Introduction to compute targets

5 minutes

In Azure Machine Learning, *Compute Targets* are physical or virtual computers on which experiments are run.

Types of compute

Azure Machine Learning supports multiple types of compute for experimentation and training. This enables you to select the most appropriate type of compute target for your particular needs.

- **Local compute** - You can specify a local compute target for most processing tasks in Azure Machine Learning. This runs the experiment on the same compute target as the code used to initiate the experiment, which may be your physical workstation or a virtual machine such as an Azure Machine Learning *compute instance* on which you are running a notebook. Local compute is generally a great choice during development and testing with low to moderate volumes of data.
- **Compute clusters** - For experiment workloads with high scalability requirements, you can use Azure Machine Learning compute clusters; which are multi-node clusters of Virtual Machines that automatically scale up or down to meet demand. This is a cost-effective way to run experiments that need to handle large volumes of data or use parallel processing to distribute the workload and reduce the time it takes to run.
- **Attached compute** - If you already use an Azure-based compute environment for data science, such as a virtual machine or an Azure Databricks cluster, you can attach it to your Azure Machine Learning workspace and use it as a compute target for certain types of workload.

ⓘ **Note**

In Azure Machine Learning studio, you can create another type of compute named *inference clusters*. This kind of compute represents an Azure Kubernetes Service cluster and can only be used to deploy trained models as inferencing services. We'll explore deployment later, but for now we'll focus on compute for experiments and model training.

The ability to assign experiment runs to specific compute targets helps you implement a flexible data science ecosystem in the following ways:

- Code can be developed and tested on local or low-cost compute, and then moved to more scalable compute for production workloads.
- You can run individual processes on the compute target that best fits its needs. For example, by using GPU-based compute to train deep learning models, and switching to lower-cost CPU-only compute to test and register the trained model.

One of the core benefits of cloud computing is the ability to manage costs by paying only for what you use. In Azure Machine Learning, you can take advantage of this principle by defining compute targets that:

- Start on-demand and stop automatically when no longer required.
- Scale automatically based on workload processing needs.

Next unit: Create compute targets

Continue >