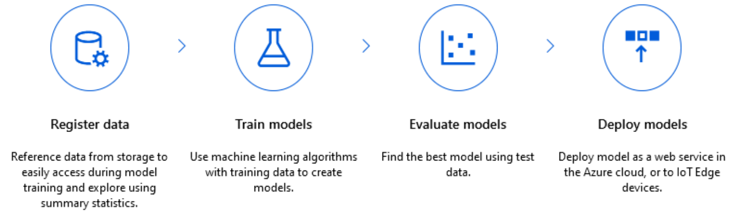
Comparing Azure Machine Learning Service and Azure Databricks

Colby Ford

When it comes to machine learning in Microsoft Azure, there are two main contenders for running your experiments: Azure Machine Learning Service and Azure Databricks. In this post, we will discuss the strengths and capabilities of each service and why you might choose one over the other for all or part of your machine learning workflow. We know that the machine learning workflow and lifecycle doesn't only include training models - there's data prep and performance tracking and, when you're finished modeling, there's also the process of putting your model to work.

Azure Machine Learning Service (AMLS) is Microsoft's homegrown solutions to supporting your end-to-end machine learning lifecycle in Azure. AMLS is a newer service on Azure that's continually getting new features. Currently you can use either the Python SDK or the R SDK to interact with the service or you can use the Designer for a low-code foray into machine learning.

AMLS includes functionality to keep track of datasets, experiments, pipelines, models, and API endpoints. Plus, you can easily provision notebook virtual machines, training clusters, and inference clusters right from the site (and, of course, from the SDK).



Register data

Reference data from storage to easily access during model training and explore using summary statistics.

Train models

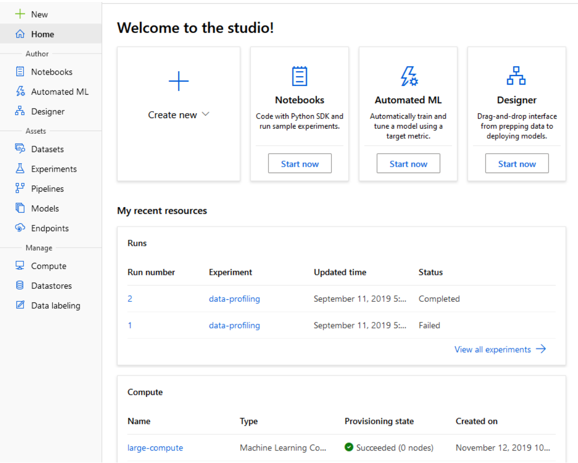
Use machine learning algorithms with training data to create models.

Evaluate models

Find the best model using test data.

Deploy models

Deploy model as a web service in the Azure cloud, or to IoT Edge devices.



Welcome to the studio!

Notebooks

Code with Python SDK and run sample experiments.

Automated ML

Automatically train and tune a model using a target metric.

Designer

Drag-and-drop interface from prepping data to deploying models.

My recent resources

Runs

Run number	Experiment	Updated time	Status
2	data-profiling	September 11, 2019 5:...	Completed
1	data-profiling	September 11, 2019 5:...	Failed

Compute

Name	Type	Provisioning state	Created on
large-compute	Machine Learning Co...	✓ Succeeded (0 nodes)	November 12, 2019 10...

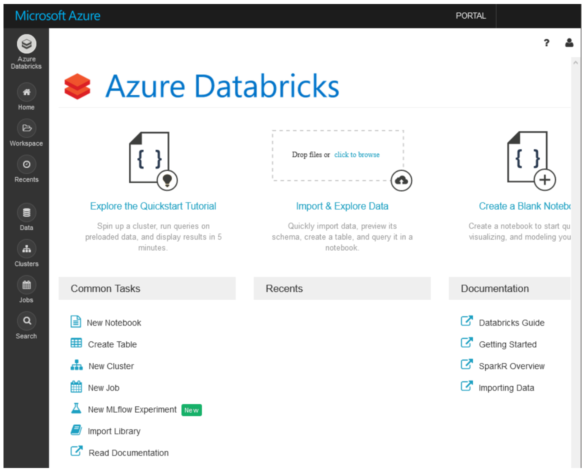
AMLS allows users to use virtually any machine learning package with the service and also includes functionality for automated machine learning (AutoML) and hyperparameter tuning (HyperDrive).

To get started, visit ml.azure.com.

Azure Databricks

If you've been a reader of the BlueGranite blog for any amount of time, you know we never go too long without talking about Azure Databricks. (See our other Databricks-related content here.) Databricks is a unified analytics platform that allows teams to tackle everything from data platform projects to data science solutions using Apache Spark.

One of the biggest benefits, in my opinion, around Databricks is the interface that allows for easily cluster creation, data management, and user collaboration while coding. That is, users can work together in the same notebook, but one person can be writing R in one block and another can be switching back and forth between Python and SQL somewhere else. This environment really makes it easy for users to work on data preparation for machine learning projects by being flexible in the language that's being used.



Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a notebook to start qu visualizing, and modeling you

Apache Spark prides itself with being the state-of-the-art "embarrassingly parallel" system for data processing, but it also has a pretty impressive machine learning library known as MLlib. This highly-scalable library allows for distributed machine learning model training on very large data in a cluster environment.

To learn more about Azure Databricks, click here.

What About Deep Learning?

A hot term you often hear in the AI space are "neural networks" and the use of deep learning frameworks to train them. Whether you're a TensorFlow troubadour or a PyTorch pro, both AMLS and Databricks can support these workloads. Both services allow you use to GPU-enabled virtual machines to train neural network models.

In AMLS, your experience in training neural network models will be quite familiar to your normal, local training. In fact, if you use Keras, TensorFlow, Chainer, or PyTorch, there are easy-to-use estimators in the AMLS SDK to help you along.

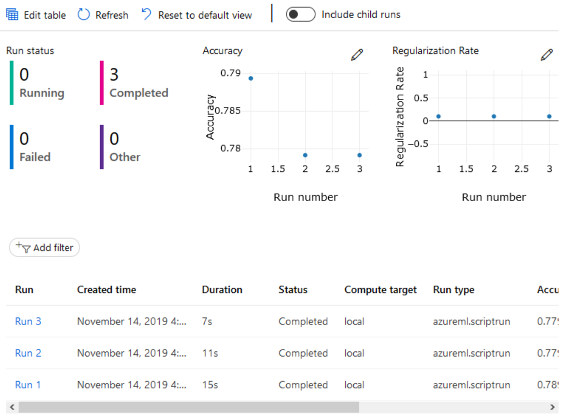
To learn more about Deep Learning in AMLS, click here.

In Databricks, however, since the underlying Spark engine isn't being used while your models are being trained on the GPU(s), you might find it less cost-effective to run your deep learning experiments in Databricks versus AMLS. (This is mainly due to the fact that you're paying a little extra for DBUs that you're not taking advantage of.) One benefit to deep learning in Databricks is the use of Horovod, a distributed training framework that works with TensorFlow, Keras, PyTorch, and MXNet. This is especially useful if your model or your data are too large to fit in memory on a single machine.

To learn more about deep learning in Databricks, click here, or for information on distributed training, click here.

But Wait! I Need To Track My Experiments!

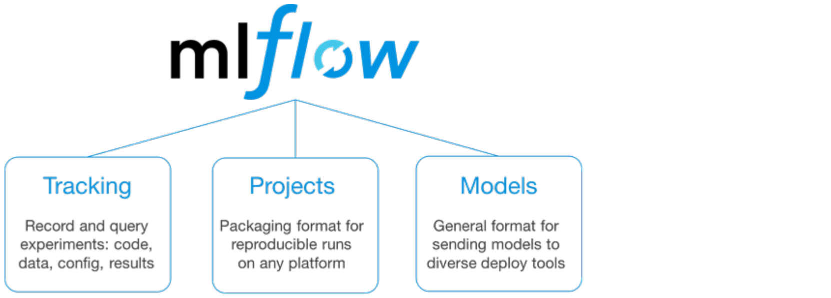
As I mentioned before, AMLS was built with this in mind from the ground up. Once you set up an experiment, AMLS will keep track of the individual runs of that experiment, including any metrics you want to keep an eye on.



Run	Created time	Duration	Status	Compute target	Run type	Accu
Run 3	November 14, 2019 4:...	7s	Completed	local	azureml.scriptrun	0.77!
Run 2	November 14, 2019 4:...	11s	Completed	local	azureml.scriptrun	0.77!
Run 1	November 14, 2019 4:...	15s	Completed	local	azureml.scriptrun	0.78!

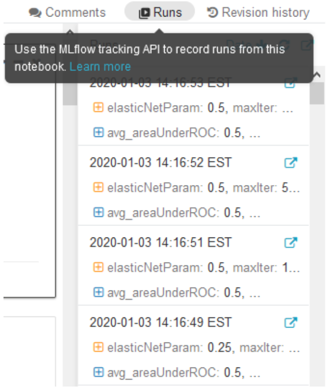
While AMLS is designed around keeping track of everything from machine learning experiments to pipelines and models, Databricks won't be left out. If you want to use Databricks as the place to train your models, but you want to use AMLS to track your results, no problemo! Since the Python SDK is just a series of libraries, you can install them on your Databricks cluster and still take advantage of AMLS.

Alternatively, though, Databricks users can use MLFlow as the tracking engine inside of Databricks. This is an open source package that come pre-installed and enable in the ML runtime versions in Databricks.


mlflow
Tracking
Record and query experiments: code, data, config, results
Projects
Packaging format for reproducible runs on any platform
Models
General format for sending models to diverse deploy tools

Source: https://databricks.com/blog/2018/06/05/introducing-mlflow-an-open-source-machine-learning-platform.html

A big benefit of MLFlow, which is seemingly simpler/less-featured than AMLS, is that MLFlow will automatically track your MLlib experiment runs with zero configuration!



Use the MLflow tracking API to record runs from this notebook. Learn more

2020-01-03 14:16:53 EST
elasticNetParam: 0.5, maxIter: ...
avg_areaUnderROC: 0.5, ...

2020-01-03 14:16:52 EST
elasticNetParam: 0.5, maxIter: 5...
avg_areaUnderROC: 0.5, ...

2020-01-03 14:16:51 EST
elasticNetParam: 0.5, maxIter: 1...
avg_areaUnderROC: 0.5, ...

2020-01-03 14:16:49 EST
elasticNetParam: 0.25, maxIter: ...
avg_areaUnderROC: 0.5, ...

Let's Talk Operationalization...

Once you have finished modeling and you have a model that you are ready to use the model in production, you might want to serve up the model as a callable API. This works well for use cases where you don't have some sort of batch scoring need, but you need your data scored ASAP.

AMLS touts super simple deployments. Within a couple clicks (or lines of SDK code), you can deploy your trained model to a Azure Container Instance or Azure Kubernetes Service-based container, complete with a URI that can be called to score your incoming data. This can even be configured with API keys or tokens. You can further customize by configuring a custom Docker image for niche use cases.

To learn more about how to deploy with AMLS, click here.

As for Databricks, you have 2 options: Take the model out of Spark and operationalize using AMLS or use MMLSpark to make an distributed web service inside your Spark cluster.

MMLSpark is a package by Microsoft that allows Spark to handle a streaming workload of data from an API endpoint. This process is great for complex data that need to be processed prior to being run through a trained model. Since the web service is distributed, there API can take full advantage of the nodes of the cluster, making this fast for heavier workloads.

To learn more about Spark Serving, click here.

Who Wins?

In short, it depends. Spark and Databricks surely wins the battle for scalability, especially in your data preparation step of the machine learning process. Plus, Spark's MLlib is also highly scalable and works well on huge datasets.

However, AMLS' new UI is really nice and the ease-of-use around the tracking capabilities, model interpretability, and model deployment are top-notch. Plus, for deep learning workflows, AMLS just seems a little easier to get going (not to mention, a little more cost-effective).

My oversimplified recommendation? Use Databricks for your heavy lifting (data prep and modeling on large datasets) and use AMLS for tracking, machine learning on normal datasets, deep learning on GPUs, and operationalization.

For your machine learning practice, the correct choice might be a blend of both. Since Databricks can use AMLS as the experiment tracker and as the place to deploy models as APIs and since AMLS can use Databricks as a compute target, the cooperativity between these services is obviously a strength that allows both services to shine!

Want to learn more?

BlueGranite is a top Azure Databricks partner, winning 2018 U.S. System Integrator Partner of the Year award for Databricks. We're also an elite Microsoft partner, helping clients build and deploy modern data platform, modern BI, and machine learning & AI solutions using Power BI and Azure data services. We'd be happy to help you explore Azure Databricks and Azure Machine Learning further. Contact us today to find out how!