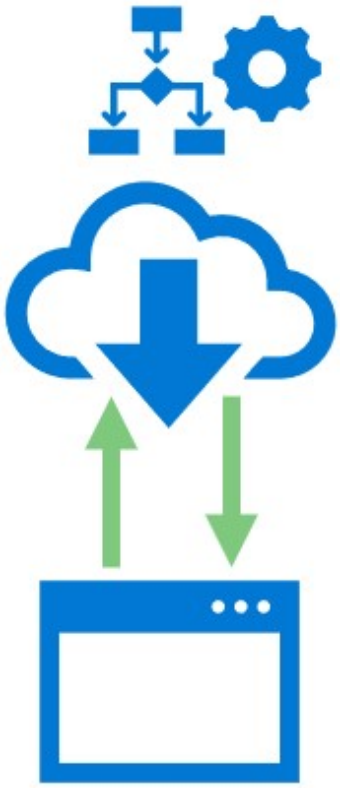
Highlight Note

Publishing a Service Endpoint

After creating and modifying an inference pipeline, you can publish an endpoint through which client applications can consume it as a web service.

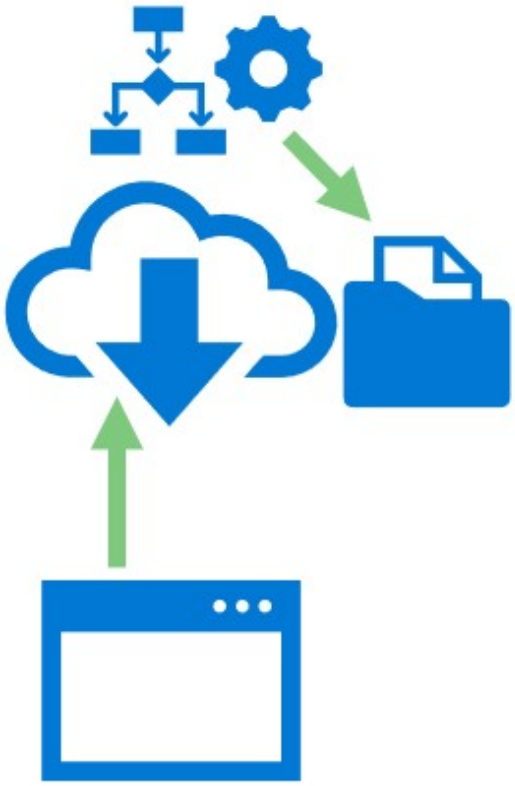
Deploying a Real-Time Inference Pipeline



For real-time inferencing, you must deploy the pipeline as a service on an Azure Kubernetes Services (AKS) compute target. The deployed pipeline service can then be accessed through an HTTP REST endpoint.

Publishing a Batch Inference Pipeline

If you have created a batch inference pipeline, you can publish an HTTP endpoint through which the pipeline can be initiated. It will run on the Azure Machine Learning training compute target you have selected for the inference pipeline.



Note: It's important to note that batch inference pipelines are run on *training* compute, even when published as consumable services.