RATE LIMITING

In 2019, Metrc API system performance was impacted in several states by a surge in calls from Third Party Integrators (TPI). To address this problem, Metrc collaborated with high-volume licensees to improve integration, reconfigured several architectural components, and upgraded internal system resources.

The corrective efforts were successful in resolving the immediate issue, but to further improve system performance and the long-term stability of Metrc's API, we implemented a rate limiting solution.

Rate limiting is a best practice used by software platforms worldwide, including Twitter, LinkedIn, Slack, Quick Books, and Google. It effectively limits the number of API calls licensees can make in a given amount of time. This ensures no single user or company can overwhelm the system. Metrc's rate limiting approach was phased in to give integrators ample time to adjust their systems.

The changes were implemented on the following timetable.

RATE LIMIT DETAILS:

First Rate Limit

- Go-Live Date: August 25, 2020 (Excluding CA) / October 12, 2020 (CA)
- Estimated Impact: 24,000 calls (3% of total) from eight integrators (1.5% of total), based on call monitoring since 8/7/2020
- Rate Limits (all implemented simultaneously):
  - 300 GET calls per second per facility,
  - 900 GET calls per second per integrator key,
  - 10 concurrent GET calls per facility, and
  - 30 concurrent GET calls per integrator
- Total Allowable Daily Calls per Integrator:
  - 77.76 million per day per integrator key
- Non-Affected Calls: Rate limits do not apply to PUT, POST, or DELETE calls.

Second Rate Limit

- Go-Live Date: October 12, 2020 (Excluding CA) / January 25, 2021 (CA)
- Rate Limits (all implemented simultaneously):
  - 100 GET calls per second per facility,
  - 300 GET calls per second per integrator key,
  - 10 concurrent GET calls per facility, and
  - 30 concurrent GET calls per integrator
- Total Allowable Daily Calls per Integrator:
  - 25.92 million per day per integrator key
- Non-Affected Calls: Rate limits do not apply to PUT, POST, or DELETE calls.

Third Rate Limit

- Go-Live Date: January 21, 2021 / Date TBD for CA
- Rate Limits (all implemented simultaneously):
  - 50 GET calls per second per facility,
  - 150 GET calls per second per integrator key,
  - 10 concurrent GET calls per facility, and
  - 30 concurrent GET calls per integrator
- Total Allowable Daily Calls per Integrator:
  - 12.96 million per day per integrator key
- Non-Affected Calls: Rate limits do not apply to PUT, POST, or DELETE calls.

TECHNICAL INFORMATION

All requests denied by rate limiting will contain the following:

- HTTP Status: 429 Too Many Requests
- HTTP Header: Retry-After <wait-period-in-seconds>

All requests denied after reaching the maximum concurrent calls will contain the following:

- HTTP Status: 429 Too Many Requests

See more information about this HTTP status code here

All Metrc sandboxes have been updated with the Rate Limits that match the production environments

FAQS

How did Metrc arrive at these specific limits?

- We first looked at the best practices for administration of APIs to determine the type and scale of rate limits. We looked at a variety of leading platforms' API policies, including Twitter, Quick Books, and Google. Rate limits in these platforms vary in implementation, but, for example, Twitter has rate limits of 15 calls and 180 calls per 15 minutes, depending on the type of calls – which would be extremely low and ineffective for integration with Metrc. Others implement a quota system with varying weights per endpoint.

- We then analyzed integrator traffic on our own API to determine what a realistic rate limit could be for them – and we believed would help stabilize our system performance. For example, some integrators were making, on average, over 15,000 calls per minute, causing serious strain on our servers.

- Lastly, we determined a target rate limit that would balance these two priorities: improved system performance and the ability for integrators to realistically meet it over the course of 90 days. We then defined three incremental milestones that integrators would need to hit over time. These milestones represent gradually decreasing rate limits that are designed to help integrators update their platforms over time and ultimately comply with the target rate limit: 50 calls per second per facility. The schedule for the first round of rate limiting can be found under Rate Limit Details towards the top of this page.

The Maryland instance of Metrc already has a rate limit implemented; will the new rate limits affect it?

- Yes. The new rate limits detailed above will replace the current rate limits in Maryland.

We think there are a few things that Metrc can do to help integrators lower the number of API calls we need to make. Is Metrc considering this?

- Yes. But right now, we are focusing our resources on implementing this first round of rate limiting over the next 90 days. Following that, we may will continue to work closely with the TPI community to evaluate and prioritize additional development work that may help integrators optimize their API calls, such as returning newly created IDs.

What do you mean by, "implemented simultaneously"? For example, if I am an integrator with 5 facilities as customers, am I limited to 900 GET calls per second or am I also getting 300 calls per second per facility yielding either 1500 GET calls or maybe even 2400 GET calls (5*300 for facilities + 900 for integrator)?

- There is an upper limit per Integrator of 900 GET calls per second currently, independent of the number of facilities. This will reduce to 150 GET calls per second per integrator in the third rate limit.

Is the concurrent GET calls per TPI a per state rate?

- Yes, all rate limits are independently tracked within each State's instance of Metrc.

Are the limits per API endpoint? Or are these inclusive to all GET endpoints?

- They are inclusive of all GET endpoints.

Is the "per facility" rate limit shared across integrators? Will I be affected by another integrator's usage at that facility?

- No, it's a facility limit that is specific to each individual integrator AND their facility.

Is there any plan in the next 6-12 months to rate limit anything else outside of GET requests?

- Not yet. But any additional restrictions would be based on what we see over the next few months and how the system is utilized. The key reason we are limiting GET calls is because it currently has the biggest performance impact on Metrc.

Are there any plans to build up any other part of API to help mitigate rate limits, such as webhooks?

- Yes, Metrc has a list of requests that we are prioritizing, such as webhooks and returning objects.