Sistemas Big data

Introducción

03/10/2023

/\* ------------------------------------------------------------------------------------------------- \*/

06/10/2023

El profesor especialista se llama Ricardo

Conceptos:

Pipeline: conjunto de procesos y herramientas. similar al concepto de ETL (Extract, Transform and Load). Su objetivo principal es analizar grandes volumenes de datos de manera eficiente. Está compuesto de los siguientes procesos:

1. Ingesta de Datos: los datos pueden obtenerse de diferentes fuentes (CSV, BD, HTTP, API) Capturar de manera eficiente los datos. Herramientas de captura como Kafka, SISS
2. Procesamiento de Datos: **El profesor prefiere hacer un preanálisis (validación de datos) como comprobar que un numero es un número, una fecha es una fecha.** Después viene el proceso de transformación y limpieza (seleccionar solamente aquellos datos necesarios).
3. Almacenamiento de Datos:
4. Procesamiento y Análisis Avanzado: Aquí puede entrar la Ia a través del Machine Learning, para hacer predicción de compras, etc…
5. Entrega de resultados: realizar reportes o análisis de las conclusiones sacadas de los datos

Dibujo en pizarra de lo que seria un pipeline (flujo de datos)

1. Datos (sources): CSV, DB, HTTP, Rest
2. Captura Kafka, Python, Pentaho, SISS
3. Preanálisis (validación de datos)
4. Transformación y limpieza: seleccionar solo lo que nos interesa
5. Machine Learning
6. Reporte Resultados

La línea entre Big Data y Bussniess Inteligence es fina. Solo se diferencia en el volumen de datos (a partir de 10 TB de datos)

Hay dos arquitecturas: funcional y técnica

/\*--------------------------------------------------------------------------------------------------------------------\*/

10/10/2023

AWS

AWS Academy para sistemas Big Data (los miércoles)

200 servicios de AWS

Esta cuenta no tiene coste,

2 cursos guiados. Múltiples “Learner Lab”

* Learner Lab es como un sandbox. Hasta 100 euros de límite
* Laboratorios guiados, cursos paso a paso muy controlados

Laboratorio guiado: AWS Academy Cloud Foundations. Estos laboratorios se pueden hacer tantas veces como quieras, no tiene límite de dinero, pero sí de tiempo. En cada laboratorio existe un entorno preparado (usuarios, servicios, etc…)

Learner lab: sale el dinero y el tiempo máximo (el laboratorio no se reinicia cuando termina el tiempo, pero hay que darle a “start lab” otra vez)

Para los laboratorios siempre hay que usar el mismo rol “lab-role”

El profesor puede tomar control de los laboratorios

17/10/2023

Primera toma contacto con AWS

Hoy vamos a ver infraestructura y seguridad

Infraestructura global de AWS

Existen diferentes regiones donde se pueden utilizar los servicios

Una región ees una zona geográfica:

* Hay replica de datos entre regioneas (controlada por el usuario)
* La comunicación entre redes va a través de la red interna de AWS

Cada región tiene niveles de redudnacnia y conectividad

Un región normalmente consta de dos o más zonas de disponibilidad

Se determina la región en:

* Gobernanza de datos y requisitos legales (cada región tiene sus propias leyes)
* Proximidad de los clientes (latencia)
* Servicios disponibles dentro de la región (no todas las regiones tiene los mismos servicios)
* Costes (varían según la región

Cada región tiene varias zonas de disponibilidad (replica de datos). Hay 69 zonas de disponibilad en todo el mundo. **AWS recomienda recplicar los datos en diversas zonas de disponibilidad**

**Los centros de datos de AWS están diseñados para una gran seguridad**

**Características**

* Elasticidad y escalabilaidad
* Tolerancia a errores
* Alta disponibilidad

Servicios básicos de AWS

* Informática virtual (máquinas, balanceo de carga)
* Redes
* Almacenamiento (objetivo, bloque, archivo)

Almacenamiento

Amazon Simple Storage Service (S3)

Glacier (como S3 pero archivos que se mueven muy poco, copias de seguridad)

Informática

EC2

AWS Lambda

EC2 AutoScaling

Bases de datos

RDS (servicio relacional)

Aurora

Redshilt (data warehouse)

Amazon DynamoDB (NoSQL)

Redes

VPC

Enrutado

VPN

Seguridad

Administración de costos

Informe de uso y costo de AWS

Administración y gobernanza

Consola de administración de AWS

Inferaz de línea de comandos de AWS (AWS CLI)

CloudWatch (Monitorización y alertas)

Seguridad en la nube de AWS

Modelo de responsabilidad compartida

* Sguriad de la infraestructura:
* Gestión de servicios compartidos: AWS es responsable de las redes, el HW y todo lo que sea interno

Responsabilididad del usuario

* Seguridad de la aplicación
* Gestión de datos
* Gesión de identidad y acceso

**AWS es responsable de la seguridad de la nube. El usuario es responsable de la seguridad dentro de la nube.**

Actividad

Actualización y parches en el sistema operativo de instancia EC2 -> usuario

Infraestructura de la virtualización AWS

Responsabilidad en materia de seguridad y características del servicio

* IaaS: Le das la máquina y haces lo que quieras con ellas (Ejemplo: EC2 y VPC)
* PaaS: Te dan máquinas con un Software ya preinstalado (RDS)
* SaaS: Se ofrece un Software que se accede desde una API o una web. Solamente

**El próximo día haremos un laboratorio guiado creando usuarios roles y demas**