**LABORATORIO BIDDATA**

### Temario del laboratorio

* Introducción a  Apache Spark Stream
* Apache Spark Streaming

**Recursos:**

* Link curso AWS: <https://awsacademy.instructure.com/courses/73687>
* Link readme: <https://labs.vocareum.com/web/2538519/939488.0/ASNLIB/public/docs/lang/es-es/README.html>
* EMR: <https://aws.amazon.com/es/emr/>

#### Introducción

#### En este laboratorio, se utilizar PySpark en un entorno de AWS para procesamiento de datos distribuidos. Utilizaremos el servicio de Amazon EMR (Elastic MapReduce) para crear el clúster de PySpark y procesar un conjunto de datos.

#### Para este laboratorio se proporciona un entorno de pruebas sobre AWS Academic, con Learner Lab, para ello debe aceptar la invitación al curso enviado a su correo electrónico, o ingresar al [link curso AWS](https://awsacademy.instructure.com/courses/46305). AWS Academy Learner Lab es una plataforma de aprendizaje en línea que brinda acceso a laboratorios prácticos y recursos de aprendizaje en la nube de AWS. Está diseñado para ayudar a los estudiantes a adquirir habilidades prácticas en tecnologías de la nube de AWS en un entorno de aprendizaje seguro y controlado. Si es la primera vez que ingresa a la plataforma, se recomienda familiarizarse con su entorno de trabajo, para ello es aconsejable leer el [link de readme](https://labs.vocareum.com/web/2538519/939488.0/ASNLIB/public/docs/lang/es-es/README.html).

#### Par el procesamiento de BigData, se utilizará Amazon EMR (Elastic MapReduce) que es un es un servicio de plataforma (PaaS) de AWS que permite a los usuarios procesar grandes cantidades de datos distribuidos utilizando frameworks de procesamiento distribuido, como Apache Hadoop y Apache Spark. EMR proporciona una plataforma de procesamiento de datos distribuida administrada en la nube, que permite a los usuarios crear, configurar y administrar clústeres de procesamiento de datos de manera fácil y rápida sin tener que preocuparse por la gestión de la infraestructura subyacente. Para más información puedes visitar el link de [AWS EMR](https://aws.amazon.com/es/emr/) y revisa las [características EMR](https://aws.amazon.com/es/emr/features/)

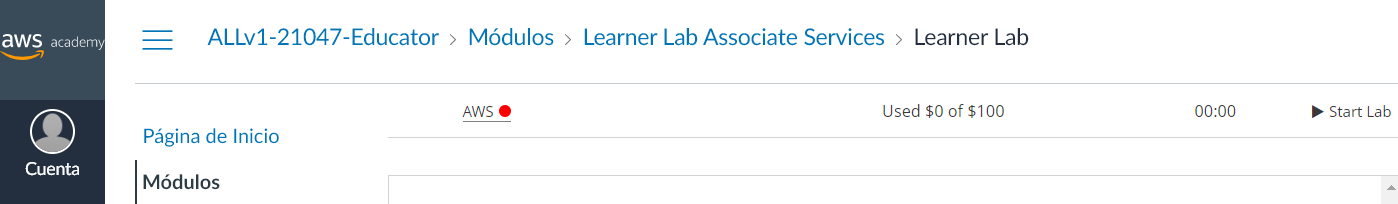
#### EMR, estará configurado en un clúster corriendo en Amazon Elastic Compute Cloud (Amazon EC2). EC2 es un servicio de infraestructura de AWS que permite a los usuarios lanzar y escalar servidores virtuales en la nube de AWS. Amazon EC2 proporciona una capacidad de cómputo escalable en la nube, lo que permite a los usuarios crear y ejecutar aplicaciones y servicios en línea en una infraestructura escalable y segura. Para más información ver la [página de EC2](https://aws.amazon.com/es/ec2/)

**Desarrollo**

**Configuración**

* Accede al panel de control de AWS y navega hasta la página de Amazon EMR.

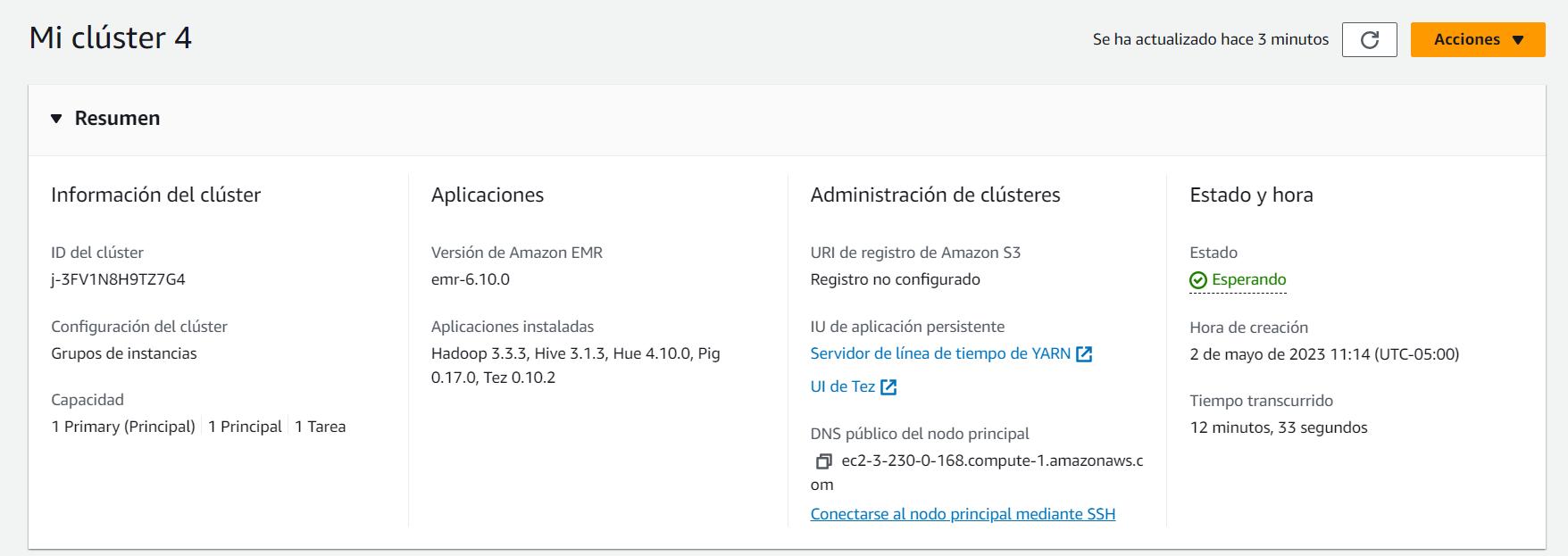
1) inicializar el lab (**clic en Start lab**)



2) Acceder a los servicios de AWS, clic en el recuadro de la imagen, o acceder a siguiente [link](https://us-east-1.console.aws.amazon.com/console/services?region=us-east-1)

3) Acceder al servicio de EMR desde la página de servicios, o clic en el [link](https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusters)

* Una vez en la página de gestión de EMR, Haz clic en el botón "Create cluster" (Crear clúster).
* En la sección de Configuración de software, selecciona la opción "Spark" como aplicación principal.
* En la sección configuración de clúster, configure un master y un nodo. Para cada servidor elije la instancia de EC2 ajusta las opciones según las necesidades. Por ejemplo, puedes seleccionar una instancia m4.large para el nodo master y dos instancias m4.xlarge para el nodo esclavo. Recuerde todo tiene un precio. revise en el valor en [link](https://aws.amazon.com/es/emr/pricing/)
* En la sección " Configuración y permisos de seguridad”, crea las claves de acceso para SSH. Asegúrate de almacenar la clave en un lugar seguro. Luego la usaremos. [sshKey](https://en.wikipedia.org/wiki/Ssh-keygen)
* Analice cada una de las opciones de configuración y comprenda y compare esto con lo visto en las clases teóricas.
* Es importante dejar que el clúster termine, un tiempo después de inactividad
* Haz clic en el botón "Create cluster" (Crear clúster) para crear el clúster.
* El clúster creado debe terne una imagen parecida a



**Carga de datos**

#### Para este laboratorio vamos a usar los dataset de la anterior evaluación. Para tal efecto debemos cargar los datos al clúster de Amazon EMR utilizando la herramienta de línea de comandos de AWS (AWS CLI) o la herramienta de transferencia de archivos SCP. Recomienda usar SCP con la clave anteriormente creada y luego conectarse con SSH para verificar. Como ejemplo doy la conexión a los servidores creados como ejemplo, recuerde debe cambiar la dirección IP y las Claves

#### ssh -i ~/test-key.pem [hadoop@ec2-3-230-0-168.compute-1.amazonaws.com](mailto:hadoop@ec2-3-230-0-168.compute-1.amazonaws.com)

#### Si no puede acceder revise la configuración de redes y seguridad y permita que se acceda desde el puerto 22.

#### Cargue y ejecute cada script

#### Para esta prueba se usa los script del examen 2 para ello guarda el archivo en el nodo principal del clúster de Amazon EMR, por medio de SCP.

#### Ejemplo Ejecuta el archivo 4\_ videogames.py utilizando el siguiente comando: spark-submit. Spark-submit es un comando en la línea de comandos utilizado para enviar una aplicación de Spark al clúster de Spark para su ejecución.

#### spark-submit 4\_videogames.py

#### Revise los resultados, y analice y cambie los procesos que considera pertinente