EMMAGATZEMATGE I GESTIÓ DE DADES

En aquest article, explicarem quines maneres tenim d'emmagatzemar la informació en un clúster de *big data*, la importància i el repte que suposa realitzar una bona elecció per a què tant el funcionament com el rendiment siguin els esperats.

En els últims 15 anys, han anat emergint molts tipus de bases de dades i sistemes d'emmagatzematge, tots ells vàlids en funció de necessitats concretes del nostre sistema. Per això, és necessari que sapiguem diferenciar cadascuna de les alternatives que es presenten a continuació:

Sistemes de fitxers distribuïts

- HDFS. És l'elecció principal quan necessitem emmagatzemar i processar fitxers en un clúster de forma distribuïda. Pot desplegar-se tant *on premise* com en plataformes de *cloud* i un dels seus principals atractius és el conjunt d'eines que s'han construït sobre aquest sistema de fitxers. Sobre la mateixa arquitectura, es pot establir la capa de transformació amb operacions MapReduce, utilitzant alguna eina de *data warehouse* com Hive, Impala o HBase o un *framework* de processament com Apatxe Spark.
- Malgrat que l'elecció més utilitzada pot ser HDFS, en l'actualitat, estan cobrant molta importància els sistemes d'emmagatzematge en el núvol (*cloud*), els quals proporcionen serveis PaaS (Platform as a Service) que, de manera ràpida i amb una configuració mínima, ens permeten començar a fer tasques d'adquisició de dades per a la nostra aplicació. Cada proveïdor disposa del seu sistema de fitxers i ha creat un *marketplace* de serveis compatibles, de manera que puguem treballar els fitxers emmagatzemats amb les eines que ens habiliten cadascun d'ells. Com a principals alternatives, podem trobar el servei S3 a AWS, Cloud Storage a Google Cloud i Azure Storage.

Bases de dades relacionals

Són útils a l'hora de realitzar operacions transaccionals sobre la informació emmagatzemada. Entre les principals alternatives destaquen:

- Hive: és una eina de *data warehouse* construïda sobre HDFS, que proporciona una interfície a l'usuari o usuària per poder llançar consultes i anàlisis sobre un esquema relacional. Implementa operacions de MapReduce sobre el clúster i el seu objectiu és el de treballar amb grans *datasets*. Malgrat penalitzar una mica en el temps d'execució i recursos reservats, és capaç de realitzar operacions molt costoses sobre grans volums d'informació.
- Impala: és una eina molt similar a Hive, que també funciona sobre HDFS. En aquest cas, permet als usuaris i usuàries executar consultes SQL amb una latència molt baixa, gràcies a la gestió d'una metadada interna i d'un processament paral·lel. S'utilitza per fer tasques descriptives i analítiques sobre la informació d'HDFS.
- És interessant conèixer quines alternatives ens ofereixen les plataformes en el núvol. En aquest cas, destaquem RDS a AWS, Cloud SQL a Google Cloud i Azure SQL Database.

Bases de dades no relacionals

Són la millor elecció, quan la informació emmagatzemada no segueix un esquema fix i es volen potenciar característiques d'escalabilitat, replicació i tolerància a fallades en el nostre sistema. Entre les principals opcions trobem:

- HBase. Treballa sobre el sistema de fitxers HDFS. És una base de dades que prioritza la compressió de la informació, és ràpida en operar en memòria i amb una estructura de clau/valor. És adequada per trobar un bon compromís entre operacions de lectura i escriptura sobre grans conjunts de dades amb un *throughput* (quantitat d'esdeveniments que es processen per unitat de temps) i latència reduïts.
- MongoDB. És una base de dades no relacional independent, és a dir que, a diferència d'Hbase, no funciona sobre cap sistema d'emmagatzematge propi com HDFS. És una elecció ideal quan volem emmagatzemar informació documental no estructurada, fàcilment indexable i amb la possibilitat de replicació de les dades, o quan volem realitzar operacions de balanceig de càrrega, emmagatzematge de fitxers i realitzar transformacions i agregacions de les dades emmagatzemades fàcilment.
- ElasticSearch. És un motor de cerca indexat molt potent a l'hora de realitzar cerques escalables de text, emmagatzematge de sèries temporals o l'adquisició de *logs* o diverses col·leccions de text. ElasticSearch és una eina inclosa dins d'un *stack* anomenat ELK, que proporciona, entre d'altres, una eina d'adquisició de dades (Logstash), una capa d'emmagatzematge (ElasticSearch) i una de visualització (Kibana).
- Neo4j. És una base de dades de grafs, que destaca per la seva utilitat a l'hora de trobar relacions entre les entitats de les nostres dades i extreure, d'una manera senzilla i gràfica, el màxim valor a la informació emmagatzemada.

Totes aquestes alternatives ens ajudaran a dissenyar l'arquitectura de *big data* més adequada per emmagatzemar i processar la nostra informació.

Recordem: és essencial saber per endavant com venen estructurades les dades, què volem fer amb elles (transformacions i anàlisis) i com ho farem (de manera distribuïda, escalable, indexada, mitjançant claus, etc.). Una vegada realitzat aquesta petita anàlisi prèvia, ja només és qüestió de desplegar la solució que més s'adeqüi a les nostres necessitats.

Descobreix tot el que Barcelona Activa pot fer per a tu

Acompanyament durant tot el procés de recerca de feina

barcelonactiva.cat/treball

Suport per posar en marxa la teva idea de negoci

barcelonactiva.cat/emprenedoria

Serveis a les empreses i iniciatives socioempresarials

barcelonactiva.cat/empreses

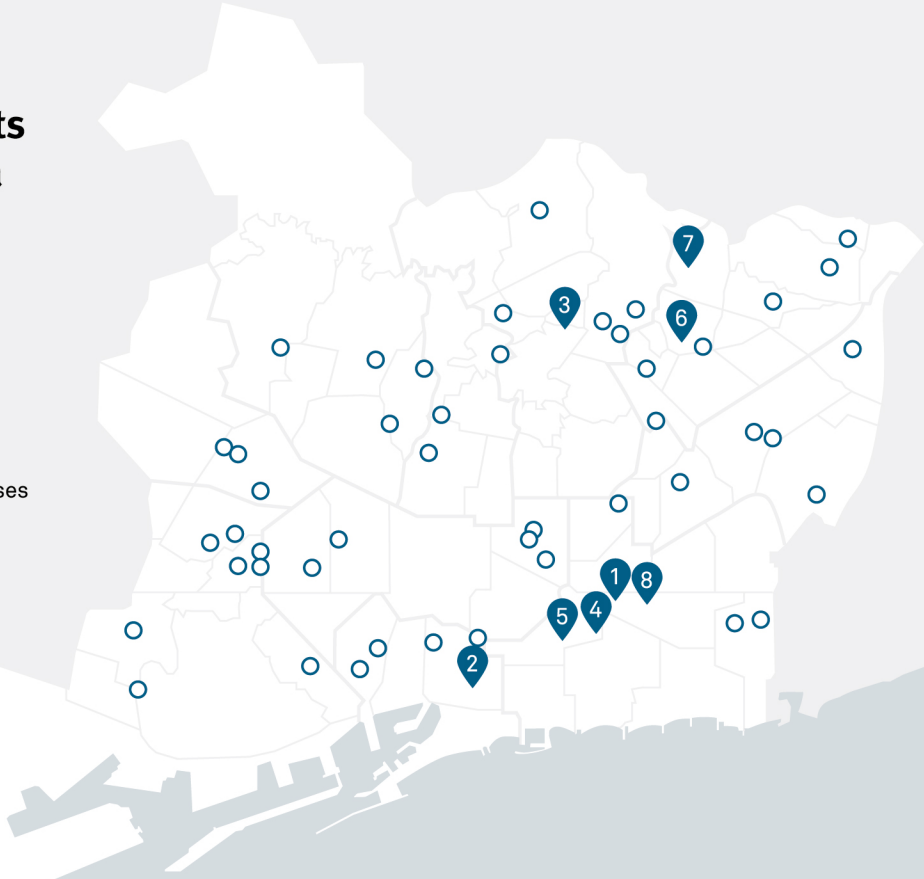
Formació tecnològica i gratuïta per a la ciutadania

barcelonactiva.cat/cibernarium

Xarxa d'equipaments de Barcelona Activa

1. Seu Central Barcelona Activa
   Porta 22
   Centre per a la Iniciativa Emprenedora Glòries
   Incubadora Glòries
2. Convent de Sant Agustí
3. Ca n'Andalet
4. Oficina d'Atenció a les Empreses
   Cibernàrium
   Incubadora MediaTIC
5. Incubadora Almogàvers
6. Parc Tecnològic
7. Nou Barris Activa
8. innoBA

○ Punts d'atenció a la ciutat



Barcelona Activa

Ajuntament de Barcelona

Darrera actualització 2020

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional

Segueix-nos a les xarxes socials:

barcelonactiva.cat/cibernarium
barcelonactiva
barcelonactiva
company/barcelona-activa