What are compute targets in Azure Machine Learning?

Article • 09/30/2022 • 8 minutes to read

A *compute target* is a designated compute resource or environment where you run your training script or host your service deployment. This location might be your local machine or a cloud-based compute resource. Using compute targets makes it easy for you to later change your compute environment without having to change your code.

In a typical model development lifecycle, you might:

1. Start by developing and experimenting on a small amount of data. At this stage, use your local environment, such as a local computer or cloud-based virtual machine (VM), as your compute target.
2. Scale up to larger data, or do distributed training by using one of these training compute targets.
3. After your model is ready, deploy it to a web hosting environment with one of these deployment compute targets.

The compute resources you use for your compute targets are attached to a workspace. Compute resources other than the local machine are shared by users of the workspace.

Training compute targets

Azure Machine Learning has varying support across different compute targets. A typical model development lifecycle starts with development or experimentation on a small amount of data. At this stage, use a local environment like your local computer or a cloud-based VM. As you scale up your training on larger datasets or perform distributed training, use Azure Machine Learning compute to create a single- or multi-node cluster that autoscales each time you submit a job. You can also attach your own compute resource, although support for different scenarios might vary.

Compute targets can be reused from one training job to the next. For example, after you attach a remote VM to your workspace, you can reuse it for multiple jobs. For machine learning pipelines, use the appropriate pipeline step for each compute target.

You can use any of the following resources for a training compute target for most jobs.

Not all resources can be used for automated machine learning, machine learning pipelines, or designer. Azure Databricks can be used as a training resource for local runs and machine learning pipelines, but not as a remote target for other training.

Training targets	Automated machine learning	Machine learning pipelines	Azure Machine Learning designer
Local computer	Yes		
Azure Machine Learning compute cluster	Yes	Yes	Yes
Azure Machine Learning compute instance	Yes (through SDK)	Yes	Yes
Azure Machine Learning Kubernetes	Yes	Yes	Yes
Remote VM	Yes	Yes	
Apache Spark pools (preview)	Yes (SDK local mode only)	Yes	
Azure Databricks	Yes (SDK local mode only)	Yes	
Azure Data Lake Analytics		Yes	
Azure HDInsight		Yes	
Azure Batch		Yes	

Tip

The compute instance has 120GB OS disk. If you run out of disk space, **use the terminal** to clear at least 1-2 GB before you **stop or restart** the compute instance.

Compute targets for inference

When performing inference, Azure Machine Learning creates a Docker container that hosts the model and associated resources needed to use it. This container is then used in a compute target.

The compute target you use to host your model will affect the cost and availability of your deployed endpoint. Use this table to choose an appropriate compute target.

Compute target	Used for	GPU support	Description
Local web service	Testing/debugging		Use for limited testing and troubleshooting. Hardware acceleration depends on use of libraries in the local system.
Azure Machine Learning endpoints	Real-time inference Batch inference	Yes	Fully managed computes for real-time (managed online endpoints) and batch scoring (batch endpoints) on serverless compute.
Azure Machine Learning Kubernetes	Real-time inference Batch inference	Yes	Run inferencing workloads on on-premises, cloud, and edge Kubernetes clusters.
Azure Container Instances (SDK/CLI v1 only)	Real-time inference Recommended for dev/test purposes only.		Use for low-scale CPU-based workloads that require less than 48 GB of RAM. Doesn't require you to manage a cluster. Supported in the designer.

> **Note**
>
> When choosing a cluster SKU, first scale up and then scale out. Start with a machine that has 150% of the RAM your model requires, profile the result and find a machine that has the performance you need. Once you've learned that, increase the number of machines to fit your need for concurrent inference.

> **Note**
>
> Container instances require the SDK or CLI v1 and are suitable only for small models less than 1 GB in size.

Learn where and how to deploy your model to a compute target.

Azure Machine Learning compute (managed)

A managed compute resource is created and managed by Azure Machine Learning. This compute is optimized for machine learning workloads. Azure Machine Learning compute clusters and compute instances are the only managed computes.

You can create Azure Machine Learning compute instances or compute clusters from:

- Azure Machine Learning studio.
- The Python SDK and the Azure CLI:
  - Compute instance.
  - Compute cluster.
- An Azure Resource Manager template. For an example template, see Create an Azure Machine Learning compute cluster .

When created, these compute resources are automatically part of your workspace, unlike other kinds of compute targets.

Capability	Compute cluster	Compute instance
Single- or multi-node cluster	✓	Single node cluster
Autoscales each time you submit a job	✓	
Automatic cluster management and job scheduling	✓	✓
Support for both CPU and GPU resources	✓	✓

> **Note**
>
> When a compute *cluster* is idle, it autoscales to 0 nodes, so you don't pay when it's not in use. A compute *instance* is always on and doesn't autoscale. You should **stop the compute instance** when you aren't using it to avoid extra cost.

Supported VM series and sizes

> **Note**
>
> H-series virtual machine series will be retired on August 31, 2022. Create compute

instance and compute clusters with alternate VM sizes. Existing compute instances and clusters with H-series virtual machines will not work after August 31, 2022.

When you select a node size for a managed compute resource in Azure Machine Learning, you can choose from among select VM sizes available in Azure. Azure offers a range of sizes for Linux and Windows for different workloads. To learn more, see VM types and sizes.

There are a few exceptions and limitations to choosing a VM size:

- Some VM series aren't supported in Azure Machine Learning.
- There are some VM series, such as GPUs and other special SKUs, which may not initially appear in your list of available VMs. But you can still use them, once you request a quota change. For more information about requesting quotas, see Request quota increases. See the following table to learn more about supported series.

Supported VM series	Category	Supported by
DDSv4	General purpose	Compute clusters and instance
Dv2	General purpose	Compute clusters and instance
Dv3	General purpose	Compute clusters and instance
DSv2	General purpose	Compute clusters and instance
DSv3	General purpose	Compute clusters and instance
EAv4	Memory optimized	Compute clusters and instance
Ev3	Memory optimized	Compute clusters and instance
ESv3	Memory optimized	Compute clusters and instance
FSv2	Compute optimized	Compute clusters and instance
FX	Compute optimized	Compute clusters
H	High performance compute	Compute clusters and instance
HB	High performance compute	Compute clusters and instance
HBv2	High performance compute	Compute clusters and instance

Supported VM series	Category	Supported by
HBv3	High performance compute	Compute clusters and instance
HC	High performance compute	Compute clusters and instance
LSv2	Storage optimized	Compute clusters and instance
M	Memory optimized	Compute clusters and instance
NC	GPU	Compute clusters and instance
NC Promo	GPU	Compute clusters and instance
NCv2	GPU	Compute clusters and instance
NCv3	GPU	Compute clusters and instance
ND	GPU	Compute clusters and instance
NDv2	GPU	Compute clusters and instance
NV	GPU	Compute clusters and instance
NVv3	GPU	Compute clusters and instance
NCasT4_v3	GPU	Compute clusters and instance
NDasrA100_v4	GPU	Compute clusters and instance

While Azure Machine Learning supports these VM series, they might not be available in all Azure regions. To check whether VM series are available, see Products available by region .

> **Note**
>
> Azure Machine Learning doesn't support all VM sizes that Azure Compute supports. To list the available VM sizes, use one of the following methods:
>
> - **REST API**
> - The **Azure CLI extension 2.0 for machine learning** command, **az ml compute list-sizes**.

If using the GPU-enabled compute targets, it is important to ensure that the correct

CUDA drivers are installed in the training environment. Use the following table to determine the correct CUDA version to use:

GPU Architecture	Azure VM Series	Supported CUDA versions
Ampere	NDA100_v4	11.0+
Turing	NCT4_v3	10.0+
Volta	NCv3, NDv2	9.0+
Pascal	NCv2, ND	9.0+
Maxwell	NV, NVv3	9.0+
Kepler	NC, NC Promo	9.0+

In addition to ensuring the CUDA version and hardware are compatible, also ensure that the CUDA version is compatible with the version of the machine learning framework you are using:

- For PyTorch, you can check the compatibility by visiting Pytorch's previous versions page .
- For Tensorflow, you can check the compatibility by visiting Tensorflow's build from source page .

Compute isolation

Azure Machine Learning compute offers VM sizes that are isolated to a specific hardware type and dedicated to a single customer. Isolated VM sizes are best suited for workloads that require a high degree of isolation from other customers' workloads for reasons that include meeting compliance and regulatory requirements. Utilizing an isolated size guarantees that your VM will be the only one running on that specific server instance.

The current isolated VM offerings include:

- Standard_M128ms
- Standard_F72s_v2
- Standard_NC24s_v3
- Standard_NC24rs_v3*

*RDMA capable

To learn more about isolation, see Isolation in the Azure public cloud.

Unmanaged compute

An unmanaged compute target is *not* managed by Azure Machine Learning. You create this type of compute target outside Azure Machine Learning and then attach it to your workspace. Unmanaged compute resources can require additional steps for you to maintain or to improve performance for machine learning workloads.

Azure Machine Learning supports the following unmanaged compute types:

- Your local computer
- Remote virtual machines
- Azure HDInsight
- Azure Batch
- Azure Databricks
- Azure Data Lake Analytics
- Azure Container Instance
- Kubernetes

For more information, see set up compute targets for model training and deployment

Next steps

Learn how to:

- Deploy your model to a compute target