Incident Jan. 21, 2022: Elevated Errors in OMS

Availability: **Order listing was unavailable**

% of clients affected: **80%**

Incident duration: **123 minutes**

Symptoms

From 13h08 to 15h11 UTC, most shopkeepers would receive errors while on the order listing page. Sales were not affected.

Summary

We reported an increase of errors in OMS APIs to list orders. This was later identified as degradation in some database clusters that started to present elevated processing time and latency. We started implementing measures to reduce the pressure on the clusters, such as blocking some APIs and scaling up the clusters. Some of these measures took very long due to the database size. All of the clusters recovered but one. To fix it, we provisioned a failover database cluster and started to split traffic between both.

Timeline

[2022-01-21 13:08 UTC]. Order listing alert was triggered and the investigation started.

[2022-01-21 13:16 UTC]. The issue was identified in our database and we started taking measures to fix it.

[2022-01-21 13:22 UTC]. The problem was partially fixed, but 40% of the accounts still had issues.

[2022-01-21 13:40 UTC]. We started recycling the remaining degraded clusters.

[2022-01-21 13:46 UTC]. The number of clients with order listing errors decreased to 20%.

[2022-01-21 14:00 UTC]. We provisioned our failover infra to support the degraded cluster.

[2022-01-21 14:50 UTC]. Traffic was split to use the failover infrastructure.

[2022-01-21 15:11 UTC]. Indexing and order listing went back to normal.

Mitigation strategy

Our alarms showed that our OMS cluster was not running smoothly and then noticed that the servers were showing high CPU usage percentages. The failover cluster was provisioned to reduce the workload by splitting the traffic to help the degraded cluster get healthy again. Once we were sure the degraded cluster was fixed, we moved the entire traffic back to it.

Follow-up actions: preventing future failures

As follow-ups to this incident, we will work on improving our alarms so that we detect service degradations faster. We will also improve our database resilience and recovery time in order to reduce the total impact.