Incident 04 Apr. 2021: Degraded sales due to payments Gateway outage

Availability: **around 20% of orders lost**

% of users affected: **All stores**

Duration of incident: **1 hour and 35 minutes**

Symptom

At 8:25 PM UTC, our monitoring services detected anomalies in our checkout system and we started working to mitigate the impact on sales. We detected a degraded event system used in our Payments Gateway, causing some payment authorizations to be delayed. Post-purchase operations were affected since events related to cancellations, settlements and status updates were also delayed in the event system.

Summary

The issue started after one of the services responsible for managing events hit 100% capacity, as shown in the internal dashboard. This resulted in events that were sent to this part of the system, to not be processed at the expected rate.

That caused the checkout connection to timeout and some sales were lost, accounting for around 20% of all orders.

Timeline

[2021-04-04 20:25 UTC] Our Payments Gateway event system had a partial outage.

[2021-04-04 22:00 UTC] The limit of the event system was increased so it wouldn't remain at 100% capacity, and a fallback system was implemented.

[2021-04-04 22:02 UTC] The event system load was relieved, leading metrics to return to normal and ultimately fixing the issue.

[2021-04-05 23:30 UTC] All incomplete orders and pending status updates waiting in queue were processed.

Mitigation Strategy

At 22:00 UTC, we redirected traffic from the degraded event system to others that were working properly. This mitigated the direct impact on sales and success rate for orders events was re-established.

However, some events got stuck in the unhealthy system. We therefore started working to recover them. The cause behind the issue was the high usage of disk resources reserved for specific operations. This led to metrics showing the host as healthy, but the service itself was not properly processing the events. We contacted our cloud provider to increase pre-configured limits for this system. As soon as this configuration was applied and the service was restarted, all the delayed events were properly processed. At 23:30 UTC all order related events were up-to-date.

Follow up

We recognize that our alert systems were not sensitive enough to help us quickly identify the system integrations' health, which directly impacted our customer's performance and sales. To prevent future instabilities, we have configured new alarms and anomaly detection measures to prevent the load from reaching critical limits before taking action.

We will also work on improving our load balancing algorithm to identify when one of the services is not properly working and redirect traffic automatically to other healthy systems.