Varonis Atlas - Anthropic Compliance API Documentation

The Anthropic Compliance API integration allows Varonis Atlas to pull Claude activity logs from Claude Enterprise & Claude Platform, process that activity through runtime policies, and make the resulting events available in AI Investigation for monitoring, alerting, session review, export, and downstream analysis.

This integration is a **pull-based log ingestion source** whereby Varonis Atlas periodically calls the secure Anthropic Compliance API. From this intelligence, organizations gain centralized AI monitoring, detection, and response capabilities over Claude usage throughout enterprise environments.

By ingesting Claude chat activity and evaluating it against Atlas runtime policies, the integration surfaces sensitive data exposure, policy violations, and risky user behavior for investigation, alerting, reporting, and audit readiness. Teams can also visualize incidents in context and maintain continuous oversight of AI activity without disrupting end-user workflows.

Prerequisites

Before configuring this integration, make sure you have:

1. **An Anthropic Compliance API key**
    - The API key must have permission to export chat activity through the Anthropic Compliance API.
2. **Access to the Varonis Atlas Admin Console**
    - You must be able to add and configure log source integrations.
3. **A target project**
    - Select the Varonis Atlas project where ingested Anthropic activity should appear.
4. **An available Data Plane**
    - Select the Data Plane that should pull, process, and forward the Anthropic activity into Varonis Atlas.

Configure the Anthropic Compliance API Integration

1. Open Log Sources

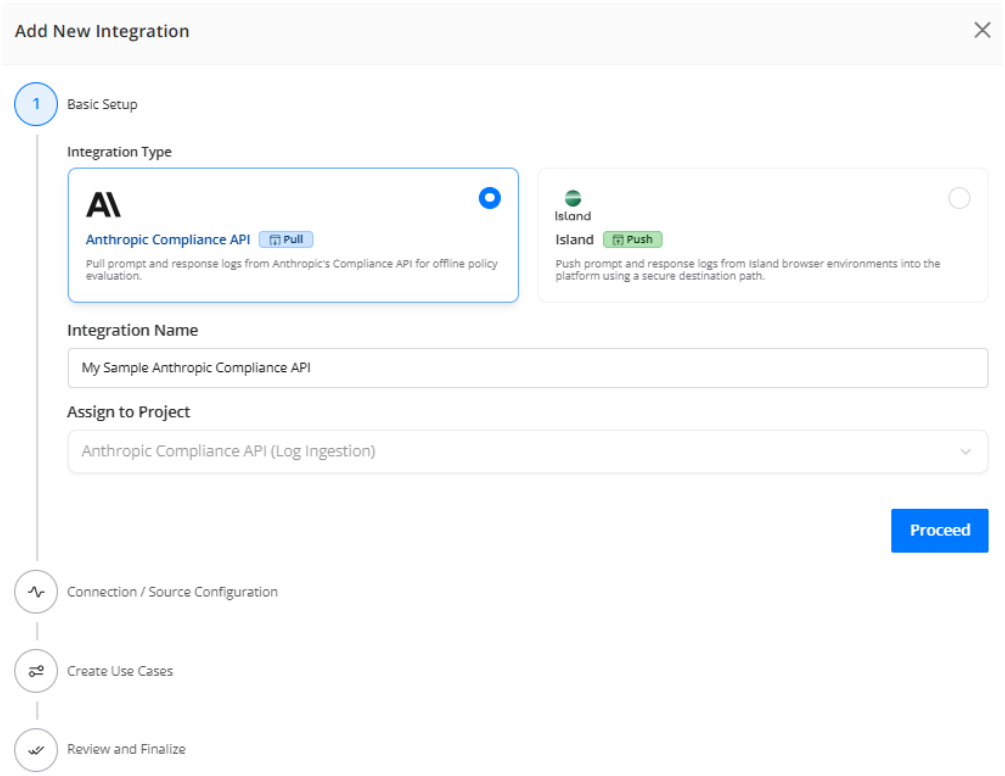
In Varonis Atlas, open the **Admin Console** and navigate to **Log Sources**.

Click **Add New Integration**.

2. Select Anthropic Compliance API

From the list of available integrations, select **Anthropic Compliance API**.

This integration uses a **pull mechanism**, meaning Varonis Atlas will call the Anthropic Compliance API on a recurring schedule to retrieve new activity.



3. Name the Integration

Enter a unique and meaningful name for the integration.

Use a name that helps administrators understand what environment, business unit, or Anthropic account this integration represents.

Example names:

- Anthropic Compliance API - Production
- Claude Activity - Engineering
- Anthropic Logs - Corporate Workspace

4. Assign the Integration to a Project

Select the Varonis Atlas project where the Anthropic activity should appear.

All ingested events from this integration will be associated with the selected project and will be visible in the runtime and investigation workflows for that project.

5. Select the Data Plane

Choose the Data Plane that should handle this integration.

The selected Data Plane will be responsible for pulling events from the Anthropic Compliance API and processing them through Varonis Atlas runtime policies.

6. Configure the Anthropic API Key

Enter the Anthropic API key that Varonis Atlas should use to retrieve chat activity.

The API key must have the required permissions to export activity through the Anthropic Compliance API.

7. Finalize the Configuration

Review the configuration and save the integration.

After the integration is created, it will appear in the **Log Sources** table. Varonis Atlas will begin syncing activity from the Anthropic Compliance API.

How Ongoing Sync Works

Once configured, Varonis Atlas calls the Anthropic Compliance API every **5 minutes** to retrieve new chat activity.

For each sync cycle, Varonis Atlas will:

1. Pull a new batch of Anthropic activity
2. Ingest any new requests and responses

3. Process the activity through applicable runtime policies
4. Generate alerts or issues for policy violations
5. Store the events in AI Investigation and runtime observability workflows
6. Make the activity available for search, session review, reporting, export, and dataset creation

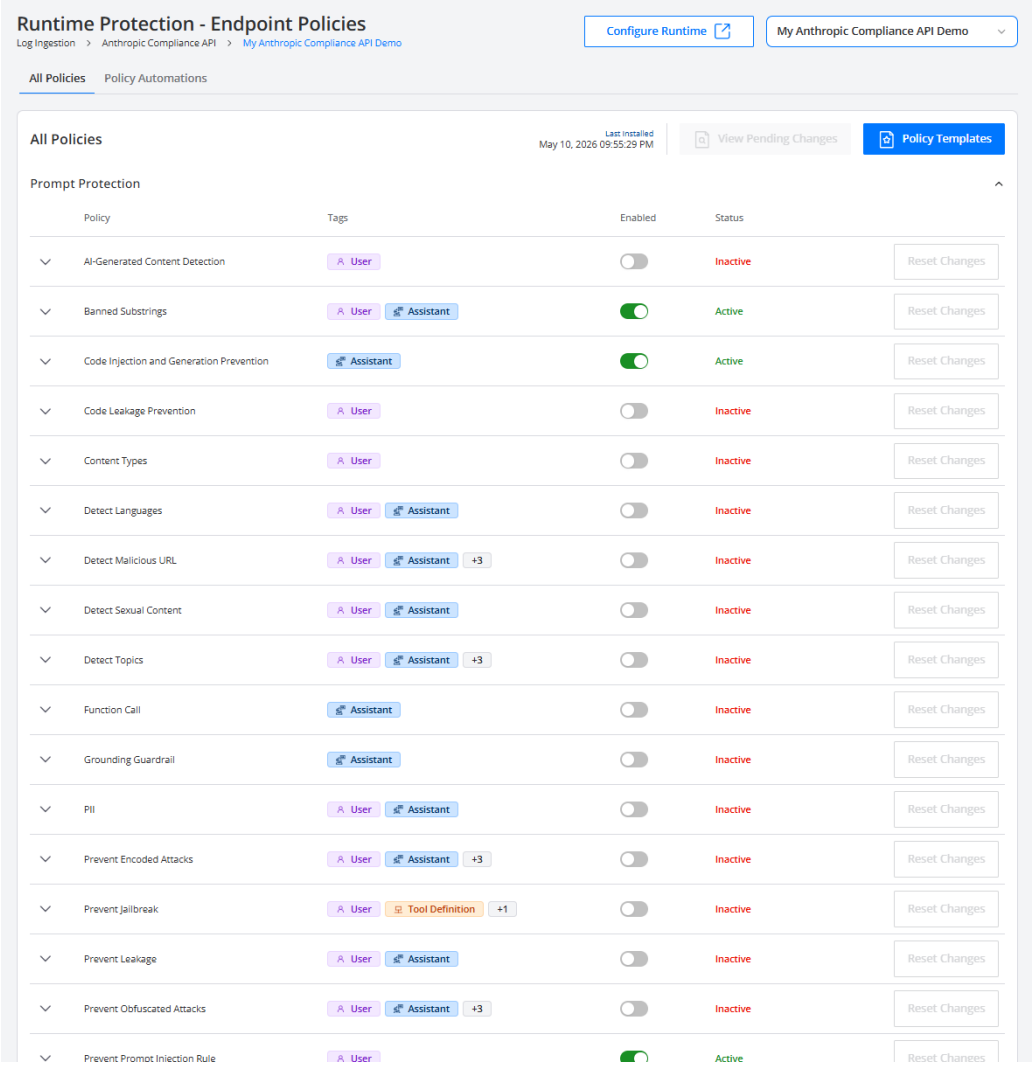
The ongoing sync continues every 5 minutes until the integration is **paused** or **deleted**.

Apply Runtime Policies

After the integration is configured, open the **Runtime Policies** page to apply policies to the Anthropic log source.

Most chat-based runtime policies can be applied to Anthropic Compliance API activity, including policies that detect:

- Sensitive data or PII in prompts or responses
- Code leakage
- Prompt injection attempts
- Toxic or inappropriate messages
- Jailbreak attempts
- Other policy-defined content risks



Because this integration is based on historical logs rather than inline request enforcement, only **logging** and **alerting** actions are available.

If a policy is configured with a **block** or **modify** action, Varonis Atlas treats that action as an **alert** for Anthropic Compliance API activity. The original Claude interaction has already occurred, so Varonis Atlas cannot block or modify it retroactively.

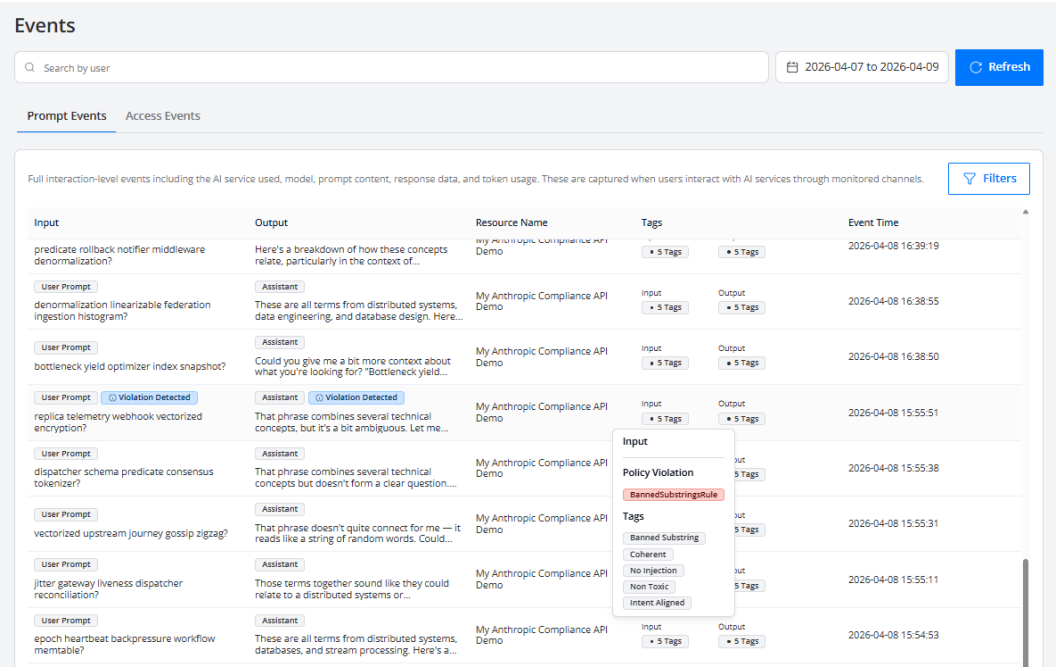
View Activity in AI Investigation

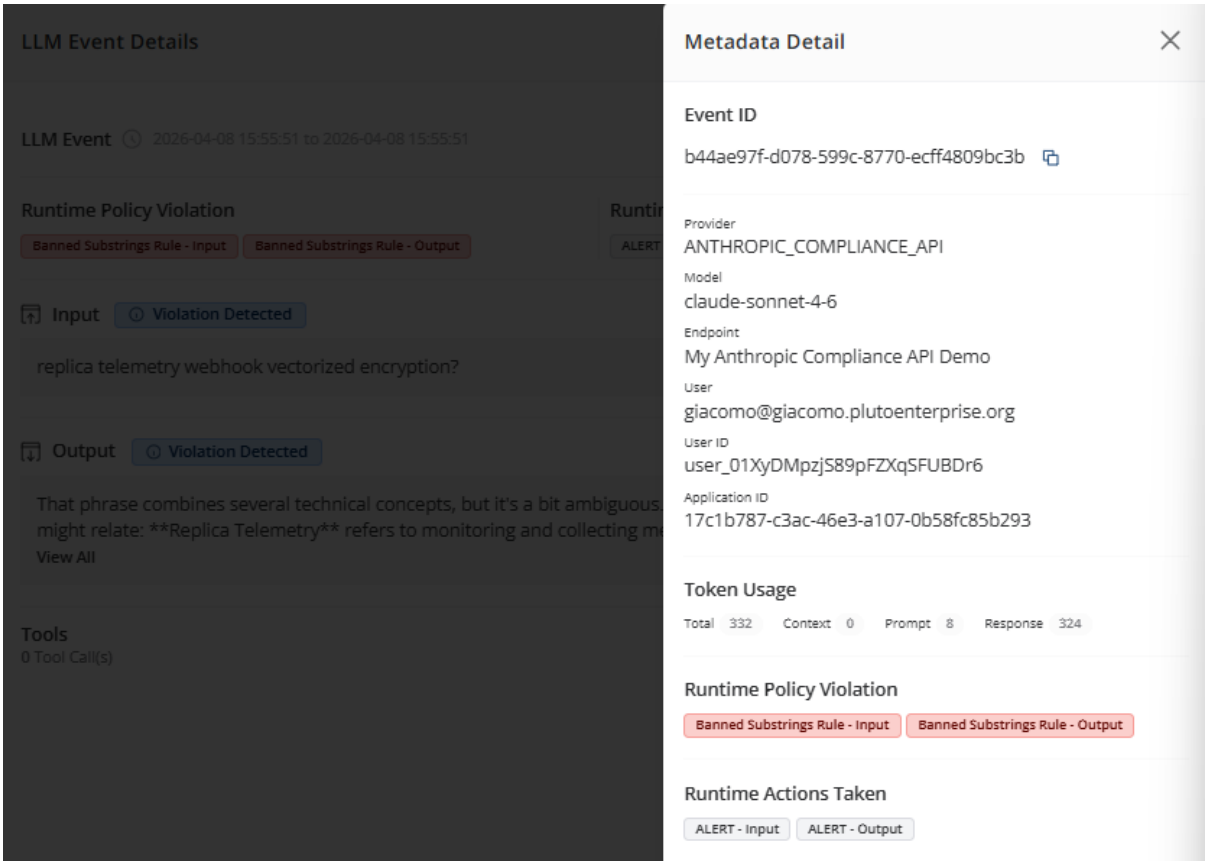
Ingested Anthropic activity appears in **AI Investigation**, where users can review activity, investigate violations, and understand the surrounding context.

Users can view:

- Individual prompt and response events
- Policy violations
- Generated issues
- User metadata, including user email and user ID
- Request metadata, including model and token usage
- Session metadata
- File or artifact context associated with the activity, where available

Varonis Atlas can show metadata and context related to files or artifacts, but this integration does **not** provide access to the actual file contents.





Review Sessions

Anthropic activity is grouped into sessions where session information is available.

In the **AI Investigation Sessions** view, users can display the full chat in chronological order. This allows investigators to review policy violations in the context of surrounding session activity rather than as isolated events.

This is useful for investigating activity such as:

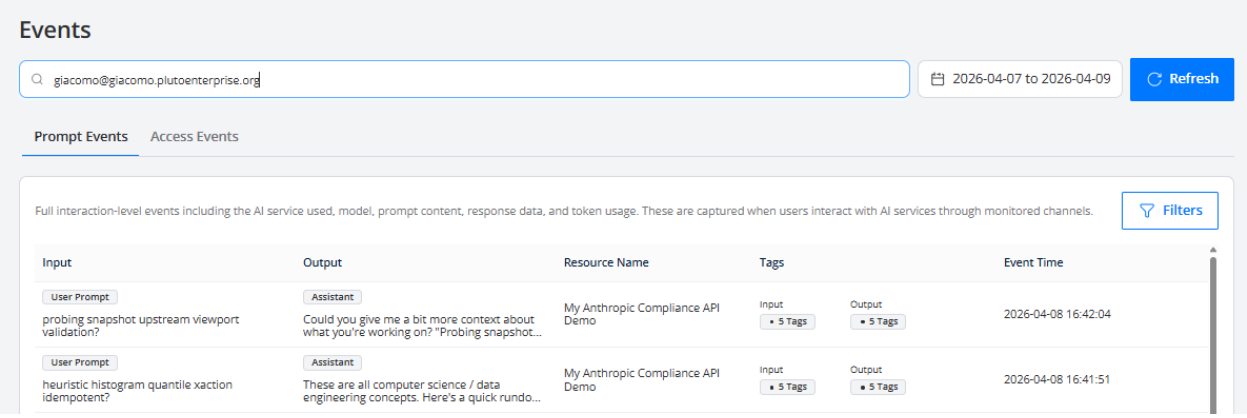
- Repeated jailbreak attempts
- Suspicious prompt patterns
- Sensitive data exposure
- Policy violations that only become clear when viewed alongside surrounding messages
- User behavior across a session

Session-based policies can be applied to sessions (chats) to automatically detect these events.

Search by User

Users can search by **user ID** to view events associated with a specific user.

This allows teams to investigate user-specific activity, review policy violations, and understand how a particular user is interacting with Anthropic systems over time.



Export and Reuse Events

Anthropic Compliance API activity supports the same observability and investigation workflows as other ingested runtime activity.

Users can:

- View individual events
- Review full sessions
- See policy violations
- Investigate generated issues
- Export events to CSV or JSON
- Include events in reports
- Compile prompts into datasets for future Evaluations or Red Teaming

Historical and ongoing activity is displayed using the correct event timestamps, so users can analyze Anthropic activity in the appropriate time window.

Backfill Historical Activity

In addition to the ongoing 5-minute sync, users can initiate a backfill to pull and process historical Anthropic activity.

Start a Backfill

To start a backfill:

1. Open the Anthropic Compliance API integration.
2. Open the integration detail drawer.
3. Navigate to the **Backfill** tab.
4. Enter the required start date and time.
5. Enter the required end date and time.
6. Start the backfill job.

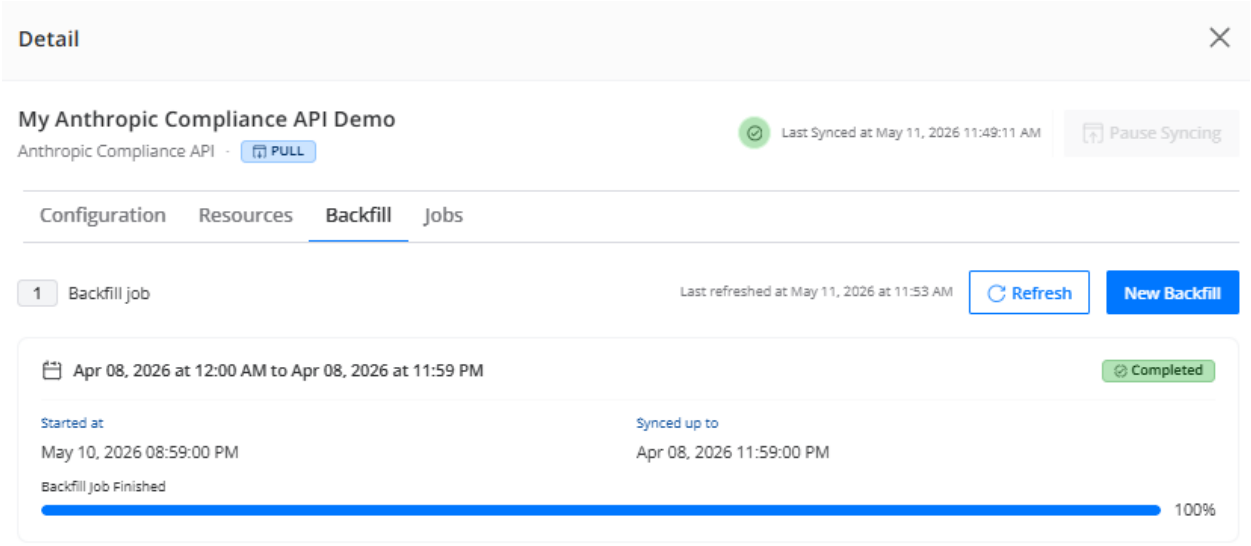
Varonis Atlas will pull and process activity for the selected time range.

Monitor Backfill Progress

Backfill progress can be reviewed from the integration detail drawer.

The **Backfill** tab shows backfill jobs that have been initiated for the integration.

The **Jobs** tab shows the status of individual processing batches.



Historical Activity Behavior

Backfilled Anthropic activity supports the same functionality as ongoing synced activity.

Historical events can be:

- Processed through runtime policies
- Logged into AI Investigation
- Displayed in sessions
- Used to generate alerts and issues
- Exported to CSV or JSON
- Included in reports
- Compiled into datasets for Evaluations or Red Teaming

Backfilled events are displayed according to their original event timestamps, so they appear in the correct historical time frame.

Limitations

The Anthropic Compliance API integration currently has the following limitations:

1. **No inline blocking or modification**
    - This integration processes logs after activity has already occurred.
    - Block and modify actions are treated as alerts.
2. **No file content inspection**
    - Varonis Atlas can display available metadata and context related to files or artifacts, but the integration does not provide the actual file contents.
3. **No use-case based separation yet**
    - Additional separation of activity by use case or other attributes is not currently supported.
    - Future support is planned to enable more granular observability and policy application by attributes such as user groups.

Future Enhancements

Varonis Atlas plans to expand this integration over time, including support for more granular policy application and observability based on additional attributes such as user groups.

Varonis Atlas also intends to leverage DSPM capabilities and data classification insights to help identify potential sensitive data sharing through files or other artifacts when activity falls outside approved policy.