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**

20/10/2023

Clase con Ricardo

Hadoop

Gestiona el flujo de datos de una pipeline. Importante para Big Data

Cloudera. Ejercicios de Hadoop, máquina virutal con Hadoop instalado, tutotiales de como levantar servicios, etc…

¿Qué es Big Data?

Capturar, almacenar y procesar grandes cantidades de datos en tiempo y costes. A partir de 30 terabytes de datos

¿Qué es Business Intelligent?

Permite a una empresa tomar decisiones en base a datos que han sido tratados

¿Qué es Hadoop?

Por muchos considerados la herramienta que revolucionó BigData. Nació a partir de GFS (Google File System), es OpenSource (Apache). Su objetivo es transformar HW de bajo coste en un servicio que permite gestionar BigData.

¿Qué es Cloudera?

Empresa tecnológica de BigData basado en Hadoop. Almacenamietno y procesamiento de grande volumen de datos

¿Qué ofrece Cloudera?

No es necesario tener todo el ecosistema de Hadoop, es modular y se puede utilizar solo lo que haga falta

Explora: Análisis de datos con Apache Impala (Data Warehouse). A través de SQL

Modelado:

Proorciona los datos: Hbase para aplicaciones on-line

Herramientas

* Cloudera Manaager: gestión clusters Big Data
* Hue: interfaz web para Hadoop.
* Apache Sqoop: traducir datos entre hadoop y una base de datos relacional
* Floom: Liempeza de datos
* Zookeper: Servicio centralizado para mantener conexiones entre servidores con disntintas tecnologías
* Hive: consulta sobre HDFS mediante HQL. Muy parecido a SQL
* Impala: motor de consultas SQL
* Oozie: Planificador de workflows
* Hbase: base de datos no relacional, creada sobre sistema de ficheros HDFS
* Apache Hadoo: incluye servicios HDFS, Yarn y Hadoop MapReduce. Yarn gestiona los procesos paralelizados. HDFS es el sistema de ficheros
* Spark: framework mas utilizado. Se puede utilizar en muchos lenguajes (Scala, Python, R, Java). Tiene herramientas de paralelización, Machine Learning…
* Kafka

¿Qué es Hadoop? ¿Cuál es su arquitectura?

Framework para procesar grandes conjuntos de datos. Diseño escalable. Se utiliza principalmente y su principal razón de existir es paralelizar el trabajo, o si no sería imposible trabajar con Big data. **Computación distribuida, Escalable, confiable**

Incluye:

* Hadoop Common:
* HDFS (Hadoop Distributed File System)
* Yarn
* Map Reduce: procesa grandes cojuntos de datos en paralelo. Para ello crea un entorno distribuido.

HDFS

Sistema de archivos distribuido de Hadoop. Es el componen de almacenamiento de Hadoop. Sus propiedad principales son:

* Grandes volúmenes de datos: capaz de almacenar petabytes
* Acesso a datos: se basa en la filosifia de
* Rentable: HDFS se puede ejecutar en un Hardware

Bloques de datos

Divide un archivo en unidades mas pequeñas de 128MB que se dividen aleatoriamente.

¿Por qué no varios bloques de 10KB? Porque el volumen de datos es demasiado grandes y quedarían demasiados “paquetitos”. Ventajas de almacenar en bloques:

* Si un archivo es demasiado grande puede no caber en un disco
* Evitar el estrangulamiento de una máquina

HDFS: Manejo básico

Comandos parecidos a los de Linux

27/10/2023

Hadoop.

Mapreduce

Framework para crear aplicaciones que procesan grandes datos en paralelo, utilizando grandes grupos de hardware básico. Su trabajo consiste en

* Input: una tabla, un texto. Los datos se fragmentan y son procesados por las tareas mamp
* IDkkkk

Los nodos de computación y almacenamiento son usualmente los mismos nodos. Este permite disponer de un ancho de banda muy grande.

“Si tienes un bloque de información de 1TB y haces un select, pagas por todo el espacio, aunque tengas un where. Hay que tener particiones para así pagar solo por esa partición

Un programa MapReduce se divide en tres componentes principales:

* Driver: punto de entrada y parametrización
* Mapper: lee, filtra y transforma el dato para la clave valor
* Reducer:
  + Shuffle: lee los mapeos y los convierte en la entrada de los reducers
  + Sort: los pares clave-valor se ordenan por clave valor
  + Reducer

Cuantos maps

Las buenas ptracticas sugieren utilizar enter 10 y 100 maps por nodo. Se puede llegar a 300 maps si las tareas son muy ligeras de CPU

Cuantos reducers

El numero de reducers esta entre 0.95 1.75 multiplicacdo por (numero de nodos \* numero de contenedores máximos por nodo)

No utilizar el 100% de la memoria para poder actuar en caso de fallo.

Limitaciones:

* Si se generan muchas claves en la fase de map, el sort puede ir muy lento
* Para volúmenes pequeños no interesa hacer un paralelismo. Esta herramienta solo tiene sentido con mucha capacidad.
* No es una buena herramienta para grafos
* Saturación de red en la parte de shuffle si la entrada y el mapeo es muy grande
* Dificultad a la hora de implementar (por separado)
* Se tiene que procesar datos por batch. No se puede procesar en tiempo real (streaming)

Formatos de ficheros en Hadoop

Formatos

* Textfile (formato predeterminado)
* Sequence File
* Parquet (parecido a CSV)
* Avro
* Orc

Hive