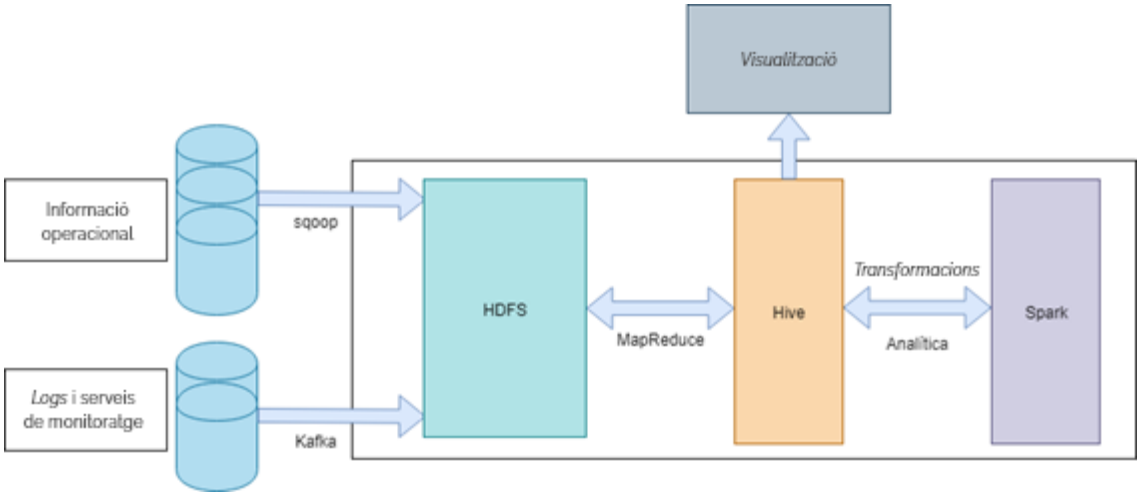
EXEMPLE D'UTILITZACIÓ D'UN *FRAMEWORK*

En aquest article veurem un cas pràctic de la utilització d'Hadoop. Per a això, representarem un *data warehouse* corporatiu (EDW, *enterprise data warehouse*), és a dir, un repositori unificat per a totes les dades que recullen els diversos processos de l'empresa. En la següent imatge es pot apreciar l'esquema lògic de l'arquitectura, que està dividit en diverses capes:

- Orígens de dades: es tenen fonts de dades de caràcter operacional (factures, albarans de compra, inventariat de productes, etc.) i d'altres que provenen dels sistemes informàtics (logs dels servidors o monitoratge de sensors). Quant a la tipologia, tindrem fonts corresponents a bases de dades, fitxers de text o cues de Kafka (servei big data que proveeix l'arquitectura d'un sistema de cues per al processament de streams en temps real).
- Emmagatzematge: es disposarà d'un clúster distribuït d'HDFS, que s'encarregarà d'emmagatzemar tots els fitxers dels orígens de dades.
- Processament: aquesta capa s'encarregarà d'executar operacions de MapReduce per netejar, transformar i analitzar la informació emmagatzemada.
- Visualització: s'utilitzaran diverses eines de visualització (MicroStrategy, QlikView o Power BI) que permeten representar quadres de comandament i exportar informes en format de text o fulls de càlcul.



Dins del flux de la dada, destacarem les etapes següents:

Etapa d'adquisició de dades

Per tal que puguem portar-nos la informació més recent dels orígens de les dades, s'utilitzaran tres serveis:

- Sqoop: és una interfície que permet transferir dades de manera automàtica entre bases de dades relacionals i HDFS. En aquest cas, podrem obtenir tota la informació operacional de manera incremental (la més recent) al nostre sistema de fitxers de Hadoop.
- Flume: servei distribuït encarregat de recol·lectar, agregar i moure grans quantitats de dades de tipus *log*. Té una arquitectura molt senzilla, utilitza pocs recursos, és robust i està enfocat a fluxos de dades en *streaming*. Ens servirà de molta ajuda a l'hora de portar-nos a HDFS els orígens de dades de tipus *log*.
- Kafka: és un servei de processament en *streaming*. A través de la gestió de cues és capaç de proporcionar fluxos de dades en temps real amb un *throughput* alt i baixa latència. Serà la base per obtenir la informació procedent de monitoratge de sensors i inserir-la en el nostre sistema de fitxers distribuïts.

Etapa d'emmagatzematge

Com a nucli del *data warehouse*, HDFS és la pedra angular que proveeix a tots els serveis de la persistència necessària per registrar la informació corporativa, estat dels serveis en execució i resultats de les anàlisis descriptives i de predicció.

La informació vindrà organitzada en les capes següents:

- *Landing zone*: emmagatzema els fitxers nous i dades crues, és a dir, sense cap mena de tractament.
- *Cleansing zone*: conté la informació una vegada que se li han aplicat processos de neteja, validació i eliminació de duplicats.
- *Transformed zone*: registra el model de dades tractat, que conté l'estructura final de cadascuna de les taules de detall i agregades, que serviran d'origen als processos analítics i als *frameworks* de processament. Els resultats analítics també poden ser emmagatzemats en aquesta capa.

Etapa de processament

Per dur a terme tant la translació i transformació d'informació entre les capes del nostre *data warehouse*, com la posterior analítica, s'utilitzaran aquestes dues eines:

- Tasques MapReduce: utilitzarem Hive com a interfície SQL per portar-nos les dades de la capa de *landing zone* a la de *cleansing zone*. Hive és un servei construït sobre HDFS que realitza operacions de MapReduce per dur a terme les consultes programades. A més, ofereix un model relacional de base de dades, que pot ser utilitzat per un *framework* de processament de dades com Spark o Storm, o per eines de visualització de dades per a la seva posterior explotació.
- *Jobs* de Spark: el clúster de Spark executarà totes les tasques de transformació de la dada, completant el model relacional amb noves taules, agregant i creuant la informació

dels diferents orígens de les dades. D'altra banda, també s'encarregarà, a través del seu mòdul de *machine learning*, de realitzar models predictius i analítics que mostrin els resultats buscats.

Etapa de visualització

Finalment i no menys important, cal representar tota la informació adquirida en el nostre sistema i els resultats de les anàlisis en diversos quadres de comandament. Diferenciarem dos tipus:

- *Dashboards* en temps real: molt útils a l'hora de mostrar l'estat actual dels *logs* de servidors i el monitoratge de sensors. Es podrà apreciar a simple vista, si algun dels valors supervisats supera un llindar preestablert, de manera que es puguin prevenir avaries o possibles sobrecarregues en algun dels nodes del clúster. També són essencials a l'hora de controlar la gestió i dimensionament de recursos assignats a les nostres màquines.
- *Dashboards* sota demanda: aquests quadres de comandament s'actualitzaran de manera periòdica o manual, quan es requereixi saber l'estat de les KPI del nostre negoci. Aquestes visualitzacions representaran el resultat dels algoritmes i anàlisis dutes a terme pel nostre motor de processament, oferint informació de tendències, prediccions i anàlisis de la informació actual.

Per implementar aquestes funcionalitats, es disposa de moltes eines. Entre les més conegudes, es poden trobar MicroStrategy, QlikView o Power BI.

A través de totes aquestes capes es completa el flux de la dada. Recordem que la dada ha de ser adquirida i emmagatzemada en el nostre sistema de fitxers, ha de ser netejada i analitzada pel nostre *framework* de processament i, finalment, ha de ser representada per la nostra eina de *reporting* preferida.

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

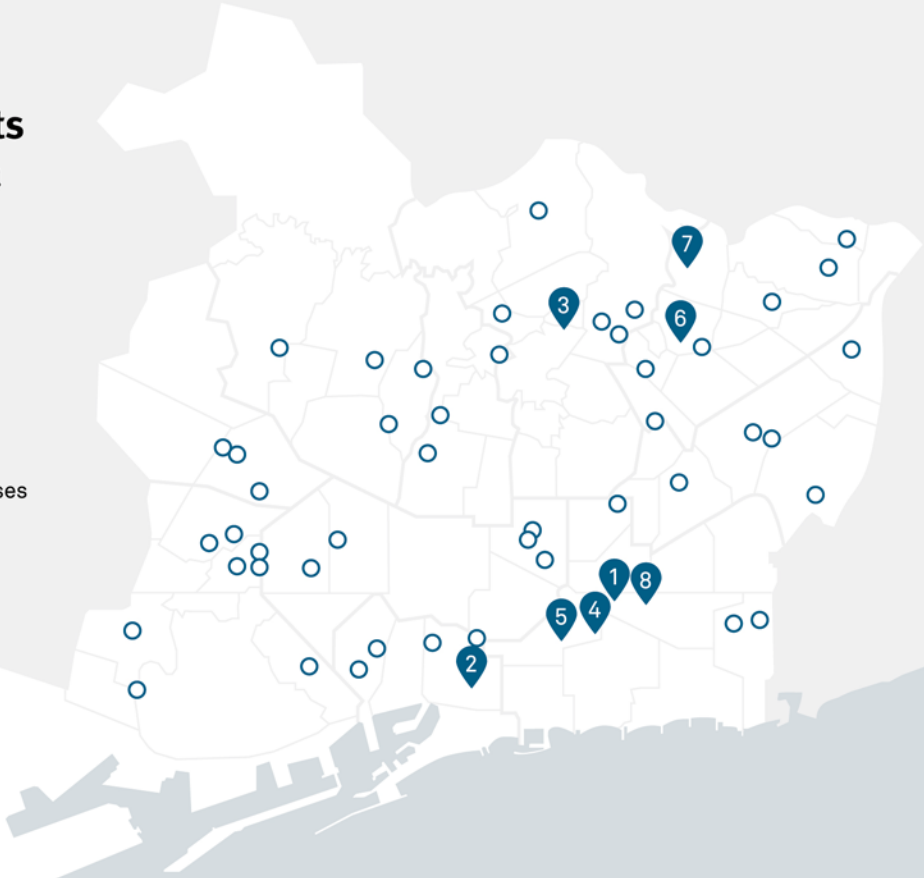
Formació tecnològica i gratuïta per a la ciutadania

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