

DATA STREAMING Y SERVICIOS EN LA NUBE

AWS



**UNIVERSIDAD
DE ANTIOQUIA**

1 8 0 3

Magister - Efraín Alberto Oviedo
alberto.oviedo@udea.edu.co

**UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS**

AGENDA

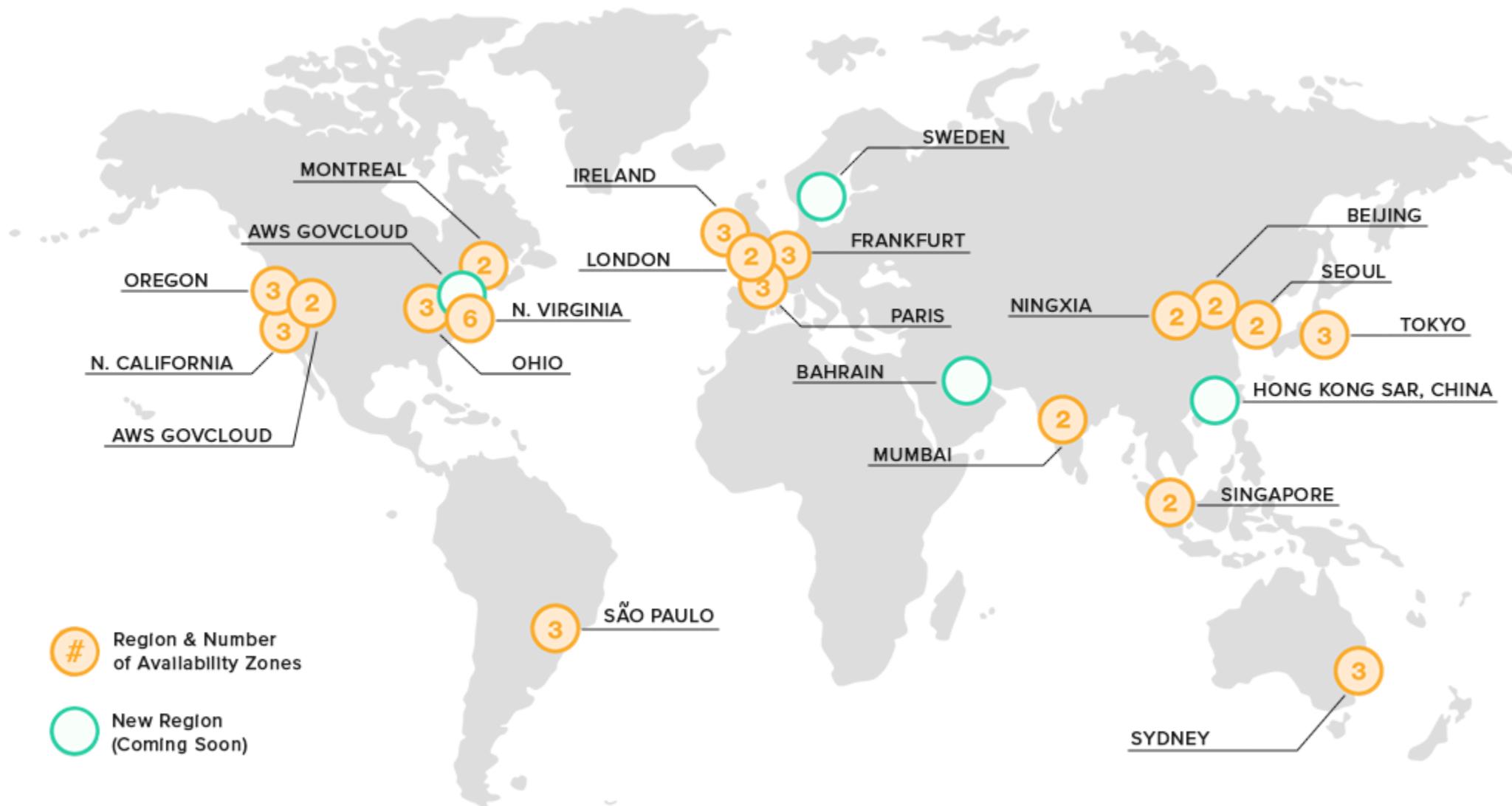
- 1. AWS**
2. Nube Privada Virtual (VPC)
3. Procesamiento (EC2)
4. Almacenamiento (S3)
5. BigData (EMR)
6. Monitoreo (CloudWatch)
7. Ejercicios

AWS (Amazon Web Services)

- Ofrece un amplio conjunto de productos globales basados en la nube
- Desde 2006 empieza a ofrecer servicios de Infraestructura de TI en forma de servicios Web
- Dispone de mas de 200 servicios
- Tiene mas de un millón de clientes en mas de 90 países
- Opera en 42 zonas de disponibilidad en 16 regiones geográficas



Regiones



Enterprise Customers



LIONSGATE®



NOKIA

AUTODESK

Schneider
Electric

COMCAST



UBISOFT

Unilever

The
New York
Times



Adobe

TOSHIBA

NASDAQ OMX

NETFLIX

MEDIACORP

Crear Cuenta

1. Ingrese a: <https://aws.amazon.com/es/>

2. Ingrese en la opción “Inicie sesión en la consola”



Consola de administración de AWS

Todo lo que necesita para acceder a la nube de AWS y administrarla, en una interfaz web

Iniciar sesión nuevamente

3. Ingrese en la opción “Crear una cuenta en AWS”
4. Llene el formulario y haga clic en Verificar la dirección de correo electrónico

Crear Cuenta

Iniciar sesión

Usuario raíz

Propietario de la cuenta que realiza tareas que requieren acceso ilimitado. [Más información](#)

Usuario de IAM

Usuario de una cuenta que realiza tareas diarias. [Más información](#)

Dirección de email del usuario raíz

nombredeusuario@ejemplo.com

Siguiente

Al continuar, acepta el [Contrato de cliente de AWS](#) u otro acuerdo para los servicios de AWS y el [Aviso de privacidad](#). Este sitio utiliza cookies esenciales. Consulte nuestro [Aviso de cookies](#) para obtener más información.

¿Es nuevo en AWS?

3

[Crear una cuenta de AWS](#)

Registrarse en AWS

Dirección de correo electrónico del usuario raíz

Se utiliza para la recuperación de cuentas y algunas funciones administrativas

Nombre de la cuenta de AWS

Elija un nombre para la cuenta. Podrá cambiarlo en la configuración de la cuenta después de registrarse.

4

Verificar la dirección de correo electrónico

0

[Iniciar sesión en una cuenta de AWS existente](#)

Crear Cuenta

5. Revise en su correo, identifique e ingrese el código de verificación

6. Haga clic en Verificar

7. Asigne una contraseña

8. Haga clic en Continuar

Registrarse en AWS

Confirme que es usted

Garantizar que esté seguro, es lo que hacemos.

Hemos enviado un correo electrónico con un código de verificación a efrain.oviedo@upb.edu.co. ([¿No es usted?](#))

Introdúzcalo a continuación para confirmar su correo electrónico.

Código de verificación

5

6

Verificar

[Volver a enviar el código](#)

¿No ha recibido el código?

- Los códigos pueden tardar hasta 5 minutos en llegar.
- Revise su carpeta de spam.

Registrarse en AWS

Cree la contraseña

Hemos verificado su identidad. X
La dirección de correo electrónico se ha verificado correctamente.

La contraseña proporciona acceso de inicio de sesión a AWS, por lo que es importante que este proceso se realice de forma correcta.

Contraseña de usuario raíz

.....

Confirmar la contraseña del usuario raíz

.....

8

Continuar (paso 1 de 5)

Crear Cuenta

9. Seleccione tipo de cuenta Personal e ingrese su información de contacto. Acepte los términos de contrato y haga clic en continuar

10. Ingrese la información de facturación y haga clic en verificar y continuar

Registrarse en AWS

Información de contacto

¿Cómo tiene previsto utilizar AWS?

- Empresarial: para su trabajo, escuela u organización
 Personal: para sus propios proyectos

¿A quién debemos contactar para consultar sobre esta cuenta?

Nombre completo

Número de teléfono

+1 ▾ 222-333-4444

País o región

Estados Unidos ▾

Dirección

Apartamento, suite, unidad, edificio, planta, etc.

Ciudad

Registrarse en AWS

Información de facturación

10

Número de tarjeta de crédito o débito



AWS acepta todas las tarjetas de crédito y débito principales. Para obtener más información sobre las opciones de pago, consulte nuestras [preguntas frecuentes](#)

Fecha de vencimiento

Mes ▾ Año ▾

Nombre del titular de la tarjeta

Crear Cuenta

11. Confirme su identidad
(preferiblemente por
mensaje de texto)

12. Ingrese el código de
verificación y haga clic en
continuar

13. Seleccione plan de
soporte: Soporte de nivel
Básico y haga clic en
finalizar registro

Registrarse en AWS

Confirme su identidad

Para poder utilizar la cuenta de AWS, debe verificar su número de teléfono. Cuando continúe, el sistema automatizado de AWS se comunicará con usted para proporcionarle un código de verificación.

¿Cómo prefiere que le envíemos el código de verificación?

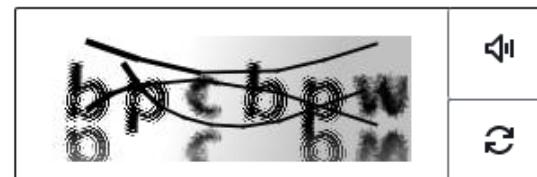
- Mensaje de texto (SMS)
- Llamada de voz

Código de país o región

Estados Unidos (+1) ▾

Número de teléfono móvil

Comprobación de seguridad



Registrarse en AWS

Confirme su identidad

Verificar código

12

Continuar (paso 4 de 5)

¿Tiene algún problema? A veces, se necesitan hasta 10 minutos para recibir el código de verificación. Si ha transcurrido más tiempo del mencionado, [vuelva a la página anterior](#) e inténtelo de nuevo.

Iniciar Sesión

Ingrese su usuario y contraseña para acceder a la consola de AWS

Servicios Búsqueda [Alt+S] Oregón ▾ Docencia ▾

Página de inicio de la Consola Información Restablecer al diseño predeterminado + Agregar widgets

Presentamos el nuevo widget Aplicaciones. Encuéntrelo en la parte inferior de la página de inicio de la consola.

Visitados recientemente Información

- AWS Cost Explorer
- Support
- IAM
- AWS Budgets
- EC2
- EMR

Le damos la bienvenida a AWS

Introducción a AWS  Conozca los aspectos fundamentales y encuentre información valiosa para sacar el máximo provecho de AWS.

Formación y certificación  Aprenda de expertos de AWS, mejore sus habilidades y aumente sus conocimientos.

¿Cuáles son las novedades de AWS?  Descubra los nuevos servicios, características y regiones de AWS.

Iniciar Sesión – AWS Academy

1. Ingrese en el menú de módulos
2. Haga clic en iniciar el Laboratorio de aprendizaje de AWS Academy

The screenshot shows the AWS Academy dashboard. On the left, there is a sidebar with icons for Cuenta, Tablero, Cursos, Calendario, and Bandeja de entrada. The 'Cursos' icon is highlighted with a red circle and the number '1'. The main content area has a 'Página de Inicio' header. Below it, there are two sections: 'Laboratorio de aprendizaje de AWS Academy. Cumplimiento y Seguridad' and 'Laboratorio de aprendizaje de AWS Academy'. The 'Iniciar el Laboratorio de aprendizaje de AWS Academy' button in the second section is highlighted with a red rectangle and the number '2'.

Página de Inicio

Módulos

Foros de discusión

Cuenta

Tablero

Cursos

Calendario

Bandeja de entrada

Laboratorio de aprendizaje de AWS Academy. Cumplimiento y Seguridad

Cómo usar de manera eficaz el Laboratorio de aprendizaje de Academy

Evaluación de conocimientos del módulo
100 pts | Obtener al menos 70.0

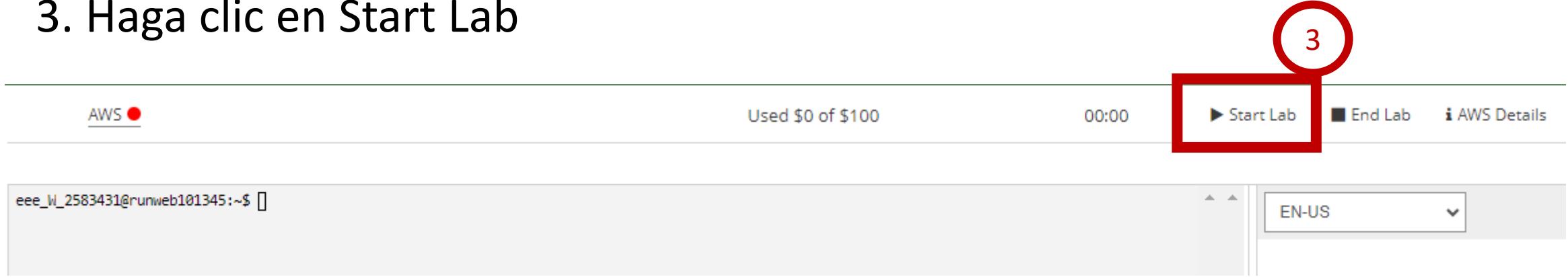
Laboratorio de aprendizaje de AWS Academy

Iniciar el Laboratorio de aprendizaje de AWS Academy

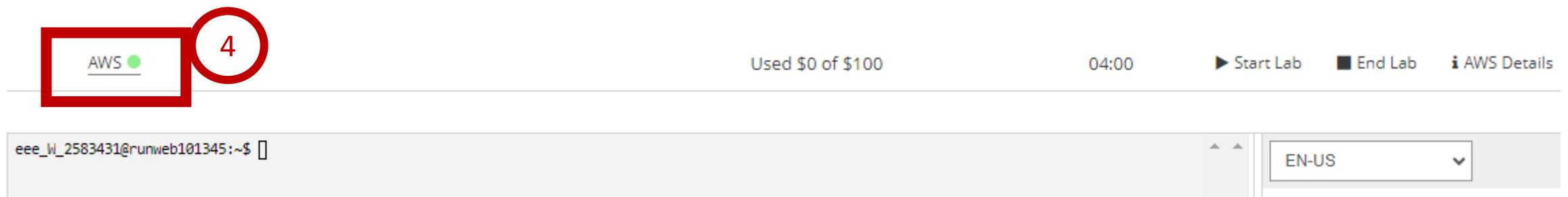
Recursos de los laboratorios de aprendizaje de AWS Academy

Iniciar Sesión – AWS Academy

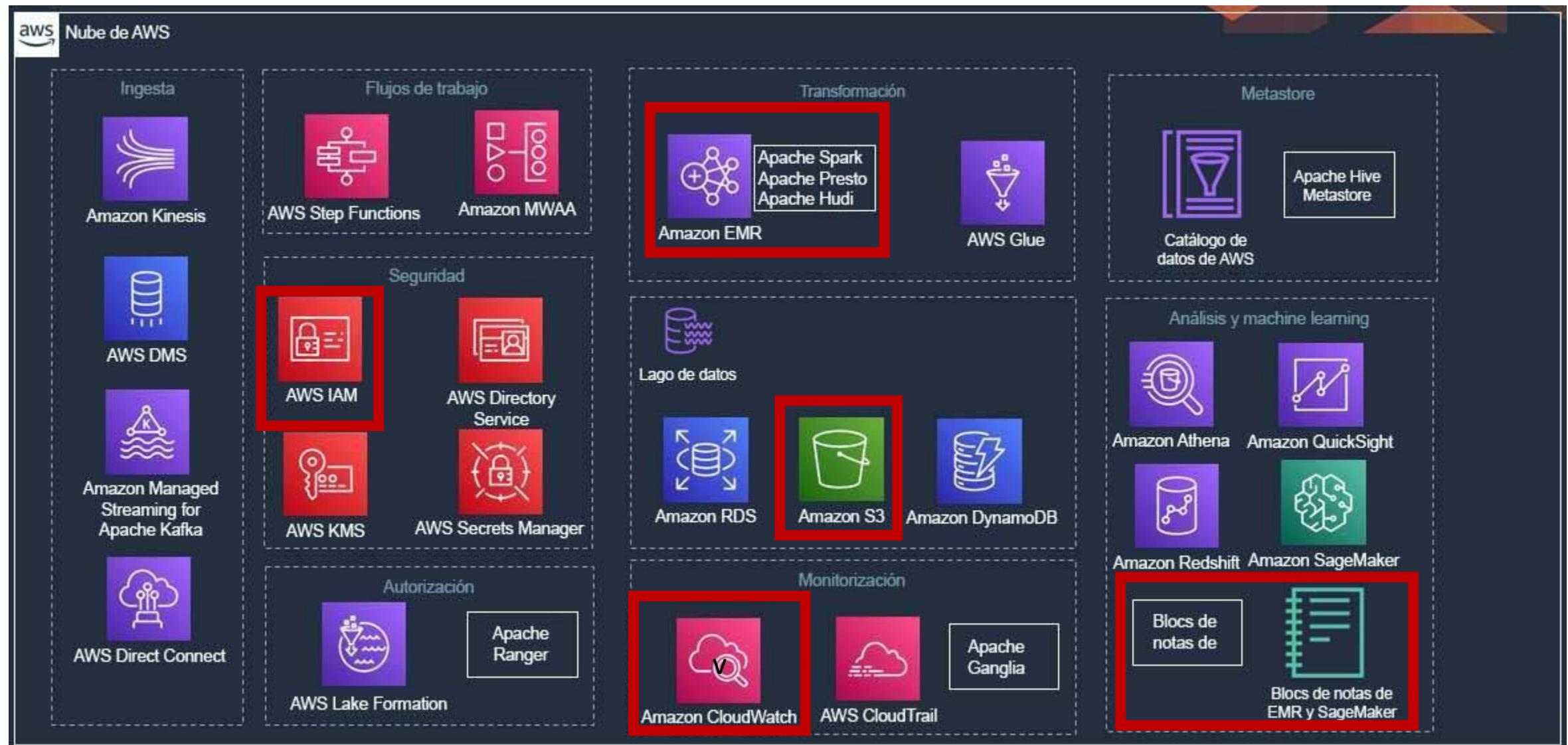
3. Haga clic en Start Lab



4. Haga clic en AWS para iniciar sesión



Servicios AWS



Capa Gratuita



Pruebas gratuitas

Las ofertas de prueba gratuita a corto plazo se inician a partir de la fecha en la que se activa un servicio en particular

MACHINE LEARNING	NOVEDAD
Nivel gratuito	PRUEBA GRATUITA
Amazon SageMaker	
2 meses	

prueba gratuita

Machine learning para todos los científicos de datos y desarrolladores.

250 horas al mes de ml.t3.medium en los



Gratis para siempre

Estas ofertas del nivel gratuito no caducan y están disponibles para todos los clientes de AWS



12 meses de uso gratuito

Disfrute de estas ofertas durante 12 meses después de su fecha de registro inicial en AWS

HERRAMIENTAS PARA DESARROLLADORES	NOVEDAD
Nivel gratuito	GRATUITO PARA SIEMPRE
Amazon CloudWatch	

10

alarmas y métricas personalizadas

Monitoreo de recursos y aplicaciones en la nube de AWS.

COMPUTACIÓN	12 MESES GRATIS
Amazon EC2	
750 horas	al mes

Capacidad de cómputo de tamaño variable en la nube.

ALMACENAMIENTO	12 MESES GRATIS
Amazon S3	

5 GB

de almacenamiento estándar

Infraestructura de almacenamiento de objetos segura, duradera y escalable.

https://aws.amazon.com/es/free/?all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc&awsf.Free%20Tier%20Types=*all&awsf.Free%20Tier%20Categories=*all

Calculadora de Precios



Comentarios

Español

Comuníquese con el departamento de ventas

Calculadora de precios de AWS

Realice una estimación del costo de su solución de arquitectura.

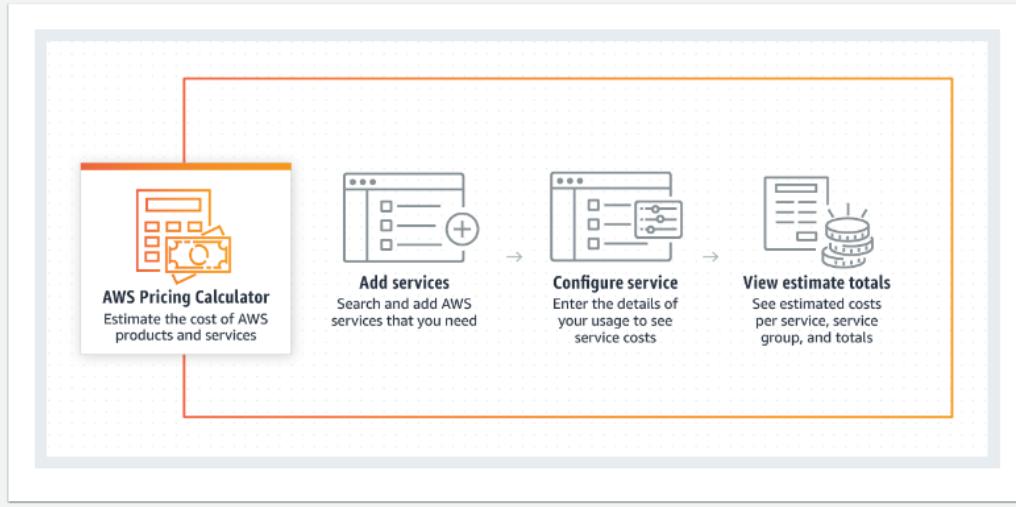
Configure una estimación de costos que se adapte a sus necesidades únicas personales o de su negocio con los productos y servicios de AWS.

Cree una estimación

Comience su estimación sin compromiso y explore los servicios y precios de AWS para sus necesidades de arquitectura.

[Crear una estimación](#)

Cómo funciona



Más recursos

[Guía del usuario](#)

[Preguntas frecuentes](#)

[Suposiciones y variaciones de precios](#)

[¿Necesita ayuda con las estimaciones? Conecte con un experto certificado de AWS en AWS IQ](#)

AWS Modernization Calculator for Microsoft Workloads

Estimate the cost of transforming Microsoft workloads to a modern architecture that uses open source and cloud-native services deployed on AWS.

[Get started](#)

IAM (Identity and Access Management)

- Administrador de usuarios y acceso a los recursos y servicios de AWS
- Permite configurar las funcionalidades permitidas para cada tipo de usuario
- Puede asignar credenciales de acceso temporales
- Entrega información para analizar el acceso de los usuarios
- Algunas funcionalidades (como los NoteBooks) no están permitidas para el usuario principal (usuario raíz) de la cuenta

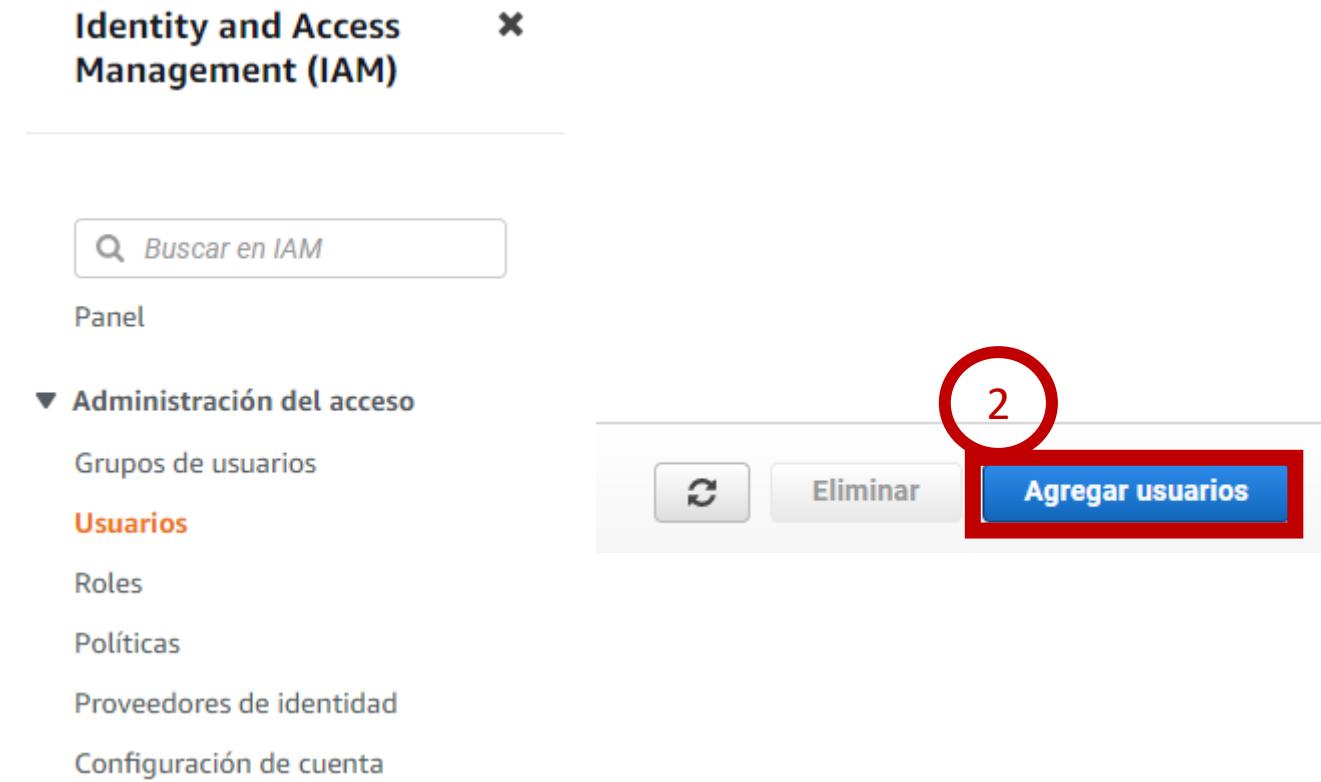
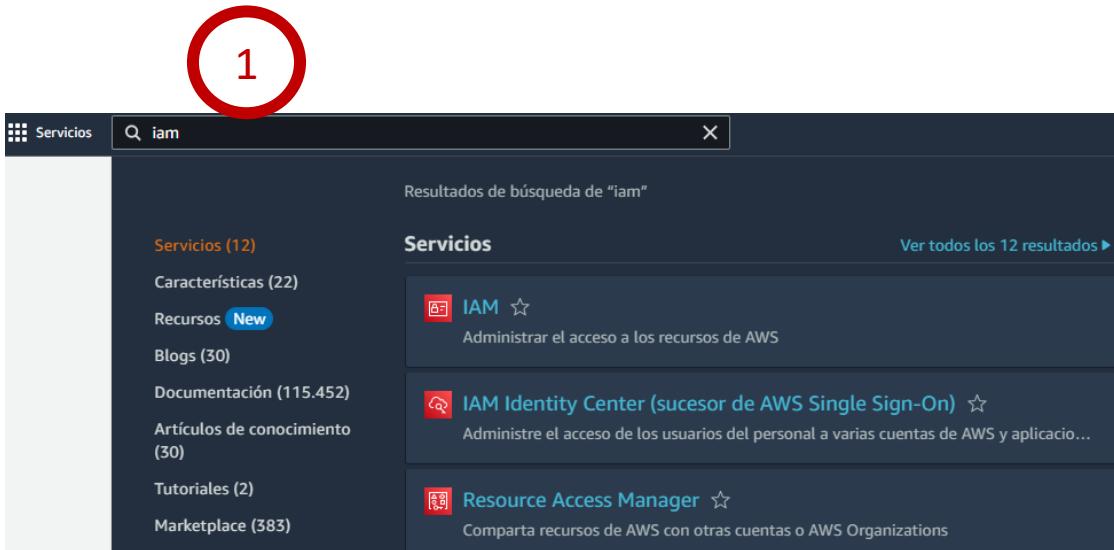


IAM (Identity and Access Management)



IAM (Identity and Access Management)

1. En la barra de búsqueda digite IAM
2. En la opción de Usuarios haga clic en Agregar usuarios



IAM (Identity and Access Management)

3. Asigne un nombre al usuario
4. Seleccione tipo de credenciales: por contraseña
5. Agregue una contraseña personalizada
6. Haga clic en crear un grupo
7. Asígnale un nombre al grupo
8. Configure los permisos del grupo de usuario
[AmazonElasticMapReduceFullAccess](#)
9. Haga clic en crear usuario

IAM (Identity and Access Management)

Establecer los detalles del usuario

Puede añadir varios usuarios a la vez con los mismos permisos y el mismo tipo de acceso. [Más información](#)

Nombre de usuario*

operador

+ Añadir otro usuario

3

Seleccionar el tipo de acceso de AWS

Seleccione cómo estos usuarios accederán principalmente a AWS. Si elige únicamente el acceso mediante programación, NO evitará que los usuarios accedan a la consola por medio de un rol asumido. Las claves de acceso y las contraseñas generadas automáticamente se proporcionan en el último paso. [Más información](#)

Seleccione el tipo de credenciales
de AWS*

Clave de acceso: acceso mediante programación

Habilita una ID de clave de acceso y una clave de acceso secreta para el SDK, la CLI y la API de AWS, además de otras herramientas de desarrollo.

Contraseña: acceso a la consola de administración de AWS

Habilita una contraseña que permite a los usuarios iniciar sesión en la consola de administración de AWS.

4

Contraseña de la consola*

Contraseña generada automáticamente

Contraseña personalizada

5

Mostrar contraseña

Requerir el restablecimiento de
contraseña

El usuario debe crear una contraseña nueva en el próximo inicio de sesión

Los usuarios obtienen automáticamente la política [IAMUserChangePassword](#) que les permite cambiar su propia contraseña.

Establecer permisos



Añadir un usuario al grupo



Copiar permisos de un usuario existente



Asociar directamente las políticas existentes

Añada un usuario a un grupo existente o cree uno. El uso de grupos es una práctica recomendada para administrar los permisos de un usuario por funciones de trabajo. [Más información](#)

Añadir un usuario al grupo

Crear un grupo

Actualizar

Crear un grupo

Cree un grupo y seleccione las políticas que desea asociar a este. El uso de grupos es una práctica recomendada para administrar los permisos. [Más información](#)

Nombre de grupo

emr

7

Crear una política

Actualizar

Filtrar políticas ▾

elasticmapreduce

	Nombre de la política ▾	Tipo	Utilizado como
<input type="checkbox"/>	AmazonElasticMapReduceRole	Administrado por AWS	Permissions policy (1)
<input type="checkbox"/>	AmazonElasticMapReduceReadOnlyAccess	Administrado por AWS	Ninguna
<input type="checkbox"/>	AmazonElasticMapReducePlacementGroupPolicy	Administrado por AWS	Ninguna
<input checked="" type="checkbox"/>	AmazonElasticMapReduceFullAccess	Administrado por AWS	Ninguna
<input type="checkbox"/>	AmazonElasticMapReduceEC2Role	Administrado por AWS	Permissions policy (1)
<input type="checkbox"/>	AmazonElasticMapReduceforAutoScalingRole	Administrado por AWS	Permissions policy (1)
<input type="checkbox"/>	AmazonElasticMapReduceEditorsRole	Administrado por AWS	Permissions policy (1)

8

IAM (Identity and Access Management)

Añadir usuario(s)

1 2 3 4 5

Correcto

Ha creado correctamente los usuarios que se muestran a continuación. Puede ver y descargar las credenciales de seguridad de los usuarios. También puede enviar a los usuarios un correo electrónico con instrucciones para iniciar sesión en la consola de administración de AWS. Esta es la última vez que las credenciales estarán disponibles para descargarlas. Sin embargo, puede [operador](#) cualquier momento.

Los usuarios con acceso a la consola de administración de AWS pueden iniciar sesión en:

<https://067205227321.signin.aws.amazon.com/console>

 Descargar .csv

	Usuario	
	operator	Enviar instrucciones de inicio de sesión por correo electrónico Enviar correo electrónico



Iniciar sesión

Usuario raíz

Propietario de la cuenta que realiza tareas que requieren acceso ilimitado. [Más información](#)

Usuario de IAM

Usuario de una cuenta que realiza tareas diarias. [Más información](#)

ID de cuenta (12 dígitos) o alias de cuenta

Siguiente

Al continuar, acepta el [Contrato de cliente de AWS](#) u otro acuerdo para los servicios de AWS y el [Aviso de privacidad](#). Este sitio utiliza cookies esenciales. Consulte nuestro [Aviso de cookies](#) para obtener más información.

AGENDA

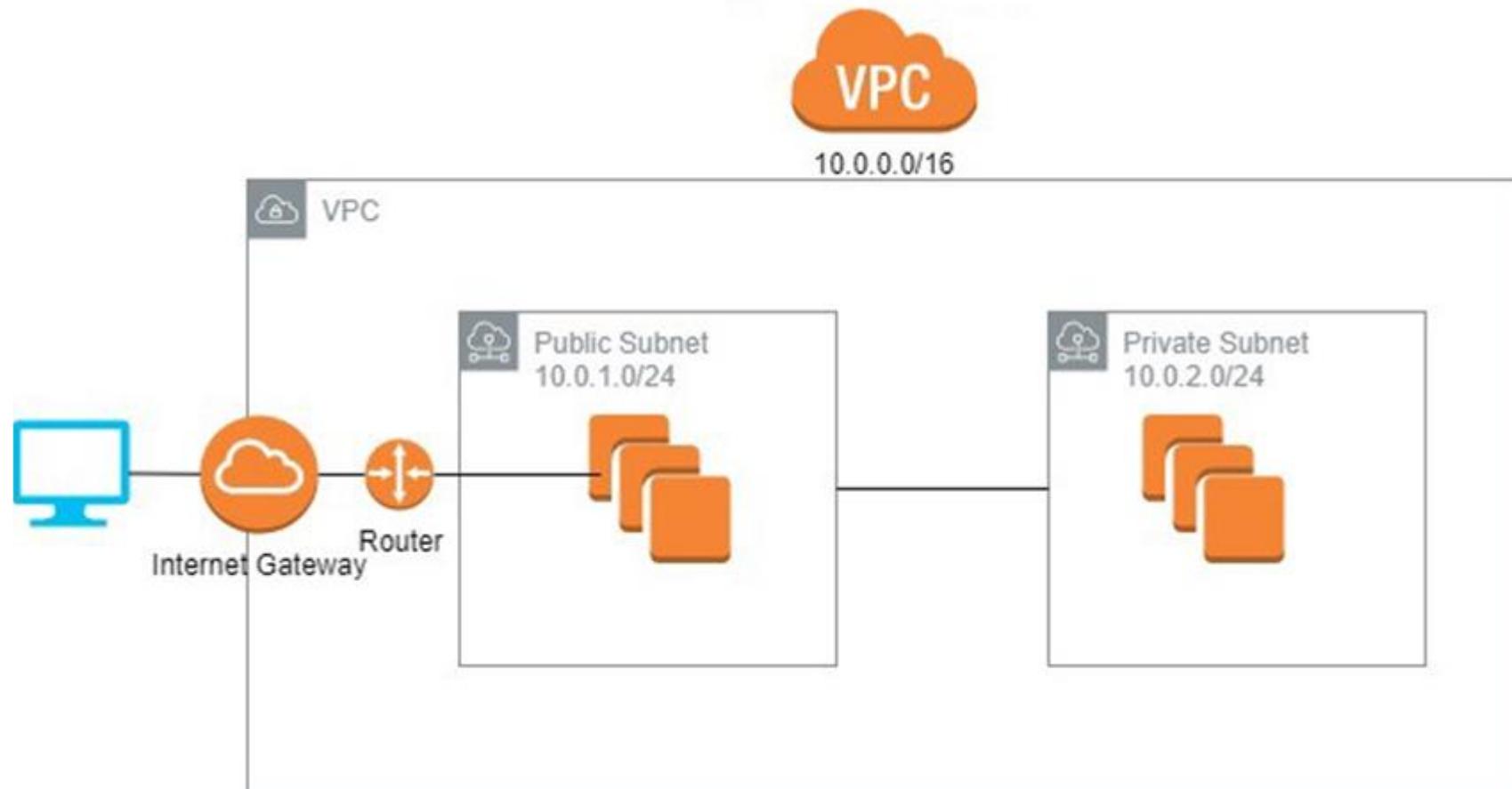
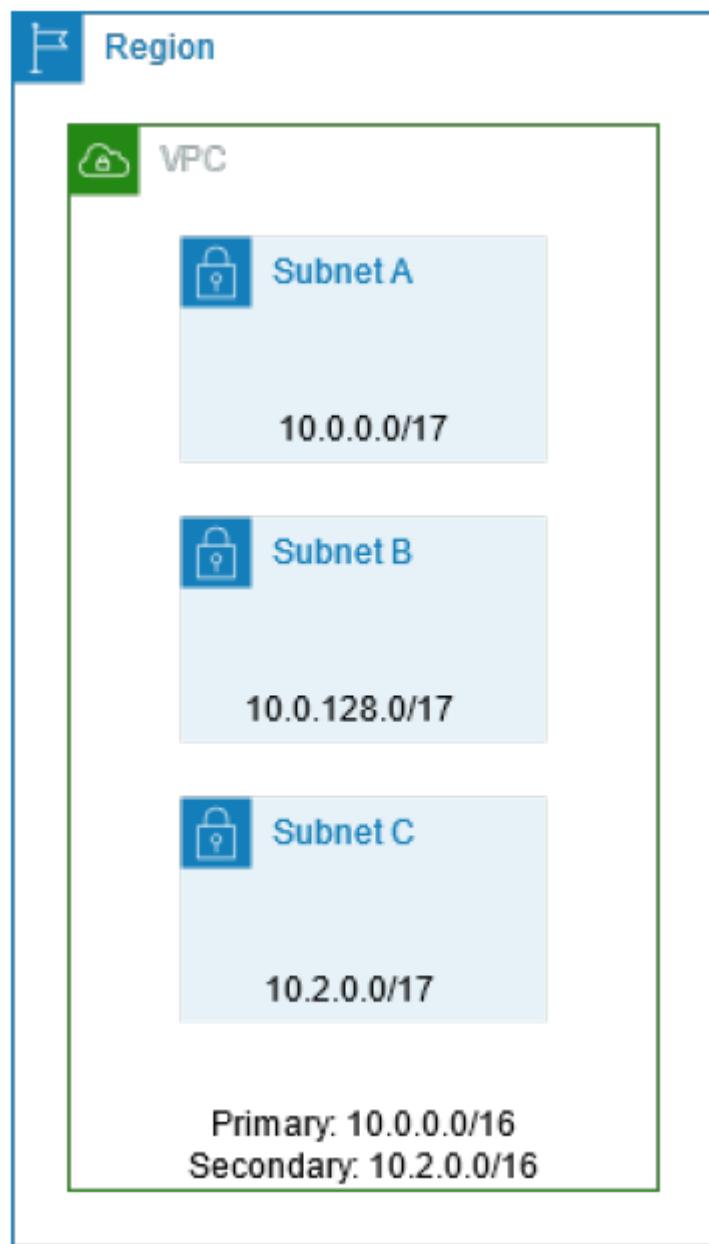
1. AWS
- 2. Nube Privada Virtual (VPC)**
3. Procesamiento (EC2)
4. Almacenamiento (S3)
5. BigData (EMR)
6. Monitoreo (CloudWatch)
7. Ejercicios

VPC (Virtual Private Cloud)

- Permite aprovisionar una sección aislada de forma lógica de la nube
- Dentro de la red virtual puede lanzar recursos AWS controlando direccionamiento IP y accesibilidad
- Dentro de la VPC puede crear sub redes
 - Pública: Servidores web con acceso a Internet
 - Privada: Bases de datos, aplicaciones internas

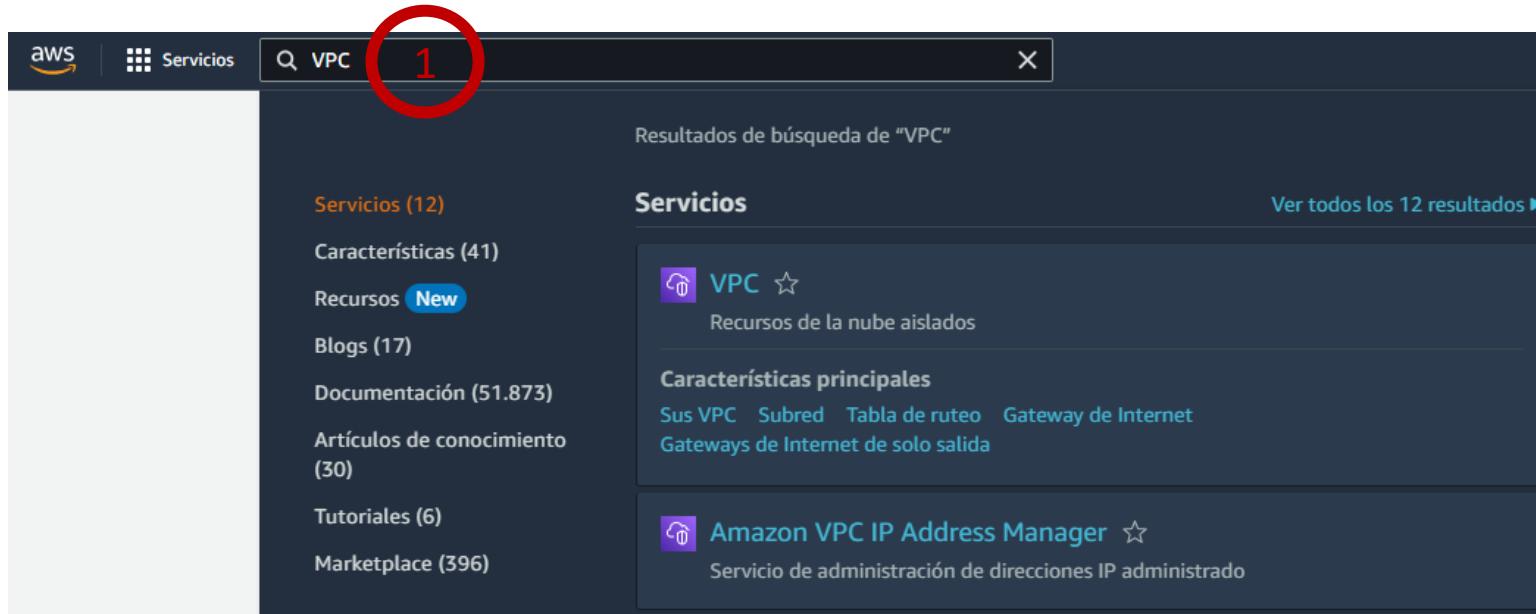


VPC



Crear VPC

1. En la barra de búsqueda digite VPC
2. Haga clic en Crear VPC
3. En la sección de recursos que se van a crear seleccione VPC y mas
4. Asigne 1 sub red pública y 1 sub red privada



Crear VPC

Crear VPC Información

Una VPC es una parte aislada de la nube de AWS que contiene objetos de AWS, como instancias de Amazon EC2. Deslizar el ratón sobre un recurso para resaltar los recursos relacionados.

Configuración de la VPC

Recursos que se van a crear [Información](#)
Cree únicamente el recurso de VPC o la VPC y otros recursos de red.

Solo la VPC VPC y más 3

Generación automática de etiquetas de nombre [Información](#)
Ingrese un valor para la etiqueta Nombre. Este valor se utilizará para generar automáticamente etiquetas Nombre para todos los recursos de la VPC.

Generar automáticamente

Bloque de CIDR IPv4 [Información](#)
Determine la IP inicial y el tamaño de la VPC mediante la notación CIDR.

10.0.0.0/16 65.536 IPs

Bloque de CIDR IPv6 [Información](#)
 Sin bloque de CIDR IPv6
 Bloque de CIDR IPv6 proporcionado por Amazon

Tenencia [Información](#)
Predeterminado

Número de zonas de disponibilidad (AZ) [Información](#)
Elija la cantidad de zonas de disponibilidad en las que desea aprovisionar subredes. Le recomendamos que tenga al menos dos para incrementar la disponibilidad.

1 2 3

Vista previa

Presentación de la nueva experiencia de creación de VPC
Hemos diseñado la nueva experiencia de creación de VPC para facilitar su uso. Ahora puede visualizar los recursos que se crearán.

- Novedad: edite la etiqueta de nombre de los recursos individuales. Desmarque "Auto-generate" (Generar automáticamente) y defina cada etiqueta de nombre.

Díganos qué piensa al respecto.

```
graph LR; VPC[test_vpc] --- Subred1[us-west-2a]; VPC --- Subred2[us-west-2b]; Subred1 --- TablaPublica1[Tabla de enrutamiento pública sin etiquetas]; Subred1 --- TablaPrivada1[Tabla de enrutamiento privada sin etiquetas]; Subred2 --- TablaPublica2[Tabla de enrutamiento pública sin etiquetas]; Subred2 --- TablaPrivada2[Tabla de enrutamiento privada sin etiquetas]
```

Visualizar las VPC

Sus VPC (1/2) [Información](#)

Filtrar las VPC

C [Acciones](#) [Crear VPC](#)

	Name	ID de la VPC	Estado	CIDR IPv4	CIDR IPv6	Conjunto de opciones	Tabla de enrutamiento
<input type="checkbox"/>	-	vpc-04b931b7545d1c0fc	Available	172.31.0.0/16	-	dopt-0113425a95b88...	rtb-06d9cb6b56da2d8be
<input checked="" type="checkbox"/>	test_vpc	vpc-0d26d791b86dd407e	Available	10.0.0.0/16	-	dopt-0113425a95b88...	rtb-0a5c04f78a58446a8

vpc-0d26d791b86dd407e / test_vpc [Acciones](#)

[Detalles](#) [Información](#)

ID de la VPC <input type="text" value="vpc-0d26d791b86dd407e"/>	Estado Available	Nombres de host de DNS Habilitado	Resolución de DNS Habilitado
Tenencia Default	Conjunto de opciones de DHCP dopt-0113425a95b88f273	Tabla de enrutamiento principal rtb-0a5c04f78a58446a8	ACL de red principal acl-02c38c845dfb98c9c
VPC predeterminada No	CIDR IPv4 10.0.0.0/16	Grupo IPv6 -	CIDR IPv6 (grupo de bordes de red) -
Métricas de uso de direcciones de red Desactivado	Grupos de reglas del firewall de DNS de Route 53 Resolver -	ID de propietario <input type="text" value="067205227321"/>	

[CIDR](#) [Registros de flujo](#) [Etiquetas](#)

[CIDR](#) [Información](#)

Tipo de dirección	CIDR	Grupo de borde de red	Grupo	Estado
IPv4	10.0.0.0/16	-	-	Associated

AGENDA

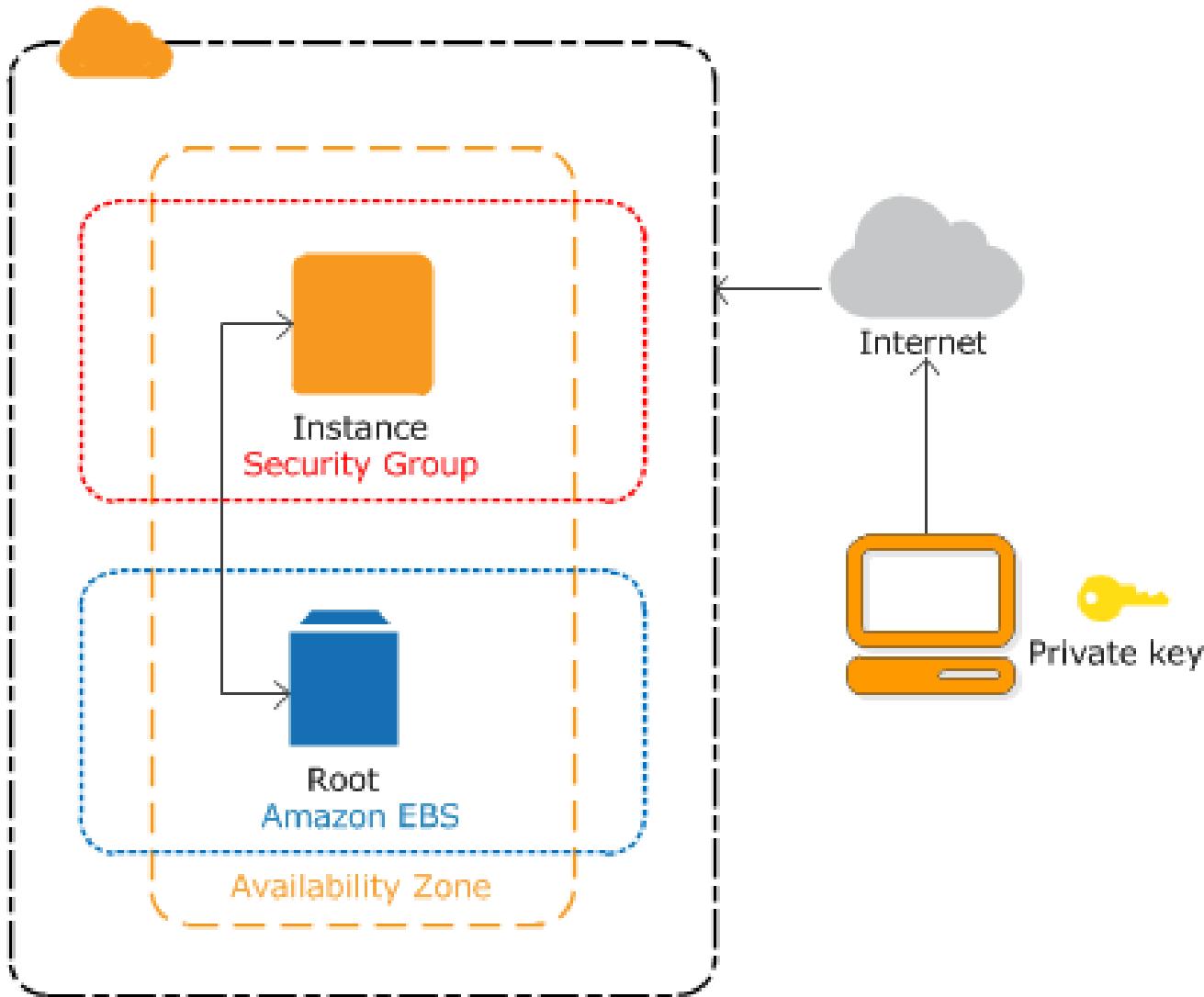
1. AWS
2. Nube Privada Virtual (VPC)
- 3. Procesamiento (EC2)**
4. Almacenamiento (S3)
5. BigData (EMR)
6. Monitoreo (CloudWatch)
7. Ejercicios

EC2 (Elastic Compute Cloud)

- Servicio web que proporciona capacidad de computación en la nube segura y de tamaño variable
- En cuestión de unos cuantos minutos puede crear e iniciar nuevas instancias
- Permite escalar la capacidad ya sea para reducir o para aumentar según su necesidad. Así puede pagar por la capacidad que realmente necesita
- Cuenta con instancias bajo demanda, reservadas y spot



EC2 (Elastic Compute Cloud)



Tipos de Instancias EC2

General Purpose	Compute Optimised	Memory Optimised	Accelerated Computing	Storage Optimised
 ARM based core and custom silicon	 Compute - CPU intensive apps and DBs	 RAM - Memory intensive apps and DB's	 Processing optimised - Machine Learning	 High Disk Throughput - Big data clusters
 Tiny - Web servers and small DBs		 Xtreme RAM - For SAP/Spark	 Graphics Intensive - Video and streaming	 IOPS - NoSQL DBs
 Main - App servers and general purpose		 High Compute and High Memory - Gaming	 Field Programmable - Hardware acceleration	 Dense Storage - Data Warehousing

<https://aws.amazon.com/es/ec2/instance-types/>

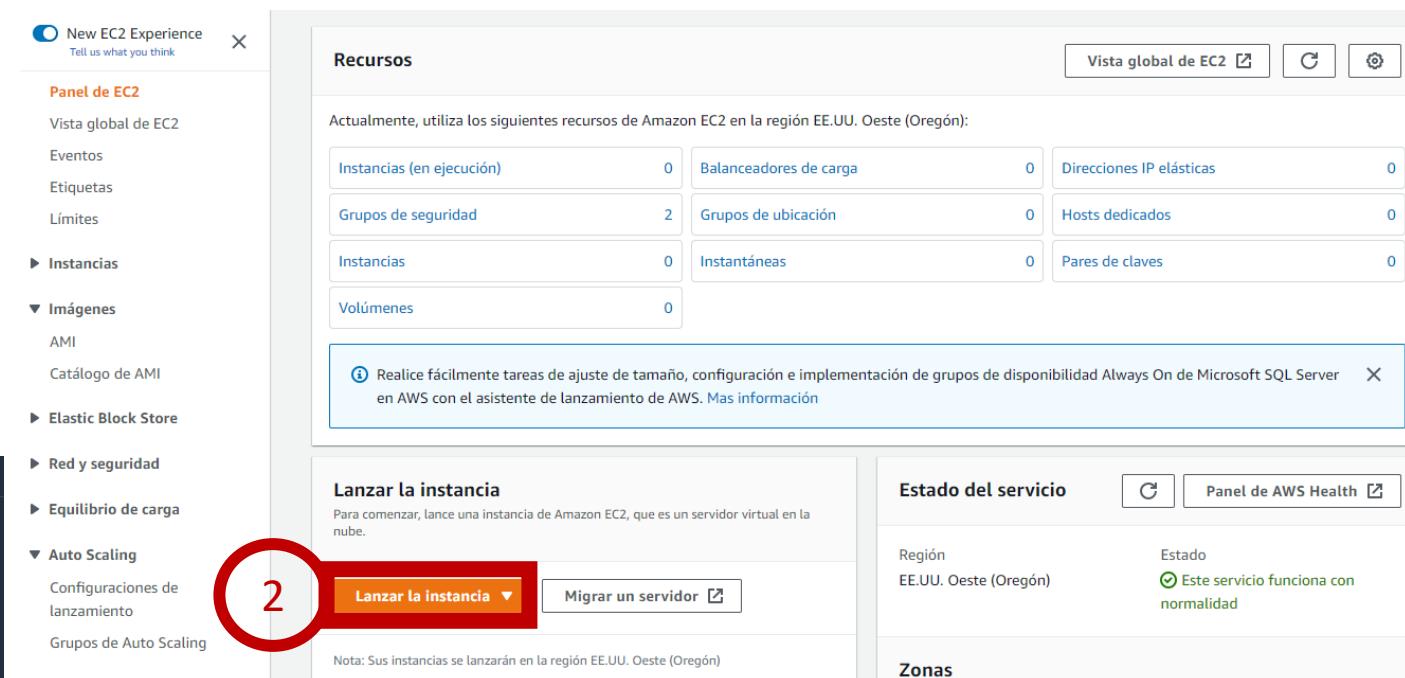
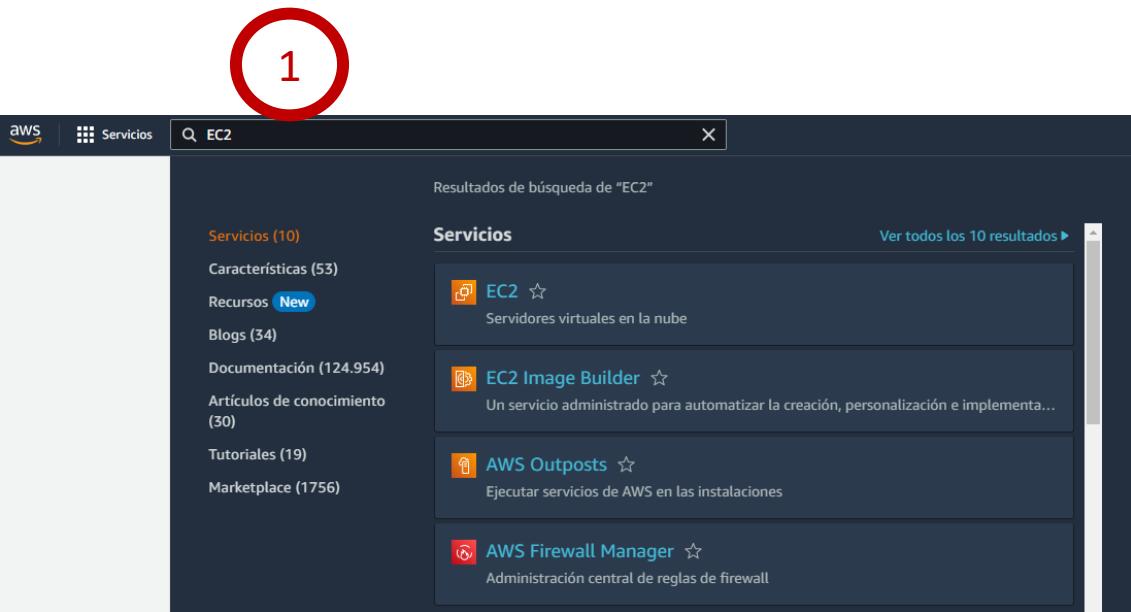
Precios bajo demanda EC2

Nombre de la instancia ▲	Tarifa por hora bajo demanda ▼	vCPU ▼	Memoria ▼	Almacenamiento ▼
a1.medium	0,0255 USD	1	2 GiB	Solo EBS
a1.large	0,051 USD	2	4 GiB	Solo EBS
a1.xlarge	0,102 USD	4	8 GiB	Solo EBS
a1.2xlarge	0,204 USD	8	16 GiB	Solo EBS
u-12tb1.112xlarge	109,20 USD	448	12288 GiB	Solo EBS
u-6tb1.112xlarge	54,60 USD	448	6144 GiB	Solo EBS
u-6tb1.56xlarge	46,40391 USD	224	6144 GiB	Solo EBS
p4d.24xlarge	32,7726 USD	96	1152 GiB	8 x 1000 SSD
u-3tb1.56xlarge	27,30 USD	224	3072 GiB	Solo EBS
x1e.32xlarge	26,688 USD	128	3904 GiB	2 x 1920 SSD

<https://aws.amazon.com/es/ec2/pricing/on-demand/>

Crear EC2

1. En la barra de búsqueda digite EC2
2. Haga clic en Lanzar instancia
3. Asigne un nombre
4. Seleccione una imagen
5. Cree un nuevo par de claves



Crear EC2

Lanzar una instancia Información

Amazon EC2 le permite crear máquinas virtuales, o instancias, que se ejecutan en la nube de AWS. Comience rápidamente siguiendo los sencillos pasos que se indican a continuación.

Nombre y etiquetas Información

Nombre 3

Agregar etiquetas adicionales

▼ Imágenes de aplicaciones y sistemas operativos (Amazon Machine Image) Información

Una AMI es una plantilla que contiene la configuración de software (sistema operativo, servidor de aplicaciones y aplicaciones) necesaria para lanzar la instancia. Busque o examine las AMI si no ve lo que busca a continuación.

Busque en nuestro catálogo completo que incluye miles de imágenes de sistemas operativos y aplicaciones

Recientes Inicio rápido

- Amazon Linux 4
- macOS
- Ubuntu
- Windows
- Red Hat
- S
- >
- Buscar más AMI
- Incluidas las AMI de AWS, Marketplace y la comunidad

Amazon Machine Image (AMI)

Amazon Linux 2 AMI (HVM) - Kernel 5.10, SSD Volume Type
ami-094125af156557ca2 (64 bits (x86)) / ami-0f96a89e4a6cf08cc (64 bits (Arm))
Virtualización: hvm Habilitado para ENA: true Tipo de dispositivo raíz: ebs

Apto para la capa gratuita ▾

▼ Par de claves (inicio de sesión) Información

Puede utilizar un par de claves para conectarse de forma segura a la instancia. Asegúrese de que tiene acceso al par de claves seleccionado antes de lanzar la instancia.

Nombre del par de claves - *obligatorio* C Crear un nuevo par de claves 5

Crear par de claves

Los pares de claves le permiten conectarse a la instancia de forma segura.

Escriba el nombre del par de claves a continuación. Cuando se lo pida, almacene la clave privada en una ubicación segura y accesible de su equipo. **Lo necesitará más adelante para conectarse a la instancia.** [Más información](#)

Nombre del par de claves El nombre puede incluir hasta 255 caracteres ASCII. No puede incluir espacios al principio ni al final.

Tipo de par de claves RSA Par de claves públicas y privadas cifradas por RSA

ED25519 Los pares de claves privadas y públicas cifradas ED25519 (no se admite para instancias de Windows)

Formato de archivo de clave privada .pem Para usar con OpenSSH

.ppk Para usar con PUTTY

Cancelar **Crear par de claves**

6. Asigne una VPC
7. Habilite la asignación automática de IP
8. Cree un grupo de seguridad
9. Lance la instancia

Crear EC2

▼ Configuraciones de red [Información](#)

VPC - **obligatorio** [Información](#)
vpc-0d26d791b86dd407e (test_vpc)
10.0.0.0/16

Subred [Información](#)
subnet-02a69a7fbab914859
VPC: vpc-0d26d791b86dd407e Propietario: 067205227321
Zona de disponibilidad: us-west-2a Direcciones IP disponibles: 4091
CIDR: 10.0.0.0/20

Asignar automáticamente la IP pública [Información](#)
Habilitar

Firewall (grupos de seguridad) [Información](#)
Un grupo de seguridad es un conjunto de reglas de firewall que controlan el tráfico de la instancia. Agregue reglas para permitir que un tráfico específico llegue a la instancia.

Crear grupo de seguridad 8 Seleccionar un grupo de seguridad existente

Nombre del grupo de seguridad - **obligatorio**
launch-wizard-2

Este grupo de seguridad se agregará a todas las interfaces de red. El nombre no se puede editar después de crear el grupo de seguridad. La longitud máxima es de 255 caracteres. Caracteres válidos: a-z, A-Z, 0-9, espacios y _:/() #,@[]+= &; {}! \$*

Descripción - **obligatorio** [Información](#)
launch-wizard-2 created 2022-11-24T14:54:06.871Z

Reglas de grupos de seguridad de entrada

▼ Regla del grupo de seguridad 1 (TCP, 22, 0.0.0.0/0) 9 Eliminar

Tipo Información	Protocolo Información	Intervalo de puertos Información
ssh	TCP	22
Tipo de origen Información	Origen Información	Descripción - <i>optional</i> Información
Cualquier lugar	<input type="text"/> Agregue CIDR, lista de prefijos	por ejemplo, SSH para Admin Desk

▼ Resumen

Número de instancias [Información](#)
1

Imagen de software (AMI)
Canonical, Ubuntu, 22.04 LTS, ...[más información](#)
ami-017fec1353bcc96e

Tipo de servidor virtual (tipo de instancia)
t2.micro

Firewall (grupo de seguridad)
Nuevo grupo de seguridad

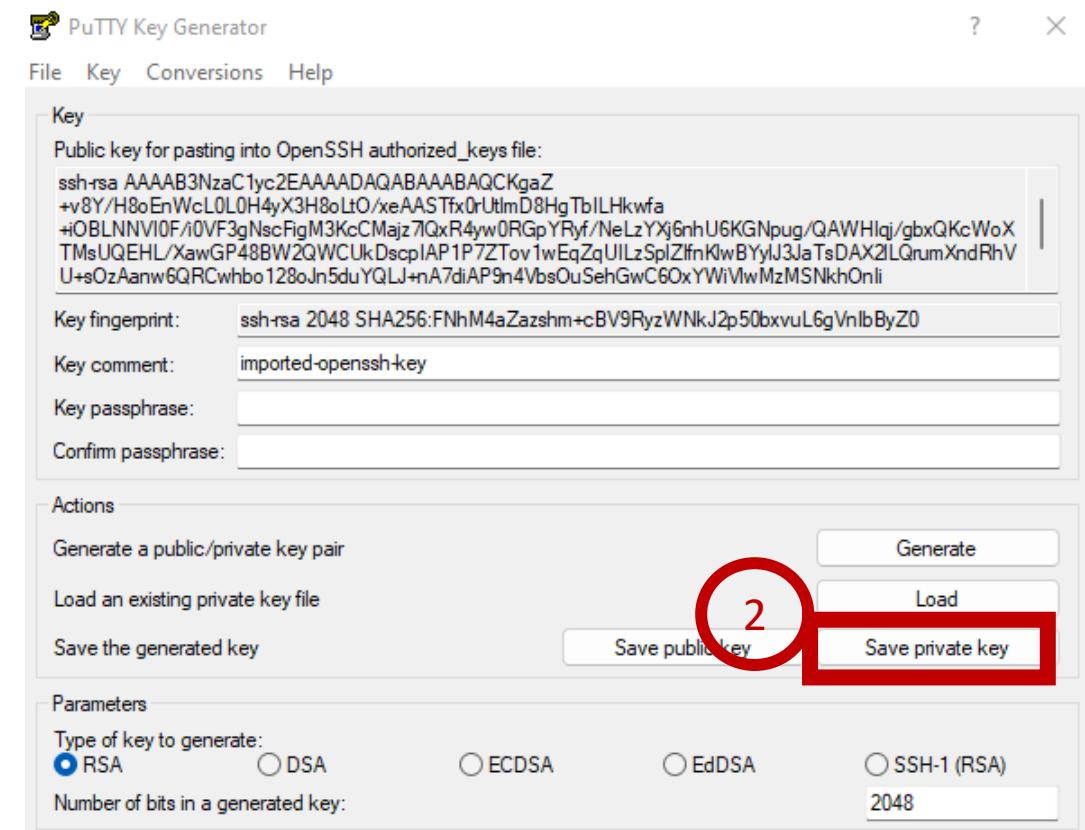
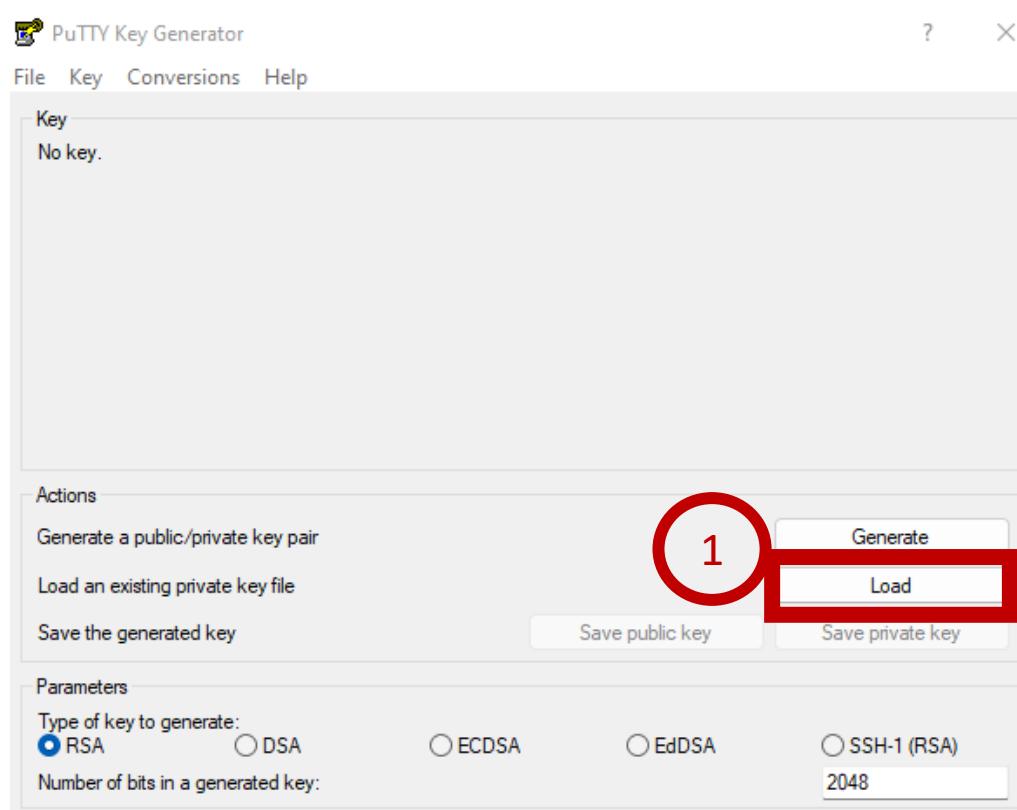
Almacenamiento (volúmenes)
1 volumen(es): 8 GiB

i **Nivel gratuito:** El primer año incluye 750 horas de uso de **instancias t2.micro** (o **t3.micro** en las regiones en las que **t2.micro** no esté disponible) en las AMI del nivel gratuito al mes, 30 GiB de almacenamiento de EBS, 2 millones de E/S, 1 GB de instantáneas y 100 GB de ancho de banda a Internet.

Cancelar Lanzar instancia

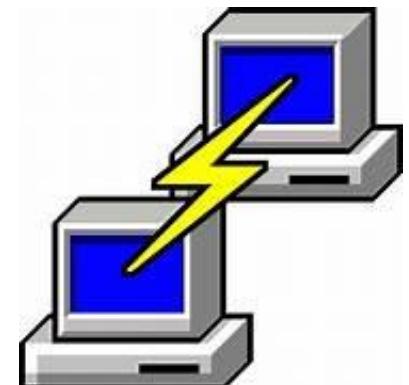
Creando clave Privada para Putty (Cliente SSH)

1. Abra Putty Key Generator y cargue la clave (*.pem) que descargó desde AWS
2. Haga clic en save private Key



Acceso a la Instancia desde Putty

- Conectarse a una instancia le permite interactuar con el sistema operativo para instalar las aplicaciones que desea utilizar
- Para conectarse desde Windows puede utilizar un cliente SSH como Putty siguiendo estos pasos:
 1. Identifique la IP pública de su instancia
 2. Ingrese el usuario@ip_pública en Putty
 - Instancias de AWS: ec2-user
 - Instancias Ubuntu: ubuntu
 3. Cargue las llaves de seguridad en Connection - SSH - Auth
 4. Inicie la conexión haciendo clic en Open



Acceso a la Instancia desde Putty

Instancias (1) [Información](#)

Find instancia by attribute or tag (case-sensitive)

C Conectar Estado de la instancia Acciones Lanzar instancias

Name	ID de la instancia	Estado de la instancia	Tipo de inst...	Comprobación ...	Estado de la alarma	Zona de dispon...	DNS de IPv4 pública	Dirección IPv4 pública
test_ec2	i-064fc17e0e6eed3e7	En ejecución	t2.micro	Inicializando	Sin alarmas	us-west-2a	ec2-35-162-151-62.us...	35.162.151.62

1

2

3

4

The screenshot illustrates the process of connecting to an AWS EC2 instance using PuTTY. It shows three main windows: the AWS CloudWatch Instances console, the PuTTY Configuration session window, and the PuTTY terminal window.

- AWS CloudWatch Instances:** Shows a single instance named "test_ec2" in the "En ejecución" state. The "Dirección IPv4 pública" column is highlighted with a red circle labeled "1".
- PuTTY Configuration Session:** The "Session" category is selected. The "Host Name" field contains "ubuntu@35.162.151.62" and the "Port" field is set to "22". This is highlighted with a red circle labeled "2".
- PuTTY Configuration Terminal:** The "Auth" category is selected. The "Private key file for authentication" field has a "Browse..." button highlighted with a red circle labeled "3".
- PuTTY Terminal:** The terminal window shows the Ubuntu login prompt: "ubuntu@ip-10-0-11-22: ~". The "Open" button at the bottom of the terminal window is highlighted with a red circle labeled "4".

EC2 – servidor LAMP

- LAMP se refiere a la instalación de los componentes básicos para crear un sitio web.
Linux – Apache – MySQL – PHP
- Siga la guía disponible en el siguiente enlace:

https://docs.aws.amazon.com/es_es/AWSEC2/latest/UserGuide/ec2-lamp-amazon-linux-2023.html

- Desde un navegador acceda a la dirección IP pública de su instancia EC2



EC2 - Wordpress



Productos ▾ Características ▾ Recursos ▾ Planes y Precios

Te damos la bienvenida al creador de páginas web más popular del mundo.

El 43 % de Internet está creado con WordPress. Hay más blogueros, pequeños negocios y grandes empresas de la lista Fortune 500 que usan WordPress que usuarios del resto de alternativas juntas. Únete a los millones de personas que han elegido WordPress.com.

[Empieza a crear tu página web](#)

<https://wordpress.com/es/>

EC2 - Wordpress

Vamos a crear una instancia EC2 de AWS a partir de una imagen de Wordpress

1. Cree una nueva instancia de EC2
2. En la sección de Imágenes de instancias y sistemas operativos seleccione la opción Buscar mas AMI
3. En el cuadro de búsqueda escriba bitnami-wordpress
4. Seleccione la opción bitnami-wordpress-x.x.x
5. Seleccione tipo de instancia t2.micro
6. Asigne claves de inicio de sesión
7. Asigne la VPN
8. Cree un nuevo grupo de seguridad
9. Lance la instancia

EC2 - Wordpress

Recientes Inicio rápido

Amazon Linux macOS Ubuntu Windows Red Hat S >

Buscar más AMI 2

Incluidas las AMI de AWS, Marketplace y la comunidad

Q wordpress X ▾

AMI de inicio rápido (0) Mis AMI (0) AMI de AWS Marketplace (390) AMI de la comunidad (500)

AMI de uso común Creadas por mí AWS y AMI de terceros de confianza Publicadas por cualquiera

▼ Linux/UNIX

Todos los de Linux/UNIX
 Amazon Linux
 CentOS
 Debian
 Fedora
 Gentoo
 macOS
 openSUSE

bitnami-wordpress-5.9.3-31-r01-linux-debian-11-x86_64-hvm-ebs-nami
ami-05ae896e953d55016
This image may not be the latest version available and might include security vulnerabilities. Please check the latest, up-to-date, available version at <https://bitnami.com/stacks>.

Plataforma: Debian Arquitectura: x86_64 Propietario: 979382823631
Fecha de publicación: 2022-06-27 Tipo de dispositivo raíz: ebs Virtualización: hvm
Habilitado para ENA: Sí

4 Seleccionar

EC2 - Wordpress

5

Tipo de instancia [Información](#)

Tipo de instancia

t2.micro Apto para la capa gratuita [Comparar tipos de instancias](#)

El proveedor de AMI recomienda usar una instancia t3a.small (o mayor) para disfrutar de una experiencia óptima con este producto.

Par de claves (inicio de sesión) [Información](#)

Puede utilizar un par de claves para conectarse de forma segura a la instancia. Asegúrese de que tiene acceso al par de claves seleccionado antes de lanzar la instancia.

Nombre del par de claves - **obligatorio** **6**

keys_1 [Crear un nuevo par de claves](#)

Configuraciones de red [Información](#)

VPC - obligatorio [Información](#)

vpc-0d26d791b86dd407e (test_vpc) 10.0.0.0/16 **7**

Subred [Información](#)

subnet-02a69a7fbab914859 [Crear una nueva subred](#) **8**

VPC: vpc-0d26d791b86dd407e Propietario: 067205227321 Zona de disponibilidad: us-west-2a Direcciones IP disponibles: 4091 CIDR: 10.0.0.0/20

Asignar automáticamente la IP pública [Información](#)

Habilitar

Firewall (grupos de seguridad) [Información](#)

Un grupo de seguridad es un conjunto de reglas de firewall que controlan el tráfico de la instancia. Agregue reglas para permitir que un **8**

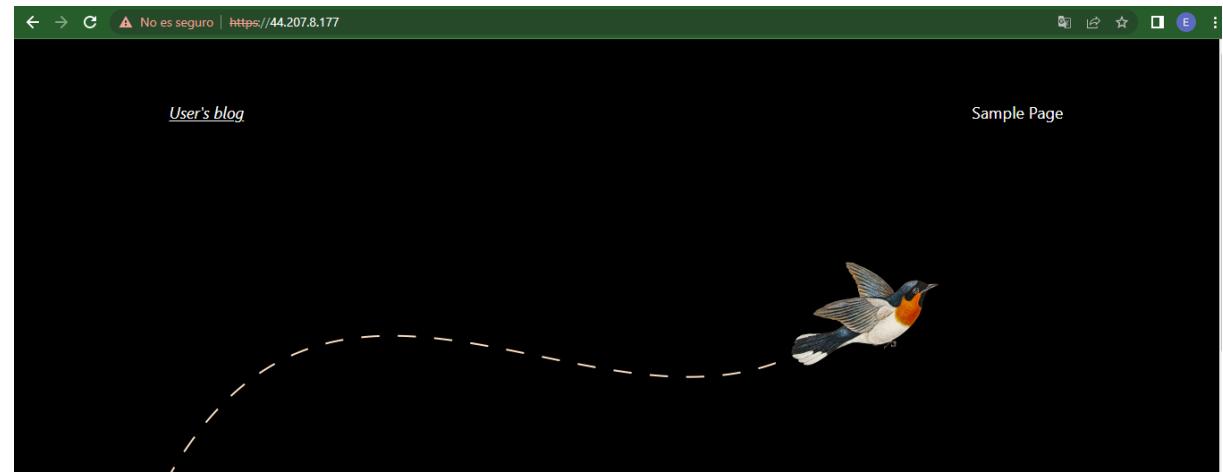
Crear grupo de seguridad Seleccionar un grupo de seguridad existente

Reglas de grupos de seguridad de entrada		
Regla del grupo de seguridad 1 (TCP, 22, 0.0.0.0/0) Eliminar		
Tipo Información	Protocolo Información	Intervalo de puertos Información
ssh	TCP	22
Tipo de origen Información	Origen Información	Descripción - optional Información
Cualquier lugar	<input type="text"/> Agregue CIDR, lista de prefijos 0.0.0.0/0 X	por ejemplo, SSH para Admin Desk
Regla del grupo de seguridad 2 (TCP, 80, 0.0.0.0/0) Eliminar		
Tipo Información	Protocolo Información	Intervalo de puertos Información
HTTP	TCP	80
Tipo de origen Información	Origen Información	Descripción - optional Información
Cualquier lugar	<input type="text"/> Agregue CIDR, lista de prefijos 0.0.0.0/0 X	por ejemplo, SSH para Admin Desk
Regla del grupo de seguridad 3 (TCP, 443, 0.0.0.0/0) Eliminar		
Tipo Información	Protocolo Información	Intervalo de puertos Información
HTTPS	TCP	443
Tipo de origen Información	Origen Información	Descripción - optional Información
Cualquier lugar	<input type="text"/> Agregue CIDR, lista de prefijos 0.0.0.0/0 X	por ejemplo, SSH para Admin Desk

EC2 - Wordpress

- Luego de que la instancia inicie, acceda desde el navegador a la ip_pública de su instancia
- Para configurar su página acceda desde el navegador a ip_pública/admin

Consulte usuario y contraseña en: Acciones – Monitoreo y solución de problemas – Obtener registro del sistema

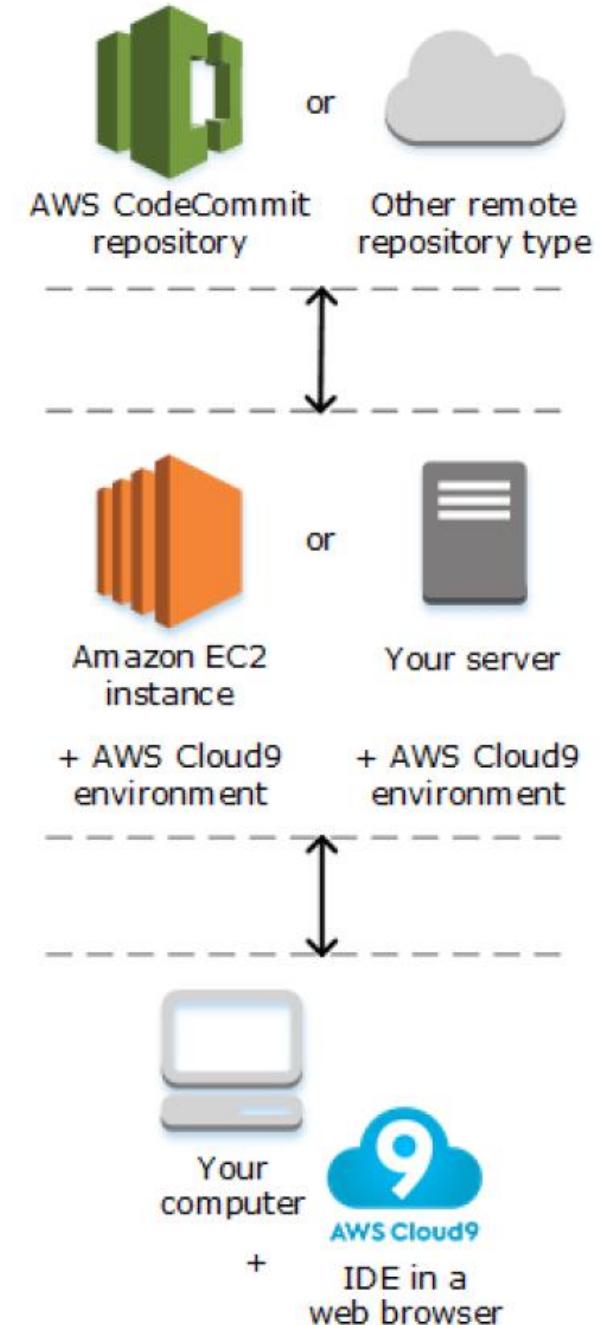


```
#####
#          Setting Bitnami application password to 'fjubskZkWgK0'
#          (the default application username is 'user')
#####
#####
```

Ingrese a la configuración de Wordpress y cree su propia página web

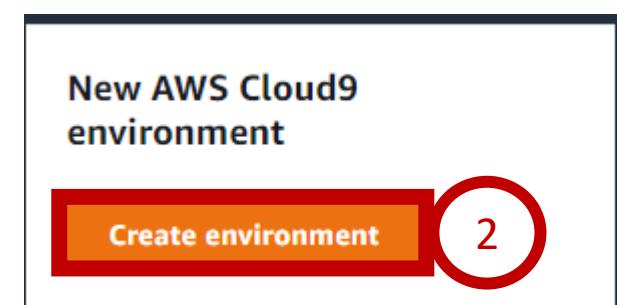
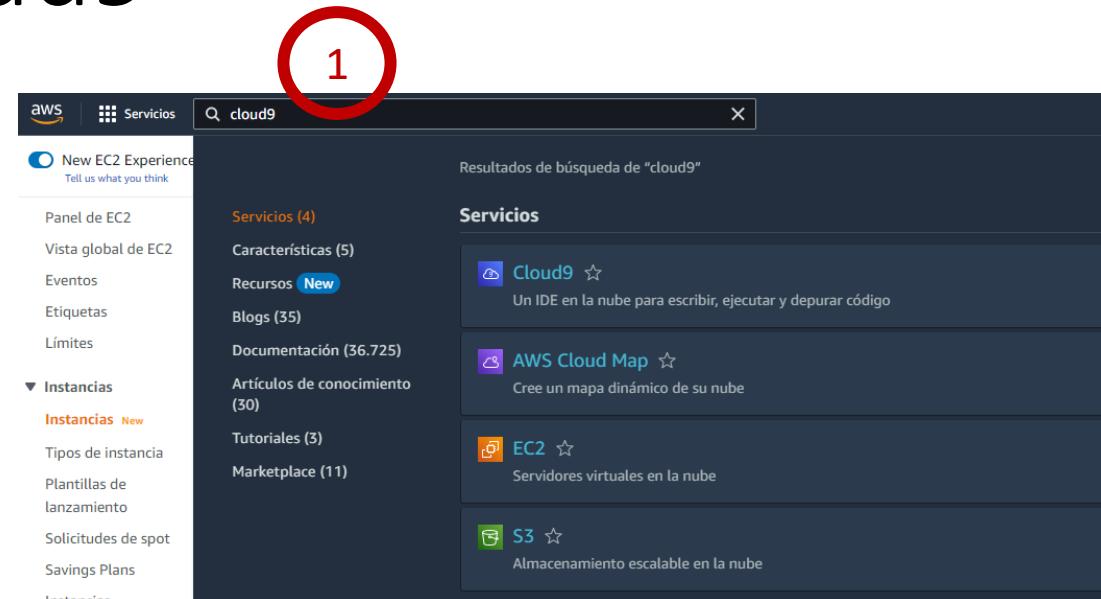
EC2 – Cloud9

- Cloud9 es un entorno de desarrollo integrado (IDE) en la nube de AWS
- Opera con varios lenguajes de programación
- Se accede a través de un navegador web
- Permite codificar, compilar, probar y depurar software



EC2 – Cloud9

1. En la barra de búsqueda digite Cloud9
2. Haga clic en Create Environment
3. Asigne un nombre
4. Seleccione New EC2 Instance
5. Seleccione tipo de instancia t2.micro
6. Asigne Secure Shell (SSH) como método de conexión
7. Asigne una VPC
8. Haga clic en Create



EC2 – Cloud9

Details

Name **3**
Limit of 60 characters, alphanumeric, and unique per user.

Description - optional

Limit 200 characters.

Environment type [Info](#) **4**
Determines what the Cloud9 IDE will run on.

New EC2 instance
Cloud9 creates an EC2 instance in your account. The configuration of your EC2 instance cannot be changed by Cloud9 after creation.

Existing compute
You have an existing instance or server that you'd like to use.

New EC2 instance

Instance type [Info](#) **5**
The memory and CPU of the Amazon EC2 instance that will be created for Cloud9 to run on.

t2.micro (1 GiB GiB RAM + 1 vCPU)
Free-tier eligible. Ideal for educational users and exploration.

t3.small (2 GiB GiB RAM + 2 vCPU)
Recommended for small web projects.

m5.large (8 GiB GiB RAM + 2 vCPU)
Recommended for production and most general-purpose development.

Additional instance types
Explore additional instances to fit your need.

Network settings [Info](#)

Connection
How your environment is accessed.

AWS Systems Manager (SSM)
Accesses environment via SSM without opening inbound ports (no ingress).

Secure Shell (SSH)
Accesses environment directly via SSH, opens inbound ports. **6**

VPC settings [Info](#)

Amazon Virtual Private Cloud (VPC)
The VPC that your environment will access. To allow the AWS Cloud9 environment to connect to this EC2 instance, attach an internet gateway (IGW) to your VPC. [Create new VPC](#)

vpc-0928635a3dcc4cec7 **7**
Name - test1_vpc

Subnet
Used to setup your VPC configuration. To use a private subnet, select AWS Systems Manager (SSM) as the connection type. [Create new subnet](#)

subnet-05c1b82cb78c93b57 **7**
Name - subnet_01

Tags - optional [Info](#)

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

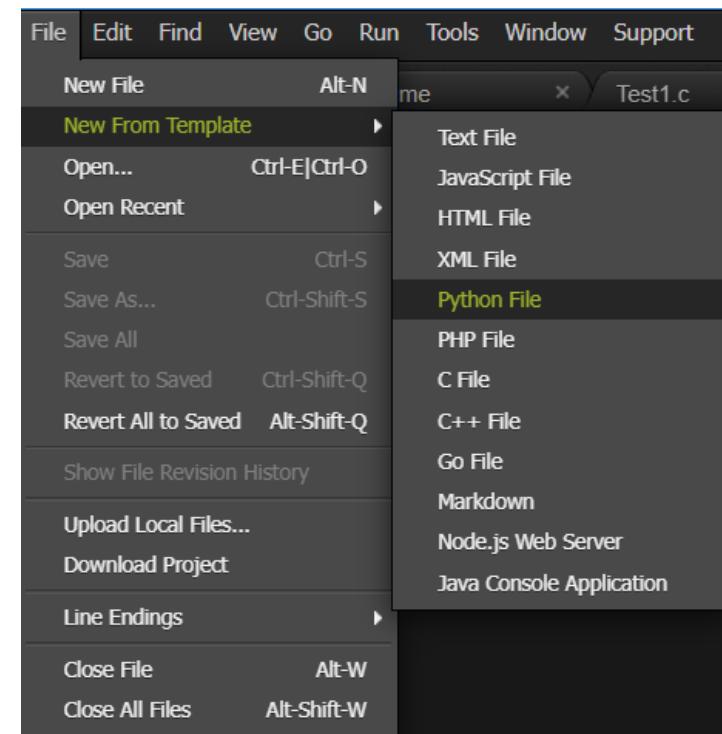
i The following IAM resources will be created in your account

- **AWSServiceRoleForAWSCloud9** - AWS Cloud9 creates a service-linked role for you. This allows AWS Cloud9 to call other AWS services on your behalf. You can delete the role from the AWS IAM console once you no longer have any AWS Cloud9 environments. [Learn more](#)

[Cancel](#) **8** [Create](#)

EC2 – Cloud9

Abra el ambiente de trabajo y cree un nuevo archivo para lenguaje Python



Ingrese el siguiente código y haga clic en run

```
1 """
2 Your module description
3 """
4 print('Calculadora sencilla en Python')
5 num1 = input('Ingrese el primer número: ')
6 num2 = input('Ingrese el segundo número: ')
7
8 sum = float(num1) + float(num2)
9
10 print('La suma de {0} y {1} es {2}'.format(num1, num2, sum))
```

AGENDA

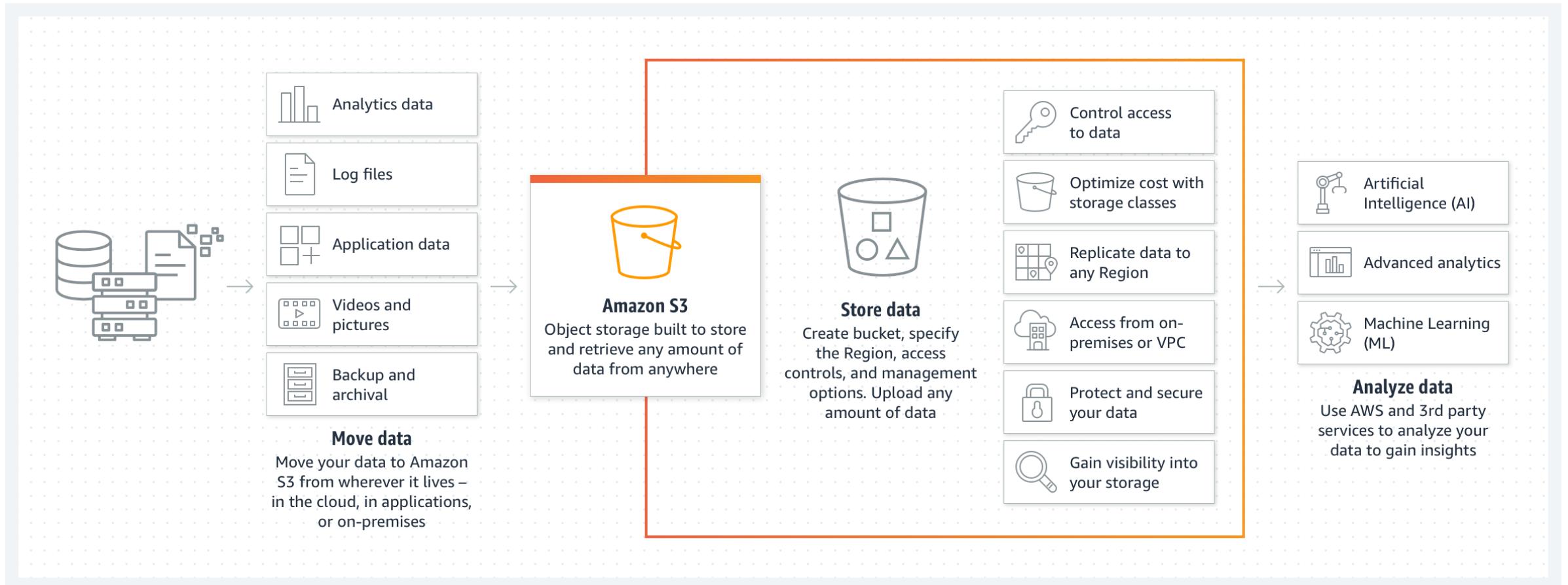
1. AWS
2. Nube Privada Virtual (VPC)
3. Procesamiento (EC2)
- 4. Almacenamiento (S3)**
5. BigData (EMR)
6. Monitoreo (CloudWatch)
7. Ejercicios

S3 (Simple Storage Service)

- Servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento.
- El almacenamiento se realiza a través de Buckets que son directorios lógicos donde se almacenan los datos como objetos
- Cada buckets debe tener un nombre único, se definen a nivel regional
- Se accede a los archivos (objetos) así:
 - `s3://nombre_bucket/carpeta/nombre_archivo`

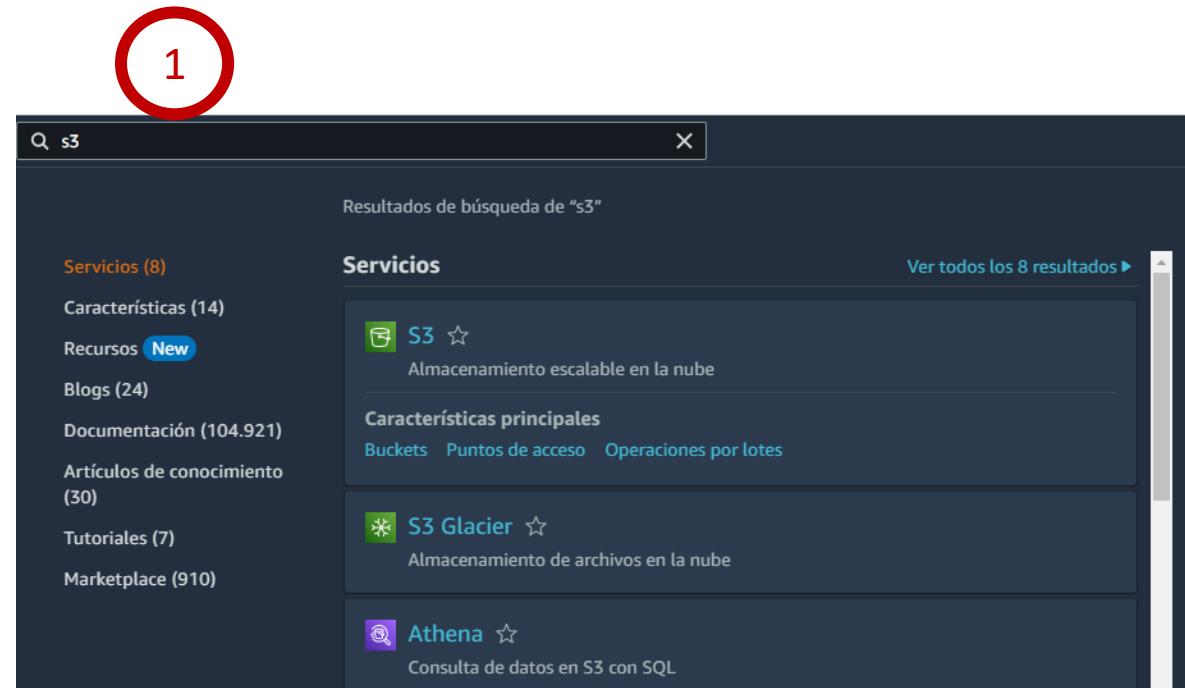


S3 (Simple Storage Service)



Crear Bucket

1. En la barra de búsqueda digite S3
2. Haga clic en Crear bucket
3. Asigne un nombre al bucket
4. Seleccione ACL Deshabilitadas
5. Haga clic en crear



Creación de un bucket

Cada objeto en S3 se almacena en un bucket. Para subir archivos y carpetas a S3, tendrá que crear un bucket donde se almacenarán los objetos.



Crear Bucket

Para subir archivos haga clic en cargar y arrastre el archivo que desea subir

Amazon S3 > Buckets > test12345bucket

test12345bucket [Info](#)

Objetos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (0)
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [índice de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones Crear carpeta Cargar

Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
No hay objetos				

No tiene objetos en este bucket.

Cargar

Cargar [Info](#)

Agregue los archivos y las carpetas que desea cargar en S3. Para cargar un archivo de más de 160 GB, utilice la CLI de AWS, el SDK de AWS o la API REST de Amazon S3. [Más información](#)

Arrastre y suelte los archivos y las carpetas que deseé cargar aquí, o elija **Add files** (Aregar archivos) o **Add folders** (Aregar carpetas).

Archivos y carpetas (0)
Se cargarán todos los archivos y las carpetas de esta tabla.

Eliminar Agregar archivos Agregar carpeta

Buscar por nombre

Nombre	Carpeta	Tipo	Tamaño
No hay archivos ni carpetas			

No ha elegido ningún archivo ni carpeta para cargar.

Crear Bucket

- Cree las siguientes carpetas:
 - input
 - output
 - script
- En la carpeta input suba el archivo: hurtos.csv

Objetos (3)		
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3		
C	<input type="checkbox"/> Copiar URI de S3	<input type="checkbox"/> Copiar URL
<input type="text"/> <i>Buscar objetos por prefijo</i>		
<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	input/	Carpeta
<input type="checkbox"/>	output/	Carpeta
<input type="checkbox"/>	script/	Carpeta

Objetos (1)		
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3		
C	<input type="checkbox"/> Copiar URI de S3	<input type="checkbox"/> Copiar URL
<input type="text"/> <i>Buscar objetos por prefijo</i>		
<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	hurtos.csv	CSV

AGENDA

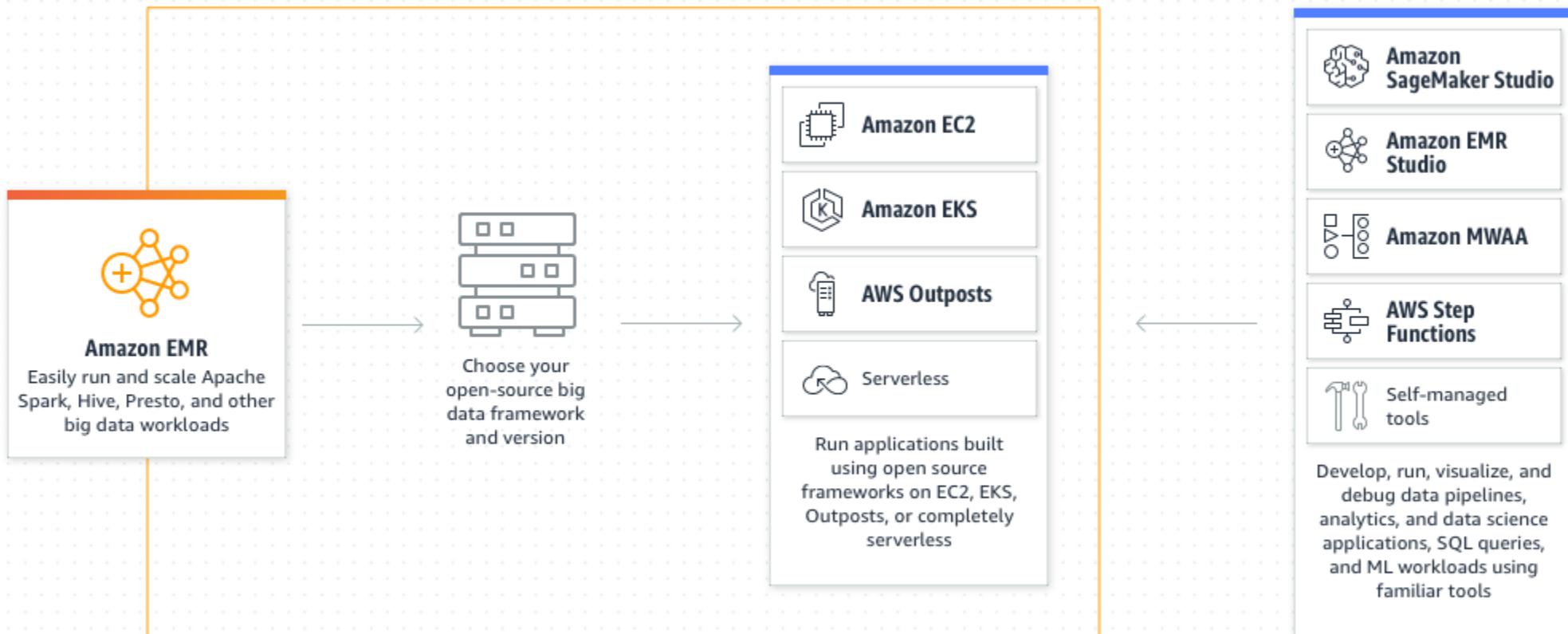
1. AWS
2. Nube Privada Virtual (VPC)
3. Procesamiento (EC2)
4. Almacenamiento (S3)
- 5. BigData (EMR)**
6. Monitoreo (CloudWatch)
7. Ejercicios

EMR (Elastic Map Reduce)

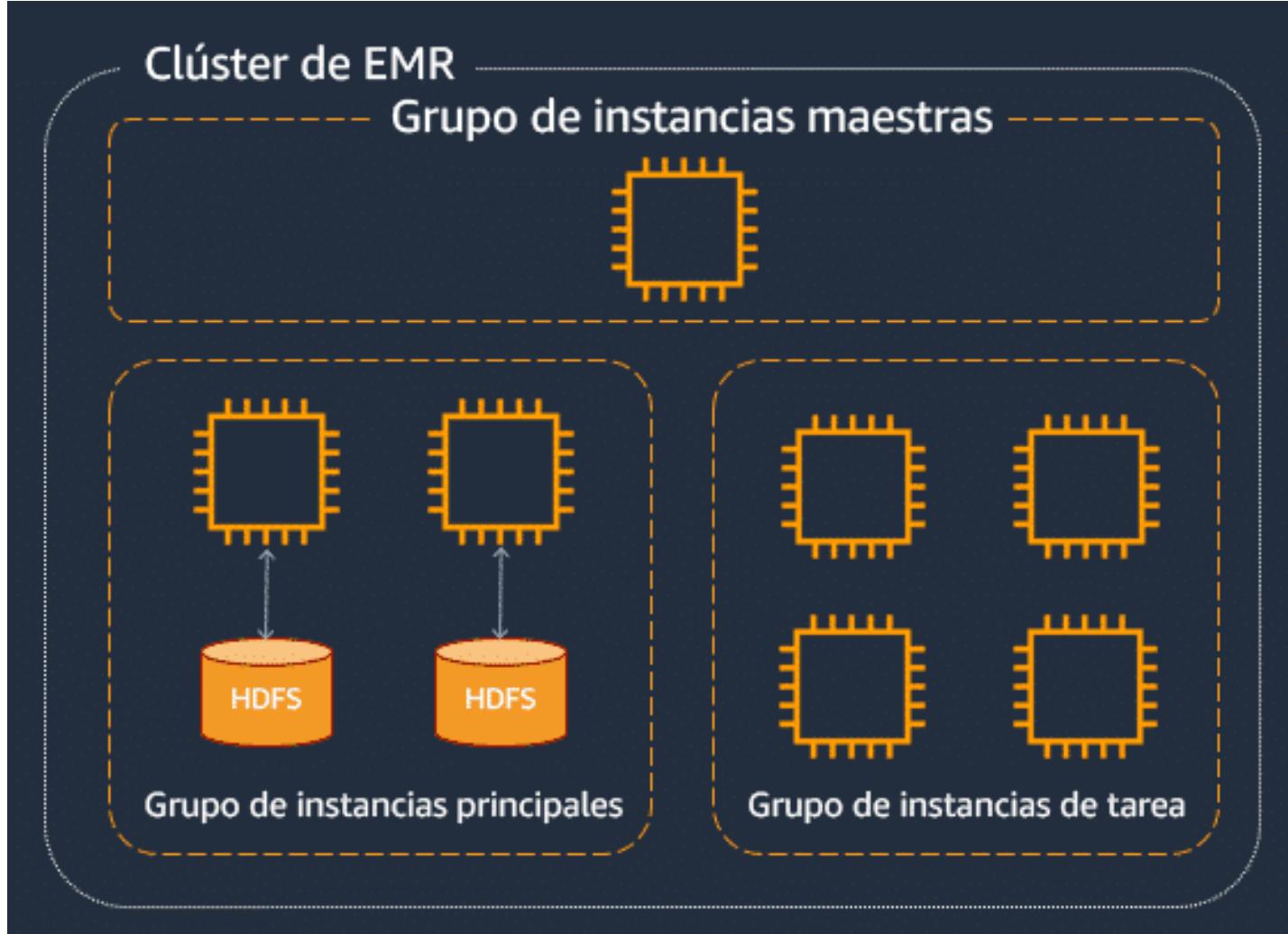
- Solución destinada al procesamiento de datos a escala de petabytes que implementa el paradigma Map-Reduce
- Simplifica la creación y operación de entornos y aplicaciones de Big Data
- Utiliza herramientas de código abierto como Hadoop, Apache Spark, Apache Hive y Presto
- Utiliza instancias EC2 para el procesamiento y es compatible con S3 para el almacenamiento



EMR

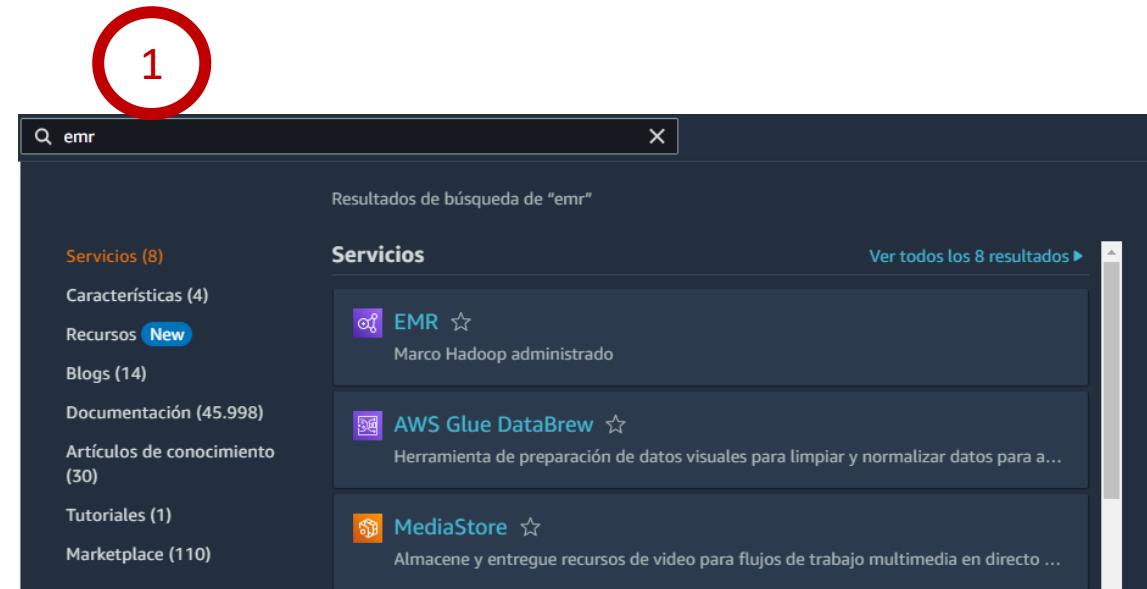


EMR



Crear EMR

1. En la barra de búsqueda digite EMR
2. Haga clic en Crear Clúster
3. Asigne un nombre al clúster
4. Habilite las herramientas a utilizar
5. Asigne el número y tipo de instancias a utilizar
6. Asigne la VPC
7. Seleccione las llaves de acceso
8. Asigne los roles de IAM
9. Haga clic en crear



Parece que no dispone de ningún clúster. Crear uno ahora:

A screenshot of the 'Create Cluster' button. The button is blue with white text and is highlighted by a red circle with the number '2'. The text on the button reads 'Crear clúster'.

Crear EMR

Crear clúster Información

Nombre y aplicaciones Información

Nombre

cluster_emr_1

Versión de Amazon EMR Información

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.

emr-6.14.0

Paquete de aplicaciones

Spark Interactive



Core Hadoop



Flink



HBase



Presto



Trino



Custom



- Flink 1.17.1
- HCatalog 3.1.3
- Hue 4.11.0
- Livy 0.7.1
- Phoenix 5.1.3
- Spark 3.4.1
- Tez 0.10.2
- ZooKeeper 3.5.10

- Ganglia 3.7.2
- Hadoop 3.3.3
- JupyterEnterpriseGateway 2.6.0
- MXNet 1.9.1
- Pig 0.17.0
- Sqoop 1.4.7
- Trino 422
- HBase 2.4.17
- Hive 3.1.3
- JupyterHub 1.5.0
- Oozie 5.2.1
- Presto 0.281
- TensorFlow 2.11.0
- Zeppelin 0.10.1

3

4

Aprovisionamiento y escalado de clústeres Información

Establezca las configuraciones de escalado y aprovisionamiento para los grupos de nodos principales y los nodos de tarea del clúster.

Elija una opción

Establecer el tamaño del clúster manualmente

Utilice esta opción si conoce los patrones de la carga de trabajo de antemano.

Utilizar escalado administrado por EMR

Supervise las métricas clave de la carga de trabajo de modo que EMR pueda optimizar el tamaño del clúster y la utilización de los recursos.

Utilizar el escalamiento automático personalizado

Para escalar mediante programación los nodos principales y los nodos de tarea, cree políticas de escalamiento automático personalizadas.

Configuración de aprovisionamiento

Establezca el tamaño del principal y tarea grupos de instancias. Amazon EMR intenta aprovisionar esta capacidad al lanzar el clúster.

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Central	m5.xlarge	1	<input type="checkbox"/>
Tarea - 1	m5.xlarge	1	<input type="checkbox"/>

5

Crear EMR

Redes [Información](#)

Virtual Private Cloud (VPC) [Información](#)

vpc-06e4b4b84091f3910 [Examinar](#) [Crear VPC](#)

Subred [Información](#)

subnet-0f8f83295bde7b2c5 [Examinar](#) [Crear subred](#)

Configuración de seguridad y par de claves de EC2: opcional [Información](#)

Configuración de seguridad
Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.

[Elegir una configuración de segur](#) [C](#) [Examinar](#) [Crear configuración de seguridad](#)

Par de claves de Amazon EC2 para el protocolo SSH al clúster [Información](#)

vockey [X](#) [Examinar](#) [Crear par de claves](#)

6

7

Roles de Identity and Access Management (IAM) [Información](#)
Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

Rol de servicio de Amazon EMR [Información](#)

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

Elegir un rol de servicio existente

Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

Crear un rol de servicio

Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

Rol de servicio

EMR_DefaultRole



8

Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

Elegir un perfil de instancia existente

Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

Crear un perfil de instancia

Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

Perfil de instancia

EMR_EC2_DefaultRole



9

Crear clúster

Crear EMR

cluster_emr_1

Se ha actualizado hace menos de un minuto



Terminar

Clonar en AWS CLI

Clonar

▼ Resumen

Información del clúster

ID del clúster
j-ATR17BHKW223

Configuración del clúster
Grupos de instancias

Capacidad
1 Primary (Principal) | 1 Principal | 1 Tarea

Aplicaciones

Versión de Amazon EMR
emr-6.14.0

Aplicaciones instaladas
Hadoop 3.3.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0,
JupyterHub 1.5.0, Livy 0.7.1, Spark 3.4.1

Administración de clústeres

Destino del registro en Amazon S3
aws-logs-136728460221-us-east-1/elasticmapreduce

DNS público del nodo principal
[ec2-44-198-55-55.compute-1.amazonaws.com](#)
[Conectarse al nodo principal mediante SSH](#)
[Conectarse al nodo principal mediante SSM](#)

Estado y hora

Estado
🕒 Comenzando

Hora de creación
10 de noviembre de 2023 9:43 (UTC-05:00)

Tiempo transcurrido
2 minutos

▼ Resumen

Información del clúster

ID del clúster
j-ATR17BHKW223

Configuración del clúster
Grupos de instancias

Capacidad
1 Primary (Principal) | 1 Principal | 1 Tarea

Aplicaciones

Versión de Amazon EMR
emr-6.14.0

Aplicaciones instaladas
Hadoop 3.3.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0,
JupyterHub 1.5.0, Livy 0.7.1, Spark 3.4.1

Administración de clústeres

Destino del registro en Amazon S3
aws-logs-136728460221-us-east-1/elasticmapreduce

IU de aplicación persistente
[Servidor de historial de Spark](#)
[Servidor de línea de tiempo de YARN](#)
[UI de Tez](#)

Estado y hora

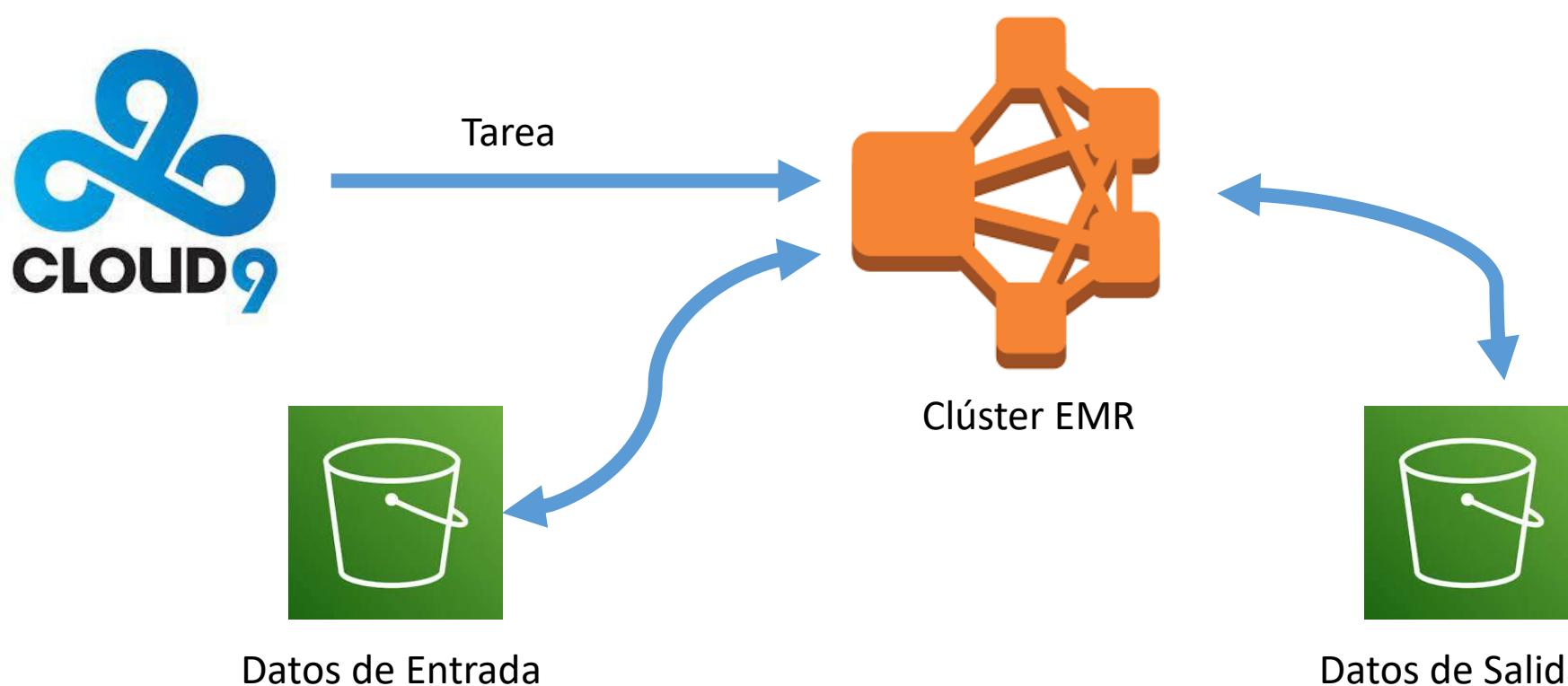
Estado
🕒 Esperando

Hora de creación
10 de noviembre de 2023 9:43 (UTC-05:00)

Tiempo transcurrido
13 minutos, 31 segundos

EMR – Acceso por consola

Objetivo: Lanzar una tarea al clúster desde Cloud9 en la que se procese un archivo leído desde S3 y se almacene el resultado en S3



EMR – Acceso por consola

1. Crear un bucket en AWS S3 y añadir directorios: input y output
2. En el directorio input del bucket subir el archivo hurtos.csv

test12345bucket [Info](#)

Objetos | Propiedades | Permisos | Métricas | Administración | Puntos de acceso

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener permisos de forma explícita. [Más información](#)

Copiar URI de S3 Copiar URL Descargar Abrir Eliminar

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo	Última modificación
<input type="checkbox"/>	input/	Carpeta	-
<input type="checkbox"/>	output/	Carpeta	-

Amazon S3 > Buckets > test12345bucket > input/ > Cargar

Cargar [Info](#)

Agreeve los archivos y las carpetas que desea cargar en S3. Para cargar un archivo de más de 160 GB, utilice la CLI de AWS, el SDK de AWS o la API REST de Amazon S3. [Más información](#)

Arrastre y suelte los archivos y las carpetas que deseé cargar aquí, o elija **Add files** (Agregar archivos) o **Add folders** (Agregar carpetas).

Archivos y carpetas (1 Total, 4.8 MB) Eliminar Agregar archivos Agregar carpeta

Se cargarán todos los archivos y las carpetas de esta tabla.

<input type="checkbox"/>	Nombre	Carpeta	Tipo	Tamaño
<input type="checkbox"/>	accidentalidad.csv	-	text/csv	4.8 MB

Destino

Destino
s3://test12345bucket/input/

Detalles del destino
Los ajustes del bucket que afectan a los objetos nuevos almacenados en el destino especificado.

Permisos
Conceder acceso público y acceso a otras cuentas de AWS.

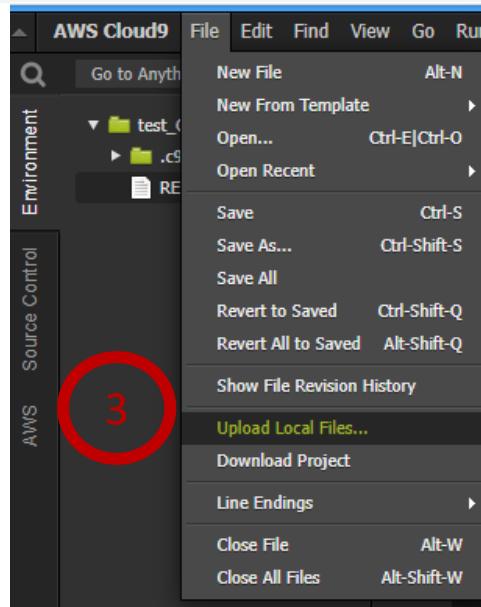
Propiedades
Especifique la clase de almacenamiento, los ajustes de cifrado, las etiquetas y mucho más.

Cancelar Cargar

EMR – Acceso por consola

3. Desde la interfaz de cloud9 subir el archivo de llaves keys_1.pem
4. Desde la consola de cloud9 modificar permisos de keys_1.pem

Environments (1)					
Name	Cloud9 IDE	Environment type	Connection	Permission	Owner ARN
test_Cloud9	Open	EC2 instance	Secure Shell (SSH)	Owner	arn:aws:iam::067205227321:root



3

A screenshot of a terminal window in AWS Cloud9. The command 'chmod 400 keys_1.pem' is being run, and the output shows the file permissions being changed. A red circle with the number 4 is drawn around the command line.

```
bash - "ip-172-31-29" x Immediate + ec2-user:~/environment $ chmod 400 keys_1.pem
```

4

EMR – Acceso por consola

5. Permitir conexión desde la dirección IP privada de la instancia Cloud9 en el clúster EMR. Para esto debes ir a EC2 seleccionar la instancia Cloud9 y tomar nota de la dirección IPV4 privada

Resumen de instancia de i-016c42d74175a899d (aws-cloud9-test-cloud9-7472536d0b944472b68a0f7e78fdf6c3) [Información](#)  [Conectar](#) [Estado de la instancia ▾](#) [Acciones ▾](#)
Se ha actualizado hace less than a minute

ID de la instancia  i-016c42d74175a899d (aws-cloud9-test-cloud9-7472536d0b944472b68a0f7e78fdf6c3)	Dirección IPv4 pública  34.222.161.66 dirección abierta 	Direcciones IPv4 privadas  10.0.5.50
Dirección IPv6 -	Estado de la instancia  En ejecución	DNS de IPv4 pública  ec2-34-222-161-66.us-west-2.compute.amazonaws.com dirección abierta 

EMR – Acceso por consola

6. Crear una regla de entrada en los grupos de seguridad del clúster EMR para permitir acceso SSH desde la IPV4 de Cloud 9

Resumen Historial de aplicaciones Monitorización Hardware Configuración

Resumen

ID: j-221S1NYGKIU9D
Fecha de creación: 2022-11-24 14:09 (UTC-5)
Tiempo transcurrido: 6 minutos
Terminar automáticamente: Cluster waits
Protección contra la terminación: Cambiar
Etiquetas: – Ver todo / Editar
DNS público principal: ec2-34-220-176-183.us-west-2.compute.amazonaws.com 
Connect to the Master Node Using SSH

Application user interfaces

Servicio de historial:  Spark history server, YARN timeline server
Conexiones:  Not Enabled Habilitar conexión web

Seguridad y acceso

Nombre de la clave: keys_1
Perfil de instancia EC2: EMR_EC2_DefaultRole
Función de EMR: EMR_DefaultRole
Función de Auto Scaling: EMR_AutoScaling_DefaultRole

Visibilidad:  Grupos de seguridad para sg-0e4dde440dfa6eee5 (ElasticMapReduce-principal: master) 
Grupos de seguridad para sg-06ff14a92aa64112a (ElasticMapReduce-slave) principal y tarea: 

Grupos de seguridad (2) Información

Filtrar grupos de seguridad

search: sg-0e4dde440dfa6eee5  Quitar los filtros

<input type="checkbox"/>	Name	ID del grupo de segu...	Nombre del grupo de segu...
<input type="checkbox"/>	–	sg-06ff14a92aa64112a	ElasticMapReduce-slave
<input type="checkbox"/>	–	sg-0e4dde440dfa6eee5	ElasticMapReduce-master

Editar reglas de entrada

Agregar regla

SSH TCP 22 Personaliza... 10.0.5.50/32 

C Administrar etiquetas  Editar reglas de entrada < 1 > 

C Cancelar Previsualizar los cambios Guardar reglas

EMR – Acceso por consola

7. Tomar nota de la información de la conexión SSH del clúster en la pestaña Resumen – DNS público principal

Resumen Historial de aplicaciones Monitorización Hardware Configuraciones Eventos Pasos Acciones de arranque

Resumen

ID: j-221S1NYGKIU9D
Fecha de creación: 2022-11-24 14:09 (UTC-5)
Tiempo transcurrido: 14 minutos
Terminar automáticamente: Cluster waits
Protección contra la Act. Cambiar terminación:
Etiquetas: – Ver todo / Editar
DNS público principal: ec2-34-220-176-183.us-west-2.compute.amazonaws.com 
Connect to the Master Node Using SSH

7

Application user interfaces

Servicio de historial:  Spark history server, YARN timeline server
Conexiones:  Not Enabled Habilitar conexión web

Detalles de las configuraciones

Etiqueta de la versión: emr-5.36.0
Distribución Hadoop: Amazon 2.10.1
Aplicaciones: Hue 4.10.0, Presto 0.267, Spark 2.4.8
URI de registro: s3://aws-logs-067205227321-us-west-2/elasticmapreduce/ 
Vista coherente de EMRFS: Deshabilitados
ID de AMI personalizada: –
Versión de Amazon Linux: 2.0.20221004.0 [Más información](#) 

Redes y hardware

Zona de disponibilidad: us-west-2a
ID de subred: [subnet-02a69a7fbab914859](#) 
Maestro: En ejecución 1 m5.xlarge
Principal: En ejecución 1 m5.xlarge
Tarea: –
Cluster scaling: Not enabled
Terminación automática: Terminar si permanece inactivo durante 1 hora

EMR – Acceso por consola

8. Desde la consola de Cloud9 conectarse por SSH al clúster EMR con el siguiente comando

```
ssh -i "nombre_llaves" hadoop@ "DNS público principal EMR"
```

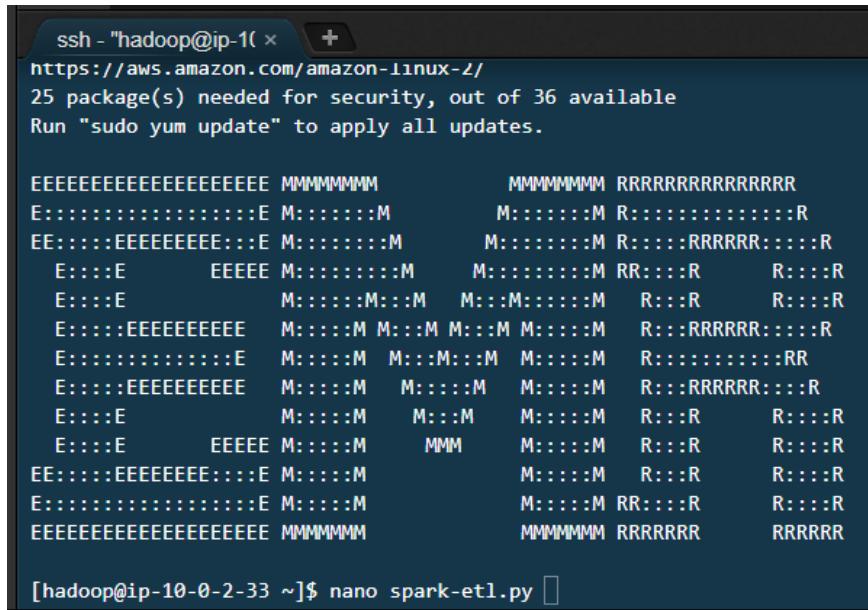
```
bash - "ip-10-0-5-50. x  Immediate  +  
ec2-user:~/environment $ ssh -i keys_1.pem hadoop@ec2-34-220-176-183.us-west-2.compute.amazonaws.com
```

```
hadoop@ip-10-0-14 ~] Immediate +  
ec2-user:~/environment $ ssh -i keys_1.pem hadoop@ec2-34-220-176-183.us-west-2.compute.amazonaws.com  
The authenticity of host 'ec2-34-220-176-183.us-west-2.compute.amazonaws.com (10.0.14.162)' can't be established.  
ECDSA key fingerprint is SHA256:RkDdu1f1oEKBHALPeFIzebqXHTDQINaIJFV9Ku+12w.  
ECDSA key fingerprint is MD5:dc:60:55:93:4c:23:92:4e:f6:56:b2:92:d8:20:de:44.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added 'ec2-34-220-176-183.us-west-2.compute.amazonaws.com,10.0.14.162' (ECDSA) to the list of known hosts.  
Last login: Thu Nov 24 19:18:41 2022  
  
_ _| _ _|_) )  
_ | ( _ / Amazon Linux 2 AMI  
_ \_\_||_|  
  
https://aws.amazon.com/amazon-linux-2/  
25 package(s) needed for security, out of 36 available  
Run "sudo yum update" to apply all updates.  
  
EEEEEEEEEEEEEEEEEE MMWWWWWW MWWWWWWM RRRRRRRRRRRRRRRR  
E:::::::::::E E M:::::M M:::::M R:::::R R:::::R  
EE:::::EEEEEEEEE:::E M:::::M M:::::M R:::::R RRRRRRR:::::R  
 E:::E EEEEEEE M:::::M M:::::M RR:::::R R:::::R  
 E:::::E M:::::M M:::::M M:::::M R:::R R:::::R  
 E:::::EEEEEEEEEE M:::::M M:::::M M:::::M R:::::RRRRRR:::::R  
 E:::::::::::E M:::::M M:::::M M:::::M R:::::RRRRRRRR:::::R  
 E:::::EEEEEEEEE M:::::M M:::::M M:::::M R:::::RRRRRRRR:::::R  
 E:::::E M:::::M M:::::M M:::::M R:::R R:::::R  
 E:::::E EEEEEEE M:::::M MMW M:::::M R:::R R:::::R  
EE:::::EEEEEEEEE:::E M:::::M M:::::M R:::::R R:::::R  
E:::::::::::E E M:::::M M:::::M RR:::::R R:::::R  
EEEEEEEEEEEEEEEEEE MMWWWWWW MWWWWWWM RRRRRRRR RRRRRR  
  
[hadoop@ip-10-0-14-162 ~]$
```

EMR – Acceso por consola

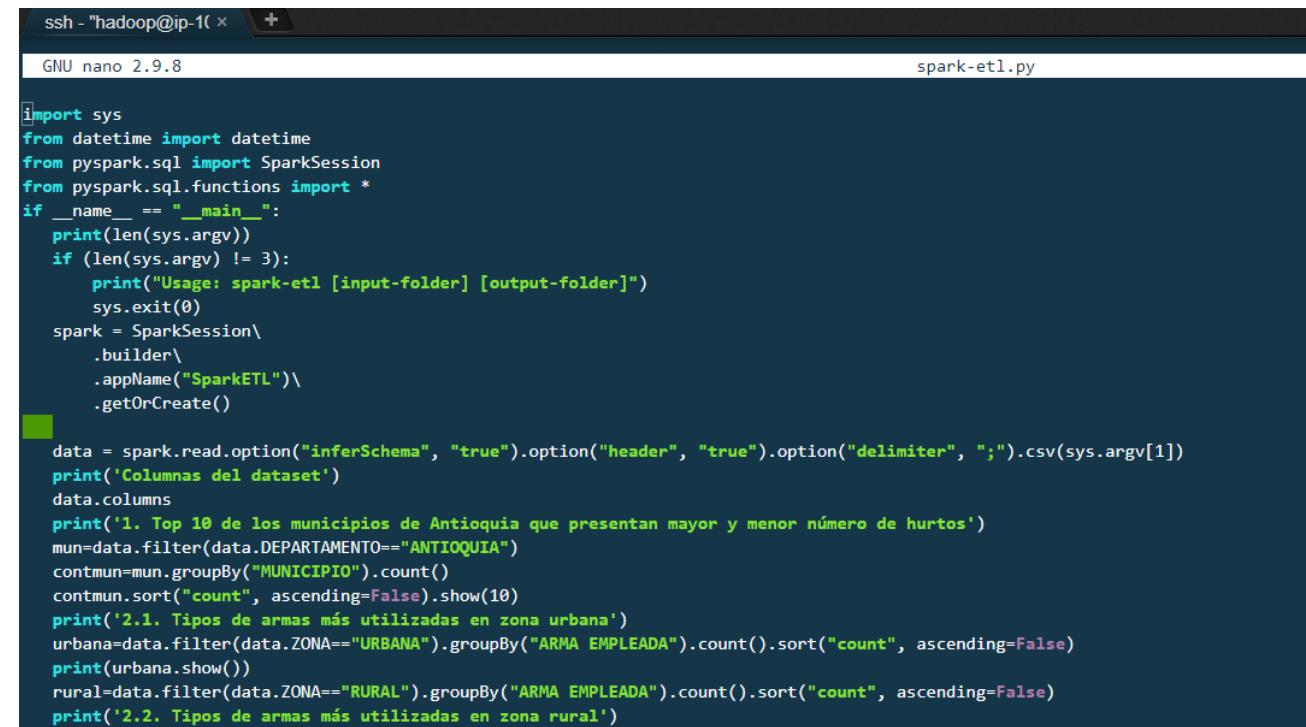
9. Desde la terminal SSH con EMR, crear un archivo donde se ejecutará el código
 - Usa el comando nano spark-etl.py para crear el archivo
 - Copia el contenido del archivo script_hurtos.txt al archivo spark-etl.py
 - Guarda el archivo pulsando CTRL + X luego confirma con Y, luego presiona ENTER

```
ssh - "hadoop@ip-10-0-2-33 ~]$ nano spark-etl.py
[REDACTED]
```



```
import sys
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
if __name__ == "__main__":
    print(len(sys.argv))
    if (len(sys.argv) != 3):
        print("Usage: spark-etl [input-folder] [output-folder]")
        sys.exit(0)
    spark = SparkSession\
        .builder\
        .appName("SparkETL")\
        .getOrCreate()
    data = spark.read.option("inferSchema", "true").option("header", "true").option("delimiter", ";").csv(sys.argv[1])
    print('Columnas del dataset')
    data.columns
    print('1. Top 10 de los municipios de Antioquia que presentan mayor y menor número de hurtos')
    mun=data.filter(data.DEPARTAMENTO=="ANTIOQUIA")
    contmun=mun.groupBy("MUNICIPIO").count()
    contmun.sort("count", ascending=False).show(10)
    print('2.1. Tipos de armas más utilizadas en zona urbana')
    urbana=data.filter(data.ZONA=="URBANA").groupBy("ARMA EMPLEADA").count().sort("count", ascending=False)
    print(urbana.show())
    rural=data.filter(data.ZONA=="RURAL").groupBy("ARMA EMPLEADA").count().sort("count", ascending=False)
    print('2.2. Tipos de armas más utilizadas en zona rural')
```

```
ssh - "hadoop@ip-10-0-2-33 ~]$ nano spark-etl.py
[REDACTED]
```



```
import sys
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
if __name__ == "__main__":
    print(len(sys.argv))
    if (len(sys.argv) != 3):
        print("Usage: spark-etl [input-folder] [output-folder]")
        sys.exit(0)
    spark = SparkSession\
        .builder\
        .appName("SparkETL")\
        .getOrCreate()
    data = spark.read.option("inferSchema", "true").option("header", "true").option("delimiter", ";").csv(sys.argv[1])
    print('Columnas del dataset')
    data.columns
    print('1. Top 10 de los municipios de Antioquia que presentan mayor y menor número de hurtos')
    mun=data.filter(data.DEPARTAMENTO=="ANTIOQUIA")
    contmun=mun.groupBy("MUNICIPIO").count()
    contmun.sort("count", ascending=False).show(10)
    print('2.1. Tipos de armas más utilizadas en zona urbana')
    urbana=data.filter(data.ZONA=="URBANA").groupBy("ARMA EMPLEADA").count().sort("count", ascending=False)
    print(urbana.show())
    rural=data.filter(data.ZONA=="RURAL").groupBy("ARMA EMPLEADA").count().sort("count", ascending=False)
    print('2.2. Tipos de armas más utilizadas en zona rural')
```

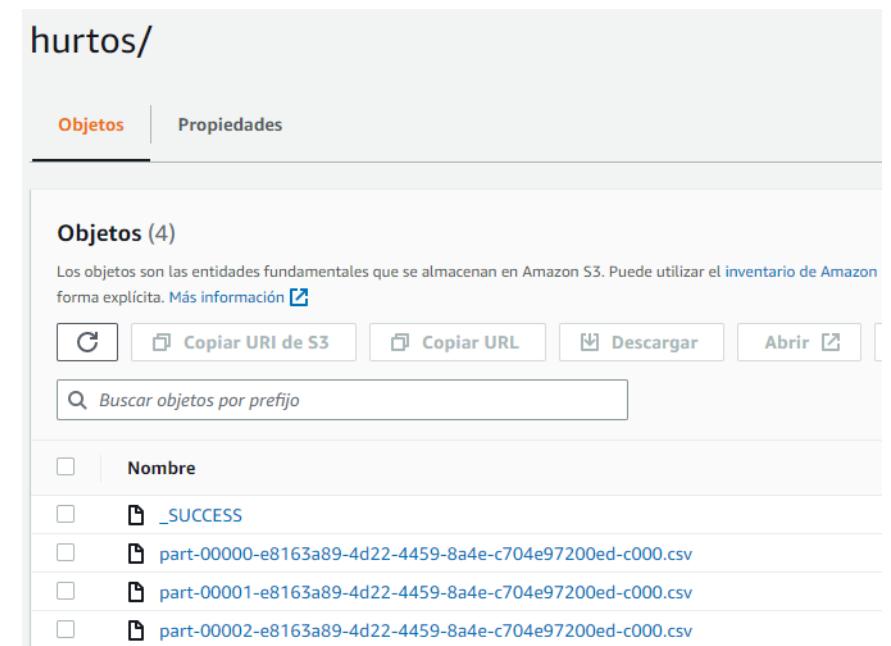
EMR – Acceso por consola

10. Desde la terminal Cloud9 conectada a EMR ejecuta el siguiente comando para ejecutar el script almacenado en el paso anterior

```
spark-submit spark-etl.py s3://“nombre_bucket”/input/ s3://“nombre_bucket”/output/hurtos
```

Puedes visualizar el resultado en la consola y en el bucket de S3

```
ssh - "hadoop@ip-10-0-2-33 ~$ spark-submit spark-etl.py s3://test12345bucket/input/ s3://test12345bucket/output/hurtos
3
22/11/29 23:12:02 INFO SparkContext: Running Spark version 2.4.8-amzn-2
22/11/29 23:12:02 INFO SparkContext: Submitted application: SparkETL
22/11/29 23:12:02 INFO SecurityManager: Changing view acls to: hadoop
22/11/29 23:12:02 INFO SecurityManager: Changing modify acls to: hadoop
22/11/29 23:12:02 INFO SecurityManager: Changing view acls groups to:
22/11/29 23:12:02 INFO SecurityManager: Changing modify acls groups to:
22/11/29 23:12:02 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view per
with modify permissions: Set(hadoop); groups with modify permissions: Set()
```



Lanzar tareas desde la interfaz de AWS

1. En S3 cree la carpeta script y suba el archivo hurtos.py
2. Ingrese a la ventana de configuración del clúster y seleccione la pestaña pasos
3. Haga clic en agregar paso

The screenshot shows the AWS Step Functions console interface. At the top, there is a navigation bar with tabs: Propiedades, Acciones de arranque, Instancias (hardware), Pasos (highlighted in blue), Aplicaciones, Configuraciones, Monitorización, Eventos, and Etiquetas (0). Below the navigation bar, there is a section titled "Pasos (0) Información" with the subtext "Cada paso es una unidad de trabajo que contiene instrucciones para manipular los datos para su procesamiento por software instalado en el clúster." There are buttons for "Actualizar tabla", "Cancelar pasos", "Cerrar paso", and "Agregar paso". A red circle with the number "2" is drawn around the "Pasos" tab. A red box with the number "3" is drawn around the "Agregar paso" button. At the bottom, there is a table header with columns: ID de paso, Estado, Nombre, Archivos de registro, Hora de creación (UTC-0...), Hora de inicio (UTC-05:...), and Tiempo transcurrido. The table body displays the message "No hay coincidencias" and "No se encuentra ninguna coincidencia".

Lanzar tareas desde la interfaz de AWS

4. En tipo de paso seleccione la opción JAR Personalizado

Agregar paso [Información](#)

Configuración de pasos

Tipo

JAR personalizado

Agrega un paso que le permite escribir un script personalizado para procesar los datos utilizando el lenguaje de programación Java.

4

Programa de transmisión

Agrega un paso que utiliza la entrada estándar para ejecutar scripts de asignación/reducción y enviar los resultados a la salida estándar.

Aplicación de Spark

Agrega un paso que envía el trabajo al marco de Spark en el clúster.

Programa de Hive

Agrega un paso que envía un script de Hive para las interacciones de almacenamiento de datos.

Script de shell

Solucione los problemas que se presentan con el clúster.

Nombre

hurtos

5

Ubicación de JAR

La ubicación de JAR puede ser una ruta en S3 o una base de nombre completo en el classpath.

command-runner.jar

6



[Ver](#)

[Explorar S3](#)

[Explorar S3](#)

Argumentos - opcional [Información](#)

Se pasan a la función principal en el archivo JAR. Si el archivo JAR no especifica una categoría principal en su archivo de manifiesto, puede especificar otro nombre de categoría como primer argumento.

spark-submit

s3://bktest1192023/script/hurtos.py

s3://bktest1192023/input/hurtos.csv

s3://bktest1192023/output/hurtos/

7

5. Asigne un nombre al paso

6. En la ubicación del JAR escriba:
command-runner.jar

7. En los argumentos indique:

spark-submit

ruta_s3_script_a_ejecutar

ruta_s3_data

ruta_s3_resultado

8. Haga clic en agregar paso

Lanzar tareas desde la interfaz de AWS

Paso
Pendiente

Pasos (1) Información

Cada paso es una unidad de trabajo que contiene instrucciones para manipular los datos para su procesamiento por software instalado en el clúster.

Pasos simultáneos: 1 

	ID de paso	Estado	Nombre	Archivos de registro
<input type="checkbox"/> 	s-050317010QOHH3JMFUG9	 Pending	hurtos	No se han creado registros aún 

Paso
completado

Pasos (1) Información

Cada paso es una unidad de trabajo que contiene instrucciones para manipular los datos para su procesamiento por software instalado en el clúster.

Pasos simultáneos: 1 

	ID de paso	Estado	Nombre	Archivos de registro
<input type="checkbox"/> 	s-050317010QOHH3JMFUG9	 Completed	hurtos	controller syslog stderr stdout 

Ejemplo Streaming WordCount

Objetivo: Crear un cluster que se conecte como cliente a un servidor a través de un puerto TCP y cuente cuantas veces se repite cada una de las palabras emitidas.



Ejemplo Streaming WordCount

1. Desde la consola de Cloud9 conectarse al clúster EMR con el siguiente comando
ssh -i "nombre_llaves" hadoop@"DNS público principal EMR"
2. Abrir una segunda terminal en Cloud9 para instalar netcat y abrir el puerto 8083 para publicar las palabras

```
sudo yum install netcat  
nc -l 8083
```
3. En la terminal de conexión con el clúster, crear el archivo **StreamingWordCount.py** para realizar el conteo de las palabras. Tome nota de la dirección IP de la instancia Cloud9 y modifíquela en el archivo
4. Lanzar el contador de palabras con el siguiente comando

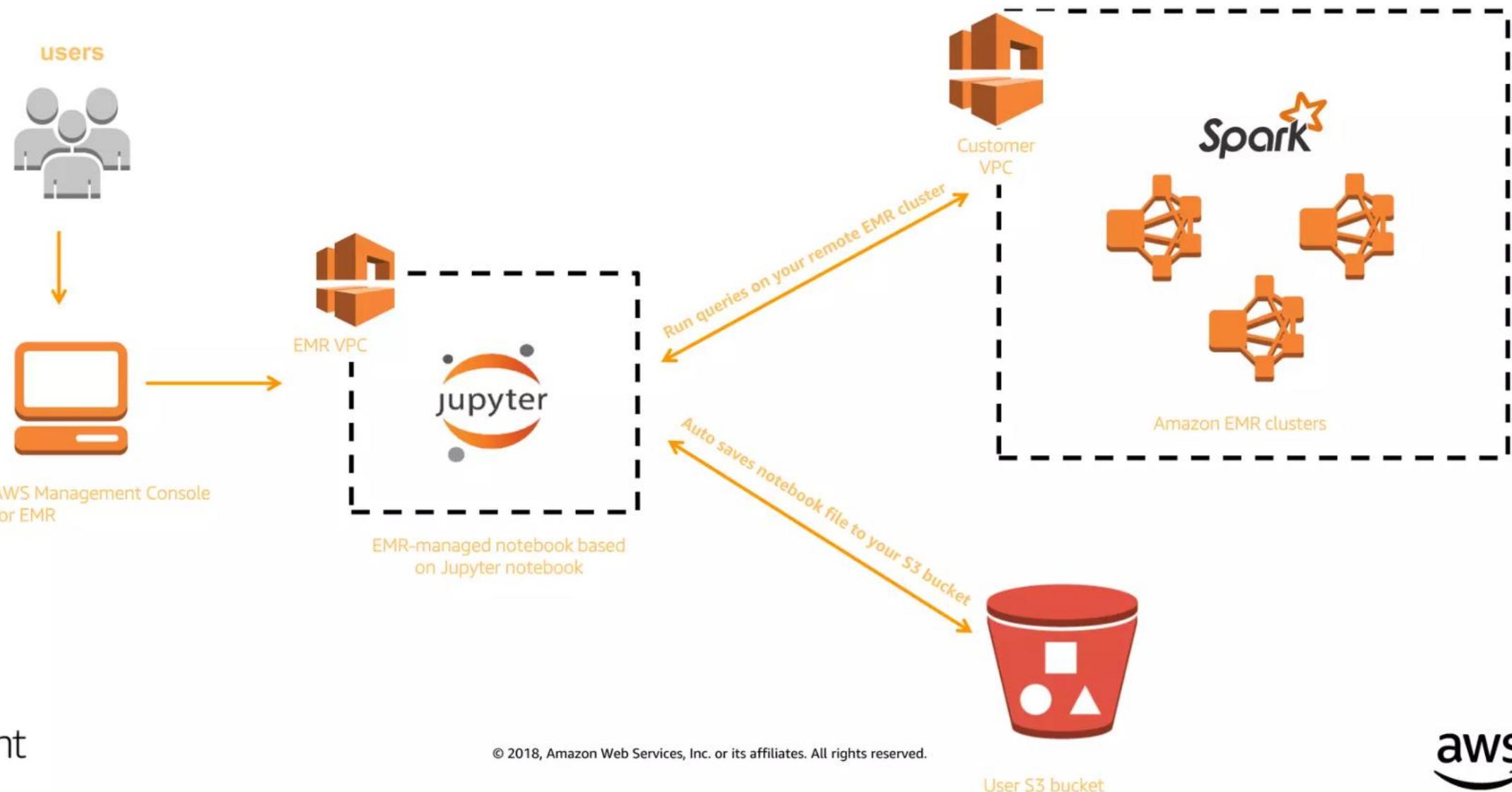
```
spark-submit StreamingWordCount.py
```

Time: 2023-11-10 15:36:40

```
-----  
('hola', 6)  
('curso', 2)  
('streaming', 2)  
('aws', 6)  
(' ', 1)  
('udea', 6)
```

Nota: Debe permitir conexiones por el puerto TCP 8083 en la instancia de Cloud9

NoteBook



The screenshot shows the AWS Management Console with the 'Amazon EMR' service selected. The left sidebar has sections for 'EMR sin servidor', 'EMR en EC2' (expanded to show 'Clústeres', 'Blocos de notas y repositorios de Git', 'Eventos', and 'Bloquear el acceso público'), and 'Configuraciones de seguridad'. Below that is 'EMR en EKS' (expanded to show 'Clústeres virtuales') and 'EMR Studio' (expanded to show 'Introducción', 'Studios' which is highlighted with a red circle, and 'WorkSpaces (Cuadernos)').

aws | Servicios | X

Amazon EMR

EMR sin servidor

▼ EMR en EC2

- Clústeres
- Blocs de notas y repositorios de Git
- Eventos
- Bloquear el acceso público
- Configuraciones de seguridad

▼ EMR en EKS

- Clústeres virtuales

▼ EMR Studio

- Introducción
- Studios**
- WorkSpaces (Cuadernos)

Crear EMR Studio y WorkSpace

1. Desde el menú de EMR haga clic en Studios y luego haga clic en Crear Studio
2. Seleccione Personalizado en Opciones de configuración
3. Indique un nombre para su NoteBook
4. Seleccione el rol de servicio LabRole
5. Asigne un nombre al espacio de trabajo (WorkSpace)
6. Asigne la VPC y subredes
7. Haga clic en Crear un Studio
8. Desde el menú de EMR haga clic en WorkSpaces, seleccione el espacio de trabajo creado y desde el menú de acciones haga clic en Detener

Crear EMR Studio y WorkSpace

Crear un Studio Información

Opciones de configuración Información

Cargas de trabajo interactivas

Trabajos por lotes

Personalizado

2

Configuración de Studio Información

Nombre del Studio

Studio_1

3

Utilice hasta 256 caracteres (alfanuméricos, guiones o guiones bajos).

Rol de servicio para permitir que Studio acceda a sus recursos de AWS

LabRole

4



[Ver detalles del permiso](#)

Configuración del espacio de trabajo Información

Nombre del espacio de trabajo

Studio_1_Workspace_1

5

Permitir la colaboración

[Eliminar](#)

Utilice hasta 256 caracteres (alfanuméricos, guiones o guiones bajos).

▼ Redes y seguridad - *opcional*

VPC Información

Seleccione una VPC para que su estudio la utilice cuando se comunique con los clústeres de EMR. Para usar claves de condición como las del ejemplo [políticas de roles de servicio para Amazon EMR](#), debe etiquetar la VPC con la clave `for-use-with-amazon-emr-managed-policies` y el valor `true`. Para administrar etiquetas, utilice [Panel de VPC](#).

vpc-0d71c84d6397f50ea (emr-vpc)



6

Subredes Información

Seleccione las subredes que su estudio puede usar cuando se comunique con los clústeres de EMR. Para usar claves de condición como las del ejemplo [políticas de roles de servicio para Amazon EMR](#), debe etiquetar cada subred con la clave `for-use-with-amazon-emr-managed-policies` y el valor `true`. Para administrar etiquetas, utilice [Panel de VPC](#).

[Seleccionar subredes](#)



subnet-0c3894eb5bec0ca9b (emr-subnet-public1-us-east-1a) X

10.0.0.0/20 - us-east-1a - 4088 direcciones IP disponibles

subnet-04a96df116926a352 (emr-subnet-private1-us-east-1a) X

10.0.128.0/20 - us-east-1a - 4091 direcciones IP disponibles

i Puede ejecutar los comandos del cuaderno desde un espacio de trabajo de EMR Studio. Debe conectar el espacio de trabajo a un EMR existente en un clúster de EC2, a un EMR en un clúster virtual de EKS o a una aplicación de EMR sin servidor.

[Cancelar](#)

[Crear un Studio](#)

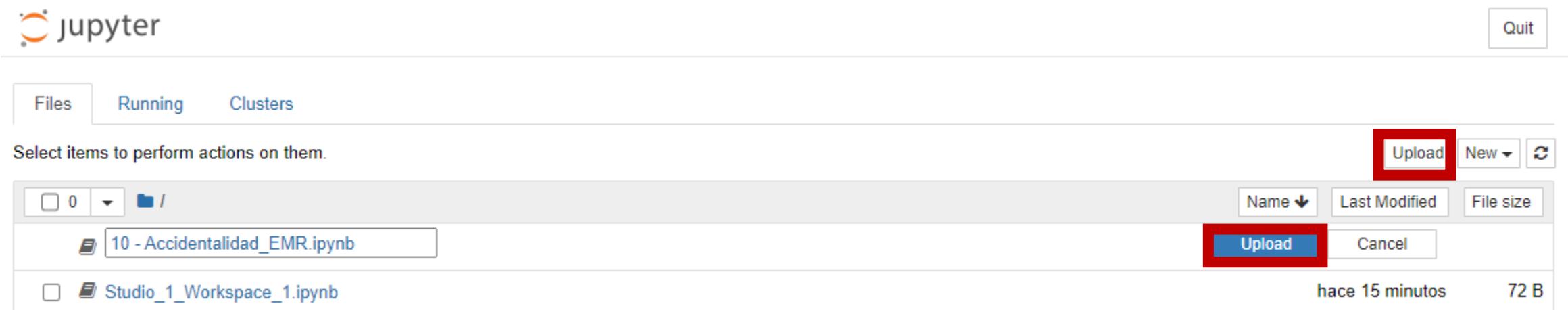
7

[Crear el estudio y lanzar el espacio de trabajo](#)

Crear EMR Studio y WorkSpace

9. Con el WorkSpace inactivo, haga clic en asociar Clúster
10. Seleccione la opción Lanzamiento en Jupyter
11. Seleccione el clúster actual en ejecución
12. Haga clic en asociar el clúster y lanzar

Finalmente cargue el notebook 10 - Accidentalidad_EMR.ipynb y ejecútelo



Crear EMR Studio y WorkSpace

Configuraciones de seguridad

EMR en EKS

Clústeres virtuales

EMR Studio

Introducción

Studios

WorkSpaces (Cuadernos)

WorkSpaces (Cuadernos) (1/1) Información

Acciones ▲

Asociar el clúster

Iniciar

Detener **8**

Eliminar

Ver los detalles

Nombre de

Studio_1_Workspace_1

Studio_1

Asociar el clúster

Lanzar en JupyterLab

Lanzamiento en Jupyter **10**

Clúster de EMR

Elija un clúster de EMR para adjuntarlo a su WorkSpace.

j-123HIU82TYYQ4 (cl2) **11**

WorkSpaces (Cuadernos) (1/1) Información

Acciones ▼

Asociar el clúster **9**

Lanzamiento rápido

Los Cuadernos de EMR ahora son EMR Studio workspaces. Puede organizar y ejecutar cuadernos interactivos en Workspaces.

Todos los Studios

Buscar espacios de trabajo por nombre, Studio, estado o

Nombre del WorkSpace

Studio_1_Workspace_1

Nombre del Studio

Studio_1

Estado

Inactivo

Grupo de seguridad del clúster

Elija el grupo de seguridad que se comunicará entre el WorkSpace y el clúster de Amazon EMR adjunto que se ejecuta en Amazon EC2.

sg-0f6decc5a1a15ad26 (DefaultEngineSecurityGroup)

Grupo de seguridad de WorkSpace

Elija el grupo de seguridad que permitirá al WorkSpace dirigir el tráfico a Internet y habilitar la vinculación de repositorios Git al WorkSpace.

sg-0dc836a7520a2d01c (DefaultWorkspaceSecurityGroupGit)

Cancelar

Asociar el clúster y lanzar **12**

Editar NoteBook

jupyter accidentalidad Last Checkpoint: hace 3 horas (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

In [1]:

```
import sys
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

Starting Spark application

ID YARN Application ID Kind State S

ID	YARN Application ID	Kind	State	S
0	application_1669825125359_0001	pyspark	idle	

SparkSession available as 'spark'.

In [3]:

```
spark = SparkSession\
    .builder\
    .appName("SparkETL")\
    .getOrCreate()
```

In [4]:

```
data = spark.read.option("inferSchema", "true").option("header", "true").csv("s3://test12345bucket/input/accidentalidad.csv")
```

▶ Spark Job Progress

In [7]:

```
data.take(3)
```

▶ Spark Job Progress

```
[Row(RADICADO=Decimal('1565221'), FECHA='01/01/2017', HORA='00:10:00', DIA='DOMINGO ', CLASE='Atropello', DIRECCION='CL 68 CR 87', TIPO_GEOCOD='EPM con Interior', GRAVEDAD='HERIDO', BARRIO='Palenque', COMUNA='Robledo', DISENO='Tramo de via'), Row(RADICADO=Decimal('1565189'), FECHA='01/01/2017', HORA='00:20:00', DIA='DOMINGO ', CLASE='Choque', DIRECCION='CL 44 CR 93', TIPO_GEOCOD='EPM sin Interior', GRAVEDAD='HERIDO', BARRIO='Campo Alegre', COMUNA='La América', DISENO='Tramo de via'), Row(RADICADO=Decimal('1565182'), FECHA='01/01/2017', HORA='00:20:00', DIA='DOMINGO ', CLASE='Choque', DIRECCION='CR 16 CL 56', TIPO_GEOCOD='Mal la vial', GRAVEDAD='HERIDO', BARRIO='Villatina', COMUNA='Villa Hermosa', DISENO='Tramo de via')]
```

```
import sys
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

```
spark = SparkSession\
    .builder\
    .appName("SparkETL")\
    .getOrCreate()
```

NoteBook - Accidentalidad

La secretaría de movilidad de la alcaldía de Medellín ha recolectado datos relacionados con la accidentalidad vial del año 2017 (Disponible en <https://www.datos.gov.co>)

A partir de los datos, obtenga la siguiente información

1. Porcentaje de accidentes en cada día de la semana
2. Barrio en el que se presenta el mayor y el menor número de accidentes
3. Comuna en la que se presenta mayor número de accidentes con muertos
4. Día de la semana en que se presentaron mayor cantidad de accidentes con muertos
5. Hora en la que se presentó mayor cantidad de accidentes con heridos

Machine Learning

- El Machine Learning o aprendizaje automático es una disciplina orientada a crear sistemas que puedan aprender por sí solos, con el fin de extraer información no trivial de grandes volúmenes de datos por medio de la identificación de patrones complejos.
- Spark implementa el aprendizaje automático a través del módulo MLlib que cuenta con un gran número de algoritmos que permiten crear modelos para el aprendizaje automático.
- Pueden identificarse dos grandes ramas en el aprendizaje automático, a saber, el aprendizaje supervisado y el aprendizaje NO supervisado.

MLlib Spark

MLlib types, algorithms and utilities

This lists functionality included in `spark.mllib`, the main MLlib API.

- [Data types](#)
- [Basic statistics](#)
 - summary statistics
 - correlations
 - stratified sampling
 - hypothesis testing
 - random data generation
- [Classification and regression](#)
 - linear models (SVMs, logistic regression, linear regression)
 - [naive Bayes](#)
 - [decision trees](#)
 - [ensembles of trees](#) (Random Forests and Gradient-Boosted Trees)
 - [isotonic regression](#)
- [Collaborative filtering](#)
 - alternating least squares (ALS)
- [Clustering](#)
 - k-means
 - Gaussian mixture
 - power iteration clustering (PIC)
 - latent Dirichlet allocation (LDA)
 - streaming k-means
- [Dimensionality reduction](#)
 - singular value decomposition (SVD)
 - principal component analysis (PCA)
- [Feature extraction and transformation](#)
- [Frequent pattern mining](#)
 - FP-growth
- [Optimization \(developer\)](#)
 - stochastic gradient descent
 - limited-memory BFGS (L-BFGS)
- [PMML model export](#)

Aprendizaje Supervisado



Predicción Discreta o Clasificación

Estudio de categorías pre-definidas para catalogar nuevos elementos.



Ejemplo: Predecir el comportamiento de pago de clientes en una entidad financiera:
BUENOS CLIENTES y MALOS CLIENTES.

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	CLASE
1	10	alto		56	Cliente Oro
2	45	bajo		54	Cliente Plata
3	23	medio		34	Cliente Bronce
4	54	alto		24	Cliente Bronce
5	21	medio		43	Cliente Oro
6	54	medio		23	Cliente Oro
7	74	alto		65	Cliente Bronce
8	46	alto		47	Cliente Plata
9	43	bajo		83	Cliente Plata
10	34	bajo		59	Cliente Bronce

Histórico o Conjunto de Entrenamiento



ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	CLASE
11	21	medio		43	?
12	54	medio		23	?
13	74	alto		65	?
14	46	alto		47	?
15	43	bajo		83	?
16	34	bajo		59	?

Datos futuros

Aprendizaje Supervisado



Predictión Continua o Regresión

Estudio de datos con el objetivo de predecir un evento numérico futuro.



Ejemplos: Estimar la expectativa de vida de un cliente.

- Predecir ventas futuras (series de tiempo)

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	PREDICCIÓN
1	10	alto		56	34
2	45	bajo		54	42
3	23	medio		34	15
4	54	alto		24	64
5	21	medio		43	36
6	54	medio		23	74
7	74	alto		65	34
8	46	alto		47	2
9	43	bajo		83	6
10	34	bajo		59	4

Histórico o Conjunto de Entrenamiento



Predicción de un número continuo

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	PREDICCIÓN
11	21	medio		43	?
12	54	medio		23	?
13	74	alto		65	?
14	46	alto		47	?
15	43	bajo		83	?
16	34	bajo		59	?

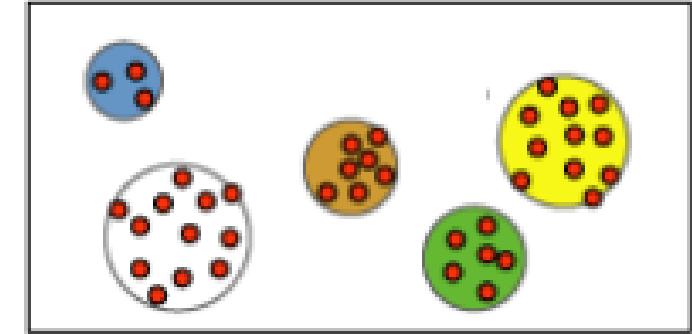
Datos futuros

Aprendizaje NO Supervisado



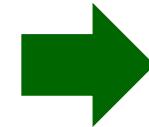
Agrupamiento / Clustering

Organizar una población de datos heterogénea en un número de clúster homogéneos.

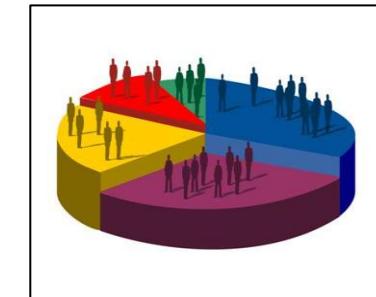


Ejemplos: Diseñar estrategias de mercadeo según el tipo de cliente. Detección de anomalías identificando datos que se alejen de los centroides de agrupación.

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27

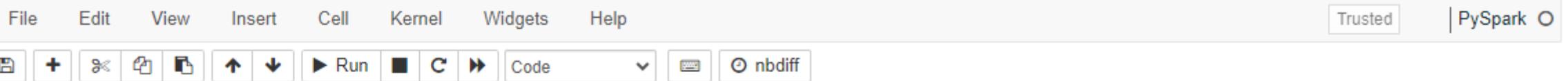


Descripción en grupos



Machine Learning

jupyter 11 - Machine Learning Last Checkpoint: hace una hora (autosaved)



```
In [ ]: import sys
from pyspark.sql import SparkSession
from pyspark.sql import functions
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.ml.stat import Correlation
from pyspark.ml.feature import VectorAssembler
```

```
In [ ]: spark = SparkSession\
    .builder\
    .appName("Spark_ML")\
    .getOrCreate()
```

Finalizar EMR

Los EMR tienen una protección contra la terminación que debe ser desactivada para poder finalizarlos

Terminar clústeres X

El clúster j-221S1NYGKIU9D (emr_ec2) tiene la protección contra la terminación activada. Para terminar este clúster, primero tiene que desactivar la protección contra la terminación.

Protección contra la terminación: Act. Cambiar

Los datos o trabajos pendientes que se encuentren en estos clústeres se perderán, como los datos almacenados en HDFS. Esta acción no se puede deshacer.

[Cancelar](#) [Finalizar](#)

Terminar clústeres X

El clúster j-221S1NYGKIU9D (emr_ec2) tiene la protección contra la terminación activada. Para terminar este clúster, primero tiene que desactivar la protección contra la terminación.

Protección contra la terminación: Act. Desactivado ✓ ✗

Los datos o trabajos pendientes que se encuentren en estos clústeres se perderán, como los datos almacenados en HDFS. Esta acción no se puede deshacer.

[Cancelar](#) [Finalizar](#)

AGENDA

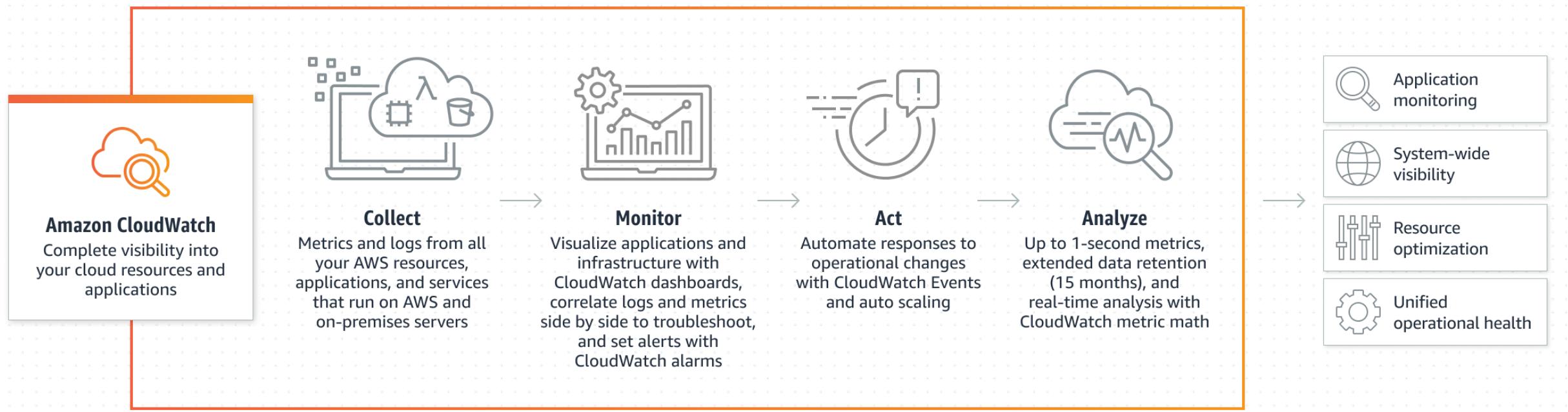
1. AWS
2. Nube Privada Virtual (VPC)
3. Procesamiento (EC2)
4. Almacenamiento (S3)
5. BigData (EMR)
- 6. Monitoreo (CloudWatch)**
7. Ejercicios

Cloud Watch

- Herramienta para monitorear recursos y aplicaciones de AWS
- Administra y visualiza registros, métricas y eventos en tiempo real
- Permite optimizar los recursos con el fin de reducir costos
- Establece acciones a ejecutar cuando se cumplan los límites definidos para los recursos y servicios de AWS

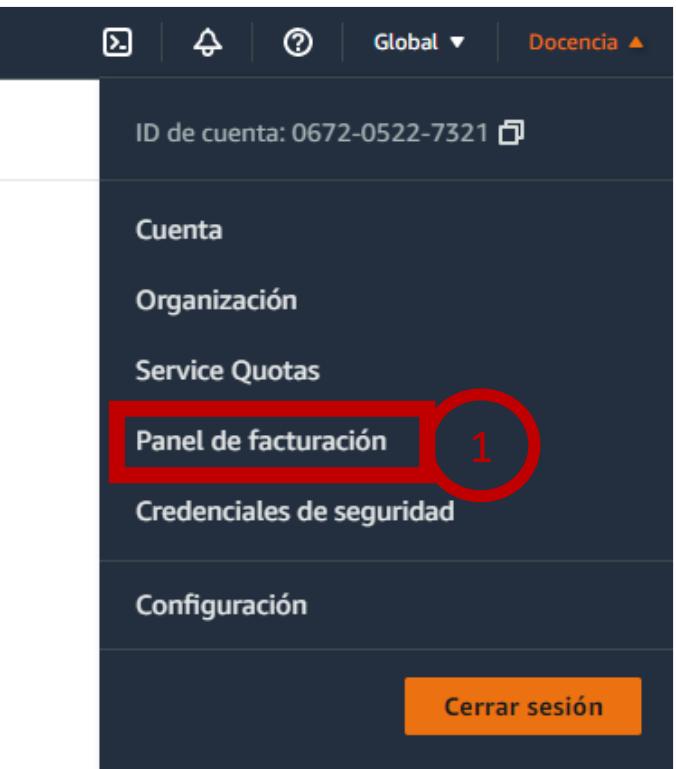


Cloud Watch



Activar Alertas Facturación

1. En la esquina superior derecha despliegue el menú y haga clic en panel de facturación
2. En el menú de Facturación seleccione la opción preferencias de facturación
3. Habilite las opciones: recibir alertas de uso de nivel gratuito y alertas de facturación
4. Asigne un correo electrónico
5. Haga clic en administre las alertas de facturación



Activar Alertas Facturación

Inicio

Facturación

Facturas

Pagos

Créditos

Órdenes de compra

Cost & usage reports

Cost categories

Etiquetas de asignación de costos

Free tier

Billing Conductor ▾

Cost Management

Cost explorer

Budgets

Budgets reports

Planes de ahorro ▾

Preferencias

Preferencias de facturación

Métodos de pago

Facturación unificada ▾

Configuración fiscal

Preferencias

▼ Preferencias de facturación

Reciba la factura en PDF por correo electrónico
Active esta característica para recibir una versión en PDF de la factura por correo electrónico. Las facturas suelen estar disponibles en los tres primeros días del mes.

Desactivar el crédito compartido
Cuando el crédito compartido está desactivado, los créditos solo se aplicarán a la cuenta del propietario y no se compartirán entre cuentas de la misma familia de facturación.
[Descargue el historial de preferencias de crédito compartido.](#)

► [Descuento compartido de RI y Savings Plans](#) ⓘ

▼ Preferencias de administración de costos

Reciba alertas de uso de nivel gratuito
Active esta característica para recibir alertas por correo electrónico cuando su uso del servicio de AWS se aproxime a los límites del uso de nivel Gratuito de AWS o los supere. Si desea recibir estas alertas en una dirección de correo electrónico que no sea la dirección principal asociada a esta cuenta, especifíquela a continuación.

Dirección de correo electrónico:

Reciba alertas de facturación
Active esta característica para supervisar automáticamente los cargos recurrentes y por uso de AWS. Esto facilita el seguimiento y la administración de los gastos en AWS. Puede configurar alertas para recibir notificaciones por correo electrónico cuando los cargos alcancen un límite específico. Una vez habilitada, esta preferencia no se puede desactivar.
[Administre las alertas de facturación](#) | [Compruebe la nueva característica de presupuestos](#)

5 **Informes de facturación detallados [heredados]**

Guarde las preferencias

3

4

5

Crear Alarma

1. Seleccione la opción Crear Alarmas
 2. Haga clic en el botón crear Alarma
 3. Haga clic en Seleccione la métrica
 4. Seleccione métricas de Facturación

Empezar a usar CloudWatch

1

No tiene alarmas, métricas ni paneles de interés configurados. Una vez que los configure, se mostrarán aquí. [Ir a la página de inicio](#)

 [Crear alarmas](#)

Configure alarmas en cualquiera de sus métricas para recibir una notificación cuando su métrica exceda el límite especificado.

 [Crear un panel predeterminado](#)

Cree y asigne un nombre a cualquier panel de CloudWatch **CloudWatch-Default** para mostrarlo aquí.

 [Ver los registros](#)

Lleve a cabo una monitoreo utilizando sus archivos de registro personalizados, de aplicación y de sistema existentes.

 [Ver los eventos](#)

Escriba reglas para indicar los eventos de interés para la aplicación y las acciones automatizadas que se deben desencadenar.

Especifique la métrica y las condiciones

Métrica

Gráfico

Vista previa de la métrica o de la expresión de la métrica, y límite de la alarma.

3

Selezione una métrica

Cancelar Siguiente

Crear Alarma

Seleccionar una métrica

Gráfico sin título

1
0.5
0

El gráfico de CloudWatch está vacío.
Seleccione algunas métricas para mostrarlas aquí.

13:45 14:00 14:15 14:30 14:45 15:00 15:15 15:30 15:45 16:00 16:15 16:30

Examinar Consulta Métricas diagramadas Opciones Origen Agregar matemática Agregar consulta

Métricas (94)

Search for any metric, dimension, resource id or account id

Gráfico con SQL Búsqueda en gráficos

▼ Espacios de nombres personalizados

AWSLicenseManager/licenseUsage 6

▼ Espacios de nombres de AWS

Facturación 2 Registras 2 S3 2 Uso 82

Cancelar Seleccionar una métrica individual para continuar

4

Crear Alarma

5. Seleccione la opción Cargo total estimado
6. Seleccione divisa en USD
7. Haga clic en Seleccionar una métrica

Métricas (2)

Todo > Facturación

Por servicio	1
Cargo total estimado	1

5

Métricas (1)

Todo > Facturación > Cargo total estimado Gráfico con SQL Búsqueda en gráficos

Nombre de métrica
Divisa (Currency) 1/1
USD

6

7

Crear Alarma

8. Seleccione un límite estático cuando sea mayor o igual a 10 USD y Haga clic en siguiente

Condiciones

Tipo de límite

Estático Utilice un valor como límite

Detección de anomalías Utilice una banda como límite

Cuando EstimatedCharges sea...

Defina la condición de la alarma.

Mayor > límite

Mayor/Igual >= límite

Menor/Igual <= límite

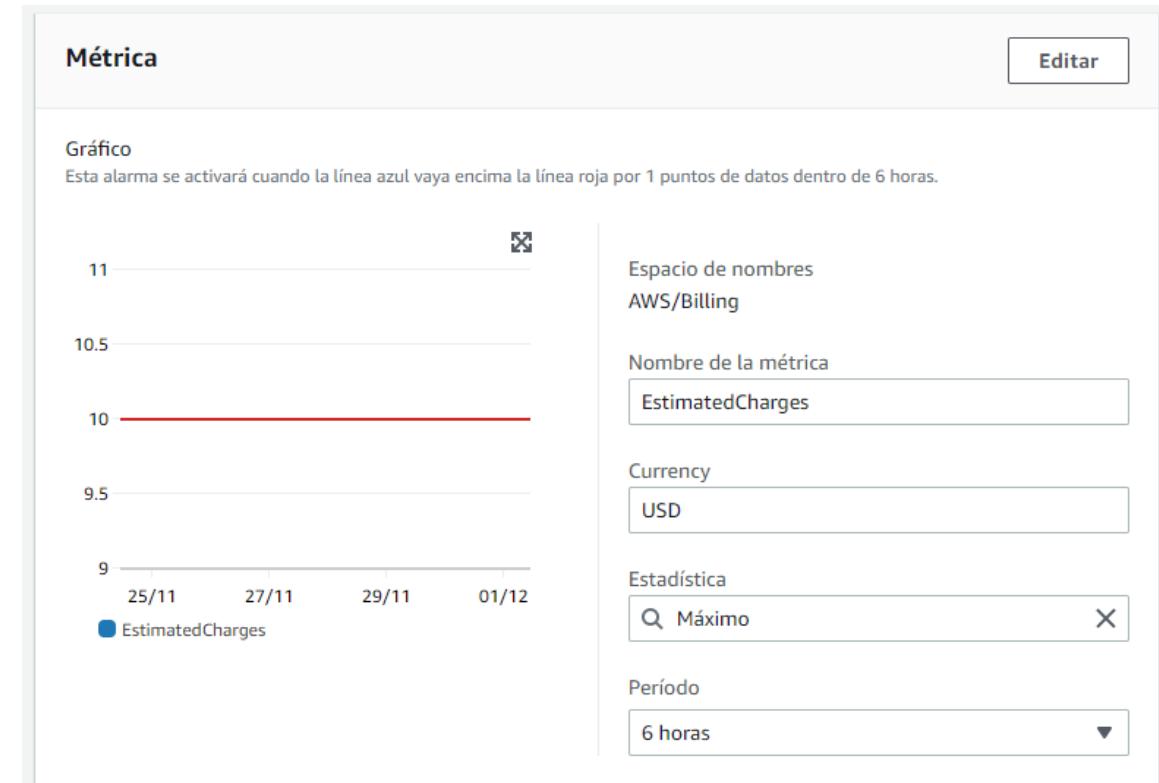
Menor < límite

que...

Defina el valor del límite

Debe ser un número

► Configuración adicional



Crear Alarma

9. Cree un nuevo tema para las notificaciones, asígnele un nombre y un correo electrónico y haga clic en crear un tema
10. Haga clic en siguiente

Notificación

Activador de estado de alarma
Definir el estado de alarma que activará esta acción.

En modo alarma
La métrica o expresión se encuentra fuera del límite definido.

CORRECTO
La métrica o expresión está dentro del límite definido.

Datos insuficientes
La alarma se acaba de iniciar o no hay suficientes datos disponibles.

Enviar una notificación al siguiente tema de SNS
Defina el tema de SNS (Simple Notification Service) que recibirá la notificación.

Seleccione un tema de SNS existente

Crear un tema nuevo

Usar ARN del tema para notificar a otras cuentas

Crear un nuevo tema...
El nombre del tema debe ser único.

Los nombres de los temas de SNS solo pueden contener caracteres alfanuméricos, guiones (-) y guiones bajos (_).

Puntos de enlace de correo electrónico que recibirán la notificación...
Añada una lista de direcciones de correo electrónico separadas por comas. Cada dirección se agregará como una suscripción al tema anterior.

usuario1@ejemplo.com, usuario2@ejemplo.com

Crear un tema 9

Agregar notificación

Acción de Auto Scaling

Acción de EC2
Esta acción solo está disponible para métricas por instancia de EC2.

Acción de Systems Manager [Información](#)

Esta acción creará un Incidente o un OpsItem en System Manager cuando la alarma esté en estado **En modo alarma**.

Agregar acción de Systems Manager

Siguiente 10

Cancelar **Anterior**

Crear Alarma

11. Asigne un nombre a la alarma y haga clic en siguiente
12. Haga clic en crear alarma

Agregar nombre y descripción

Nombre y descripción

Nombre de la alarma
PagoMensual

Descripción de la alarma - *opcional*
Descripción de la alarma
Hasta 1024 caracteres (0/1024)

[Cancelar](#) [Anterior](#) **Siguiente**

11

Paso 2: configurar las acciones [Editar](#)

Acciones

Notificación
Cuando En modo alarma, enviar una notificación a "CostoMensual"

Paso 3: agregar el nombre y la descripción [Editar](#)

Nombre y descripción

Nombre
PagoMensual

Descripción
-

[Cancelar](#) [Anterior](#) **Crear alarma**

12

Crear Alarma

Las acciones de la alarma creada permanecerán con una advertencia hasta que confirme la suscripción en el correo electrónico

The screenshot shows the AWS CloudWatch Metrics Alarms console with a single alarm listed:

Nombre	Estado	Última actualización del estado	Condiciones	Acciones
PagoMensual	Datos insuficientes	2022-12-01 11:57:18	EstimatedCharges >= 10 para 1 puntos de datos dentro de 6 horas	Acciones habilitadas Advertencia

A tooltip is displayed over the 'Acciones' button, stating: "Nuevo - Ahora puede desactivar las acciones de alarma compuesta asignando una alarma supresora." (New - Now you can deactivate composite alarm actions by assigning a suppressor alarm.)

Suscripción confirmada

The screenshot shows the AWS CloudWatch Metrics Alarms console with the same alarm listed, but now in a 'CORRECTO' state:

Nombre	Estado	Última actualización del estado	Condiciones	Acciones
PagoMensual	CORRECTO	2022-12-01 11:58:07	EstimatedCharges >= 10 para 1 puntos de datos dentro de 6 horas	Acciones habilitadas

AGENDA

1. AWS
2. Nube Privada Virtual (VPC)
3. Procesamiento (EC2)
4. Almacenamiento (S3)
5. BigData (EMR)
6. Monitoreo (CloudWatch)
7. Ejercicios

Salarios en Ciencia de Datos

Se dispone de un dataset que contiene información relacionada con los salarios para empleados del área de ciencia de datos, que incluyen las siguientes variables

- work_year: Año en el que se pagó el salario (2020-2023)
- experience_level: Nivel de experiencia
- employment_type: Tipo de contrato (Tiempo completo, parcial, etc)
- job_title: Rol del trabajador
- salary: Salario bruto pagado
- salary_currency: Moneda en la que se pagó el salario
- salaryinusd: Salario en dólares
- remote_ratio: Cantidad de trabajo realizado de forma remota
- company_ratio: País donde está ubicada la oficina principal de la compañía
- company_size: Tamaño de la compañía

Dataset disponible en <https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023>

Salarios en Ciencia de Datos

A partir del dataset entregado, resuelva las siguientes inquietudes:

1. Es cierto que los empleos que se pagan en euros son mejor remunerados que los pagados en dólares
2. Cual es el promedio de salario por tamaño de empresa
3. Son los empleos remotos los mejor remunerados
4. Top 10 de las profesiones con salarios mas altos y mas bajos
5. Países en los que se presenta el mayor y el menor salario

Salarios en Ciencia de Datos

Pasos a seguir:

- Cree un bucket de S3 donde suba el dataset de salarios en ciencia de datos (ds_salaries.csv)
- Cree un clúster de EMR
- Cree un notebook para ejecutarse en el clúster de EMR donde cargue los datos de S3 y resuelva las inquietudes planteadas
- Verifique el comportamiento de las métricas de Cloud Watch relacionadas con los componentes creados