datanami
DATA SCIENCE • AI • ADVANCED ANALYTICS
(https://www.datanami.com)

Select Language ▼
Translation Disclaimer (/termsofuse.html#translation)
Search this site Search

Home (https://www.datanami.com/) About (https://www.datanami.com/about/)
Resources (https://www.datanami.com/whitepaper/) Events (https://www.datanami.com/events/)
Subscribe (https://www.datanami.com/subscribe/)

Follow Datanami: (https://www.facebook.com/pages/Datanami/12476054763
(https://www.twitter.com/datanami) (https://www.linkedin.com/groups/Big-Data-News-Network-4166980) (https://www.datanami.com/feed/)

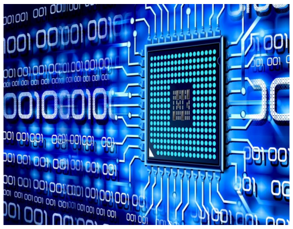
HOME (HTTPS://WWW.DATANAMI.COM/) FEATURES SECTORS APPLICATIONS TECHNOLOGIES VENDORS
JOB BANK (HTTPS://WWW.DATANAMI.COM/JOB-BANK/) EVENTS

Top Stories On

January 27, 2020

# An Open Source Alternative to AWS SageMaker

Alex Woodie



(Robert Lucian Crusitu/Shutterstock)

There's no shortage of resources and tools for developing machine learning algorithms. But when it comes to putting those algorithms into production for inference, outside of AWS's popular SageMaker, there's not a lot to choose from. Now a startup called Cortex Labs is looking to seize the opportunity with an open source tool designed to take the mystery and hassle out of productionalizing machine learning models.

Infrastructure is almost an afterthought in data science today, according to Cortex Labs (https://www.cortex.dev/) co-founder and CEO Omer Spillinger. A ton of energy is going into choosing how to attack problems with data – why, use machine learning of course! But when it comes to actually deploying those machine learning models into the real world, it's relatively quiet.

"We realized there are two really different worlds to machine learning engineering," Spillinger says. "There's the theoretical data science side, where people talk about neural networks and hidden layers and back propagation and PyTorch and Tensorflow. And then you have the actual system side of things, which is Kubernetes and Docker and Nvidia (https://www.nvidia.com/) and running on GPUs and dealing with S3 and different AWS (http://www.aws.amazon.com/) services.

Both sides of the data science coin are important to building useful systems, Spillinger says, but it's the development side that gets most of the glory. AWS has captured a good chunk of the market with SageMaker, which the company launched in 2017 and which has been adopted by tens of thousands of customers. But aside from just a handful of vendors working in the area, such as Algorithmia (http://www.algoritihmia.com/), the general data-building public has been forced to go it alone when it comes to inference.

## Scaling ML

A few years removed from UC Berkeley's computer science program and eager to move on from their tech jobs, Spillinger and his co-founders were itching to build something good. So when it came to deciding what to do, Spillinger and his co-founders decided to stick with what they knew, which was working with systems.

"We thought that we could try and tackle everything," he says. "We realized we're probably never going to be that good at the data science side, but we know a good amount about the infrastructure side, so we can help people who actually know how to build models get them into their stack much faster."

Cortex Labs' software begins where the development cycle leaves off. Once a model has been created and trained on the latest data, then Cortex Labs steps in to handle the deployment into customers' AWS accounts using its Kubernetes engine (AWS is the only supported cloud at this time; on-prem inference clusters are not supported).

This Just In Most Read

January 28, 2020
Study: Decision-Making is Driving Data Governance Programs, Not Regulatory Compliance (https://www.datanami.com/this-just-in/study-decision-making-is-driving-data-governance-programs-not-regulatory-compliance/)
DataGrail Finds Americans Expect Privacy Controls, 4 out of 5 People Want a Privacy Law to Protect Personal Data
(https://www.datanami.com/this-just-

Visit additional Tabor Communications publications
HPCwire EnterpriseAI HPCwire JAPAN
(https://hpcwire.com/)(http://www.enterpriseai.news/)(https://www.hpcwire.jp/

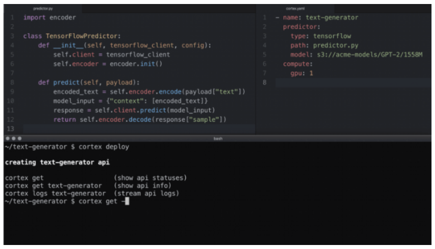
(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2016/01/shutterstock_cluster_computing_bluebay.jpg)

*(bluebay/Shutterstock.com)*

"Our starting point is a trained model," Spillinger says. "You point us at a model, and we basically convert it into a Web API. We handle all the productionalization challenges around it."

That could be shifting inference workloads from CPUs to GPUs in the AWS cloud, or vice versa. It could be we automatically spinning up more AWS servers under the hood when calls to the ML inference service are high, and spinning down the servers when that demand starts to drop. On top of its build-in AWS cost-optimization capabilities, the Cortex Labs software logs and monitors all activities, which is a requirement in today's security- and regulatory-conscious climate.

Cortex Labs is a tool for scaling real-time inference, Spillinger says. It's all about scaling the infrastructure under the hood.



```
predictor.py
import encoder

class TensorFlowPredictor:
    def __init__(self, tensorflow_client, config):
        self.client = tensorflow_client
        self.encoder = encoder.init()

    def predict(self, payload):
        encoded_text = self.encoder.encode(payload["text"])
        model_input = {"context": [encoded_text]}
        response = self.client.predict(model_input)
        return self.encoder.decode(response["sample"])
```

```
cortex.yaml
- name: text-generator
  predictor:
    type: tensorflow
    path: predictor.py
    model: s3://acme-models/GPT-2/1558M
  compute:
    gpu: 1
```

```
bash
~/text-generator $ cortex deploy

creating text-generator api

cortex get                    (show api statuses)
cortex get text-generator     (show api info)
cortex logs text-generator    (stream api logs)
~/text-generator $ cortex get -
```

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2020/01/Cortex_CLI.png)

*Cortex Labs delivers a command line interface (CLI) for managing deployments of machine learning models on AWS*

"We don't help at all with the data science," Spillinger says. "We expect our audience to be a lot better than us at understanding the algorithms and understanding how to build interesting models and understanding how they affect and impact their products. But we don't expect them to understand Kubernetes or Docker or Nvidia drivers or any of that. That's what we view as our job.

The software works with a range of frameworks, including TensorFlow, PyTorch, scikit-learn, and XGBoost. The company is open to supporting more. "There's going to be lots of frameworks that data scientists will use, so we try to support as many of them as we can," Spillinger says.

Cortex Labs' software knows how to take advantage of EC2 spot instances, and integrates with AWS services like Elastic Kubernetes Service (EKS), Elastic Container Service (ECS), Lambda, and Fargate. The Kubernetes management alone may be worth the price of admission.

"You can think about it as a Kubernetes that's been massaged for the data science use case," Spillinger says. "There's some similarities to Kubernetes in the usage. But it's a much higher level of abstraction because we're able to make a lot of assumptions about the use case."

## Open MLOps Standard

There's a lack of publicly available tools for productionalizing machine learning models, but that's not to say that they don't exist. The tech giants, in particular, have been building their own platforms for doing just this. Airbnb, for instance, has its BigHead offering (https://www.slideshare.net/databricks/bighead-airbnbs-endtoend-machine-learning-platform-with-krishna-puttaswamy-and-andrew-hoh), while Uber has talked about its system, called Michelangelo (https://eng.uber.com/michelangelo/).

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2020/01/cortex-labs-logo.png)

"But the rest of the industry doesn't have these machine learning infrastructure teams, so we decided we'd basically try to be that team for everybody else," Spillinger says.

Cortex Labs' software is distributed under an open source license and is available for download from its GitHub Web page (https://github.com/cortexlabs). Making the software open source is critical, Spillinger says, because of the need for standards in this area. There are proprietary offerings in this arena, but they don't have a chance of becoming the standard, whereas Cortex Labs does.

Contributors

Alex Woodie

Editor in Chief

George Leopold

Contributing Editor

Steve Conway

Hyperion Research

Tiffany Trader

Contributing Editor

“We think that if it’s not open source, it’s going to be a lot more difficult for it to become a standard way of doing things,” Spillinger says.

Cortex Labs isn’t the only company talking about the need for standards in the machine learning lifecycle. Last month, Cloudera (http://www.cloudera.com/) announced its intention to push for standards (https://www.datanami.com/2019/12/11/its-time-for-mlops-standards-cloudera-says/) in machine learning operations, or MLOps. Anaconda (http://www.anaconda.com/), which develops a data science platform, also is backing

Eventually, the Oakland, California-based company plans to develop a managed service offering based on its software, Spillinger says. But for now, the company is eager to get the tool into the hands of as many data scientists and machine learning engineers as it can.

**Related Items:**

It’s Time for MLOps Standards, Cloudera Says (https://www.datanami.com/2019/12/11/its-time-for-mlops-standards-cloudera-says/)

Machine Learning Hits a Scaling Bump (https://www.datanami.com/2019/12/12/machine-learning-hits-a-scaling-bump/)

Inference Emerges As Next AI Challenge (https://www.datanami.com/2017/11/02/inference-emerges-next-ai-challenge/)

**Share this:**

(https://www.datanami.com/2020/01/27/an-open-source-alternative-to-aws-sagemaker/?share=reddit)

✉ (https://www.datanami.com/2020/01/27/an-open-source-alternative-to-aws-sagemaker/?share=email)

Tags: AI deployment (https://www.datanami.com/tag/ai-deployment/), Docker (https://www.datanami.com/tag/docker/), ec2 (https://www.datanami.com/tag/ec2/), inference (https://www.datanami.com/tag/inference/), Kubernetes (https://www.datanami.com/tag/kubernetes/), machine learning inference (https://www.datanami.com/tag/machine-learning-inference/), MLOps (https://www.datanami.com/tag/mlops/), Omer Spillinger (https://www.datanami.com/tag/omer-spillinger/), PyTorch (https://www.datanami.com/tag/pytorch/), TensorFlow (https://www.datanami.com/tag/tensorflow/)