CASE STUDY

Improving the performance of a conversational AI platform with Advance dedicated servers

OVH
Innovation for Freedom

sentione

120+
physical Elasticsearch nodes

70TB+
of migrated data

x6
more performance

EXECUTIVE SUMMARY

SentiOne, recognised by Deloitte as one of the fastest-growing technology companies in Central Europe, delivers a unique conversational AI platform that allows you to monitor online discussions, engage with your audience and automate customer service across all web channels.

While most AI companies that provide chatbots struggle with training resources, SentiOne uses real-life internet discussions to train their NLU (Natural Language Understanding) engine. Thanks to a vast dataset, collected while developing an in-depth social listening tool, SentiOne is able to train exceptionally accurate deep learning engines.

> *"Over the years of our collaboration with OVH, we've been able to scale our business smoothly and process terabytes of data, giving quality insights to our customers."*
>
> **Michał Brzezicki, Co-Founder, Chief Technical Officer at SentiOne**

With a growing amount of data to gather and analyse in real time, the company has faced many infrastructure and software challenges. For one, they have experienced degradation of performance with their Elasticsearch cluster, which is crucial for a smooth customer experience with the platform. Furthermore, SentiOne also noticed disproportionate usage of certain clusters, which led to some of them overloading, while others remained idle.

To improve the speed of their conversational AI platform, SentiOne undertook a huge maintenance plan to migrate the hardware and update the software at the same time. Michał Brzezicki, Co-Founder and CTO at SentiOne, spoke to us to share the details of how this complex project was managed by his team.

THE CHALLENGE

SentiOne's team of IT experts listed several potential reasons that could lead to the Elasticsearch cluster's degradation, including insufficient CPU power, underperforming storage, hardware failures, software configuration, and connectivity and network issues.

Network

Elasticsearch is very prone to connectivity issues. When a network instability occurs between two nodes in the cluster, the whole cluster slows down. To maintain reliable and secure connections between the nodes, SentiOne decided to use vRack: a private network solution, designed by OVH.

The network issues were almost permanently solved once this solution was enabled, as vRack allowed SentiOne to create stable connections between nodes, independent from the public network.

To gain a deeper understanding of their infrastructure's health, SentiOne's team uses Grafana with Sensu plugin to store all system and Elasticsearch status values. With both public and vRack interfaces, no anomalies were recorded, so the team ruled out the network as a source of performance issues.

CPU

While further investigating the potential causes of performance issues, the team looked into CPU usage and thread management. The general CPU usage in the cluster was approaching 80%, which was considered appropriate, as it indicated a proper utilisation of resources in terms of costs. However, the team noticed that the cluster was noticeably slower during rebalancing.

During slowdowns, the CPU usage was low and restricted to just one core. A deeper analysis indicated that most of the CPU time was spent on IO operations. As the team had already excluded the network as a cause, they therefore started to investigate the storage.

Storage

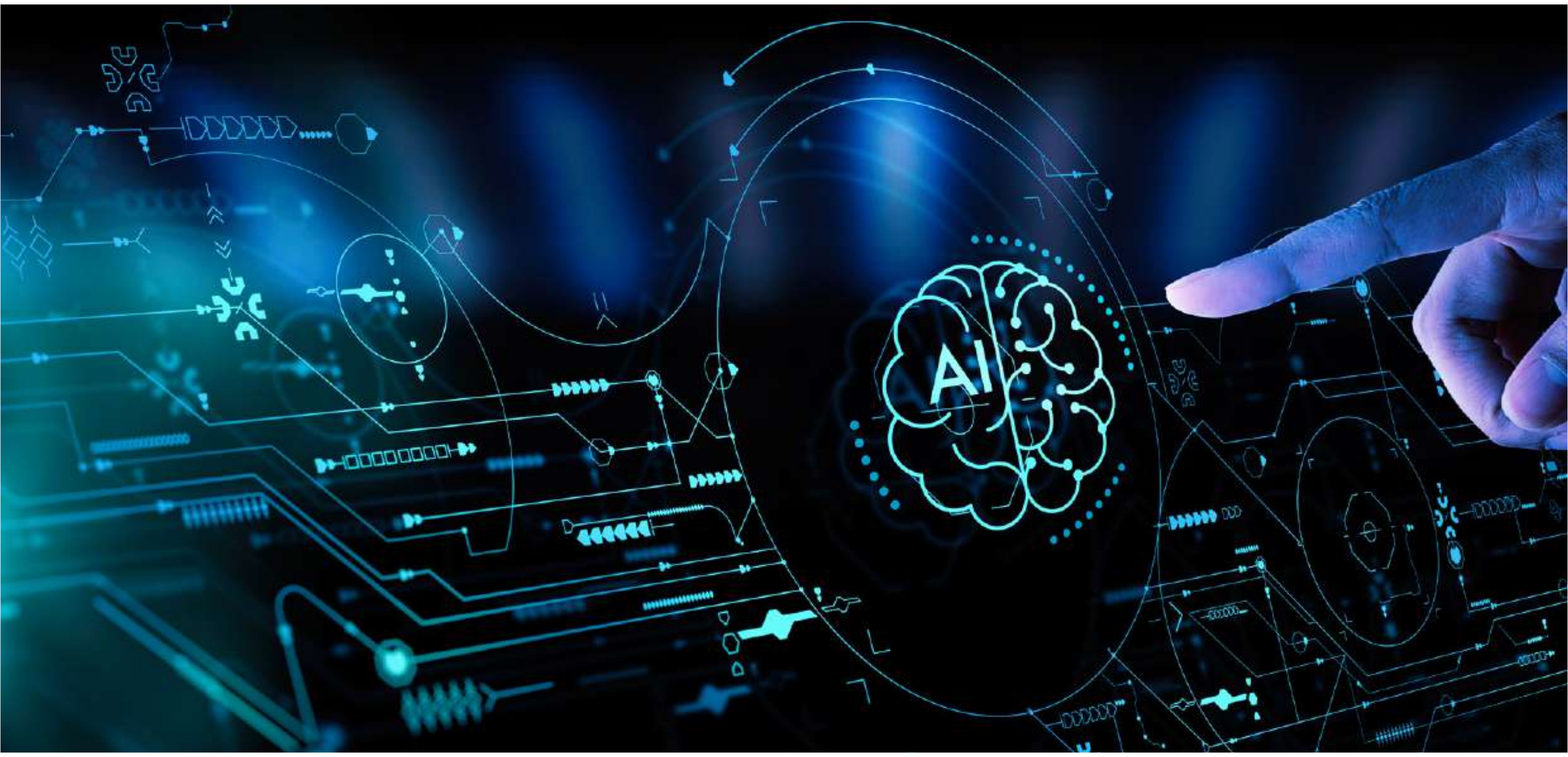
Creating a Grafana dashboard for the IO Time metric from Sensu provided SentiOne's team with the answers. The resulting graph clearly showed how much time CPU spent on IO operations in a given period. It turned out that this figure could be almost 100%!

This also explained why the cluster was slowing down during rebalancing. If a random node from which they copied a shard was under high IO load, copying the shard would only make things worse.

Elasticsearch cluster configuration

SentiOne's infrastructure was based on "hot and cold" nodes. The hot nodes hosted popular, regularly-queried indices, while the cold ones contained rarely-accessed data. Though adding more replicas for hot indices to balance the load seemed like an obvious solution, this proved not to be the case. Firstly, queries were assigned to shards randomly, so if a heavy query was repeated multiple times (e.g. different aggregations of the same query), then all replicas were overloaded with similar or identical computations. Additionally, the Elasticsearch version installed on SentiOne's cluster at the time didn't provide Adaptive Replica Selection (ARS). It became clear that adding new nodes to the cluster would not scale the performance unless the team updated the software.

Having identified the bottlenecks that had caused the degraded performance of the Elasticsearch cluster, the IT team could start to plan software, architecture and hardware changes to enhance the responsiveness of the platform.



THE SOLUTION

Replacing the hardware and scaling clusters horizontally seemed like the most obvious option, but due to having a big but limited budget, SentiOne started by sanitising the cluster usage.

Before the team was ready to migrate to a new dedicated server infrastructure, there were several other actions to perform.

Smarter data partitioning

First, SentiOne engineers analysed user queries and how they affected the cluster performance. The relationship between query time and a number of queried shards and their size is straightforward. The bigger the shards, the longer each query takes, and when more shards are queried, the longer the overall process takes.

To improve the query time, SentiOne decided to divide the data into smaller shards, no bigger than 30GB. Once queried, the data of a single shard could fit into the file system cache in RAM, at least in theory.

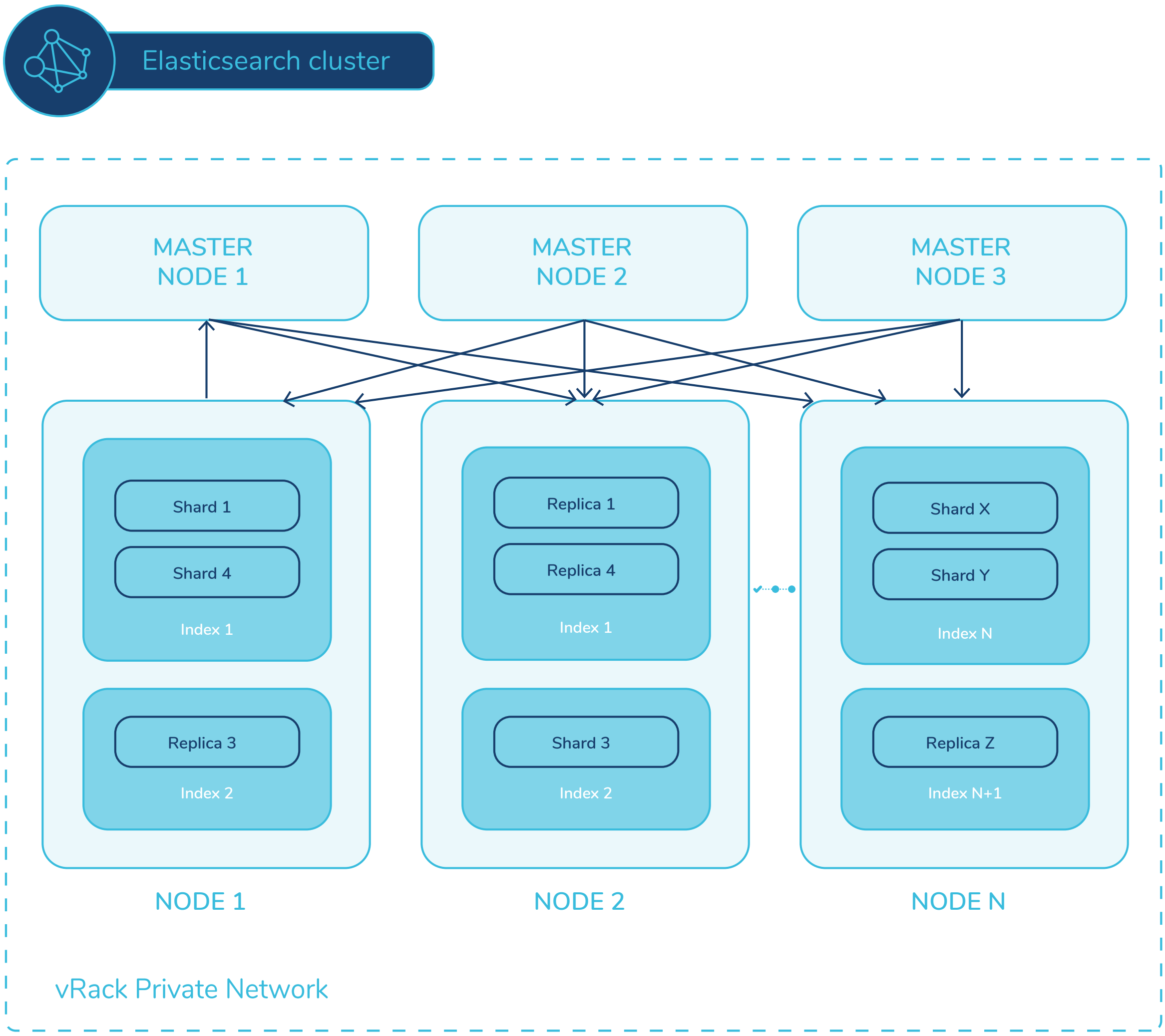
After investigating which timeframes are queried by users, SentiOne has also changed the way data is split into indices. They divided data according to customers' behaviour, minimising the number of shards per query, and shortened the span from monthly to weekly indices.

Application

SentiOne's team revised the points in which the Elasticsearch cluster is queried, optimised calls and changed the UI, so users would not have the impression of the system "freezing". They also introduced checks that would block excessively heavy queries, based on execution times within a moving timeframe. Historically, a single heavy query could degrade the performance of the whole cluster, so this safety check was a necessity, even though it would limit some users.

Architecture changes

With the new version of the cluster, SentiOne abandoned the “hot and cold” approach for several reasons. First of all, it is difficult to make a clear differentiation between hot and cold indices. At which point can you call an index “hot”? Second, the cold nodes had very low IO usage with spikes of CPU usage, while the hot nodes were overloaded by the IO but had unconsumed CPU resources. This was an obvious waste of resources and, in turn, a waste of money. Finally, finding the right balance of hot and cold nodes proved to be a tricky task, leading to more unnecessary costs.



New hardware

The previous infrastructure consisted of three master nodes, 82 hot nodes, and 42 cold nodes, all equipped with the same storage solution: two Intel® SSD DC S4500 drives. As the main goal of the hardware migration was to improve storage performance, SentiOne selected Advance range servers – specifically ADV-2 models with NVMe drives – for their new data nodes.

NVMe drive technology is a combination of solid-state drives, PCIe buses, and NVMe protocol, designed to bridge the performance gap between compute and storage.

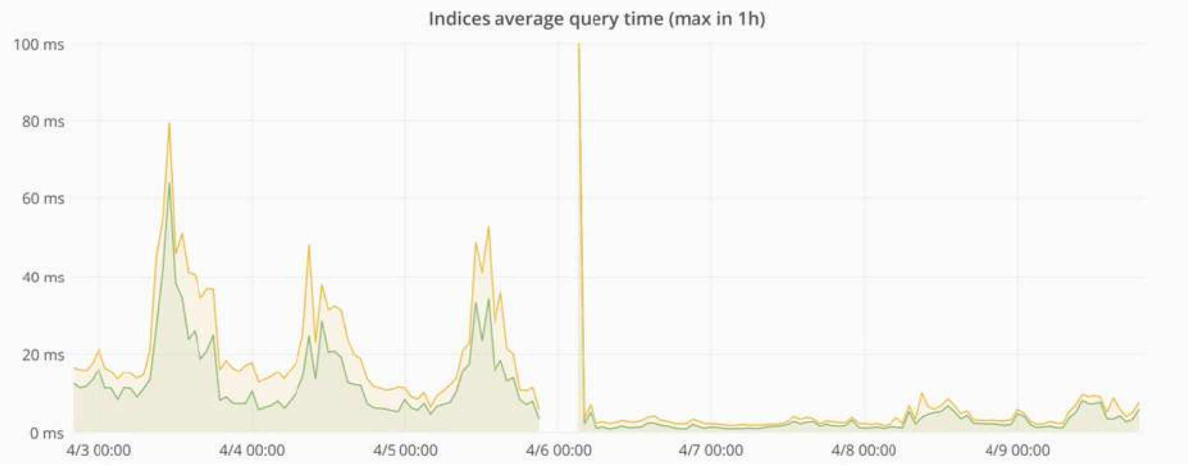
The interface between the SSDs can significantly influence the total latency the user will experience at the application level, while two elements — the physical interface (PCIe) and the protocol (NVMe) — determine the overall performance. A PCIe 3.0 x4 link offers almost five times more throughput than the fastest SATA, and nearly three times more than the best SAS interface. The NVMe protocol offers up to 64 thousand queues, ensuring the interface is capable of coping with the high number of I/O threads that are generated by multicore CPUs.

To deploy the new cluster of 121 data nodes and three master nodes, SentiOne spent two months developing, testing and planning. Though the migration and Elasticsearch upgrade was not a trouble-free process, and resulted in six hours of cluster downtime, it nonetheless delivered the expected benefits

THE RESULT

SentiOne's omnichannel platform monitors billions of discussions on thousands of web sources. To provide enterprises with audience insights and AI customer service automation, it collects, processes and analyses massive amounts of data. Thus, storage performance plays a crucial role.

The new cluster, built on ADV-2 dedicated servers and equipped with NVMe drives, was a game changer for SentiOne. The hardware migration has brought an immediate performance boost, in that the cluster is now six times faster.



The average query time of two random indices, with approximately 6x performance boost.

The Elasticsearch version update – which included enabling the ARS feature – has allowed the cluster to work more evenly, and in times of higher load, share it equally across multiple nodes. The team measured that during regular operations the load approaches 50%, and does not exceed 80% at stress, which gives SentiOne a healthy reserve for future scaling and new features.

> *"Budget-wise, the migration process turned out to be very beneficial. Costs of running the cluster have been reduced by about 5% with a significant improvement in performance."*
>
> **Michał Brzezicki, Co-Founder, Chief Technical Officer at SentiOne**

OVH is a global, hyper-scale cloud provider that offers businesses industry-leading performance and value. Founded in 1999, the group manages and maintains 28 datacentres across four continents, deploys their own fibre-optic global network and controls the entire hosting chain. Relying on their own infrastructures, OVH offers simple and powerful solutions and tools that put technology at the service of business, and revolutionise the way that our more than one million customers around the world work. Respect for individuals, freedom and equal opportunities for access to new technology have always been firmly rooted principles of the company. “Innovation for Freedom”.

OVH
Innovation for Freedom

ovh.com OVH ovhcom OVH