NUCLEUS RESEARCH

GUIDEBOOK

# TENSORFLOW ON AWS

PROGRAM: DATA ANALYTICS
DOCUMENT R206 • December 2017

ANALYSTS
Rebecca Wettemann, Barbara Peck

Nucleus Research Inc. 100 State Street Boston, MA 02109 **NucleusResearch.com** Phone: +1 617.720.2000

## THE BOTTOM LINE

**The machine learning (ML) and deep learning (DL) research space is an emerging field for data scientists seeking to apply ML and DL capabilities to complex analytics problems such as recommendations and natural language and image recognition and processing.** DL is a subset of ML , where algorithms can learn (as opposed to task-specific algorithms) and are often compared to the information processing and communication patterns of a biological nervous system. In analyzing the experiences of ML and DL researchers supporting more than 388 unique projects across multiple disciplines and domains, Nucleus found that Amazon Web Services (AWS) is the platform of choice for TensorFlow-based DL, with 88 percent of TensorFlow projects running on AWS.

•••

## THE SITUATION

The field of machine learning(ML) and deep learning (DL) tools, technologies, and techniques is still emerging, as researchers in academia, startups, and large firms alike seek to harness ML and DL capabilities to address specific data and business challenges. DL is most commonly defined as the application of self-learning algorithms that can process information in both supervised and unsupervised situations to address complex analytical problems. The most common areas where DL capabilities are being used include video and image analysis, recommender analysis (for applications such as e-commerce and CRM), and speech recognition.

To better understand the current state of the deep learning research environment; challenges faced by researchers; models, frameworks, and libraries being used; and deployment models including the use of public clouds, Nucleus conducted extensive research with DL and ML practitioners including those in academia and the private sector. Private sector experts came from traditional large corporations, AI and DL startups and services companies, and large technology companies that are leveraging DL and other analytics as a product and service differentiator. Our research included in-depth interviews with 32 individuals or teams performing DL analyses today, representing more than 388 unique customers.

To better understand the current DL landscape, we asked respondents about a number of topics including:

- The business challenges and goals of their research
- The frameworks, methods, tools, and data libraries being used
- How their end customers (when applicable) were benefiting from DL
- Where and how they were deploying their models and why
- The relative strengths and weaknesses of different frameworks and deployment strategies

We found that TensorFlow is the most commonly-used DL framework today. Of the total projects analyzed:

- Thirty-six percent were using only TensorFlow to support their DL efforts
- Forty-three percent were using TensorFlow as well as other frameworks (such as MXNet, Caffe, PyTorch, or other proprietary tools or frameworks)
- Only 21 percent were not using TensorFlow at all.

Many of the projects using more than one framework still considered TensorFlow to be their primary framework for DL, meaning that roughly four out of five DL projects today are being supported by TensorFlow.

We also found that while many researchers are still running DL experiments on on-premise servers today, momentum is building for moving these workloads to the cloud. As users seek to productize their DL offerings or move beyond experiments to a scalable production environment, cloud is increasingly the deployment model of choice. Of our recent analysis, in all, fewer than 10 percent of projects were being run on premise.

**Momentum is building for moving DL workloads to the cloud, as users seek to productize their DL offerings or move beyond experiments to a scalable production environment.**

When it comes to cloud, Amazon Web Services (AWS) was the overwhelming platform choice for TensorFlow:

- Of 388 projects, 80 percent using TensorFlow and other frameworks are running exclusively on AWS
- 88% using only TensorFlow are running exclusively on AWS
- Only one in 10 of TensorFlow projects are running on another platform.

**Nine out of 10 TensorFlow projects are running on AWS.**

# WHY TENSORFLOW ON AWS

In analyzing the experience of DL researchers, Nucleus found there were five main reasons why they chose AWS to run their analyses instead of other cloud providers: the breadth of the Amazon platform, cost, their relationship with Amazon as a partner, compute locations, and Amazon's peer resources and community.

## BREADTH OF AMAZON CAPABILITIES

Nucleus found that researchers looking beyond small test projects chose Amazon because of its data storage, support for multiple frameworks, and other resources that could support large-scale production DL projects. Users said:

- *"It really comes down to what you use for programming language. [Compared to other providers] Amazon is significantly more broad in terms of what they'll support.."*
- *"With AWS you get support for bringing these things [ML applications] to production, a full set of features, integration with the rest of your stack, the infrastructure and tooling makes it a lot easier."*
- *"As an individual developer it is easier to use what you know. AWS is good at so many things, and you get better integration when doing more than a point project."*
- *"We're already using AWS for other capabilities. [Other providers] are building tools to catch up, but AWS continues to enhance its functionality."*

Nucleus also found that organizations' existing relationship with Amazon made using AWS for TensorFlow a natural extension of their existing footprint (particularly their existing investments in storage, databases, and analytics), making adding compute for DL relatively seamless.

Although Amazon SageMaker, announced in late November 2017, was not available to the users Nucleus analyzed given the timing of the release, Nucleus expects that the service will provide even greater benefits to TensorFlow users. SageMaker is a fully managed service that enables developers and data scientists to build, train, and deploy machine learning models at any scale without worrying about any of the heavy lifting that typically slows down the ML/DL workflow. The service allows TensorFlow users to author models in a hosted Jupyter notebook environment (or import existing models), rapidly train across virtually unlimited CPUs or GPUs, automatically tune models, and deploy and host models with minimal coding. Nucleus expects users of Amazon SageMaker will see reductions in both the time and engineering resources required to get TensorFlow models into production.

## COST VERSUS VALUE

Although pricing strategies and discounting continues to evolve, DL experts found Amazon's pricing attractive, particularly for those reselling their own services on the AWS TensorFlow platform. Users said:

- *"We look at the lowest cost because we resell the layer on top. We want to provide the lowest cost to our customers."*
- *"A startup knows that their infrstructure is going to need to scale and they need to commit as little money as possible. Our conclusion was for price and features that I couldn't beat AWS."*

### User example: SAAS decisioning platform

One DL expert Nucleus interviewed had developed a software-as-a-service (saas) platform for decision making for marketing and creative teams. The company provides a model as a platform built on Tensorflow to teams that is used to analyze internal and external content to develop recommended content calendars and investments in the type of content likely to appeal to specific engagement goals. It chose AWS as its cloud platform because "*it scales perfectly, it's affordable, and it's reliable*."

## AMAZON RELATIONSHIP AND CAPABILITIES

Nucleus also found that organizations' existing relationship with Amazon made using AWS for TensorFlow a natural extension of their existing footprint, making adding compute for DL relatively seamless. Most already had a large amount of data stored in Amazon S3 or one of AWS's other storage or database services, and they spoke about the importance of conducting analysis where the data already exists in order to maximize speed and cost efficiency. Users pointed out that machine learning and deep learning were only part of their analytics workflow, and that AWS provided them with a broad range of options for data lake storage, security, and analytics tools. AWS is also the only cloud provider that currently offers compute instances based on NVIDIA's latest V100 GPUs, which provide significant performance benefits for both training and running ML/DL models.. Users said:

- *"It would depend on where your customers are and what's collocated closer to your GPUs."*
- *"The data needs to be where the compute is and if you're not doing that you're creating a future problem for yourself."*

### User example: Data science consultancy

Another DL expert Nucleus interviewed had helped a number of clients, mostly in the nonprofit sector, develop and run DL models to support their research. The

consultancy used a number of different DL frameworks and other tools, and found AWS was its platform of choice for production projects because of AWS's flexibility and support for multiple frameworks, saying, *"It comes down to what you're using. Amazon is significantly more broad in terms of what they'll support."*

### PEER COMMUNITY AND RESOURCES

Users also cited the peer community of AWS users exploring DL projects and the resources and support provided by AWS as a reason for running their Tensorflow projects on AWS. The DL field is still largely in the experimentation phase, with startups, existing technology firms, and academics alike continuing to evolve their data processes and platforms. As the DL research community is relatively small and evolving, many look to other leaders in the field – as well as the cloud platform providers – to provide not just guidance on data sets and processes but on how to best leverage cloud computing to bring their projects beyond the test phase. Nucleus found that AWS's thought leadership and peer community was another significant factor driving DL practitioners to choose AWS for their TensorFlow analysis. Users said:

- *"Peers are a huge component. You look and see people pushing the state of the art – that goes a long way toward trust and assurance that you're going to get what you need out of it."*
- *"TensorFlow has poor documentation, but the AWS community more than compensates."*
- *"AWS is well documented and easy to use – there's just more information on how to use it than [other providers]."*
- *"AWS is better by far. It can handle more data faster, and has the best community support and the best roadmap."*

## KEY BENEFIT AREAS

Key benefits DL experts running their projects on AWS cited included:

- Increased productivity. Availability of support and resources for AWS, as well as AWS's contributions to the TensorFlow community, made it easier for new users to get projects up and running. Additionally, AWS's support for a broader set of tools and datasets made experimentation easier.
- Improved technology management. The scalability and reliability of AWS, as well as the ability to support projects with near-local compute, enabled them to manage both the cost and performance of their TensorFlow models. Additionally, organizations that already had an AWS relationship found they were able to extend into ML and DL to support new projects.

- Increased profits. For those commercializing their DL projects, the reliability and availability of AWS as well as their ability to provide a commercial-grade platform to customers made it relatively easier to attract customers and monetize their efforts.

## CONCLUSION

Although the field of deep learning and machine learning is still evolving, those in the field are looking to competing resources that provide not just support for a particular framework but one that can support the evolving needs of their research from small projects to commercially viable applications. In our analysis we found that the breadth of Amazon's capabilities beyond TensorFlow, cost-value proposition, reputation and trust as an enterprise computing partner, and peer community and resources made it the cloud platform of choice for DL professionals. As Amazon continues to build out its global infrastructure (with $P_3$ instances) as well as extended capabilities such as SageMaker specifically designed to support DL research, Nucleus expects it will gain more momentum as the dominant cloud platform for DL research.