# ANALYTICS

## AMAZON EMR

### AMAZON EMR FEATURES

Amazon EMR es un servicio web que permite a las empresas, los investigadores, los analistas de datos y los desarrolladores procesar grandes cantidades de datos.

EMR utiliza un marco de Hadoop alojado que se ejecuta en Amazon EC2 y Amazon S3.

Con EMR puede ejecutar análisis a escala de petabytes a menos de la mitad del costo de las soluciones locales tradicionales y más de 3 veces más rápido que Apache Spark estándar.

Puede ejecutar cargas de trabajo en instancias Amazon EC2, en clústeres de Amazon Elastic Kubernetes Service (EKS) o en las instalaciones mediante EMR en AWS Outposts.

Se ejecuta en una zona de disponibilidad dentro de una Amazon VPC. Soporta Apache Spark, HBase, Presto y Flink.

Más utilizado para el análisis de registros, análisis financiero o actividades de extracción, traducción y carga (ETL). Un paso es una tarea programática para realizar algún proceso en los datos (por ejemplo, contar palabras).

Un clúster es una colección de instancias EC2 aprovisionadas por EMR para ejecutar los pasos.

EMR es un buen lugar para implementar Apache Spark, un procesamiento distribuido de código abierto utilizado para cargas de trabajo de big data que utiliza el almacenamiento en caché en memoria y la ejecución optimizada de consultas.

También puede iniciar clústeres de Presto. Presto es un motor de consulta SQL distribuido de código abierto diseñado para consultas analíticas rápidas en grandes conjuntos de datos.

EMR lanza todos los nodos para un clúster determinado en la misma zona de disponibilidad de Amazon EC2.

Puede obtener acceso a Amazon EMR mediante la consola de administración de AWS, las herramientas de línea de comandos, los SDK o la API de EMR.

Con EMR tiene acceso al sistema operativo subyacente (puede SSH en).

## AMAZON KINESIS

### AMAZON KINESIS FEATURES

Amazon Kinesis facilita la recopilación, el procesamiento y el análisis de datos de streaming en tiempo real para que pueda obtener información oportuna y reaccionar rápidamente a la nueva información.

Kinesis es una colección de servicios para procesar flujos de datos diversos.

Los datos se procesan en "fragmentos", y cada fragmento puede ingerir 1000 registros por segundo. Hay un límite predeterminado de 500 fragmentos, pero puede solicitar un aumento a fragmentos ilimitados. Un registro consta de una clave de partición, un número de secuencia y un blob de datos (hasta 1 MB).

Almacén de datos transitorio: retención predeterminada de 24 horas, pero se puede configurar durante un máximo de 7 días. Hay cuatro tipos de servicio de Kinesis, y estos se detallan a continuación.



### KINESIS VIDEO STREAMS

Kinesis Video Streams facilita la transmisión segura de vídeo desde dispositivos conectados a AWS para análisis, aprendizaje automático (ML) y otros procesos.

Almacena, cifra e indexa de forma duradera los flujos de datos de vídeo y permite el acceso a los datos a través de API fáciles de usar.

Los productores proporcionan flujos de datos.

Almacena datos durante 24 horas de forma predeterminada, hasta 7 días.

Almacena datos en fragmentos: 5 transacciones por segundo para lecturas, hasta una velocidad de lectura máxima de 2 MB por segundo y 1000 registros por segundo para escrituras de hasta un máximo de 1 MB por segundo.

Los consumidores reciben y procesan datos. Puede tener varios fragmentos en una secuencia.

Admite el cifrado en reposo con cifrado del lado del servidor (KMS) con una clave maestra de cliente. Kinesis Video Streams no aparece mucho en los exámenes de AWS.

### KINESIS DATA STREAMS

Kinesis Data Streams le permite crear aplicaciones personalizadas que procesan o analizan datos de streaming para necesidades especializadas.

Kinesis Data Streams permite el procesamiento en tiempo real de big data de streaming.

Kinesis Data Streams es útil para mover rápidamente los datos de los productores de datos y luego procesarlos continuamente.

Kinesis Data Streams **almacena datos** para su posterior procesamiento por parte de las aplicaciones (diferencia clave con Firehose, que entrega datos directamente a los servicios de AWS).

Los casos de uso comunes incluyen:

* Registro acelerado y ingesta de alimentación de datos.
* Métricas e informes en tiempo real.
* Análisis de datos en tiempo real.
* Procesamiento de flujo complejo.

El siguiente diagrama ilustra la arquitectura de alto nivel de Kinesis Data Streams.

* Los productores envían continuamente datos a Kinesis Data Streams.
* Los consumidores procesan los datos en tiempo real.
* Los consumidores pueden almacenar sus resultados mediante un servicio de AWS como Amazon DynamoDB, Amazon Redshift o Amazon S3.
* Las aplicaciones de Kinesis Streams son consumidores que se ejecutan en instancias EC2.
* Los fragmentos son grupos o registros de datos identificados de forma única en una secuencia.
* Los registros son las unidades de datos almacenadas en Kinesis Stream.



**Consumers process the data and**

**save to destinations**

Destinations

Stream

Amazon S3

Amazon DynamoDB

Amazon RedShift

Amazon EMR

Kinesis Firehose

**Analytics Tools**

Un productor crea los datos que componen el flujo. Los productores se pueden utilizar a través de lo siguiente:

* Kinesis Streams API.
* Kinesis Producer Library (KPL).
* Kinesis Agent.

Un registro es la unidad de datos almacenada en un flujo de datos de Amazon Kinesis.

Un registro se compone de un número de secuencia, una clave de partición y un blob de datos.

De forma predeterminada, se puede acceder a los registros de una secuencia durante un máximo de 24 horas desde el momento en que se agregan a la secuencia (se puede aumentar a 7 días habilitando la retención de datos extendida).

Un blob de datos son los datos de interés que su productor de datos agrega a un flujo de datos.

El tamaño máximo de un blob de datos (la carga útil de datos antes de la codificación Base64) dentro de un registro es de 1 megabyte (MB).

Un fragmento es la unidad de rendimiento base de un flujo de datos de Amazon Kinesis.

Un fragmento proporciona una capacidad de entrada de datos de 1 MB/seg y salida de datos de 2 MB/seg. Cada fragmento puede admitir hasta 1000 registros PUT por segundo.

Una corriente se compone de uno o más fragmentos.

La capacidad total de la corriente es la suma de las capacidades de sus fragmentos.

Kinesis Data Streams admite el resharding, que le permite ajustar el número de fragmentos en su flujo para adaptarse a los cambios en la velocidad del flujo de datos a través del flujo.

Hay dos tipos de operaciones de reharding: división de particiones y combinación de particiones.

* En una división de fragmentos, se divide un solo fragmento en dos fragmentos.
* En una combinación de fragmentos, se combinan dos fragmentos en un solo fragmento.

Forma

Descripción generada automáticamente con confianza media

La división aumenta el número de fragmentos en el flujo y, por lo tanto, aumenta la capacidad de datos del flujo.

La división aumenta el costo de su transmisión (usted paga por fragmento).

La fusión reduce el número de fragmentos en el flujo y, por lo tanto, disminuye la capacidad de datos (y el costo) del flujo.

Los consumidores son las instancias EC2 que analizan los datos recibidos de un flujo. Los consumidores son conocidos como aplicaciones de Amazon Kinesis Streams.

Cuando la velocidad de datos aumente, agregue más fragmentos para aumentar el tamaño de la secuencia. Elimine los fragmentos cuando la velocidad de datos disminuya.

Las claves de partición se utilizan para agrupar datos por fragmento dentro de una secuencia. Kinesis Streams utiliza claves maestras de KMS para el cifrado.

Para leer o escribir en una secuencia cifrada, las aplicaciones de productor y consumidor deben tener permiso para acceder a la clave maestra.

Kinesis Data Streams se replica de forma sincrónica en tres zonas de disponibilidad.

### KINESIS DATA FIREHOSE

Kinesis Data Firehose es la forma más fácil de cargar datos de streaming en almacenes de datos y herramientas de análisis. Captura, transforma y carga datos de streaming.

Permite análisis casi en tiempo real con herramientas y paneles de inteligencia empresarial existentes. Kinesis Data Streams se puede utilizar como fuente(s) de Kinesis Data Firehose.

Puede configurar Kinesis Data Firehose para transformar sus datos antes de entregarlos. Con Kinesis Data Firehose no necesita escribir una aplicación ni administrar recursos. Firehose puede agrupar, comprimir y cifrar datos antes de cargarlos.

Firehose replica sincrónicamente los datos en tres zonas de disponibilidad a medida que se transportan a los destinos. Cada flujo de entrega almacena registros de datos durante un máximo de 24 horas.

Una fuente es donde sus datos de transmisión se generan y capturan continuamente. Un flujo de entrega es la entidad subyacente de Amazon Kinesis Data Firehose.

Un registro son los datos de interés que su productor de datos envía a un flujo de entrega. El tamaño máximo de un registro (antes de la codificación Base64) es de 1000 KB.

Un destino es el almacén de datos donde se entregarán sus datos. Los destinos de firehose incluyen:

* Amazon S3.
* Amazon Redshift.
* Amazon Elasticsearch Service.
* Splunk.

Los productores proporcionan flujos de datos. Sin fragmentos, totalmente automatizado.

Puede cifrar datos con una clave existente de AWS Key Management Service (KMS). El cifrado del lado del servidor se puede utilizar si se utiliza Kinesis Streams como origen de datos.

Firehose puede invocar una función de AWS Lambda para transformar los datos entrantes antes de entregarlos a un destino.

Para los destinos de Amazon S3, los datos de streaming se entregan a su bucket de S3. Si la transformación de datos es

habilitado, puede realizar una copia de seguridad opcional de los datos de origen en otro bucket de Amazon S3:

Destinations



Firehose **captures**, **transforms** and

**loads** streaming data



Optional Lambda transformation



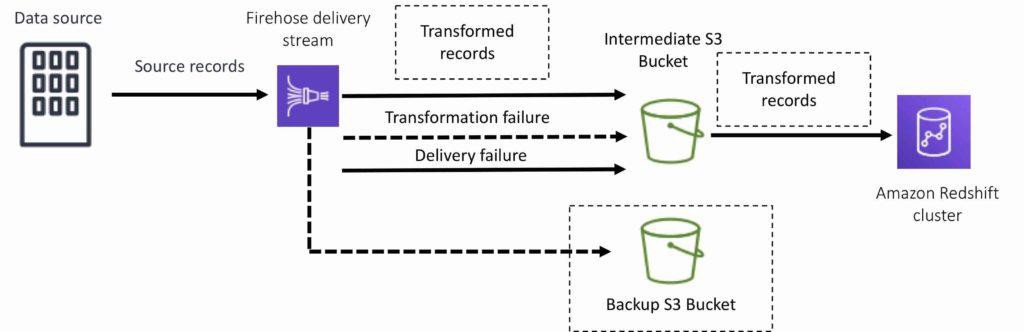
Amazon S3

Amazon RedShift

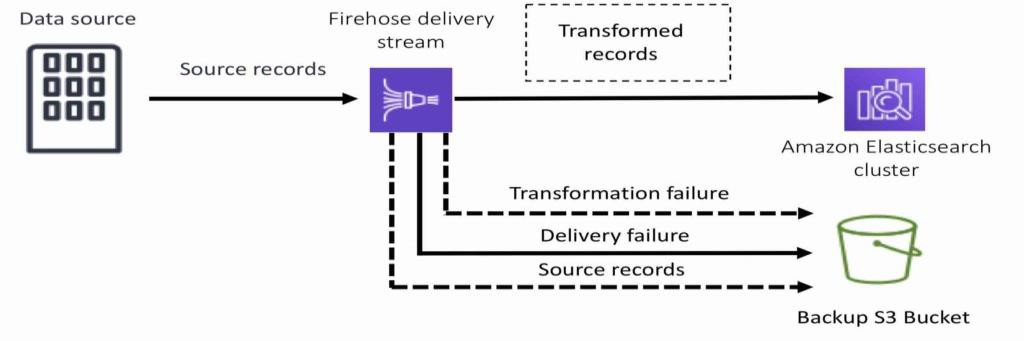
Amazon Elasticsearch

Analytics Tools

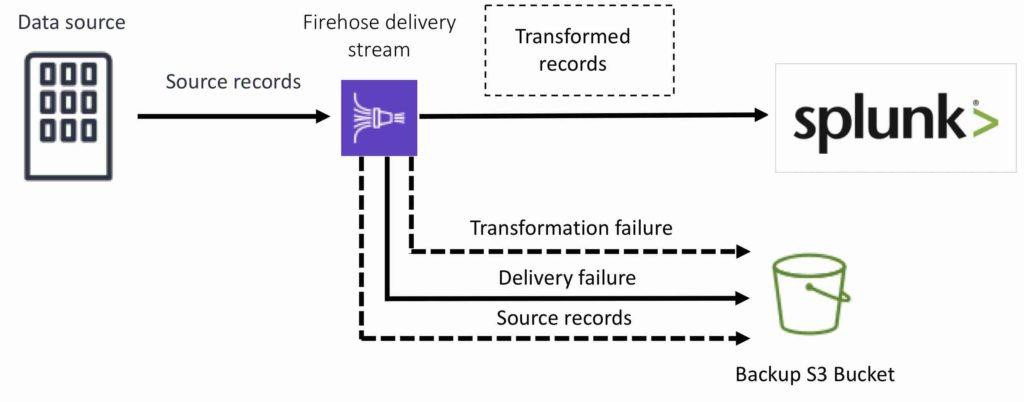
Para los destinos de Amazon Redshift, los datos de streaming se entregan primero en el bucket de S3. A continuación, Kinesis Data Firehose emite un comando **COPY** de Amazon Redshift para cargar datos desde el bucket de S3 al clúster de Amazon Redshift. Si la transformación de datos está habilitada, puede realizar opcionalmente una copia de seguridad de los datos de origen en otro bucket de Amazon S3:



Para los destinos de Amazon Elasticsearch, los datos de streaming se entregan a su clúster de Amazon ES y, opcionalmente, se puede realizar una copia de seguridad en su bucket de S3 simultáneamente.:



Para los destinos de Splunk, los datos de streaming se entregan a Splunk y, opcionalmente, se puede realizar una copia de seguridad en su bucket de S3 simultáneamente.:



### KINESIS DATA ANALYTICS

Amazon Kinesis Data Analytics es la forma más sencilla de procesar y analizar datos de streaming en tiempo real. Puede utilizar consultas SQL estándar para procesar secuencias de datos de Kinesis.

Proporciona análisis en tiempo real. Casos de uso:

* Generar análisis de series temporales.
* Feed paneles de control en tiempo real.
* Crear alertas y notificaciones en tiempo real.

Cree y ejecute rápidamente código SQL potente en fuentes de streaming. Puede ingerir datos de Kinesis Streams y Kinesis Firehose.

Salida a S3, RedShift, Elasticsearch y Kinesis Data Streams. Se encuentra sobre Kinesis Data Streams y Kinesis Data Firehose.

Una aplicación de Kinesis Data Analytics consta de tres componentes:

* Entrada: la fuente de transmisión para su aplicación.
* Código de aplicación: una serie de instrucciones SQL que procesan la entrada y producen la salida.
* Salida: uno o más flujos en la aplicación para contener resultados intermedios.

Kinesis Data Analytics admite dos tipos de entradas: fuentes de datos de streaming y fuentes de datos de referencia:

* Un origen de datos de streaming es un origen de datos generados continuamente que se leen en la aplicación para su procesamiento.
* Un origen de datos de referencia son datos estáticos que la aplicación utiliza para enriquecer los datos procedentes de orígenes de streaming.

Puede configurar destinos para conservar los resultados.

Admite Kinesis Streams y Kinesis Firehose (S3, RedShift, Elasticsearch) como destinos.

IAM se puede utilizar para proporcionar a Kinesis Analytics permisos para leer registros de orígenes y escribir en destinos.

### SECURITY

Controle el acceso/autorización mediante políticas de IAM. Cifrado en vuelo mediante endpoints HTTPS.

Cifrado en reposo mediante KMS.

Posibilidad de cifrar / descifrar datos en el lado del cliente.

Puntos de conexión de VPC disponibles para que Kinesis acceda dentro de una VPC.

### SQS VS SNS VS KINESIS

SQS:

* Los consumidores extraen datos.
* Los datos se eliminan después de ser consumidos.
* Puede tener tantos trabajadores (consumidores) como necesite.
* No es necesario aprovisionar el rendimiento.
* No hay garantía de pedido (excepto con colas FIFO).
* Retraso de mensaje individual. SNS:
* Enviar datos a muchos suscriptores.
* Hasta 10.000.000 de suscriptores.
* Los datos no se conservan (se pierden si no se eliminan).
* Pub/sub.
* Hasta 10.000.000 de temas.
* No es necesario aprovisionar el rendimiento.
* Se integra con SQS para un patrón de arquitectura de distribución.

Kinesis:

* Los consumidores extraen datos.
* Tantos consumidores como necesites.
* Posibilidad de reproducir datos.
* Diseñado para big data, análisis y ETL en tiempo real.
* Pedido a nivel de fragmento.
* Los datos caducan después de X días.
* Debe aprovisionar el rendimiento.

## AMAZON ATHENA

### AMAZON ATHENA FEATURES

Amazon Athena es un servicio de consulta interactivo que facilita el análisis de datos en Amazon S3 mediante SQL estándar.

Athena no tiene servidor, por lo que no hay infraestructura que administrar y solo paga por las consultas que ejecuta.

Athena es fácil de usar: simplemente apunte a sus datos en Amazon S3, defina el esquema y comience a realizar consultas con SQL estándar.

Amazon Athena utiliza Presto con compatibilidad completa con SQL estándar y funciona con una variedad de formatos de datos estándar, incluidos CSV, JSON, ORC, Apache Parquet y Avro.

Si bien Amazon Athena es ideal para consultas rápidas y ad-hoc y se integra con Amazon QuickSight para facilitar la visualización, también puede manejar análisis complejos, incluidas uniones grandes, funciones de ventana y matrices.

Amazon Athena utiliza un catálogo de datos administrado para almacenar información y esquemas sobre el

bases de datos y tablas que cree para los datos almacenados en Amazon S3.

Con Amazon Athena, no tiene que preocuparse por administrar o ajustar clústeres para obtener rapidez

rendimiento.

Athena está optimizado para un rendimiento rápido con Amazon S3.

Athena ejecuta automáticamente las consultas en paralelo, de modo que obtiene resultados de consultas en segundos, incluso en grandes conjuntos de datos.

La mayoría de los resultados se entregan en cuestión de segundos.

Con Athena, no hay necesidad de trabajos complejos de ETL para preparar los datos para el análisis. Esto hace que sea fácil para cualquier persona con habilidades SQL analizar rápidamente conjuntos de datos a gran escala.

Athena está integrado de forma inmediata con [AWS Glue](https://aws.amazon.com/glue/) Data Catalog, lo que le permite crear un repositorio de metadatos unificado en varios servicios, rastrear orígenes de datos para descubrir esquemas y rellenar su catálogo con definiciones de tablas y particiones nuevas y modificadas, y mantener el control de versiones de esquemas.

También puede utilizar las capacidades ETL totalmente administradas de Glue para transformar datos o convertirlos en formatos de columnas para optimizar el costo y mejorar el rendimiento.

### USE CASES

Los servicios de consulta como Amazon Athena, los almacenes de datos como Amazon Redshift y los sofisticados marcos de procesamiento de datos como Amazon EMR abordan diferentes necesidades y casos de uso.

Amazon Redshift proporciona el rendimiento de consultas más rápido para cargas de trabajo de inteligencia empresarial e informes empresariales, en particular aquellas que involucran SQL extremadamente complejo con múltiples uniones y subconsultas.

Amazon EMR simplifica y resulta rentable ejecutar marcos de procesamiento altamente distribuidos como Hadoop, Spark y Presto en comparación con las implementaciones locales. Amazon EMR es flexible: puede ejecutar aplicaciones y código personalizados, así como definir parámetros específicos de computación, memoria, almacenamiento y aplicaciones para optimizar sus requisitos analíticos.

Amazon Athena proporciona la forma más sencilla de ejecutar consultas ad-hoc para datos en S3 sin necesidad de configurar ni administrar ningún servidor.

La siguiente tabla muestra el caso de uso principal y las situaciones para utilizar algunos servicios de consulta y análisis de AWS:

|  |  |  |
| --- | --- | --- |
| **AWS**  **Service** | **Primary Use Case** | **When to use** |
| Amazon Athena | Query | Ejecute consultas interactivas con datos directamente en Amazon S3 sin preocuparse por formatear datos o administrar la infraestructura. Se puede utilizar con otros servicios como Amazon RedShift |
| Amazon RedShift | Data Warehouse | Extraiga datos de muchas fuentes, formatéelos y organícelos, almacénelos y admita consultas complejas y de alta velocidad que producen informes comerciales. |
| Amazon | Data | Marcos de procesamiento altamente distribuidos como Hadoop, Spark y Presto. |
| EMR | Processing | Ejecute una amplia variedad de tareas de procesamiento de datos de escalabilidad horizontal para aplicaciones como aprendizaje automático, análisis de gráficos, transformación de datos, transmisión de datos. |
| AWS Glue | ETL Service | Transforme y mueva datos a varios destinos. Se utiliza para preparar y cargar datos para análisis. La fuente de datos puede ser S3, RedShift u otra base de datos. Glue Data Catalog puede ser consultado por Athena, EMR y RedShift Spectrum |

### BEST PRACTICES

[Best practices](https://aws.amazon.com/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/) para el rendimiento con Athena:

* **Divida sus datos**: divida la tabla en partes y mantenga los datos relacionados juntos en función de los valores de columna, como la fecha, el país, la región, etc. Athena admite particiones de Hive.
* **Agrupar sus datos**: la partición de sus datos es agrupar los datos dentro de una sola partición.

1. **Usar compresión**: AWS recomienda usar Apache Parquet o Apache ORC.

* **Optimice el tamaño de los** archivos: las consultas se ejecutan de manera más eficiente cuando la lectura de datos se puede paralelizar y cuando los bloques de datos se pueden leer secuencialmente.
* **Optimice la** generación de almacenes de datos en columnas: Apache Parquet y Apache ORC son almacenes de datos en columnas populares.
* **Optimizar ORDER BY**– La cláusula ORDER BY devuelve los resultados de una consulta en orden de ordenación.
* **Optimizar GROUP BY**: el operador GROUP BY distribuye filas basadas en las columnas GROUP BY a los nodos de trabajo, que contienen los valores GROUP BY en la memoria.
* **Usar funciones aproximadas**: para explorar grandes conjuntos de datos, un caso de uso común es encontrar el recuento de valores distintos para una determinada columna utilizando COUNT(DISTINCT column).
* **Incluya solo las columnas que necesita**: al ejecutar sus consultas, limite la instrucción SELECT final a solo las columnas que necesita en lugar de seleccionar todas las columnas.

### PRICING

Con Amazon Athena, solo paga por las consultas que ejecuta.

Se le cobrará en función de la cantidad de datos analizados por cada consulta.

Puede obtener importantes ahorros de costos y ganancias de rendimiento al comprimir, particionar o convertir sus datos a un formato de columnas, ya que cada una de esas operaciones reduce la cantidad de datos que Athena necesita escanear para ejecutar una consulta.

## AWS GLUE

### AWS GLUE FEATURES

AWS Glue es un servicio de pago por uso, extracción, transformación y carga (ETL) totalmente administrado que automatiza los pasos lentos de la preparación de datos para el análisis.

AWS Glue detecta y perfila automáticamente los datos a través de Glue Data Catalog, recomienda y genera código ETL para transformar los datos de origen en esquemas de destino.

AWS Glue ejecuta los trabajos etL en un entorno Apache Spark totalmente administrado y escalable para cargar los datos en su destino.

AWS Glue también le permite configurar, orquestar y supervisar flujos de datos complejos. Puede crear y ejecutar un trabajo etl con unos pocos clics en la consola de administración de AWS.

Simplemente apunte AWS Glue a los datos almacenados en AWS y AWS Glue descubrirá los datos y almacenará los metadatos asociados (por ejemplo, definición de tabla y esquema) en AWS Glue Data Catalog.

Una vez catalogados, los datos se pueden buscar, consultar y estar disponibles para ETL de forma inmediata.

AWS Glue consta de un catálogo de datos que es un repositorio central de metadatos, un motor ETL que puede generar automáticamente código Scala o Python y un programador flexible que maneja la resolución de dependencias, el monitoreo de trabajos y los reintentos.

Juntos, estos automatizan gran parte del trabajo pesado indiferenciado involucrado con el descubrimiento, categorización, limpieza, enriquecimiento y movimiento de datos, para que pueda pasar más tiempo analizando sus datos.

### AWS GLUE CRAWLERS

Puede utilizar un rastreador para rellenar AWS Glue Data Catalog con tablas. Este es el método principal utilizado por la mayoría de los usuarios de AWS Glue.

Un rastreador puede rastrear varios almacenes de datos en una sola ejecución.

Una vez finalizado, el rastreador crea o actualiza una o más tablas en el catálogo de datos. Los trabajos de extracción, transformación y carga (ETL) que defina en AWS Glue utilicen estas tablas de Data Catalog como orígenes y destinos.

El trabajo ETL lee y escribe en los almacenes de datos especificados en las tablas Catálogo de datos de origen y destino.

Los rastreadores de AWS Glue se conectan a un almacén de datos de origen o destino, avanzan a través de una lista priorizada de clasificadores para determinar el esquema de los datos y, a continuación, crean metadatos en AWS Glue Data Catalog.

Los metadatos se almacenan en tablas de un catálogo de datos y se utilizan en el proceso de creación de trabajos ETL.

Puede ejecutar rastreadores según una programación, bajo demanda o activarlos en función de un evento para asegurarse de que los metadatos estén actualizados.

AWS Glue genera automáticamente el código para extraer, transformar y cargar datos.

Simplemente apunte AWS Glue a un origen y un destino, y AWS Glue creará scripts ETL para transformar, acoplar y enriquecer los datos.

El código se genera en Scala o Python y se escribe para Apache Spark.

AWS Glue ayuda a limpiar y preparar los datos para el análisis al proporcionar una transformación de aprendizaje automático denominada FindMatches para la deduplicación y la búsqueda de registros coincidentes.

### USE CASES

Utilice AWS Glue para descubrir las propiedades de los datos, transformarlos y prepararlos para el análisis.

Glue puede descubrir automáticamente datos estructurados y semiestructurados almacenados en lagos de datos en  [Amazon S3,](https://aws.amazon.com/s3/) almacenes de datos en  [Amazon Redshift](https://aws.amazon.com/redshift/) y varias bases de datos que se ejecutan en AWS.

Proporciona una vista unificada de los datos a través del catálogo de datos de Glue que está disponible para ETL, consultando e informando mediante servicios como [Amazon Athena,](https://aws.amazon.com/athena/) [Amazon EMR](https://aws.amazon.com/emr/) y [Amazon Redshift Spectrum.](https://aws.amazon.com/redshift/)

Glue genera automáticamente código Scala o Python para trabajos ETL que puede personalizar aún más utilizando herramientas con las que ya está familiarizado.

AWS Glue no tiene servidor, por lo que no hay recursos informáticos que configurar y administrar.

## AMAZON OPENSEARCH

### AMAZON OPENSEARCH FEATURES

Amazon OpenSearch Service es el sucesor de Amazon Elasticsearch Service.

Amazon OpenSearch Service es un conjunto de aplicaciones y análisis de búsqueda distribuidos y de código abierto basado en Elasticsearch.

Elasticsearch es un motor de búsqueda y análisis distribuido basado en Apache Lucene.

Elasticsearch es un motor de búsqueda popular comúnmente utilizado para análisis de registros, búsqueda de texto completo, inteligencia de seguridad, análisis de negocios y casos de uso de inteligencia operativa.

Con OpenSearch puede realizar análisis de registros de forma interactiva, realizar monitoreo de aplicaciones en tiempo real, búsqueda de sitios web, análisis de métricas de rendimiento y más.

 Puede elegir entre una variedad de opciones de motor de código abierto para su clúster de OpenSearch. Las opciones incluyen la última versión de OpenSearch y muchas versiones de ALv2 Elasticsearch.

### DEPLOYMENT AND MONITORING

Se puede crear un clúster de OpenSearch mediante la consola de administración de AWS, la API o la AWS CLI. Especifique el número de instancias, los tipos de instancia y las opciones de almacenamiento.

Las actualizaciones en contexto se pueden realizar sin tiempo de inactividad.

Proporciona monitoreo y alertas integrados con notificaciones automáticas.

Puede configurar alertas mediante los paneles de Kibana u OpenSearch y la API de REST. Las notificaciones se pueden enviar a través de webhooks personalizados, Slack, Amazon SNS y Amazon Chime. OpenSearch Service admite varios lenguajes de consulta, como:

* Domain-Specific Language (DSL).
* SQL queries with OpenSearch SQL.
* OpenSearch Piped Processing Language (PPL). OpenSearch integrates with open-source tools including:
* Logstash.
* OpenTelemetry.
* ElasticSearch APIs.

### OPENSEARCH IN AN AMAZON VPC

Los dominios de OpenSearch Services se pueden lanzar en una Amazon VPC.

El uso de una VPC permite una comunicación segura entre el servicio OpenSearch y otros servicios dentro de la VPC.

A continuación, se indican algunas de las formas en que los dominios de VPC difieren de los dominios públicos. Cada diferencia se describe más adelante con más detalle.

1. Debido a su aislamiento lógico, los dominios que residen dentro de una VPC tienen una capa adicional de seguridad en comparación con los dominios que utilizan puntos de conexión públicos.

* Si bien los dominios públicos son accesibles desde cualquier dispositivo conectado a Internet, los dominios VPC requieren algún tipo de VPN o proxy.
* En comparación con los dominios públicos, los dominios de VPC muestran menos información en la consola. En concreto, la ficha **Estado del clúster** no incluye información de partición y la ficha Índices no está presente**.**
* Los extremos de dominio adoptan diferentes formas (https:// search-domain-name vs. https:// **vpc**- domain-name).
* No puede aplicar directivas de acceso basadas en IP a dominios que residen dentro de una VPC porque los grupos de seguridad ya aplican directivas de acceso basadas en IP.

Tenga en cuenta las siguientes limitaciones:

* Si inicia un nuevo dominio dentro de una VPC, no podrá cambiarlo posteriormente para usar un punto de enlace público. Lo contrario también es cierto.
* Puede iniciar su dominio dentro de una VPC o usar un punto de enlace público, pero no puede hacer ambas cosas.
* No puede iniciar su dominio dentro de una VPC que utilice arrendamiento dedicado. Debe utilizar una VPC con el arrendamiento establecido en **Predeterminado**.
* Después de colocar un dominio dentro de una VPC, no puede moverlo a una VPC diferente, pero puede cambiar la configuración de las subredes y del grupo de seguridad.
* Para acceder a la instalación predeterminada de OpenSearch Dashboards para un dominio que reside dentro de una VPC, los usuarios deben tener acceso a la VPC.

### THE ELK STACK

ELK es un acrónimo que describe una combinación popular de proyectos: Elasticsearch, Logstash y Kibana.

La pila ELK le brinda la capacidad de agregar registros de todos sus sistemas y aplicaciones, analizar estos registros y crear visualizaciones.

ELK es útil para visualizar datos de monitoreo de aplicaciones e infraestructura, solución de problemas, análisis de seguridad y más.

### SECURITY

Los dominios de OpenSearch Service ofrecen cifrado de datos en reposo. Utiliza AWS KMS para el almacenamiento y la administración de claves de cifrado. El cifrado utiliza AES-256.

El cifrado también cifra las comunicaciones de nodo a nodo mediante TLS 1.2.

El cifrado de nodo a nodo es opcional y se puede habilitar a través de la consola, la CLI o la API. Una vez que el cifrado de nodo a nodo está habilitado, no se puede deshabilitar. En su lugar, debe crear un nuevo

desde una instantánea sin esta configuración habilitada.

Amazon OpenSearch Service admite tres tipos de políticas de acceso:

* [Resource-based policies](https://docs.aws.amazon.com/opensearch-service/latest/developerguide/ac.html#ac-types-resource)
* [Identity-based policies](https://docs.aws.amazon.com/opensearch-service/latest/developerguide/ac.html#ac-types-identity)
* [IP-based policies](https://docs.aws.amazon.com/opensearch-service/latest/developerguide/ac.html#ac-types-ip)

El control de acceso detallado ofrece capacidades adicionales dentro de Amazon OpenSearch Service. El control de acceso detallado ofrece los siguientes beneficios:

* Control de acceso basado en roles.
* Seguridad a nivel de índice, documento y campo.
* OpenSearch Dashboards multi-tenancy.
* Autenticación básica HTTP para paneles de OpenSearch y OpenSearch. OpenSearch Service admite la autenticación a través de SAML y Amazon Cognito.

## ANALYTICS QUIZ QUESTIONS

Answers and explanations are provided below after the last question in this section.

###### Question 1:

A user is testing a new service that receives location updates from 5,000 rental cars every hour. Which service will collect data and automatically scale to accommodate production workload?

1. Amazon EC2
2. Amazon Kinesis Firehose
3. Amazon EBS
4. Amazon API Gateway

###### Question 2:

Which AWS service can be used to prepare and load data for analytics using an extract, transform and load (ETL) process?

1. AWS Lambda
2. Amazon Athena
3. AWS Glue
4. Amazon EMR

###### Question 3:

A Solutions Architect is designing a solution for a financial application that will receive trading data in large volumes. What is the best solution for ingesting and processing a very large number of data streams in near real time?

1. Amazon EMR
2. Amazon Kinesis Firehose
3. Amazon Redshift
4. Amazon Kinesis Data Streams

###### Question 4:

You have recently enabled Access Logs on your Application Load Balancer (ALB). One of your colleagues would like to process the log files using a hosted Hadoop service. What configuration changes and services can be leveraged to deliver this requirement?

1. Configure Access Logs to be delivered to DynamoDB and use EMR for processing the log files
2. Configure Access Logs to be delivered to S3 and use Kinesis for processing the log files
3. Configure Access Logs to be delivered to S3 and use EMR for processing the log files
4. Configure Access Logs to be delivered to EC2 and install Hadoop for processing the log files

###### Question 5:

A Solutions Architect is designing the messaging and streaming layers of a serverless application. The messaging layer will manage communications between components and the streaming layer will manage real-time analysis and processing of streaming data.

The Architect needs to select the most appropriate AWS services for these functions. Which services should be used for the messaging and streaming layers? (choose 2)

1. Use Amazon Kinesis for collecting, processing and analyzing real-time streaming data
2. Use Amazon EMR for collecting, processing and analyzing real-time streaming data
3. Use Amazon SNS for providing a fully managed messaging service
4. Use Amazon SWF for providing a fully managed messaging service
5. Use Amazon CloudTrail for collecting, processing and analyzing real-time streaming data

### ANALYTICS - ANSWERS

###### Question 1 answer: B

Explanation:

What we need here is a service that can streaming collect streaming data. The only option available is Kinesis Firehose which captures, transforms, and loads streaming data into “destinations” such as S3, RedShift, Elasticsearch and Splunk.

Amazon EC2 is not suitable for collecting streaming data.

EBS is a block-storage service in which you attach volumes to EC2 instances, this does not assist with collecting streaming data (see previous point).

Amazon API Gateway is used for hosting and managing APIs not for receiving streaming data.

###### Question 2 answer: C

Explanation:

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics.

Amazon Elastic Map Reduce (EMR) provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

AWS Lambda is a serverless application that runs code as functions in response to events.

###### Question 3 answer: D

Explanation:

Kinesis Data Streams enables you to build custom applications that process or analyze streaming data for specialized needs. It enables real-time processing of streaming big data and can be used for rapidly moving data off data producers and then continuously processing the data. Kinesis Data Streams stores data for later processing by applications (key difference with Firehose which delivers data directly to AWS services).

Kinesis Firehose can allow transformation of data and it then delivers data to supported services.

RedShift is a data warehouse solution used for analyzing data. EMR is a hosted Hadoop framework that is used for analytics.

###### Question 4 answer: C

Explanation:

Access Logs can be enabled on ALB and configured to store data in an S3 bucket. Amazon EMR is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. EMR utilizes a hosted Hadoop framework running on Amazon EC2 and Amazon S3.

Neither Kinesis nor EC2 provide a hosted Hadoop service.

You cannot configure access logs to be delivered to DynamoDB.

###### Question 5 answer: A,C

Explanation:

Amazon Kinesis makes it easy to collect, process, and analyze real-time streaming data. With Amazon Kinesis Analytics, you can run standard SQL or build entire streaming applications using SQL.

Amazon Simple Notification Service (Amazon SNS) provides a fully managed messaging service for pub/sub patterns using asynchronous event notifications and mobile push notifications for microservices, distributed systems, and serverless applications.

Amazon Elastic Map Reduce runs on EC2 instances so is not serverless.

Amazon Simple Workflow Service is used for executing tasks not sending messages. Amazon CloudTrail is used for recording API activity on your account.