Understanding latency and throughput in UBA and Splunk Cloud environments

Interconnected on-premises Splunk User Behavior Analytics (UBA) and Splunk Cloud Platform deployments require additional planning to account for increased network latency, more effectively distributed resource utilization across nodes, and reduced query times.

You can contact your Splunk team for an in-depth analysis of your specific environment.

Fundamental to any UBA deployment, whether on-premises or in Splunk Cloud Platform is a regular copy of events from Cloud to UBA. However, Cloud deployments typically have higher network latency than on-premises deployments. The focus of this guide is minimizing and mitigating the impact of this increased latency.

Sources and symptoms of latency

The event copy has two significant sources of latency - the query to identify events, and the network transfer.

\[
\text{Total latency} = \text{query latency} + \text{network copy latency}
\]

Under ideal conditions, an event should be available in a UBA index within 2 minutes of being ingested by Splunk Cloud Platform. This rapid ingestion and indexing allows you to quickly respond to threats and anomalous behavior. If indexing is taking longer than 2 minutes, latency may be a contributing factor.

Root cause of indexing latency

Data is transferred between UBA and Splunk on a 60-second schedule. If the previous transfer is still running when the next transfer is scheduled to start, the second transfer is delayed.

High latency increases the likelihood of delayed transfers, and delayed transfers cause indexing delays. Contiguous delays then compound each other.
Because each data source has its own transfer process, it’s possible for some data sources to be delayed while others are not. If the rate of inbound events drops, it’s possible for the delay to be reduced or eliminated as transfers catch up.

**Understanding network latency**

High network latency between UBA and Splunk Cloud Platform is the most significant source of indexing delays.

Network latency is a measure of the delay required to establish a connection between hosts. Latency is typically measured as round trip time (RTT) in milliseconds (ms).

Latency increases as physical distance between hosts increases. In-line network devices like switches, routers, firewalls, WAFs, DDOS protection devices and services can also increase latency. Latency is the primary driver of per-connection throughput.

Splunk Cloud Platform maintains a single connection to UBA for each data source. The number of supported data sources increases with the number of nodes in a UBA deployment. See [What size cluster do I need?](#) for more details.

It’s important to understand that single connection throughput is the maximum throughput for a single UBA data source.

---

The simple formula for calculating per-connection throughput is:

\[
\text{TCP window size (KiB) / RTT (\mu s)}
\]

---

**Example**

\[
\begin{align*}
\text{TCP window size} & = 64 \text{ KiB} \\
\text{Latency} & = 70 \text{ ms} / 1000 = 0.07 \mu s \\
\text{Single connection throughput} & = 64 / 0.07 = 914 \text{ KiB/sec}
\end{align*}
\]

This example is based on a hypothetical deployment of UBA on the West Coast of the USA and Splunk Cloud Platform on the East. Here a single data source that can use up to 914 KiB/sec. Using this throughput you can determine the maximum possible throughput between Cloud and UBA.

<table>
<thead>
<tr>
<th>Number of UBA Nodes</th>
<th>Supported Data Sources</th>
<th>Maximum Potential Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>(914 KiB * 6) = 5484 KiB/sec</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>(914 KiB * 10) = 9140 KiB/sec</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>(914 KiB * 12) = 10968 KiB/sec</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>(914 KiB * 24) = 21936 KiB/sec</td>
</tr>
</tbody>
</table>

---

The information provided in Splunk Lantern is intended for informational and educational purposes only. All information is provided in good faith, however, Splunk disclaims any and all representations and warranties, express and implied, regarding the information provided, including without limitation any warranties and representations regarding the completeness, adequacy or accuracy of the information. You agree to take full responsibility for the results arising from the use of the information provided.
Required throughput for a data source

The minimum throughput required for a data source is a function of event size and event frequency.

The required throughput for a single data source = \( \frac{\text{event size (bytes)}}{1000} \times \text{count of events/second} \).

---

**Estimating the required throughput for a data source**

- Average event size = 425 bytes
- Average events per second = 2000

\[
\left( \frac{425}{1000} \right) \times 2000 = 850 \text{ KiB/sec}
\]

For existing deployments, you can use this query on Splunk Cloud Platform to get the EPS statistics for the last 30 days:

```splunk
| tstats count AS eps where index=* earliest=-30d@d groupby index, sourcetype _time span=1s
| stats count AS NumSeconds max(eps) perc99(eps) perc90(eps) avg(eps) AS avg_eps BY index, sourcetype
| addinfo
| eval PercentageOfTimeWithData = NumSeconds / (info_max_time - info_min_time)
| fields - NumSeconds info*
| eval EffectiveAverage = avg_eps * PercentageOfTimeWithData
| fieldformat PercentageOfTimeWithData = round(PercentageOfTimeWithData*100,2) . "%"
```

---

**Query latency**

The first step of the event copy is a query that returns the events to be copied.

Query runtime is the second most significant contributor to delayed transfers. Query latency is no higher for Splunk Cloud Platform deployments than on-premises, however any reductions made to query latency can be used to offset increased network latency.

Using Splunk UBA Kafka ingestion reduces UBA query runtime by at least 50% and up to 75%. You can read more about Kafka ingestion later on in this guide, or jump straight to that section.

---

**Monitoring data source latency**

You can monitor the delay for each data source with a query:
If any data source has a sustained delay greater than 60 seconds then check our recommendations for help on remediating this.

You should also set up an alert for continuous monitoring, see Configure email alerts to your Splunk UBA deployment administrators for details on how to do this.