Monitor and collect data from ML web service endpoints

06/09/2020 • 5 minutes to read • +2

In this article

Prerequisites
Web service metadata and response data
Use Python SDK to configure
Use Azure Machine Learning studio to configure
Evaluate data
Export data for further processing and longer retention
Example notebook
Next steps

APPLIES TO: ✓ Basic edition ✓ Enterprise edition (Upgrade to Enterprise edition)

In this article, you learn how to collect data from and monitor models deployed to web service endpoints in Azure Kubernetes Service (AKS) or Azure Container Instances (ACI) by enabling Azure Application Insights via

- Azure Machine Learning Python SDK
- Azure Machine Learning studio at https://ml.azure.com

In addition to collecting an endpoint's output data and response, you can monitor:

- Request rates, response times, and failure rates
- Dependency rates, response times, and failure rates
- Exceptions

Learn more about Azure Application Insights.

Prerequisites

- If you don't have an Azure subscription, create a free account before you begin. Try the free or paid version of Azure Machine Learning today
- An Azure Machine Learning workspace, a local directory that contains your scripts, and the Azure Machine Learning SDK for Python installed. To learn how to get these prerequisites, see How to configure a development environment

- A trained machine learning model to be deployed to Azure Kubernetes Service (AKS) or Azure Container Instance (ACI). If you don't have one, see the Train image classification model tutorial

Web service metadata and response data

> ⓘ **Important**
>
> Azure Application Insights only logs payloads of up to 64kb. If this limit is reached then you may see errors such as out of memory, or no information may be logged.

To log information for a request to the web service, add `print` statements to your score.py file. Each `print` statement results in one entry in the trace table in Application Insights, under the message `STDOUT`. The contents of the `print` statement will be contained under `customDimensions` and then `Contents` in the trace table. If you print a JSON string, it produces a hierarchical data structure in the trace output under `Contents`.

You can query Azure Application Insights directly to access this data, or set up a continuous export to a storage account for longer retention or further processing. Model data can then be used in the Azure Machine Learning to set up labeling, retraining, explainability, data analysis, or other use.

Use Python SDK to configure

Update a deployed service

1. Identify the service in your workspace. The value for `ws` is the name of your workspace

   Python Copy

   ```python
   from azureml.core.webservice import Webservice
   aks_service= Webservice(ws, "my-service-name")
   ```

2. Update your service and enable Azure Application Insights

   Python Copy

   ```python
   aks_service.update(enable_app_insights=True)
   ```

Log custom traces in your service

If you want to log custom traces, follow the standard deployment process for AKS or ACI in the How to deploy and where document. Then use the following steps:

1. To send data to Application Insights during inference, update the scoring file by adding print statements. To log more complex information, such as the request data and the response, us a JSON structure. The following example score.py file logs the time the model is initialized, the input and output during inference, and the time any errors occur:

   ⓘ **Important**

   Azure Application Insights only logs payloads of up to 64kb. If this limit is reached, you may see errors such as out of memory, or no information may be logged. If the data you want to log is larger 64kb, you should instead store it to blob storage using the information in **Collect Data for models in production**.

   Python Copy

```
import pickle
import json
import numpy
from sklearn.externals import joblib
from sklearn.linear_model import Ridge
from azureml.core.model import Model
import time

def init():
    global model
    #Print statement for appinsights custom traces:
    print ("model initialized" + time.strftime("%H:%M:%S"))

    # note here "sklearn_regression_model.pkl" is the name of the model
registered under the workspace
    # this call should return the path to the model.pkl file on the
local disk.
    model_path = Model.get_model_path(model_name =
'sklearn_regression_model.pkl')

    # deserialize the model file back into a sklearn model
    model = joblib.load(model_path)


# note you can pass in multiple rows for scoring
```

```
def run(raw_data):
    try:
        data = json.loads(raw_data)['data']
        data = numpy.array(data)
        result = model.predict(data)
        # Log the input and output data to appinsights:
        info = {
            "input": raw_data,
            "output": result.tolist()
            }
        print(json.dumps(info))
        # you can return any datatype as long as it is JSON-
serializable
        return result.tolist()
    except Exception as e:
        error = str(e)
        print (error + time.strftime("%H:%M:%S"))
        return error
```

2. Update the service configuration

Python Copy

```
config = Webservice.deploy_configuration(enable_app_insights=True)
```

3. Build an image and deploy it on AKS or ACI.

Disable tracking in Python

To disable Azure Application Insights, use the following code:

Python Copy

```
## replace <service_name> with the name of the web service
<service_name>.update(enable_app_insights=False)
```

Use Azure Machine Learning studio to configure

You can also enable Azure Application Insights from Azure Machine Learning studio when you're ready to deploy your model with these steps.

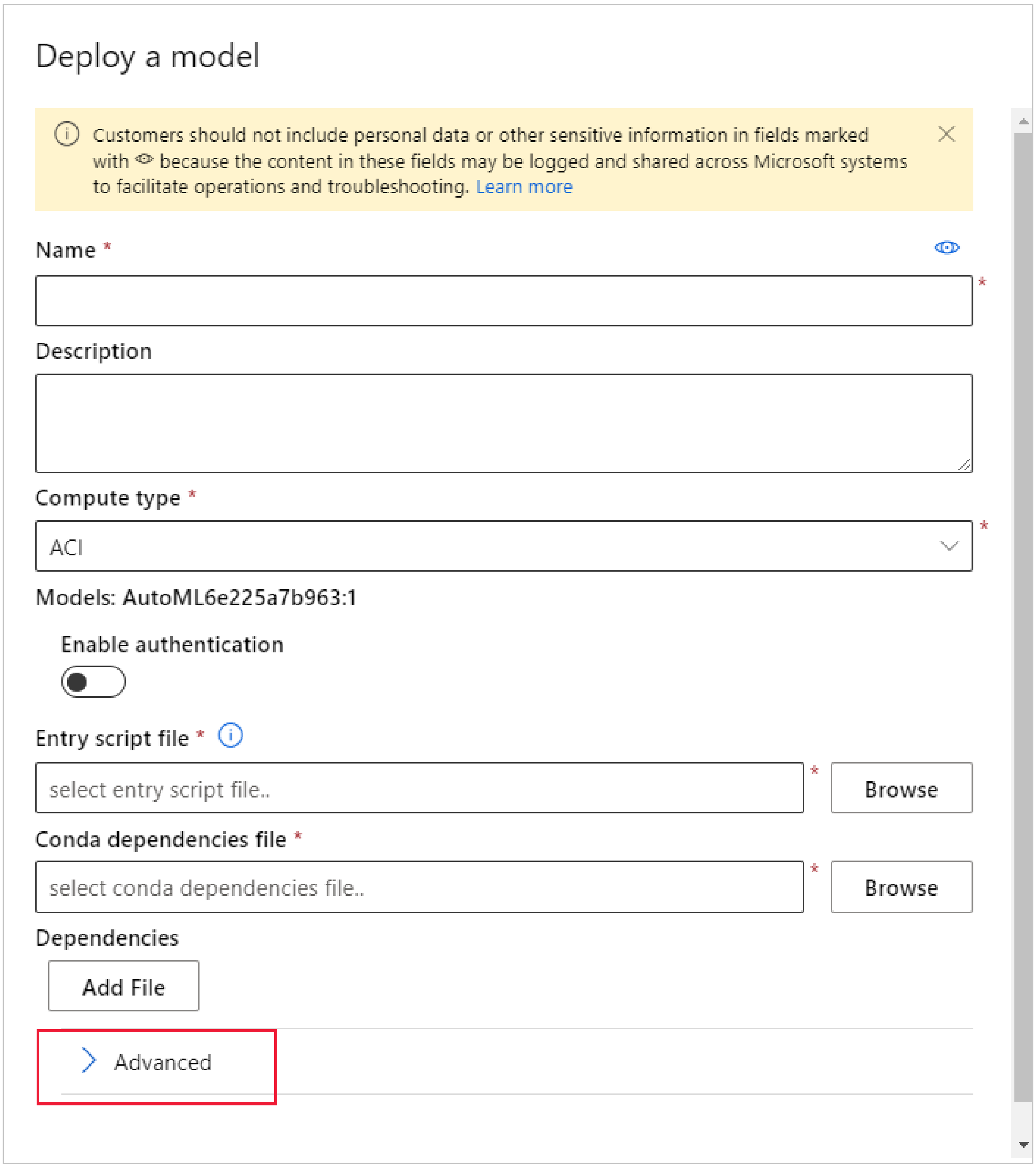
1. Sign in to your workspace at https://ml.azure.com/

2. Go to **Models** and select which model you want to deploy

3. Select **+Deploy**

4. Populate the **Deploy model** form

5. Expand the **Advanced** menu



Deploy a model

Customers should not include personal data or other sensitive information in fields marked with because the content in these fields may be logged and shared across Microsoft systems to facilitate operations and troubleshooting. Learn more

Name *

Description

Compute type *

ACI

Models: AutoML6e225a7b963:1

Enable authentication

Entry script file *

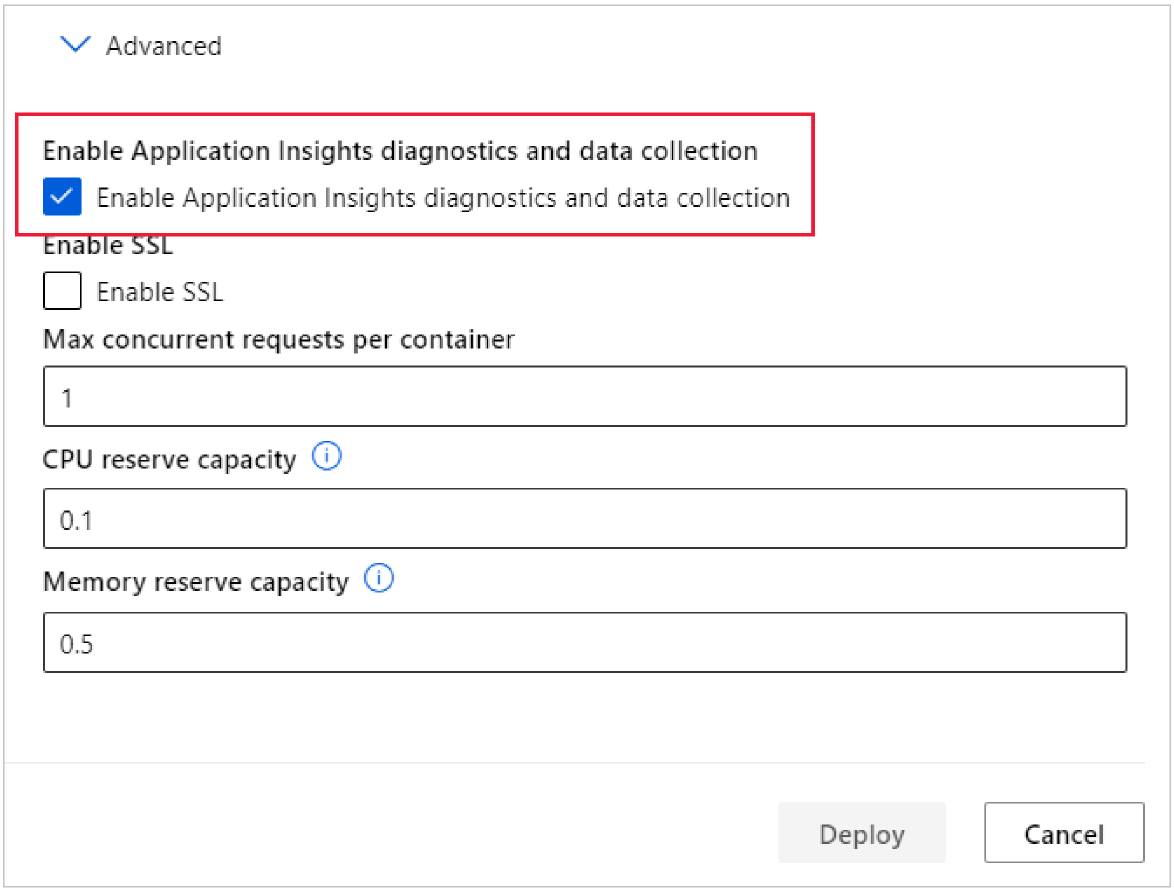
select entry script file..

Conda dependencies file *

select conda dependencies file..

Dependencies

6. Select **Enable Application Insights diagnostics and data collection**



Advanced

Enable Application Insights diagnostics and data collection

☑ Enable Application Insights diagnostics and data collection

Enable SSL

□ Enable SSL

Max concurrent requests per container

1

CPU reserve capacity

0.1

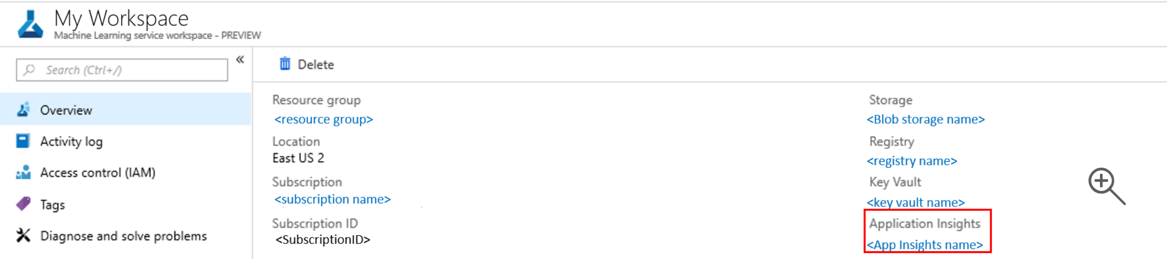
Memory reserve capacity

0.5

Evaluate data

Your service's data is stored in your Azure Application Insights account, within the same resource group as Azure Machine Learning. To view it:

1. Go to your Azure Machine Learning workspace in the Azure portal and click on the Application Insights link



My Workspace
Machine Learning service workspace - PREVIEW

Resource group
<resource group>

Location
East US 2

Subscription
<subscription name>

Subscription ID
<SubscriptionID>

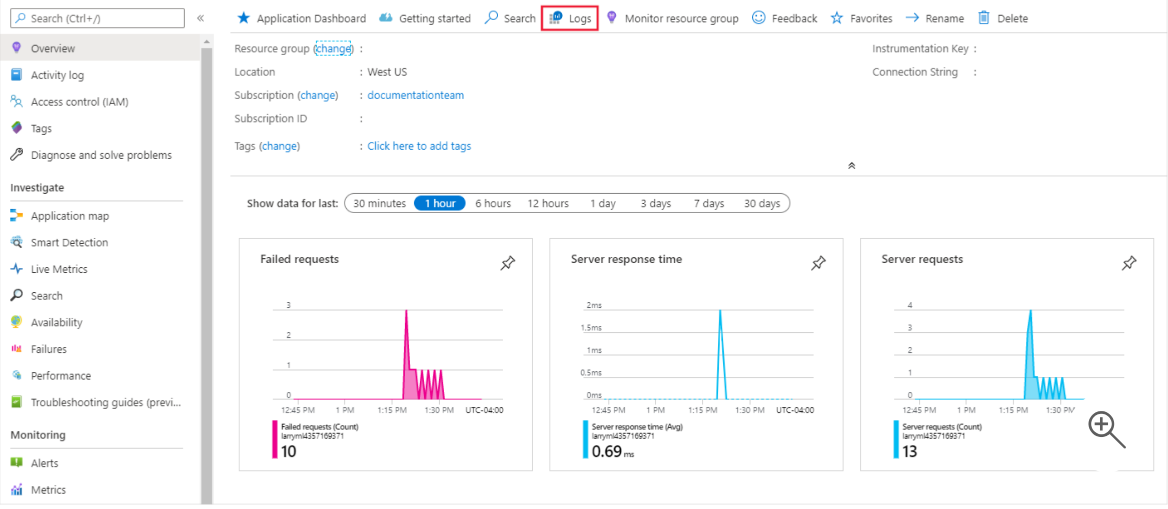
Storage
<Blob storage name>

Registry
<registry name>

Key Vault
<key vault name>

Application Insights
<App Insights name>

2. From the **Overview** tab or the **Monitoring** section in the list on the left, select **Logs**.



Resource group (change) :

Location : West US

Subscription (change) : documentationteam

Subscription ID :

Tags (change) : Click here to add tags

Instrumentation Key :

Connection String :

Failed requests (Count) larryml4357169371 10

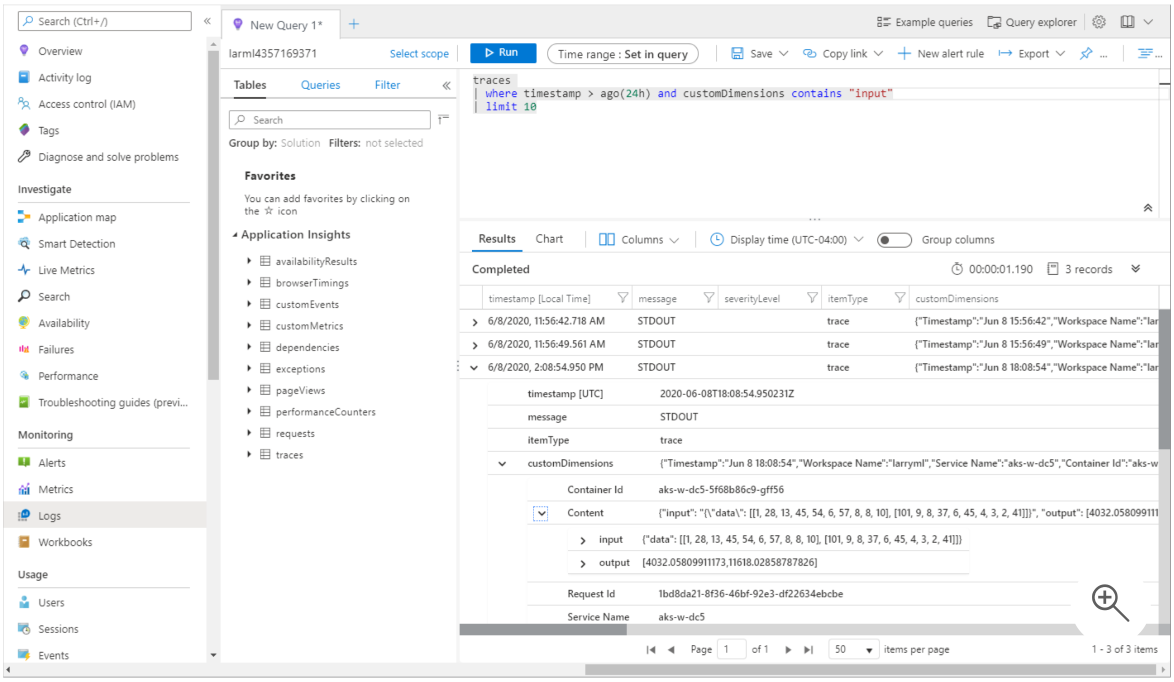
Server response time (Avg) larryml4357169371 0.69 ms

Server requests (Count) larryml4357169371 13

3. To view information logged from the score.py file, look at the **traces** table. The following query searches for logs where the **input** value was logged:

Kusto

```
traces
| where customDimensions contains "input"
| limit 10
```



larml4357169371

Group by: Solution Filters: not selected

Application Insights
- availabilityResults
- browserTimings
- customEvents
- customMetrics
- dependencies
- exceptions
- pageViews
- performanceCounters
- requests
- traces

```
traces
| where timestamp > ago(24h) and customDimensions contains "input"
| limit 10
```

Completed 00:00:01.190 3 records

timestamp [Local Time]	message	severityLevel	itemType	customDimensions
6/8/2020, 11:56:42.718 AM	STDOUT		trace	{"Timestamp":"Jun 8 15:56:42","Workspace Name":"lar
6/8/2020, 11:56:49.561 AM	STDOUT		trace	{"Timestamp":"Jun 8 15:56:49","Workspace Name":"lar
6/8/2020, 2:08:54.950 PM	STDOUT		trace	{"Timestamp":"Jun 8 18:08:54","Workspace Name":"lar

timestamp [UTC]	2020-06-08T18:08:54.950231Z
message	STDOUT
itemType	trace
customDimensions	{"Timestamp":"Jun 8 18:08:54","Workspace Name":"larryml","Service Name":"aks-w-dc5","Container Id":"aks-w
Container Id	aks-w-dc5-5f68b86c9-gff56
Content	{"input": "{\"data\": [[1, 28, 13, 45, 54, 6, 57, 8, 8, 10], [101, 9, 8, 37, 6, 45, 4, 3, 2, 41]]}", "output": [4032.058099111
input	{"data": [[1, 28, 13, 45, 54, 6, 57, 8, 8, 10], [101, 9, 8, 37, 6, 45, 4, 3, 2, 41]]}
output	[4032.05809911173,11618.02858787826]
Request Id	1bd8da21-8f36-46bf-92e3-df22634ebcbe
Service Name	aks-w-dc5

To learn more about how to use Azure Application Insights, see What is Application Insights?.

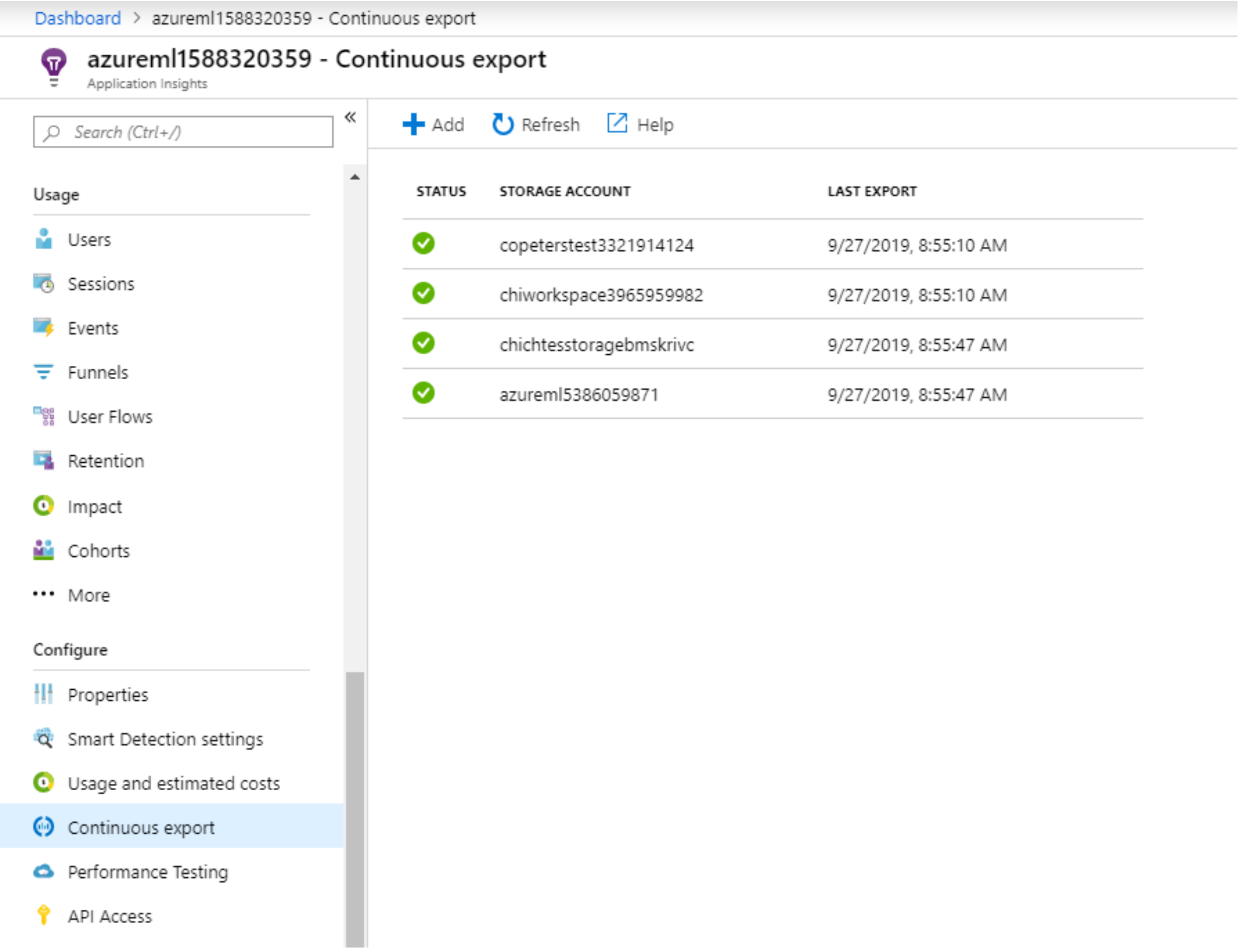
Export data for further processing and longer retention

ⓘ **Important**

Azure Application Insights only supports exports to blob storage. Additional limits of this export capability are listed in **Export telemetry from App Insights**.

You can use Azure Application Insights' continuous export to send messages to a supported storage account, where a longer retention can be set. The data is stored in JSON format and can be easily parsed to extract model data.

Azure Data Factory, Azure ML Pipelines, or other data processing tools can be used to transform the data as needed. When you have transformed the data, you can then register it with the Azure Machine Learning workspace as a dataset. To do so, see How to create and register datasets.



STATUS	STORAGE ACCOUNT	LAST EXPORT
✓	copeterstest3321914124	9/27/2019, 8:55:10 AM
✓	chiworkspace3965959982	9/27/2019, 8:55:10 AM
✓	chichtesstoragebmskrivc	9/27/2019, 8:55:47 AM
✓	azureml5386059871	9/27/2019, 8:55:47 AM

Example notebook

The enable-app-insights-in-production-service.ipynb notebook demonstrates concepts in this article.

Learn how to run notebooks by following the article Use Jupyter notebooks to explore this service.

Next steps

- See how to deploy a model to an Azure Kubernetes Service cluster or how to deploy a model to Azure Container Instances to deploy your models to web service endpoints, and enable Azure Application Insights to leverage data collection and endpoint monitoring
- See MLOps: Manage, deploy, and monitor models with Azure Machine Learning to learn more about leveraging data collected from models in production. Such data can help to continually improve your machine learning process