Introducció al *Big Data*

Ajuntament de Barcelona

Barcelona Activa

Índex

1 INTRODUCCIÓ AL *BIG DATA*

Durant aquest mòdul farem els nostres primers passos pel món del *big data*, introduirem conceptes clau per entendre què és, com funciona, on es desplega, per què té sentit tant per a un client o una clienta com per a una empresa i quan és necessari respecte als sistemes tradicionals.

Començarem entenent la importància de la dada, ja que és el precursor de qualsevol sistema d'adquisició, anàlisi i explotació. Fins ara, la *business intelligence* (BI) intentava oferir solucions als problemes corporatius, optimitzar els processos interns i d'aquesta manera guanyar avantatge sobre la seva competència. A causa de la pròpia evolució dels sistemes informàtics, el creixement del volum d'informació que s'ha de processar és exponencial. Aquests sistemes tradicionals (BI), que fan ús de sistemes d'emmagatzematge i eines que, per les seves característiques, no estan preparades per gestionar tanta quantitat de dades, han quedat obsolets per a les demandes i necessitats de qualsevol organització. Per aquest motiu, han nascut altres estratègies com el *big data*, que centra els seus esforços a cobrir les necessitats de l'ara i del demà, perquè unes de les seves principals característiques són la seva escalabilitat i tolerància a errors.

Per tant, un sistema *big data* ha de cobrir les etapes principals dels processos habituals de BI, és a dir, ha d'abastar tot el flux de la dada, des que es genera un registre, fins que és emmagatzemat en el nostre sistema, netejat i normalitzat per a la seva posterior exploració i representat en un informe o quadre de comandament. Durant aquest primer mòdul, veurem les diferències entre els sistemes actuals i els tradicionals, repassant les bases de dades i tècniques d'anàlisi d'informació que utilitza cadascun.

Estructura del mòdul 1:

- Què és el *big data*?
  Presentarem i entendrem què és el *big data*, explicant el flux de la dada i alguns casos d'ús pràctics que podem veure en el nostre dia a dia.

- Origen i utilitat de les dades
  Veurem detalladament els tipus de dades que es recullen en el *big data*, els diferents orígens i la forma en la qual s'utilitzen per afegir valor a les organitzacions o empreses.

- Les 4 V del *big data*
  Entendrem què són les 4 V del *big data* i per què són necessàries.

- Conceptes bàsics relacionats
  Abordarem conceptes bàsics que tenen relació amb el *big data*: bases de dades i mineria de dades.

- Altres conceptes relacionats
  Es descriuran conceptes relacionats amb l'analítica de dades i la intel·ligència de negoci i com es conceben dins d'un sistema de big *data*.

- Resum d'idees clau
  Repassarem els conceptes vistos en el primer mòdul del curs.

- Exercici
  Realitzarem un cas pràctic sobre dades reals d'empreses, per poder aplicar els conceptes estudiats.

1.1 QUÈ ÉS EL *BIG DATA*?

Hola!

En aquest vídeo analitzarem què és el *big data*, intentarem entendre com funciona i mostrarem per què està tan de moda avui dia. És una nova manera de veure el món, i ningú està fora del seu abast!

El primer que ens pot venir al cap quan llegim o escoltem *big data* pot ser quelcom com "dades grans" o almenys quelcom relacionat amb gestionar moltíssima informació, amb una supercomputadora al darrere gestionant molts uns i zeros sobre una pantalla negra (tal com recordarem a *Matrix*), i no aniríem mal encaminats! Al cap i a la fi, *big data* és una manera de gestionar moltes dades, totes diferents entre elles i a una gran velocitat.

Però, com funciona això del *big data*? Podem distingir quatre etapes principals:

1. Adquisició de dades: el *big data* està format per molts conjunts de dades, que provenen, entre d'altres, de bases de dades, fitxers o serveis al núvol (per exemple: una xarxa social, la bossa o l'aplicació de meteorologia). Cal tenir en compte que es poden arribar a recollir *terabytes*, o fins i tot *petabytes* d'informació. Els procediments tradicionals d'ETL (*Extract, Transform and Load,* processos de gestió de la dada de *business intelligence*) no són capaços de gestionar múltiples orígens de dades diferents ni processar tal quantitat d'informació de manera eficaç, mentre que els processos *big data* sí que estan preparats per dur a terme aquestes tasques de manera ràpida i veraç.

2. Transformació: totes aquestes fonts d'informació tenen estructures de dades diferents, ja que algunes són fitxers de text pla sense cap jerarquia i d'altres són dades que segueixen un esquema fix. En qualsevol cas, un procés de *big data* és capaç de processar tota aquesta informació, revisar la seva integritat i qualitat (per exemple, validació d'esquemes, normalització i de duplicació) i transformar-la per poder ser interpretada d'una manera senzilla (agregació i filtratge de dades).

3. Emmagatzematge: tota la informació adquirida s'emmagatzema en bases de dades distribuïdes, relacionals o no relacionals, o en sistemes de fitxers clusteritzats (que veurem més endavant). Aquestes solucions poden residir en un conjunt de servidors locals *on premise* o, cada vegada més comú, al núvol, on proveïdors com Google, Amazon o Microsoft ens proporcionen aquests sistemes a cop de clic i en només segons.

4. Exploració: per acabar, una vegada que es té la informació emmagatzemada i ben organitzada, podem crear els nostres propis quadres de comandament per analitzar-la en temps real, o bé aplicar algun model d'aprenentatge automàtic i intel·ligència artificial.

Tot això està molt bé, però com pot ajudar-me a mi o a una empresa a millorar el dia a dia? Fàcil, el *big data* està present quan utilitzem una plataforma *streaming*, com Netflix o Spotify, per recomanar-nos noves sèries o cançons en funció dels nostres gustos. Està present a les xarxes socials per mostrar-nos temes, missatges o fotos d'interès. També a Google Maps per indicar-nos quina és la ruta més ràpida per arribar a la nostra destinació, o a l'aplicació del banc per gestionar els nostres estalvis i categoritzar les nostres despeses.

Altres casos d'ús molt comuns solen estar enfocats a millorar els processos de negoci a les empreses. Per exemple, les companyies de venda *online* utilitzen algorismes de predicció per optimitzar l'estoc en funció de les cerques a internet, xarxes socials, tendències, meteorologia, etc. En el món de la logística, s'està aconseguint optimitzar les rutes de repartiment, així com el seguiment de les mercaderies, gràcies a la informació de trànsit generada pels nostres dispositius GPS o la que cedeix trànsit des dels seus servidors.

També podem apreciar l'impacte del *big data* en la millora del rendiment esportiu, on es poden determinar patrons i estils de joc de diferents esports, de manera que es pugui millorar la planificació dels entrenaments en funció de la condició física o l'estratègia utilitzada en un partit.

I això és només el començament del que ens pot oferir aquesta tecnologia. Ens veiem en el següent vídeo!

1.2 ORIGEN I UTILITAT DE LES DADES

En aquest article parlarem dels tipus de dades que existeixen en l'univers del *big data* i quines són les seves principals fonts. Mostrarem exemples perquè ens anem familiaritzant a poc a poc amb formats, sintaxis i perquè anem comprenent com afecta cada cas d'ús al negoci i com pot millorar la nostra qualitat de vida.

Tipus de dades

En primer lloc, analitzarem els diferents tipus de dades amb els quals es treballa. Com a primera aproximació, imaginarem que treballem amb molts fitxers independents i que

cadascun d'ells pot comprendre un dels formats que es descriuen a continuació, els quals es categoritzen en funció de la seva estructura.

- Estructurades
  - Aquest tipus de dades tenen la seva longitud, format i mida ben definits. S'emmagatzemen en taules, fulls de càlcul o bases de dades. A la taula que es mostra a continuació, es poden apreciar les últimes transaccions bancàries d'una persona particular. La informació ve organitzada pels atributs "Tipus", "Origen" i "Concepte", i cada registre té una mètrica, corresponent a l'"Import" de l'operació:

TIPUS	ORIGEN	CONCEPTE	IMPORT
Pagament amb targeta	Compte corrent	Gasolinera	55
Transferència	Compte d'estalvi	Balanç de despeses	500
Pagament amb targeta	Compte corrent	Supermercat	30

- Semiestructurades
  - Els registres d'un origen semiestructurat no segueixen una definició estàndard com els estructurats, és a dir, no tenen assignat un format comú, però sí que presenten una organització basada en metadades. Gràcies a ella, s'estableixen objectes i les seves relacions, que moltes vegades estan acceptats per convenció en formats coneguts com a HTML, JSON o XML. En l'exemple que es mostra a continuació, es pot observar un document tipus JSON, en el qual s'ha creat un objecte "Operació" i els seus elements niats, que representen transaccions bancàries i les seves propietats:

```
{
   "Operació": [
   {
      "Tipus": "Pagament targeta",
      "Origen": "Compte corrent",
      "Concepte": "Gasolinera",
      "Import": "55 €"
   },
   {
      "Tipus": "Transferència",
      "Origen": "Compte estalvi",
      "Concepte": "Balanç despeses",
      "Import": "500 €"
   },
   {
      "Tipus": "Pagament targeta",
      "Origen": "Compte corrent",
      "Concepte": "Supermercat",
      "Import": "30 €"
   },
}
```

- No estructurades

- Tal com indica el seu nom, els conjunts de dades que comprenen informació no estructurada són aquells que no segueixen un format específic. Solen associar-se a documents de text, imatges, vídeos o *e-mails,* entre d'altres. Seguint amb els exemples anteriors, la taula que descriu les transaccions bancàries podria estar emmagatzemada bé en una fulla de càlcul (.xlsx), en format de text avançat (.docx) o documents PDF. La seva explotació és més complexa que fitxers de text pla i requereix de programari específic, moltes vegades sota llicència, o d'algoritmes avançats, com pot ser el processament d'imatges o fitxers d'àudio.

Tipus de dades per origen

Els sistemes *big data* han arribat per donar resposta al gran volum d'informació disponible que gestionem, de manera que es pugui processar de manera ràpida i crear un valor que ofereixi solucions a qualsevol tipus d'escenari. Fins ara, les bases de dades tradicionals (Oracle, MySQL, SQL Server, SAP, etc.) han pogut donar una solució parcial al processament de la informació, però, ara, els sistemes moderns, gràcies a les noves arquitectures i capacitats de processament, són capaços d'adquirir, transformar i analitzar les dades de múltiples orígens en temps real. Aquestes fonts poden provenir de l'app de missatgeria del nostre *smartphone*, dels sensors que monitoren la geolocalització, temperatura, posició, etc. de diferents dispositius, de correus electrònics o de *logs* de servidors. Tots aquests nous focus d'informació formen un nou i complex univers de dades, que, combinats, multipliquen el seu valor, pel fet de ser emmagatzemats de manera conjunta al mateix lloc.

Podríem considerar alguns dels exemples que es descriuen a continuació com els orígens més comuns dels sistemes de *big data*.

- Generats per persones. Són aquelles fonts que generem de manera o no conscient, bé sigui amb l'ús de les eines quotidianes o amb diversos registres sanitaris, bancaris, etc.
  - Missatgeria instantània
  - Notes de veu i enregistraments d'àudio
  - Correus electrònics
  - Registres electrònics (mèdics, Seguridad Social, bancaris, sistemes privats, etc.)
  - Documents electrònics (DNI, passaport, permís de conduir, targetes de fidelització, etc.)

- Xarxes socials i fonts web. Avui dia, gran part de la població publica imatges, redacta històries o anècdotes, interactua amb altres persones a través de comentaris o reaccions i, per descomptat, utilitza motors de cerca per accedir a qualsevol tipus d'informació. Tot això s'emmagatzema en les famoses *cookies*, en l'historial del nostre navegador i, sobretot, en els servidors dels portals als quals accedim. Aquesta informació és la més valuosa per a les empreses, perquè amb ella poden traçar tendències de mercat, recomanacions personalitzades de pel·lícules, roba o productes d'interès per generar publicitat o registres d'activitat. Per tant, trobem:

- Qualsevol tipus de xarxa social
- Motors de cerca (Google, Bing, Yahoo!, etc.)
- Informació sobre clics a vincles i elements
- RRSS (fonts de dades de Twitter, publicacions a Facebook, xarxes socials de blogs, premsa...)
- Contingut web (pàgines, imatges, enllaços, etc.)

- **Comunicació entre màquines (*machine-to-machine*, M2M).** També conegut per molts com a IoT (*internet of things*). Es tracta de fonts físiques i automatitzades que utilitzen sistemes de radiofreqüència, protocols com Bluetooh, ZigBee, WiFi, RDIF o GPS per monitorar una certa activitat.

  - Targetes d'accés i dispositius de seguiment RFID
  - Dispositius de geolocalització a través de senyals GPS
  - Altres sensors (parquímetres, màquines expenedores, caixers, etc.)

- **Transaccions.** Esdeveniments generats per dispositius de telecomunicacions (mòbils, senyals de radiocomunicació...) o esdeveniments únics de pagaments bancaris o vendes.

  - Registres de comunicacions (crides, missatgeria, VoIP, etc.)
  - Registres de facturació (pagaments amb targeta, pagament *online*, registre de comptador de llum intel·ligent, etc.)
  - Registres de vendes o comandes *online* (utilització de comerç electrònic i d'altres aplicacions com Glovo o Wallapop)

- **Biomètrics.** Són aquells que ajuden a la identificació unívoca d'una persona en funció de la base de trets físics o de conducta. Empleats principalment en sistemes d'accés i seguretat.

  - Reconeixement facial
  - Informació genètica (ADN)
  - Empremtes dactilars
  - Escaneig de retina

1.3 LES 4 V DEL *BIG DATA*

Hola de nou!

En aquest vídeo veurem una de les bases del *big data*. És molt important tenir clar com identificar una tecnologia *big data* d'una altra que no ho és. Són conceptes molt bàsics però que se solen confondre entre ells. No et preocupis, per a això som aquí, no?

Qualsevol sistema de *big data* es construeix sobre quatre pilars, les conegudes 4 V del *big data*: volum, varietat, velocitat i veracitat.

Volum

Avui dia, les dades es generen automàticament pels servidors, les xarxes de comunicació, les interaccions personals, el monitoratge IoT, etc. Això comporta un volum massiu d'informació que ha de ser gestionat per un sistema escalable que sigui capaç d'emmagatzemar i processar tota aquesta quantitat d'elements. Cal tenir en compte que es tenen uns 25.000 milions de dispositius connectats a la xarxa i estem en un creixement exponencial. Per això, el *big data* no només ha de ser capaç d'adquirir aquestes dades aquí i ara, sinó que ha d'estar preparat per anar creixent a mesura que així ho fan les seves fonts.

Però, en què es tradueix això? Perquè ens en fem una idea, una arquitectura de *big data* estàndard ha de poder acceptar fluxos de centenars de *megabytes* o *gigabytes* al segon i poder emmagatzemar fins a *petabytes* d'informació, considerant un petabyte (PB) com 1024 *terabytes* (TB). De manera global, s'estan tractant fins a 40 *zettabytes* de dades, o el que és el mateix, 40 milions de *petabytes* o 40.000 milions de *terabytes*, una xifra marejadora i que sembla l'avantsala del que està per arribar.

Varietat

Aquí podem apreciar l'evolució de la tecnologia, ja que fa uns anys els sistemes tradicionals només processaven informació relacional emmagatzemada en bases de dades. En canvi, ara es pot treballar amb tot tipus de fitxers i estructures. Exemples? Doncs vídeos, fotografies, arxius de so, correus electrònics, fitxers de text o sistemes de monitoratge, entre d'altres. Això provoca que calgui realitzar una enginyeria complexa per al tractament, transformació i homogeneïtzació de cadascuna de les fonts, amb la finalitat que l'anàlisi posterior sigui eficaç i senzill.

Velocitat

Aquesta característica defineix la rapidesa amb la qual es processen les dades amb les que ha d'interactuar la nostra eina de *big data*. Per fer-ho, el nostre sistema haurà d'admetre un flux de dades continu molt gran i de diverses fonts en temps real. Aquest és un dels majors reptes de les empreses i de l'enginyeria de dades. El fet de poder processar la informació en temps real permet a les organitzacions entendre la seva informació amb més claredat, poder crear anàlisis predictives, generar alarmes de l'estat dels seus serveis i prendre decisions

que aporten solucions estratègiques essencials i molt competitives.

Veracitat

Quan parlem de veracitat ens referim a la qualitat de la dada, la seva disponibilitat, el biaix, soroll i la possible alteració que hagi pogut patir a causa de factors aliens al nostre sistema. És la característica més difícil de controlar i un dels principals maldecaps dels responsables del *big data*. Cal saber què i com s'emmagatzema, és a dir, si la informació que ve d'origen és útil en el seu estat natural o si és necessari aplicar-li algun tipus de procés de neteja per garantir que la dada és vàlida, o bé, completar aquesta informació per poder donar-li un valor afegit. Aquest punt és tan important com els altres i cobra vital importància a l'hora d'avaluar i analitzar la informació per oferir les millors solucions estratègiques

En resum, aquests quatre conceptes són la base sobre la qual es construeix qualsevol arquitectura de *big data* i tots són igual d'importants, ja que només es concep un sistema d'aquestes característiques si és capaç d'analitzar un volum massiu de dades en temps real, de múltiples fonts simultàniament i amb una qualitat mínima reconeguda. No sempre es pot aconseguir la millor opció per als quatre elements, però dependrà de nosaltres aconseguir la solució de compromís que millor s'adapti als requisits del projecte.

Fins aviat!

1.4 CONCEPTES BÀSICS RELACIONATS

Hola de nou

Durant els pròxims minuts aprendrem quins tipus de bases de dades comprèn l'univers del *big data*. També introduirem altres conceptes com la mineria i l'arquitectura o modelatge de dades. Tot a punt? Comencem!

Bases de dades

Una base de dades és una eina que s'encarrega d'organitzar la informació. Disposa d'utilitats perquè es pugui accedir, gestionar i actualitzar les dades amb facilitat. Poden estudiar-se des de diferents perspectives, tot i que podrem parlar de dos tipus de sistemes principals en funció de la seva mutabilitat:

- Estàtics: són aquells que comprenen dades immutables, és a dir, no es poden modificar. El seu ús està enfocat principalment a la consulta de la informació sobre esdeveniments passats, facilitant la generació d'informes i la mineria de dades. També es coneixen com OLAP (*On Line Analytical Processing*). S'organitzen en cubs multidimensionals, que precarreguen la informació des d'un magatzem de dades. Els més coneguts són Cognos (IBM), SAP, Oracle Database OLAP o Microsoft Analysis Services.

- **Dinàmics**: emmagatzemen registres que sí que poden ser alterats. Són els més utilitzats. També anomenats **OLTP** (*On Line Transaction Processing*). Basats en bases de dades relacionals tradicionals, que es veuran a continuació.

També podem veure les bases de dades en funció de la seva organització:

- **Relacionals**. Són les més tradicionals i utilitzades tant pels sistemes de *business intelligence* com per alguns orígens de *big data*. Es caracteritzen per guardar la informació de manera estructurada en taules, on prèviament s'ha indicat quin tipus i longitud tindran els seus camps. Permeten relacionar els elements de diferents taules de manera senzilla i ràpida. El seu llenguatge de consulta és SQL (*Structured Query Language*).
    - Algunes d'aquestes bases de dades podrien ser les tradicionals Oracle, MySQL, SQL Server o bé altres de més modernes allotjades en plataformes al núvol com RDS (Amazon Web Services), Cloud SQL (Google Cloud Platform) o Azure SQL Database.
- **No relacionals**. També conegudes com a NoSQL (Not only SQL). Es caracteritzen per no tenir un esquema definit per a l'emmagatzematge de la informació. Això permet emmagatzemar qualsevol tipus de dada, tant text, com arxius d'àudio o vídeo, correus electrònics o PDF. A diferència de les anteriors, estan preparades per treballar en clúster (conjunt de servidors), de manera que permeten:
    - L'escalabilitat del sistema
    - Funcionar de manera distribuïda, la qual cosa comporta un processament paral·lelitzador més ràpid que el tradicional
    - Tolerància a fallades
    - Reduir costos

Per contra, són una mica més complexes que les primeres i, en molts casos, són menys flexibles a l'hora de creuar elements de diferents conjunts de dades .

Les bases de dades no relacionals serien MongoDB, DocumentDB, Apatxe Cassandra, Redis o HBase.

Finalment, estudiarem la visió de bases de dades en funció del seu contingut. Poden ser:

- **Transaccionals**: aquesta categoria engloba la majoria de bases de dades tradicionals. Asseguren la integritat de la dada, tenen baixa latència i són fiables.
- **Documentals**: guarden dades semiestructurades, principalment sota l'estàndard JSON o XML. Permeten emmagatzemar registres (documents) amb diferents camps entre si, millorant la flexibilitat de la solució. Molt enfocades a ser utilitzades com a objectes, agilitant el processament de les dades. La base de dades documental més coneguda és MongoDB.

- Clau/valor: estan dissenyades per a consultes ràpides en temps real, ja que solen estar emmagatzemades en memòria (RAM). Redis, Cassandra o DynamoDB (Amazon Web Services o AWS) són les més conegudes per a aquest propòsit.
- Columnars: emmagatzemen dades estructurades en columnes, ideals per a analítiques de dades. Exemples: HBase i Kudu.
- Gràfiques: estructuren la seva informació en nodes i arestes, de manera que pugui ser representada gràficament de manera senzilla. Neo4J és la més utilitzada per la comunitat.

Mineria de dades

Com gestionem i processem tota aquesta informació emmagatzemada perquè sigui útil i pràctica en els nostres sistemes? La mineria de dades o *data mining* és el conjunt de tècniques i eines que analitzen grans volums de dades. És molt important tenir clares les diferències entre *big data* i mineries de dades, podem entendre-ho com un tot.

- El big data s'encarrega de recaptar tota la informació a les bases de dades comentades prèviament.
- A través del data mining es netegen les dades i s'exploten, creant models analítics i predictius per al descobriment de patrons i tendències. Això fa augmentar la rendibilitat i productivitat de l'organització.

Sota aquestes anàlisis, la mineria de dades serviria, per exemple, per detectar frau, predir una malaltia o estimar la congestió de trànsit en una ruta determinada.

Ens veiem en el següent vídeo!

1.5 ALTRES CONCEPTES RELACIONATS

Avui dia, totes les empreses tenen fonts i bases de dades enormes, però no totes són capaces d'analitzar tota la informació i donar valor a les dades per millorar el seu negoci. Aquelles que no aconsegueixin superar el repte estan condemnades a prendre males decisions, a no tenir un coneixement clar de la seva posició en el mercat i a no saber predir quines necessitats hauran de cobrir a mitjà i llarg termini. Per això, és molt important que es destinin molts esforços a fer tasques d'anàlisis sobre les dades de negoci.

En primer lloc, s'ha de tenir clar **quin tipus d'anàlisi es vol realitzar**, ja que, en funció del nostre objectiu, podem trobar quatre perspectives diferents:

- Anàlisi prescriptiva: ens ajudarà en la presa de decisions, a realitzar accions concretes sobre qüestions actuals del negoci, optimitzant els recursos de l'actual operativa. Són tècniques matemàtiques que informen del que podria passar i, d'aquesta manera, suggerir decisions per millorar els indicadors de negoci. Aquesta anàlisi ajudaria a una

empresa, per exemple, a identificar una oportunitat de mercat i poder realitzar una campanya de captació de clientela adequada.

- Anàlisi predictiva: ens donarà una estimació del que succeirà. Utilitza tècniques de modelització, aprenentatge automàtic i mineria de dades per analitzar dades actuals i històriques i a partir d'elles fer prediccions d'esdeveniments futurs. Per exemple, és comú que les empreses energètiques o de telecomunicacions tinguin models que permetin predir l'ús que té cada client o clienta del seu servei, de manera que puguin proposar-los ofertes que permetin un estalvi en les seves factures.
- Anàlisi diagnòstica: establirà el perquè d'un fet. És el més senzill, ja que, utilitzant les regles de negoci i tècniques bàsiques de bases de dades, es poden obtenir els motius d'ocurrència d'un fet concret. Són essencials per determinar i explicar els resultats d'accions determinades. Per exemple, qualsevol empresa utilitza aquestes anàlisis per estudiar per què s'ha produït una disminució o augment de beneficis en un moment concret, de manera que li permeti prendre accions futures per evitar o potenciar aquesta situació.
- Anàlisi descriptiva: analitzarà què va passar en una situació passada. És una anàlisi estadística que avalua els diferents valors que poden prendre les variables del nostre model i ajudar-nos així a reproduir i traçar el camí a una situació desitjada. Els més simples ajuden a quantificar, per exemple, el volum de vendes d'una empresa en funció de productes o categories, mentre que uns altres de més complexos ajuden a determinar la distribució geogràfica d'una malaltia en un període determinat.

En qualsevol dels casos anteriors s'ha de tenir en compte que els resultats són dinàmics, és a dir, situacions actuals poden ajudar a entendre el perquè de fets històrics. Això podria alterar i afegir condicionants als nostres models perquè es vagin ajustant cada vegada més en el temps, utilitzant tècniques d'aprenentatge automàtic.

Una vegada decidit quin tipus de recerca volem realitzar sobre les dades, es pot procedir a establir una sèrie de passos que seran d'ajuda per a emmagatzemar la informació de manera ordenada i neta, la qual cosa ajudarà, en gran mesura, a la seva posterior anàlisi. Per això, existeixen les fases següents:

- Es determina l'objectiu i els KPI de negoci (*Key Performance Indicators*). Aquests indicadors permetran mesurar el resultat de manera numèrica, de manera que se sàpiga prèviament quines mètriques calcularà la nostra anàlisi. Per exemple, es pot saber el comportament i preferències de la clientela en funció de quant i quan compren, o avaluar la competència, sabent qui té més vendes i qui reté més clientela.
- S'adquireixen les dades. Aportarem al nostre sistema les diferents fonts de les quals s'han d'alimentar les nostres bases de dades. Aquesta informació podrà ser estructurada, semiestructurada o no estructurada i cadascuna d'aquestes fases passarà per un procés alternatiu que normalitzi la seva informació i l'emmagatzemi de manera que tant la seva escriptura com lectura siguin ràpides.

- Es processen les dades. Es fan les primeres tasques per organitzar la informació emmagatzemada de manera correcta. Es filtra, agrupa i creua la informació per construir un model correcte per al nostre negoci.
- Es netegen les taules, col·leccions i índexs. S'aplicaran processos de neteja que s'encarreguin de modificar o esborrar registres erronis o corruptes i duplicats.
- Anàlisi exploratòria de les dades. Aquestes primeres iteracions intenten realitzar una anàlisi descriptiva de les dades, veient quins atributs són significatius i quins valors adopten al llarg del temps, gràcies a la creació d'histogrames, tendències i altres gràfics.
- Modelatge i algoritmes. Es construeixen variables estadístiques (mitjanes, medianes, modes, desviacions, màxims, mínims, etc.), regressions i s'apliquen algoritmes d'aprenentatge automàtic i predicció.
- Explotació de la informació. Una vegada que s'ha realitzat l'anàlisi requerida, es generen informes i quadres de comandament per oferir, de manera senzilla i interactiva, el resultat de l'anàlisi i els KPI. Per a això, és comú que s'utilitzin eines de *reporting* com Power BI, MicroStrategy o QlikView.

Amb el seguiment correcte de les fases anteriors es poden obtenir uns resultats que marquin la diferència en una empresa, ja que és fonamental tant la presa de decisions com el corresponent coneixement del mercat i la situació actual de l'organització. Desafortunadament, tenir bones dades no garanteix bons resultats de la nostra anàlisi, però unes males dades sí que ens asseguren conclusions incorrectes. Per aquest motiu, és essencial dedicar recursos a conèixer què i com es té i a on es vol arribar.

1.6 IDEES CLAU: INTRODUCCIÓ AL *BIG DATA*

Durant aquest primer mòdul del curs, hem fet els nostres primers passos en el món del *big data*, hem après què és i com funciona, oferint diferents perspectives de com n'és d'útil aquesta nova tecnologia per a empreses de perfils tan diferents com les tèxtils o plataformes *streaming*.

Al seu torn, s'ha mostrat quins tipus de dades existeixen en funció de la seva estructura (estructurades, no estructurades o semiestructurades) i origen. D'aquesta manera, podem fer-nos una idea de quanta informació es genera diàriament, tant per a persones particulars a través dels seus dispositius mòbils, com per altres elements d'ús habitual com un caixer, una comanda *online* o la targeta d'accés a l'empresa.

S'ha introduït el concepte dels quatre pilars del *big data*, les 4 V sobre les quals està construït cada sistema (volum, varietat, velocitat i veracitat). Cal tenir clar que una arquitectura *big data* s'encarrega "únicament" de processar molta informació de manera molt ràpida, de diferents fonts simultàniament i oferint una qualitat i consistència adequades.

S'ha ofert una visió global sobre les bases de dades utilitzades per aquests sistemes, de quins tipus principals existeixen (estàtiques i dinàmiques), com es classifiquen en funció de la seva organització (relacionals i no relacionals) i en funció del seu contingut (transaccionals, documentals, clau/valor, columnars i gràfiques).

Finalment, apart d'introduir què és la mineria de dades i com s'integra amb el *big data*, s'han descrit els diferents tipus d'anàlisis d'informació que es poden dur a terme (prescriptiva, predictiva, diagnòstica i descriptiva), les diferents fases per les quals ha de passar una dada en el nostre sistema i per què és tan important que un negoci dediqui recursos i esforços a realitzar aquestes investigacions sobre les seves dades.

2 INFRAESTRUCTURA PEL *BIG DATA*

Al llarg d'aquest segon mòdul del curs podrem aprofundir en l'evolució que ha tingut el *big data* des dels seus orígens i com ha aconseguit cobrir una necessitat que s'anava augmentant en els últims anys. Cada vegada generem més informació i necessitem sistemes escalables que estiguin preparats, ara i a llarg termini, per suportar aquest creixement.

Al seu torn, estudiarem què és un clúster i de quines parts es compon. És essencial que comencem a consolidar aquests conceptes, perquè són la base de tots els dissenys i arquitectures de *big data*. Comprovarem com és un flux real de la dada, és a dir, des que aquest es genera en origen, com s'emmagatzema i es transforma en el nostre sistema i finalment com s'agrega per ser visualitzat en un informe o en una eina de *reporting*.

Finalment, veurem detalladament quines són les formes més comunes de gestionar i emmagatzemar la informació al clúster, de manera que siguem autònoms a l'hora de triar un tipus de base de dades o un altre en funció de les nostres necessitats. Avui dia, hi ha moltes alternatives i tipus de sistemes compatibles, per la qual cosa és important que, a part de tenir un coneixement base sobre les diferents categories d'emmagatzematge que existeixen, puguem trobar un compromís entre rendiment, escalabilitat, disponibilitat i accessibilitat del nostre disseny.

Això ens permetrà tenir una visió global de per què es necessita el *big data*, com es dissenya un sistema d'aquestes característiques, quines parts té i com funciona. És només el començament per entendre como el *big data* ofereix autonomia, valor, intel·ligència i reducció de costos al negoci.

En el mòdul 2 tractarem:

- La tecnologia i el *big data*
  Es donaran a conèixer els començaments del *big data* i el paper que suposa per a les tecnologies actuals, que són bàsiques per comprendre les següents parts del mòdul.

- Característiques de les plataformes de *big data*
  Es presentaran les principals característiques de disseny i arquitectura. Es descriurà cadascuna de les seves capes, apreciant el flux de les dades des que es genera fins que s'explota.
- Infraestructura de bases de dades: clústers
  De quins elements es formen i en quines fases es pot dividir un clúster.
- Emmagatzematge i gestió de dades
  Introducció a les diferents solucions informàtiques per a l'emmagatzematge i la gestió de dades i els seus principals reptes.
- Resum d'idees clau
  Resum dels conceptes principals.

2.1 LA TECNOLOGIA I EL *BIG DATA*

Perquè puguem entendre què és el *big data* i la seva implicació tecnològica, cal fer un breu repàs de la seva evolució i el que ha suposat per als sistemes actuals, els quals han sabut solucionar el gran problema de l'augment exponencial de la generació d'informació fins avui.

El terme *big data* va ser emprat per primera vegada per la NASA a finals de la dècada de 1990. L'agència espacial ja va advertir aleshores del gran problema que estava suposant l'augment de les dades per als seus sistemes. No va ser fins anys més tard, el 2003, quan Google va publicar el seu sistema de fitxers GFS (Google FileSystem) i el processament MapReduce, que servirien de punt de partida per a la creació del que avui coneixem com a *big data* i que començaria amb el projecte Hadoop per part de Yahoo!

Aquest invent va suposar una gran revolució: Google va aconseguir indexar en el seu motor de cerca la informació de qualsevol web d'Internet gràcies als sistemes de fitxers distribuïts. D'aquesta manera, en lloc de processar molta informació en la mateixa màquina, es va arribar a la conclusió que era molt més ràpid emmagatzemar i processar la informació en diferents servidors de manera paral·lela, aconseguint un resultat molt més ràpid i amb uns recursos no gaire costosos.

El 2006, Yahoo! i la comunitat Open Source van agafar el testimoni amb el desenvolupament de l'ecosistema Hadoop, el primer gran projecte encarregat d'emmagatzemar, processar i analitzar grans volums de dades. Hadoop és un sistema de codi obert que permet distribuir fitxers de manera senzilla en diferents nodes, per poder executar una programació paral·lela sobre aquests. Es constitueix principalment de:

- Un sistema de fitxers distribuïts anomenat HDFS (Hadoop FileSystem), permetent emmagatzemar fitxers en diferents dispositius.
- Un *framework* de processament anomenat MapReduce, que permet aïllar el programador de totes les tasques de programació en paral·lel; és a dir, és el sistema Hadoop el que s'encarrega de buscar on és cada fitxer i com s'ha de tractar per aconseguir la solució de la manera més ràpida.

Des de llavors, els dissenys i arquitectures de *big data* no han parat d'evolucionar, adoptant diferents formes en plataformes com Cloudera o Hortonworks, que s'encarregarien d'agrupar eines d'emmagatzematge, processament i anàlisi de dades per facilitar a les empreses la seva inclusió en aquesta nova tecnologia. Tant va ser així que, el 2012, Obama va ser el primer candidat a unes eleccions presidencials a utilitzar models predictius amb la finalitat d'obtenir avantatge amb un marge suficient enfront dels seus rivals i, el 2014 i 2015, l'IoT i les Smart Cities connectarien el món, ajudant al monitoratge de control de qualitat de l'aire, atenció mèdica, supervisió de trànsit, il·luminació, etc.

Moltes personalitats expertes assenyalen que el 2020 es produirà un augment estimat del 4300 % en la generació de dades anual, per la qual cosa és essencial tenir en compte el *big data* i la seva implicació tecnològica, per poder afrontar el creixement del volum i la quantitat d'informació.

2.2 CARACTERÍSTIQUES DE LES PLATAFORMES DE *BIG DATA*

Hola de nou!

En aquest vídeo parlarem del disseny i la composició d'una arquitectura de *big data*. Una arquitectura pot estar formada per cinc components o fases principals: adquisició de dades, emmagatzematge, processament de dades, visualització i administració.

Vegem-los detalladament:

1. Adquisició de dades

Aquesta primera fase intentarà connectar els diferents orígens de dades amb el nostre sistema, per així poder emmagatzemar i analitzar la informació posteriorment. Existeixen dos mètodes principals per a la recol·lecció de la informació:

- D'una banda, *batch* o per lots. És una execució periòdica d'un procés. Buscarà en l'origen de dades si existeix informació recent des de l'última connexió realitzada. És comú en fonts com a sistemes de fitxers o bases de dades.
- D'altra banda, *streaming* o en temps real. Aquest tipus de recol·lecció es connecta amb l'origen i crea un flux continu de dades, de manera que la informació es genera i emmagatzema al mateix temps. És utilitzat, amb freqüència, en sistemes de monitoratge, detecció d'anomalies, anàlisi web o predicció de tendències.

2. Emmagatzematge

Entrant en la fase d'emmagatzematge, ens trobem que les bases de dades tradicionals no són suficients per gestionar volums tan grans d'informació. Per aquesta raó, trobem diferents solucions per guardar les dades adquirides. Vegem-les:

- Sistemes de fitxers distribuïts: intenten emmagatzemar fitxers en diversos servidors. Aquests fitxers es particionen i es repliquen en les diferents màquines, oferint capacitats d'alta disponibilitat i escalabilitat al nostre sistema. El sistema de fitxers original va ser Hadoop, tot i que ara es troben moltes alternatives en plataformes al núvol com S3 (AWS), Cloud Storage (Google Cloud) i Azure Storage.
- Bases de dades no relacionals. Permeten emmagatzemar informació de manera documental (format JSON o XML) o en forma de clau o valor. Estan capacitades per gestionar grans quantitats d'informació, consumeixen pocs recursos i el seu escalament és senzill. Podem destacar bases de dades com Redis, Cassandra, MongoDB o HBase.
- Bases de dades relacionals. Tot i que són poc flexibles, consumeixen més recursos que les anteriors i tenen capacitats d'escalatge pitjor. Són una solució essencial quan es requereixen operacions transaccionals.

3. Processament de dades

Una vegada que tenim la informació correctament emmagatzemada en el nostre sistema de fitxers o bases de dades, arriba el moment de processar i analitzar la informació. Com duem a terme aquesta tasca? Això dependrà del tipus d'anàlisi que vulguem dur a terme, però totes les eines comparteixen la característica de processament en paral·lel.

Entre d'altres, podem trobar els següents serveis:

- MapReduce. És l'eina de processament original de Hadoop. Es compon d'una primera fase per recuperar la informació (Map) i una altra per executar l'operació sol·licitada (Reduce), utilitzant la capacitat de cadascuna de les màquines del nostre sistema.
- Apache Spark. Motor de processament distribuït de codi obert. És el més utilitzat actualment, proporcionant un rendiment fins a 100 vegades més ràpid que MapReduce, pel fet de poder treballar en memòria.
- Apache Storm. Processa en temps real i de manera senzilla grans quantitats de dades.

4. Visualització

Aquesta última capa del cicle de la dada permet representar l'anàlisi duta a terme per la fase de processament. Per fer-ho, tenim un gran ventall de possibilitats, entre les quals trobem les següents categories:

- Notebooks
- Llibreries gràfiques de JavaScript.

- Eines d'anàlisi gràfica, entre les quals destaquen Kibana i Grafana.
- Eines propietàries, com Tableau, QlikView, MicroStrategy o Power BI.

5. Administració

Finalment, trobem aquesta capa transversal, que s'encarrega de gestionar i monitorar els recursos del nostre sistema. Això permet conèixer en tot moment l'estat dels nostres nodes i els serveis que estem utilitzant en cadascuna de les capes.

Com podem observar, una arquitectura *big data*, malgrat ser complexa, té fases ben definides. Això ens permetrà poder-les estudiar totes de manera individual amb molta facilitat.

Fins aviat!

2.3 INFRAESTRUCTURA DE BASES DE DADES: CLÚSTERS

Hola!

En aquest vídeo explicarem com es distribueix la informació al llarg dels servidors que componen un sistema de *big data*. En primer lloc, necessitem tenir clars tres conceptes bàsics:

- Un **clúster** és un grup de servidors que treballen de manera conjunta. Cadascun dels servidors proporciona emmagatzematge, capacitat de processament i gestió de recursos al sistema.
- Un **node** és un únic servidor en el clúster. N'existeixen de dos tipus: els nodes mestres gestionen la distribució de tasques i els nodes esclaus les executen.
- Un **dimoni** és un programa en execució en un dels nodes. Cadascun realitza diferents funcions en el clúster, com el monitoratge dels recursos o la planificació de tasques.

A continuació, explicarem els tres components que formen un clúster: la capa d'emmagatzematge, la de processament i la de gestió de recursos. Junts proveeixen al sistema capacitats de processament distribuït de la informació.

Començarem per la **capa d'emmagatzematge**. Per a l'explicació, ens basarem en com funciona HDFS, el sistema d'emmagatzematge de Hadoop, ja que és el sistema de fitxers distribuïts més usat. HDFS proveeix d'un servei d'emmagatzematge redundant per a quantitats massives de dades, sense necessitat d'utilitzar servidors amb recursos de maquinari excessiu.

Els fitxers es divideixen en blocs, per defecte, amb una grandària de 128 MB. Es distribueixen en temps de càrrega. Cada bloc es replica en múltiples nodes, permetent l'accés als fitxers, encara que tinguem una caiguda d'alguna de les màquines. En aquest cas, el node mestre es

diu Namenode i emmagatzema la metadada, és a dir, sap on està situada la partició de cada fitxer. Els nodes esclaus es diuen Datanodes i emmagatzemen les diferents particions dels fitxers.

Seguim amb **la gestió de recursos.** Per a què puguem treballar amb les dades emmagatzemades en clúster de manera paral·lela, cal un servei que s'encarregui de gestionar els recursos del processament distribuït de la informació. Per a això, tenim solucions com YARN (*Yet Another Resource Negotiator*), que és la capa de processament de Hadoop, que està formada per un gestor de recursos i un planificador de tasques.

En aquest cas, el node mestre es denomina ResourceManager (RM). S'encarrega d'organitzar els recursos de manera global al llarg del clúster. Els nodes esclaus s'anomenen NodeManager (NM). La seva funció principal és reservar recursos per a cadascuna de les aplicacions sol·licitades pel ResourceManager.

Finalment, parlarem del **processament**. En aquest cas, tenim *frameworks* de processament distribuïts com Spark, Storm o Flink, i tots funcionen de manera similar. En el cas de Spark, tenim un *entry point* denominat "Context de Spark " (*Spark context*), que utilitzarà el gestor de recursos triat (YARN, per exemple) per distribuir la informació al llarg del clúster. Spark treballa en memòria, la qual cosa permet realitzar les operacions de forma molt més ràpida que altres serveis que treballin en disc. A més, permet recuperar-se en cas de caiguda d'algun dels nodes

Una vegada analitzades les tres parts d'un clúster, vegem un exemple pràctic:

Tenim una empresa el sistema de la qual intenta gestionar l'emmagatzematge dels albarans de compra de les seves comandes. Vol saber, en temps real, el nombre total de comandes i els beneficis bruts diaris. Per a això, quan sol·licitem a Hadoop que insereixi cada albarà a HDFS, el que fa és dividir els fitxers en diversos blocs i guardar-los en diferents nodes del clúster, aconseguint que la informació estigui distribuïda. D'altra banda, el *framework* de processament utilitzarà cada bloc de manera paral·lela per realitzar els càlculs desitjats, obtenint així un funcionament molt més ràpid i eficient que qualsevol sistema tradicional.

Amb això ja podem veure amb claredat les bases d'un clúster de *big data*, de quines parts està format i com es gestiona. Fins al pròxim vídeo!

2.4 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.

2.5 IDEES CLAU: INFRAESTRUCTURA PEL *BIG DATA*

Vegem quins conceptes hem après en aquest segon mòdul.

Hem repassat l'evolució **històrica del *big data*** i com ha anat integrant-se cada vegada més en el dia a dia. Google pot presumir de ser la companyia que té la xarxa distribuïda més àmplia del món, a part de ser la pionera en introduir els conceptes d'arquitectures distribuïdes. Això va suposar un abans i un després que ajuda a les companyies actualment a emmagatzemar i processar les dades d'una manera ràpida, escalable i barata.

S'ha de tenir en compte com està estructurada una **arquitectura *big data***. Per a això, és interessant que tinguem clar quin és el flux natural de la dada. En primer lloc, l'origen de dades generarà els registres necessaris, ja sigui en format de text, o en base de dades o un sistema de cues, per després ser adquirit i emmagatzemat en el nostre repositori de dades.

Posteriorment, es processarà la informació, netejant registres buits, nuls o duplicats, validant formats i aplicant transformacions i models que s'encarreguin d'oferir valor al negoci. El resultat de tot aquest procés tractarà de donar resposta als indicadors de rendiment o KPI plantejats, així com oferir una visió de l'estat actual de la informació, tendències i prediccions.

Veiem que en tot moment estan presents les 4 V del *big data*: volum, velocitat, varietat i veracitat.

Per poder dur a terme tot el **procés d'emmagatzematge, transformació i visualització de la informació** cal donar-li els mitjans sobre els quals operar. Hem vist com màquines individuals ja no poden suportar el volum de dades que es genera actualment, per la qual cosa és necessari treballar amb clúster de servidors. Un clúster no és res més que un conjunt de màquines que treballen en conjunt per oferir una capacitat de reposició i processament major, a través de la distribució de la informació i treballs paral·lels. Aquest tipus d'arquitectures estan pensades per ser escalables horitzontalment, és a dir, podem afegir més màquines al nostre clúster de manera gairebé automàtica. També proporcionen característiques d'alta disponibilitat, perquè estan preparades per continuar funcionant, encara que algun dels nodes del clúster quedi fora de funcionament temporalment.

Finalment, hem repassat quines **possibilitats d'emmagatzematge de la informació** existeixen en funció de la seva categoria, estructura i funcionalitat. Triar el sistema de magatzematge és una de les decisions més importants que s'han de realitzar sobre el disseny d'un sistema de big *data*, per la qual cosa és fonamental saber les diferències entre un sistema de fitxers distribuïts i bases de dades relacionals i no relacionals.

3 *FRAMEWORKS*

Al llarg d'aquest mòdul veurem un dels tipus de serveis més importants en *big data:* els *frameworks*. Aprendrem què són, com estan formats, en quin llenguatge es programen, quines funcionalitats desenvolupen i sota quines circumstàncies exploten el seu potencial.

És important que ens familiaritzem amb aquest tipus de conceptes i que posem el focus en com funcionen i quants tipus de *frameworks* hi ha, més que aprendre l'especialització d'un en concret. Això es deu al fet que hi ha infinitat d'alternatives *open-source* disponibles per dur a terme la nostra aplicació i totes estan en contínua evolució, per la qual cosa és probable que les solucions ofertes avui siguin substituïdes demà per d'altres de més modernes i amb millor rendiment. Per aquesta raó, al llarg del mòdul es presentarà de manera clara i concisa les particularitats dels *frameworks* en funció de la teva tipologia, amb la finalitat que puguem entendre quin és el seu funcionament, independentment del seu nom.

Es posarà l'accent en els principals *frameworks* d'emmagatzematge i processament distribuïts que existeixen ara mateix al mercat: Hadoop, Spark i Storm. Amb ells, serem

capaços de complir les bases de les 4 V del *big data* fàcilment, perquè cadascun ofereix un conjunt d'eines molt accessible que ens abstreuen d'arquitectura i funcions de baix nivell, podent centrar-nos en el nucli de la nostra aplicació i donar valor a la informació del nostre sistema.

Estructura del mòdul:

- Què és un *framework*?
  Explicació introductòria de què són i com funcionen. Coneixerem què ens aporten a la nostra aplicació, de quines parts estan compostos i com es classifiquen.
- Principals *frameworks* utilitzats
  Comparació completa entre els *frameworks* més utilitzats: Hadoop, Storm i Spark. Aprendrem a quins escenaris s'adapta millor cadascun d'ells, quins són els seus forts i les seves característiques principals.
- Exemples d'utilització de *framework*.
  Presentació d'un cas pràctic real de la utilització de Hadoop i Spark. Es mostrarà una arquitectura *big data*, de manera que puguem tenir clar com es diregeixen les dades de diverses fonts de dades, com es netegen i transformen, i com s'analitzen a través dels diferents *frameworks* de processament.
- Exercici per repassar l'après en el mòdul
  En aquest exercici, s'haurà de decidir quin *framework* és el més adequat per a cadascun dels casos d'ús que es mostren. Es posaran en pràctica els coneixements adquirits al mòdul.
- Resum d'idees clau del mòdul
  Es realitzarà un petit resum de les característiques principals, l'arquitectura i els objectius dels *frameworks* que s'han estudiat en el curs.

3.1 QUÈ ÉS UN FRAMEWORK?

Hola de nou! En aquest vídeo explicarem què és un *framework*, per què hem d'utilitzar-lo i alguns exemples que ens seran útils de cara a plantejar els nostres futurs desenvolupaments. Comencem!

Què és un *framework*? Un *framework* no és res més que un marc de treball, és a dir, un conjunt d'eines, convencions, estàndards i bones pràctiques que ens faran la vida més fàcil a l'hora de crear la nostra aplicació. En general, proporcionen funcionalitats complexes que eviten que dediquem molt temps a implementar tasques repetitives o de baix nivell. D'aquesta manera, podem focalitzar els nostres esforços a aportar valor a l'aplicació.

Per tant, podem dir que un *framework* ens ajudarà a complir els objectius següents:

- **Construcció del disseny de l'estructura bàsica per al programari desenvolupat**. Ofereix una sèrie de funcions, classes i objectes que seran utilitzats com un patró de disseny pel programador o programadora.
- **Funcionalitats bàsiques**. El *framework* és el que ofereix totes les funcionalitats de baix nivell. D'aquesta manera, no hem de reinventar la roda i crear de nou totes les singularitats fonamentals. Així evitarem crear una vegada i una altra connexions amb una base de dades, la paginació del nostre lloc web o escriure funcions de processament de text.
- **Reutilització**. Una de les característiques principals d'un *framework* és facilitar la creació de blocs de codi que permetin ser utilitzats en diferents punts de la nostra aplicació de manera comuna. S'ha d'evitar, en la mesura del possible, escriure la mateixa funcionalitat més d'una vegada al llarg del nostre programa.
- **Augment de productivitat**. La nostra principal dedicació serà la de crear pròpiament l'aplicació. No haurem de preocupar-nos ni de l'arquitectura ni de com interactuen i es distribueixen les dades en el nostre servidor o clúster. Això permet, al seu torn, que puguem migrar entre *frameworks* fàcilment.
- **Afavorir el treball en equip**. El fet de tenir una estructura comuna sobre la qual treballar permet que el codi implementat sigui més llegible. Això donarà l'oportunitat a altres persones d'entendre el funcionament del desenvolupament, de manera que puguin estendre la seva funcionalitat o corregir algun defecte de la seva operativa.
- **Bones pràctiques**. Un *framework* ofereix uns estàndards sobre els quals haurem de desplegar el nostre desenvolupament. En la majoria de casos, un bon seguiment de les bones pràctiques marcades per la comunitat permet millorar considerablement el rendiment de la nostra aplicació, així com la seva reutilització i interpretació.

Quant a la composició i arquitectura d'un *framework*, podem dir que està format per tres capes diferents, que són:

- **Infraestructura**. Defineix tasques de comunicació de xarxa, computació i emmagatzematge. Assegura, de manera transparent, a la persona usuària fet que dades de diferents formats puguin ser emmagatzemades i transferides de manera eficient, segura i escalable al nostre sistema. Comparteix objectiu principal amb les V del *big data*. Aquesta capa proveeix propietats per gestionar quantitats massives d'informació de manera escalable al costat del creixement de l'organització. Intenta optimitzar, en la mesura del possible, l'IOPS (operacions d'entrada i sortida per segon), per assegurar que la taxa de transferència i processament sigui l'adequada a les nostres necessitats.
- **Plataforma**. És la col·lecció de funcions que faciliten processar les dades amb un bon rendiment. La plataforma inclou capacitats per integrar, gestionar i aplicar tasques de processament de la informació. En entorns de *big data*, això significa que la plataforma necessita facilitar i gestionar solucions per al processament i emmagatzematge distribuït.
- **Processament**. Aquesta capa s'encarrega d'oferir funcionalitats d'accés a la informació. Les tasques executades operen sobre els *datasets* emmagatzemats de manera

distribuïda. Aquí és on resideixen les tasques de neteja, transformació, anàlisi de les dades i execució dels algorismes. Aquest processament oferirà els resultats requerits i que aporten el valor al negoci.

Finalment, classificarem els *frameworks* de processament en funció de la seva tipologia:

- *Frameworks* de processament *batch*. Són aquells que fan tasques planificades per al tractament de la informació. Aquí destaquen les operacions MapReduce d'Apache Hadoop.
- *Frameworks* de processament *streaming*. Marcs de treball que gestionen operacions de la dada amb fluxos en temps real, de manera que la informació és processada quan entra en el nostre sistema. Els principals *frameworks* d'aquest tipus són Apache Storm i Apache Samza.
- *Frameworks* híbrids. Aquests últims destaquen per funcionar tant en *batch* com en *streaming*. Apache Spark i Apache Flink són els més adequats per a aquest tipus d'escenaris.

Aquests serien els punts bàsics que tindrem en compte a l'hora de caracteritzar i triar un *framework* de *big data*. Els continuarem estudiant en els pròxims apartats del curs. Fins aviat!

3.2 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:

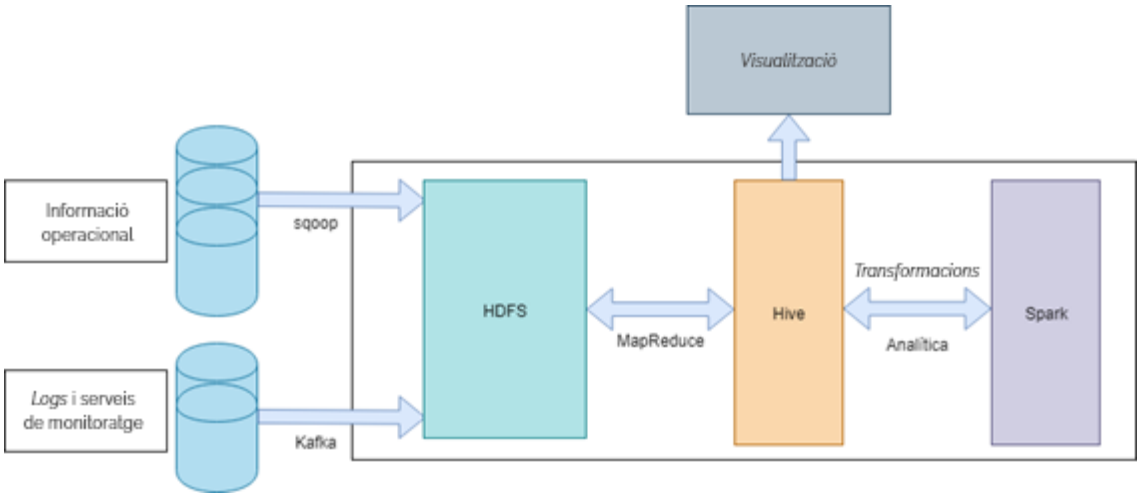
- Processament streaming: Storm ofereix millor rendiment pel fet que utilitza una metodologia micro-batch, és a dir, comprova el flux d'entrada amb més freqüència que Spark.
- Llenguatges de programació: Storm suporta més llenguatges que Spark.
- Latència: Storm proporciona millor latència amb menys 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.

3.3 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.

3.4 IDEES CLAU: *FRAMEWORKS*

Una vegada finalitzat el mòdul, és important que repassem els punts més importants.

Hem definit un ***framework*** com un conjunt d'eines i estàndards que ajuden a l'enginyeria de dades a dur a terme tasques amb molta més rapidesa i de manera òptima. Implementa metodologies i bones pràctiques per evitar que es cometin errors a l'hora de crear vincles recursius, accessos recurrents a la informació o, en general, una gestió amb un mal rendiment de les estructures de dades distribuïdes.

Un dels aspectes més importants d'un *framework* és la infraestructura que proveeix al nostre sistema. Amb una configuració adequada, el *framework* s'encarrega de dividir, distribuir i replicar els paquets al clúster, despreocupant-nos de tota la capa de baix nivell del sistema. Aquests automatismes estalvien molts costos a les empreses, perquè la instal·lació d'un ***framework*** és ràpida, l'ús és senzill i el rendiment és òptim.

S'ha vist com Hadoop és un dels *frameworks* més complets, ja que disposa de quatre capes. La d'emmagatzematge està basada en HDFS. El gestor de recursos, clau per a l'orquestració

de tasques entre mestres i esclaus, corresponent a YARN. La fase de processament s'associa amb operacions de MapReduce. I, finalment, es tenen totes les eines de suport al desenvolupament de les aplicacions i a la interconnexió de serveis de l'ecosistema de Hadoop.

A més de Hadoop, hem estudiat altres *frameworks* de processament com:

- Spark (processament streaming i híbrid): a més de tenir un component molt fort en el processament streaming, és molt més complet que Storm, perquè disposa d'eines d'analítica avançada, aplicació de models de machine learning i representació de grafs.
- Storm (principalment per a processament en temps real): Apache Storm és un framework exclusiu de tractament d'streams, parcel·la en la qual supera a Spark.

Finalment, hem introduït un exemple on es va representar una arquitectura completa d'un sistema de *big data* clàssic, amb les capes d'adquisició, emmagatzematge, processament i visualització de la informació.

Abans de tancar el mòdul, s'ha de destacar una particularitat dels *frameworks*, la seva modularitat. Pot ser que el dia de demà siguin uns altres els que substitueixin els que s'han estudiat en aquest mòdul, però tots solen estar preparats per oferir una compatibilitat suficient perquè la migració entre tecnologies no sigui massa costosa.

És per això que el coneixement més important a tenir en compte és el de quines necessitats hem de cobrir (sistemes distribuïts, alta disponibilitat, replicació, processament *batch*, temps real, analítica avançada, etc.) i quins tipus de *frameworks* existeixen en el mercat que puguin adaptar-se als nostres requisits.

4 VISUALITZACIÓ DE DADES: PROGRAMES I METODOLOGIES

En aquest mòdul donarem a conèixer algunes eines de *business intelligence* (BI) que han estat integrades dins de l'ecosistema ***big data***. Aquestes eines són útils per al negoci en la presa de decisions i en la generació de quadres de comandament, que ajudaran a comprendre l'anàlisi realitzada i la informació de negoci utilitzada en el pla estratègic de l'empresa. Repassarem serveis com Tableau, QlikView, Power BI, Kibana i Grafana.

A continuació, presentarem una sèrie de casos d'ús reals d'un disseny ***big data***. Són escenaris que s'apliquen en el nostre dia a dia i que ens ajuden a automatitzar tasques, recomanar-nos continguts personalitzats, a detectar fraus o a optimitzar els preus. Això ens ajudarà a comprendre la importància d'aquesta tecnologia i l'impacte que està tenint en la nostra societat.

Finalment, revisarem quines són les tendències de les eines de *big data* en els pròxims anys. Respondrem a preguntes com "quina serà l'evolució de les plataformes *cloud*?", "quin tipus de processament té un major recorregut?", "sobre quina analítica es focalitzaran els esforços?" o "quina fase del flux de la dada serà la més reforçada?".

Finalitzarem el mòdul amb un breu test sobre els continguts que s'han vist en el curs.

Estructura del mòdul:

- Visualització de dades
  Importància i paper que juga en *big data* la visualització de dades.
- Metodologies i programes
  Diferents metodologies: els seus usos i avantatges. Els principals programes, les metodologies que utilitza cadascun i les seves principals característiques: Tableau, QlikView, Power BI, Kibana i Grafana.
- Casos d'ús de *big data*
  Exemples reals de l'ús de *big data.*
- Tendències
  Tendències en l'ús de *big data* en els pròxims anys.
- Resum d'idees clau
  Resum d'idees clau del mòdul.
- Test sobre els continguts del curs
  10 preguntes tipus test sobre els conceptes treballats en el curs.

4.1 VISUALITZACIÓ DE DADES

Hola!

En aquesta sessió analitzarem la importància que té la visualització de dades en el món del *big data.* Qualsevol companyia que es preocupi per la qualitat de la dada i per la viabilitat del seu negoci ha de recórrer a aquest tipus d'eines.

Per què?

Molt senzill: per a les empreses és molt important saber en tot moment tant l'estat financer actual de l'organització, com la tendència de mercat en la pròxima temporada. Necessiten mètriques i idees disruptives per millorar el seu negoci. I tot això amb l'ús d'eines molt senzilles i visuals, accessibles a empreses de mida gran i petita, gràcies a les seves possibilitats d'escalabilitat.

Amb aquestes solucions de programari, les empreses ara poden donar sentit a complexos conjunts de dades de *big data* sense massa maldecaps. Aquestes solucions de *business intelligence* poden recollir, analitzar i convertir les dades en informes i quadres de comandament comprensibles per a la persona que administra els sistemes, un o una enginyera de la dada o un executiu o executiva. L'objectiu no és un altre que donar valor a la

informació emmagatzemada en els nostres sistemes i convertir-ho en beneficis per a l'empresa. Com podem aconseguir-ho? Dependrà del cas d'ús en el qual ens trobem. Per exemple:

- **Administració de sistemes.** Podem estalviar cost de màquines en la companyia si presentem un informe amb el dimensionament i ús dels servidors corporatius. Si es demostra que la capacitat està sobredimensionada, es pot recondicionar l'arquitectura perquè s'ajusti més a les nostres necessitats i així evitar pagar per recursos que no s'utilitzen.
- **Llicenciament.** Les empreses inverteixen molts diners en el llicenciament de programari d'ofimàtica, processament algorítmic, eines de *big data*, etc. Fent ús de serveis de visualització de dades és molt fàcil apreciar quin és l'ús que s'està donant a les llicències adquirides i si realment s'estan amortitzant o si és necessària l'adquisició de més quantitat.
- **Gestió de recursos.** Podem tenir el control en tot moment i, a simple vista, de l'administració de l'estoc de productes, de la gestió de la cartera de clientela i plantilla, del registre de projectes, etc. Els diferents quadres de comandament poden representar tant informació històrica, com en temps real, així com anàlisis predictives de l'estat futur dels recursos.
- **Tendències de mercat.** Gràcies a l'anàlisi prèvia que han realitzat els nostres models analítics de *big data*, som capaços de reproduir quin serà l'estat del mercat sobre la base de la nostra experiència passada i les necessitats actuals de la clientela. Aquest és un dels objectius de qualsevol empresa, ja que poder anticipar-se a la competència pot suposar un èxit en l'àmbit econòmic, en visibilitat i en captació de talent.

Aleshores, per què utilitzar eines de *business intelligence* en el nostre negoci? Perquè, tal com hem vist en els exemples, els beneficis d'aquest tipus d'eines superen amb escreix les inversions que comporten. Poden ajudar a les empreses a obtenir informació de gran valor que ajudi el creixement corporatiu, a resoldre inquietuds de negoci, a recopilar dades de màrqueting més ràpidament, a proporcionar una vista en temps real de l'organització i a permetre l'anticipació de resultats futurs utilitzant anàlisis predictives.

Cada vegada són més les empreses que utilitzen aquests serveis per al seu creixement i per això el mercat d'aquest tipus de solucions està en plena expansió. De fet, el mercat global del programari de BI (*business intelligence*) oferirà un creixement anual del 7,1 % fins al 2025. S'espera que els guanys aconsegueixin els 26 bilions de dòlars l'any 2021, pels 16,5 bilions actuals. Aquesta expansió està relacionada amb l'evolució tecnològica del ***big data***, que s'està desenvolupant en l'última dècada i que seguirà en progressió en els pròxims anys.

Les noves tendències de programari de BI han proporcionat noves capacitats a les organitzacions. El descobriment de dades, que solia ser el territori dels experts i expertes en anàlisi avançada, ara es fa més fàcil amb aquestes plataformes. Això s'aconsegueix a través de l'anàlisi visual, la qual cosa permet a les persones responsables en la presa de decisions accedir i actuar immediatament sobre les dades. Potser una de les tendències més importants en les solucions de BI és la seva provisió de suport mòbil, la seva compatibilitat

multiplataforma i la seva implementació al núvol, la qual cosa permet a les persones usuàries accedir i analitzar informació des de qualsevol dispositiu.

Creiem que sobren els motius per a què les empreses s'adhereixin a aquesta tecnologia: és el present i el futur per al creixement de qualsevol organització.

Fins ara!

4.2 METODOLOGIES I PROGRAMES

En aquest article oferirem una visió àmplia sobre les tècniques de visualització de dades, els factors que afecten l'elecció d'una gràfica i una anàlisi de les principals eines del mercat.

Què determina l'elecció d'una visualització de dades?

Una visualització és l'eina per donar sentit a la dada. Per presentar la informació i les seves correlacions de la manera més senzilla, els i les analistes utilitzen diverses tècniques: gràfics, diagrames, mapes, etc. Triar la millor tècnica i la seva disposició és el camí per fer que la dada sigui accessible a qualsevol perfil. I també al contrari, una mala representació de la informació pot comportar no explotar correctament la dada o fer-la irrellevant. Per això, es presenten cinc factors que influencien a l'hora de seleccionar una visualització:

- Públic. És molt important ajustar-se a l'audiència objectiu. Si es busca oferir una dada agregada a un client o clienta final, potser, amb la visualització senzilla, és més que suficient. Si, per contra, la representació va enfocada a algú que treballa en enginyeria de la dada, és possible que calgui oferir visualitzacions i diagrames amb més detall.
- Contingut. El tipus de dada determina la tècnica triada. Per exemple, si es vol representar una sèrie temporal, l'ideal és utilitzar un gràfic de línia, però, si volem comparar els elements d'un atribut concret, és convenient utilitzar gràfics de barres.
- Context. S'haurien d'utilitzar diferents aproximacions en funció del context. Segons l'element que s'estigui estudiant, serà important jugar amb la combinació de colors, contrastos i ombres correctes. Aquesta característica és necessària per poder mostrar diferents gràfiques superposades o que comparteixin la mateixa representació, permetent dibuixar més d'una mètrica d'un atribut amb claredat.
- Dinamisme. Cada tipus de dada i el seu nivell d'agregació influeixen a l'hora de triar la visualització. Per exemple, una sèrie temporal que s'actualitza en temps real i té un detall de segon no tindrà el mateix aspecte que una gràfica històrica amb informació mensual.
- Propòsit. La manera en la qual estan implementades les diferents visualitzacions també provoca un impacte en el missatge i objectiu del quadre de comandaments ofert. D'aquesta manera, la persona usuària que necessiti saber els KPI del seu negoci utilitzarà un quadre de comandaments senzill, que ha de representar, a simple vista i de manera senzilla, unes mètriques representatives de l'estat dels objectius de la companyia. D'altra banda, per poder crear una anàlisi complexa de la informació, pot

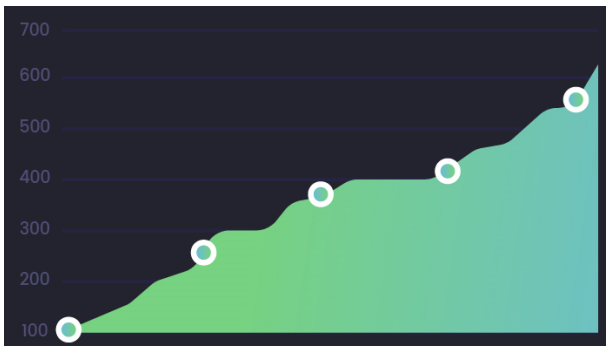
interessar afegir diferents tipus de representacions amb filtres i elements de control.

Tipus de representació

Tenint en compte aquests cinc factors, es tria entre diferents tipus de visualitzacions possibles. Aquestes són les més comunes (en el següent recurs es pot aprofundir sobre la visualització de dades http://atenciociutadana.gencat.cat/web/.content/manuals/guia-visualitzacio-dades.pdf):

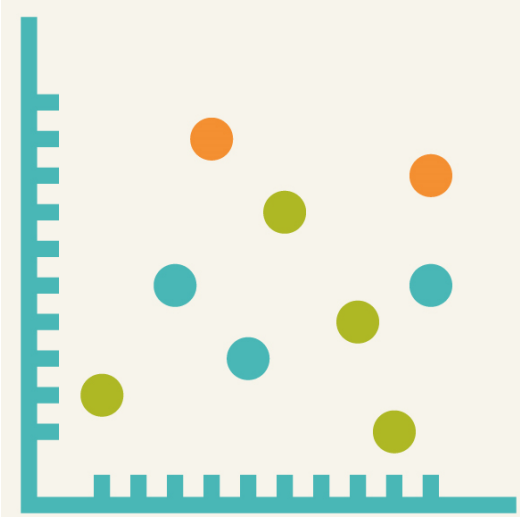
Gràfics

La manera més senzilla de mostrar un conjunt de dades és a través d'un gràfic. Poden variar entre línies i barres, que mostren una relació entre elements al llarg del temps i un pastís, que representa els elements d'un atribut de manera proporcional.



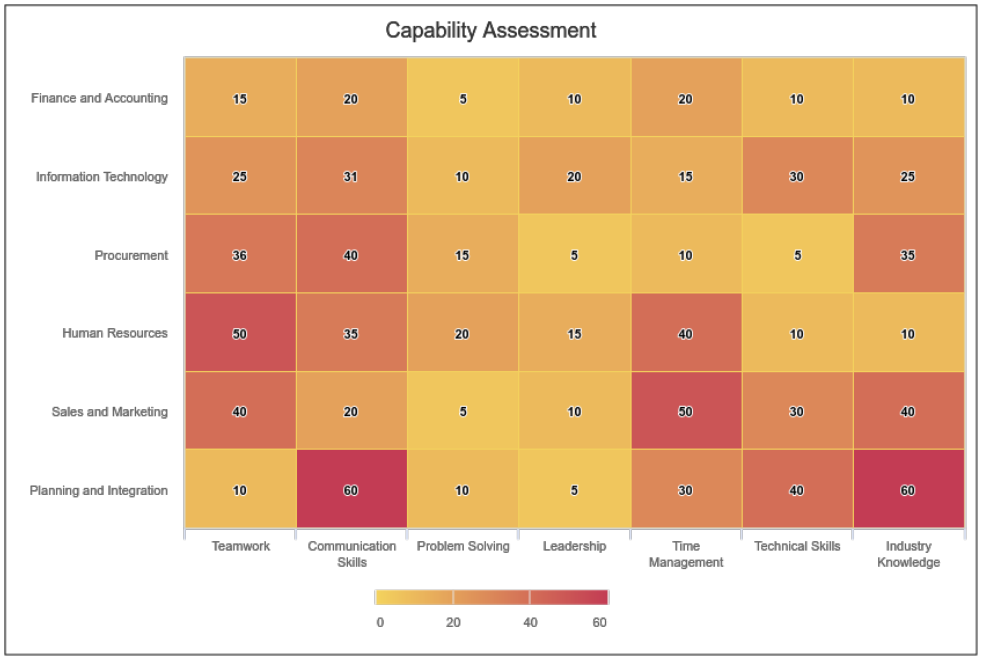
Dispersions

Permeten distribuir dos o més conjunts de dades a través d'un espai bidimensional o tridimensional. Mostren la correlació entre els conjunts de dades i la dispersió de les seves variables. Les dispersions més comunes solen tenir la forma de gràfic de bombolles o gràfic de punts (també conegudes com a parcel·la XY o simplement dispersió).



Mapes de calor

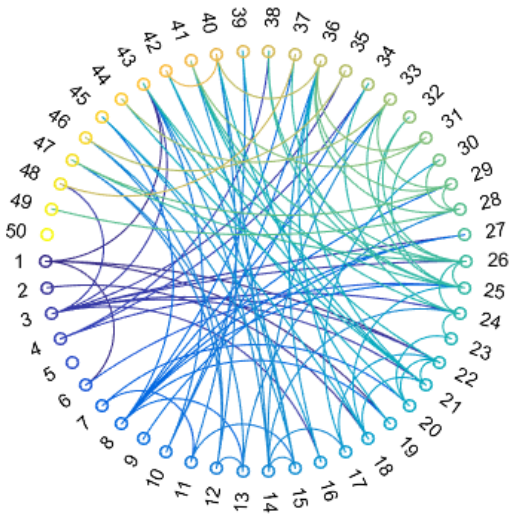
Permeten el posicionament d'elements sobre objectes rellevants o àrees, utilitzat en la creació de sinòptics, mapes geogràfics, plans, el *layout* d'una pàgina web, etc.



Diagrames i matrius

Els diagrames se solen utilitzar per demostrar relacions complexes entre les dades. N'hi ha de diversos tipus, entre els més usats: gràfics d'herència, multidimensionals o d'arbre.

Les matrius són una visualització típica del *big data*, que permeten mostrar la correlació entre múltiples *data sets* que evolucionen en temps real.



Eines de visualització

Tableau (https://www.tableau.com)

És una de les líders en aquest camp, ja que pot ser usada tant per analistes com per l'usuari o usuària final. Destaca per la seva senzilla interfície i multitud de llibreries de visualitzacions interactives. Inclou un ampli catàleg d'integració amb altres plataformes com bases de dades SQL, Hadoop o Amazon Web Services. El seu ús pot estar enfocat a petites visualitzacions ocasionals o a anàlisis exhaustives de les dades. Pot gestionar tant *streaming* de big *data* com informació estàtica.

QlikView (https://www.qlik.com)

És la major competència de Tableau . Tot i que, en general, ambdues ofereixen unes característiques bastant semblants, com a principal diferència caldria assenyalar que Qlik destaca pel seu rendiment, encara que els seus quadres de comandament són principalment estàtics. Encara no ofereix la possibilitat de crear visualitzacions en temps real.

Power BI (https://powerbi.microsoft.com)

Aquesta eina està més enfocada a oferir visualitzacions més complexes i analítica avançada. És excepcional, gràcies a la seva facilitat d'ús i interfície de *drag and drop* (agafar i deixar anar). És molt intuïtiva i disposa de moltes capacitats d'integració amb altres plataformes, especialment amb tots els serveis de Microsoft.

Permet crear informes amb visualitzacions de diferents orígens de dades alhora i també és compatible amb fonts en *streaming*.

Kibana (https://www.elastic.co/products/kibana)

Pertany a l'*stack* d'ElasticSearch i només pot treballar amb les dades d'aquesta base de dades. Per contra, pot ser la millor eina de visualització de *logs* del mercat. Ofereix moltes possibilitats d'analítica avançada, representació de grafs, generació d'alarmes de manera automàtica i utilització de models de *machine learning* dins de la mateixa eina i de manera interactiva.

Grafana (https://grafana.com)

És una de les eines de visualització de *big data* i IoT més populars, per ser *open source* i oferir un rendiment molt bo a l'hora d'explotar informació en temps real. S'integra amb més de trenta fonts diferents, incloent AWS i ElasticSearch.

Genera taulers dinàmics, permetent representar al mateix temps informació de diferents orígens de dades i amb múltiples mètriques diferents. També facilita la creació d'alertes i notificacions en funció de regles predefinides.

4.3 CASOS D'ÚS DE *BIG DATA*

Ara que ja es té una visió global i completa de com funciona i de per quines peces està format un disseny *big data*, podem representar una sèrie de casos d'ús reals que s'apliquen avui dia en el nostre entorn més pròxim:

Visió 360º de la clientela

Actualment, qualsevol empresa proveïdora de serveis ha de tenir la informació més detallada possible sobre la seva clientela. La informació s'origina en múltiples fonts, des de dades proporcionades pel mateix client o clienta, passant per l'ús i la geolocalització que realitza la persona usuària des del seu dispositiu mòbil, navegació web, converses telefòniques i intercanvi de correus electrònics amb la companyia, fins a l'encreuament amb la tendència de mercat per agrupar-lo al costat d'altres usuaris i usuàries afins, així com el contingut publicat en xarxes socials.

Tota aquesta informació permet a les companyies emprar models analítics i d'aprenentatge automàtic per poder personalitzar la interacció amb la clientela. D'aquesta manera, la companyia pot suggerir una nova estratègia d'estalvi al client o clienta, productes que li poden interessar, descomptes que pot aplicar a la seva factura o identificar aquelles persones amb més risc de causar baixa de l'empresa.

Prevenció de frau

Els sistemes de prevenció i detecció de frau amb *big data* són molt utilitzats pels bancs per alertar els seus clients i clientes de què les seves targetes estan essent utilitzades de manera fraudulenta, podent bloquejar-les de manera automàtica si es tracta d'un atac clar. Això és senzill de detectar: l'alarma s'activa quan es duen a terme diversos pagaments en diferents ubicacions en un curt període de temps o quan s'ha realitzat una compra en una ubicació molt llunyana al domicili de la persona, sense indicis que hagi realitzat cap viatge en la data indicada.

En els últims anys, aquests sistemes s'han tornat més sofisticats i han inclòs moltes millores per prevenir aquest tipus d'incidents. Per exemple, es pot associar un factor de risc a cadascuna de les compres realitzades a través de la targeta bancària, de manera que es puguin detectar possibles compres fraudulentes que se surtin del patró d'ús del client o clienta. A més d'això, els models de predicció van aprenent a mesura que succeeixen casos nous, amb la finalitat de prevenir diferents escenaris de frau. D'aquesta manera, per exemple, es poden ubicar codis postals o zones on l'índex de criminalitat és més alt.

Optimització de preus

Els negocis B2C (*business-to-consumer* o negoci a consumidor) o B2B (*business-to-business* o negoci a negoci) utilitzen tecnologies i analítiques ***big data*** per optimitzar el preu dels seus productes. Per a qualsevol companyia, l'objectiu és establir un preu als seus productes o serveis de manera que puguin maximitzar els beneficis aconseguits. Si el preu és massa alt, vendran menys i generaran pitjors beneficis. En cas contrari, si el preu és massa baix, el negoci no serà rendible.

En funció de l'històric de preus, transaccions i la resta de condicions del mercat, aquestes companyies són capaces ara d'establir un preu automàtic en funció de diverses estratègies. Les solucions *big data* poden, a més, segmentar la clientela i oferir diferents alternatives al servei ofert tenint en compte les necessitats de cada persona, posició geogràfica, grups d'edat, estatus social, etcètera.

Motors de recomanació

És un dels casos d'ús més populars, aplicat per totes les plataformes de reproducció de contingut *streaming*. Utilitzant l'historial de pel·lícules, sèries o cançons reproduïdes amb anterioritat, el sistema és capaç de recomanar contingut afí. Aquests algorismes també són utilitzats per pàgines de venda *online*, per recomanar productes semblants o del gust de la persona usuària, o per motors de cerca per oferir la publicitat més adequada a cada perfil.

4.4 TENDÈNCIES

Hola! Ara que ja coneixem l'estat actual de l'ecosistema ***big data*** i els seus usos més habituals, és hora de repassar el que hauria d'oferir aquesta tecnologia en el futur més pròxim. Som-hi!

Cloud i *big data*

Cada vegada són més les empreses que trien el *cloud* per poder emmagatzemar, transformar i analitzar la seva informació amb baix cost i de manera ràpida. Això no vol dir que totes les companyies utilitzin el núvol com a sistema principal. Els núvols híbrids són i semblen ser la solució ideal en molts escenaris. Els processos *batch*, certes automatitzacions i informació sensible continuaran executant-se en entorns *on premise*, mentre que gran part del processament en temps real i l'avaluació de models d'aprenentatge automàtic es processarà al núvol.

Els beneficis de l'ús de la computació al núvol sobre arquitectures *big data* són inqüestionables. Les companyies són capaces d'establir una arquitectura escalable i distribuïda ràpidament, reduint costos en adquisició de màquines i en el manteniment d'un *data center*. Per contra, l'èxit es troba en l'equilibri. No sempre interessarà la inversió de portar-nos tota la informació i processos al núvol. Cal recordar que, en molts casos, cada

proveïdor –vegeu Microsoft Azure, Amazon Web Services o Google Cloud– imposa els seus serveis propietaris, condicionant al fet que totes les aplicacions i desenvolupaments que es despleguin a partir de llavors es duguin a terme exclusivament sobre la seva plataforma.

La importància de l'analítica en temps real

Les fonts d'informació que proveeixen els sistemes de fluxos *streaming* i a la seva analítica seran cada vegada més demandades per les organitzacions. Aquestes dades solen oferir informació d'origen web, xarxes socials, dispositius mòbils, servidors o informació de la xarxa corporativa. Les analítiques *streaming* són el plat fort de les plataformes *cloud*. Permeten capturar la informació en temps real, ajudant a la companyia pe tal que la presa de decisions sigui ràpida i precisa.

Els casos d'ús més interessants es troben en el manteniment predictiu, ciberseguretat, optimització d'operacions, detecció de frau o aplicació de regles sobre el mercat borsari. L'evolució natural de la tecnologia provocarà que cada vegada sigui més senzill i eficient implantar models predictius i analítica *streaming* a les companyies, per la qual cosa en els pròxims anys serà molt comú veure aquest tipus d'operatives en el nostre entorn més pròxim.

Creixement del *machine learning*

Els pròxims anys ajudaran al fet que la convergència entre l'analítica tradicional i els algorismes d'aprenentatge automàtic sigui molt més integrada. Veurem més organitzacions utilitzant *machine learning* per millorar les activitats de negoci.

Fins ara, les companyies tenien equips diferents de data *science* per a avaluar els models i equips d'enginyeria de la dada per adquirir i transformar la informació. Aquests perfils professionals s'estan convertint en conjunts mixtos que seran capaços de dur a terme el flux complet de la dada amb més naturalitat.

De fet, existeixen dues tendències que estan accelerant l'ús de models de *machine learning*. La primera és que la barrera d'entrada cada vegada és menor, qualsevol persona amb coneixements de programació bàsics pot executar un model en uns senzills passos. La segona és que cada vegada són més les eines automàtiques que permeten productivitzar aquest tipus de models. Actualment, un *data science* que crea un bon algorisme d'aprenentatge automàtic ha d'esperar que l'enginyer o enginyera de la dada estableixi el flux necessari per portar-lo a producció. En els pròxims anys, els *frameworks* d'automatització permetran als *data science* realitzar aquestes operacions de manera autònoma.

Narrativa de dades i visualització

A mesura que els *data warehouses* es consoliden de manera més ordenada i optimitzada al *cloud*, es produirà una millora inevitable en la qualitat dels resultats oferts per l'analítica *big data*. Això es traduirà en què cada vegada serà més fàcil narrar històries rellevants i precises a través de quadres de comandament avançats. Això, juntament a què els models d'aprenentatge automàtic aportaran una visió de classificació, clusterització i predicció molt més precisa, serem capaces d'oferir molt més valor a la dada de la companyia, generant majors beneficis amb menys esforç.

Aquestes serien les tendències que hem de tenir en compte en els pròxims anys. Som conscients que l'evolució d'aquestes tecnologies és exponencial i hem d'estar preparats per adaptar-nos a les pròximes innovacions que ens oferiran els sistemes *big data*. Les possibilitats per millorar el nostre dia a dia són infinites!

4.5 IDEES CLAU: VISUALITZACIÓ DE DADES: PROGRAMES I METODOLOGIES

Una vegada finalitzat el mòdul, passarem a repassar els punts més importants que hem de tenir en compte.

Visualització de dades

Aquestes eines ens permeten oferir una capa de visualització a les dades emmagatzemades en el nostre *data warehouse*. Es caracteritzen per la seva facilitat d'ús, per la seva integració amb qualsevol tipus de base de dades o sistema de fitxers distribuïts, per la seva capacitat de representació de processament *batch* o temps real i per la gran varietat de recursos de què disposa a l'hora de generar quadres de comandament complexos.

Són fonamentals per a qualsevol negoci, perquè ajuden en la presa de decisions, a proporcionar l'estat actual de la informació corporativa, en la detecció de possibles amenaces o incidències i en l'anticipació de resultats. A més, aquests serveis estan evolucionant ràpidament per poder oferir solucions analítiques integrades amb l'eina, de manera que puguem mostrar alternatives de *data discovery* i prediccions automàtiques en temps de visualització.

És important recordar els cinc factors que s'han de tenir en compte a l'hora de dissenyar una visualització de dades: a quin públic està dirigida, el tipus de contingut que es vol oferir, el seu context, el dinamisme dels gràfics i el propòsit o missatge que es vol presentar.

D'altra banda, entre les eines de visualització més importants del mercat trobem: Tableau, QlikView, Power BI, Kibana i Grafana. Totes tenen els seus punts forts i haurem de triar-les acuradament en funció del nostre pressupost i necessitats.

Casos d'ús

D'entre tots els escenaris més típics dins de l'ecosistema *big data*, en destaquem quatre: visió 360º de la clientela, prevenció de frau, optimització de preus i motors de recomanació. Tots es basen en arquitectures *big data* distribuïdes i fan ús dels *frameworks* que s'han vist en el curs.

És fonamental tenir en compte la repercussió que estan tenint aquest tipus de tecnologies en el nostre dia a dia. Gràcies a l'ús de dispositius electrònics, plataformes al núvol, xarxes socials, adquisició de productes en comerços, utilització d'eines financeres, etcètera, generem una quantitat de dades que són utilitzades per empreses grans o proveïdores de serveis en les seves analítiques de dades.

Tendències

Per acabar, hem vist quina serà l'evolució natural de *big data* en els pròxims anys. S'espera que totes les eines que s'han mostrat al llarg del curs segueixin el seu desenvolupament cap a un millor rendiment i una millor economia de recursos. Algun dels punts més importants del *roadmap* seria el disseny d'arquitectures en núvols híbrids, la focalització en processament de fluxos *streaming*, la millora i creixement dels algorismes d'aprenentatge automàtic i el progrés en la visualització de dades.

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

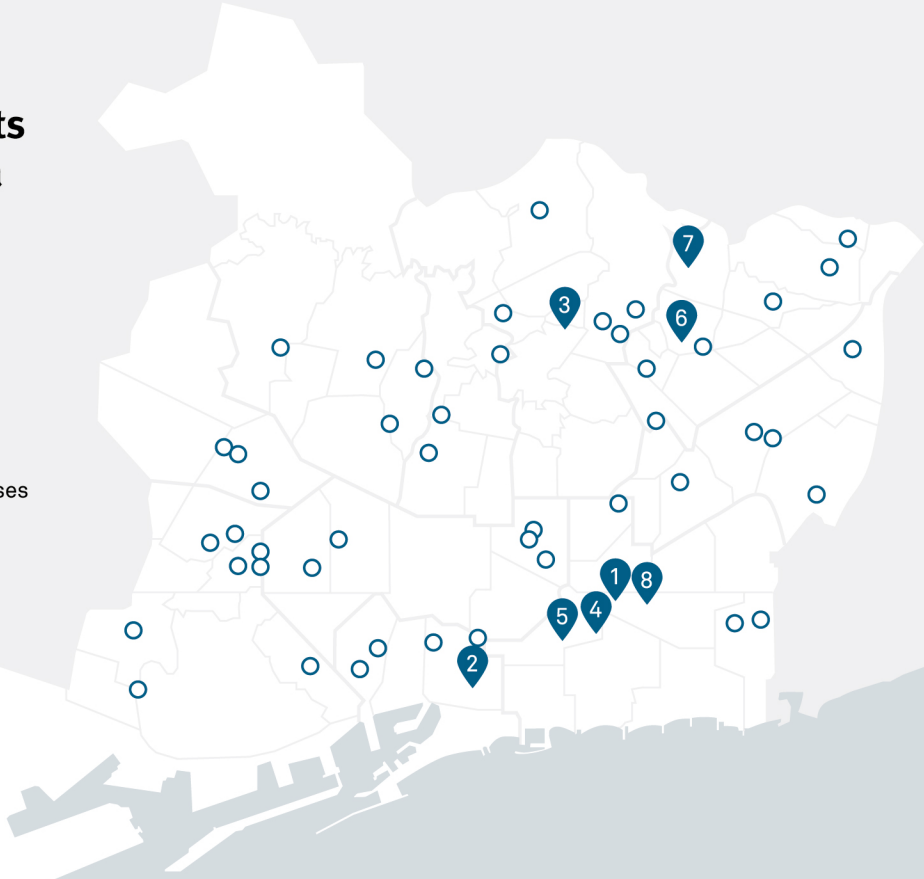
Formació tecnològica i gratuïta per a la ciutadania

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