Incident Report

Aug 3, 2023 : Degraded Performance on Master Data Search

Status Page URL	https://status.vtex.com/incidents/z5965y02gqqw
Impacted accounts	Stores that use Master Data for critical operations
Duration	12h and 37 minutes
Availability	Stores relying on Master Data to customize their sales flow (e.g. B2B Suite) had a major impact on sales. Others experienced delays in Master Data indexing and triggers.

Summary

On August 3, 2023, starting at 11:50 UTC, our monitoring systems identified a partial degradation of one sharded cluster[1] of Master Data, affecting ~30% of accounts. The mitigation process had side-effects on background operations for all accounts starting at 14:00 UTC. On August 4, 2023, at 00:27 UTC, the degraded cluster was recovered.

Symptoms

1. Errors in Search API caused by partial degradation of one sharded cluster

From 11:50 to 22:31 UTC, our monitoring systems identified an increase in errors to a group of stores. This resulted in customers experiencing difficulties while making requests to the Master Data Search API. Only stores using this cluster were affected.

2. Delays in background operations for indexing and triggers

Our mitigation process delayed some background operations to all accounts. The creation and update of documents took more time than expected to be available in the indexer and triggers execution from 14:00 UTC to 00:27 UTC. All stores were affected.

---

[1] Sharding is a strategy for distributing data across multiple databases, which can then be stored on multiple servers.

Timeline

[2023-08-03 11:50:00 UTC]	One of the sharded clusters became unavailable.
[2023-08-03 12:16:00 UTC]	Our incident response team was alerted and started investigating the issue.
[2023-08-03 13:18:00 UTC]	We identified the root cause of the issue and started performing mitigation actions.
[2023-08-03 14:00:00 UTC]	The incident response team decided to turn off indexing for all clusters, as part of our mitigation actions. As a side-effect, all accounts started experiencing delays in background operations for indexing and triggers.
[2023-08-03 22:31:00 UTC]	Impacted cluster shard was recovered. Indexing begins to recover for impacted customers.
[2023-08-03 00:27:00 UTC]	Indexing delay was fully resolved.

Mitigation strategy

We reestablished normal operations of the platform after (1) reducing the load to the affected cluster. This action caused a delay to background operations. (2) We identified the root cause that impacted the cluster and, based on this diagnosis, scaled the cluster environment to handle production load. (3) We enabled the new cluster in production with reduced load, with background processes still disabled. (4) Finally, we gradually started to process background operations until they were back to normal.

It is important to note that, during the incident, the delays in indexing and execution of triggers were an unavoidable consequence of our mitigation actions. Until this incident, we had successfully operated with a single indexing queue, shared across all sharded clusters – similar to what you would see in a supermarket with a single queue and multiple registers. While we were already working on splitting indexing in multiple queues, this upgrade was not yet production-ready.

Therefore, we had to turn off indexing for all clusters before we could restore the degraded cluster. In our follow-up actions, we will address this so that any outages can be contained within the affected clusters, instead of causing side-effects on all clusters.

Follow-up actions: preventing future failures

Moving forward, we are taking the following actions to prevent this issue to happen and speed up our response to reduce the impact that we had if something similar ever happens again:

- Apply the same configuration to prevent the issue from happening again to all Master Data indexing clusters.
- Improve the Master Data Workers architecture to reduce the impact to all stores when we have a delay to process messages by splitting indexing in multiple queues.
- Review our reindex mechanism to recover stores faster when migrating their data to new healthy clusters.