UN CASO DE ÉXITO

SentiOne utiliza servidores Advance para optimizar el rendimiento de su plataforma de conversación basada en IA

OVH
Innovation for Freedom

sentione

+120
nodos Elasticsearch físicos

+70 TB
datos migrados

x6
mejora del rendimiento

RESUMEN

SentiOne, reconocida por Deloitte como una de las empresas tecnológicas con más rápido crecimiento de toda Centroeuropa, ofrece una plataforma de conversación basada en inteligencia artificial que permite monitorizar conversaciones online, interactuar con la audiencia y automatizar el servicio de atención al cliente en todos los canales digitales.

La mayoría de las empresas de inteligencia artificial que ofrecen chatbots suelen tener dificultades para entrenar estos sistemas. Sin embargo, SentiOne utiliza conversaciones online reales para entrenar su motor de entendimiento del lenguaje natural (NLU, del inglés «natural language understanding»). Gracias al extenso conjunto de datos de que dispone, recopilados durante el desarrollo de su herramienta de escucha social, SentiOne es capaz de entrenar motores de deep learning con una precisión excepcional.

Desde que colaboramos con OVHcloud, hemos podido escalar nuestro negocio fácilmente y procesar terabytes de datos para ofrecer información relevante a nuestros clientes.

Michał Brzezicki, cofundador y CTO de SentiOne

OVH

Debido a la creciente cantidad de datos que deben recopilar y analizar en tiempo real, la empresa tuvo que hacer frente a diversos retos en materia de infraestructura y software. Por un lado, el rendimiento de su cluster Elasticsearch, que es fundamental para ofrecer una experiencia de cliente fluida en la plataforma, se había degradado. Por otro lado, la empresa también había detectado que el uso de los clusters era desigual: mientras que algunos se sobrecargaban, otros permanecían libres.

Para mejorar la velocidad de su plataforma de conversación basada en IA, SentiOne puso en marcha un completo plan de mantenimiento para migrar el hardware y actualizar el software de forma simultánea. Michał Brzezicki, cofundador y CTO de SentiOne, nos ha explicado los detalles de este complejo proyecto gestionado por su equipo.

OVH

EL DESAFÍO

El equipo informático de SentiOne comenzó analizando las distintas causas que podrían estar provocando la degradación del rendimiento del cluster Elasticsearch: potencia insuficiente de la CPU, bajo rendimiento del almacenamiento, fallos de hardware, configuración del software o problemas de conectividad y red.

Red

Elasticsearch es especialmente propenso a los problemas de conectividad: cuando la red que conecta dos nodos es inestable, todo el cluster se ralentiza. Para que las conexiones entre los nodos fueran fiables y seguras, SentiOne decidió utilizar el vRack, la solución de red privada diseñada por OVHcloud.

Una vez activado y configurado el vRack, SentiOne pudo disfrutar de conexiones estables entre los nodos al margen de la red pública, erradicando definitivamente los problemas de red.

Además, el equipo de SentiOne implementó Grafana con el plugin Sensu para almacenar todos los logs del sistema y de Elasticsearch y así conocer con mayor precisión el estado de salud de su infraestructura. Utilizando ambas interfaces, la pública y la del vRack, no se registró ninguna anomalía, por lo que el equipo descartó la red como posible origen de los problemas de rendimiento.

Procesamiento

En su estudio de las posibles causas de los problemas de rendimiento, el equipo analizó el uso de CPU y la gestión de los threads. El uso general de CPU en el cluster se acercaba al 80%, un porcentaje considerado como adecuado, ya que indica que los recursos se están aprovechando correctamente para optimizar su coste. Sin embargo, detectaron que el cluster se ralentizaba considerablemente durante el balanceo.

Cuando se producían esas ralentizaciones, el uso de CPU era muy bajo y se limitaba a un único core. Un análisis más profundo reveló que la mayor parte del tiempo de uso de la CPU correspondía a operaciones de E/S. Como ya habían descartado la red como posible causa, empezaron a investigar el almacenamiento para intentar averiguar la razón.

Storage

El equipo de SentiOne creó entonces un panel Grafana para las métricas de tiempo de E/S provenientes de Sensu. El gráfico resultante mostraba claramente el tiempo que dedicaba la CPU a las operaciones de E/S durante un determinado período de tiempo. ¡Y esta cifra alcanzaba en algunos casos el 100%!

Eso explicaba la ralentización del cluster durante los balanceos: si un nodo aleatorio desde el que se hubiera copiado un fragmento soportaba una carga elevada de E/S, intentar copiar este fragmento empeoraba aún más las cosas.

Configuración del cluster Elasticsearch

La infraestructura de SentiOne se basaba en un sistema de nodos «fríos» y «calientes». Los nodos calientes alojaban los índices más populares, es decir, aquellos que se consultaban con mayor frecuencia, mientras que los fríos contenían los datos a los que se accedía en contadas ocasiones. Aunque añadir más réplicas de los índices calientes para balancear la carga parecía la solución más evidente, en este caso no fue así. En primer lugar, las consultas se asignaban a los fragmentos de forma aleatoria, de modo que si se repetía una consulta pesada en varias ocasiones (p. ej., diferentes agregaciones de la misma consulta), todas las réplicas se sobrecargaban con cálculos similares o idénticos. Además, por aquel entonces, la versión de Elasticsearch instalada en el cluster de SentiOne no ofrecía la funcionalidad Adaptive Replica Selection (ARS). Parecía evidente que añadir nuevos nodos al cluster no permitiría mejorar el rendimiento, a menos que el equipo actualizara el software.

Una vez identificados los cuellos de botella que provocaban la degradación del rendimiento del cluster Elasticsearch, el equipo comenzó a planificar las mejoras de software, arquitectura y hardware que permitirían optimizar la capacidad de respuesta de la plataforma.

LA SOLUCIÓN

Aunque sustituir el hardware y escalar los clusters horizontalmente parecía la opción más evidente, SentiOne empezó por optimizar el uso del cluster, ya que, a pesar de disponer de un presupuesto nada desdeñable, este estaba limitado. Así que optaron por realizar otras acciones antes de migrar a una nueva infraestructura de servidores dedicados.

Una partición de datos más inteligente

Los ingenieros de SentiOne empezaron analizando las consultas de los usuarios y cómo estas afectaban al rendimiento del cluster. Existe una relación directa entre el tiempo de consulta y el número de fragmentos consultados y su tamaño. Cuanto más grandes son los fragmentos, más tiempo requiere cada consulta, y cuantos más fragmentos se consultan, más tiempo dura todo el proceso.

Para reducir el tiempo de consulta, SentiOne decidió dividir los datos en fragmentos más pequeños, no superiores a 30 GB. Una vez consultados, los datos de un único fragmento podrían guardarse en la caché del sistema de archivos en RAM, al menos en teoría.

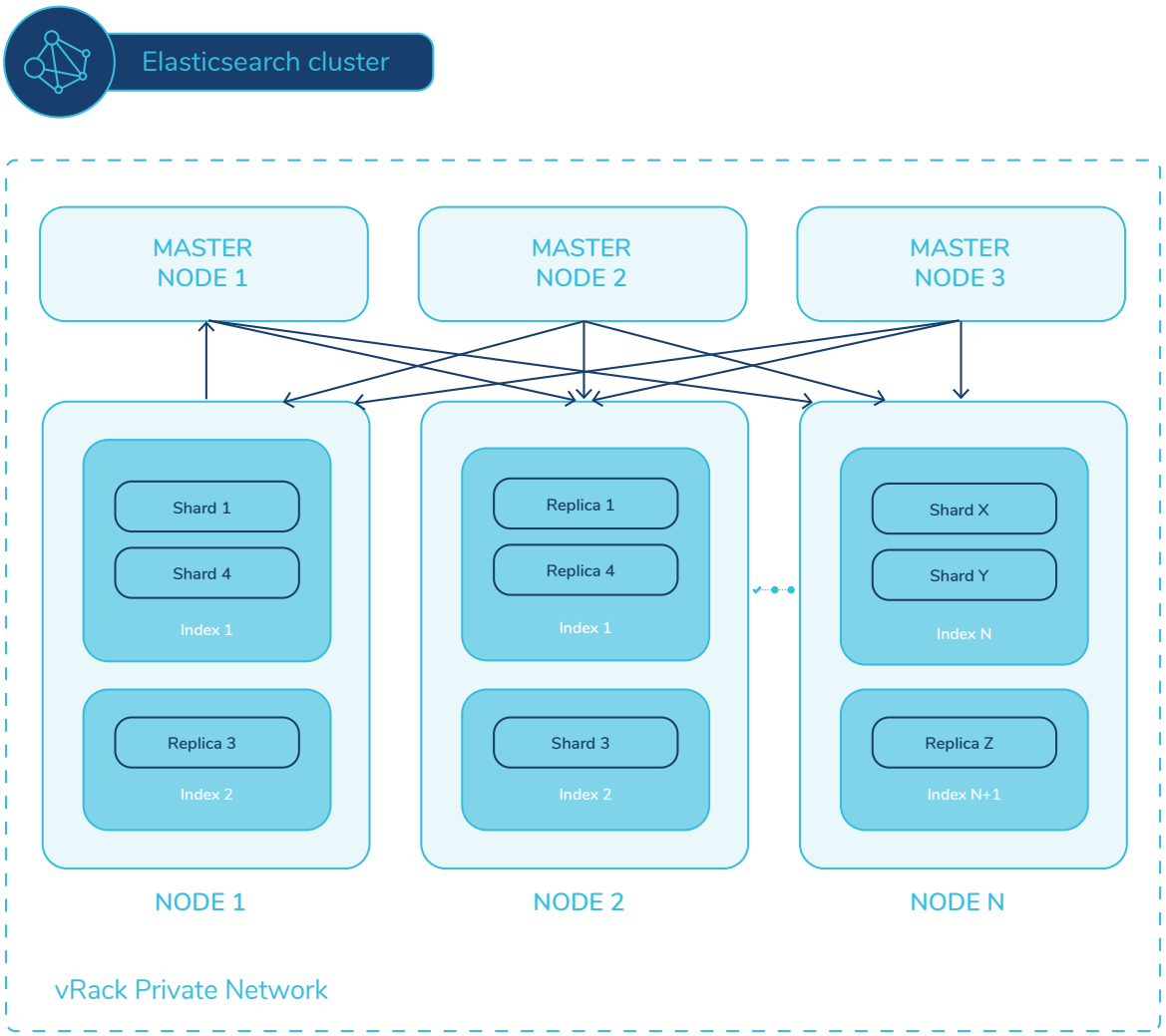
Tras estudiar qué intervalos de tiempo consultaban los usuarios, SentiOne también cambió la forma en la que se indexaban los datos, dividiéndolos en función del comportamiento del usuario para así minimizar el número de fragmentos utilizados por cada consulta, y reduciendo la extensión de los índices de mensual a semanal.

Aplicación

El equipo de SentiOne revisó los puntos en los que se consultaba el cluster de Elasticsearch, optimizó las llamadas y modificó la interfaz para que los usuarios no tuvieran la sensación de que el sistema se quedaba bloqueado. También introdujo una serie de comprobaciones para bloquear las consultas excesivamente pesadas, basándose en los tiempos de ejecución. Hasta entonces, una única consulta muy pesada podía degradar el rendimiento de todo el cluster, por lo que este control de seguridad era necesario, aunque impusiese restricciones a algunos usuarios.

Cambios en la arquitectura

Con la nueva versión del cluster, SentiOne abandonó su enfoque de frío y caliente por diversos motivos. En primer lugar, resultaba difícil distinguir entre índices fríos y calientes: ¿en qué momento un índice debía considerarse caliente? En segundo lugar, los nodos fríos tenían un uso de E/S muy bajo, con picos de carga de CPU, mientras que los nodos calientes estaban sobrecargados por las E/S, pero no consumían todos los recursos de CPU. Al no aprovechar bien los recursos, estaban gastando más dinero del necesario. Por último, conseguir un buen equilibrio entre los nodos fríos y calientes era una tarea compleja, que añadía todavía más costes superfluos.



Nuevo hardware

La arquitectura anterior estaba compuesta por tres nodos maestros, 82 nodos calientes y 42 nodos fríos, todos ellos equipados con la misma solución de almacenamiento: dos discos Intel® SSD DC S4500. El objetivo principal de la migración del hardware era mejorar el rendimiento del almacenamiento, por lo que SentiOne eligió los servidores de la gama Advance —en concreto el modelo Advance-2 con discos NVMe— para sus nuevos nodos de datos.

La tecnología NVMe, que combina combina unidades de estado sólido (SSD), buses PCIe y el protocolo NVMe, ha sido diseñada para reducir la diferencia de rendimiento entre el procesamiento y el almacenamiento.

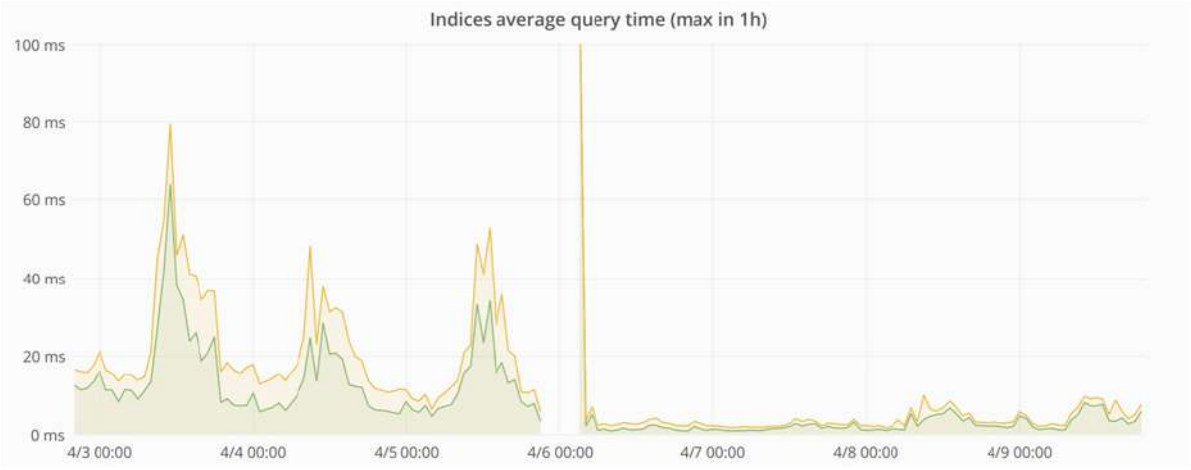
La interfaz entre los SSD puede influir de forma significativa en la latencia total que experimenta el usuario a nivel de la aplicación, mientras que la interfaz física (PCIe) y el protocolo (NVMe) determinan el rendimiento general. Una conexión PCIe 3.0 x4 proporciona una tasa de transferencia casi cinco veces mayor que la interfaz SATA más rápida y casi tres veces mayor que la mejor SAS. El protocolo NVMe ofrece hasta 64 000 colas, garantizando así que la interfaz sea capaz de hacer frente a la gran cantidad de hilos de E/S que pueden generar los procesadores multinúcleo.

Antes de desplegar el nuevo cluster, compuesto por 121 nodos de datos y 3 nodos maestros, SentiOne dedicó dos meses a las fases de desarrollo, testeo y planificación. Aunque el proceso de migración y actualización de Elasticsearch no estuvo exento de problemas (se produjo una interrupción del servicio en el cluster de seis horas), sí arrojó los resultados esperados.

EL RESULTADO

La plataforma omnicanal de SentiOne monitoriza miles de millones de conversaciones en millares de sitios web. Para poder ofrecer a las empresas información sobre audiencias y servicios de atención al cliente automatizados mediante IA, la empresa recopila, procesa y analiza inmensas cantidades de datos. El rendimiento del almacenamiento desempeña, pues, un papel decisivo.

El nuevo cluster, basado en servidores dedicados Advance-2 provistos de discos NVMe, supuso un punto de inflexión para SentiOne. La migración del hardware se tradujo en un incremento inmediato del rendimiento, multiplicando por seis la velocidad del cluster.



Tiempo medio de las consultas en el cluster Elasticsearch

La actualización de la versión Elasticsearch, que incluía la activación de la funcionalidad ARS, permitió que el cluster funcionase de forma más homogénea, repartiendo la carga de manera uniforme entre varios nodos. Según las mediciones realizadas por el equipo, la carga ahora se aproxima al 50% en condiciones normales y no supera nunca el 80%, proporcionando así un buen margen para escalar o añadir nuevas funcionalidades cuando sea necesario.

> *Desde el punto de vista económico, la migración ha sido un éxito: el coste de gestión del cluster se ha reducido en un 5% y hemos conseguido mejorar el rendimiento considerablemente.*
>
> *Michał Brzezicki, cofundador y CTO de SentiOne*

OVH es un proveedor mundial de cloud hiperescalable que ofrece a las empresas valor y prestaciones de referencia. Como líder europeo, OVH es la alternativa en cloud. El grupo, fundado en 1999, gestiona 28 datacenters ubicados en 12 localizaciones de cuatro continentes, cuenta con su propia red mundial de fibra óptica y controla la totalidad de la cadena de alojamiento. Basándose en sus infraestructuras propias, OVH proporciona herramientas y soluciones simples y potentes que permiten poner la tecnología al servicio del sector y revolucionan la forma de trabajar de más de 1,4 millones de clientes en todo el mundo. El respeto del individuo y las libertades, así como la igualdad de acceso a las nuevas tecnologías, son valores clave estrechamente ligados a la empresa. Y de ahí el lema de OVH: «Innovation for Freedom».

OVH
Innovation for Freedom

ovh.es OVH ovhcom OVH