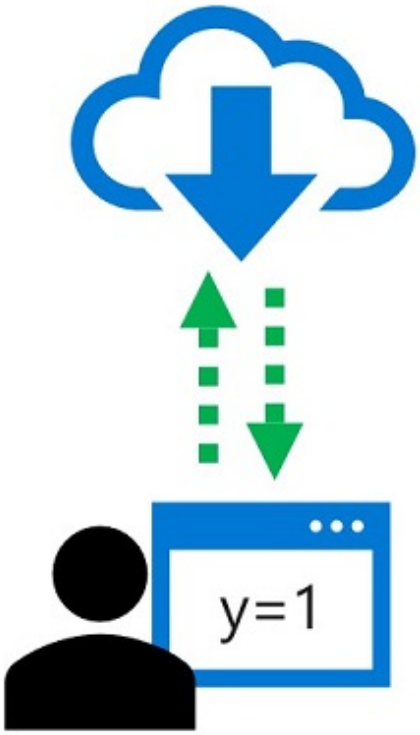
Highlight Note

What is Real-Time Inferencing?

In machine learning, *inferencing* refers to the use of a trained model to predict labels for new data on which the model has not been trained. Often, the model is deployed as part of a service that enables applications to request immediate, or *real-time*, predictions for individual or small numbers of data observations.



In Azure Machine learning, you can create real-time inferencing solutions by deploying a model as a real-time service, hosted in a containerized platform such as Azure Kubernetes Services (AKS).