Introducción © EDICIONES ROBLE, S.L.

Indice

Introducción

Introducción. Alberto Oikawa

Resulta un hecho evidente que, en nuestros días, la mayor parte de la información que mueve el mundo desarrollado —la que se emplea para la toma de decisiones, la que permite la evolución científica, el desarrollo económico, el conocimiento de nuestra especie y de muchas otras e incluso la que determina nuestras relaciones sociales— es, eminentemente, digital. Es posible que, en origen, esa información no provenga de una fuente digital, pero también es innegable que, si se desea estudiar, es decir, convertir datos en información, en cualquiera de los ámbitos en que se trabaje, es necesario digitalizarla y almacenarla. Después, la curiosidad de cada uno hace el resto.

Debido a la gran cantidad y variabilidad de las fuentes de datos digitales (o susceptibles de serlo) de las que actualmente disponemos, se presentan, de forma directamente proporcional, problemas del tipo: cómo almacenar esos datos en un mundo de recursos finitos, cómo leer/escribir esos datos en un tiempo también finito (y cada vez más valioso) y cómo obtener información a partir de ellos, al ser tantos y tan dispares. Esto es, básicamente, el Big Data.

Aunque los problemas que se nos presentan no son en su totalidad resolubles, como es habitual en el mundo real, sí hemos sido capaces de “esquivarlos” o de “minimizar” su influencia en el natural desarrollo de la teoría de la información. En este módulo, no se tratarán problemas que requieran una gran abstracción (aunque esta será la que lo resuelva todo); se tratarán problemas físicos, problemas de capacidad, problemas de manipulación, los orígenes de los mismos, herramientas de las que disponemos para tratarlos y bases necesarias para su futuro desarrollo.

Así pues, mediante la lectura de la documentación y de los ejercicios propuestos, el alumno podrá adquirir los conocimientos necesarios para afrontar problemas de tratamiento de grandes volúmenes de datos de naturaleza dispar, descubrir algunos nuevos e, incluso, solucionarlos.

Objetivos

Tras el estudio de este módulo, los alumnos habrán alcanzado los siguientes objetivos:

1. Distinguir entre procesamiento *batch* y *streaming.*
2. Comprender los conceptos principales sobre ambos tipos de procesamiento.
3. Familiarizarse con las tecnologías Big Data que los posibilitan, entre otras:
    1. Hadoop.
    2. Spark.
    3. Kafka.
4. Adquirir conocimientos sobre arquitecturas de procesamiento paralelo.
5. Adquirir conocimientos sobre arquitectura *cloud* (IaaS): AWS, etc.

IMPORTANTE: Aquí os dejamos algunas consideraciones que pueden resultaros de ayuda a la hora de afrontar los contenidos del módulo.

Consejos sobre las máquinas virtuales

Importación y funcionamiento:

- **Pasos principales** y algunos **errores** frecuentes:
  - 1.- Lo primero que debéis hacer es descargar las máquinas.
  - 2.- Para que funcionen las máquinas virtuales, es necesario que el PC donde las vayamos a montar permita la "Virtualización Hardware". Esta es una opción que suele venir en la BIOS del equipo. Puede llamarse “Virtualización Hardware" o similar en función del modelo del PC.
  - 3.- Para confirmar que vuestro PC soporta virtualización podéis usar:https://www.grc.com/securable.htm
  - 4.- Instalar VirtualBox desde https://www.virtualbox.org/wiki/Downloads
  - 5.- Importar el archivo .ova descargado dentro de Virtualbox (menú Archivo/Importar Servicio Virtualizado).
  - 6.- Si la importación os da un error estilo ERROR_ZIP_CODE muy posiblemente el archivo .ova esté corrupto. Para solucionarlo debéis descargar de nuevo las máquinas virtuales.
  - 7.- Una vez importada la máquina, procedemos a encenderla. Si no encendiera bien o diera algún error, revisar el punto 2.
  - 8.- Es importante que confirméis que vuestro sistema operativo es de 64.
- Vídeo sobre la importación de las MV:https://s3-eu-west-1.amazonaws.com/master-eoi/Importaci%C3%B3n+MV+Virtualbox.zip

El **usuario** y la **contraseña** de la máquina virtual son "**bigdata**" y "**bigdata**".

La máquina virtual tiene 3 GB de memoria. Si veis que rendimiento es malo y disponéis de memoria, podéis ampliarla. Es necesario detener la máquina virtual, no vale con guardar su estado. Posteriormente, tendréis que volver a lanzar los demonios. Para cambiar los valores tenéis que entrar en la opción de menú Máquina/Configuración y cambiarlo en Sistema.

Para evitar problemas de corrupción de HDFS, al cerrar la MV seleccionad: "Guardar el estado de la máquina".

Otros consejos importantes

Copiar-pegar comandos

En ocasiones, se ha observado que copiar y pegar comandos desde la documentación en pdf u otros formatos da muchos problemas. El motivo es que al copiar desde el pdf y pegar en la máquina virtual, cambia algunos caracteres por otros parecidos, pero no iguales. Algunos casos son "-", ";", """, etc. Sed cuidadosos en este punto.

Sistema de archivos corrupto

Otro **error muy común** es que se nos **corrompa el sistema de archivos**. Un **síntoma** de esto es que **no** arranque **DataNode** o **NameNode**.

- El síntoma es que no arrancan los todos los demonios de HDFS. Para solucionarlo, debes:

1.- Parar HDFS y YARN

stop-dfs.sh

stop-yarn.sh

2.- Borrar el contenido de /hadoop_store/hdfs/datanode y de /hadoop_store/hdfs/namenode con:

sudo rm -R /home/bigdata/hadoop_store/hdfs/datanode/*

sudo rm -R /home/bigdata/hadoop_store/hdfs/namenode/*

3.- Formatear de nuevo el sistema de archivos:

hdfs namenode -format

4.- Arrancar los demonios y revisar con jps:

start-dfs.sh

start-yarn.sh

A la hora de ejecutar los comandos...

A la hora de ejecutar los comandos, es preciso observar la ruta desde la que se ejecutan en las imágenes que los acompañan en la documentación.

Máquina virtual

En esta unidad, utilizaremos máquinas virtuales para familiarizarnos con el entorno. A este respecto dejamos a vuestra disposición dos máquinas virtuales:

OPCIÓN 1: máquina sin configurar

Se puede descargar la máquina virtual en el enlace que aparece a continuación: máquina sin configurar.

OPCIÓN 2: máquina configurada

Se puede descargar la máquina virtual en el enlace que aparece a continuación: máquina configurada.

Nota:

La versión de Ubuntu de estas máquinas es Ubuntu 17.10. Las máquinas virtuales tienen descargados los paquetes de *software* de las distintas herramientas que se van a estudiar en este módulo. Se recomienda no cambiar estas versiones, dado que existen limitaciones de compatibilidad en algún ejemplo. Por lo demás, el funcionamiento de versiones posteriores es muy similar.

Es preciso señalar que la velocidad a la que salen las distintas versiones hace difícil saber cuáles serán las versiones disponibles en el momento de realizar la formación.

En cuanto a Hadoop, la versión con la que se va a trabajar en este módulo es la 2.8.0. En este caso, Hadoop 2.8.0 es una versión estable y ampliamente implantada en el mundo profesional.

Recursos

Documentos

Máquina sin configurar
Maquina_virtual_sin_config.pdf

Máquina configurada
Maquina_virtual_configurada.pdf

Enlaces de Interés

Vídeo sobre la importación de las MV.
https://s3-eu-west-1.amazonaws.com/master-eoi/Importaci%C3%B3n%20MV%20Virtualbox.zip