Gigamon®

Gigamon Insights User and Deployment Guide

Gigamon Insights

Product Version: 6.12.03

Document Version: 2.0

(See Change Notes for document updates.)

Gigamon Inc.
3300 Olcott Street
Santa Clara, CA 95054
408.831.4000

Change Notes

When a document is updated, the document version number on the cover page will indicate a new version and will provide a link to this Change Notes table, which will describe the updates.

Document Version	Date Updated	Change Notes
2.0	4/30/2026	Reordered topics to align with the user journey, improving the overall flow and logical progression of the document.
1.0	3/16/2026	The original release of this document with 6.12.03 LA.

Contents

Gigamon Insights

NOTE: Gigamon Insights is released as a Limited Availability (LA) feature, providing you with the opportunity to evaluate its capabilities before its General Availability (GA) release. For access to LA software please contact your Gigamon account team or file a support case and ensure you have reviewed the LA Terms and Conditions. You can file support cases with Gigamon support during the LA period and provide feedback to Gigamon on your experience with Gigamon Insights.

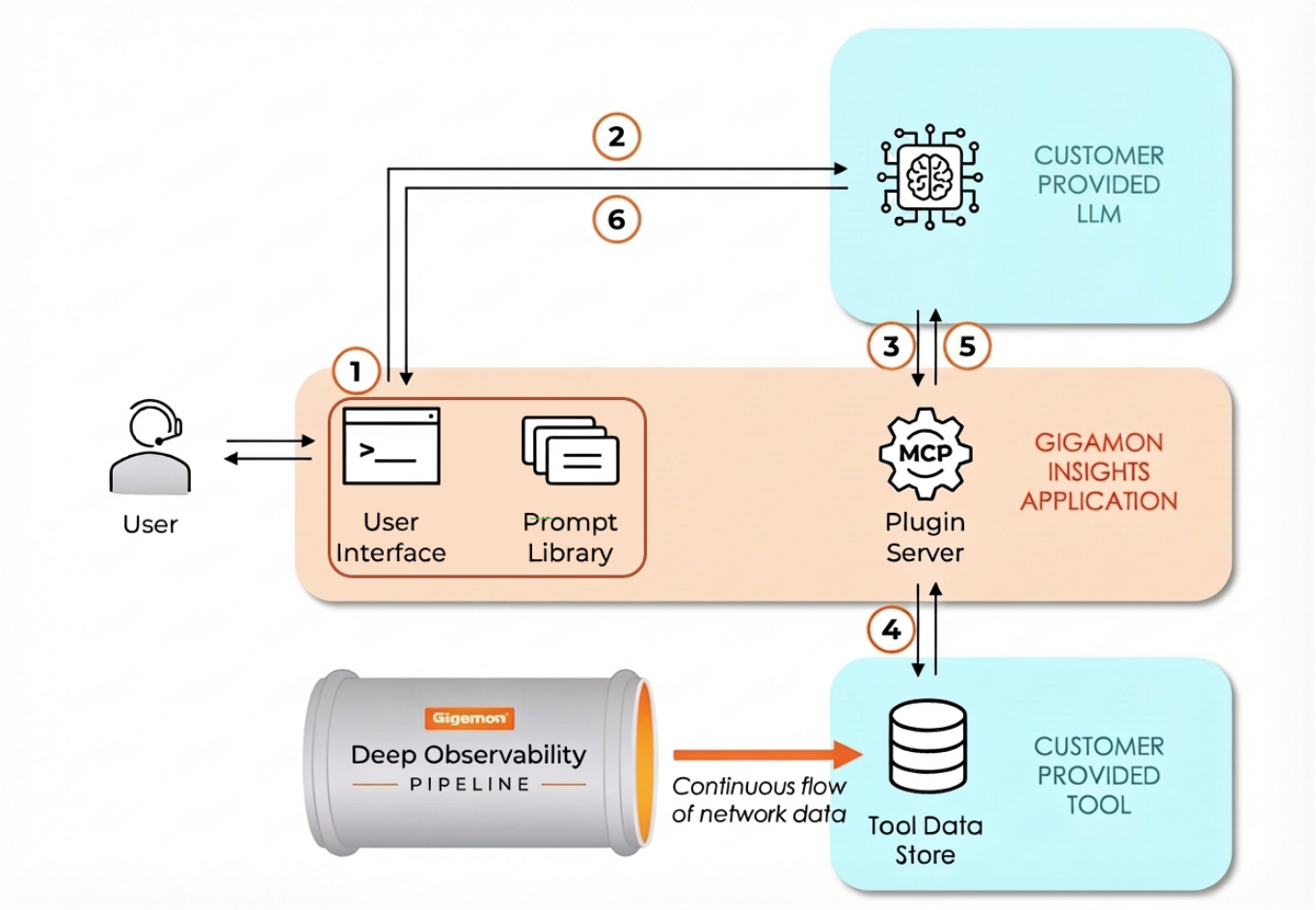
Gigamon Insights is a Generative and Agentic AI application that provides immediate, actionable answers to technical and business questions using trusted network packet data, enriched with additional context. Instead of relying on static dashboards and manual queries, you can ask questions in plain language and receive deep, context-driven responses. It uses:

- Gigamon Application Metadata Intelligence (AMI) and AMX metadata as its primary data source.
- A curated prompt library for Security, Network and Application Insights.
- Model Context Protocol (MCP) integrations with customer tools—Elastic and Splunk.
- Large Language Models (LLMs) (Anthropic Claude Sonnet and Google Gemini) configured through GigaVUE-FM.

What Gigamon Insights Can Do

- **Ask questions using natural language prompts**: Ask questions in free-form, conversational language or use predefined prompts that follow best practices and are available globally to help teams work consistently across the organization.
- **Create and save your own prompts**: Create custom prompts tailored to your specific tasks and save them for later use. Your saved prompts are private to you and are not available globally, so you can experiment and refine prompts without affecting shared content.
- **Upload a dashboard screenshot for context**: Upload one dashboard screenshot per prompt. The screenshot provides visual context that aligns with what you see in your monitoring tools, helping the system generate more relevant responses.
- **Capture and export feedback**: Provide feedback using thumbs-up or thumbs-down ratings and optional comments. Export responses as PDF or Word documents for sharing, record-keeping, and audit purposes.

How Gigamon Insights Works



1. Initial Prompt: User submits prompt, either freeform or a pre-defined prompt from the library. Passed to LLM.
2. Prompt Expansion: LLM expands the human prompt with more context and generates multiple data queries behind the scenes.
3. Agentic Calls: data queries submitted via the MCP protocol.
4. Data Queries: MCP Server queries the data store for relevant data.
5. Iterative Retrieval: LLM processes the returned data and, optionally, repeats agentic calls multiple times to access additional data and refine insights.
6. Insight: LLM returns an insightful response.

Deployment Overview

This topic gives you an at-a-glance view of the end-to-end Gigamon Insights deployment. Use it to understand the order of tasks and to jump to the detailed instructions for each phase.

The deployment includes four phases:

Plan your deployment

Before you begin, confirm that your environment meets the supportability requirements. At a high level, you need:

- A supported version of GigaVUE-FM and a supported deployment platform (VMware ESXi or AWS).
- A supported large language model (LLM) provider — Amazon Bedrock and Google Vertex AI.
- A supported MCP tool — Elastic or Splunk.
- A supported web browser to access the Gigamon Insights interface.
- The required firewall ports must be opened between GigaVUE-FM, Gigamon Insights, and the external LLM and MCP endpoints.

For the full list of supported versions, system requirements, open ports, browsers, and rules, refer to Supportability.

Prerequisites

Before you deploy Gigamon Insights, complete the following tasks in GigaVUE-FM:

- Create a user and assign the Gigamon Insights user group. Gigamon Insights uses this user account to authenticate the Gigamon Insights with GigaVUE-FM during deployment.
- Generate an API token for the user. Enter this token when you deploy the Gigamon Insights instance so it can register with GigaVUE-FM.

For instructions, refer to Prerequisites for Gigamon Insights Deployment

Deploy Gigamon Insights

Gigamon Insights is delivered as a virtual appliance. You can deploy it on either of the following platforms:

- VMware ESXi — Deploy the OVA file using the vSphere Client on a vCenter-managed host. For instructions, refer to Deploy Gigamon Insights on VMware vCenter.
- AWS — Subscribe to the Gigamon Insights offer in AWS Marketplace and launch the instance in your VPC. For instructions, refer to Deploy Gigamon Insights on AWS.

During deployment, you provide the GigaVUE-FM IP address and the API token so the appliance automatically registers with GigaVUE-FM.

Post-deployment configurations

After Gigamon Insights is deployed and registered, configure it in GigaVUE-FM so it can process prompts and return answers:

- Acquire LLM credentials for Amazon Bedrock and Google Vertex. Gigamon Insights uses these models to interpret prompts and generate responses.
- Acquire MCP server credentials for either Elastic or Splunk. The MCP server lets Gigamon Insights query your data store for the AMI and AMX metadata it needs. You can use only one MCP server at a time.
- Enable Gigamon Insights in GigaVUE-FM by uploading the YAML file with your LLM and MCP details and saving the configuration.

For instructions, refer to Gigamon Insights Post Deployment Configurations in GigaVUE-FM.

Supportability

This section describes the supported versions, third-party platforms, and hardware requirements for deploying Gigamon Insights.

Supported Gigamon Software Versions

Gigamon Insights is supported starting with version 6.12.03.

Supported third party platforms

Gigamon Insights supports the following platforms:

- VMware ESXi
- AWS

System/Infrastructure Requirements for VMware ESXi

The following table outlines the system/infrastructure requirements for deploying Gigamon Insights with VMware ESXi.

Table 1: System/Infrastructure Requirements for VMware ESXi

System/Infrastructure Requirements	
VMware Hypervisor	vSphere ESXi: v8.xx and above.
CPU	4 vCPUs
RAM	16GB
Disk Space	144GB
Network	At least one 1Gb NIC

The following table lists the supported VMware ESXi hypervisor versions for Gigamon Insights.

	vCenter Server	ESXi	GigaVUE-FM
v7.0	v7.0U3	v7.0U3	v6.12.00
v8.0	v8.0U2, v8.0U3	v8.0U2, v8.0U3	v6.12.00

Open Ports Access Requirements

NOTE: To ensure the Gigamon Insights feature functions correctly, you must establish network connectivity between GigaVUE-FM and Gigamon Insights by opening the required firewall ports listed below.

The following table describes the open ports access requirements.

Table 2: Open Ports Access Requirements

Direction	Protocol	Port Number	Service	Source CIDR	Destination	Purpose
Inbound	TCP	22	SSH	Administrator Subnet	Gigamon Insights	Allows CLI access to user-initiated management and diagnostics.
Inbound	TCP	443	HTTPS	• GigaVUE-FM • End Users	Gigamon Insights	Allows GigaVUE-FM and end users to reach Gigamon Insights to explore AMI data using prompts.
Inbound	TCP	80	HTTP	GigaVUE-FM	Gigamon Insights	Allow Gigamon Insights certificate installation.
Outbound	TCP	443	HTTPS (GigaVUE-FM, external LLM endpoints)	Gigamon Insights	LLM Service Providers (Amazon Bedrock/Google Vertex AI) **Amazon Bedrock**: https://docs.aws.amazon.com/general/latest/gr/bedrock.html **Google Vertex AI**: https://docs.cloud.google.com/ve	Allows Gigamon Insights to reach GigaVUE-FM to provide contextual

Direction	Protocol	Port Number	Service	Source CIDR	Destination	Purpose
					rtex-ai/docs/reference/rest	information for queries and to communicate with external LLM providers over HTTPS only.
Outbound	TCP	Customer-defined HTTPS port (for example, 443)	HTTPS (external MCP server)	Gigamon Insights	External MCP Servers (Splunk/Elastic)	Allows Gigamon Insights to communicate with an external MCP server over HTTPS; the port is determined by the customer's MCP server configuration.
Outbound	TCP	443	HTTPS	Gigamon Insights	**Azure**: • login.microsoftonline.com • scp1adls1.blob.core.windows.net	Allows Gigamon Insights to export the telemetry data.
Outbound	TCP	9600	Certificate provider in GigaVUE-FM	Gigamon Insights	GigaVUE-FM	Allows Gigamon Insights to communicate with GigaVUE-FM for certificate request and renewal.

Supported LLM Providers

The following models are supported:

- Anthropic Claude Sonnet 4.5 when accessed via Amazon Bedrock.

- Google Gemini 3.1 Pro when accessed via Google Vertex.

Supported MCP tools

- **Elastic**: versions 8.x and 9.x, through an Elastic MCP server. For the most recent remote MCP server implementation, Elastic version 9.2.x is recommended.
- **Splunk**: versions 8.x, 9.x, and 10.x, through a Splunk MCP server installed in the customer's Splunk deployment.

Supported Browsers

Gigamon Insights supports the following web browsers:

OS	Browser	Browser Version
Mac OSX	Google® Chrome®	v 141 onwards
	Apple® Safari®	v26.1 onwards
Linux	Google® Chrome®	v141 onwards

Rules and Notes

Keep the following guidelines when you use Gigamon Insights:

- You can connect one MCP server to a Gigamon Insights instance at a time.
- Gigamon Insights supports up to three concurrent user sessions.

What to do Next

Refer to Prerequisites for Gigamon Insights Deployment

Prerequisites for Gigamon Insights Deployment

Ensure to complete the following tasks in GigaVUE-FM before you deploy Gigamon Insights:

- Create a User and Assign to Gigamon Insights User Group
- Generate an API Token

Create a User and Assign to Gigamon Insights User Group

You must add the GigaVUE-FM user to the **Gigamon Insight Group** to access Gigamon Insights.

If you have not created a GigaVUE-FM user, follow the instructions below:

1. In GigaVUE-FM, go to **Settings** and select **Authentication** > **GigaVUE-FM User Management** > **Users**.

2. On the User page, select **New User**.
3. In the Add User page, enter the following details:
    - **Name**: Actual name of the user
    - **Username**: User name configured in GigaVUE-FM
    - **Password/Confirm Password**: Password for the user.
    - **Email**: Email ID of the user
    - **User Group**: Select the **Gigamon Insight Group** to associate the user. Ensure that the user is assigned to this group.

4. Click **Ok**. The new user is added.

Generate an API Token

Gigamon Insights uses an API token to authenticate the Gigamon Insights appliance with GigaVUE-FM. You provide this token during deployment so the appliance can register with GigaVUE-FM, fetch its configuration, and exchange contextual data for prompts.

You must generate the token by signing in to GigaVUE-FM as one of the following users:

- A user with write permissions for Third Party Orchestration
- A user assigned to **Super Admin Group** user group.

The token inherits the permissions of that user, so it grants only the access that Gigamon Insights needs.

> **NOTE:** We recommend creating a custom user role with write access to Third Party Orchestration instead of using a user who is assigned to **Super Admin Group** user group.

To create a custom role with write access to Third Party Orchestration:

1. In GigaVUE-FM go to **Settings** > **Authentication** > **GigaVUE-FM User Management** > **Roles**.
2. Click **New Role**. In the New Role page, specify the following:
    - Role Name – Enter a name for the role
    - Description – Enter a description
    - Select Permissions – From the Select Permissions tab:
        - Select **Third Party Orchestration**
        - Assign **Write** permission
3. Click **Apply**. Assign this role to the user who will generate the API token.

To generate the API token in GigaVUE-FM:

1. Log in to GigaVUE-FM as a user with the required custom role or as a user assigned to **Super Admin Group** user group.
2. In the left pane, go to **Authentication** > **GigaVUE-FM** > **User Management**.

3. On the User Management page, click the **Tokens** drop-down menu and choose **Current User Tokens** or generate a new token for the user group.
4. Copy the generated token and use it when deploying the instance.

What to do next

Refer to the following topics based on the platform on which you want to deploy Gigamon Insights:

- Deploy Gigamon Insights on VMware vCenter
- Deploy Gigamon Insights on AWS

Deploy Gigamon Insights on VMware vCenter

This section provides the steps to deploy Gigamon Insights on VMware vCenter.

Before you begin:

- Ensure that you have installed VMware vSphere Standard, Enterprise, or Enterprise Plus on compatible hardware. Refer to Hardware Requirements for the minimum hardware requirements and Open Ports Access Requirements for the required port numbers.
- Make sure the VMware ESXi host and the Gigamon Insights VM are time-synchronized. If the ESXi host time is incorrect, the Gigamon Insights VM inherits it, which can cause TLS or signature errors when Gigamon Insights connects to services such as Amazon Bedrock and prevent successful registration in GigaVUE-FM.

The Gigamon Insights software package for vSphere is distributed as an OVA file. You can download the "gigamon-insights-6.12.03.ova" file from the VÜE Community.

Use the vSphere Client to install the Gigamon Insights OVA file. You cannot deploy Gigamon Insights directly from the ESXi host. You must login to the vCenter on the vSphere client to deploy a Gigamon Insights instance.

IMPORTANT NOTE:

The OVA file must be stored in a location that is accessible to the vSphere Client.

To deploy a Gigamon Insights using OVA file:

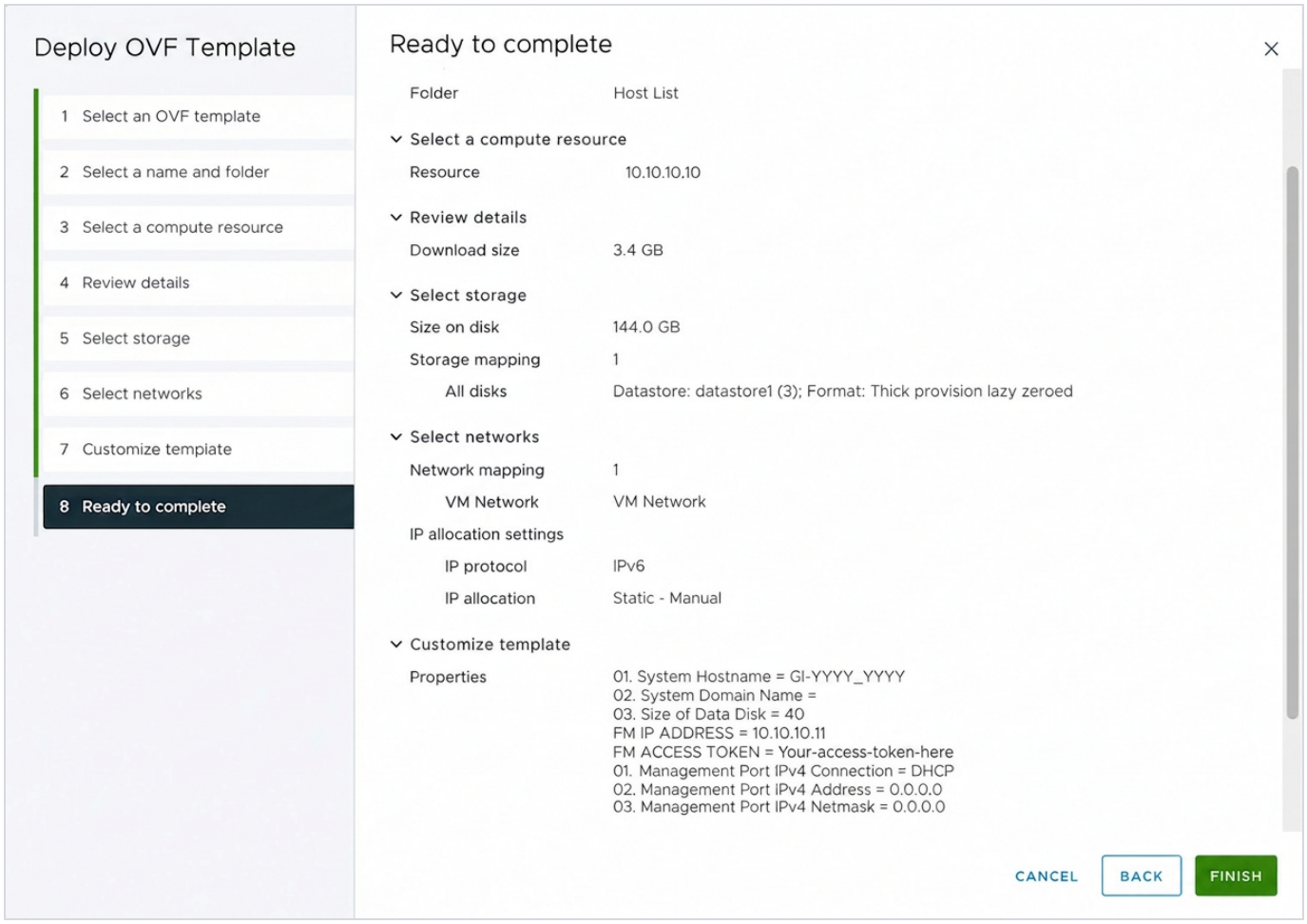
1. Log in to the VMware vCenter web interface.
2. In the vSphere Client, select an inventory object that is a valid parent object of a virtual machine, such as a data center, cluster, or ESXi host.
3. Right-click the ESXi Host, Cluster, or data center on which you want to deploy Gigamon Insights, and then select **Deploy OVF Template**.

   The **Deploy OVF Template** wizard opens.
4. In the **Select OVF Template**, choose one of the following options:

- **URL**—Enter the URL from where you want to download and install the OVF package.
- **Local file**—Click **Browse** to navigate to the OVA file available on your local machine, and then select the OVA file.

5. Click **Next**. The **Select a name and folder** page of the **Deploy OVF Template** wizard appears. Specify a unique name for the Gigamon Insights instance (for example, Insights-Prod-1), and then select a location and host to which you want to deploy the Gigamon Insights instance.
6. Click **Next**. The **Select a compute resource** page of the **Deploy OVF Template** wizard appears. Select a destination compute host for the OVF deployment. The Deploy OVF Template wizard performs validation to ensure that the selected host has all the required resources for the Gigamon Insights deployment.
7. Click **Next**. The **Review details** page of the **Deploy OVF Template** wizard appears. Verify the OVF template details.
8. Click **Next**. The **Select storage** page of the **Deploy OVF Template** wizard appears. In this page:
   a. From the **Select virtual disk format** drop-down list, select the format for the virtual disks and provisioning.
   b. Select the datastore where the virtual machine files will be stored.
9. Click **Next**. The **Select networks** page of the **Deploy OVF Template** wizard appears.
   a. Select a network that provides the necessary connectivity for the Gigamon Insights VM, including management access and any required data-plane communication.
   b. Ensure the IP Protocol drop-down remains at its default value. Altering this setting may result in vCenter-related issues.
10. Click **Next**. The **Customize template** page of the **Deploy OVF Template** wizard appears. Enter the following mandatory deployment properties:
    a. FM HOST ADDRESS - IP address or FQDN of the GigaVUE-FM instance.
    b. FM ACCESS TOKEN - Access token generated in GigaVUE-FM.

    Gigamon Insights is registered with these fields during deployment.
11. Click **Next**. The **Ready to Complete** page of the **Deploy OVF Template** wizard appears.
    a. Verify that all the settings are correct, and then click **Finish**.
    b. Monitor the deployment progress in the **Recent Tasks** pane on the vSphere Client Home page. After the operation completes, the Gigamon Insights instance is successfully deployed..



12. Go to **Actions** > **Power** > **Power on** the VM. Wait for the VM to complete its boot process.
13. You need to SSH into the VM. For first time log in, use the following default credentials:
    - Username: admin
    - Password: gigamon123A!!

    NOTE: For security reasons, you are prompted to change the password immediately after your first login. We strongly recommend that you change the default password to secure the appliance.
14. Verify that Gigamon Insights is successfully deployed by navigating in **GigaVUE-FM** > **Settings** > **System** > **Gigamon Insights**.

What to do next

Refer to Gigamon Insights Post Deployment Configurations in GigaVUE-FM.

Deploy Gigamon Insights on AWS

This section provides the steps for deploying Gigamon Insights on AWS.

To subscribe to Gigamon Insights, perform the following steps:

1. Log in to your AWS account.
2. Go to AWS Marketplace: https://aws.amazon.com/marketplace.
3. In the AWS Marketplace console, go to **AWS Marketplace** > **Discover products**.
4. In the **Search** field, type "Gigamon" and select Gigamon Insights product.
5. Select **View Purchase Options**. The terms and conditions page is displayed.
6. Review the Terms and Conditions and select **Accept Terms**.
7. Review the summary and then select **Launch your software**.
8. In the **Launch Gigamon Insights** page, enter the following details for your deployment:
   a. Select **Launch method** as Launch from EC2 console.
   b. Auto-populates the latest version in the **Software Version** field.
   c. Choose your deployment **Region**.
   d. Click **Launch from EC2** to launch.
9. In the **Launch an instance** page, perform the following:
   a. From the **EC2 Instance Type** drop-down list, select the **m5.xlarge** instance type. Refer to the Recommended and Supported Instance Types for AWS.
   b. Select your preferred **Key Pair** for secure access to the instance.
   c. From the **VPC Settings** drop-down list, select the VPC for deploying Gigamon Insights.
   d. [Optional] Enable the Auto-assign Public IP if you plan to access Gigamon Insights over the internet.
   e. In the **Subnet Settings**, select your desired Subnet.
   f. In the Configure Storage, ensure that it is:
      - 100 GiB gp2 Root volume, Not encrypted
      - 40Gib gp2 EBS volume, Not encrypted
   g. In the **Security Group Settings**, configure the security group to match your access and permissions needs. For details, refer to Security Group.

      In Inbound security group rules, add the following:
      - **SSH access**: Select **Type** as SSH to allow traffic using SSH. This is required to generate the API key.

- **HTTPS access**: Select **Type** as **HTTPS**, Port range: 443 and Source as Anywhere. This is to allow HTTPS traffic from the internet which is used for communication between GigaVUE-FM and Gigamon Insights).
- **HTTP access**: Select **Type** as **HTTP**, Port Range as 80 and Source as Anywhere. GigaVUE-FM sends http request to Gigamon Insights for verification.

h. In the **user data**, you can upload the YAML file or copy the following format to create your own YAML file:

```
#cloud-config
write_files:
 - path: /etc/fm.conf
  permissions: '0600'
  owner: root:root
  content: |
      FM_IP=<IP address of the GigaVUE-FM>
      FM_BEARER_TOKEN=<API token generated in GigaVUE-FM>
```

i. Enter the following mandatory deployment properties:

- FM_IP: IP address or FQDN of the GigaVUE-FM instance.
- FM_BEARER_TOKEN: GigaVUE-FM API token for the user.

j. Upload the YAML file and the YAML template populate values in the User Defaults.

k. Review the pre-populated default values for Gigamon Insights configuration.

l. If there are multiple LLM models defined in the YAML file, use the dropdown to select the appropriate service or model.

m. Click **Save** to apply the configuration.

What to do next

Refer to Gigamon Insights Post Deployment Configurations in GigaVUE-FM.

Gigamon Insights Post Deployment Configurations in GigaVUE-FM

After you deploy Gigamon Insights, you must complete a few setups in GigaVUE-FM before you can start using it. These steps connect Gigamon Insights to a large language model (LLM) and to your data tool. Without them, Gigamon Insights cannot read your prompts or return answers.

You must complete the following tasks:

- **Acquire credentials from a supported LLM provider**. You need credentials for AWS Bedrock and Google Vertex. Gigamon Insights uses these models to read your prompts in plain language, expand them with context, and turn them into clear, useful answers. You must set up both providers so that you can switch between them at any time from the Gigamon Insights interface.
- **Set up an MCP server on either Elastic or Splunk**. The Model Context Protocol (MCP) server lets Gigamon Insights query your data store for the AMI and AMX metadata it needs to answer your questions. You can use only one MCP server at a time, so pick the tool where your metadata is stored.

You add the LLM credentials and MCP server details to the YAML file that GigaVUE-FM uses to enable Gigamon Insights. After you finish these tasks, you can enable Gigamon Insights in GigaVUE-FM and start asking questions.

Refer to the following sections:

Acquire Amazon Bedrock Credentials

To acquire Amazon Bedrock credentials, perform the following steps:

Complete Anthropic (FTU) Form

If this is the first time your AWS Account or organization has used an Anthropic LLM you will need to complete the First Time Use form.

1. Navigate to **Bedrock** > **Model Catalog** > **Claude Sonnet 4.x** .
2. Click **Open in playground**.
3. If needed you will be prompted to complete the FTU form, otherwise you can continue configuring LLM access for Gigamon Insights.

Create IAM Policy

1. Log in to https://aws.amazon.com/console.
2. Navigate to IAM > Access Management > Policies and click **Create policy**.
3. Select an AWS service: **Bedrock**.
4. Filter Actions allowed: **Invoke**.
5. Select **InvokeModel** and **InvokeModelWithResponseStream**.
6. Under Resources, add ARNs to grant access to both the foundation model and the inference profile.
7. Click Add ARNs and add the foundation-model ARN:
    - Resource Region: Select Region (or any)

- Resource resource: **anthropic.claude-sonnet-4-5-20250929-v1:0 (or any other Sonnet 4.x model)**

8. To grant inference profile access, click **Add permissions**.
9. Select **Bedrock** as the AWS service.
10. Filter Actions by **Invoke**, and select: **InvokeModel** and **InvokeModelWithResponseStream**
    a. Under Resources, add the inference-profile ARN.
       - Resource Region: Select Region (or any)
       - Resource resource: **us.anthropic.claude-sonnet-4-5-20250929-v1:0**
    b. Go to Request conditions > **Add another condition** for the inference-profile ARN.
       - Condition key: **aws:ResourceAccount**
       - Operator: **StringEquals**
       - Value: **${aws:PrincipalAccount}**
    c. Then click **Add Condition**, followed by **Next**
11. Under Policy details, provide a Policy name: insights-llm-access (example).
12. Review all permissions and click **Create Policy** to save your new policy.

Your IAM policy is created and ready to be attached to the IAM user.

The following is an example JSON template after the required permissions are configured.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AllowInferenceProfileAccess",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel",
                "bedrock:InvokeModelWithResponseStream"
            ],
            "Resource": [
                "arn:aws:bedrock:*:*:inference-profile/us.anthropic.claude-sonnet-4-5-20250929-
v1:0"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount": "${aws:PrincipalAccount}"
                }
            }
        },
        {
            "Sid": "AllowFoundationModelAccess",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel",
                "bedrock:InvokeModelWithResponseStream"
            ],
```

```
            "Resource": [
                "arn:aws:bedrock:*::foundation-model/anthropic.claude-sonnet-4-5-20250929-v1:0"
            ]
        }
    ]
}
```

Create IAM User

1. In the search bar of the AWS Management Console, type IAM, and then select IAM (Identity and Access Management).
2. In the left navigation pane, under Access Management, click **Create Users**.
3. Enter the following information:
   a. User name: insights-llm-user (example)
   b. Under Permission options, select **Attach policies directly**. In the search bar, type the name of your customer-managed policy (the one you created earlier).
   c. Select the policy check box.
   d. Select permissions: insights-llm-access and AmazonBedrockReadOnly .

   NOTE: Use **Add Permissions** to add the AmazonBedrockReadOnly permission.
4. Click **Next**, then **Create User**. Search for user and click on the returned name.
   a. Under Summary, click on **Create access key**.
   b. For Use case select: **Other**.
   c. Click **Next**. Provide a Description tag value: insights-key.
   d. Click **Create access key**, your access key and secret access key is generated. Make a note of these access keys to add them in the YAML configuration file to enable this LLM for Gigamon Insights.

Acquire Google Vertex Credentials

To acquire Google Vertex credentials, perform the following steps:

Verify Required IAM Roles

1. In the Google Cloud Console, go to **IAM & Admin** > **IAM**.
2. In the IAM page, locate the user account:
   - Use the Filter field to search by the user's email address.
   - Select the user entry to view assigned roles.
3. Verify that the following roles are assigned:
   - Service Account Admin: Ensures the user can create, delete, and manage service accounts.

- Project IAM Admin: Ensures the user can assign roles (for example, granting Vertex AI access to service accounts).
- Service Account Key Admin: Ensures the user can create and manage service account keys.

Verify Organization Policy Setting

1. In the Google Cloud Console, go to: **IAM & Admin** > **Organization Policies**.
2. In the Policies list, use the filter field and search for: **iam.managed.disableServiceAccountApiKeyCreation**
3. Select the policy from the list and check the policy status. Ensure the policy is **Inactive**.

If it is active, follow these steps:

1. Click **Actions** > **Edit policy**. Under **Policy source**, select **Override parent's policy**, then click **Add a rule**.
2. Under **Enforcement**, select **Off**. Click **Done**
3. Click **Set policy**, then the dialog that pops up, click **Set policy again**. Gigamon Insights is now enabled to create API key in Google Vertex.

Create Gigamon Insights service account

1. Go to **IAM & Admin** > **Service Accounts**.
2. Click **Create service account**. Configure the service account as follows:
   a. **Service account name**: gi-vertexai-access
   b. **Service account ID**: gi-vertexai-access
3. Click **Create and continue**.
4. Under **Permissions**, click **Select a role** and choose **Agent Platform User** from the menu.
5. Click **Continue** followed by **Done**.

Create Gigamon Insights Service Account API Key File

1. In the Google Cloud Console, go to: **IAM & Admin** > **Service Accounts**.
2. Click on the service account that you created.
3. In the Keys tab, click **Add key**, and then select **Create new key**.
4. In the Create private key dialog, select **JSON** as the key type.
5. Click **Create**. The key file is automatically downloaded to your local system.

Open the downloaded JSON file and copy its contents. Use these details when configuring the Gigamon Insights using the YAML file.

Set up MCP Server

Configure either the Splunk MCP server or the Elastic MCP server based on your deployment, and skip the section that does not apply. For detailed steps to obtain the required endpoint and credentials, refer to the:

- Acquire Splunk MCP Server Endpoint and API Key
- Acquire Elastic Endpoint and API Key

Acquire Splunk MCP Server Endpoint and API Key

Before you begin, make sure the Splunk user account you use has the required capabilities. Add the following capabilities to the user role:

- mcp_tool_execute
- mcp_tool_admin

To acquire Splunk credentials, perform the following steps:

1. Log in to your Splunk Web UI as a user who has the required capabilities.
2. Navigate to **Apps** > **Find More Apps** > Splunk MCP Server.
3. If you are installing Splunk MCP Server for the first time, enter your Splunk user name and password when prompted, and install the app.
4. If Splunk MCP Server is already installed, open the Splunk MCP Server app.
5. On the Splunk MCP Server page, click **Create MCP Encrypted Token**.
6. In the Create MCP Encrypted Token dialog:
   a. From User, select the user account for which you want to create the token.
   b. **Audience** - enter mcp.
   c. **Expiration** - Specify when the token should expire.
   d. Select **Create**.
   e. Copy the displayed token—it will only appear once.
   f. Copy the endpoint URL displayed in the **Endpoint**.

Use the copied endpoint URL and API key in the YAML file provided in GigaVUE-FM.

Acquire Elastic Endpoint and API Key

Elastic supports two setups for acquiring the API key:

- For Elastic version 9.1 or earlier, use the local MCP server to generate the API key.
- For Elastic version 9.2 or later, use the remote MCP server to generate the API key.

Before you begin, make sure you have an Elastic account with either of the following privileges:

- manage_api_key

- admin

Acquire Elastic credentials (Local MCP Server)

To acquire credentials from Elastic 9.1 or earlier, perform the following steps:

1. Log in to the Elastic UI with an account that has the required privileges.
2. Navigate to **Stack Management** > **Security** > **API Keys**.
3. Click **Create API key**. In the Create API key page, provide the following:
   - **Name**: Enter a descriptive name for the key.
   - Expiration, privileges, and metadata: For this integration, you do not need to specify an expiration date, custom security privileges, or metadata.
4. Click **Create API key**. The UI displays the new key:
   a. Copy the API key ID—it is shown only once.

NOTE: Set the ES_URL in the YAML configuration file in the following format, replacing the placeholder with your Elastic endpoint: https://<ELASTIC_IP:<endpoint>

Acquire Elastic credentials (Remote MCP Server)

To acquire credentials from Elastic 9.2 or later (Remote MCP Server), refer to the detailed steps in Elastic Agent Builder MCP server.

Use the endpoint URL and API key in the YAML file provided in GigaVUE-FM

Enable Gigamon Insights in GigaVUE-FM

After you acquire the LLM and MCP server credentials, follow these steps to enable Gigamon Insights in GigaVUE-FM:

NOTE: To enable Gigamon Insights, you must be a member of the fm_admin or fm_super_admin user group in GigaVUE-FM.

1. Sign in to GigaVUE-FM. Navigate to **Settings** > **Systems** > **Gigamon Insights**.
2. Select the Gigamon Insights instance configured for this Gigamon-Insights and, under the **Actions** drop-down list, click **Edit**. In this page, you can configure the LLM and Server Settings.
3. Before configuring LLM and server settings, you need to confirm your terms with the Gigamon AI Usage Telemetry Data Collection & Usage Notice.

   NOTE: In this LA release, accepting AI usage telemetry is required to use Gigamon Insights. If you do not accept the usage terms, you cannot enable or access this feature.
4. [Optional] Enter your Customer ID when you accept the terms and conditions.

5. Download the YAML template to configure the LLM and MCP server. Open the template in a text editor and edit the configuration file with your environment-specific information as described below:
   a. LLM provider details (for example, Amazon Bedrock) and models you want to use.
   b. MCP server details (for example, Elastic or Splunk endpoints and credentials).
6. You can also copy the below template:

```
# List of LLM providers and models
llmModels:
  ## Uncomment the bedrock section below to use bedrock endpoint
  #bedrock:
    ## AWS Bedrock display name
    #BEDROCK_AWS_NAME: AWS Bedrock
    ## AWS region for Bedrock (e.g., us-east-2)
    #BEDROCK_AWS_DEFAULT_REGION:
    ## AWS Access Key ID for Bedrock
    #BEDROCK_AWS_ACCESS_KEY_ID:
    ## AWS Secret Access Key for Bedrock
    #BEDROCK_AWS_SECRET_ACCESS_KEY:
    ## Comma-separated list of AWS Bedrock models (e.g., us.anthropic.claude-sonnet-4-20250514-
v1:0)
    #BEDROCK_AWS_MODELS:

  ## Uncomment the google section below to use google endpoint
  #google:
    ## Google Vertex display name
    #GOOGLE_NAME: Google Vertex
    ## Paste your Google service account JSON here
    ## TIP: Wrap the entire JSON in single quotes to avoid YAML escape issues
    #GOOGLE_SERVICE_KEY_FILE:
    ## Command-seperated list of Vertex AI models to use (e.g., gemini-2.5-pro)
    #GOOGLE_MODELS:

# List of MCP Servers
# Only one MCP Server is supported at a time
mcpServers:
  ## Uncomment elastic-local-mcp-server section below to use local hosted Elastic Server
  #elastic-local-mcp-server:
    ## URL of ES Server
    #ES_URL:
    ## API Key for Elastic
    #ES_API_KEY:
    ## Set this to 'TRUE' to bypass security checks; otherwise, set it to 'FALSE'
    #ES_SSL_SKIP_VERIFY:

  ## Uncomment elastic-agent-builder section below to use remote hosted Elastic Server
  #elastic-agent-builder:
    ## URL of ES Server
    #ES_URL:
    ## API Key for Elastic
    #ES_API_KEY:

  ## Uncomment splunk-mcp-server section below to use remote hosted Splunk Server
  #splunk-mcp-server:
    ## Provide a full MCP URL for Splunk (preferred). Example:
```

```
#SPLUNK_MCP_URL: https://my-deployment.example.com/mcp/v1/
## Auth Token for Splunk
#AUTH_TOKEN:
```

7. Upload the file. In **User Defaults**, the system automatically populates the following:
   a. LLM Service and LLM Model - To use as default. This model appears preselected when you open Gigamon Insights.
   a. MCP server - To use by default (for example, the Elastic MCP server from your YAML file). Only one MCP server is supported.
8. Click **Save** to store the configuration and defaults.

What to do next

Refer to Get Started with Gigamon Insights.

Get Started with Gigamon Insights

The following section provides instructions on how to start using Gigamon Insights and using prompts effectively.

Launch Gigamon Insights

Before launching Gigamon Insights, confirm that your GigaVUE-FM user account is assigned to the Gigamon Insights user group.

To start using Gigamon Insights:

1. Enter the IP address or FQDN of the Gigamon Insights instance (configured on VMware or AWS) in a web browser. The SAML sign-in page appears.
2. Follow the prompt to access the GigaVUE-FM login page. Enter your GigaVUE-FM credentials to launch Gigamon Insights.
3. Start using the Gigamon Insights chat interface.
   - The left panel provides:
     - **New chat**: Opens the new chat.
     - **History**: View prompts and responses from the past 30 days.
     - **Prompts**: Predefined prompts are organized into three main use-case categories—security, network, and application use cases. Each category includes Structured, Hybrid, and Quick prompts, as described in the How to Use Prompts Successfully.

        - **Create Prompt**: Save personal prompts only visible to you. You can categorize these prompts.
    - **Quick Actions**: Curated best practice prompts based on commonly used workflows. You can select a category and subcategory to load the predefined prompts into the chat window.
    - **Prompt Area**: You can enter your free-form questions or use prompts from the prompt library. Additionally, you can upload a dashboard screenshot per prompt.
    - Switch LLMs: When both Amazon Bedrock and Google Vertex are configured, you can switch the active LLM between them as needed.
4. Type a prompt. You can either select a predefined prompt or type an open prompt. Review your question and any attached screenshot and submit the prompt. Gigamon Insights returns an answer.
5. Submit your feedback by selecting Thumbs-up or Thumbs-down.

How to Use Prompts Successfully

This section describes how to write effective prompts for Gigamon Insights. It focuses on three areas you control in every question:

- Prompt and context engineering – how you describe the task.
- Time window management – which period of data you analyze.
- Data coverage – how much data you scan and return.

Use these practices to reduce hallucinations, make outputs explainable, and keep results repeatable.

Prompt and context engineering

Prompt and context design directly affect answer quality and the risk of hallucinations. Large language models (LLMs) balance all instructions statistically, rather than executing them like code. Overly long or conflicting prompts can still produce confident but incorrect answers.

In Gigamon Insights, you shape your questions and task prompts to get reliable, explainable results.

Use these guidelines for any question:

- **Be clear and direct**. State the objective in a single sentence before adding details. For example:
  “Analyze authentication failures in the past 7 days to identify suspicious login behavior.”
- **Prioritize high-impact instructions**. Use concise, high-priority instructions instead of many low-impact details. Put critical requirements (for example, “use read-only queries”, “separate facts from interpretation”) early in the prompt.
- **Make uncertainty explicit**. Tell the assistant how to respond when data is incomplete. For example:
  “If the data is not sufficient, state that you cannot determine the answer from the available data.”

- **Avoid conflicting requirements**. Do not combine instructions such as “never ask for clarification” and “must always be correct” in the same prompt. Prefer one clear priority: accuracy, speed, or minimal interaction.
- **Treat prompts as iterative**. Refine your wording based on previous answers. Tighten the objective, time window, or filters until the output is consistently useful.
- **Separate facts from interpretation**. Ask the assistant to structure responses into sections such as:
  - Summary
  - Evidence (directly from tools and logs)
  - Assessment (interpretation)
  - Recommendations
- **Specify output format**. When you need a particular structure, say so: “Return: summary, evidence, assessment, and recommendations.” This improves consistency and makes results easier to review or export.

Prompt types and when to use them

Use one of three prompt styles depending on your task.

Table 1: Prompt Types

Prompt Type	Best for	Characteristics	Example
Structured	Complex analysis; regulated or security-critical workflows	Longer, explicit requirements and constraints	**Objective**: Analyze network and authentication logs from the last 24 hours to detect port scanning, host sweeping, profiling, or spoofing. **Data scope**: Use only the network flow logs and authentication logs available in Gigamon Insights. Analysis requirements: - Identify patterns that indicate scanning, sweeping, profiling, or spoofing. - Use time-based aggregation where helpful (for example, counts per source IP per minute or hour). - Correlate authentication events with network events when needed. **Output format**: • Summary • Evidence (including key query results and example records) • Assessment • Recommendations **Constraints**: • Do not infer behavior without supporting data. • Clearly separate facts from interpretation. • Call out any data gaps or limitations.
Quick/Short	Fast, exploratory checks; experienced users	One to two sentences; minimal structure	“Check the last 24 hours of authentication logs for suspicious activity.” "Check the last 24 hours of network and authentication logs for port scans, host sweeps, profiling, or spoofing. Return any suspicious behavior you find."
Hybrid	Most day-to-day investigations	Brief objective plus clear output structure	“Analyze the last 24 hours of network and authentication data to detect port scanning, host sweeping, profiling, or spoofing. **Return**: • Summary • Evidence

Prompt Type	Best for	Characteristics	Example
			• Assessment • Recommendations”

Time window management

Time window management controls which period of data the model considers. In conversation, you often use natural expressions such as “last week” or “recently”. Gigamon Insights converts those expressions into precise, inspectable time filters and can complement them with UI-based controls.

- Use conversational time for ad-hoc questions.
  Examples: “last week”, “past 7 days”, “last 4 hours”.
- Use precise windows for repeatable work.
  For repeatable queries, investigations, and audits, specify start and end timestamps with time zones.
- State the time range type.
  - Absolute: Specific dates (for example, “from 2025-05-01 to 2025-05-15 UTC”).
  - Relative: Rolling windows (for example, “past 7 days”).
  - Comparative: Period vs Period (for example, “compare last 7 days to the previous 7 days”).
- Set the granularity. Specify whether you want minute-level, hour-level, or day-level analysis, depending on data volume and the kind of pattern you expect.
- Call out business-hours or weekend rules. State whether you want to exclude weekends or focus only on business hours.
- Reuse the prompt styles. Apply structured, quick, and hybrid prompts to express time requirements clearly and consistently.

Time expression categories and examples

Gigamon Insights is designed to handle a wide range of natural time expressions. Use the patterns below as guidance.

Absolute time range (fixed dates)

Examples:

- “... from Jan 1 to Mar 31, 2025.”
- “... on October 12, 2025.”
- “... between 2025-05-01 and 2025-05-15.”
- “... dated June 2025.”
- “... Aug 18 between 2pm and 4pm.”
- “... on March 3 from 9:00–11:00 UTC.”
- “... after midnight on Jan 1.”

- "... for Q2 2025."
- "... for fiscal year 2024."
- "... for the first week of December."

Relative time range (rolling or anchored)

Examples:

- "... from the last 24 hours."
- "... in the past 7 days."
- "... over the last month."
- "... from the last hour."
- "... in the last 90 days."
- "... since Monday."
- "This week's usage."
- "Last month's usage."
- "This quarter so far."
- "Week to date."
- "... last 30 days, excluding weekends."
- "Business-hours traffic this week."
- "Nighttime usage over the past month."

Comparative time range (before vs after, period vs period)

Examples:

- "Compare this week to last week."
- "Month over month revenue."
- "Quarter over quarter growth."
- "Week over week error rates."
- "Year over year sales."
- "Compare last 30 days to the previous 30 days."
- "Compare January vs February."
- "7-day average week over week."
- "Compare the last 14 days to the 14 days before that."
- "How did today compare to the same day last week?"

Mixed or domain-specific language

Examples:

- "Are errors worse than they were last month?"
- "Is this week better or worse than usual?"
- "Spikes in the last 15 minutes."
- "What's changed recently?"

Data coverage

Data coverage determines how much data the system reads and returns, and whether that fits within the model's context window and platform limits. Gigamon Insights treats coverage as an explicit, enforceable contract rather than an assumption. The guidelines are:

- Treat each question as a coverage contract. State whether you need a sample, a bounded slice, or full coverage.
- Use simple coverage modes. Use sample, bounded, and complete coverage and map them to clear row limits and behavior.
- State what you need:
  - A summary or sample (small row limit).
  - A bounded view (filtered by time or attributes, often aggregated).
  - Full coverage (subject to safety limits, often for exports or audits).
- Expect safety caps even for "full coverage". When you ask for "full coverage" or say "show me everything", Gigamon Insights still applies hard limits (for example, a maximum number of rows per table or per user, such as 10,000 rows).
- Understand system behavior at high coverage. For resource-intensive requests, the system may:
  - Return aggregated views instead of every row.
  - Ask you to narrow the time range or filters.
  - Page results and summarize each chunk rather than returning all raw rows at once.

Debuggability and Troubleshooting

This section guides tools and procedures for diagnosing, analyzing, and resolving issues in the Gigamon Insights environment. It includes steps for generating sysdump files that capture logs and system data essential for troubleshooting and verifying the Gigamon Insights service status.

Generate Sysdump files

You can create a sysdump using the following steps:

1. Log in to the Gigamon Insights

   ```
   ssh admin@<gi-ip-or-hostname>
   ```

2. Switch to sudo access.

   ```
   sudo su
   ```

3. Navigate to the sysdump script directory

   ```
   cd /gi-data/gi/bin
   ```

4. Run the sysdump script

```
./generate-sysdump.sh
```

The script creates a sysdump archive and displays the file path where the archive is stored.

```
[admin@GI-61204 ~]$ cd /gi-data/gi
[admin@GI-61204 gi]$ cd bin
[admin@GI-61204 bin]$ ./generate-sysdump.sh
This script requires root privileges to collect all system information.
Please run as root or with sudo.
Attempting to re-run with sudo...
Starting system diagnostic capture to /gi-data/sysdumps/sysdump-gi-6.12.04_20260427_110233
Capturing Host information...
Capturing Docker information...
Capturing Application configurations...
Capturing Container details...
Capturing Network information...
Capturing Logs...
Capturing System information...
Capturing Metrics...
Capturing FM gi information...
Successfully created encrypted archive /gi-data/sysdumps/sysdump-gi-6.12.04_20260427_110233.tar.gz.gpg
Use 'gpg /gi-data/sysdumps/sysdump-gi-6.12.04_20260427_110233.tar.gz.gpg' to decrypt.
```

5. Download the generated sysdump file from the specified location using the scp command.
6. Follow the instructions in Contacting Technical Support to send the logs to the support team.

Verify Gigamon Insights Service Status

To verify whether Gigamon Insights is active, run the following command:

```
systemctl status gi
```

```
[admin@GI-61204 gi]$ systemctl status gi
● gi.service - Gigamon Insight Service
   Loaded: loaded (/etc/systemd/system/gi.service; enabled; vendor preset: disabled)
   Active: active (exited) since Fri 2026-04-24 11:24:21 UTC; 2 days ago
  Process: 2325 ExecStart=/gi-data/gi/bin/start.sh (code=exited, status=0/SUCCESS)
 Main PID: 2325 (code=exited, status=0/SUCCESS)
    Tasks: 0 (limit: 101911)
   Memory: 0B
   CGroup: /system.slice/gi.service
```

AI Telemetry Usage

For the LA release of Gigamon Insights, you must enable AI Usage Telemetry. This will periodically share non-sensitive AI usage data with Gigamon for the purpose of improving AI quality. Gigamon is committed to security and privacy. Any Personally Identifiable Information (PII) is sanitized and no data is shared with third-parties. All possible security measures are also implemented to protect and isolate the data collected.

Telemetry collection is designed to minimize data storage, exclude AI-generated responses, and protect user privacy by automatically removing personally identifiable information (PII).

The system collects the following categories of telemetry.

User Prompts (after PII Scrubbing)

The text you enter as a prompt is processed by an automated PII scrubbing pipeline, and any sensitive or identifiable information is removed before it is used. Only the sanitized version of the prompt is stored. PII scrubbing process detects and removes or masks commonly sensitive identifiers, including:

- IP addresses
- MAC addresses
- Email addresses
- Phone numbers
- Government identifiers (SSN)
- Credit card numbers
- Street/mailing addresses

Detected PII elements (for example, IP addresses or email addresses) are replaced with generic placeholders such as <IP>, <MAC_ADDRESS>, or <EMAIL_ADDRESS>.

For example,

Original Prompt: Search all available metadata logs for any records associated with the specified Email address, IP address or MAC address, including historical and current entries. Search Criteria:

- IP Address: 10.0.1.162
- MAC Address: 20:e5:2a:b6:93:f1
- Email Address: john.doe@acme.com

Stored Telemetry Version: Search all available metadata logs for any records associated with the specified Email address, IP address or MAC address, including historical and current entries. Search Criteria:

- IP Address: <IP>
- MAC Address: <MAC_ADDRESS>
- Email Address: <EMAIL_ADDRESS>

Operational Metadata

Operational metadata describes how each AI request is processed and includes:

- Message ID – Unique identifier for the AI request.
- Conversation ID – Identifier for the chat session.
- Parent Message ID – Identifier for the preceding message.
- Tenant ID/Tenant code – Unique deployment identifiers assigned during setup.
- Timestamps – When the message was created and last updated.
- Provider identifier – AI or LLM provider (for example, Bedrock).
- Model identifier – Model used for inference.
- Token counts – Number of input (prompt), output (completion), and total tokens.

Operational metadata and the scrubbed prompt text are stored as structured JSON records.

User Feedback

If you choose to rate or comment on an AI response, the following optional telemetry may be collected:

- User rating – Thumbs up or thumbs down.
- Optional comment – Free text feedback about the response.

Feedback comments also undergo PII scrubbing before storage. Feedback telemetry is used to identify areas where AI responses can be improved.