# Infraestructuras tecnológicas para *big data*

PID_00288154

Remo Suppi Boldrito

Tiempo mínimo de dedicación recomendado: 1 hora

Universitat Oberta de Catalunya

**Remo Suppi Boldrito**

Ingeniero de Telecomunicaciones. Doctor en Informática por la Universitat Autònoma de Barceloa (UAB). Profesor del Departamento de Arquitectura de Computadores y Sistemas Operativos en la UAB.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Josep Jorba Esteve

Primera edición: febrero 2022

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Remo Suppi Boldrito
Producción: FUOC

# Índice

# 1. Introducción

En esta asignatura veremos los conceptos fundamentales de las infraestructuras tecnológicas para el *big data*, cómo se clasifican, su despliegue y su análisis, complementado con un conjunto de casos de uso que permitirán al estudiantado analizar y experimentar con las infraestructuras reales que luego utilizará en un entorno profesional.

La asignatura está organizada en cinco módulos y cuatro retos (PEC). En el módulo «Introducción a las infraestructuras para *big data*» se realizará una **introducción conceptual a las infraestructuras para *big data***, donde se analizarán el concepto de *incremento de las prestaciones* y el de *big data*, así como los aspectos de organización y los modelos de procesamiento (clúster, nube –*cloud*–, IoT/sensores, *edge*). Este módulo también incluirá una incursión en la interacción y el procesamiento para describir los aspectos fundamentales de las API y los entornos funcionales.

El módulo «Virtualización e hipervisores» se centrará en el **concepto de *virtualización* y los hipervisores**; se hará un repaso de diferentes hipervisores (KVM, VirtualBox, HyperV, Proxmox, VMware) y los contenedores (Docker y LXC).

Los **sistemas de almacenamiento distribuido y las redes** serán tratados en el módulo «Sistema de almacenamiento distribuido y redes», especialmente centrado en las redes de almacenamiento (NAS, SAN, GlusterFS, HDFS) y las redes definidas por software SDN (OpenFlow, OSwitch, LinuxBridge).

Las **arquitecturas de software para *big data*** describen las características de las infraestructuras de software para el procesamiento de datos masivos; concretamente se centran en dos de las grandes herramientas *open-source* para el procesamiento de datos masivos como son Hadoop y Spark, y se analizan diferentes distribuciones integradas –que incluyen además otras herramientas del ecosistema Hadoop– como son BigTop y BigData Europe. Además, se explican los cambios de licencia (a restringida) de distribuciones a las que hasta el 2021 se podía acceder solamente con un registro, como son Cloudera-Hortonworks (muy ultizadas anteriormente en entornos *open-source* y académicos). Estos contenidos conforman el módulo «Arquitecturas de software para *big data*».

Finalmente, el módulo «Infraestructura como código (IaC). Monitorización y seguridad» presentará la **infraestructura como código (IaC)** y su despliegue en Terraform y Ansible. También se tratarán aspectos relacionados con la monitorización y la seguridad.

Los **retos** (PEC) estarán orientados a que el estudiantado afiance los conceptos tratados en cada apartado y experimente sobre las infraestructuras que luego encontrará en el entorno laboral. Los contenidos de cada reto se muestran a continuación:

**Reto 1. Despliegue de un servicio *proxy balanced* sobre un clúster virtualizado**

- **Objetivo**: el estudiantado deberá construir un clúster virtualizado de 4 máquinas virtuales (MV) en un esquema Master-Worker sobre OpenNebula, y desplegar un servicio de proxy que balancee la carga entre los nodos trabajadores (*workers*).

- Estudio/revisión de los módulos «Introducción a las infraestructuras para *big data*» y «Virtualización e hipervisores» (guía, bibliografía, recursos externos).

**Reto 2. Despliegue de un servicio de archivos distribuido**

- **Objetivo**: el estudiantado deberá construir un clúster virtualizado de 5 MV para crear un sistema de archivos distribuido que sea tolerante a fallos, y probar su rendimiento desde una 6ta MV. También se deberán hacer pruebas de fiabilidad y analizar el tiempo de respuesta en la lectura/escritura de grandes archivos.

- Estudio/revisión del módulo «Sistema de almacenamiento distribuido y redes» (guía, bibliografía, recursos externos).

**Reto 3. Despliegue de un clúster de *big data* con herramientas adecuadas para el procesamiento de los datos estáticos y en flujo de datos (*streams*)**

- **Objetivos**: el estudiantado deberá desplegar un clúster de *big data* y realizar pruebas de concepto para medir prestaciones y escalabilidad entre 2, 4, 6 nodos sobre conjuntos de datos superiores a 500 Mb.

- Estudio/revisión del módulo «Arquitecturas de software para *big data*» (guía, bibliografía, recursos externos).

**Reto 4. Despliegue de una infraestructura sobre un proveedor público utilizando dos herramientas que usen metodologías/procedimeintos diferentes**

- **Objetivo**: el estudiantado deberá utilizar diferentes herramientas IaC para desplegar un clúster sobre máquinas virtuales y sobre contenedores de un servicio *proxy balanced* sobre un proveedor público, y analizar el rendimiento con un test web (*benchmark*).

- Estudio/revisión del módulo «Infraestructura como código (IaC). Monitorización y seguridad» (guía, bibliografía, recursos externos).