Learn how to work with compute targets in Azure Machine Learning. Compute targets allow you to run your machine learning workloads. Explore how and when you can use a compute instance or compute cluster.

**Learning objectives**

In this module, you'll learn how to:

* Choose the appropriate compute target.
* Work with compute instances and clusters.
* Manage installed packages with environments.

-- INTRODUCTION

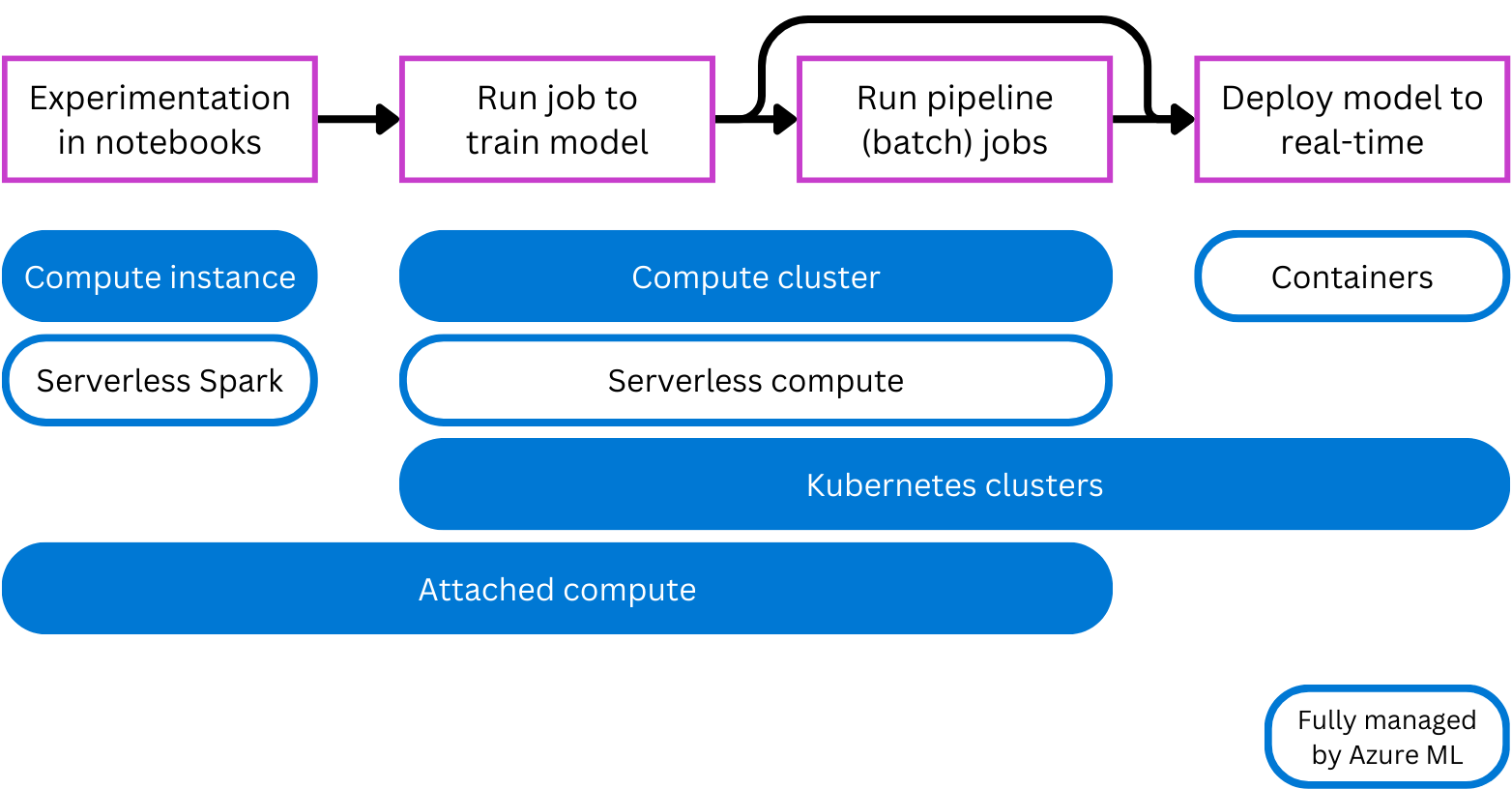
En Azure Machine Learning, puedes escalar tus proyectos de machine learning al utilizar computación en la nube, lo que evita las limitaciones de entrenar modelos en un dispositivo local. Los recursos de computación administrados en la nube de Azure te permiten ahorrar tiempo en la gestión de infraestructura. Ya sea que trabajes en notebooks para experimentación o ejecutes scripts en producción, la computación en Azure Machine Learning facilita la ejecución de tus cargas de trabajo a escala.

-- Choose the appropriate compute target

En Azure Machine Learning, los *compute targets* son dispositivos físicos o virtuales donde se ejecutan trabajos, y puedes elegir entre diferentes tipos según tus necesidades:

* **Instancia de computación**: Funciona como una máquina virtual y se usa principalmente para ejecutar notebooks, ideal para experimentación.
* **Clústeres de computación**: Conjunto de máquinas virtuales escalables automáticamente, rentable para procesar grandes volúmenes de datos y permite procesamiento en paralelo.
* **Clústeres de Kubernetes**: Basados en Kubernetes, ofrecen mayor control sobre la configuración y permiten usar AKS (Azure Kubernetes Service) en la nube o Arc Kubernetes en instalaciones locales.
* **Compute adjunto**: Permite conectar recursos de computación existentes como máquinas virtuales de Azure o clústeres de Databricks.
* **Compute sin servidor**: Computación totalmente administrada y bajo demanda para trabajos de entrenamiento.

Azure Machine Learning te permite gestionar tu propio *compute* o usar computación administrada por Azure.



**-- Guía para elegir el tipo de *compute* en Azure Machine Learning**

1. **Para experimentación**:
   * Si estás desarrollando un modelo y experimentando con un subconjunto pequeño de datos en un *notebook*, una **instancia de computación** en la nube (similar a una máquina virtual) es ideal, ya que se ejecuta continuamente y facilita el trabajo interactivo.
   * Para ejecutar código en Spark, puedes optar por **Spark serverless compute** para aprovechar el procesamiento distribuido.
2. **Para producción**:
   * Una vez listo para producción, el modelo suele requerir entrenamiento continuo. Aquí, los **clústeres de computación** resultan útiles, ya que escalan automáticamente en función de la carga y son eficientes para procesar grandes volúmenes de datos.
   * Como alternativa, puedes usar **compute sin servidor** para entrenamientos sin preocuparte por la gestión de infraestructura.
3. **Para despliegue**:
   * **Predicciones en lote**: Los **clústeres de computación** o **compute sin servidor** son adecuados para ejecutar trabajos en pipeline, escalando bajo demanda.
   * **Predicciones en tiempo real**: Para predicciones en tiempo real, los **contenedores** son la opción recomendada, siendo ligeros y eficientes en costos. Azure Machine Learning puede gestionar los contenedores automáticamente o puedes usar clústeres de Kubernetes para personalizar la configuración.

**Crear y usar una instancia de computación en Azure Machine Learning**

Para ejecutar código en notebooks, puedes usar una **instancia de computación** administrada por Azure Machine Learning. Puedes crear esta instancia desde el Azure Machine Learning Studio, la CLI de Azure o el SDK de Python.

**Crear una instancia de computación con el SDK de Python**

Texto

Descripción generada automáticamentePara crear una instancia de computación usando el SDK de Python, utiliza el siguiente código:

**Nota**: El nombre de la instancia debe ser único dentro de una región de Azure. Si el nombre ya existe, se mostrará un mensaje de error para que intentes con otro nombre.

**Crear una instancia de computación con un script**

Usar un script para crear la instancia permite instalar automáticamente paquetes, herramientas o software necesarios, y también clonar repositorios en la instancia. Esto es útil para crear entornos de desarrollo consistentes cuando varios usuarios necesitan la misma configuración.

**Asignación y uso de una instancia de computación en Azure Machine Learning**

**Asignar una instancia de computación a un usuario**

* Una **instancia de computación** solo puede asignarse a un usuario, ya que no soporta cargas de trabajo en paralelo.
* Si tienes permisos, puedes asignar la instancia a otro usuario al crearla.

**Minimizar el tiempo de computación**

* Para reducir costos, inicia la instancia solo cuando estés trabajando activamente en el notebook, y detén la instancia cuando no estés ejecutando código.
* Puedes programar la instancia para que se inicie y detenga en horarios específicos, o configurarla para que se apague automáticamente después de un período de inactividad.

**Consejo**: Programar el apagado automático al final de cada día puede evitar costos adicionales si olvidas detener la instancia.

**Usar una instancia de computación**

* Puedes utilizar la experiencia de notebooks integrada en el **Azure Machine Learning Studio** para acceder y ejecutar código en la instancia.
* Si prefieres trabajar en **Visual Studio Code** para un control de versiones más sencillo, puedes conectar la instancia para ejecutar celdas de notebooks de forma remota desde VS Code.

-- Crear y usar un clúster de computación en Azure Machine Learning

Después de la experimentación, al preparar el código para producción, es ideal utilizar **clústeres de computación** para ejecutar scripts en entornos escalables. Los clústeres son útiles para manejar cargas de trabajo que requieren escalabilidad, como entrenamiento de modelos o procesamiento de grandes volúmenes de datos.

**Crear un clúster de computación con el SDK de Python**

Texto

Descripción generada automáticamentePara crear un clúster de computación, usa el siguiente código:

**Parámetros clave**:

* **size**: Especifica el tipo de máquina virtual (CPU o GPU).
* **max\_instances**: Máximo de nodos para escalar.
* **tier**: Prioridad baja para reducir costos (disponibilidad no garantizada).

**Usar un clúster de computación**

El clúster se puede utilizar en tres escenarios principales:

1. **Ejecutar un trabajo en pipeline** usando el Designer.
2. **Trabajo de AutoML**.
3. **Ejecutar un script como un trabajo**.

El clúster escala automáticamente cuando recibe un trabajo y se apaga al finalizar, permitiendo también el entrenamiento de múltiples modelos en paralelo (ideal para AutoML).

**Configurar el clúster como destino de computación al ejecutar un script**

Puedes definir el clúster como destino de computación al enviar un trabajo con el siguiente código:

Texto

Descripción generada automáticamente

Una vez que el trabajo se envía, el clúster se escala y comienza a ejecutar el script cuando los nodos están listos. Puedes monitorizar el trabajo en el enlace proporcionado.

Texto

Descripción generada automáticamente