to provide monitoring of the data transformation, model training, and hosting components for the prediction endpoint
- An AWS SageMaker notebook server to enable data exploration by using a Jupyter notebook
- AWS Identity and Access Management (IAM) to enforce the principle of least privilege on each processing component. The IAM role and policy restrict access to only the resources that are necessary
- A demo scenario that builds and updates a predictive model for daily Amazon Elastic Compute Cloud (Amazon EC2) Spot pricing

Real-World Example

Effective Data Lakes: Challenges and Design Patterns
http://amzn.to/AWS-ML-DataLakes

Data lakes are emerging as the most common architecture built in data-driven organizations today. We walk through patterns to solve data lake challenges, like real-time ingestion, choosing a partitioning strategy, file compaction techniques, database replication to your data lake, handling mutable data, machine learning integration, security patterns, and more.

Blog:
Amazon Sagemaker Neo Helps Detect Objects & Classify Images on Edge Devices

By Satadal Bhattacharjee and Kimberly Madia

Available online at: http://amzn.to/AWS-ML-blog-Sagemaker

Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services in Japan and an APN Premium Consulting Partner. NRI is increasingly getting requests to help customers optimize inventory and production plans, reduce costs, and create better customer experiences. To address these demands, NRI is turning to new sources of data, specifically videos and photos, to help customers better run their businesses.

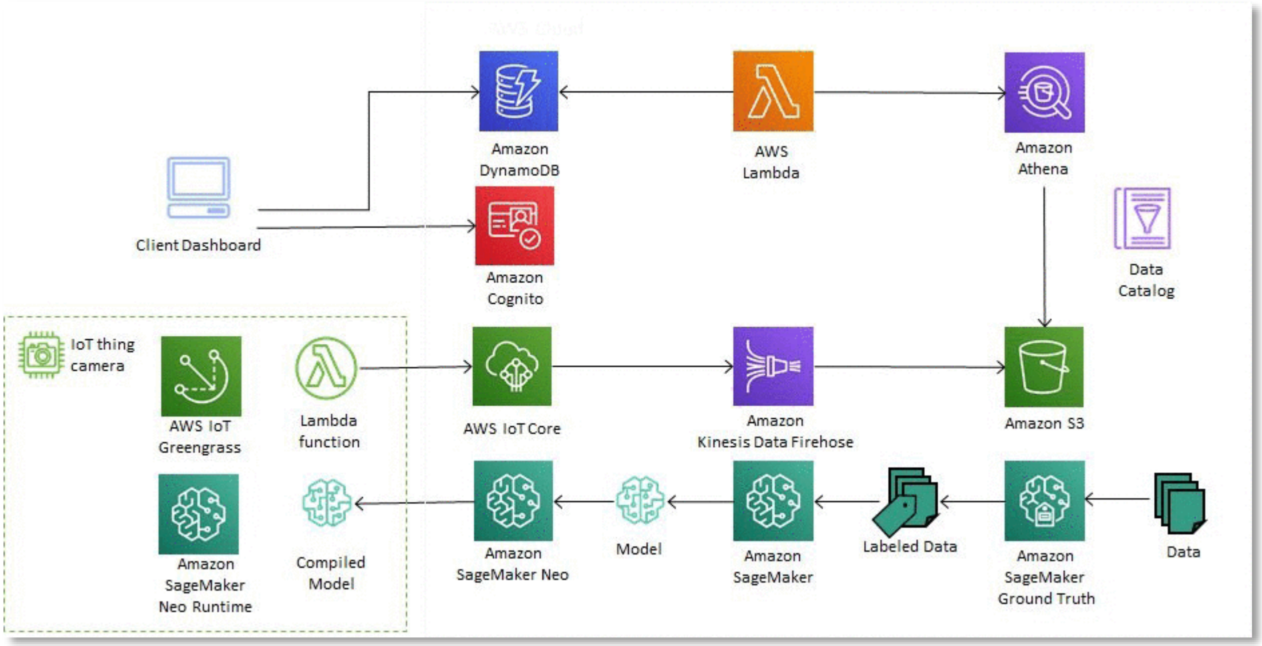
For example, NRI is helping Japanese convenience stores use data from in-store cameras to monitor inventory. And, NRI is helping Japanese airports to optimize people flow based on traffic patterns observed inside the airport.

In these scenarios, NRI needed to create a machine learning models that detects objects. NRI needed to detect goods (drinks, snacks, paper products, etc.) and people that leave stores for retailers, and commuters for airports.

NRI turned to Acer and AWS to meet their goals. Acer aiSage, is an edge computing device that uses computer vision and AI to provide real-time insights. Acer aiSage makes use of Amazon SageMaker Neo, a service that lets you train models that detect objects and classify images once and run them anywhere, and AWS IoT Greengrass, a service that brings local compute, messaging, data caching, sync, and machine learning inference capabilities to edge devices.

“One of our customers, Yamaha Motor Co., Ltd., is evaluating AI-based store analysis and smart store experience.” said Shigekazu Ohmoto, Senior Managing Director, NRI. “We knew that we had to build several computer vision models for such a solution. We built our models using MXNet GluonCV, compiled the models with Amazon SageMaker Neo, and then deployed the models on Acer’s aiSage through AWS IoT Greengrass. Amazon SageMaker Neo reduced the footprint of the model by abstracting out the ML framework and optimized it to run faster on our edge devices. We leverage full AWS technology stacks including edge side for our AI solutions.”

Here is how object detection and image classification works at NRI:



Amazon SageMaker is used to train, build, and deploy the machine learning model. Amazon SageMaker Neo makes it possible to train machine learning models once and run them anywhere in the cloud and at the edge.

Amazon SageMaker Neo optimizes models to run up to twice as fast, with less than a tenth of the memory footprint, with no loss in accuracy. You start with a machine learning model built using MXNet, TensorFlow, PyTorch, or XGBoost and trained using Amazon SageMaker. Then, choose your target hardware platform. With a single click, Amazon SageMaker Neo compiles the trained model into an executable.

The compiler uses a neural network to discover and apply all of the specific performance optimizations to make your model run most efficiently on the target hardware platform. You can deploy the model to start making predictions in the cloud or at the edge.

At launch, Amazon SageMaker Neo was available in four AWS Regions: US East (N. Virginia), US West (Oregon), EU (Ireland), Asia Pacific (Seoul). As of May 2019, SageMaker Neo is now available in Asia Pacific (Tokyo), Japan.

To learn more about Amazon SageMaker Neo, see the Amazon SageMaker Neo webpage.

Real-World Example

Run ML Models at the Edge with AWS Greengrass ML
http://amzn.to/AWS-ML-Greengrass

Learn how you can integrate your Machine Learning models into an edge device using AWS Greengrass and make inference on your data. We will review the AWS IoT services that enable AWS Greengrass ML Inference to operate, showcase how you would go about setting up your edge device, and finally demo local edge object recognition.

Solution:

Fraud Detection Using Machine Learning

Available online at: http://amzn.to/AWS-ML-solution-fraud

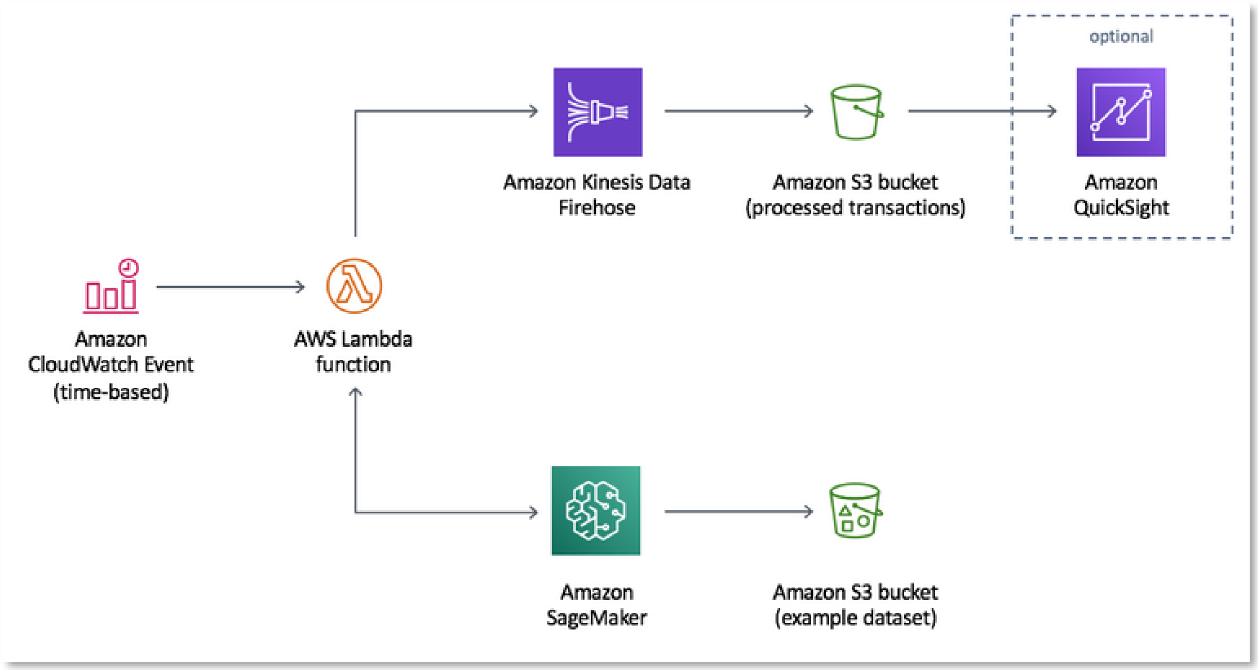
What does this AWS Solution do?

Fraud Detection Using Machine Learning deploys a machine learning (ML) model and an example dataset of credit card transactions to train the model to recognize fraud patterns. The model is self-learning which enables it to adapt to new, unknown fraud patterns.

You can use this solution to automate the detection of potentially fraudulent activity, and the flagging of that activity for review. The solution is easy to deploy and includes an example dataset but you can modify the solution to work with any dataset.

AWS Solution Overview

Fraud Detection Using Machine Learning enables you to execute automated transaction processing on an example dataset or your own dataset. The included ML model detects potentially fraudulent activity and flags that activity for review. The diagram below presents the architecture you can automatically deploy using the solution’s implementation guide and accompanying AWS CloudFormation template.


optional
Amazon Kinesis Data Firehose
Amazon S3 bucket (processed transactions)
Amazon QuickSight
Amazon CloudWatch Event (time-based)
AWS Lambda function
Amazon SageMaker
Amazon S3 bucket (example dataset)

Fraud Detection Using Machine Learning Architecture

This solution includes an AWS CloudFormation template that deploys an example dataset of credit card transactions contained in an Amazon Simple Storage Service (Amazon S3) bucket and an Amazon SageMaker endpoint with an ML model that will be trained on the dataset.

The solution also deploys an Amazon CloudWatch Events rule that is configured to run every minute. The rule is configured to trigger an AWS Lambda function that processes transactions from the example dataset and invoke the Amazon SageMaker endpoint which predicts whether those transactions are fraudulent based on the trained ML model. An Amazon Kinesis Data Firehose delivery stream loads the processed transactions into another Amazon S3 bucket for storage.

Once the transactions have been loaded into Amazon S3, you can use analytics tools and services, including Amazon QuickSight, for visualization, reporting, ad-hoc queries, and more detailed analysis.

By default, the solution is configured to process transactions from the example dataset. To use your own dataset, you must modify the solution. For more information, see the deployment guide.

Real-World Example

FICO: Fraud Detection and Anti-Money Laundering with AWS Lambda and AWS Step Functions

http://amzn.to/AWS-ML-TMA-FICO

Sven from FICO explains how they use a combination of AWS Lambda and AWS Step Functions to architect an on-demand solution for fraud detection and anti-money laundering.

Video:

Viz.ai Uses Deep Learning to Analyze CT Scans and Save Lives

Available online at: http://amzn.to/AWS-ML-TMA-VIZ

When someone suffers a stroke, every seconds counts and could mean the difference between being ok, permanent brain damage, or even death. With that taken into consideration, the current process for deciding how to treat the victim of a stroke just seems nonsensical. Typically, once a victim is rushed to the closest hospital, they are given a CT scan. A specialized doctor then needs to review the results of the scan, and that doctor is often located at a different hospital. The multi-step process of sending the scan, it being reviewed by the doctor, and the returning of their analysis, can take extremely long, with countless neurons dying in the patient's brain every second. Viz.ai is on a mission to reduce that timeline by leveraging specially trained deep learning models. CTO David Golan talks about how they're looking to save lives by leveraging AWS and deep learning.

Whitepaper:

Power Machine Learning at Scale

Mapping Parellelized Modeling-to-HPC Infrastrcure on AWS

Available online at: http://amzn.to/AWS-ML-whitepaper

Abstract

This white paper presents best practices for executing machine learning (ML) workflows at scale on AWS. It provides an overview of end-to-end considerations, challenges, and recommended solutions for architecting an infrastructure appropriate for ML use cases.

The intended audience for this white paper includes IT groups, enterprise architects, data scientists, and others interested in understanding the technical recommendations for executing parallelized modeling at scale using High Performance Computing (HPC) infrastructure on AWS.

Introduction

Businesses are generating, storing, and analyzing more data than ever before. With the latest advances in machine learning (ML), there is a drive to use these vast datasets to build business outcomes. Although ML algorithms have been used for more than 20 years, recent increased momentum in ML adoption is due to advancements in algorithmic frameworks and the compute infrastructure used to run the algorithms, including computational accelerators. The resulting ability to iteratively and automatically apply complex mathematical calculations to datasets at scale within relatively short timeframes, has added a new dimension of possibilities for ML analytics. Enterprises increasingly rely on machine learning to automate tasks, provide personalized services to their end users, and increase efficiency of their operations by gathering and analyzing data from a variety of deployed devices and sensors. To leverage economies of scale and increased agility, companies are moving their ML workloads to the AWS Cloud. For example, Marinus Analytics is using face search and recognition technologies to combat human trafficking. Petabytes of patient data stored and analyzed in the Philips HealthSuite digital platform is being analyzed using ML methodologies. TuSimple is using ML in the AWS Cloud for the perceptual distance capabilities of their L4 autonomous driving systems.

The present capabilities to scale and accelerate ML workloads are based on High Performance Computing (HPC) methodologies and applications. Modern HPC can use Graphics Processing Units (GPUs) for general purpose computing (GPGPUs), massively parallel data storage, and low-latency, high-bandwidth network communication to solve compute and memory intensive problems. These problems are common to many scientific research domains, including climate research, fluid dynamics, and life sciences. These disciplines share computational needs with areas of ML, such as deep learning (DL). The computational demands of DL workloads make them ideal candidates to benefit from HPC methods.

(Read the full paper at https://d1.awsstatic.com/whitepapers/aws-power-ml-at-scale.pdf)

Real-World Example

Build, Train, and Deploy Machine Learning Models at Scale Using AWS
http://amzn.to/AWS-ML-video-Sagemaker

Learn how AdTech companies use AWS services like Glue, Athena, Quicksight, and EMR to analyze your Google DoubleClick Campaign Manager data at scale without the burden of infrastructure or worries about server maintenance. We'll live-process a click stream so you can see how Machine Learning can help maximize your revenue by finding the most optimal path of a campaign and we'll look at a real world demo from A9's Advertising Science Team of how they use the data to build Look-alike Model in their projects.