ScienceLogic

Monitoring vLLM

vLLM Monitoring PowerPack version 100

Table of Contents

Chapter 1

Introduction

Overview

This manual describes how to collect and analyze metrics from vector Language Learning Model (vLLM) deployments in SL1 using the Dynamic Applications in the "vLLM Monitoring" PowerPack.

The following sections provide an overview of vLLM and the "vLLM Monitoring" PowerPack:

This chapter covers the following topics:

NOTE: ScienceLogic provides this documentation for the convenience of ScienceLogic customers. Some of the configuration information contained herein pertains to third-party vendor software that is subject to change without notice to ScienceLogic. ScienceLogic makes every attempt to maintain accurate technical information and cannot be held responsible for defects or changes in third-party vendor software. There is no written or implied guarantee that information contained herein will work for all third-party variants. See the End User License Agreement (EULA) for more information.

What Does the vLLM Monitoring PowerPack Monitor?

To monitor vLLM deployments using SL1, you can install the *vLLM Monitoring*PowerPack. This PowerPack enables you to align Dynamic Applications to vLLM deployments to collect data.

The *vLLM Monitoring*PowerPack includes:

- Dynamic Applications that discover and collect data from vLLM.

Installing the vLLM Monitoring PowerPack

Before completing the steps in this manual, you must import and install the latest version of the *vLLM Monitoring* PowerPack.

TIP: By default, installing a new version of a PowerPack overwrites all content from a previous version of that PowerPack that has already been installed on the target system. You can use the ***Enable Selective PowerPack Field Protection*** setting in the **Behavior Settings** page (System > Settings > Behavior) to prevent new PowerPacks from overwriting local changes for some commonly customized fields. For more information, see the section on *Global Settings*.

IMPORTANT: Ensure that you are running version 12.1.2 or later of SL1 before installing this PowerPack. For details on upgrading SL1, see the relevant *SL1 Platform Release Notes*.

To download and install the PowerPack:

1. Search for and download the PowerPack from the **PowerPacks** page (Product Downloads > PowerPacks & SyncPacks) at the ScienceLogic Support Site.
2. In SL1, go to the **PowerPacks** page (System > Manage > PowerPacks).
3. Click the **[Actions]** button and choose *Import PowerPack*. The **Import PowerPack** dialog box appears.
4. Click **[Browse]** and navigate to the PowerPack file from step 1.
5. Select the PowerPack file and click **[Import]**. The **PowerPack Installer** modal displays a list of the PowerPack contents.
6. Click **[Install]**. The PowerPack is added to the **PowerPacks** page.

NOTE: If you exit the **PowerPack Installer** modal without installing the imported PowerPack, the imported PowerPack will not appear in the **PowerPacks** page. However, the imported PowerPack will appear in the **Imported PowerPacks** modal. This page appears when you click the **[Actions]** menu and select *Install PowerPack*.

Chapter 2

Overview

The following sections describe how to collect and analyze metrics from vLLM (vector Language Learning Model) deployments in SL1 using the "vLLM Monitoring" PowerPack:

This chapter covers the following topics:

Prerequisites for Monitoring vLLM Deployments

To configure the SL1 system to monitor vLLM deployments using the *vLLM Monitoring* PowerPack, you must meet the following requirements and have the following information:

- You must install version 102 of the "Low-code Tools" PowerPack. This will allow you to create a universal credential to use when aligning the Dynamic Applications in this PowerPack to the vLLM device.
- The vLLM metrics endpoint must be enabled and verified to be working. For information about enabling and verifying the vLLM metrics endpoint, see the vLLM documentation on production metrics.
- The exposed `/metrics` endpoints must be reachable by SL1 through one of the normal methods of authentication supported by Low-Code Tools. See https://docs.sciencelogic.com/dev-docs for more information about the supported methods of authentication.
- A device in SL1 with the IP address where the `/metrics` endpoint is exposed. For example, you can have a physical device or virtual device that represents the vLLM model server, as long as it has an IP address and that IP works for `http://<ip>:<port>/metrics` calls.

NOTE: A virtual device can only be used if an IP address is configured on it.

Creating a Credential for vLLM

To configure SL1 to monitor vector Language Learning Model (vLLM) deployments, you must first create a universal type credential. This credential allows the Dynamic Applications in the *vLLM Monitoring* PowerPack to communicate with vLLM deployments.

> **NOTE:** You must install the "Low Code Tools" PowerPack, version 102 or greater to create a universal credential for aligning the Dynamic Applications in this PowerPack to your virtual or physical device.

To configure a universal credential to access a vLLM deployment:

1. Go to the **Credentials** page (Manage > Credentials).

> **NOTE:** To configure the universal credential, you must use the default SL1 user interface, not the classic user interface.

1. Click **[Create New]** and select *Create Low-code tools: rest v102 Credential*. The **Create Credential** modal page appears.
2. Supply values in the following fields:
    - ***Name***. Type a name for your credential.
    - ***All Organizations***. Toggle on (blue) to align the credential to all organizations, or toggle off (gray) and then select one or more specific organizations from the from the ***What organization manages this service?*** drop-down field to align the credential with those specific organizations.
    - ***Authentication Type***. Select the appropriate authentication type. Depending on the authentication type selected, you may need to provide additional information. For more information, see https://docs.sciencelogic.com/dev-docs.
    - ***URL***. Type the URL of your vLLM deployment.
3. Click **[Save & Close]**.

Aligning Dynamic Applications to vLLM Deployments

If you have already discovered the vLLM instance as a physical device, you can align the vLLM Dynamic Applications to that device. If you do not have a physical device for the vLLM instance, you must create a virtual device and then manually align Dynamic Applications to the virtual device.

Manually Aligning vLLM Dynamic Applications to the Physical Device

To manually align the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications to the physical device:

1. Go to the **Devices** page (Devices > Classic Devices, or Registry > Devices > Device Manager in the classic SL1 user interface).
2. Locate your vLLM physical device and click its wrench icon ().
3. In the **Device Investigator**, click the **[Collections]** tab.
4. Click the **[Actions]** button at the top of the page, then click the **[Add Dynamic Application]** button.
5. In the **Align Dynamic Application** modal, locate and select the "vLLM Metrics Config" Dynamic Application.
6. Under ***Credentials***, select the vLLM credential you created and click **[Save]**.
7. Repeat steps 4-6 for the "vLLM Metrics Performance" Dynamic Application.
8. Click **[Save]**.

Creating a vLLM Virtual Device

If you do not have a physical device to align the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications to, you must create a ***virtual device*** that represents the vLLM deployment. A virtual device is a user-defined container that represents a device or service that cannot be discovered by SL1. You can use the virtual device to store information gathered by policies or Dynamic Applications.

If you want to discover more than one vLLM account, you must create a virtual device for each account that you want to use.

To create a virtual device that represents your vLLM deployment:

1. Go to the **Device Manager** page (Devices > Classic Devices, or Registry > Devices > Device Manager in the classic SL1 user interface).
2. Click **[Actions]** and select *Create Virtual Device* from the menu. The **Virtual Device** modal page appears.
3. Enter values in the following fields:
    - ***Device Name***. Enter a name for the device.
    - ***Organization***. Select the organization for this device. The organization you associate with the device limits the users that will be able to view and edit the device. Typically, only members of the organization will be able to view and edit the device.
    - ***Device Class***. Select *Virtual Device | Content Verification*.
    - ***Collector***. Select the collector group that will monitor the device.
4. Click **[Add]** to create the virtual device.
5. Repeat these steps for each vLLM deployment that you want to use.

Manually Aligning the vLMM Dynamic Applications to the Virtual Device

After creating the vLLM virtual device, you must manually align the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications to the virtual device.

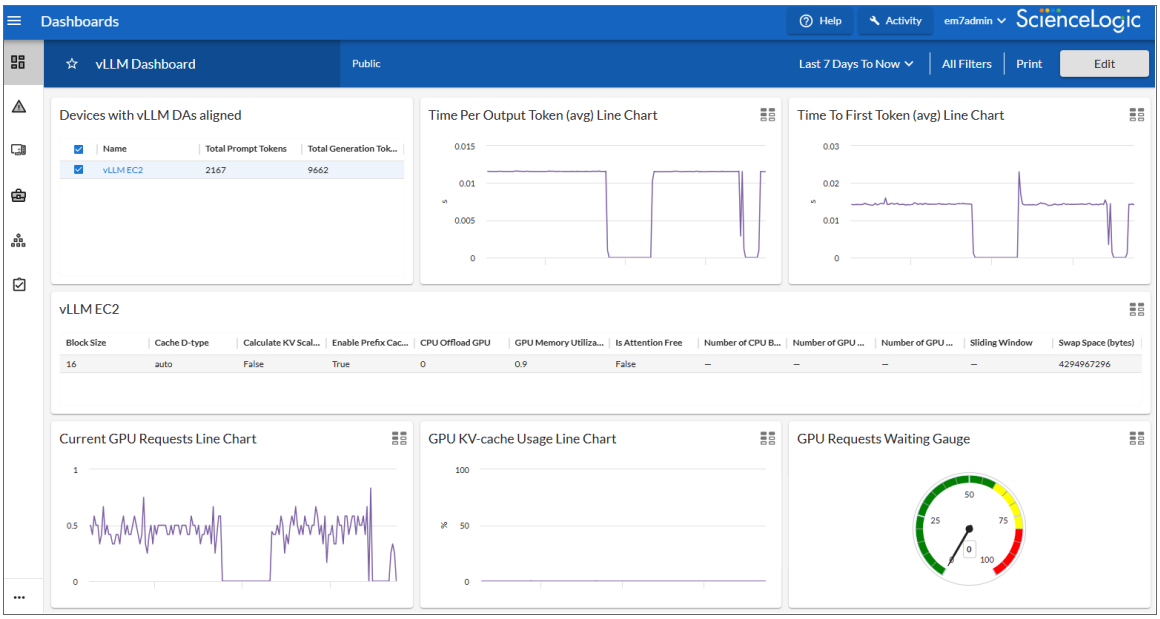
To manually align the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications:

1. Go to the **Devices** page (Devices > Classic Devices, or Registry > Devices > Device Manager in the classic SL1 user interface).
2. Locate your vLLM virtual device and click its wrench icon ().
3. In the **Device Investigator**, click the **[Collections]** tab.
4. Click the **[Actions]** button at the top of the page, then click the **[Add Dynamic Application]** button.
5. In the **Align Dynamic Application** modal, locate and select the "vLLM Metrics Config" Dynamic Application.
6. Under ***Credentials***, select the vLLM credential you created and click **[Save]**.
7. Repeat steps 4-6 for the "vLLM Metrics Performance" Dynamic Application.
8. Click **[Save]**.

vLLM Dashboard

The "vLLM Monitoring" PowerPack includes the "vLLM Dashboard" that you can use to view various metrics for devices aligned to the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications. The "vLLM Dashboard" contains the following widgets:

- ***Devices with vLLM DAs aligned***. Lists the devices that are aligned to the "vLLM Metrics Config" and "vLLM Metrics Performance" Dynamic Applications
- ***Time Per Output Token (avg) Line Chart***. Displays the average time (in seconds) it takes to generate each token of output, providing insight into device processing speed and efficiency
- ***Time to First Token (avg) Line Chart***. Displays the average time in seconds it takes from receiving a request to generating the first token of output, providing an indication of the initial response latency
- ***Current GPU Requests Line Chart***. Displays the number of in-progress requests currently being processed.
- ***GPU KV-cache Usage Line Chart***. Displays the percentage of the GPU KV cache (key-value memory) that is currently in use.
- ***GPU Requests Waiting Gauge***. Displays the number of requests currently waiting in the queue to be processed, indicating the level of demand on the system and potential delays in processing.


Dashboards
Help
Activity
em7admin
ScienceLogic
vLLM Dashboard
Public
Last 7 Days To Now
All Filters
Print
Edit
Devices with vLLM DAs aligned
Name
Total Prompt Tokens
Total Generation Tok...
vLLM EC2
2167
9662
Time Per Output Token (avg) Line Chart
0.015
0.01
0.005
0
Time To First Token (avg) Line Chart
0.03
0.02
0.01
0
vLLM EC2
Block Size
Cache D-type
Calculate KV Scal...
Enable Prefix Cac...
CPU Offload GPU
GPU Memory Utiliza...
Is Attention Free
Number of CPU B...
Number of GPU ...
Number of GPU ...
Sliding Window
Swap Space (bytes)
16
auto
False
True
0
0.9
False
4294967296
Current GPU Requests Line Chart
1
0.5
0
GPU KV-cache Usage Line Chart
100
50
0
GPU Requests Waiting Gauge
50
25
75
0
100

Copyrights and Trademarks

ScienceLogic, the ScienceLogic logo, and EM7 are trademarks of ScienceLogic, Inc. in the United States, other countries, or both.

Below is a list of trademarks and service marks that should be credited to ScienceLogic, Inc. The ® and ™ symbols reflect the trademark registration status in the U.S. Patent and Trademark Office and may not be appropriate for materials to be distributed outside the United States.

- ScienceLogic™
- EM7™ and em7™
- Simplify IT™
- Dynamic Application™
- Relational Infrastructure Management™

The absence of a product or service name, slogan or logo from this list does not constitute a waiver of ScienceLogic's trademark or other intellectual property rights concerning that name, slogan, or logo.

Please note that laws concerning use of trademarks or product names vary by country. Always consult a local attorney for additional guidance.

Other

If any provision of this agreement shall be unlawful, void, or for any reason unenforceable, then that provision shall be deemed severable from this agreement and shall not affect the validity and enforceability of any remaining provisions. This is the entire agreement between the parties relating to the matters contained herein.

In the U.S. and other jurisdictions, trademark owners have a duty to police the use of their marks. Therefore, if you become aware of any improper use of ScienceLogic Trademarks, including infringement or counterfeiting by third parties, report them to Science Logic's legal department immediately. Report as much detail as possible about the misuse, including the name of the party, contact information, and copies or photographs of the potential misuse to: legal@sciencelogic.com. For more information, see https://sciencelogic.com/company/legal.

800-SCI-LOGIC (1-800-724-5644)

International: +1-703-354-1010