

Incident Report

Jul 9, 2024 : Degraded performance for checkout operations

Status Page URL	https://status.vtex.com/incidents/01J2CBD1MRWQTA4M4VD3HE2Z3D
Impacted accounts	All accounts
Duration	Period 1: 78 minutes Period 2: 91 minutes
Availability	Sales flow completely interrupted

Summary

On July 09, 2024 our inventory database experienced two periods of unavailability. These outages occurred from 17:57 UTC to 19:15 UTC and from 19:29 UTC to 21:02 UTC. During these times, users might have encountered issues with navigation, search, and checkout, depending on the store implementation.

Symptoms

Our automated monitoring system identified that VTEX's order flow experienced high levels of errors due to a failure in our Inventory highly available database system. This resulted in two periods of blockage in the global sales flow, with shoppers experiencing failures while placing their orders. This system is accessed by multiple internal modules and its unavailability resulted in various errors in different integrations and checkout solutions.

After the full recovery of the affected system, there were lingering side-effects on merchant experience as order indexing and inventory related requests were queued while the database system was unavailable. With queues going back to expected levels during the next few hours.

Timeline

[2024-07-09 17:57 UTC]	Our incident response team was notified of increased latency in the inventory database.
[2024-07-09 18:00 UTC]	Recently updated applications were rolled back to previous versions.
[2024-07-09 18:21 UTC]	The team scaled the application, restarted the distributed cache and restarted the distributed database system, with no effect on error rates.
[2024-07-09 18:43 UTC]	Given that our fast mitigation procedure didn't yield positive results, all connections to the inventory system were dropped and the team initiated automated creation of a new environment to gradually move traffic to.
[2024-07-09 18:53 UTC]	1% of traffic was directed to the new environment and the team was able to observe successful connections to the database.
[2024-07-09 18:58 UTC]	The team started to gradually route more traffic to the new environment in order to cache resource-intensive queries.
[2023-07-09 19:15 UTC]	Our mitigation strategy fully recovered the sales flow. The response team began to restart stopped applications connecting to the database.
[2023-07-09 19:29 UTC]	Our automated notification systems alerted us to another increase in errors in the inventory database.
[2023-07-09 19:42 UTC]	The response team decided to recreate the database environment with increased capacity.
[2023-07-09 20:30 UTC]	The team started to gradually route traffic to the new environment in order to cache resource-intensive queries.
[2023-07-09 21:02 UTC]	Our incident response team confirmed sales flow was fully reestablished.

Mitigation strategy

Our preliminary investigation suggests an increase in IO operations correlated to usual infrastructure movements, causing resource starvation. The programmed automated response to this sudden increase in load resulted in blocked internal connections to the service, even after our automatic recovery system was activated.

Observing the sudden increase in connection errors, the response team opted for a mitigation strategy that required blocking traffic and a full restart of the database cluster with a gradual increase in traffic to guarantee cache integrity.

As the response team quickly acted on the decided strategy and traffic was gradually sent to the restarted application, they still had not been able to pinpoint the source of the problem and restarting part of the applications resulted in another spike in failures. This caused the second checkout outage described.

Armed with a better understanding of the source of failures, our team was able to repeat the mitigation strategy, leaving suspected applications out of the procedure, which recovered the sales flow back to its expected state.

Follow-up actions: preventing future failures

Multiple teams are working together with our cloud providers to fully understand the source of the sudden increase in errors, as well as why our previous automated procedures were not able to mitigate the problem. In parallel to this, we have scaled the capabilities of the whole application to guarantee resilience and increased observability to the affected database so similar behavior can be detected before it can affect the sales flow. Our engineering team is also working on improving our testing procedures to include testing for similar situations in our development pipeline. We'll also further shard, as in separate with stronger boundaries, our highly available databases and applications to ensure future incidents are kept contained to a smaller subset of shopper sessions.

The VTEX team recognizes the impact of this incident on our ecosystem and apologizes for any effect it might have had. Our focus is on maintaining the trust of our community and improving the most resilient platform for digital commerce.