Understanding Data Collection in Sumo Logic

1.1 What is Data Collection?

Data Collection in Sumo Logic is the process of **gathering logs, metrics, and events from various sources across your environment** — such as servers, applications, network devices, and cloud services — and **sending them securely to the Sumo Logic Cloud** for storage, analysis, and visualization.

In simple terms, **data collection is how your raw operational data reaches Sumo Logic** so that analysts and engineers can monitor system health, investigate incidents, and derive insights.

1.2 Purpose of Data Collection

The main goal of data collection is to **enable centralized visibility** across your IT infrastructure. By collecting all relevant data into a single cloud platform, organizations can:

- Detect and investigate security threats
- Monitor application performance and availability
- Troubleshoot system and network issues faster
- Meet compliance and auditing requirements
- Gain real-time operational insights through dashboards and alerts

How Data Collectors Work in Sumo Logic

2.1 Overview

A **Data Collector** in Sumo Logic is the **component responsible for gathering data from different systems and sending it securely to the Sumo Logic Cloud** for analysis.
It acts as the **bridge** between your data sources (servers, applications, or cloud services) and the Sumo Logic platform.

Collectors ensure that data is continuously transmitted in real time, even from geographically distributed environments, while maintaining reliability and security.

2.2 The Role of a Collector

The collector’s job is to:

1. **Connect to your data source** (e.g., log files, syslog streams, cloud APIs).
2. **Read and process** that data in a defined format.
3. **Attach metadata** (e.g., source name, host, category).
4. **Securely transmit** the data to the **Sumo Logic Cloud ingestion endpoint**.

Once the data reaches Sumo Logic, it becomes searchable, indexable, and usable for dashboards, alerts, and analytics.

Differences Between Data Collectors in Sumo Logic

2.3 Overview

Sumo Logic provides **two main types of data collectors** — **Installed Collectors** and **Hosted Collectors**.
Both serve the same purpose: to collect data and send it to the Sumo Logic Cloud. However, they differ in **how** they are deployed, **what kind of data** they collect, and **where** they operate.

Understanding these differences is important for designing an efficient and scalable data collection strategy.

2.4 Installed Collector

Definition

An **Installed Collector** is a **lightweight agent installed directly on a machine** (such as a server, workstation, or virtual machine).
It runs as a background service and is responsible for collecting **local data** such as log files, system metrics, or custom scripts.

Key Features

- Installed on **on-premises systems** or **VMs**.
- Can collect data from **files, directories, syslog, or scripts**.
- Works well in **hybrid or private data centers**.
- Requires **local installation and maintenance**.
- Supports **local buffering** — stores data temporarily if the network is down.
- Sends data securely to Sumo Logic using **HTTPS/TLS**.

Common Use Cases

- Collecting `/var/log/syslog` from Linux servers.
- Gathering Windows Event Logs from domain controllers.
- Reading web server logs (e.g., Apache, Nginx) stored locally.
- Monitoring custom applications that generate logs on-premise.

Example

You install an Installed Collector on a Linux machine and configure a source to read:

```
/var/log/apache2/access.log
```

It continuously reads this file and sends the logs to Sumo Logic Cloud.

2.5 Hosted Collector

Definition

A **Hosted Collector** is a **cloud-based collector managed entirely within the Sumo Logic platform**.
You don’t install any software; instead, you configure it in the Sumo Logic UI to collect data from **cloud services or remote endpoints** using APIs or integrations.

Key Features

- Managed by **Sumo Logic Cloud** (no local installation).
- Ideal for **cloud-native or SaaS environments**.
- Collects data through **APIs, webhooks, or cloud integrations**.
- Zero maintenance — no need to update or monitor locally.
- Highly scalable and automatically available across regions.

Common Use Cases

- Collecting AWS CloudTrail or CloudWatch logs.
- Ingesting Azure Activity Logs or Office 365 audit data.
- Integrating with GCP, Kubernetes, or SaaS applications.

Example

You create a **Hosted Collector** and link it to your AWS account to collect CloudTrail logs via an API integration — all configuration happens in the cloud.

2.6 How a Collector Works (Step-by-Step)

Below is the typical lifecycle of how a collector operates:

Step 1: Configuration

- The collector is created either manually or automatically through an integration.
- You define the **data sources** it will monitor (e.g., log files, directories, syslog, APIs).

Step 2: Data Ingestion

- The collector starts **reading data** from its assigned sources in real time.
- It can handle multiple sources simultaneously.

Step 3: Metadata Tagging

- Before sending data, the collector attaches **metadata** such as:
  - `_sourceHost` (hostname)
  - `_sourceCategory` (data type or system)
  - `_collector` (collector name)

This metadata helps organize and filter logs later in the Sumo Logic UI.

Step 4: Secure Transmission

- Data is **compressed, encrypted (TLS/HTTPS)**, and transmitted to Sumo Logic's **ingestion endpoint** in the cloud.
- If the network connection is temporarily lost, the collector **queues data locally** and **resends it** when connectivity is restored — ensuring no data loss.

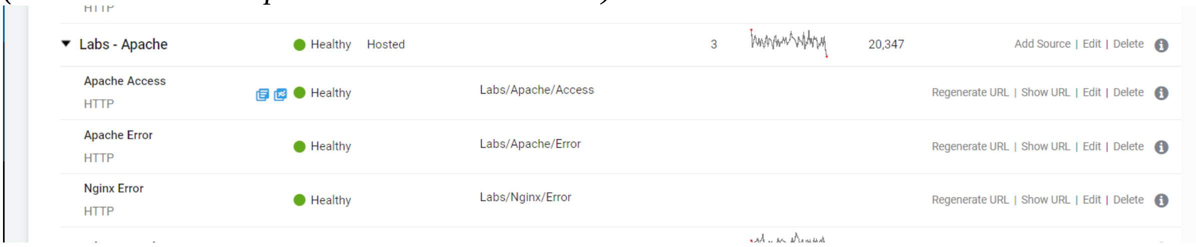
Step 5: Cloud Processing

- Sumo Logic receives the data, indexes it, and makes it searchable through the **Search**, **Dashboards**, or **Metrics** interfaces.

Exploring the Labs-Apache Collector in Sumo Logic

In this section, we will review the **Labs-Apache Collector** configured in the Sumo Logic environment. The screenshots below demonstrate the collector's details as displayed under **Manage Data → Collection** in the Sumo Logic UI.

(Screenshot: Labs-Apache Collector Overview)



▼ Labs - Apache	● Healthy	Hosted		3	20,347	Add Source | Edit | Delete
Apache Access HTTP	● Healthy		Labs/Apache/Access			Regenerate URL | Show URL | Edit | Delete
Apache Error HTTP	● Healthy		Labs/Apache/Error			Regenerate URL | Show URL | Edit | Delete
Nginx Error HTTP	● Healthy		Labs/Nginx/Error			Regenerate URL | Show URL | Edit | Delete

---

3.1 Understanding the Columns and Their Meanings

After navigating to the **Collection** page, several columns provide insights into the configuration and health of the collectors and sources. Below is an explanation of each column as it appears for the **Labs-Apache Collector**.

---

a. Health Status

- The **Labs-Apache Collector** is displayed as **Healthy**.

- In Sumo Logic, the **Health** column uses **color-coded indicators** to show the current state:
  - □ **Healthy** – Collector and Sources are running normally.
  - □ **Warning** – Some issues may be affecting data ingestion.
  - ● **Error** – Collector or Source has failed to collect data.

This allows you to quickly assess the operational status of all collectors in one view.

---

b. Type of Collector

- The **Type** column shows that the **Labs-Apache Collector** is a **Hosted Collector**.
- This means the collector is managed within the **Sumo Logic Cloud**, rather than installed locally on a server.
- Hosted Collectors are ideal for collecting data from **cloud services, APIs, or remote sources**.

---

c. Sources Configured

- The **Labs-Apache Collector** has **3 sources** configured under it.
- Each **Source** defines *what kind of data* is being collected — for example:
  - Apache access logs
  - Apache error logs
  - System or custom logs
- The **Status** of each source indicates whether it's active, paused, or disabled.

---

d. Source Category

- The **SourceCategory** for this collector is configured as **`labs/apache/access`**.
- The **Source Category** is an important metadata tag that helps organize and search data within Sumo Logic.
- Analysts can use this tag in queries, for example:
- `_sourceCategory=labs/apache/access`

  to retrieve all logs collected by this Apache source.

---

e. Log Message Graph

- The **Last Hour** column displays a **graph** showing the **number of log messages ingested per minute** over the past hour.
- This visual metric helps analysts confirm that data ingestion is continuous and within expected rates.

f. Log Message Count

- The final column, **Messages**, displays the **total number of log messages ingested in the past hour**.
- In this example, the **Labs-Apache Collector** shows **20,347 log messages received**, indicating successful and active data collection.

3.2 Summary

The table below summarizes the observed properties of the **Labs-Apache Collector**:

Attribute	Observation
Collector Name	Labs-Apache
Health Status	Healthy
Collector Type	Hosted
Number of Sources	3
Source Category	labs/apache/access
Last Hour Graph	Shows continuous ingestion
Messages Received	20,347 log messages in the past hour

✅ **Conclusion:**
The **Labs-Apache Hosted Collector** is functioning properly, successfully ingesting Apache web server logs in real time. The system health indicators and message graph confirm that data collection and transmission to Sumo Logic Cloud are active and reliable.