nuclio learning

Análisis de datos para Big Data en AWS

Prof: Arturo Lorenzo Hernández.

Arturo Lorenzo Hernández

Arquitecto de Datos en Iskaypet

Estudié Ingeniería Informática en la Universidad de Murcia, España.
Llevo más de 3 años trabajando en el mundo de la ciencia de datos.
Empecé como ingeniero de datos y actualmente desempeño tareas de arquitectura de datos en AWS.

Grupo IskayPet®
The best for animals, the best for all

PREGUNTA

¿Qué vas a conseguir con este curso?

- Conocer los conceptos principales en Big Data y Análisis de datos
- Dominar los recursos cloud de AWS para el análisis de datos Big Data
- Poner en práctica los conocimientos aprendidos con un proyecto real
- Guía para conseguir trabajo a día de hoy con estas tecnologías

nuclio learning

ÍNDICE

Análisis de datos Big Data en AWS

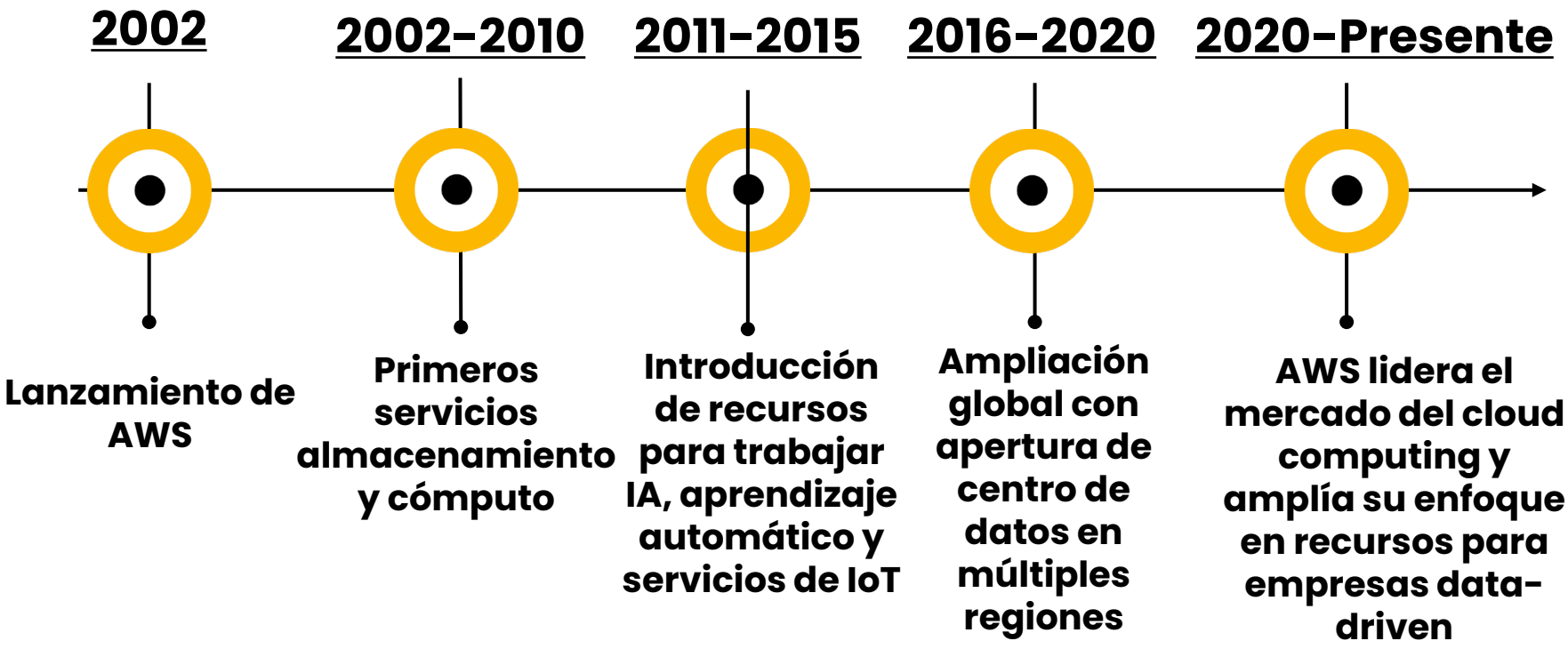
nuclio learning

01

Fundamentos de AWS y Big Data

Introducción a AWS

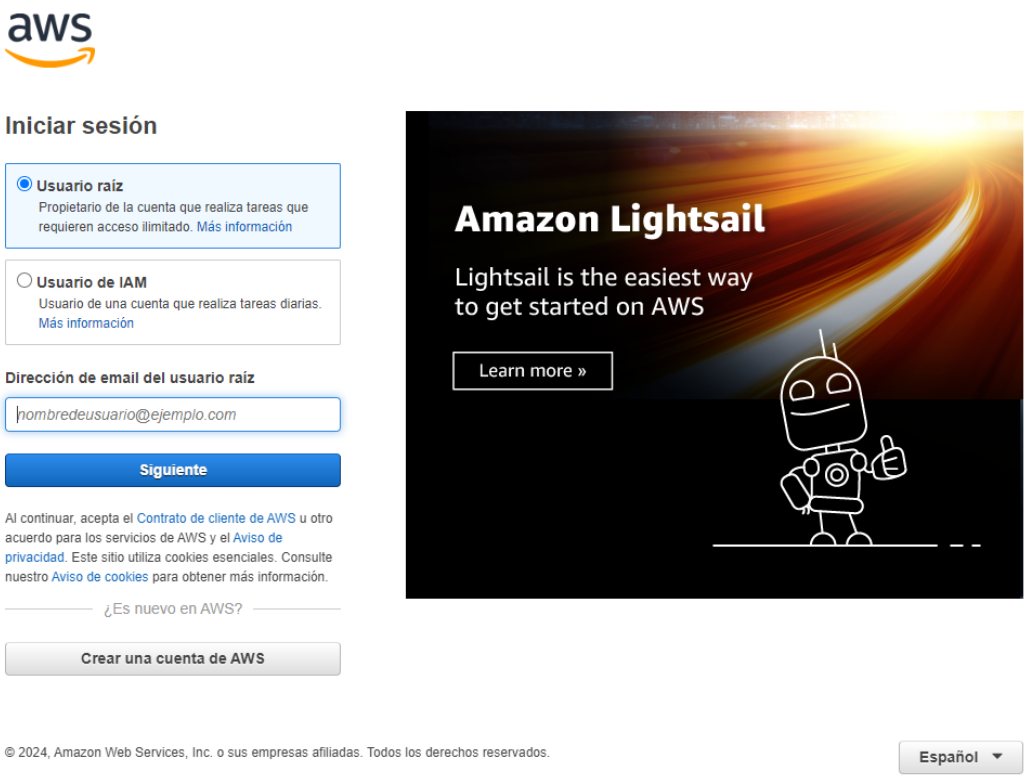
nuclio learning


2002
Lanzamiento de AWS
2002-2010
Primeros servicios almacenamiento y cómputo
2011-2015
Introducción de recursos para trabajar IA, aprendizaje automático y servicios de IoT
2016-2020
Ampliación global con apertura de centro de datos en múltiples regiones
2020-Presente
AWS lidera el mercado del cloud computing y amplía su enfoque en recursos para empresas data-driven

Categorías de recursos

Computación	- Amazon EC2 - Amazon Lambda - Amazon ECS
Almacenamiento	- Amazon S3 - Amazon EBS
Bases de datos	- Amazon RDS - Amazon DynamoDB - Amazon Redshift
Analítica	- Amazon Glue - Amazon EMR - Amazon Kinesis - Amazon Athena
Servicios de mensajería	- Amazon SNS - Amazon SQS

¿Creamos nuestra primera cuenta en AWS?



nuclio learning

01

Fundamentos de AWS y Big Data

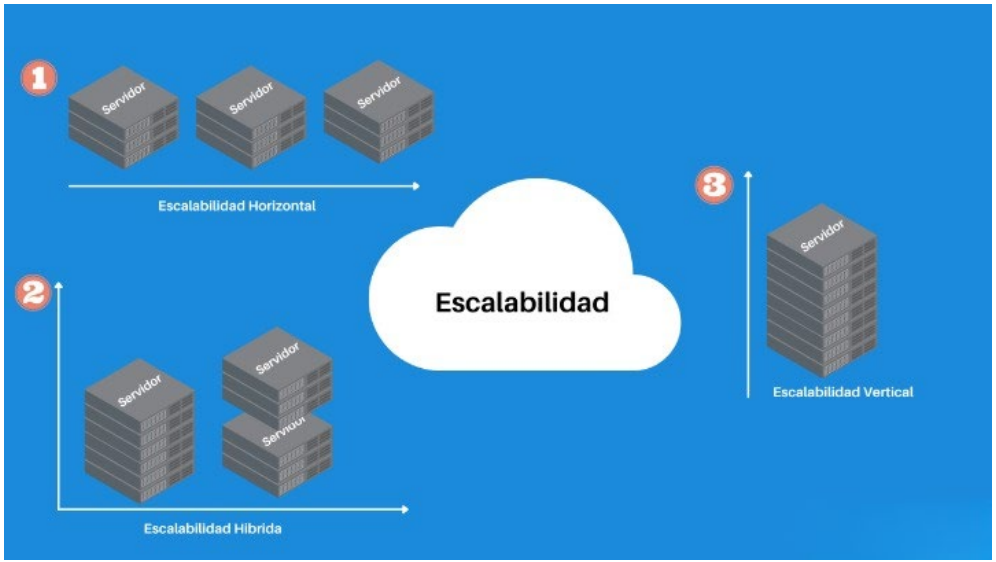
¿Cómo se relaciona el análisis de datos con AWS?

nuclio learning

Escalabilidad, elasticidad y pago por uso

Escalabilidad

Puedes aumentar o disminuir el uso de los recursos en la nube con un par de clicks y sin necesidad de comprar infraestructura física.



Escalabilidad, elasticidad y pago por uso

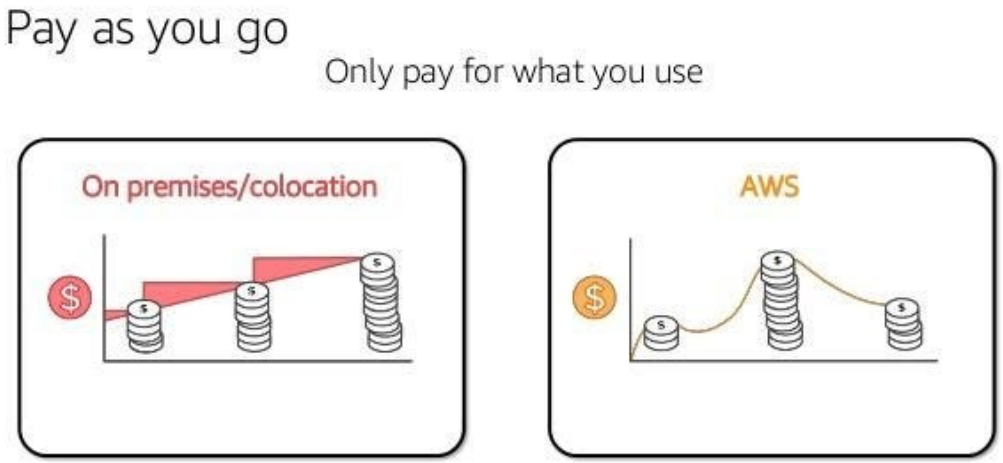
Elasticidad

Los recursos en la nube te permiten que tus servidores escalan de forma automática según la demanda, muy potente cuando no sabes qué uso van a tener tus servidores.

Escalabilidad, elasticidad y pago por uso

Pago por uso

Los usuarios solo pagan por los recursos que consumen, si solo quieres tener un servicio encendido un par de días solo vas a pagar por ese consumo.



Recursos top para análisis de datos

Amazon S3

AWS Lambda

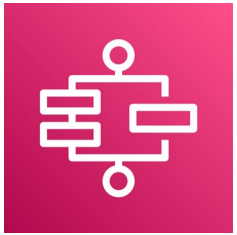
AWS Glue

Recursos top para análisis de datos

Amazon Athena

Quicksight

Step functions



01

Fundamentos de AWS y Big Data

Análisis de datos en el contexto de Big Data

nuclio learning

- La **cantidad y variedad de formato de datos** han aumentado mucho en los últimos años lo que ha llevado a las empresas a la necesidad de utilizar nuevas técnicas y tecnologías que antes no existían, para poder recopilar, limpiar y analizar los datos de forma masiva.

- Big Data se refiere a **conjuntos de datos extremadamente grandes y complejos** que son difíciles de procesar y analizar utilizando herramientas de gestión de datos tradicionales, por ello se requieren de herramientas avanzadas.

Características principales Big Data (5Vs)

Volumen

Cantidad masiva de datos generados cada segundo desde diversas fuentes.

Velocidad

Rapidez con la que se generan y procesan los datos para cumplir con las demandas.

Valor

Utilidad y relevancia de los datos para la toma de decisiones empresariales.

Variedad

Diversidad de tipos y fuentes de datos, tanto estructurados como no estructurados.

Veracidad

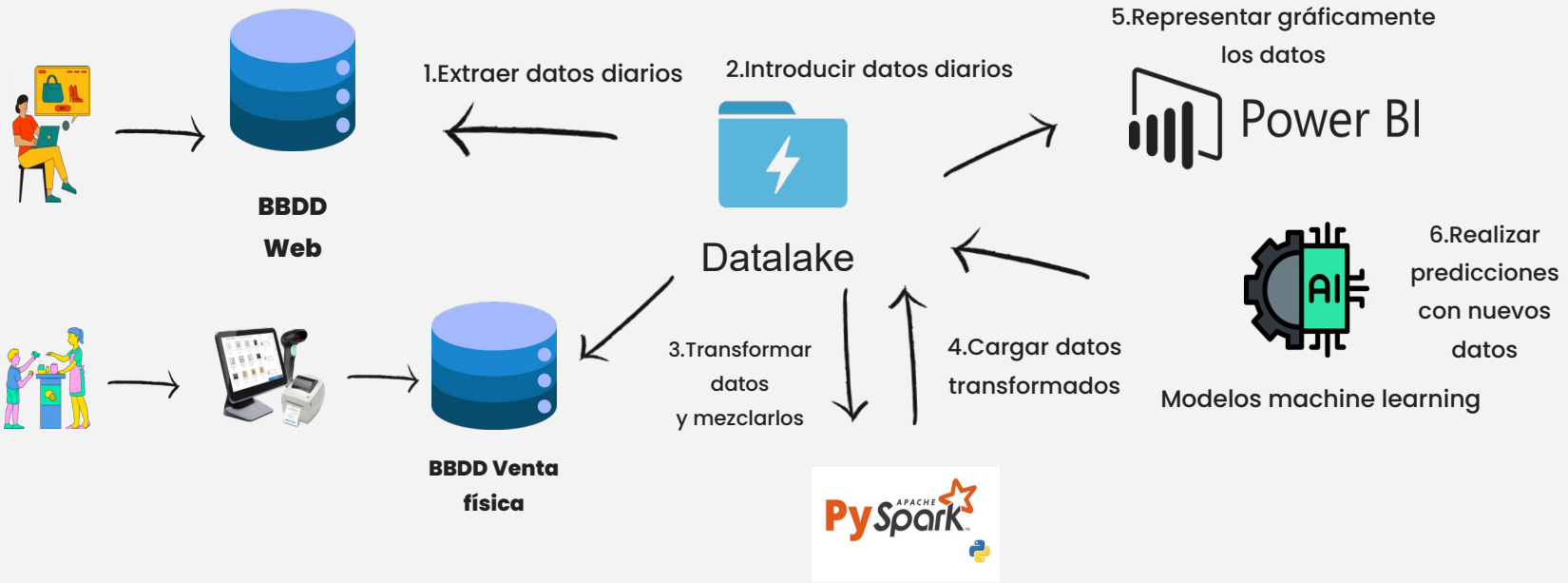
Calidad y precisión de los datos, importante para asegurar resultados fiables.

- AWS ofrece servicios de análisis de datos completos, seguros, escalables y rentables. **Los servicios de análisis de AWS se adaptan a todas las necesidades de análisis de grandes volúmenes de datos:** movimiento de datos, almacenamiento, lagos de datos, análisis de macrodatos, machine learning...

- **El análisis de datos ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades**. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más fidelizados.

- Optimización de inventarios
- Optimización de campañas de marketing digital
- Detección de tendencias de mercado bursátil
- Optimización de Precios
- Reducción del Churn de clientes
- Optimización de la cadena de suministro
- Personalización de contenidos en medios digitales
- Análisis de sentimientos en redes sociales

Flujo de datos de compra en una empresa



02

Almacenamiento de datos en AWS

Amazon S3: Almacenamiento escalable y seguro

nuclio learning

- **Amazon S3 (Simple Storage Service) es un servicio de almacenamiento de objetos altamente escalable, duradero y seguro** que permite almacenar y recuperar cualquier cantidad de datos desde cualquier lugar en la web.

- Amazon S3 es como un sistema de ficheros tradicional de cualquier ordenador, permite almacenar *objetos* (ficheros) en *buckets* (directorios) a cualquier escala ya que los datos siguen un patrón de arquitectura distribuida.

Algunas características

- Un bucket debe contener un nombre único de forma global.
- El nombre de un bucket se escribe en minúscula sin barras bajas y debe tener entre 3-63 caracteres.
- Un objeto en S3 contiene una key que se define por la ruta completa en el bucket hacia ese objeto (prefijo + nombre del objeto).
- El máximo de tamaño de un objeto es de 5TB, si se quiere subir un objeto de más peso se tiene que usar la opción de *muti-part upload*

Clases almacenamiento S3

- **S3 Standard**: Uso para datos accedidos frecuentemente, muy poca latencia de recuperación de datos.

- **S3 Standard Infrequent Access**: Uso para datos con menos frecuencia de acceso pero que se necesita un rápido acceso cuando se requieren.

- **S3 Glacier**: Almacenamiento low-cost para archivos de backup o archive.
    - *Glacier Instant Retrieval*: Recuperación de datos en milisegundos, recomendado para acceso a datos de forma trimestral.
    - *Glacier Flexible Retrieval*: Recuperación de datos entre 1 minuto y 12 horas dependiendo de la clase a la que recupere.
    - *Glacier Deep Archive*: Recuperación de datos entre 12 y 48 horas.

Clases almacenamiento S3

- **S3 Intelligent-Tiering**: Uso para datos que no sabes qué tipo de frecuencia de acceso van a tener, los datos se mueven entre clases de forma inteligente.
  - Standard Frequent: Tier por defecto.
  - Standard Infrequent Access: Objetos no recuperados en 30 días.
  - Glacier instant retrieval: Objetos no recuperados en 90 días.
  - Glacier flexible retrieval: Configurable para objetos no recuperados entre 90 y +700 días.
  - Glacier deep archive: Configurable para objetos no recuperados entre 180 y +700 días.

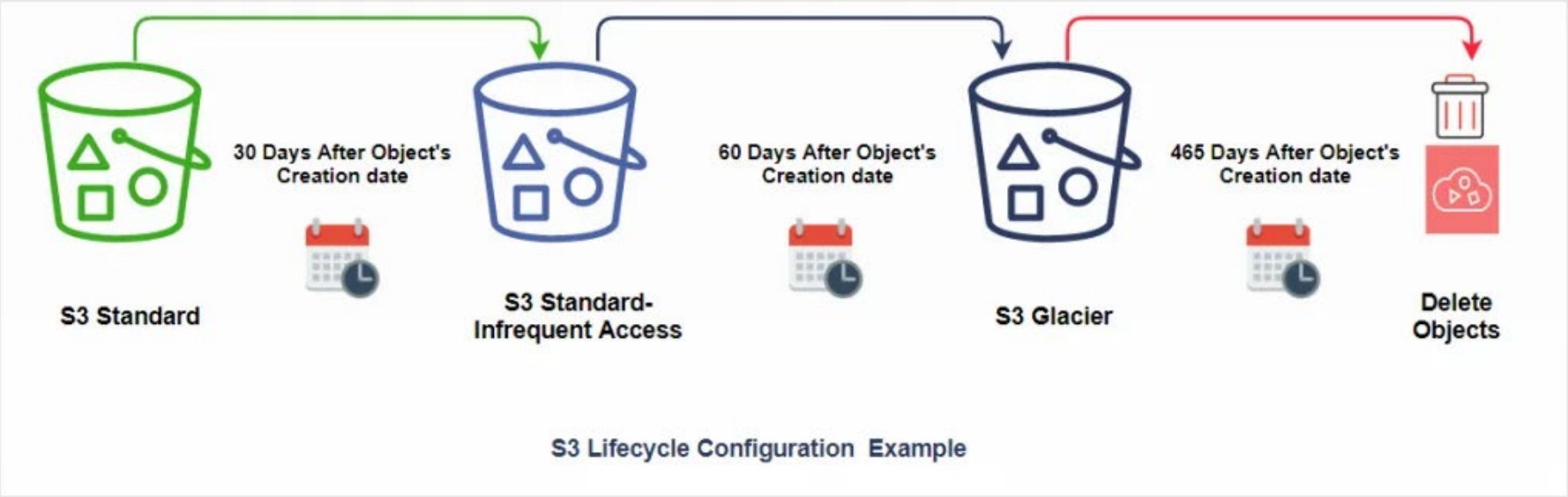
Elegir una clase u otra depende de las necesidades actuales y las diferencias principales es el tiempo de recuperación y el precio de una clase u otra.

Ciclo de vida en S3

- El ciclo de vida en Amazon S3 se refiere a las **políticas de ciclo de vida** (Lifecycle Policies) que **gestionan la transición y expiración de objetos** almacenados en S3 de manera automatizada.

- Estas políticas permiten a los usuarios definir reglas para **trasladar objetos entre diferentes clases de almacenamiento** y eliminar objetos innecesarios, optimizando costos y gestionando eficientemente los datos a lo largo del tiempo.

Ciclo de vida en S3



Versionado de datos en S3

- El versionado en Amazon S3 es una característica que permite mantener **múltiples versiones de un objeto en un bucket**. Esto facilita la protección contra la pérdida accidental de datos y la recuperación de versiones anteriores de objetos.

- Supongamos que tienes un bucket llamado "mis-datos-importantes" y subes un archivo llamado "datos.csv". Si habilitas el versionado y luego cargas una nueva versión de "datos.csv", S3 mantendrá ambas versiones del archivo. **Puedes listar las versiones del objeto y recuperar la versión que necesites.**

02

Almacenamiento de datos en AWS

Casos de uso de servicios de almacenamiento en Big Data

nuclio learning

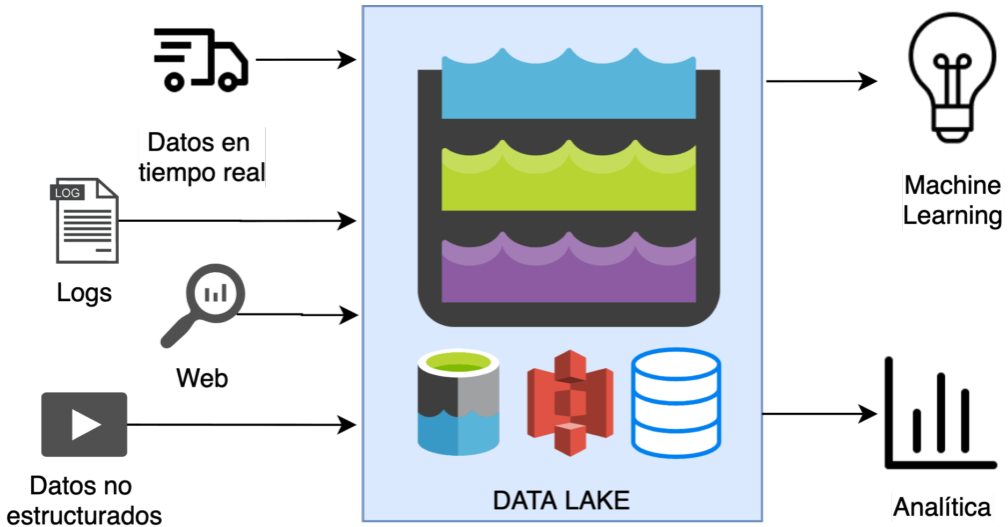
Casos de uso de S3

- Backup y almacenamiento
- Disaster recovery
- Archive
- Application hosting
- Media hosting
- Software delivery
- Data lakes y big data analytics
- Static website

Datalake para análisis Big Data

Datalake

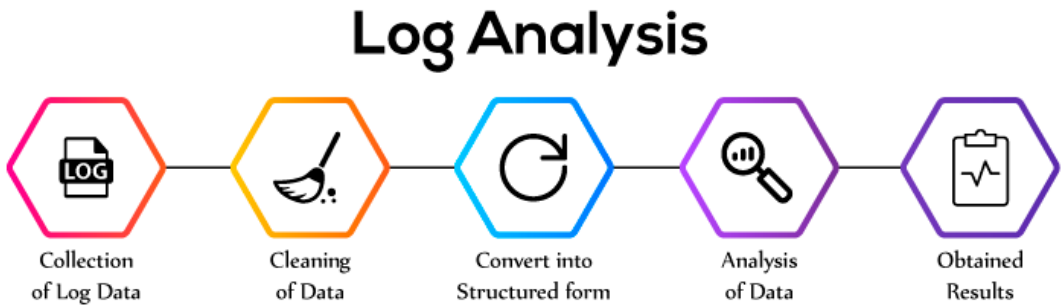
Un Data Lake es un almacenamiento centralizado que permite guardar todos los datos de una organización sin importar su tipo o formato. Es flexible, escalable y facilita el acceso y análisis de grandes volúmenes de datos, proporcionando una base poderosa para la toma de decisiones basadas en datos.



Almacenamiento de logs

Análisis de logs

El análisis de logs es fundamental en Big Data porque proporciona una visión detallada y en tiempo real del estado del sistema, comportamiento del usuario y el rendimiento de las aplicaciones. Esta información es vital para tomar decisiones informadas que beneficien tanto a la operación técnica como a la estrategia de negocio.



03

Ingesta de datos en S3

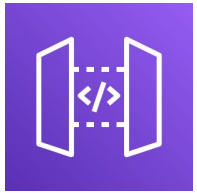
AWS Lambda: Ejecución de código serverless

nuclio learning

- En AWS (Amazon Web Services), una función lambda **es un servicio serverless que te permite ejecutar código sin tener que aprovisionar ni administrar servidores**. Puedes utilizar funciones lambda para ejecutar código en respuesta a eventos como cambios en datos, acciones del usuario o invocaciones de API.

- Estas funciones se pueden configurar para ejecutarse automáticamente en respuesta a eventos específicos, como subida de archivos a un bucket de Amazon S3, inserción de datos en una tabla de Amazon DynamoDB, o invocación de una API Gateway.

Integraciones más usadas con AWS Lambda



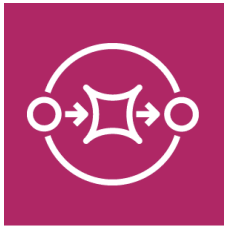
API Gateway

Kinesis

S3

Eventbridge

DynamoDB



Simple Queue Service (SQS)

Algunas características

- El pago se hace por petición y tiempo de cómputo. El servicio es muy barato, casi gratis en la mayoría de casos de uso.
- Se puede programar con muchos lenguajes, como Node.js, Python, Java, C# o Ruby.
- Se puede configurar con una memoria RAM de 128MB - 10GB
- Capacidad de almacenamiento en disco temporal de 512MB - 10GB
- Máximo de ejecuciones concurrentes de 1000 y máximo de tiempo de ejecución de 15 minutos.

03. ¿Qué utilidad tiene una función lambda en el análisis de datos?

- Preprocesamiento de datos
- Ingesta de datos desde distintos orígenes
- Transformaciones sencillas de pocos datos
- Orquestación de flujo de trabajo
- Manejo de metadatos (estados de ejecuciones)
- Envíos de correos electrónicos con información analítica (Lambda + Athena + SNS)
- Automatización de configuraciones para procesos big data

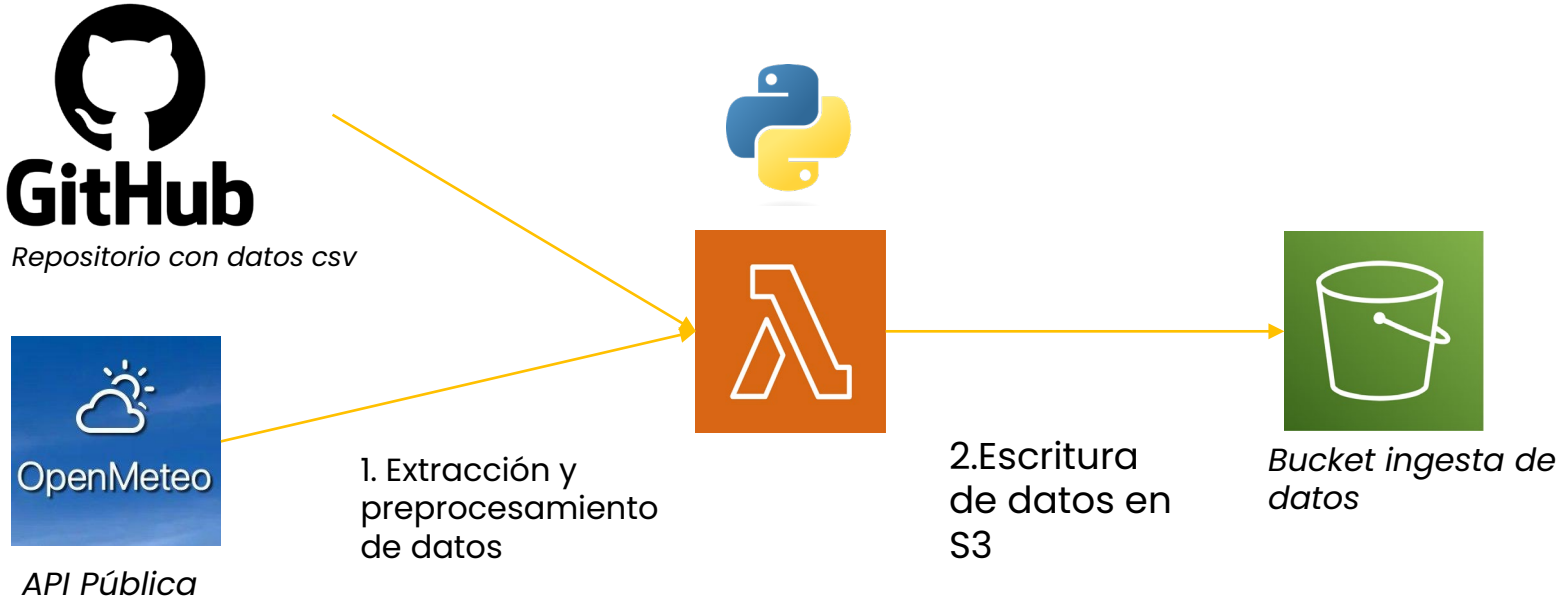
03

Ingesta de datos en S3

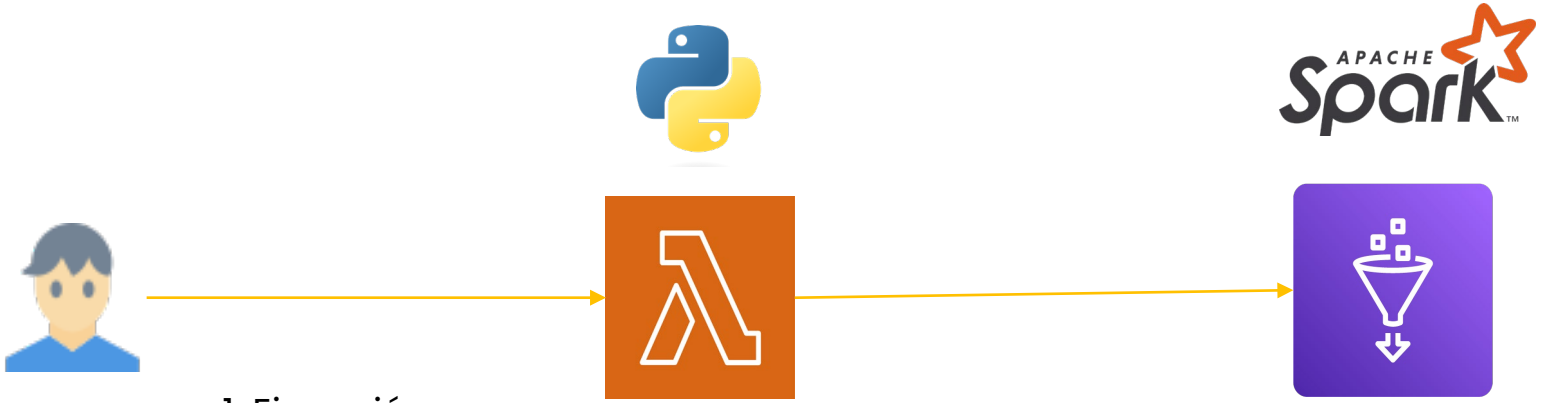
Casos de uso reales de funciones lambda en análisis de datos

nuclio learning

Ingesta de datos en datalake



Orquestación de procesos big data



1. Ejecución con información sobre el tamaño de proceso spark

2. Cálculo de recursos (memoria y cpu)

3. Envío de proceso spark con configuraciones optimizadas a Glue

04

PROCESAMIENTO DE DATOS EN AWS

Amazon Glue Notebooks: Procesamiento distribuido de datos con PySpark

nuclio learning

- El procesamiento de datos distribuido es una metodología en la cual **las tareas de procesamiento de datos se dividen y se ejecutan en múltiples nodos o máquinas dentro de una red**. Esto se hace para manejar grandes volúmenes de datos de manera más eficiente y rápida.

- La necesidad de esta metodología surge del problema que se tenía al intentar procesar datos con una única máquina, aunque tuviese paralelismo multinúcleo y multihilo, para el procesamiento de gigabytes, terabytes y petabytes se hace una tarea imposible para los tiempos que se requieren.

Un poco de historia:

Los primeros días: Computación centralizada

En las décadas de **1950 a 1970**, **el procesamiento de datos se realizaba principalmente en grandes mainframes**. Estas máquinas centralizadas eran capaces de manejar grandes cantidades de datos, pero tenían limitaciones en cuanto a escalabilidad y costos.

Surgimiento del procesamiento distribuido

Durante las décadas de **1980 y 1990**, **la aparición de las redes de computadoras marcó el inicio de la transición hacia el procesamiento distribuido**. Se comenzaron a experimentar con clústeres de computadoras, donde varias máquinas trabajaban juntas para completar tareas de procesamiento de datos.

Un poco de historia:

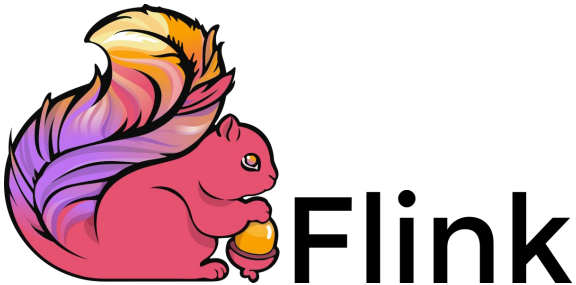
La revolución de Hadoop

Inspirado por el Google File System (GFS), Doug Cutting y Mike Cafarella crearon el framework Hadoop en los 2000s, **esta herramienta utiliza el Hadoop Distributed File System (HDFS) y el modelo de programación MapReduce**, que dividía las tareas en sub-tareas más pequeñas que podían ser procesadas en paralelo por múltiples nodos.

Expansión y diversificación

En la década de 2010, se produjo una expansión y diversificación significativa en las tecnologías de procesamiento de datos. **Apache Spark fue introducido en 2014, ofreciendo procesamiento en memoria que era mucho más rápido que Hadoop** y capaz de manejar tanto datos en lotes como en tiempo real.

Tecnologías más usadas



Características de Spark más top

- Apache Spark es conocido por su **procesamiento en memoria**, realiza operaciones a velocidades mucho más altas que los sistemas tradicionales.
- **Ofrece un modelo de programación unificado** con APIs en Java, Scala, Python y R, facilitando el desarrollo en diversos lenguajes y soportando procesamiento batch, consultas interactivas, análisis en tiempo real y machine learning.
- Spark **se integra fácilmente con Hadoop**, utilizando HDFS y otros componentes del ecosistema Hadoop.
- Los RDDs (Resilient Distributed Datasets) en Spark son **estructuras de datos distribuidas** que permiten el procesamiento paralelo de grandes conjuntos de datos, ofreciendo tolerancia a fallos mediante su capacidad de reconstrucción automática.

Conceptos clave de Spark

Concepto	Descripción
SparkSession	Punto de entrada principal para trabajar con datos estructurados en Spark. Se utiliza para crear DataFrames, leer archivos, configurar opciones de Spark, entre otras tareas.
SparkContext	Contexto principal de Spark que sirve para establecer configuraciones mayormente. Se utiliza para crear RDDs, acceder a configuraciones de Spark, administrar recursos y controlar la ejecución de aplicaciones Spark en un clúster
Driver	El proceso principal de Spark que coordina las operaciones y programa la ejecución del trabajo en el clúster.

Conceptos clave de Spark

Conceptos	Descripción
Executors	Procesos que realizan el trabajo real de procesamiento de datos en el clúster, ejecutando tareas asignadas por el Driver.
Cluster manager	Gestor de clúster que coordina la asignación de recursos y la administración de los nodos en el clúster de Spark
RDD (Resilient Distributed Dataset)	Un conjunto de datos distribuido e inmutable que puede ser procesado en paralelo en un clúster de nodos Spark.

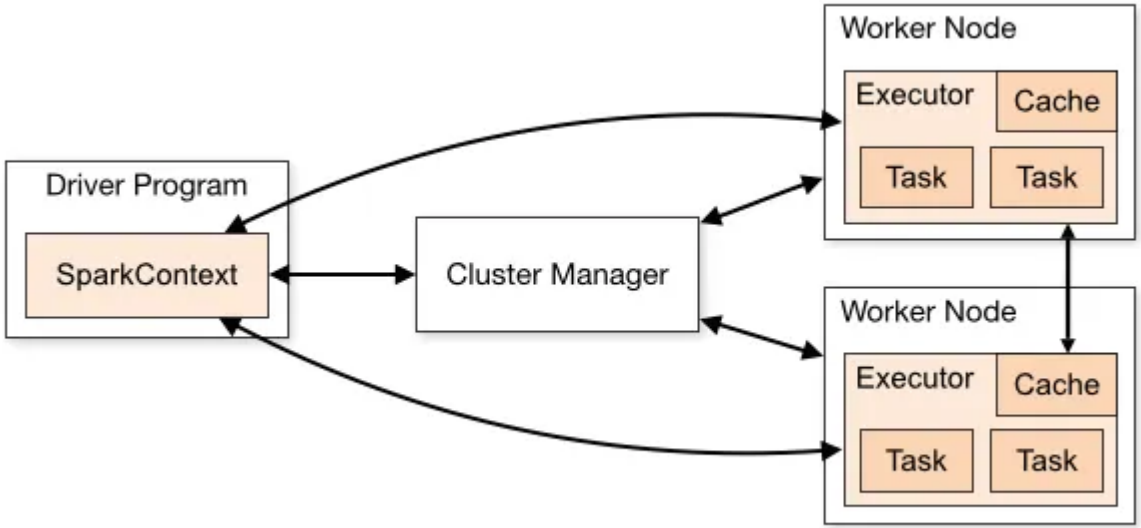
Conceptos clave de Spark

Conceptos	Descripción
DataFrames	Una abstracción de datos distribuida que organiza los datos en columnas con un esquema definido y permite realizar operaciones SQL y análisis de datos
Datasets	Similar a los DataFrames, pero con tipos de datos fuertemente tipados, lo que proporciona más seguridad y optimización de rendimiento.
Transformaciones	Operaciones que transforman un conjunto de datos en otro como ‘map’, ‘filter’, ‘join’, etc...

Conceptos clave de Spark

Conceptos	Descripción
Acciones	Operaciones que inician la ejecución de un plan y devuelven resultados al Driver, como ‘collect’, ‘count’, ‘write’ etc.

Arquitectura de Spark



- Glue Notebooks es una herramienta de Amazon Web Services (AWS) que forma parte de AWS Glue, un **servicio de procesamiento de datos distribuido serverless que facilita la preparación y carga de datos para análisis**.

- AWS Glue Notebooks permite a los usuarios **crear, editar y ejecutar scripts interactivos en Jupyter notebooks**, una popular plataforma de código abierto para la computación interactiva y colaborativa.

Beneficios de Glue Notebooks

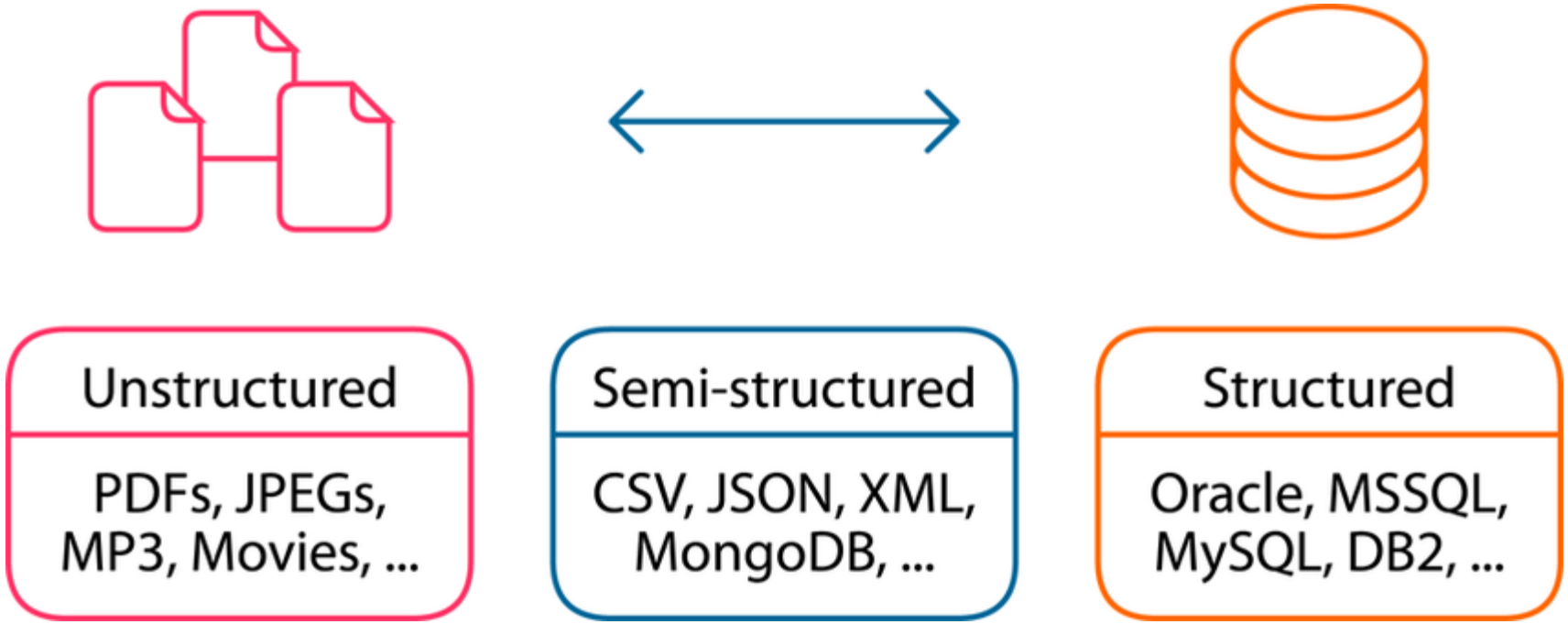
- **Agilidad**: Facilita el desarrollo rápido de soluciones de datos con resultados inmediatos, mejorando la productividad de los equipos.
- **Escalabilidad**: Permite manejar grandes volúmenes de datos sin preocuparse por la infraestructura, adaptándose a las necesidades crecientes.
- **Eficiencia**: Reduce significativamente el tiempo necesario para la preparación y transformación de datos, optimizando los flujos de trabajo.
- **Colaboración**: Proporciona un entorno compartido donde los equipos de datos pueden trabajar juntos de manera efectiva, mejorando la cooperación y la comunicación.

04

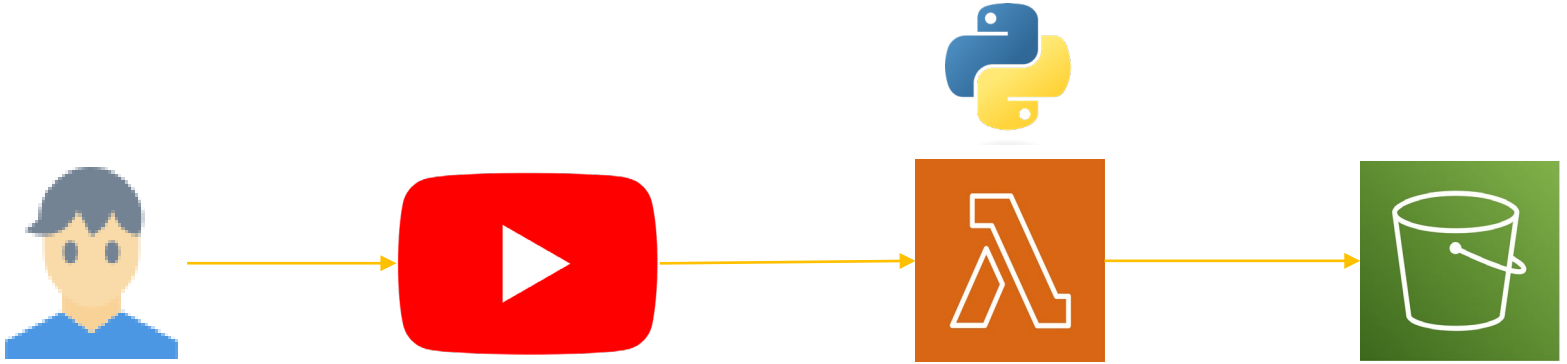
PROCESAMIENTO DE DATOS EN AWS

Ejemplos prácticos de procesamiento de datos en AWS.

nuclio learning


Unstructured
PDFs, JPEGs, MP3, Movies, ...
Semi-structured
CSV, JSON, XML, MongoDB, ...
Structured
Oracle, MSSQL, MySQL, DB2, ...

PROCESAMIENTO DE IMÁGENES

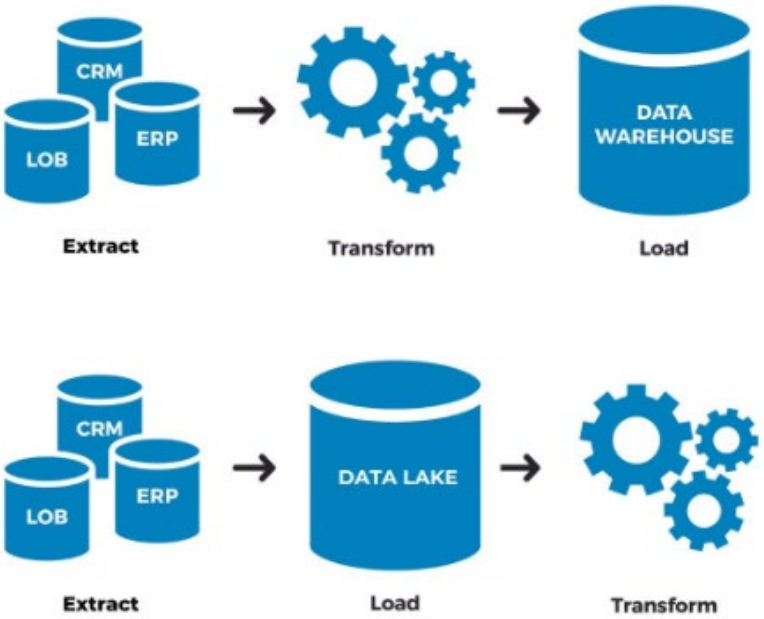


1. Usuario sube vídeo con foto de miniatura

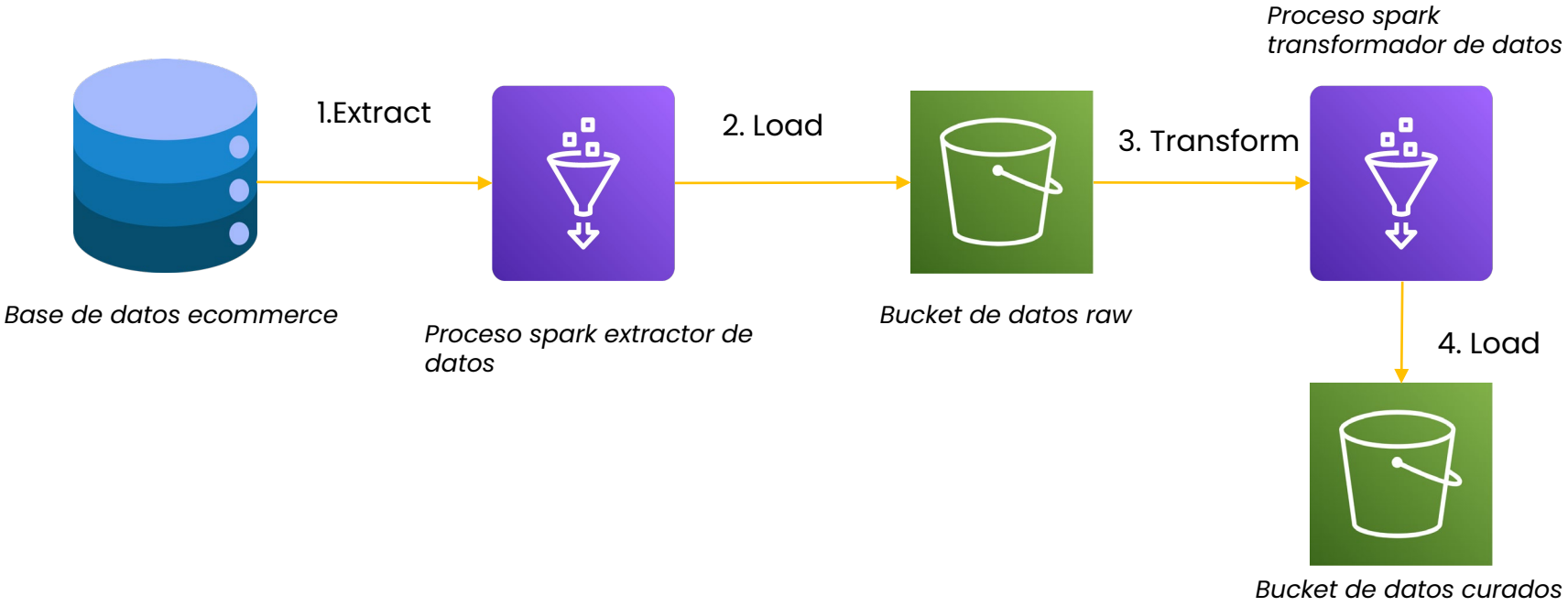
2. Extracción de imagen de miniatura y reducción de dimensión a thumbnail

3. Carga de foto redimensionada en bucket de S3

ETL vs ELT


CRM
LOB
ERP
DATA WAREHOUSE
Extract
Transform
Load
CRM
LOB
ERP
DATA LAKE
Extract
Load
Transform

CASO DE USO ELTL



05

CONSULTA DE DATOS EN AWS

Amazon Glue Data Catalog: Meta Almacén de esquemas y bases de datos.

nuclio learning

Meta almacén de datos

Meta almacén

Un meta almacén de datos, o catálogo de datos, es esencial en Big Data porque organiza y describe los ficheros del datalake.

Facilita la búsqueda y el descubrimiento de datos y el cumplimiento de políticas. Además, ayuda a los equipos a trabajar juntos de manera más eficiente al compartir una visión clara y accesible de los datos.

¿Por qué es importante un meta almacén?

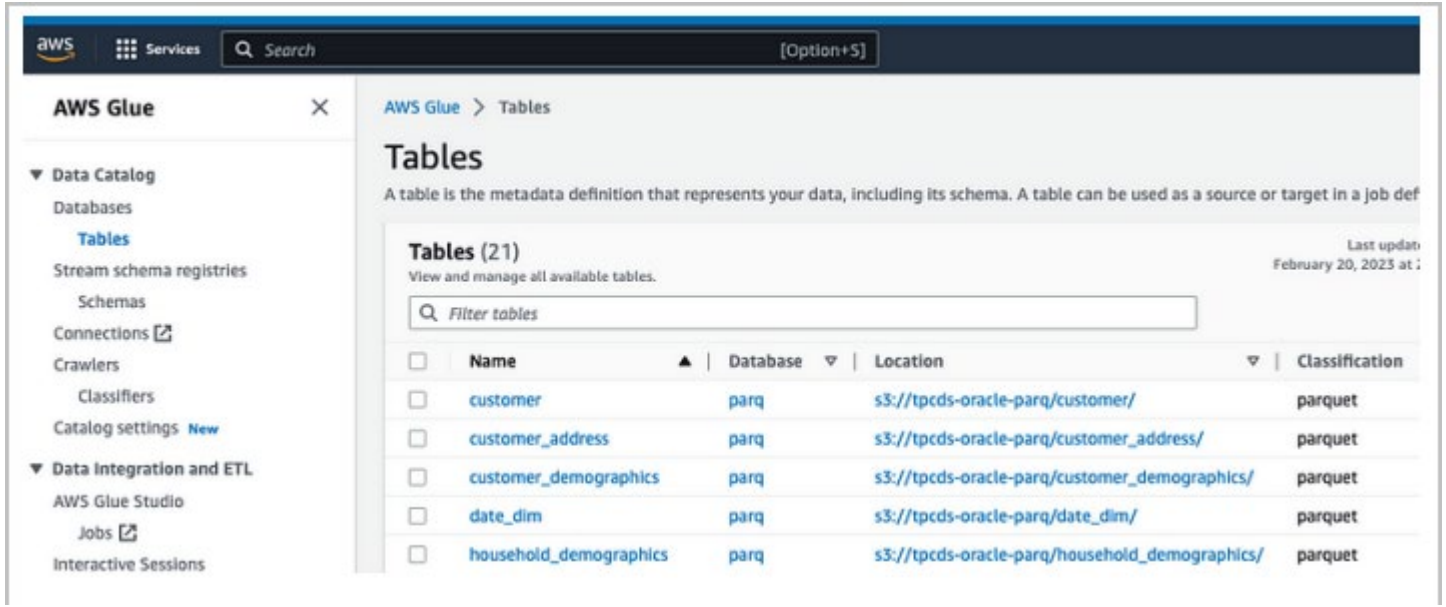
La necesidad de un meta almacén en big data surge por varias razones:

- No tenemos el concepto de base de datos, tablas ni esquema.
- Para referirnos a una “tabla” de la única forma que podemos es a través de su ruta de S3.

- Glue Data Catalog es un servicio de AWS que **centraliza y gestiona metadatos de diversas fuentes de datos, la principal es S3**. Permite descubrir y organizar datos, creando una base de metadatos con información sobre ubicaciones, esquemas y versionados de tablas, facilitando así llevar un registro de todos los cambios hechos en cada tabla.

- Además ofrece **integración con otros servicios de AWS como Athena y Redshift** para poder explorar los datos con queries SQL y permitiendo un análisis más amigable.

Ejemplo de tablas en Glue Data Catalog



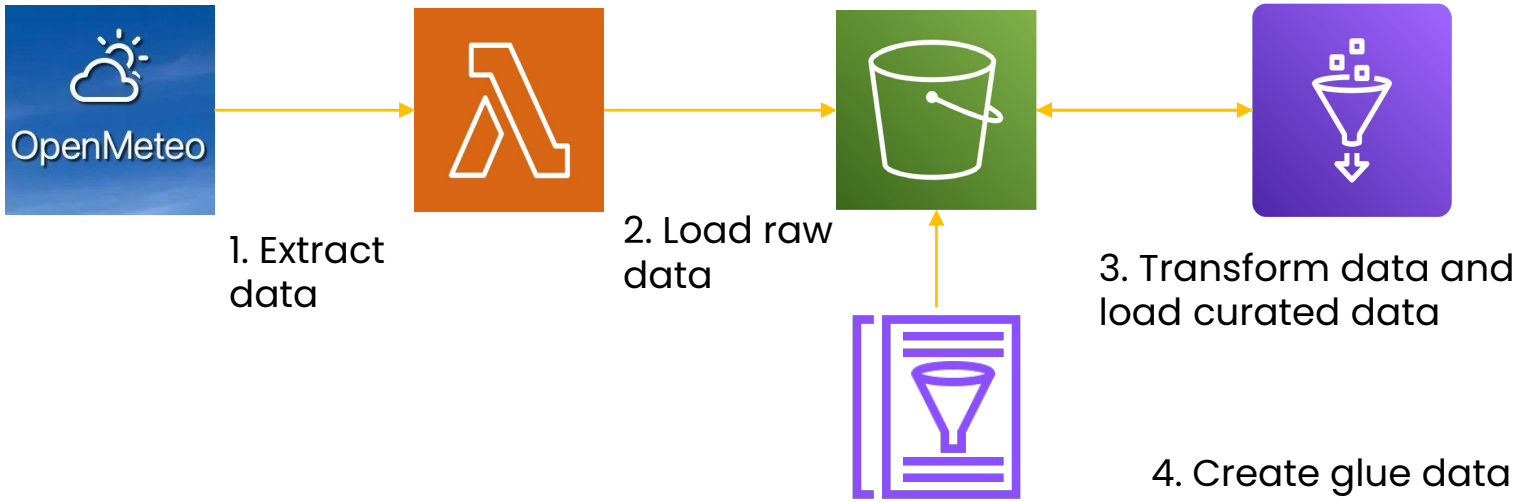
05

CONSULTA DE DATOS EN AWS

Amazon Athena: Consultas ad hoc en datos almacenados en S3

nuclio learning

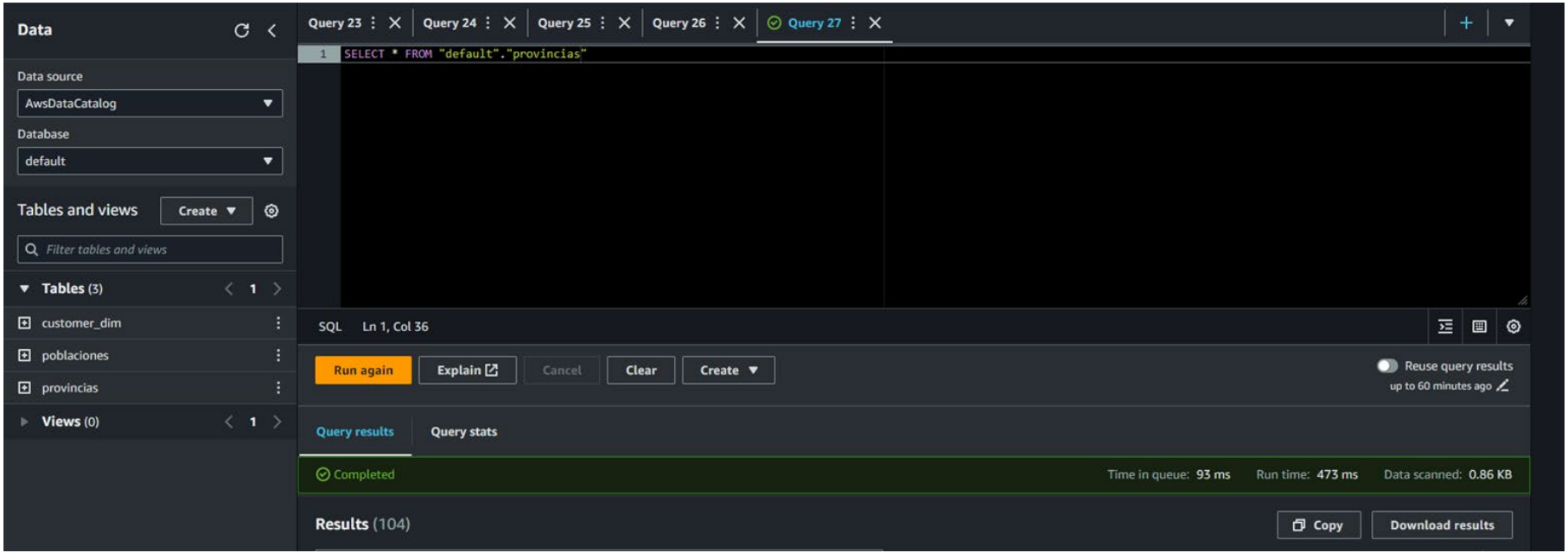
¿Cómo analizamos los datos transformados?



- Amazon Athena es un **servicio de consulta interactiva que permite analizar datos en Amazon S3** utilizando SQL estándar. Es serverless, por lo que no requiere administrar infraestructura, y cobra solo por las consultas ejecutadas.

- Athena se **integra con AWS Glue Data Catalog para gestionar metadatos y facilitar el análisis de datos** estructurados, semiestructurados y no estructurados. Es ideal para análisis ad-hoc y puede procesar grandes volúmenes de datos rápidamente..

Ejemplo de query en Amazon Athena

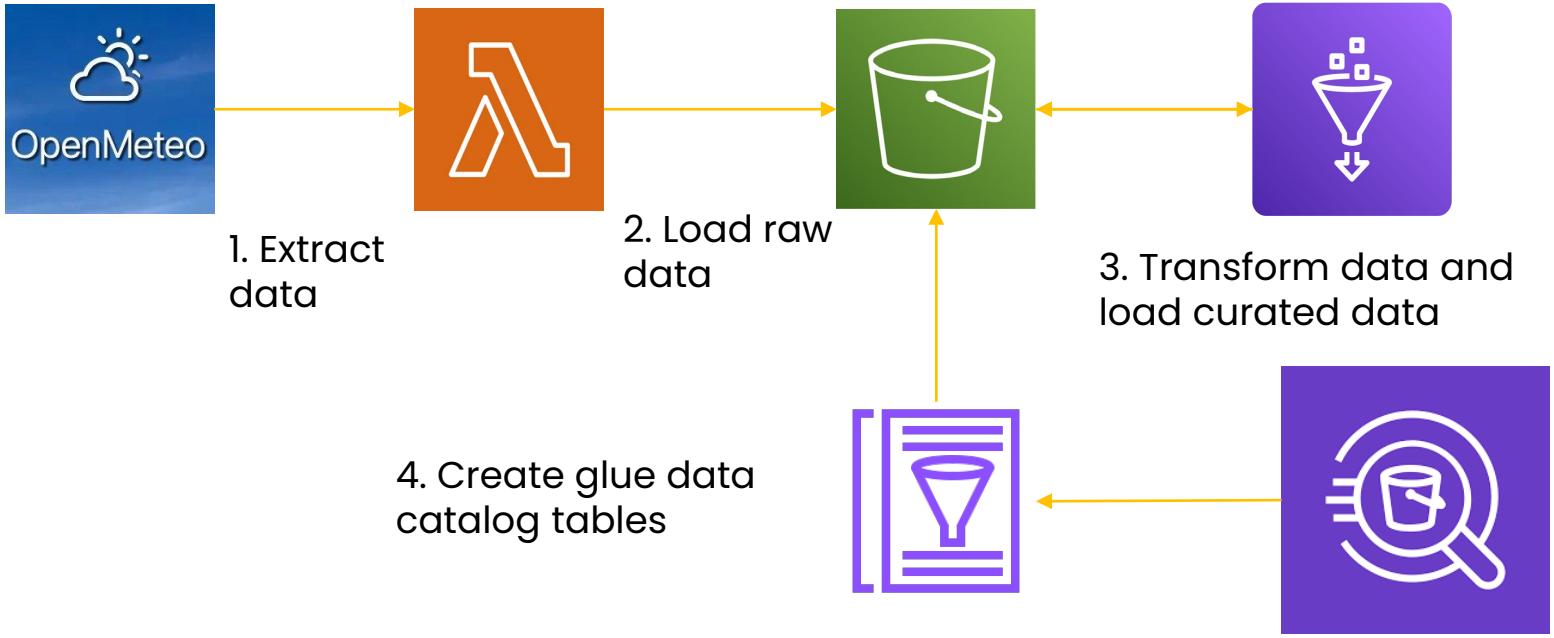


Características de Athena

- Aunque por defecto la fuente de datos de Athena es Glue Data Catalog, también ofrece multitud de conectores hacia otras fuentes.
- Tiene un precio de 5.00$ por cada TB de datos escaneado.
- Soporta lectura de ficheros parquet, csv, json, orc y avro.
- Permite descarga de resultados de queries en formato csv, guardado de queries y exportación de resultados hacia un bucket de S3.
- El lenguaje de consultas es SQL, está construido sobre el motor de consultas presto.

  Podemos encontrar las funciones built-in en este enlace de AWS
  https://docs.aws.amazon.com/athena/latest/ug/functions.html

¿Cómo analizamos los datos transformados?



06

VISUALIZACIÓN DE DATOS Y COMUNICACIÓN DE RESULTADOS

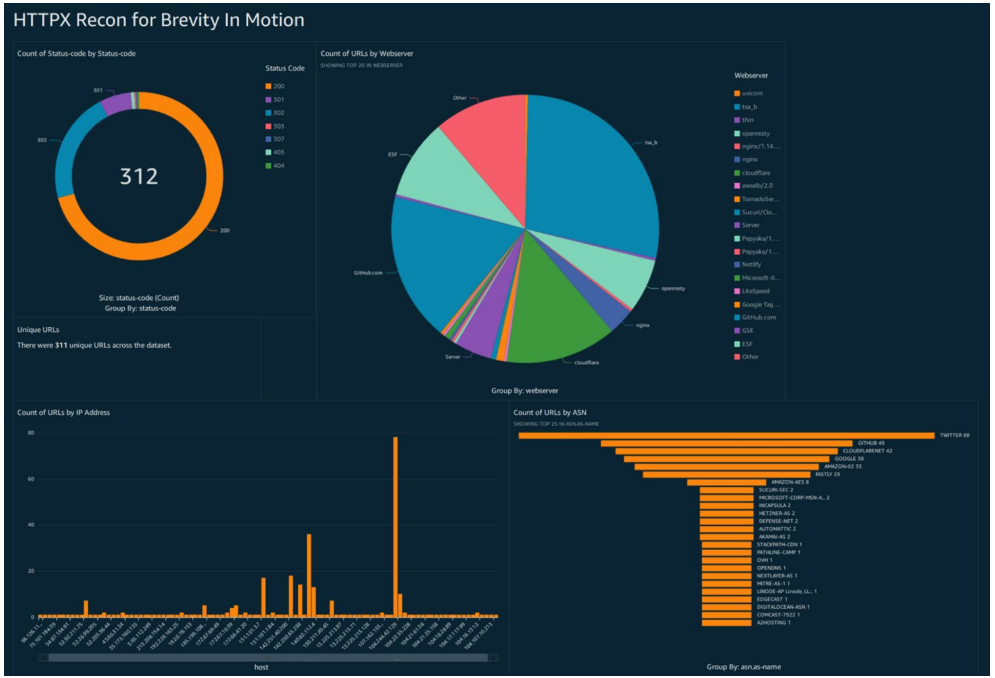
Amazon QuickSight: Herramienta de reporting

nuclio learning

Business Intelligence

Dashboard

Un dashboard es una herramienta visual que muestra datos clave y métricas de rendimiento en un formato fácilmente comprensible. Facilitan la toma de decisiones, mejoran la eficiencia operativa, permiten un seguimiento continuo de objetivos y estrategias, y ayudan a identificar tendencias y problemas rápidamente.



nuclio learning

Tipos de gráficos más comunes

Barras

Comparar diferentes categorías o grupos, mostrando cantidades o frecuencias de manera clara

Pastel

Representan proporciones o partes de un todo, mostrando la contribución de cada categoría a un total

Líneas

Visualizar tendencias a lo largo del tiempo, permitiendo un seguimiento fácil de cambios y patrones

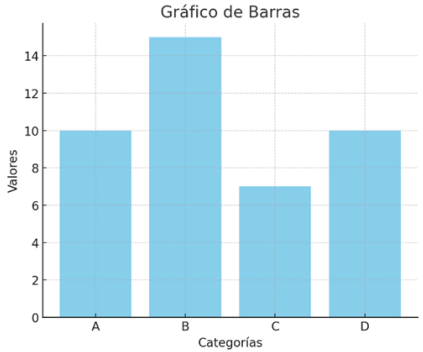
Área

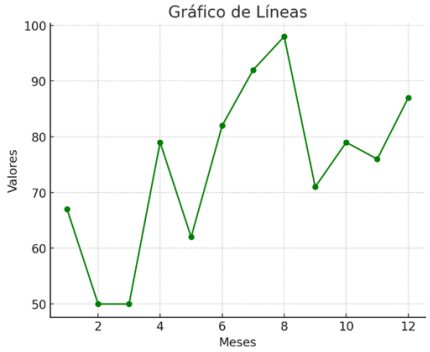
Similares a los gráficos de líneas, pero con áreas sombreadas debajo de las líneas. Utilizados para mostrar acumulaciones

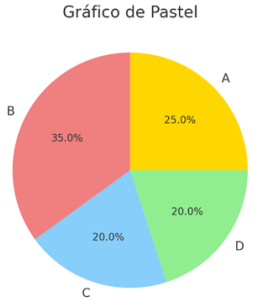
Dispersión

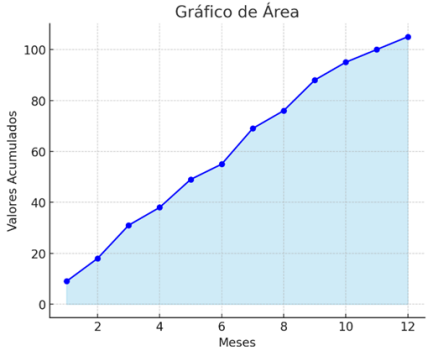
Muestran la relación entre dos variables mediante puntos distribuidos en el gráfico.

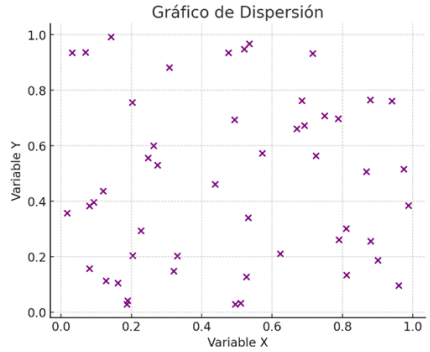
Tipos de gráficos más comunes











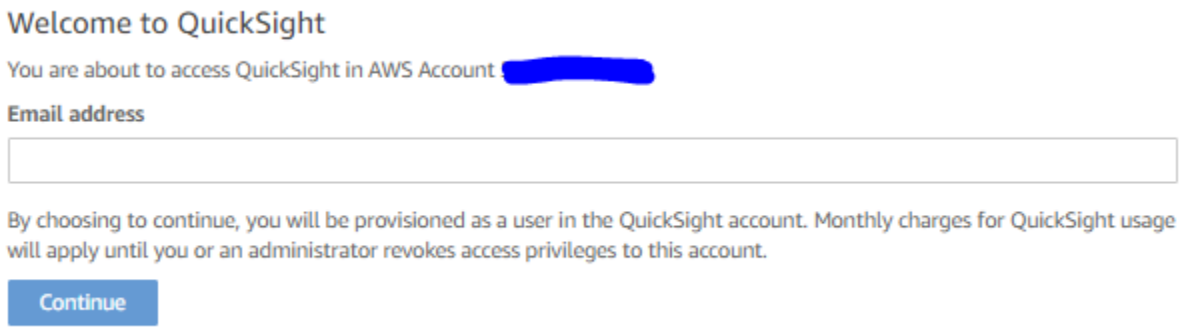
- Amazon QuickSight es un servicio de inteligencia empresarial basado en la nube ofrecido por Amazon Web Services (AWS). **Permite a los usuarios crear y publicar dashboards** que incluyen visualizaciones de datos, gráficos y tablas, accesibles desde cualquier dispositivo.

- Además, QuickSight se integra fácilmente con otros servicios de AWS y diversas fuentes de datos, facilitando la consolidación y análisis de información heterogénea. Su capacidad de autoescalado y su modelo de pago por usuario lo convierten en una opción flexible y económica para empresas de todos los tamaños.

Características de Quicksight

- Existen dos modelos de Quicksight la versión standard con usuarios individuales y la versión enterprise que ofrece grupos de usuarios.
- En la versión enterprise exiten dos roles diferentes, autores con un precio de 24$/mes y los lectores con un precio de 3$/mes
- Quicksight puede graficar datos de cualquier recurso de almacenamiento de AWS y además de fuentes externas como bases de datos on-premise, Jira, Salesforce, ficheros CSVs.
- Ofrece una herramienta que hace los reportes rápidos y escalables llamada SPICE que ofrece cómputo en memoria de los datos almacenados en Quicksight.

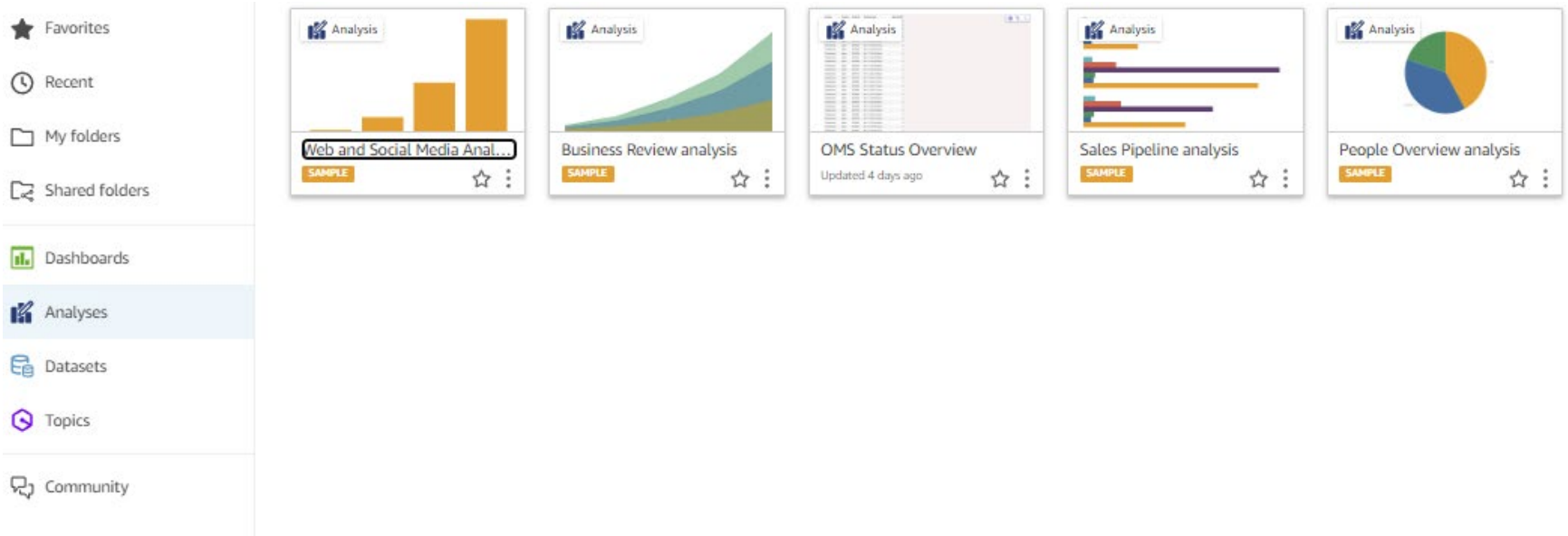
Para crear un usuario en Quicksight tenemos que acceder al recurso desde el portal y **registrarnos con nuestro correo electrónico** asociado a la cuenta de AWS.



Conceptos clave de Quicksight

- **Datasets:** Conjuntos de datos importados o conectados que se utilizan para crear visualizaciones y análisis en QuickSight.
- **Analyses:** Entornos interactivos donde los usuarios exploran y visualizan datos utilizando gráficos y tablas
- **Dashboards:** Paneles interactivos y compartibles que presentan visualizaciones y KPIs derivados de los análisis.
- **Security & Permissions:** Configuraciones que controlan el acceso y los permisos de usuarios y grupos para asegurar la protección de los datos y recursos.

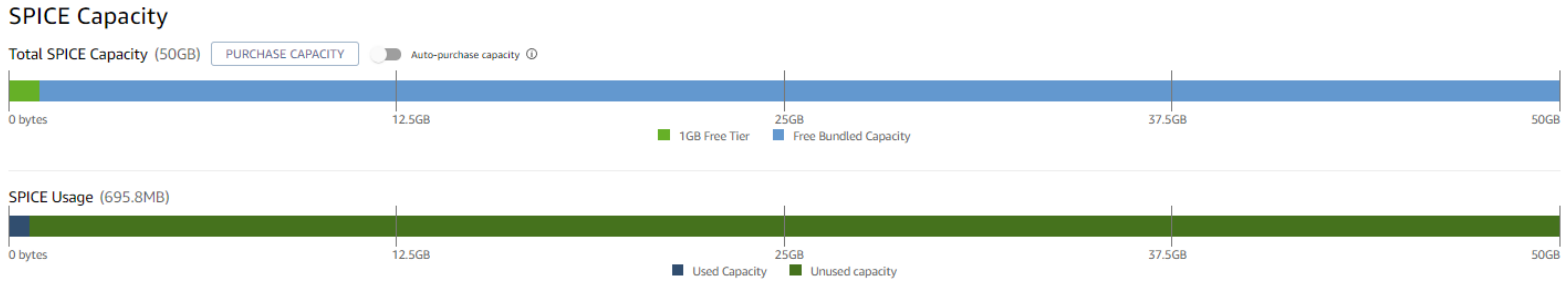
Página principal Quicksight



Mejora de performance en Quicksight

SPICE (Super-fast, Parallel, In-memory Calculation Engine) es el motor de análisis en memoria de Amazon QuickSight que permite realizar consultas rápidas y análisis interactivos.

Mejora los reportes al proporcionar tiempos de respuesta más rápidos, capacidad de manejar grandes volúmenes de datos y reducción de carga en las bases de datos fuentes



06

VISUALIZACIÓN DE DATOS Y COMUNICACIÓN DE RESULTADOS

Casos de uso típicos de dashboard en empresas

nuclio learning

Dashboard ventas por producto

Dashboard de entregas de mercancía

nuclio learning

Dashboard recruitment en RRHH

nuclio learning

07

ORQUESTACIÓN DE FLUJOS DE TRABAJO

AWS Step functions: Orquestación de flujos de trabajo

nuclio learning

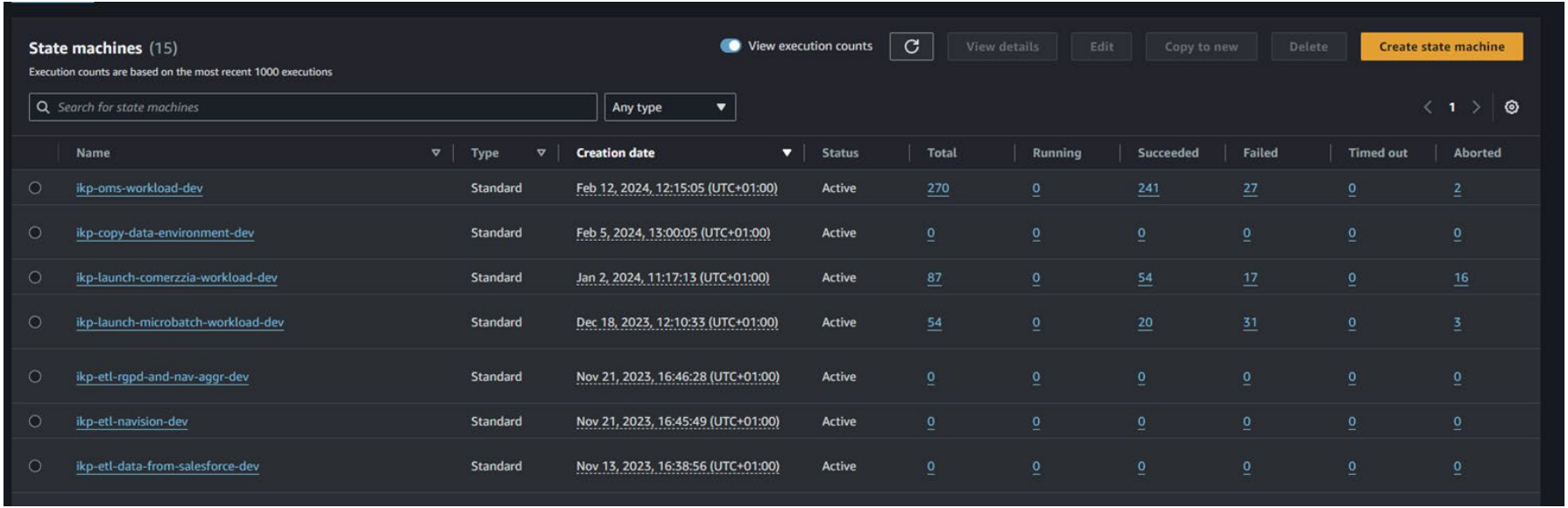
- La orquestación de un flujo de trabajo es la **coordinación y ejecución secuencial de varias tareas** o procesos interrelacionados.

- AWS Step Functions es un servicio que facilita la **orquestación de estos flujos de trabajo al permitir definir, ejecutar y monitorear pasos individuales de manera eficiente.**

  Ofrece una manera sencilla de coordinar tareas de los distintos servicios de AWS, lo que simplifica el desarrollo y la automatización de procesos complejos como puede ser el de procesamiento de datos.

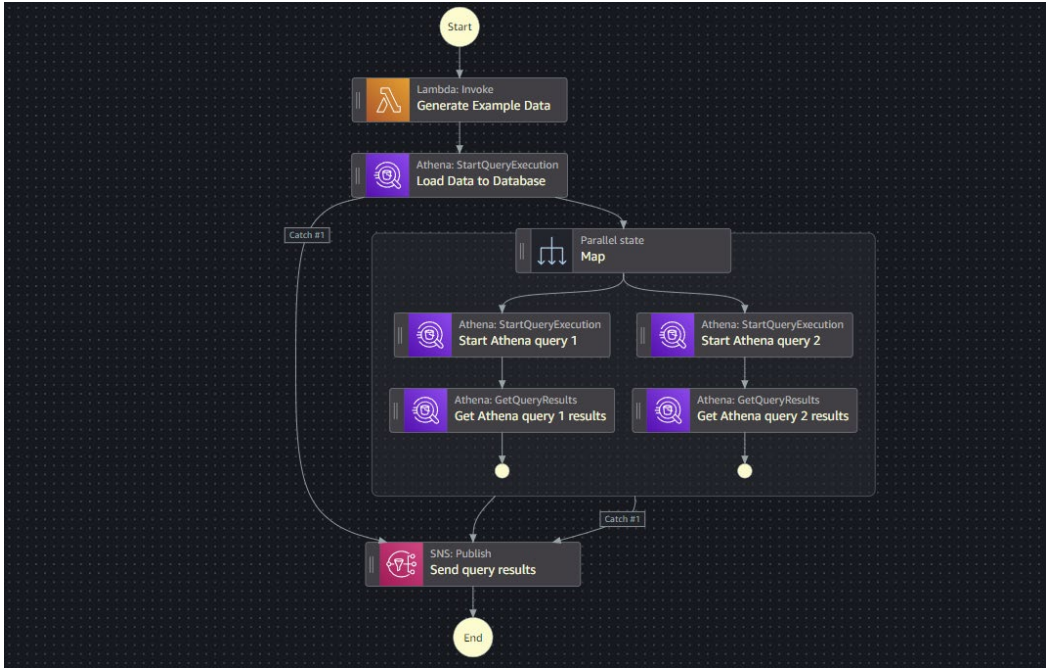
- Flujos de trabajo de ETL
- Análisis de datos en tiempo real
- Generación de informes automatizados
- Procesamiento de datos de eventos
- Generación de facturas automáticas ecommerce
- Gestión de llamadas a microservicios
- Preprocesamiento y entrenamiento de datos para modelos de machine learning

Página principal del servicio Step Functions



Name	Type	Creation date	Status	Total	Running	Succeeded	Failed	Timed out	Aborted
ikp-oms-workload-dev	Standard	Feb 12, 2024, 12:15:05 (UTC+01:00)	Active	270	0	241	27	0	2
ikp-copy-data-environment-dev	Standard	Feb 5, 2024, 13:00:05 (UTC+01:00)	Active	0	0	0	0	0	0
ikp-launch-comerzzia-workload-dev	Standard	Jan 2, 2024, 11:17:13 (UTC+01:00)	Active	87	0	54	17	0	16
ikp-launch-microbatch-workload-dev	Standard	Dec 18, 2023, 12:10:33 (UTC+01:00)	Active	54	0	20	31	0	3
ikp-etl-rgpd-and-nav-aggr-dev	Standard	Nov 21, 2023, 16:46:28 (UTC+01:00)	Active	0	0	0	0	0	0
ikp-etl-navision-dev	Standard	Nov 21, 2023, 16:45:49 (UTC+01:00)	Active	0	0	0	0	0	0
ikp-etl-data-from-salesforce-dev	Standard	Nov 13, 2023, 16:38:56 (UTC+01:00)	Active	0	0	0	0	0	0

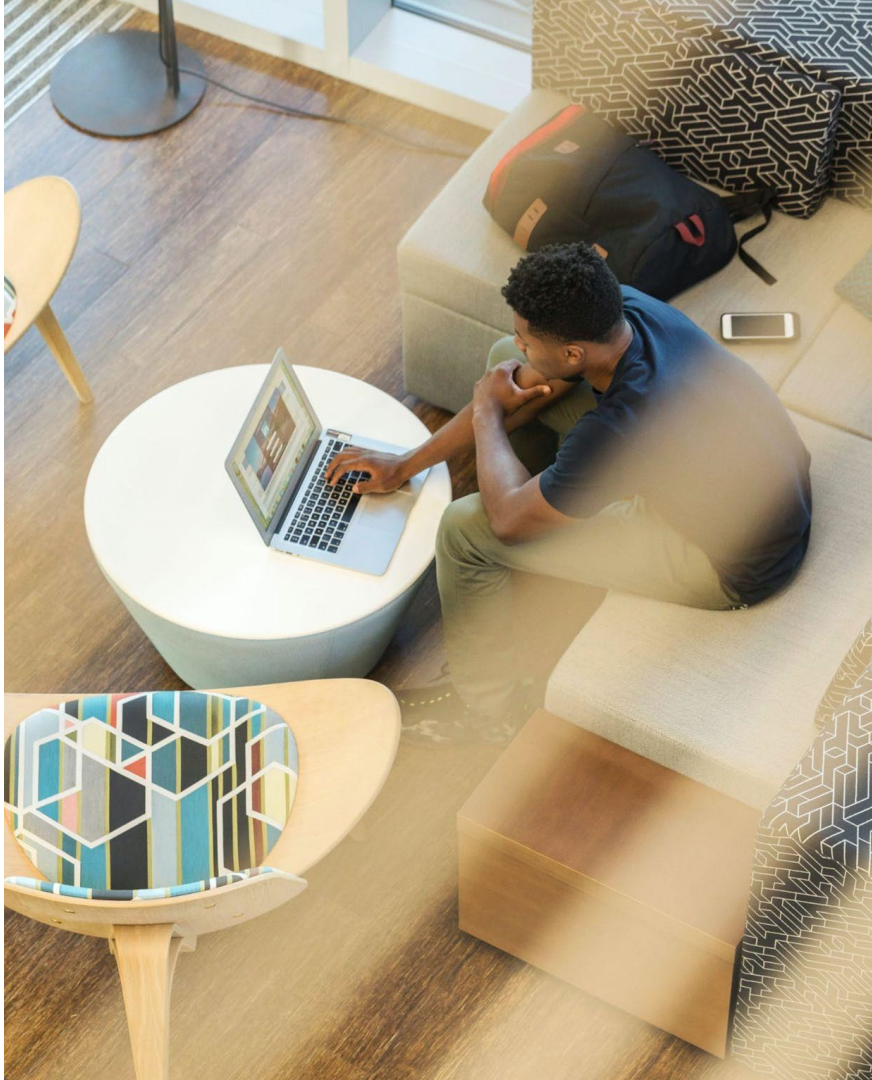
Orquestación de ingesta y consulta de datos



08

RECOMENDACIONES FINALES Y PRÓXIMOS PASOS

Siguientes pasos como analista de datos en cloud



nuclio learning

Roles más comunes en ciencia de datos

- Ingeniero de datos
- Arquitecto de datos
- Analista de datos
- Cientifico de datos
- Ingeniero de machine learning

1. Enfócate en un sólo rol:

Lo más importante para este sector es aprender las tecnologías más demandadas para uno de los roles.

Céntrate en esas tecnologías enfocadas a un puesto de trabajo.

2. Linkedin:

La plataforma para encontrar trabajo más grande del mundo es Linkedin, si no tienes un perfil atractivo y no eres activo en la plataforma compartiendo tus logros o pensamientos te va a costar más encontrar un buen trabajo.

3. Piensa en grande:

Intenta siempre conectar las ideas que hay detrás de cada tecnología y qué problema resuelve en las empresas, de esta manera te será mucho más sencillo avanzar en tu carrera y encontrar trabajos cada más interesantes.

4. Síndrome del impostor:

No te abrumes por lo complejo y lo profundo que es el mundo tecnológico, todos empezamos por la base y nos bloqueamos horas y horas con problemas que después nos parecen muy sencillos.

La frustración es parte de la tecnología.

5. La historia importa

La ciencia de datos tiene una historia muy corta por el momento, esto es una ventaja ya que si entendemos como empezaron estas tecnologías nos costará mucho menos entender el presente.

6. Date tiempo

Aprender una nueva profesión lleva tiempo y en este sector no es diferente, es más, cuesta lo suyo.

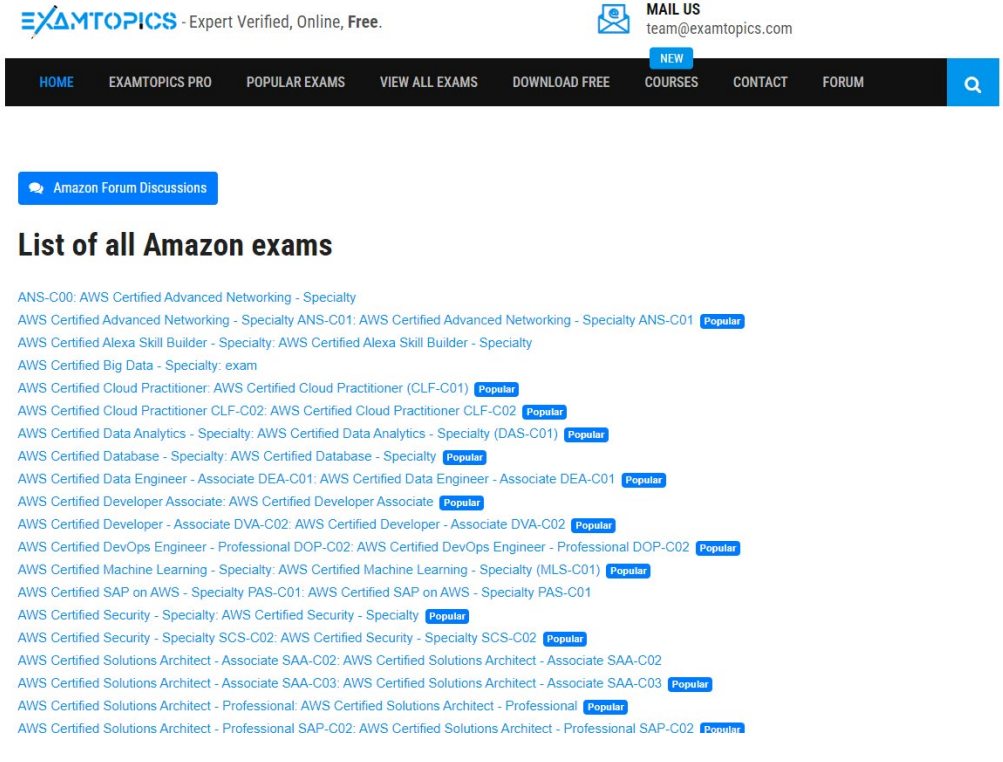
Rodéate de los mejores profesores y mentores para acortar el proceso lo más rápido posible.

- Para validar tus conocimientos en cloud antes las empresas **es muy recomendable conseguir las certificaciones de pago** que ofrecen las distintas plataformas cloud.

- Cada plataforma cloud tiene sus **caminos de certificaciones dependiendo de que rol vas a desempeñar**, no son las mismas certificaciones para un ingeniero de datos que para un científico de datos o para un desarrollador web. Además estas empresas te ofrecen contenido teórico y práctica gratis para estudiarlas.

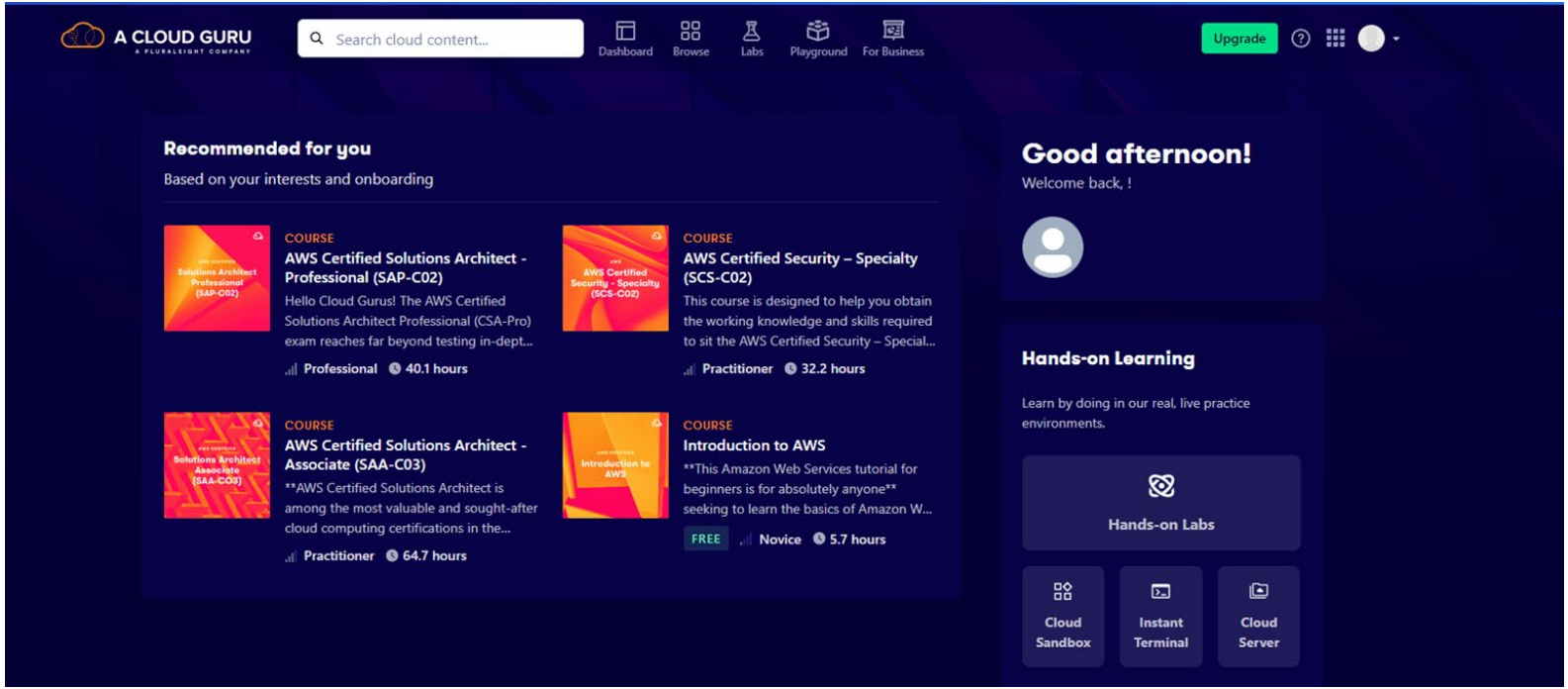
  Las certificaciones no son muy caras, pero **podemos encontrar muchas veces Vouchers que ofrecen descuentos** sobre todo para la nube de Microsoft, Azure.

Páginas para practicar exámenes gratis

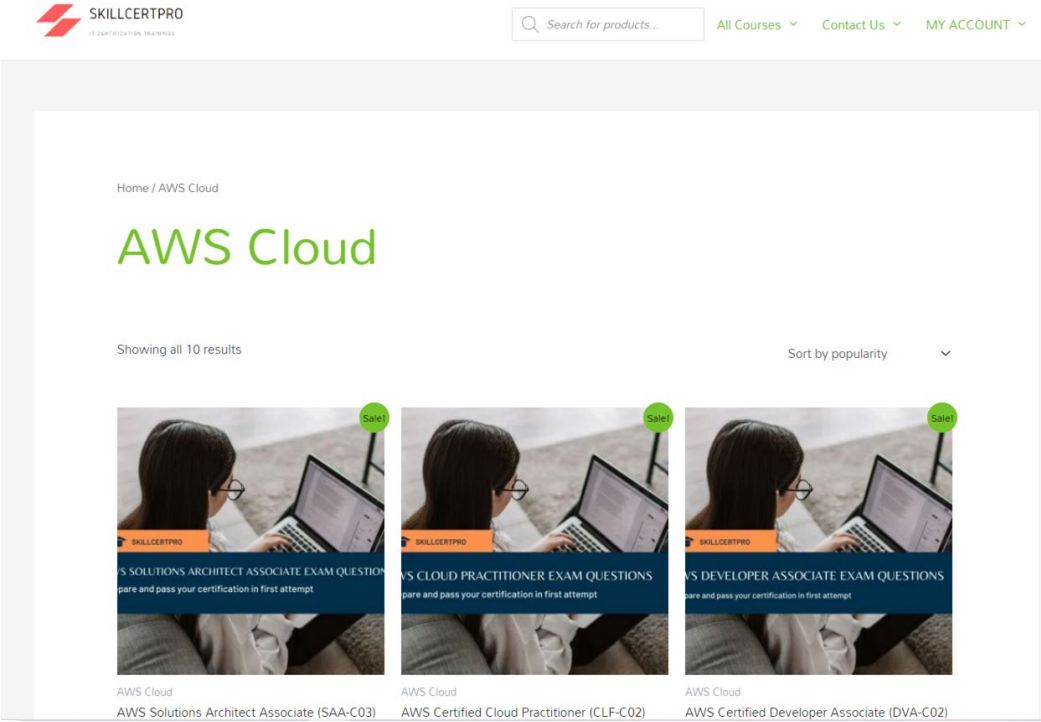


nuclio learning

Páginas para practicar exámenes gratis



Páginas para practicar exámenes gratis



nuclio learning

Caminos de certificaciones en AWS

nuclio learning

Caminos de certificaciones en AWS

IA/ML

Ingeniero de machine learning

Investigue, cree y diseñe sistemas de inteligencia artificial (IA) para automatizar los modelos predictivos y diseñe sistemas, modelos y esquemas de machine learning.

Opcional para profesionales de TI y la nube

Caminos de certificaciones en AWS

Desarrollo

Ingeniero de desarrollo de software

Desarrolle, cree y administre software en todas las plataformas y dispositivos.

Opcional para profesionales de TI y la nube

Operaciones

Administrador de sistemas

Instale, actualice y administre los componentes de cómputo y el software e integre los procesos de automatización.

Opcional para profesionales de TI y la nube

Para profundizar

Ingeniero de la nube

Implemente y opere la infraestructura de cómputo en red de una organización e implemente sistemas de seguridad para proteger los datos.

Opcional para profesionales de TI y la nube

Para profundizar

nuclio learning

Caminos de certificaciones en AWS

nuclio learning

Caminos de certificaciones en AWS

Seguridad

Ingeniero de seguridad en la nube

Diseñe una arquitectura de seguridad informática y desarrolle diseños detallados de ciberseguridad. Desarrolle, ejecute y realice un seguimiento del rendimiento de las medidas de seguridad para proteger la información.

Opcional para profesionales de TI y la nube

Arquitecto de seguridad en la nube

Diseñe e implemente soluciones empresariales en la nube que apliquen la gobernanza para identificar, comunicar y minimizar los riesgos empresariales y técnicos.

Opcional para profesionales de TI y la nube

Redes

Ingeniero de redes

Diseñe e implemente redes informáticas y de información, como redes de área local (LAN), redes de área amplia (WAN), intranets, extranets, etc.

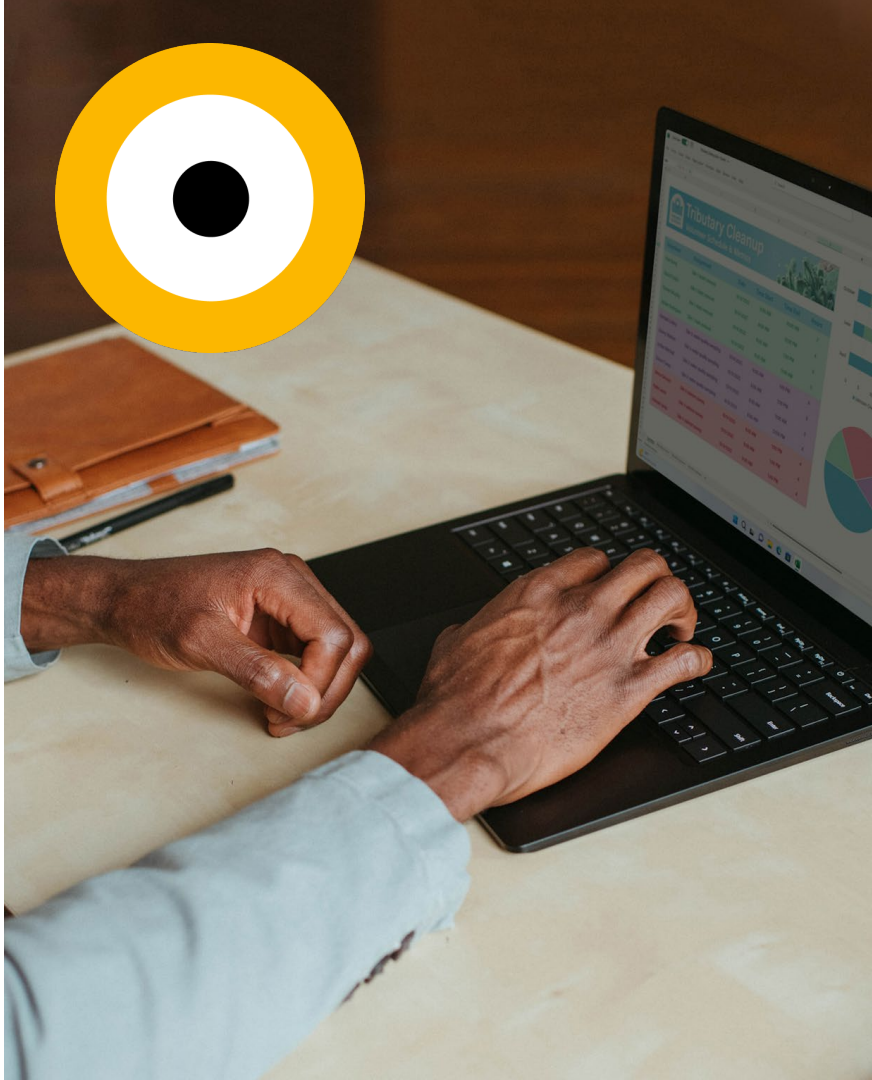
Opcional para profesionales de TI y la nube

nuclio learning

¡Gracias!

tumail@nucliolearning.com
+34 000 000 000
nucliolearning.com

nuclio learning



nuclio learning