**Deploy a model to a batch endpoint**

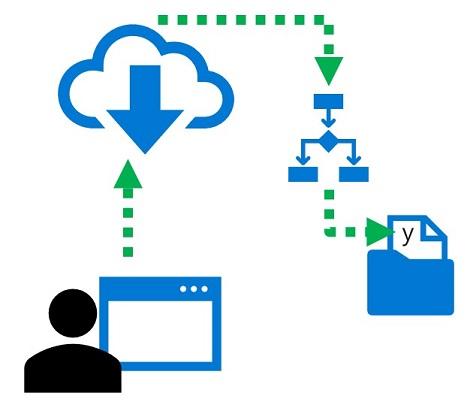
Learn how to deploy models to a batch endpoint. When you invoke a batch endpoint, you'll trigger a batch scoring job.

**Learning objectives**

In this module, you'll learn how to:

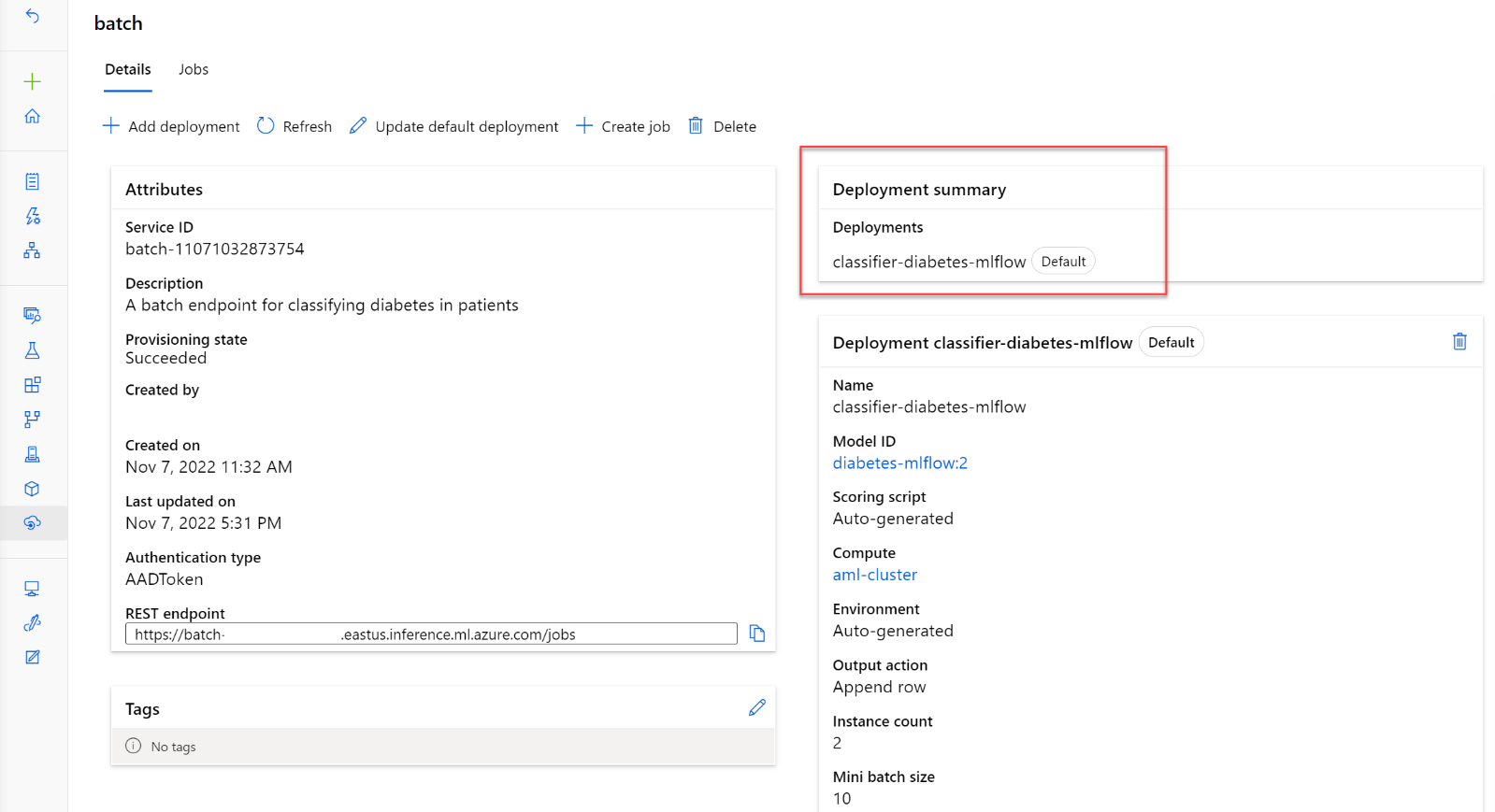
* Create a batch endpoint.
* Deploy your MLflow model to a batch endpoint.
* Deploy a custom model to a batch endpoint.
* Invoke batch endpoints.

--- Se ha entrenado un modelo para predecir las ventas de productos utilizando Azure Machine Learning. Cada mes, se desea utilizar este modelo para pronosticar las ventas del mes siguiente.

En muchos escenarios de producción, las tareas prolongadas que manejan grandes volúmenes de datos se realizan como operaciones por lotes. En el contexto del aprendizaje automático, la inferencia por lotes se utiliza para aplicar de manera asincrónica un modelo predictivo a múltiples casos y escribir los resultados en un archivo o base de datos.

--- Para obtener predicciones por lotes, puedes desplegar un modelo en un endpoint. Este endpoint es un endpoint HTTPS que se puede invocar para activar un trabajo de puntuación por lotes. La ventaja de este tipo de endpoint es que permite integrar la puntuación por lotes con servicios existentes, como Azure Synapse Analytics o Azure Databricks, facilitando así la ingesta y transformación de datos.

Cada vez que se invoca el endpoint, se envía un trabajo de puntuación por lotes al espacio de trabajo de Azure Machine Learning, utilizando generalmente un clúster de computación para puntuar múltiples entradas. Los resultados se pueden almacenar en un datastore conectado al espacio de trabajo.



**Crear un Endpoint por Lotes**

Para desplegar un modelo en un endpoint por lotes, primero debes crear el endpoint utilizando la clase BatchEndpoint. Los nombres de los endpoints deben ser únicos dentro de una región de Azure. El siguiente comando crea un endpoint:

Texto

Descripción generada automáticamente

**Desplegar un Modelo en un Endpoint por Lotes**

Puedes desplegar múltiples modelos en un endpoint por lotes. Al invocar el endpoint, se usará el despliegue predeterminado a menos que se especifique lo contrario.

**Usar Clusters de Computación para Despliegues por Lotes**

El tipo de computación ideal para despliegues por lotes es el clúster de computación de Azure Machine Learning. Si deseas que el trabajo de puntuación por lotes procese los nuevos datos en lotes paralelos, debes provisionar un clúster de computación con más de una instancia máxima.

Para crear un clúster de computación, utiliza la clase AMLCompute:

Texto

Descripción generada automáticamente

Este código configurará un clúster de computación dedicado que puede escalar según sea necesario para el procesamiento por lotes.

--- Una forma sencilla de desplegar un modelo en un endpoint por lotes es utilizando un modelo MLflow, ya que Azure Machine Learning generará automáticamente el script de puntuación y el entorno necesarios.

**Registrar un Modelo MLflow**

Antes de desplegar un modelo MLflow en un endpoint por lotes, debes registrarlo en el espacio de trabajo de Azure Machine Learning. Esto evita la necesidad de un script de puntuación y entorno personalizados. Para registrar un modelo, utiliza la clase Model, especificando el tipo como MLFLOW\_MODEL. El siguiente código registra el modelo desde una ruta local:

Texto

Descripción generada automáticamente

Asegúrate de que la carpeta model contenga el archivo MLmodel, que describe cómo se puede cargar y utilizar el modelo.

**Desplegar un Modelo MLflow en un Endpoint por Lotes**

Para desplegar un modelo MLflow en un endpoint por lotes, utiliza la clase BatchDeployment. Al hacerlo, especificas cómo deseas que se comporte el trabajo de puntuación por lotes. La ventaja de usar un clúster de computación es que puedes ejecutar el script de puntuación en varias instancias de manera paralela.

Los parámetros que puedes configurar en el despliegue incluyen:

* **instance\_count**: Cantidad de nodos de computación a utilizar para generar predicciones.
* **max\_concurrency\_per\_instance**: Número máximo de ejecuciones paralelas del script de puntuación por nodo de computación.
* **mini\_batch\_size**: Número de archivos pasados por cada ejecución del script de puntuación.
* **output\_action**: Qué hacer con las predicciones: summary\_only o append\_row.
* **output\_file\_name**: Archivo al que se agregarán las predicciones, si eliges append\_row para output\_action.

El siguiente código despliega un modelo MLflow en un endpoint por lotes:

Texto

Descripción generada automáticamente

Este código configura el despliegue del modelo para ejecutar la puntuación por lotes y manejar las predicciones de manera eficiente.

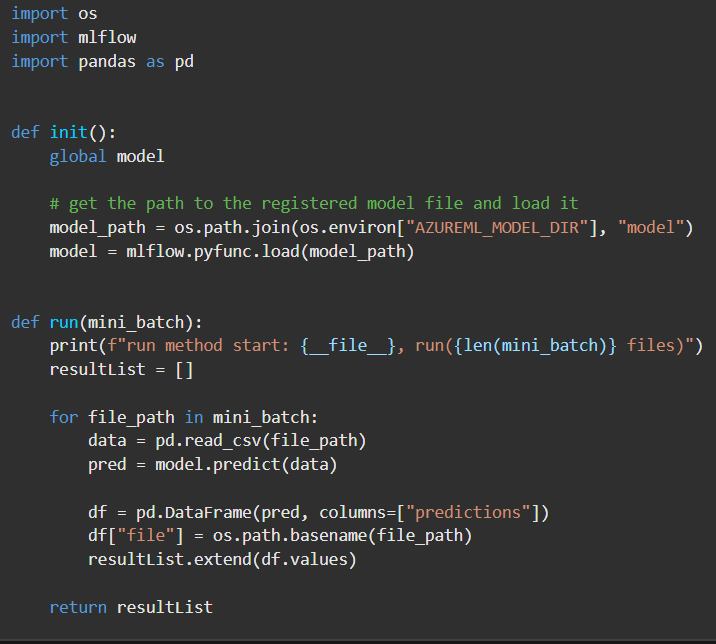
--- Para desplegar un modelo en un endpoint por lotes sin usar el formato de modelo MLflow, es necesario crear un script de puntuación y un entorno de ejecución.

**Crear el Script de Puntuación**

El script de puntuación es un archivo que lee los nuevos datos, carga el modelo y realiza la puntuación. Debe incluir dos funciones:

* **init()**: Se llama una vez al inicio del proceso, ideal para preparaciones costosas, como cargar el modelo.
* **run()**: Se llama para cada mini lote para realizar la puntuación. Este método debe devolver un DataFrame de pandas o una lista/array.

Un ejemplo de script de puntuación es el siguiente:



**Crear un Entorno**

El despliegue necesita un entorno de ejecución que incluya todas las dependencias requeridas. Puedes crear un entorno utilizando una imagen de Docker con dependencias de Conda o un Dockerfile. Asegúrate de incluir la biblioteca azureml-core, necesaria para los despliegues por lotes.

Define las dependencias en un archivo conda.yaml:

Texto

Descripción generada automáticamente

Luego, crea el entorno con el siguiente código:

Texto

Descripción generada automáticamente

**Configurar y Crear el Despliegue**

Finalmente, puedes configurar y crear el despliegue utilizando la clase BatchDeployment. Aquí tienes el código para hacerlo:

Texto

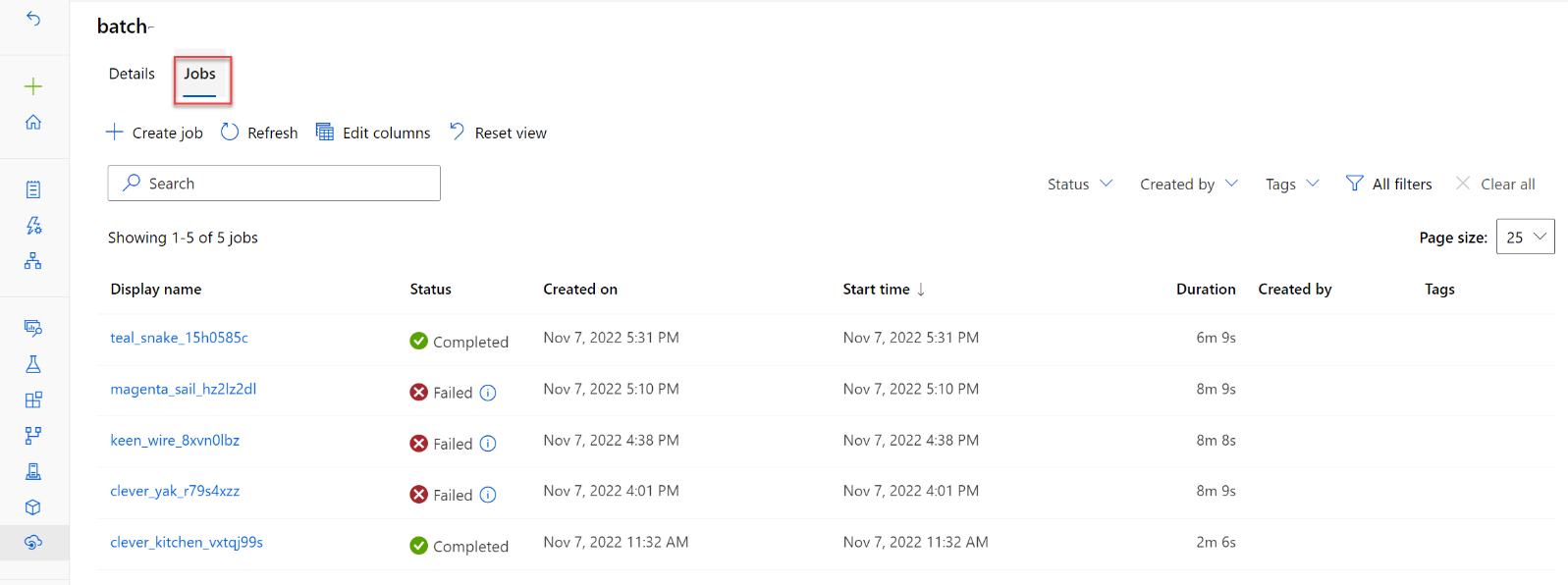
Descripción generada automáticamente

--- **Invocar el Trabajo de Puntuación por Lotes**

Al invocar un endpoint por lotes, inicias un trabajo de pipeline en Azure Machine Learning que espera un parámetro de entrada que apunte al conjunto de datos que deseas puntuar. Para preparar los datos para las predicciones por lotes, primero registra una carpeta como un activo de datos en el espacio de trabajo de Azure Machine Learning.

Texto

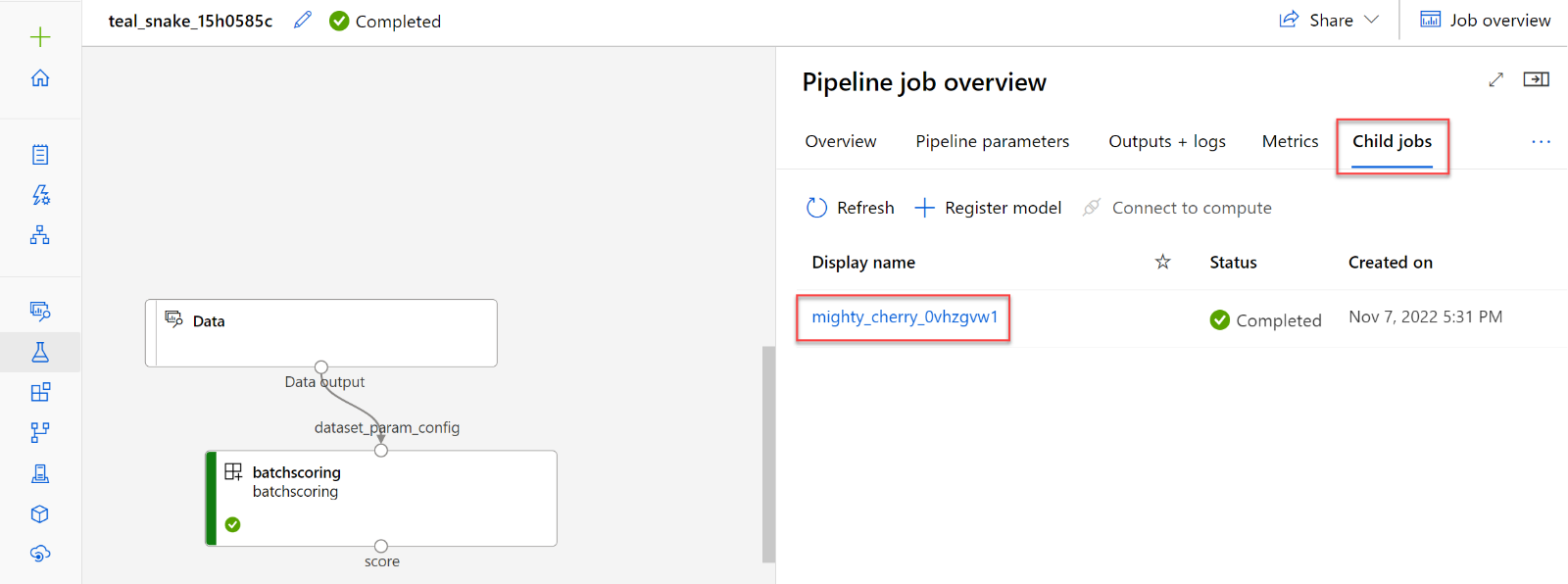
Descripción generada automáticamentePuedes usar el activo de datos registrado como entrada al invocar el endpoint por lotes mediante el SDK de Python:



Puedes monitorear la ejecución del trabajo de pipeline en el estudio de Azure Machine Learning. Todos los trabajos desencadenados por la invocación del endpoint por lotes aparecerán en la pestaña de trabajos del endpoint.

**Solucionar Problemas en un Trabajo de Puntuación por Lotes**

El trabajo de puntuación por lotes se ejecuta como un trabajo de pipeline. Para solucionar problemas, revisa los detalles y los registros de salida del trabajo de pipeline.



Si necesitas solucionar problemas en el script de puntuación, selecciona el trabajo hijo y revisa sus salidas y registros. Navega a la pestaña **Outputs + logs**. En la carpeta logs/user/, encontrarás tres archivos útiles para la solución de problemas:

* **job\_error.txt**: Resume los errores en tu script.
* **job\_progress\_overview.txt**: Proporciona información general sobre el número de mini lotes procesados hasta el momento.
* **job\_result.txt**: Muestra errores al llamar a las funciones init() y run() en el script de puntuación.

Estos registros te ayudarán a identificar y corregir cualquier problema en el proceso de puntuación por lotes.

