Infraestructures tecnològiques per a *big data*

PID_00288153

Remo Suppi Boldrito

Temps mínim de dedicació recomanat: 1 hora

Universitat Oberta de Catalunya

Remo Suppi Boldrito

Enginyer de Telecomunicacions. Doctor en Informàtica per la Universitat Autònoma de Barcelona (UAB). Professor del Departament d'Arquitectura de Computadors i Sistemes Operatius de la UAB.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Josep Jorba Esteve

Primera edició: febrer 2022

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Remo Suppi Boldrito
Producció: FUOC

Índex

1. Introducció

En aquesta assignatura veurem els conceptes fonamentals de les infraestructures tecnològiques per al *big data* (dades massives), com es classifiquen, el seu desplegament i la seva anàlisi, complementats amb un conjunt de casos d'ús que permetran a l'estudiantat analitzar i experimentar amb les infraestructures reals que després utilitzarà en un entorn professional.

L'assignatura s'organitza en cinc mòduls i quatre reptes (PAC). En el mòdul «Introducció a les infraestructures per a *big data*» es farà una **introducció conceptual a les infraestructures per a *big data***, en la qual s'analitzarà el concepte d'*increment de les prestacions* i el de *big data*, així com els aspectes d'organització i els models de processament (clúster, núvol –*cloud*–, IoT/sensors, *edge*). Aquest mòdul també farà una incursió en la interacció i en el processament per descriure els aspectes fonamentals de les API i dels entorns funcionals.

El mòdul «Virtualització i hipervisors» se centrarà en el **concepte de *virtualització* i en els hipervisors**; s'hi durà a terme un repàs de diferents hipervisors (KVM, VirtualBox, HyperV, Proxmox, VMware) i dels contenidors (Docker i LXC).

Els **sistemes d'emmagatzematge distribuït i les xarxes** es tractaran en el mòdul «Sistema d'emmagatzematge distribuït i xarxes», especialment centrat en les xarxes d'emmagatzematge (NAS, SAN, GlusterFS, HDFS) i en les xarxes definides per programari SDN (OpenFlow, OSwitch, LinuxBridge).

Les **arquitectures de programari per a *big data*** descriuen les característiques de les infraestructures de programari per al processament de dades massives. Concretament, se centren en dues de les grans eines *open-source* per al processament de dades massives, com són Hadoop i Spark, i s'analitzen diferents distribucions integrades –que inclouen, a més, altres eines de l'ecosistema Hadoop– com BigTop i BigData Europe. També s'expliquen els canvis de llicència (a restringida) de distribucions a les quals es podia accedir fins al 2021 només amb un registre, com són les de Cloudera-Hortonworks (molt utilitzades anteriorment en entorns *open-source* i acadèmics). Aquests continguts conformen el mòdul «Arquitectures de programari per a *big data*».

Finalment, el mòdul «Infraestructura com a codi (IaC). Monitoratge i seguretat» presentarà la **infraestructura com a codi (IaC)** i el seu desplegament en Terraform i Ansible. També es tractaran aspectes relacionats amb el monitoratge i la seguretat.

Els **reptes** (PAC) estaran orientats a aconseguir que l'estudiantat consolidi els conceptes tractats en cada apartat i experimenti sobre les infraestructures que després trobarà en l'entorn laboral. Els continguts de cada repte es mostren a continuació:

Repte 1. Desplegament d'un servei *proxy balanced* sobre un clúster virtualitzat

- **Objectiu**: l'estudiantat haurà de construir un clúster virtualitzat de quatre màquines virtuals (MV) en un esquema Master-Worker sobre OpenNebula i desplegar un servei de *proxy* que balancegi la càrrega entre els nodes treballadors (*workers*).

- Estudi i revisió dels mòduls «Introducció a les infraestructures per a *big data*» i «Virtualització i hipervisors» (guia, bibliografia, recursos externs).

Repte 2. Desplegament d'un servei d'arxius distribuït

- **Objectiu**: l'estudiantat haurà de construir un clúster virtualitzat de 5 MV per crear un sistema d'arxius distribuït que toleri errades, i provar-ne el rendiment des d'una sisena MV. També caldrà fer proves de fiabilitat i analitzar el temps de resposta en la lectura/escriptura de grans arxius.

- Estudi i revisió del mòdul «Sistema d'emmagatzematge distribuït i xarxes» (guia, bibliografia, recursos externs).

Repte 3. Desplegament d'un clúster de *big data* amb eines adequades per al processament de les dades estàtiques i en flux de dades (*streams*)

- **Objectius**: l'estudiantat haurà de desplegar un clúster de *big data* i fer proves de concepte per mesurar prestacions i escalabilitat entre 2, 4, 6 nodes sobre conjunts de dades superiors a 500 Mb.

- Estudi i revisió del mòdul «Arquitectures de programari per a *big data*» (guia, bibliografia, recursos externs).

Repte 4. Desplegament d'una infraestructura sobre un proveïdor públic utilitzant dues eines que emprin metodologies i procediments diferents

- **Objectiu**: l'estudiantat haurà d'utilitzar diferents eines IaC per desplegar un clúster sobre màquines virtuals i contenidors d'un servei *proxy balanced* sobre un proveïdor públic i analitzar-ne el rendiment amb un test web (*benchmark*).

- Estudi i revisió del mòdul «Infraestructura com a codi (IaC). Monitoratge i seguretat» (guia, bibliografia, recursos externs).