PRINCIPALS *FRAMEWORKS* UTILITZATS

En aquest article es descriuran els tres principals *frameworks* de processament d'informació que existeixen actualment: Hadoop, Apache Spark i Apache Storm.

Hadoop

Hadoop és un *framework open source* de processament i emmagatzematge distribuït, utilitzat en aplicacions de *big data* i executat sobre sistemes clusteritzats. En la majoria de casos és el centre d'ecosistemes encarregats d'oferir analítica avançada i predictiva, *data mining* i aplicacions de *machine learning*.

Està format principalment per quatre components:

- HDFS (*Hadoop Distributed File System*): sistema de fitxers que gestiona l'emmagatzematge i l'accés distribuït a la informació sobre diversos nodes d'un clúster.
- YARN (*Yet Another Resource Manager*): és el gestor de recursos d'un clúster de Hadoop, responsable d'assignar els recursos del sistema a les diferents aplicacions i tasques executades.
- MapReduce: és el *framework* de processament utilitzat en aplicacions *batch* per a moure grans volums d'informació en sistemes Hadoop.
- Eines Hadoop: és el conjunt d'utilitats i llibreries que proporcionen les capacitats necessàries per donar suport i interconnectar tots els serveis de l'ecosistema Hadoop.

El funcionament de Hadoop es basa en dos components principals: el primer és el **sistema de fitxers** (HDFS), que s'encarrega de dividir les dades en diferents nodes, replicar-los per oferir alta disponibilitat a l'aplicació i gestionar la informació i l'estat del clúster; el segon component, **MapReduce**, processa les dades en cadascun dels nodes paral·lelament i calcula el resultat de cada tasca.

Hadoop és important perquè:

- Pot emmagatzemar i processar grans quantitats de dades estructurades i no estructurades ràpidament.
- El processament està protegit davant de caigudes del sistema. D'aquesta manera, si un node queda fora de servei, la tasca és redirigida automàticament a altres nodes disponibles per tal que la computació distribuïda no falli.
- Les dades no necessiten ser preprocessades una vegada són emmagatzemades. Les organitzacions poden emmagatzemar tota la informació que desitgin, incloent dades no estructurades (text, vídeo, imatges, etc.) i decidir després què fer-ne.
- És escalable horitzontalment. En cas de necessitat, es poden afegir més nodes al clúster per emmagatzemar o processar més informació.

Apache Spark

Spark és un motor de processament distribuït de propòsit general. Destaca per la seva versatilitat, perquè els quatre mòduls pels quals està format permeten la seva compatibilitat amb molts escenaris d'analítica avançada, tant en *batch*, com en *streaming*, per aplicar algoritmes i models de predicció o representació de grafs. Està optimitzat per treballar en memòria i pot aconseguir una velocitat de processament fins a 100 vegades major que amb MapReduce, podent manipular *petabytes* de dades al mateix temps. Suporta els llenguatges de programació Java, Scala, Python i R.

Els casos d'usos més típics d'Apache Spark són:

- Processament streaming: processament de logs, dades de sensors i, en general, qualsevol stream de dades com el clickstreaming (provinent de fonts web), xarxes socials, monitoratge de sistemes, transaccions financeres, etc. Les dades provenen de manera contínua en un stream i s'emmagatzemen i processen en el mateix instant en el qual entren en el sistema. Aquest tipus de processament és molt útil per a anàlisis de sentiment, tractament de telemetria en mitjans de transport i logística o sistemes de recomanació en aplicació de música i streaming de vídeo.
- Machine learning: la capacitat d'Spark per treballar amb les dades en memòria i executar consultes de manera recursiva i escalable, converteix aquest framework en una excel·lent opció per executar algoritmes d'aprenentatge automàtic. D'aquesta manera, es poden oferir respostes de tendències de mercat o comportament, predicció d'esdeveniments o detecció de frau.
- Analítica interactiva: proporciona molta flexibilitat a l'hora d'executar consultes a bases de dades de manera ràpida, sense necessitat que estiguin preestablertes pel sistema de manera estàtica. Spark ajuda a refrescar la informació d'un quadre de comandament dinàmic o a dur a terme tasques de data discovery per donar respostes a preguntes de negoci.
- Integració de la dada: Spark és una peça fonamental a l'hora de fer tasques de consistència de la informació, reduint considerablement els costos i temps de processament dels processos corporatius. És molt utilitzat a l'hora de crear processos ETL per extreure diferents orígens de dades i completar jobs de neteja, normalització i càrrega de resultats en el sistema de destinació.

Apache Storm

És un sistema de computació en temps real, *open source*, tolerant a errors i distribuït. A diferència d'Apache Spark, està molt enfocat a processament d'*streams* i esdeveniments en temps real. Inclou el seu propi gestor de recursos, mentre que Spark necessita la utilització de YARN o Mesos per a l'orquestració de tasques.

Atès que comparteix moltes singularitats amb Spark, és important tenir clares les seves diferències i a què està enfocat cada *framework* per poder triar l'un o l'altre correctament en funció de les nostres necessitats:

- Processament streaming: Storm ofereix millor rendiment pel fet que utilitza una metodologia micro-batch, és a dir, comprova el flux d'entrada amb més freqüència que Spark.
- Llenguatges de programació: Storm suporta més llenguatges que Spark.
- Latència: Storm proporciona millor latència amb menys restriccions.
- Cost de desenvolupament: Storm no suporta que el mateix codi sigui utilitzat per a processament batch i de temps real, mentre que Spark si que ofereix aquesta possibilitat. És molt important en entorns multiprocessament.
- Throughput. Spark pot processar fins a 100 k de registres per segon, fins a 10 vegades més que Storm.

Com es pot apreciar en l'anàlisi de l'article, tots els *frameworks* ofereixen un potencial enorme i tots tenen les seves particularitats, per la qual cosa és molt important saber quin tipus d'emmagatzematge i processament precisa la nostra organització per poder triar el motor de processament que més s'ajusti 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

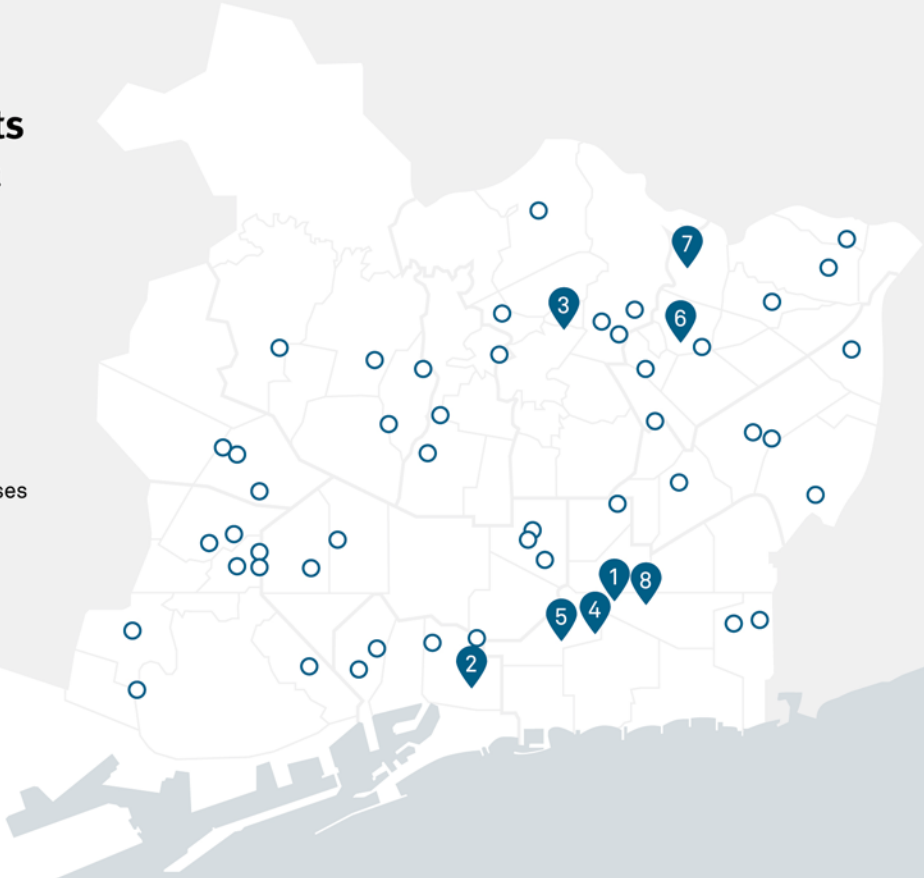
Formació tecnològica i gratuïta per a la ciutadania

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