Monitor online endpoints

Article • 10/19/2022 • 6 minutes to read

In this article, you learn how to monitor Azure Machine Learning online endpoints. Use Application Insights to view metrics and create alerts to stay up to date with your online endpoints.

In this article you learn how to:

- ✓ View metrics for your online endpoint
- ✓ Create a dashboard for your metrics
- ✓ Create a metric alert

Prerequisites

- Deploy an Azure Machine Learning online endpoint.
- You must have at least Reader access on the endpoint.

Metrics

Use the following steps to view metrics for an online endpoint or deployment:

1. Go to the Azure portal .
2. Navigate to the online endpoint or deployment resource.

   online endpoints and deployments are Azure Resource Manager (ARM) resources that can be found by going to their owning resource group. Look for the resource types **Machine Learning online endpoint** and **Machine Learning online deployment**.

3. In the left-hand column, select **Metrics**.

Available metrics

Depending on the resource that you select, the metrics that you see will be different. Metrics are scoped differently for online endpoints and online deployments.

Metrics at endpoint scope

- Request Latency
- Request Latency P50 (Request latency at the 50th percentile)
- Request Latency P90 (Request latency at the 90th percentile)
- Request Latency P95 (Request latency at the 95th percentile)
- Requests per minute
- New connections per second
- Active connection count
- Network bytes

Split on the following dimensions:

- Deployment
- Status Code
- Status Code Class

Bandwidth throttling

Bandwidth will be throttled if the limits are exceeded for *managed* online endpoints (see managed online endpoints section in Manage and increase quotas for resources with Azure Machine Learning). To determine if requests are throttled:

- Monitor the "Network bytes" metric
- The response trailers will have the fields: `ms-azureml-bandwidth-request-delay-ms` and `ms-azureml-bandwidth-response-delay-ms`. The values of the fields are the delays, in milliseconds, of the bandwidth throttling.

Metrics at deployment scope

- CPU Utilization Percentage
- Deployment Capacity (the number of instances of the requested instance type)
- Disk Utilization
- GPU Memory Utilization (only applicable to GPU instances)
- GPU Utilization (only applicable to GPU instances)
- Memory Utilization Percentage

Split on the following dimension:

- InstanceId

Create a dashboard

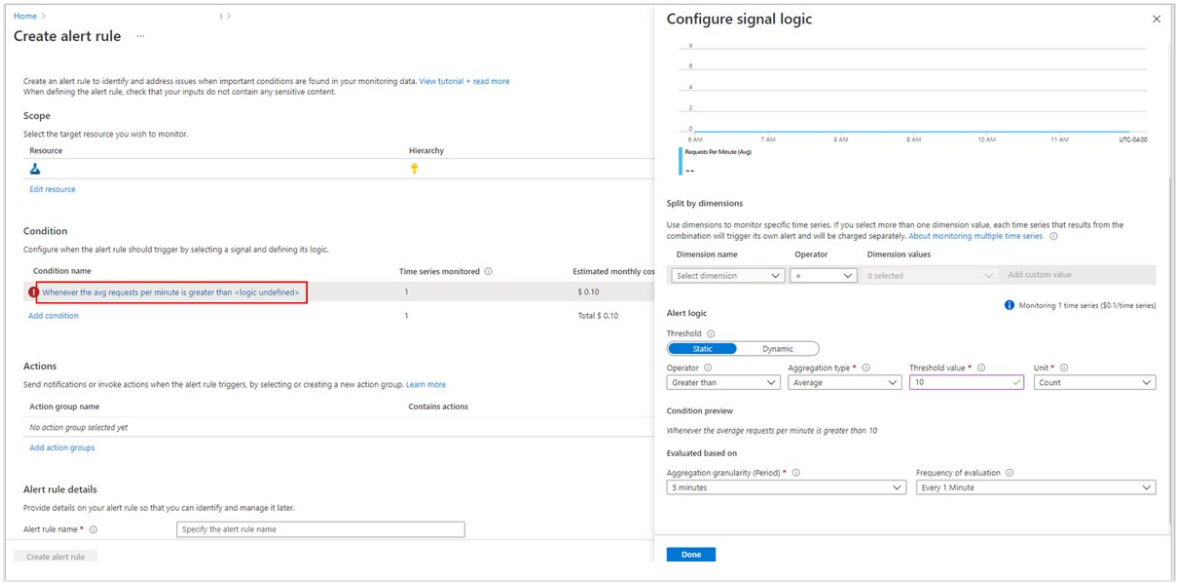
You can create custom dashboards to visualize data from multiple sources in the Azure portal, including the metrics for your online endpoint. For more information, see Create custom KPI dashboards using Application Insights.

Create an alert

You can also create custom alerts to notify you of important status updates to your online endpoint:

1. At the top right of the metrics page, select **New alert rule**.

2. Select a condition name to specify when your alert should be triggered.



3. Select **Add action groups** > **Create action groups** to specify what should happen when your alert is triggered.

4. Choose **Create alert rule** to finish creating your alert.

Logs

There are three logs that can be enabled for online endpoints:

- **AMLOnlineEndpointTrafficLog** (preview): You could choose to enable traffic logs if you want to check the information of your request. Below are some cases:
  - If the response isn't 200, check the value of the column “ResponseCodeReason” to see what happened. Also check the reason in the "HTTPS status codes" section of the Troubleshoot online endpoints article.
  - You could check the response code and response reason of your model from the column “ModelStatusCode” and “ModelStatusReason”.
  - You want to check the duration of the request like total duration, the request/response duration, and the delay caused by the network throttling. You could check it from the logs to see the breakdown latency.
  - If you want to check how many requests or failed requests recently. You could also enable the logs.
- **AMLOnlineEndpointConsoleLog**: Contains logs that the containers output to the console. Below are some cases:
  - If the container fails to start, the console log may be useful for debugging.
  - Monitor container behavior and make sure that all requests are correctly handled.
  - Write request IDs in the console log. Joining the request ID, the AMLOnlineEndpointConsoleLog, and AMLOnlineEndpointTrafficLog in the Log Analytics workspace, you can trace a request from the network entry point of an online endpoint to the container.
  - You may also use this log for performance analysis in determining the time required by the model to process each request.
- **AMLOnlineEndpointEventLog** (preview): Contains event information regarding the container’s life cycle. Currently, we provide information on the following types of

events:

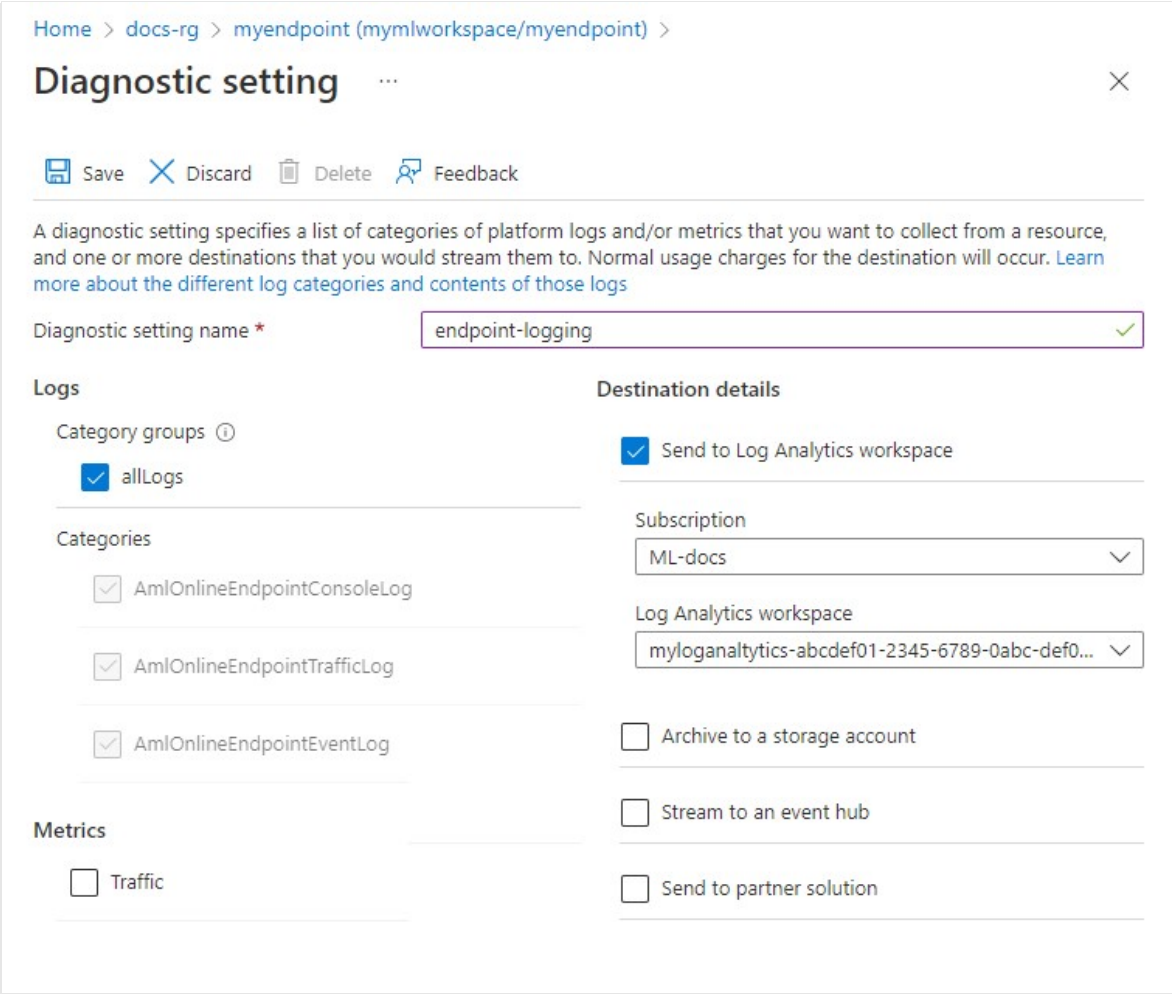
Name	Message
BackOff	Back-off restarting failed container
Pulled	Container image "<IMAGE_NAME>" already present on machine
Killing	Container inference-server failed liveness probe, will be restarted
Created	Created container image-fetcher
Created	Created container inference-server
Created	Created container model-mount
Unhealthy	Liveness probe failed: <FAILURE_CONTENT>
Unhealthy	Readiness probe failed: <FAILURE_CONTENT>
Started	Started container image-fetcher
Started	Started container inference-server
Started	Started container model-mount
Killing	Stopping container inference-server
Killing	Stopping container model-mount

How to enable/disable logs

> **Important**
>
> Logging uses Azure Log Analytics. If you do not currently have a Log Analytics workspace, you can create one using the steps in **Create a Log Analytics workspace in the Azure portal**.

1. In the Azure portal , go to the resource group that contains your endpoint and then select the endpoint.
2. From the **Monitoring** section on the left of the page, select **Diagnostic settings** and then **Add settings**.

3. Select the log categories to enable, select **Send to Log Analytics workspace**, and then select the Log Analytics workspace to use. Finally, enter a **Diagnostic setting name** and select **Save**.

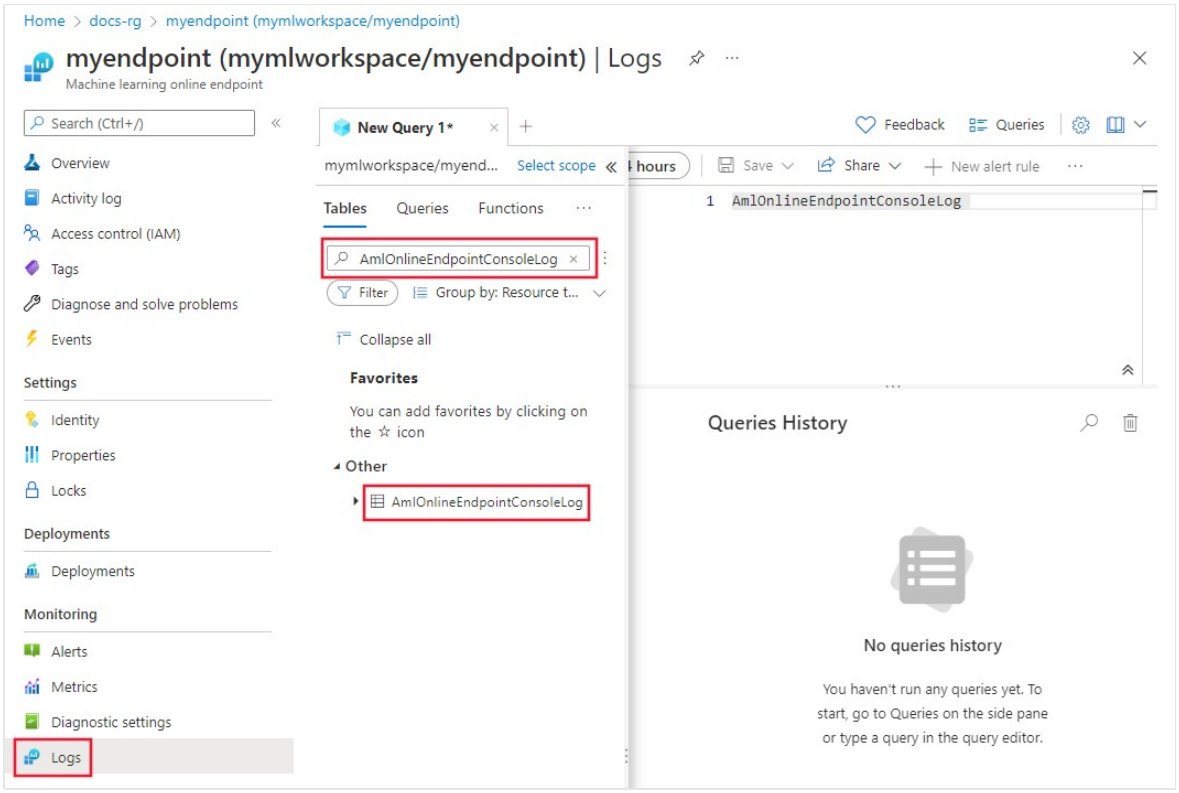


> **Important**
>
> It may take up to an hour for the connection to the Log Analytics workspace to be enabled. Wait an hour before continuing with the next steps.

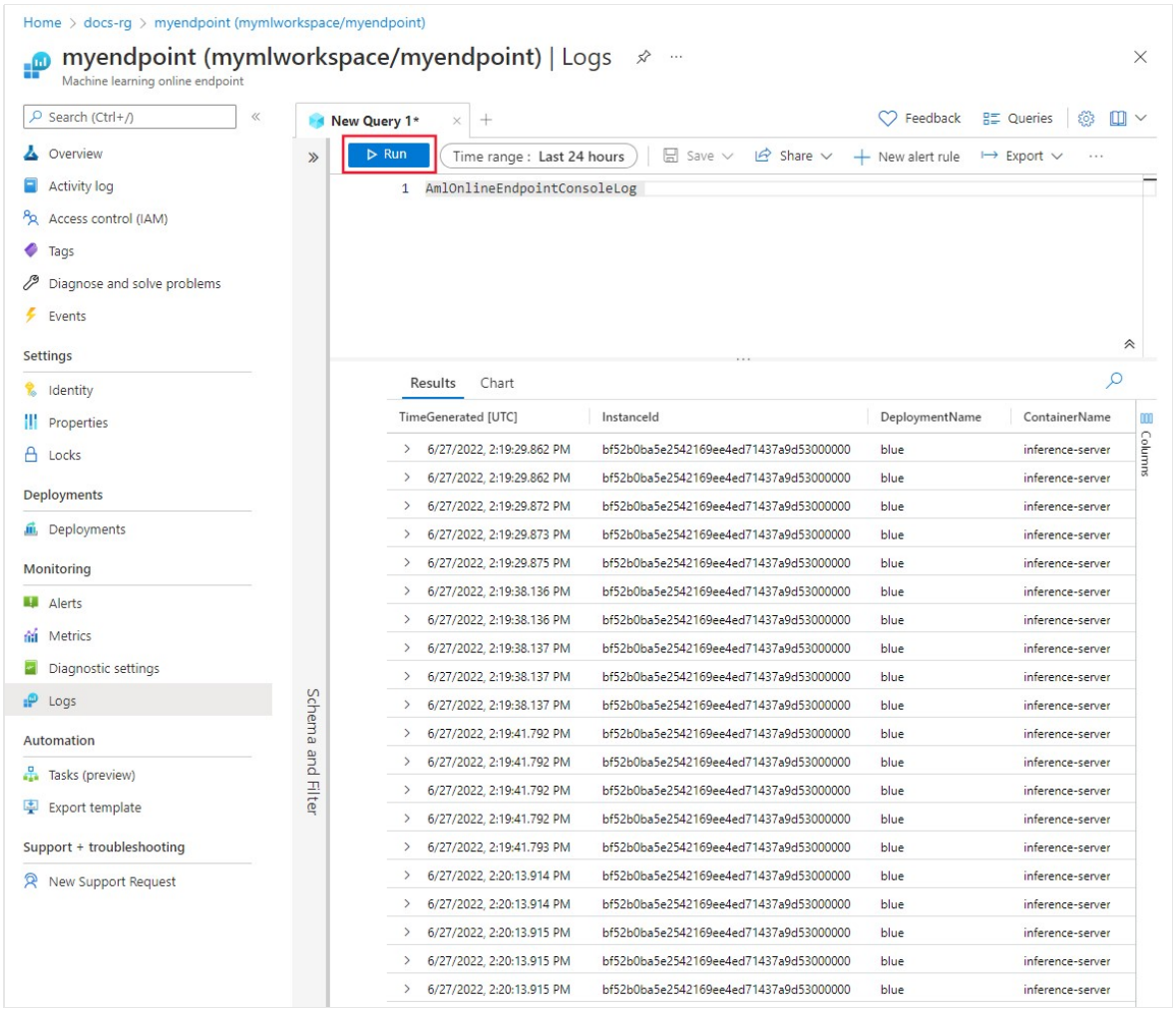
4. Submit scoring requests to the endpoint. This activity should create entries in the logs.

5. From either the online endpoint properties or the Log Analytics workspace, select **Logs** from the left of the screen.

6. Close the **Queries** dialog that automatically opens, and then double-click the **AmlOnlineEndpointConsoleLog**. If you don't see it, use the **Search** field.

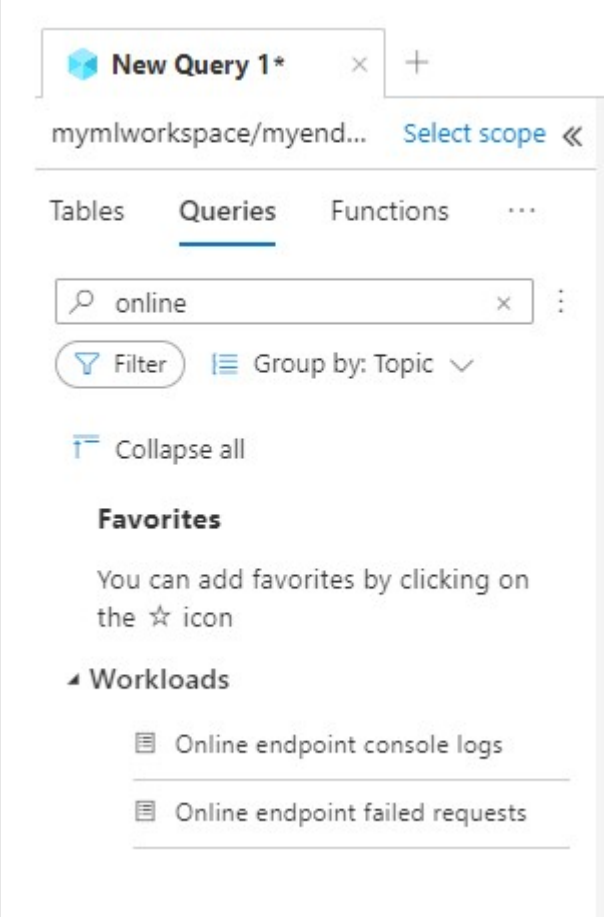

Home > docs-rg > myendpoint (mymlworkspace/myendpoint)
myendpoint (mymlworkspace/myendpoint) | Logs
Machine learning online endpoint
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Events
Settings
Identity
Properties
Locks
Deployments
Deployments
Monitoring
Alerts
Metrics
Diagnostic settings
Logs
New Query 1*
Feedback
Queries
mymlworkspace/myend...
Select scope
hours
Save
Share
New alert rule
Tables
Queries
Functions
AmlOnlineEndpointConsoleLog
Filter
Group by: Resource t...
Collapse all
Favorites
You can add favorites by clicking on the ☆ icon
Other
AmlOnlineEndpointConsoleLog
1 AmlOnlineEndpointConsoleLog
Queries History
No queries history
You haven't run any queries yet. To start, go to Queries on the side pane or type a query in the query editor.

7. Select **Run**.



Example queries

You can find example queries on the **Queries** tab while viewing logs. Search for **Online endpoint** to find example queries.



Log column details

The following tables provide details on the data stored in each log:

AMLOnlineEndpointTrafficLog (preview)

Field name	Description
Method	The requested method from client.
Path	The requested path from client.
SubscriptionId	The machine learning subscription ID of the online endpoint.
WorkspaceId	The machine learning workspace ID of the online endpoint.
EndpointName	The name of the online endpoint.
DeploymentName	The name of the online deployment.
Protocol	The protocol of the request.

Field name	Description
ResponseCode	The final response code returned to the client.
ResponseCodeReason	The final response code reason returned to the client.
ModelStatusCode	The response status code from model.
ModelStatusReason	The response status reason from model.
RequestPayloadSize	The total bytes received from the client.
ResponsePayloadSize	The total bytes sent back to the client.
UserAgent	The user-agent header of the request.
XRequestId	The request ID generated by Azure Machine Learning for internal tracing.
XMSClientRequestId	The tracking ID generated by the client.
TotalDurationMs	Duration in milliseconds from the request start time to the last response byte sent back to the client. If the client disconnected, it measures from the start time to client disconnect time.
RequestDurationMs	Duration in milliseconds from the request start time to the last byte of the request received from the client.
ResponseDurationMs	Duration in milliseconds from the request start time to the first response byte read from the model.
RequestThrottlingDelayMs	Delay in milliseconds in request data transfer due to network throttling.
ResponseThrottlingDelayMs	Delay in milliseconds in response data transfer due to network throttling.

AMLOnlineEndpointConsoleLog

Field Name	Description
TimeGenerated	The timestamp (UTC) of when the log was generated.
OperationName	The operation associated with log record.
InstanceId	The ID of the instance that generated this log record.
DeploymentName	The name of the deployment associated with the log record.

Field Name	Description
ContainerName	The name of the container where the log was generated.
Message	The content of the log.

AMLOnlineEndpointEventLog (preview)

Field Name	Description
TimeGenerated	The timestamp (UTC) of when the log was generated.
OperationName	The operation associated with log record.
InstanceId	The ID of the instance that generated this log record.
DeploymentName	The name of the deployment associated with the log record.
Name	The name of the event.
Message	The content of the event.

Next steps

- Learn how to view costs for your deployed endpoint.
- Read more about metrics explorer.