1.4 - Monitor Azure AI services

Overview

Azure AI services enable you to integrate artificial intelligence into your applications and services. It's important to be able to monitor Azure AI Services in order to track utilization, determine trends, and detect and troubleshoot issues.

After completing this module, you will be able to:

- Monitor Azure AI services costs.
- Create alerts and view metrics for Azure AI services.
- Manage Azure AI services diagnostic logging.

Introduction

Azure AI services provides a cloud-based platform for building artificial intelligence capabilities into your applications. Like any software service, you should monitor AI services to **track costs, identify utilization trends, and detect potential issues.**

After completing this module, you'll be able to:

- Monitor Azure AI services costs.
- Create alerts and view metrics for Azure AI services.
- Manage Azure AI services diagnostic logging.

Monitor cost

One of the main benefits of using cloud services is that you can gain cost efficiencies by only paying for services as you use them. Some Azure AI services resources offer a free tier with restrictions on use, which is useful for development and testing; and one or more billed tiers that incur charges based on transactions. The specific billing rate depends on the resource type.

Plan costs for AI services

Before deploying a solution that depends on AI services, you can estimate costs by using the Azure Pricing Calculator.

To use the pricing calculator to estimate AI services costs, create a new estimate and select **Azure AI services** in the **AI + Machine Learning** category. Then select the specific AI service API you plan to use (for example, *Azure AI Text Analytics*), the region where you plan to provision it, and the pricing tier of the instance you plan to use; and fill in the expected usage metrics and support option. To create an estimate that includes multiple AI services APIs, add additional **Azure AI services** products to the estimate.

After you've created an estimate, you can save it. You can also export it in Microsoft Excel format.

View costs for AI services

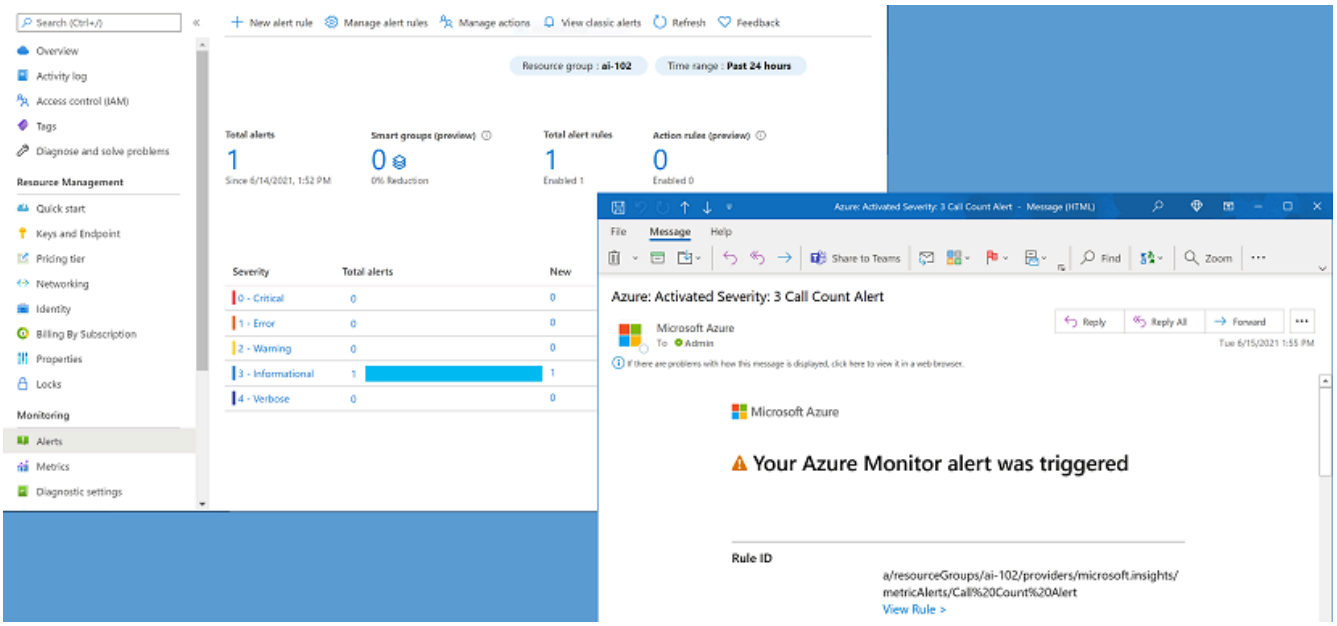
In common with other Azure resources, you can view details of accumulated costs for AI services resources in the Azure portal.

To view costs for AI services, sign into the Azure portal and select your subscription. You can then view overall costs for the subscription by selecting the **Cost analysis** tab. To view only costs for AI services, add a filter that restricts the data to reflect resources with a **service name** of **Cognitive Services**.

> Note: For more information, see Plan and manage costs for Azure AI services in the AI services documentation.

Create alerts

Microsoft Azure provides alerting support for resources through the creation of *alert rules*. You use alert rules to configure notifications and alerts for your resources based on events or metric thresholds. These alerts will ensure that the correct team knows when a problem arises.



Alert rules

To create an alert rule for an Azure AI services resource, select the resource in the Azure portal and on the **Alerts** tab, add a new alert rule. To define the alert rule, you must specify:

- The *scope* of the alert rule - in other words, the resource you want to monitor.
- A **condition** on which the alert is triggered. The specific trigger for the alert is based on a **signal type**, which can be **Activity Log** (an entry in the activity log created by an **action performed on the resource,** such as regenerating its subscription keys) or **Metric** (a **metric threshold** such as the number of errors exceeding 10 in an hour).
- Optional *actions*, such as sending an email to an administrator notifying them of the alert, or running an **Azure Logic App** to address the issue automatically. *Azure Logic Apps* is a cloud-based platform for creating and running automated workflows that integrate your apps, data, services, and systems.
- *Alert rule details*, such as a name for the alert rule and the resource group in which it should be defined.

> Note: For more information, see Overview of alerts in Microsoft Azure in the Azure documentation.

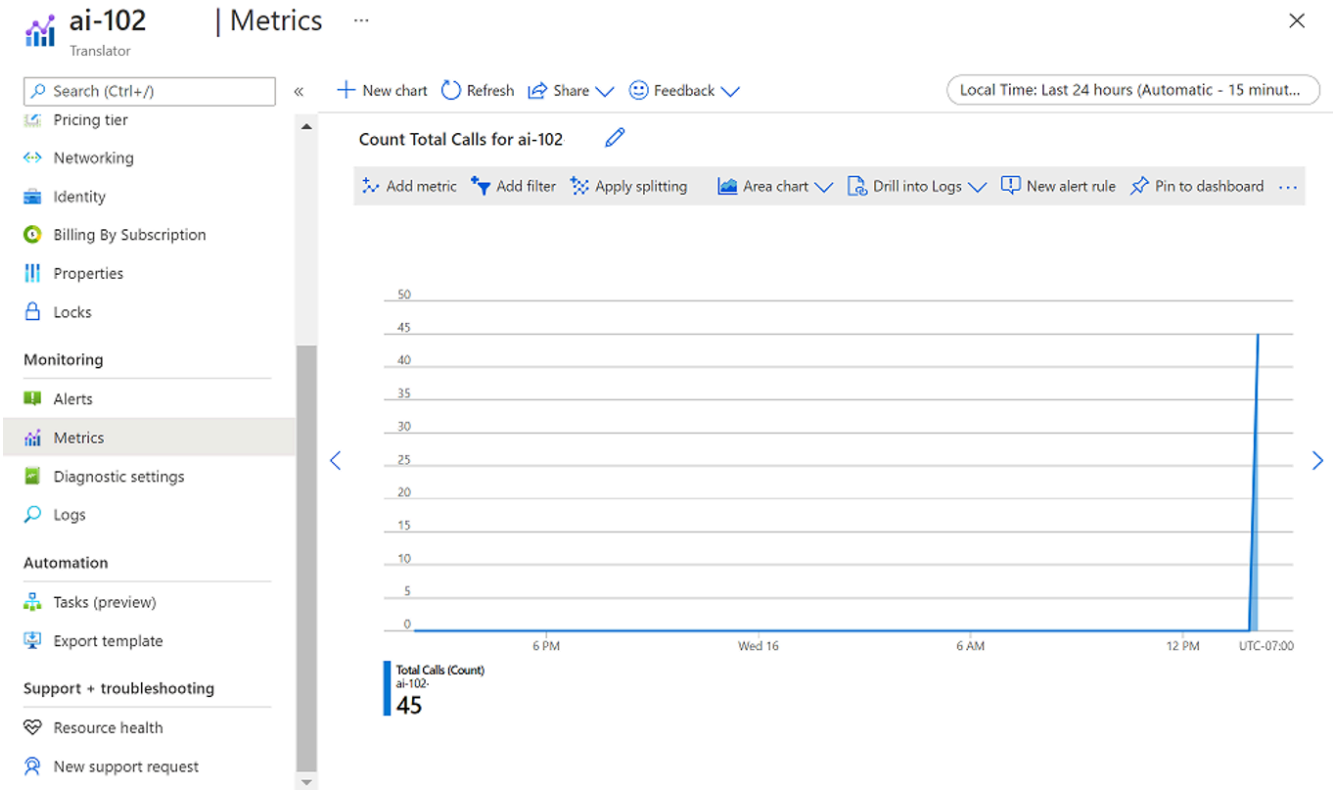
View Metrics

Azure Monitor collects metrics for Azure resources at regular intervals so that you can track indicators of resource utilization, health, and performance. The specific metrics gathered depend on the Azure resource. In the case of Azure AI services, Azure Monitor collects metrics relating to **endpoint requests, data submitted and returned, errors, and other useful measurements.**

View metrics in the Azure portal

You can view metrics for an individual resource in the Azure portal by selecting the resource and viewing its **Metrics** page. On this page, you can add resource-specific metrics to charts. By default an empty chart is created for you, and you can add more charts as required.

For example, the following image shows the **Metrics** page for an AI services resource, showing the **count of total calls to the service** over a period of time.

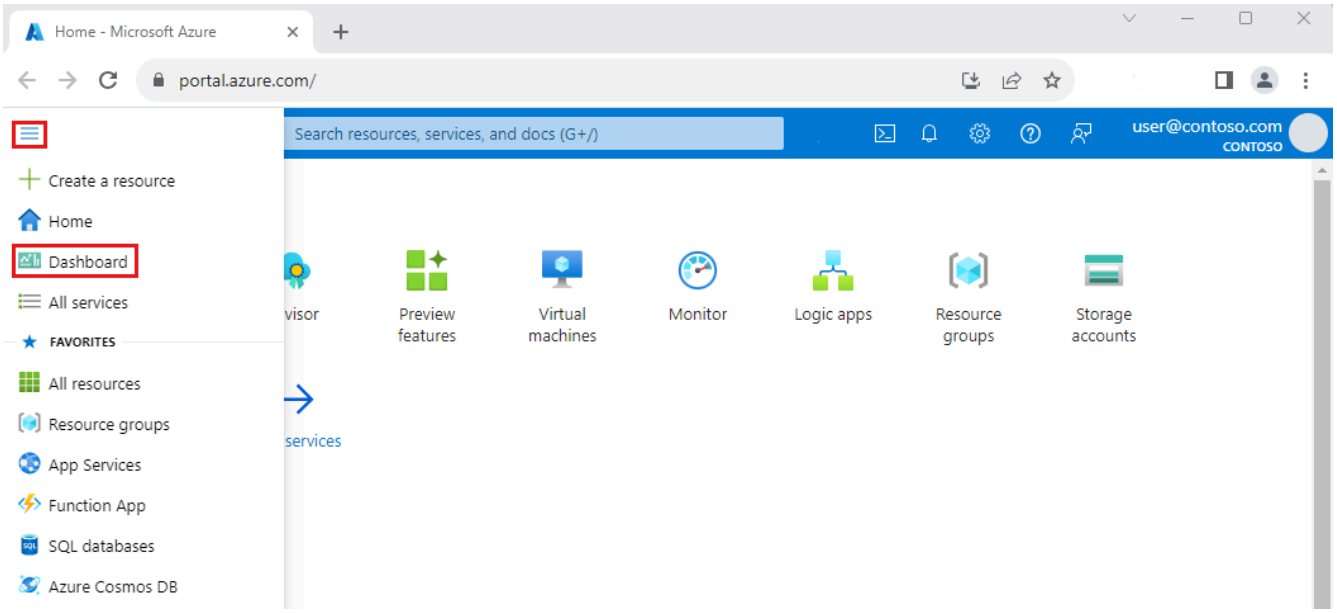


You can add multiple metrics to a chart and choose appropriate aggregations and chart types. When you're satisfied with chart, you can *share* it by exporting it to Excel or copying a link to it, and you can *clone* it to create a duplicate chart in the **Metrics** page - potentially as a starting point for a new chart that shows the same metrics in a different way.

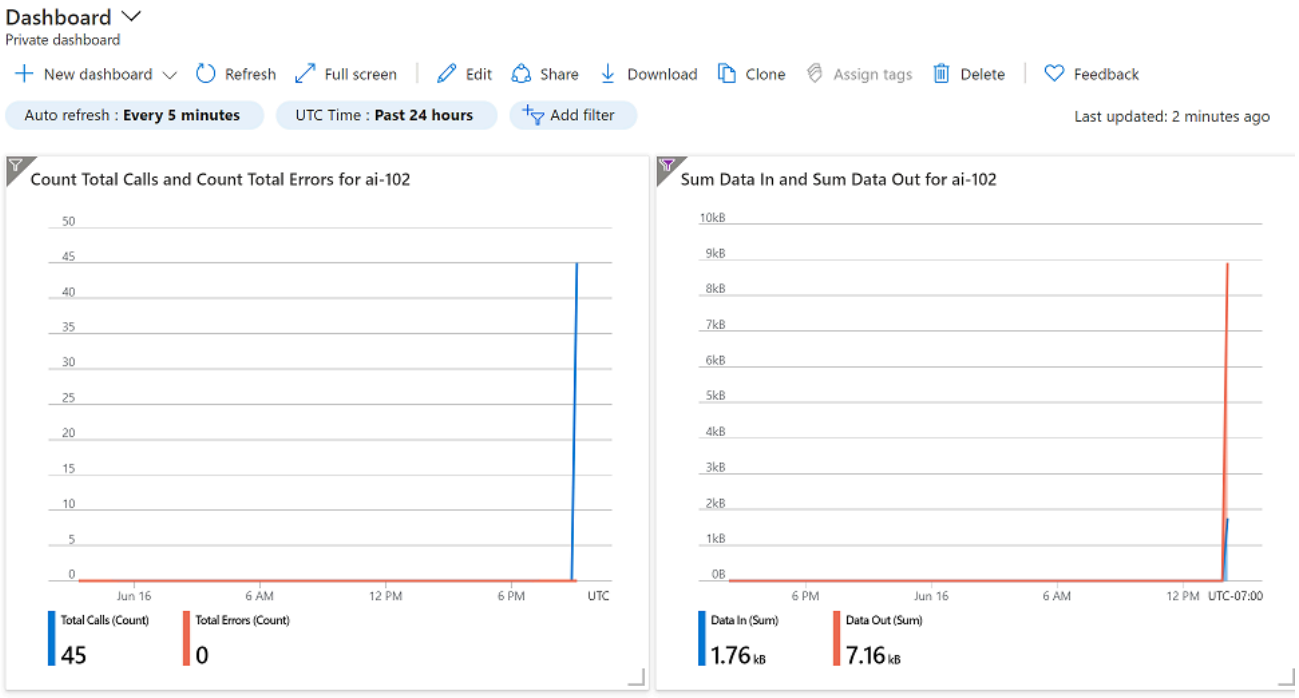
Add metrics to a dashboard

In the Azure portal, you can create *dashboards* that consist of **multiple visualizations from different resources in your Azure environment** to help you gain an overall view of the health and performance of your Azure resources.

To create a dashboard, select **Dashboard** in the Azure portal menu (your default view may already be set to a dashboard rather than the portal home page). From here, you can add up to 100 named dashboards to encapsulate views for specific aspects of your Azure services that you want to track.



You can add a range of tiles and other visualizations to a dashboard, and when viewing metrics for a specific resource in a chart, as described previously, you can add the chart to a new or existing dashboard. In the following image, two charts showing metrics for an AI services resource have been added to a dashboard.



> Note: For more information about dashboards, see Create a dashboard in the Azure portal in the Azure documentation.

Manage diagnostic logging

Diagnostic logging enables you to capture rich operational data for an Azure AI services resource, which can be used to analyze service usage and troubleshoot problems.

(i) Create resources for diagnostic log storage

To capture diagnostic logs for an AI services resource, you need a **destination** for the log data.

You can use **Azure Event Hubs** as a destination in order to then forward the data on to a custom telemetry solution, and you can connect directly to some third-party solutions; but in most cases you'll use one (or both) of the following kinds of resource within your Azure subscription:

- **Azure Log Analytics** - a service that enables you to query and visualize log data within the Azure portal.
- **Azure Storage** - a cloud-based data store that you can use to store log archives (which can be exported for analysis in other tools as needed).

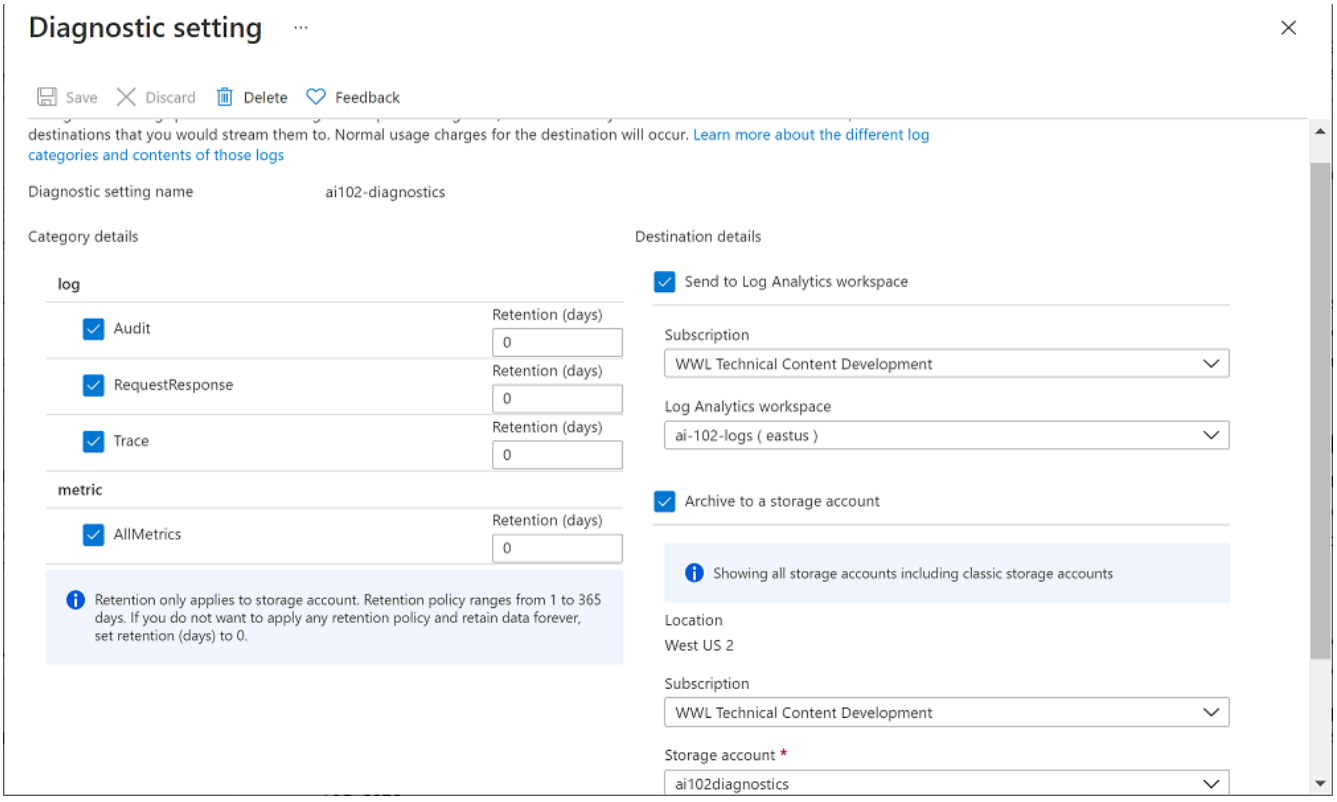
You should create these resources before configuring diagnostic logging for your AI services resource. If you intend to archive log data to Azure Storage, **create the Azure Storage account in the same region as your AI services resource.**

(ii) Configure diagnostic settings

With your log destinations in place, you can configure diagnostic settings for your AI services resource. You define diagnostic settings on the **Diagnostic settings** page of the blade for your AI services resource in the Azure portal. When you add diagnostic settings, you must specify:

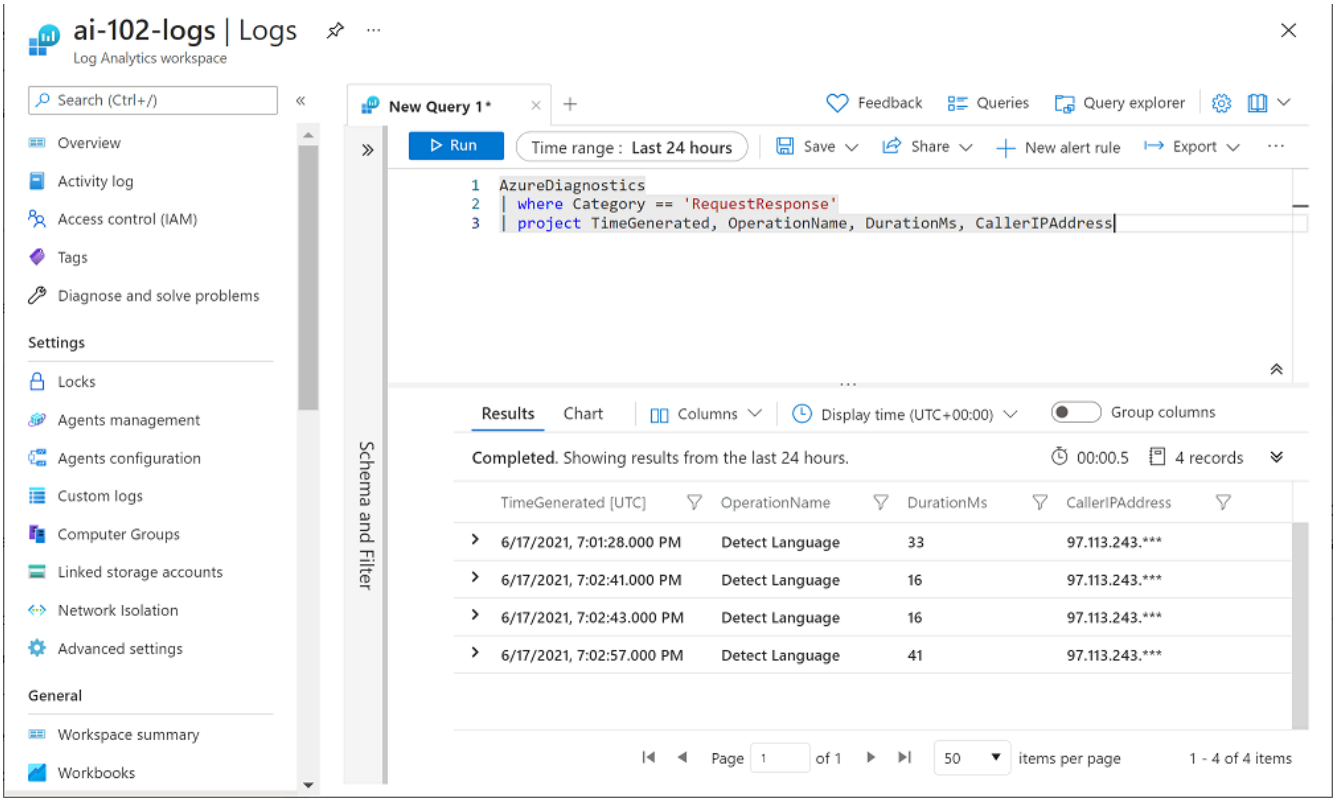
- A name for your diagnostic settings.
- The **categories of log event data that you want to capture.**
- Details of the **destinations in which you want to store the log data.**

In the following example, the diagnostic settings store all available log data and metrics in Azure Log Analytics and Azure Storage.



(iii) View log data in Azure Log Analytics

It can take an hour or more before diagnostic data starts flowing to the destinations, but when the data has been captured, you can view it in your **Azure Log Analytics** resource by running queries, as shown in this example.



> Note: For more information, see Enable diagnostic logging for Azure AI services in the Azure AI services documentation.

Exercise

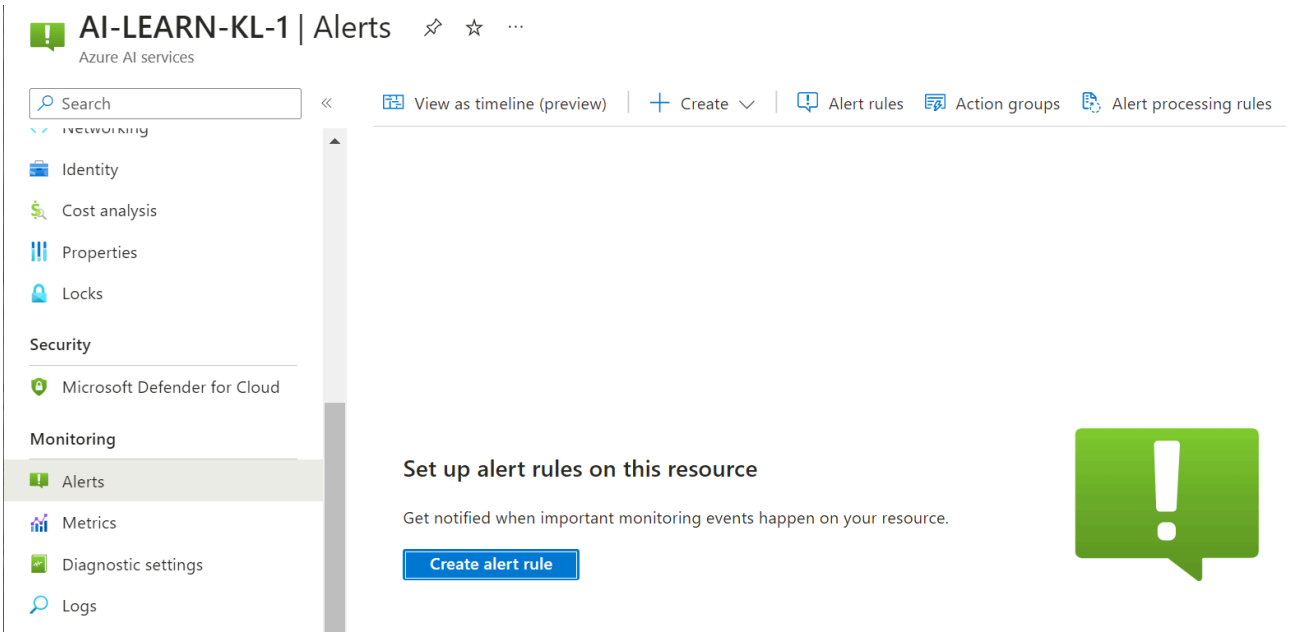
Monitor Azure AI Services

Azure AI Services can be a critical part of an overall application infrastructure. It's important to be able to monitor activity and get alerted to issues that may need attention.

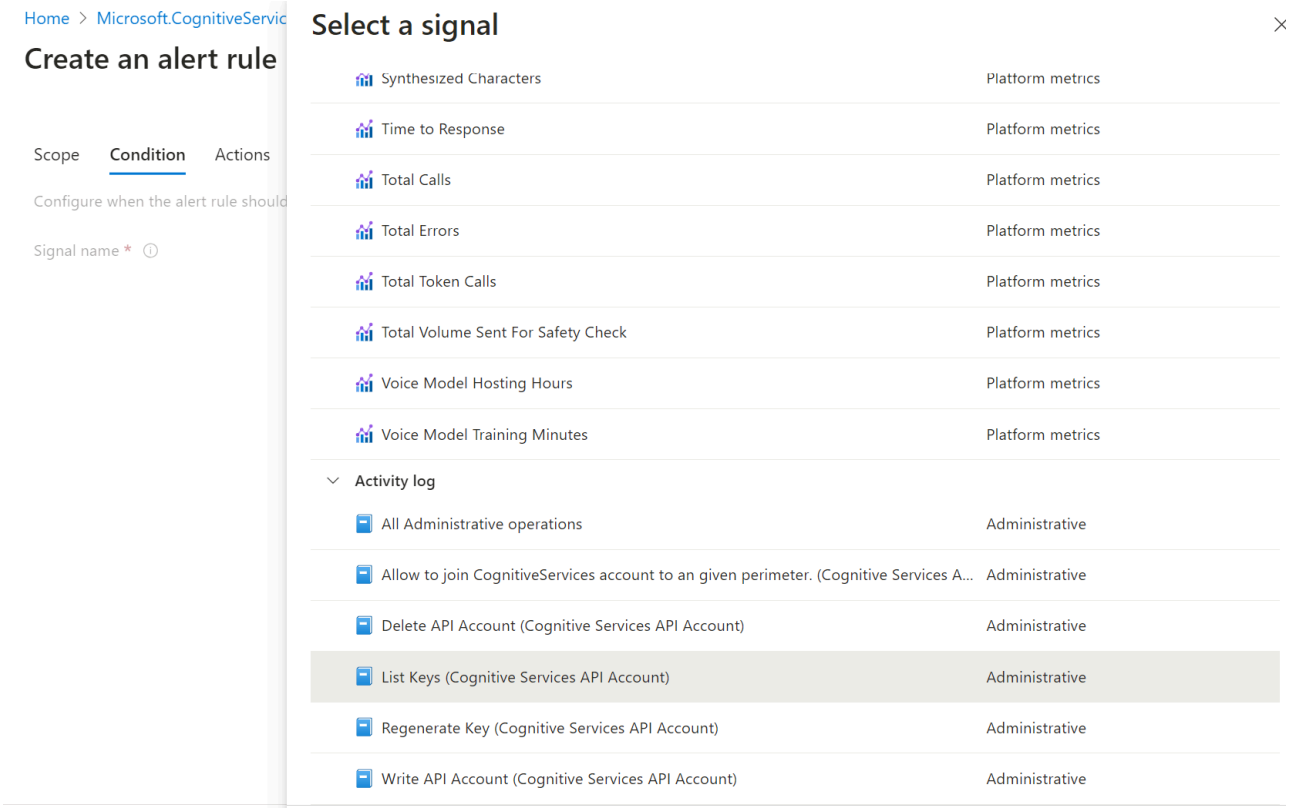
Configure an alert

Let's start monitoring by defining an alert rule so you can detect activity in your Azure AI services resource.

1. In the Azure portal, go to your Azure AI services resource and view its **Alerts** page (in the **Monitoring** section).

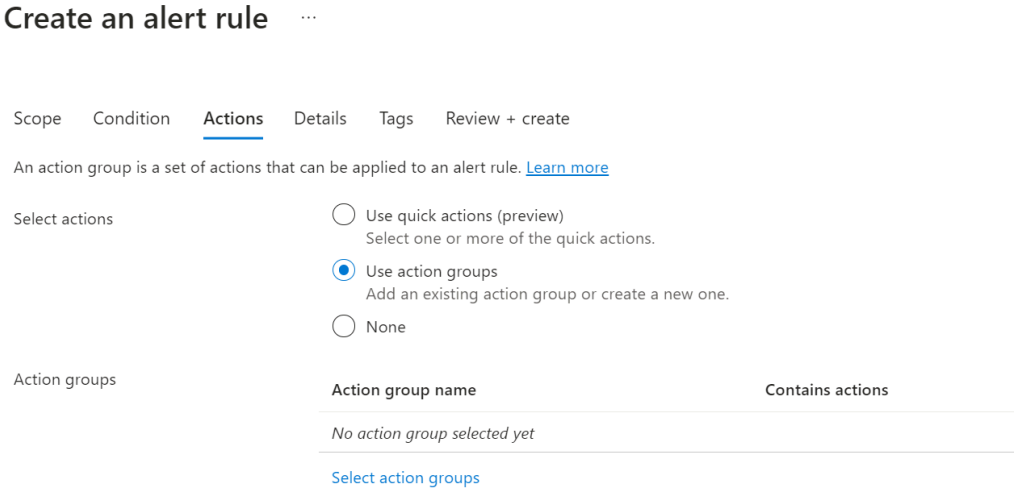


2. Select **+ Create** dropdown, and select **Alert rule**
3. In the **Create an alert rule** page, under **Scope**, verify that the your Azure AI services resource is listed. (Close **Select a signal** pane if open)
4. Select **Condition** tab, and select on the **See all signals** link to show the **Select a signal** pane that appears on the right, where you can select a signal type to monitor.
5. In the **Signal type** list, scroll down to the **Activity Log** section, and then select **List Keys (Cognitive Services API Account)**. Then select **Apply**.

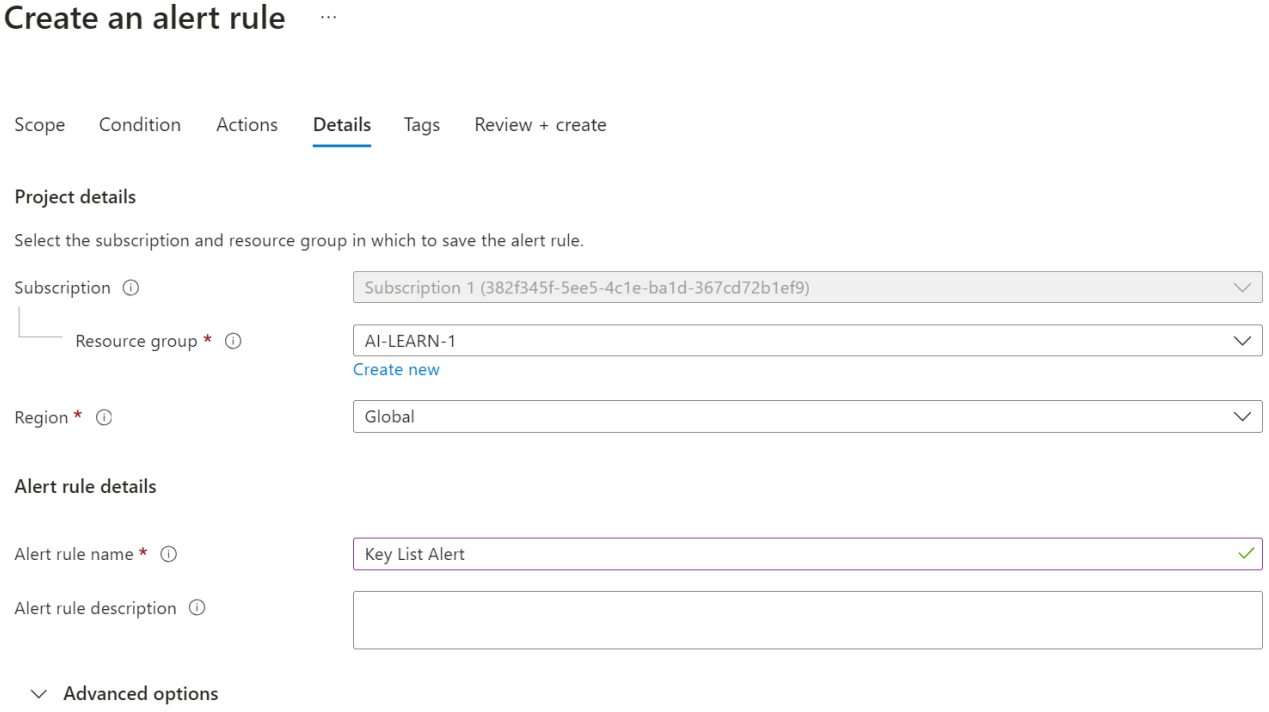


6. Review the activity over the past 6 hours.
7. Select the **Actions** tab. Note that you can specify an *action group*. This enables you to **configure automated actions when an alert is fired** - for example, sending an email notification. We won't do

that in this exercise; but it can be useful to do this in a production environment.



8. In the **Details** tab, set the **Alert rule name** to **Key List Alert**.



9. Select **Review + create**.
10. Review the configuration for the alert. Select **Create** and wait for the alert rule to be created.
11. Now you can use the following command to get the list of Azure AI services keys, replacing `<resourceName>` with the name of your Azure AI services resource, and `<resourceGroup>` with the name of the resource group in which you created it.

```
az cognitiveservices account keys list --name <resourceName> --resource-group
<resourceGroup>
```

The command returns a list of the keys for your Azure AI services resource.

> **Note** If you haven't logged into Azure CLI, you may need to run `az login` before the list keys command will work.

12. Switch back to the browser containing the Azure portal, and refresh your **Alerts page**. You should see a **Sev 4** alert listed in the table (if not, wait up to five minutes and refresh again).

13. Select the alert to see its details.

Visualize a metric

As well as defining alerts, you can view metrics for your Azure AI services resource to monitor its utilization.

1. In the Azure portal, in the page for your Azure AI services resource, select **Metrics** (in the **Monitoring** section).
2. If there is no existing chart, select **+ New chart**. Then in the **Metric** list, review the possible metrics you can visualize and select **Total Calls**.
3. In the **Aggregation** list, select **Count**. This will enable you to monitor the total calls to you Azure AI Service resource; which is useful in determining how much the service is being used over a period of time.
4. To generate some requests to your Azure AI service, you will use **curl** - a command line tool for HTTP requests. In your editor, open **rest-test.cmd** and edit the **curl** command it contains (shown below), replacing `<yourEndpoint>` and `<yourKey>` with your endpoint URI and **Key1** key to use the Text Analytics API in your Azure AI services resource.

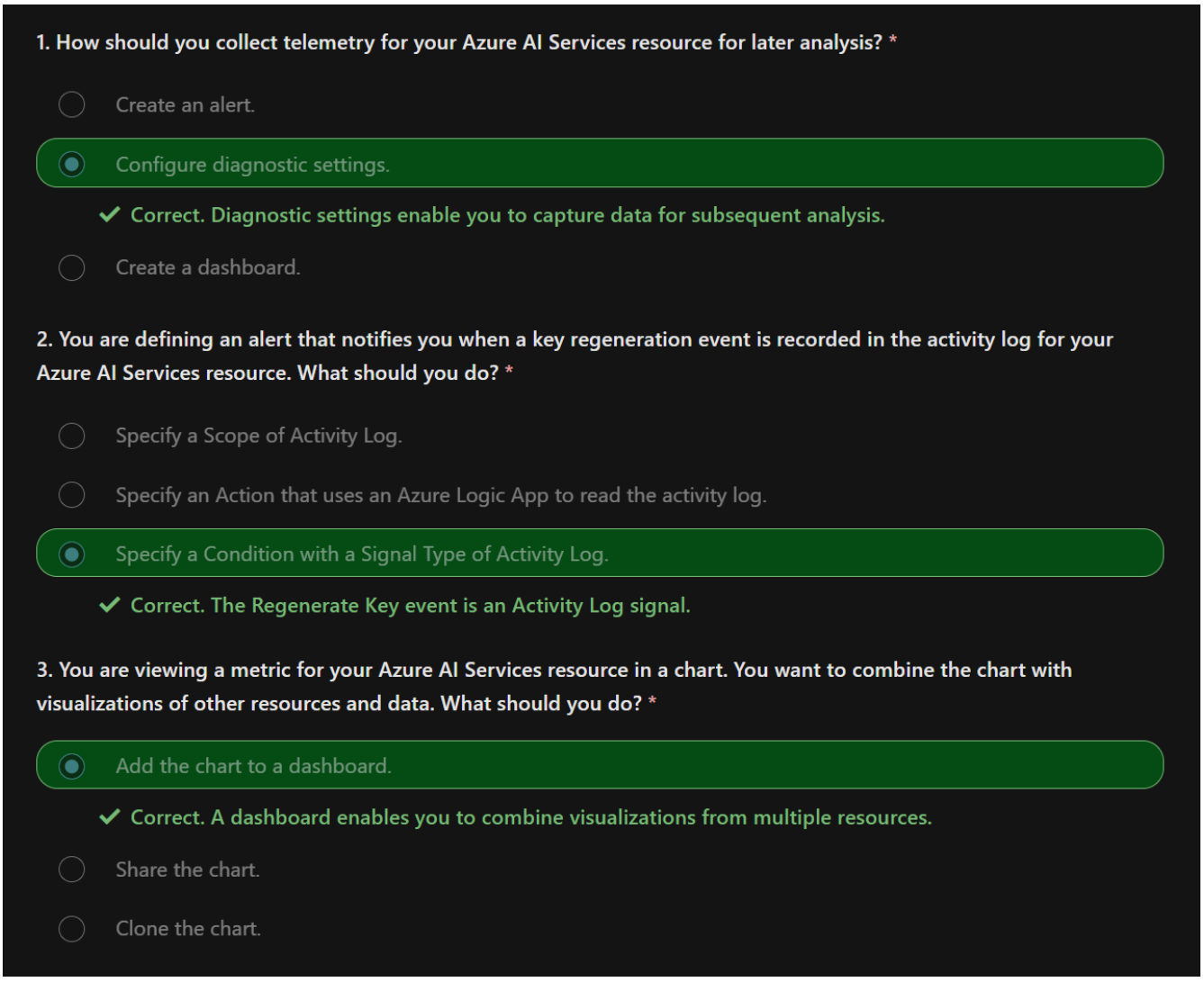
   ```
   curl -X POST "<your-endpoint>/language/:analyze-text?api-version=2023-04-01" -H
   "Content-Type: application/json" -H "Ocp-Apim-Subscription-Key: <your-key>" --data-
   ascii "{'analysisInput':{'documents':[{'id':1,'text':'hello'}]}, 'kind':
   'LanguageDetection'}"
   ```

5. Save your changes and then run the following command:

   ```
   ./rest-test.cmd
   ```

   The command returns a JSON document containing information about the language detected in the input data (which should be English).
6. Re-run the **rest-test** command multiple times to generate some call activity (you can use the **^** key to cycle through previous commands).
7. Return to the **Metrics** page in the Azure portal and refresh the **Total Calls** count chart. It may take a few minutes for the calls you made using *curl* to be reflected in the chart - keep refreshing the chart until it updates to include them.

Knowledge Check



Summary

In this module, you learned how to:

- Monitor Azure AI services costs.
- Create alerts and view metrics for Azure AI services.
- Manage Azure AI services diagnostic logging.

To learn more about Azure AI services, refer to the AI services documentation.

✍️ Compiled by Kenneth Leung (2025)