codurance | CRAFT AT HEART

AWS OpenSearch: ¿Qué es y por qué implementarlo?

Ignacio Gonzalez

codurance.com

Indice

Introducción

Los humanos producimos datos de manera continua: aplicaciones y sistemas recopilan información y, además de monitorear nuestras aplicaciones, registran problemas relacionados con los controles de salud de nuestros sistemas. Ahora vivimos en una era en la que contamos con herramientas para procesar grandes volúmenes de información en tiempo real, lo que nos capacita para tomar decisiones altamente informadas y emprender acciones cuando sea necesario, ya sea para nuestros negocios o incluso en nuestras vidas. Esta guía presenta una de esas herramientas que recientemente se volvieron sumamente populares por distintas razones: el Servicio OpenSearch de Amazon, un motor de búsqueda de texto completo con capacidades que transformarán la forma en que tú y tu organización procesan información.

Motores de búsqueda

Los motores de búsqueda ofrecen servicios que le permiten al usuario encontrar información a través de queries y recuperar información relevante de sus índices en función del input del usuario. Actualmente, existen diferentes tipos de motores de búsqueda, como los motores de búsqueda web, de imágenes, de vídeos, de noticias, entre otros.

Un motor de búsqueda de texto completo es un buscador diseñado para recuperar información basada en documentos escritos, permitiéndole a los usuarios buscar documentos a través de palabras o frases. Por lo general, los motores de búsqueda clasifican los resultados según la relevancia de la información proporcionada. La relevancia de los resultados depende de diferentes algoritmos y otros factores como la proximidad de las palabras, la frecuencia o el análisis contextual. Estos buscadores utilizan índices invertidos (una lista de palabras que señalan el documento en el que se encuentran) y estrategias de tokenización, así como filtrado, para reducir el scope de los queries.

El servicio de Amazon OpenSearch

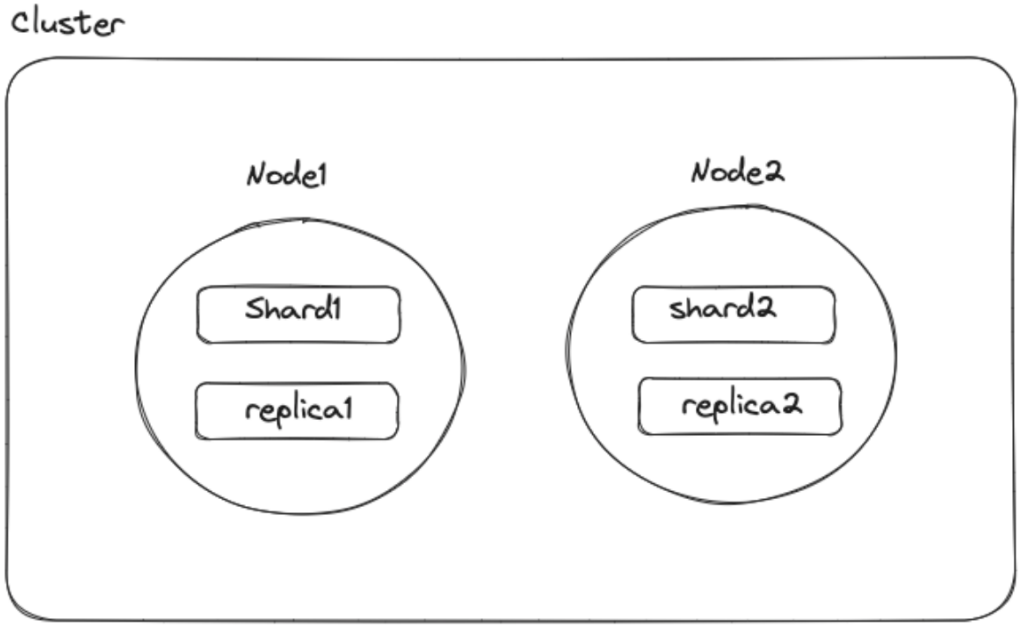
Amazon OpenSearch, previamente conocido como Amazon Elasticsearch (no hay que confundirlo con Elasticsearch de Elastic.co, ya que este último pertenece a otra empresa, y Amazon OpenSearch es una versión derivada de Elasticsearch de Elastic.co), es un servicio que permite almacenar grandes cantidades de datos y realizar operaciones sobre ellos. Basado en Apache Lucene, Amazon OpenSearch es de código abierto y ofrece las mismas funcionalidades tanto de un motor de búsqueda como de un motor de análisis (frecuentemente utilizados para búsquedas de texto completo, análisis de registros, monitoreo y exploración de datos). Datos de entrada como métricas, registros, trazas, etc., pueden ser ingresados al servicio OpenSearch para sucesivamente ser analizados y obtener insights en tiempo real. Gracias al poder de Amazon OpenSearch, puedes desplegar, operar y escalar clústeres de OpenSearch en la nube a mayor escala.

Algunas de las ventajas de este servicio incluyen:

- Escalabilidad: totalmente basada en tus necesidades ya que admite diferentes tamaños de clúster.
- Disponibilidad: tolerancia a fallos mediante la distribución de nodos en múltiples zonas de disponibilidad.
- Seguridad: proporciona integración con AWS IAM, para que el administrador pueda definir roles si es necesario.
- Monitoreo y alertas: se integra con AWS CloudWatch, permitiéndole así al usuario monitorear diferentes métricas de rendimiento, configurar alarmas y recibir notificaciones.
- Integración con otros servicios de AWS.

Adicionalmente, Amazon OpenSearch Serverless ofrece un servicio que reduce el tiempo y esfuerzo implicado en la gestión de la infraestructura de OpenSearch. Aunque esto te facilitará el uso de una herramienta tan potente de búsqueda y análisis, debes saber que también hará que pierdas cierto grado de control y personalización.

En el contexto de Amazon OpenSearch, es relevante señalar los siguientes conceptos:



- **Clúster:** es una colección de uno o más nudos.
- **Node:** almacena tus datos y procesa las solicitudes de búsqueda.
- **Índices:** la forma en que se organiza la información es mediante el uso de índices, siendo cada índice una colección de documentos JSON.
- **Shards:** los índices se dividen en fragmentos (shards) para una distribución equitativa de los datos entre los nodos de un clúster.

“Por ejemplo, un índice de 400 GB podría ser demasiado grande para que cualquier nodo individual en tu clúster pueda procesarlo. Pero si lo divides en diez fragmentos

o 'shards', cada uno de 40 GB, OpenSearch puede distribuir los fragmentos entre diez nodos y trabajar con cada fragmento de manera individual.

Por defecto, OpenSearch crea una réplica de cada fragmento (shard) primario. Si divides tu índice en diez fragmentos, por ejemplo, OpenSearch también creará diez réplicas de los mismos. Estos fragmentos de réplica sirven de respaldo en caso de que un nodo falle. A pesar de ser solo una parte de un índice de OpenSearch, cada fragmento es en realidad un índice completo de Lucene. Este detalle es importante, ya que cada instancia de Lucene es un proceso en ejecución que consume CPU y memoria. Tener más fragmentos no siempre es mejor. Dividir un índice de 400 GB en 1,000 fragmentos, por ejemplo, ejercería una carga innecesaria en tu clúster. Una buena regla general es mantener el tamaño del fragmento entre 10 y 50 GB." ***Link***

El origen de Amazon OpenSearch

El origen de Amazon OpenSearch se remonta a 2010, cuando Elastic.co creó Elasticsearch de código abierto, un motor de búsqueda y análisis distribuido construido sobre Apache Lucene. Este fue diseñado para proporcionar capacidades escalables y eficientes de búsqueda de texto completo, así como exploración de datos y otras características.

Reconociendo su utilidad, en 2015, Amazon Web Services decidió introducir el Servicio de Amazon Elasticsearch, ofreciendo el servicio Elasticsearch en la nube y facilitando a los usuarios la implementación y operación de clústeres Elasticsearch.

En 2021, la plataforma de AWS anunció el lanzamiento del Servicio de Amazon OpenSearch, que sucede al Servicio de Amazon Elasticsearch. Este proporciona las mismas funcionalidades, siendo de código abierto e independiente de Elastic.co.

Usos del servicio Amazon OpenSearch

Algunos de los usos más populares e interesantes de este servicio incluyen funciones como el análisis de registros, ya que OpenSearch está bien equipado con herramientas para procesar grandes cantidades de registros con el fin de realizar búsquedas, visualizar y monitorear mediante la configuración de alarmas para detectar anomalías.

Adicionalmente, se puede utilizar para monitorear aplicaciones mediante el análisis de métricas y rendimiento utilizando los datos ingeridos, permitiéndole a los usuarios realizar controles de salud e identificar problemas en sus aplicaciones.

La capacidad de búsqueda, descubrimiento y visualización de información también es ampliamente reconocida a través de la función de búsqueda de texto completo. Esta función posibilita a los usuarios construir motores de búsqueda para cualquier

tipo de aplicación, lo que, a su vez, le permite a los clientes buscar información. Estas funcionalidades abren la puerta a otras funciones más genéricas, como el análisis de negocios e inteligencia. Al tener acceso a diferentes vistas y métricas de tus datos, puedes llegar a conclusiones informadas y tomar acciones sobre el dominio en el que estás trabajando.

Todo esto puede realizarse de manera instantánea, ya que OpenSearch es capaz de metabolizar y analizar datos en tiempo real. Dentro de la función de búsqueda y descubrimiento, es relevante destacar algunas de las herramientas disponibles para el usuario:

- Búsqueda de texto completo o 'full-text search'
- Filtrado
- Búsqueda facetada
- Autocompletar
- Puntuación de relevancia o 'relevance scoring'

En cuanto respecta a la visualización, es importante tener en cuenta que Amazon OpenSearch no utiliza Kibana, sino una herramienta similar llamada OpenSearch Dashboards, que esencialmente persigue el mismo propósito.

Escalabilidad y Confiabilida

En lo que respecta a características como la escalabilidad, es relevante destacar la escalabilidad elástica, la cual consiste en la capacidad de escalar el tamaño de tu clúster de acuerdo con la carga de trabajo que esté gestionando. Esta "elasticidad" le permite a los usuarios adaptarse a distintas cantidades de tráfico en diversos contextos. Amazon OpenSearch sigue una estrategia que implica almacenar información en distintos fragmentos dentro de un clúster, lo cual no solo mejora el rendimiento, sino que, al tratarse de un servicio gestionado, también automatiza el proceso. En este sentido, Amazon OpenSearch se encarga del aprovisionamiento de la infraestructura, la configuración y el mantenimiento.

Otros aspectos importantes a destacar son la confiabilidad, la durabilidad y la disponibilidad de los datos. Esto se debe a que Amazon OpenSearch garantiza una copia de seguridad automática, replicando datos y almacenándolos en nodos diferentes para evitar pérdida de datos. El monitoreo y las alarmas son otras dos opciones adicionales vinculadas a la confiabilidad que funcionan de manera excelente para detectar anomalías y recibir notificaciones al respecto.

Como se mencionó anteriormente, Amazon OpenSearch también proporciona un servicio totalmente gestionado, lo que implica que AWS se encarga de mantener la infraestructura y realizar otras tareas de mantenimiento, como aplicar parches, realizar actualizaciones, y mantener el hardware, etc.

El despliegue Multi-AZ es otra característica popular en varios servicios de AWS, que permite a los usuarios desplegar clústeres en diferentes Zonas de Disponibilidad (AZ, por sus siglas en inglés), mejorando así la tolerancia a fallos y evitando interrupciones. Esto, combinado con otras funciones como copias de seguridad automáticas que pueden ser almacenadas en buckets de S3, ofrece protección para tus datos.

Por último, pero no menos significativo, el soporte de AWS, la documentación y los foros son excelentes recursos para recibir asistencia y obtener información sobre resolución de determinados problemas.

Consideraciones de costo y uso

Este servicio sigue un esquema de pago por uso que variará según el tipo de servicio, el tamaño de almacenamiento y los números de instancias. Es crucial evaluar minuciosamente las instancias antes de elegirlas, ya que deben estar configuradas de manera óptima para lograr el rendimiento máximo, considerando variables como CPU, memoria y carga de trabajo de red. En lo que respecta a las opciones de almacenamiento, existen dos alternativas disponibles: EBS y S3. Por lo general, se recomienda usar EBS para baja latencia, mientras que S3 ofrece una capacidad más económica para datos que se acceden con menor frecuencia.

AWS también cobra por la transferencia de datos, así que es crucial tener en cuenta el volumen de datos a ingresar: solo pagas por los recursos consumidos por tu carga de trabajo. OpenSearch Ingestion te cobrará únicamente por la computadora necesaria para ingresar, transformar y enrutar datos en un pipeline de Ingestion de OpenSearch.

Las instancias reservadas pueden adquirirse por uno o tres años, lo que posiblemente resulte en ahorros durante ese período, dependiendo de tu estrategia y como esta se compare con las instancias on-demand. También es importante mencionar la estructura de precios para UltraWarm y almacenamiento en frío. UltraWarm permite conservar grandes cantidades de datos manteniendo la eficiencia, mientras que el almacenamiento en frío representa una opción de almacenamiento más económica para datos con un acceso menos frecuente en AWS S3, y solo pagas por la capacidad informática cuando es necesario.

AWS también ofrece un nivel gratuito que consta de 750 horas al mes para una instancia ‘t2.small.search’ o ‘t3.small.search’, típicamente utilizadas para pruebas, así como 10 GB al mes de EBS. Para obtener más información, puedes consultar este link.

Seguridad y cumplimiento de la normativa

El cumplimiento de las normas y la seguridad son temas sumamente relevantes en el ámbito empresarial. Nadie puede ignorar estos dos aspectos sin enfrentar problemas a largo plazo cuando se desee interactuar con otras entidades (ya sea dentro del mismo país o en un contexto industrial similar).

Cuando hablamos de la seguridad de datos, IAM es el servicio de AWS que te permite definir de manera detallada los accesos. Este servicio está íntimamente vinculado con Amazon OpenSearch para garantizar que solo el personal autorizado pueda acceder a la plataforma. Otra opción consiste en aislar tu clúster en tu propia red privada y controlar su tráfico mediante Virtual Private Cloud (VPC), creando grupos de seguridad.

Amazon OpenSearch también admite cifrado en reposo y en tránsito. En reposo significa que puedes habilitar el cifrado para tus datos utilizando AWS Key Management Service. En tránsito, tienes la opción de utilizar el cifrado SSL/TLS para asegurar la comunicación entre clientes y clústeres.

Otro aspecto relevante es que la retención de datos se puede personalizar a distintos períodos en tus copias de seguridad.

Al mismo tiempo, la conformidad y auditoría están disponibles mediante la integración con otros servicios, como AWS CloudTrail, o características incorporadas, como los registros de acceso, que proporcionan información sobre el uso dentro de los clústeres. Además, CloudWatch puede ser configurado para recibir rastreos o registros, facilitando información detallada sobre las acciones realizadas durante el uso del servicio.

Amazon OpenSearch también cumple con regulaciones como GDPR, HIPAA e ISO 27001, lo que contribuye a cumplir con requisitos específicos de la industria en cuanto a la privacidad de datos. Esto demuestra que AWS ha alcanzado un alto nivel de control de seguridad sobre sus infraestructuras.

Ejemplo de caso de uso: Migración a OpenSearch para un e-commerce minorista

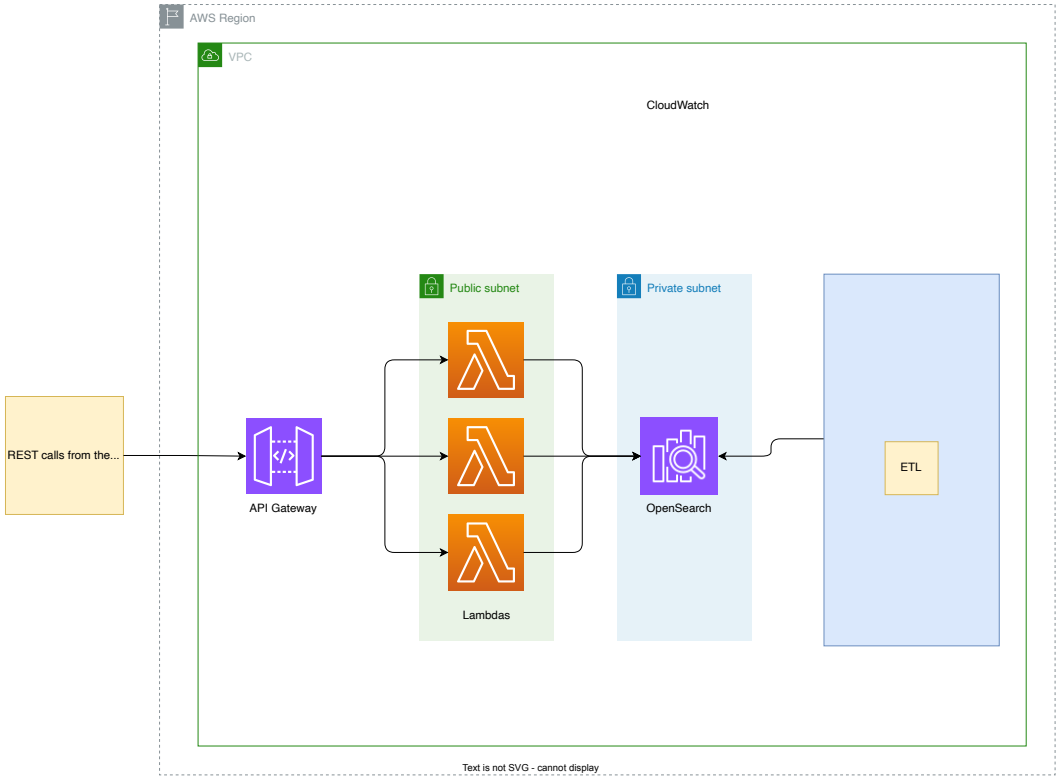
Imagina que una empresa minorista de e-commerce necesita mejorar su sistema de búsqueda y recomendación de productos, ya que el actual es lento y no proporciona la información más precisa. En este contexto específico, Amazon OpenSearch sería beneficioso por varias razones. En primer lugar, la empresa podría indexar sus productos en su catálogo, permitiéndole a sus clientes realizar consultas, aplicar filtros y ordenar los resultados. Amazon OpenSearch ofrece las funciones de un motor de búsqueda de texto completo y proporciona recomendaciones personalizadas mediante el análisis del historial de navegación y compras de los clientes, junto con los atributos de los productos.

Hemos abordado previamente el tema de la escalabilidad y la confiabilidad, lo que, en este escenario, posibilitaría el crecimiento del catálogo de la empresa mientras la instancia de OpenSearch se ajusta a dicho crecimiento.

Otro motivo válido para contratar Amazon OpenSearch serían las capacidades de alerta y monitoreo que ofrece, permitiendo la configuración de alarmas y notificaciones ante la presencia de anomalías o problemas.

Arquitectura genérica para el caso de uso hipotético.

El diagrama ilustra la arquitectura recomendada para implementar AWS OpenSearch en el caso de uso mencionado anteriormente. La utilización de OpenSearch como plataforma de registro y monitoreo (como alternativa a AWS CloudWatch) constituye un caso de uso distinto, por lo que sería necesario contemplar una arquitectura diferente.



En el escenario contemplado, el cual emplea OpenSearch como un motor de búsqueda accesible para los usuarios finales, se deben considerar los siguientes componentes principales:

- AAWS OpenSearch como el motor de búsqueda.
- Un proceso ETL para alimentar los índices del motor de búsqueda con los datos a buscar.
- Una arquitectura de front-end para exponer la funcionalidad de búsqueda

a los navegadores de los usuarios.
- Una solución de registro y monitoreo.

El motor de OpenSearch se encuentra en el centro y debe implementarse en una red con las siguientes características:

- A Una subred privada: el motor de búsqueda no debería desplegarse en una subred pública, siguiendo la misma lógica de cualquier base de datos que se pueda implementar en AWS.
- OpenSearch debe estar abierto para recibir llamadas desde el front-end y el componente ETL.

La configuración ETL representada en este diagrama es un marcador de posición muy genérico, ya que cada empresa tiene su propia red de componentes de backend. El componente ETL debe encargarse de insertar, actualizar y eliminar documentos en los índices de OpenSearch mediante la llamada a las APIs de indexación. Aunque en el diagrama este componente se muestra como parte de la VPC de OpenSearch, con las configuraciones de red apropiadas, también puede implementarse en otra VPC.

Las búsquedas realizadas por los usuarios finales idealmente deberían procesarse en milisegundos, y la configuración adecuada para los componentes sin servidor es la siguiente: una API Gateway y una o más funciones Lambda son las elecciones correctas. Esta configuración llamará a las APIs de funciones de búsqueda de OpenSearch. Adicionalmente, AWS CloudWatch se podría utilizar como una solución pre-configurada para el registro y monitoreo.

Caso de uso: Migración a OpenSearch en la industria de viajes y hospitalidad

Codurance respaldó a un cliente en la migración de su anterior motor de búsqueda a AWS OpenSearch. El antiguo motor de búsqueda era un producto SaaS, por lo que el proyecto de migración consistió en la reimplementación de las funcionalidades necesarias con tecnologías de AWS. La nueva infraestructura incluye componentes serverless (AWS API Gateway, Lambda y Step Functions) y AWS OpenSearch.

Entre los beneficios obtenidos destacamos que:

- La empresa tenía una gran influencia del stack de servicios y herramientas de AWS, por lo que contar con AWS OpenSearch aumentó la interoperabilidad con el resto de los servicios.
- Ahora el nivel de escalabilidad es enorme debido a que AWS puede manejar una gran cantidad de datos.
- Tener a AWS como entidad de mantenimiento del servicio resulta de gran ayuda en las tareas de aprovisionamiento, mantenimiento, aplicación de parches y actualizaciones.

- Dado que AWS es una entidad sólida, el nivel de seguridad también es sobresaliente.
- Ahora la empresa cuenta con un ecosistema popular que le permite al equipo buscar recursos en caso de necesidad.

Conclusión

En Codurance, nos esforzamos constantemente por mejorar nuestros sistemas, y Amazon OpenSearch se destaca como una herramienta extraordinariamente poderosa que no solo nosotros, sino cualquier organización, podría aprovechar para obtener insights valiosos y dar sentido a sus datos. Amazon OpenSearch ofrece diversas funciones, como análisis en tiempo real, agregaciones y consultas complejas, junto con estrategias para monitorear la salud del sistema.

La escalabilidad y el rendimiento mejorado son elementos esenciales en cualquier sistema. La naturaleza distribuida de OpenSearch facilita la gestión de grandes volúmenes de datos de manera impecable, adaptándose dinámicamente al aumento en la cantidad de información ingresada. Amazon OpenSearch es compatible con diversos formatos de datos, permitiendo a los usuarios recopilar información de diversas fuentes y presentando una notable compatibilidad con otros servicios de AWS.

Con una comunidad activa, se encuentra cada vez más respaldo, además del ya disponible, constituyendo otro aspecto positivo a considerar al evaluar herramientas como esta. Amazon OpenSearch transformará la forma en que analizas e ingieres datos en tus sistemas.

codurance | CRAFT AT HEART

Software | Equipos | Procesos | Comunidad

hello@codurance.com @codurance codurance.com