Dades massives i mineria de dades socials, conceptes i eines bàsiques

PID_00275694

Jordi Morales i Gras

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Jordi Morales i Gras

Doctor en Sociologia per la Universitat del País Basc, professor d'Anàlisi de Xarxes, Machine Learning i Big Data, i soci-director de Network Outsight, empresa especialitzada en l'anàlisi sociològica de Big Data.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Andrea Rosales

Primera edició: setembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Morales i Gras
Producció: FUOC

Índex

Introducció

Vivim en un món cada cop més digitalitzat. L'aparició d'internet, la seva expansió massiva a partir de la dècada dels noranta del segle XX per tot el planeta —amb diferències substancials entre països i hemisferis— i la seva colonització posterior de la vida quotidiana mitjançant desenes de dispositius connectats han facilitat l'aparició del que anomenem *big data*. Però, contràriament al que hom podria suposar, quan parlem de *bigdata* o *dades massives* no ens estem referint merament a una qüestió de mida o pes de les dades. En realitat, estem caracteritzant un paradigma comunicatiu nou i propi del segle XXI, i que comporta implicacions de tota mena: tecnològiques, socials, culturals, legals i polítiques.

El canvi més important que ha comportat el procés de digitalització del món ha estat l'augment de la traçabilitat de l'activitat humana i no humana, la qual cosa ha resultat en un augment dels sistemes de predicció automatitzada. Actualment, gran part d'allò que fa o deixa de fer una persona, una comunitat, una màquina o un sistema complex pot quedar registrat en una base de dades per a la seva explotació posterior. D'aquesta manera, han proliferat i proliferen dispositius *smart* que són capaços de registrar la nostra activitat, predir el nostre comportament o els nostres desigs i fer-nos tot tipus de propostes per satisfer-los amb la màxima rapidesa possible. Per exemple, avui ja estem acostumats al fet que el mòbil ens proposi la millor ruta per a arribar a la feina o que el nostre *marketplace* preferit ens digui el tipus de sabates que compra la gent que ha adquirit els mateixos pantalons que nosaltres. També els edificis i les ciutats —i ben aviat els estats i les regions— s'han apuntat al paradigma *smart*, que els seus analistes i gestors basen en algoritmes d'intel·ligència artificial cada dia més complexos que els acompanyen en la presa de decisions.

Per tot això, durant els últims anys, el volum de dades accessibles, que poden ser processades i analitzades ha crescut exponencialment, la qual cosa ha implicat una autèntica revolució per al coneixement. D'aquesta capacitat d'acumular dades, n'han derivat diferents reptes:

- **Reptes tecnològics i d'enginyeria.** Com es poden emmagatzemar volums cada cop més grans i diversos de dades? Com es poden estructurar i ordenar? Com es poden processar i disposar per a l'anàlisi a una velocitat acceptable i, si és possible, en temps real?

- **Reptes científics i interpretatius.** Com es pot generar coneixement a partir de tantes dades tan diverses? Quines tècniques disponibles cal adaptar? Quines tècniques noves cal inventar? Quines són les preguntes clau que cal respondre en el nou paradigma emergent?

- **Reptes estratègics i de mercat.** Com es pot traduir el coneixement en intel·ligència competitiva? Com es pot afegir valor a les dades? Com es poden dissenyar sistemes d'indicadors que facilitin la presa de bones decisions en períodes breus?

En aquesta assignatura ens centrem fonamentalment en les necessitats i els reptes del tercer tipus, els que neixen dels contexts de gestió de dades i dels problemes derivats de la generació d'intel·ligència o coneixement aplicat. Per començar, visitarem i revisarem el concepte de *big data*, que és l'element central del paradigma comunicatiu actual. Tot seguit, n'explorarem la cadena de valor: ho farem en termes generals en primer lloc i, finalment, en el cas específic de les dades provinents dels *social media*. Ens centrarem en aquest punt en els diferents processos de mineria de dades socials (*social data mining*) que poden aportar valor a una estratègia de gestió dels *social media*.

1. El concepte de *big data*

Quan parlem de *big data* el primer que cal dir és que la mida importa, però que no ho és tot. És evident que quan l'adjectiu que acompanya a *data* és *big* s'està denotant certa centralitat de l'aspecte del volum; però també és cert que hi ha altres qüestions que són tan importants com el volum a l'hora de caracteritzar aquest paradigma comunicatiu.

A l'inici del segle, l'analista de dades Doug Laney va definir les tres *v* del *big data*: volum, velocitat i varietat (2001). A aquella definició inicial, diversos analistes han afegit nous conceptes que comencen amb *v* (taula 1). Si ens atenem estrictament a les *v* que han estat identificades en treballs acadèmics —i deixem de banda contribucions fetes des de blogs o empreses de programari com SAS o Oracle— podem destacar les cinc *v* i la *c* d'Abiodun Oguntimilehini Emmanuel Ojo Ademola (2014), les set *v* d'M.Ali-ud-din Khan, *et. al.* (2014), o les nou *v* i la *c* de Ripon Patgirii Arif Ahmed (2016).

Taula 1. Les *v* (i la *c*) del *big data* segons...

Doug Laney (2001)	Abiodun Oguntimilehin i Emmanuel Ojo Ademola (2014)	M.Ali-ud-din Khan *et. al.* (2014)	Ripon Patgiri i Arif Ahmed (2016)
Volum	Volum	Volum	Volum
Varietat	Varietat	Varietat	Varietat
Velocitat	Velocitat	Velocitat	Velocitat
	Variabilitat	**Validesa**	Valor
	Valor	**Veracitat**	Veracitat
	Complexitat	**Volatilitat**	Validesa
		Valor	**Visualització**
			Virtualitat
			Variabilitat/Volatilitat
			Complexitat

Font: elaboració pròpia

Nota

Els conceptes en negreta són els que no es troben en columnes anteriors.

És evident que no totes les *v* i les *c* identificades tenen el mateix valor, i també ho és que algunes apunten a aspectes força similars. El «valor» és precisament el component que comparteixen les tres conceptualitzacions del *big data* fetes durant la segona dècada del segle XIX; això és així perquè el paradigma comunicatiu del *big data* ha crescut com un paradigma fortament orientat al mercat i a l'oportunitat d'afegir valor a les dades.

1.1. Volum

L'ús intensiu de les noves tecnologies en tots els àmbits de la vida i la gran capacitat a l'hora de traçar i registrar l'activitat humana i no humana tenen com a conseqüència la generació d'una gran quantitat de dades.

Mai en tota la història de la humanitat havíem generat tantes dades com fins ara. Per la naturalesa descentralitzada d'internet, és impossible conèixer amb exactitud el volum de dades que es produeixen cada dia.

Reflexió

Aproximacions dutes a terme el 2018 (Marr, Bernard, 2018) apunten al fet que diàriament es generen prop de 2.5 quintilions de *bytes* o, el que és el mateix, prop de 2.500 milions de *gigabytes*. Si féssim servir discs de Blu Ray d'una capa (25 GB per disc) per emmagatzemar totes les dades i els apiléssim, cada dia sumaríem 1.200 metres de dades: 4 cops la torre Eiffel de Paris. I per si no fos prou, s'estima que la xifra anterior es duplica cada dos anys.

El paradigma de les dades massives es caracteritza per un augment exponencial i amb caràcter permanent en el volum de les dades produïdes. La recollida i l'emmagatzematge de les dades actualment impliquen un repte de primer ordre, com també l'adquisició de les competències necessàries per analitzar-les. Precisament per això, és important entendre que la característica fonamental de les dades massives no és el volum, perquè el volum de les dades és relatiu a la capacitat de computació de cada moment: el que avui considerem una «gran» quantitat de dades és molt probable que esdevinguin engrunes d'aquí uns quants anys.

1.2. Varietat

La gran quantitat de dispositius que són intensivament utilitzats per a la producció de dades té com a conseqüència directa una enorme varietat en la tipologia, els formats i l'estructura de les dades que es generen. A més de dades quantitatives (xifres) i corpus documentals (texts), el paradigma del *big data* es nodreix de fotografies, vídeos, àudios, coordenades geogràfiques i un llarg etcètera.

Segons el mode d'estructuració de les dades i la manera com s'emmagatzemen, és habitual classificar-les en dades estructurades, no estructurades i semiestructurades (José Luis Gómez García, i Jordi Conesa i Caralt, 2015).

Anomenem **dades estructurades** les dades que es poden integrar en una base de dades relacional (una base de dades MySQL, PostgreSQL o MS Access).

Una base de dades estructurada (taula 2) tindrà una sèrie de camps o atributs predefinits que seran equivalents per a cada fila o cas. Això vol dir que haurem planificat el format de la dada que hi emmagatzemarem abans de fer-ho: el tipus de dada, la seva posició, longitud, etc.

Taula 2. Exemple de base de dades estructurada

Firstname	Lastname	Gender	Age	Occupation	Salary	Marital status
Lucas	Phillips	Male	30	Journalist	125598	Single
Maddie	Farrell	Female	23	Florist	39017	Married
Blake	Perry	Male	18	Electrician	109020	Married
Belinda	Mason	Female	22	Chef	109829	Single
Edwin	Morgan	Male	19	Lawyer	195222	Single
Amber	Hall	Female	24	Account	157511	Single
Melanie	Dixon	Female	20	Composer	40231	Single
Tyle	Alexander	Male	26	Producer	46289	Married
Edward	Carter	Male	30	Historian	140207	Single
Daisy	Holmes	Female	26	Architect	118097	Married
Melanie	Clark	Female	29	Interior Designer	169336	Single
Freddie	Russell	Male	28	Natgenatucuab	106862	Married
Dale	Higgins	Male	27	Account	150590	Married
Arnold	Cameron	Male	21	Teacher	148883	Married
Ryan	Higgins	Male	20	Firefighter	185389	Single
Walter	Morgan	Male	25	Singer	68184	Married
Adrian	Myers	Male	24	Medic	198430	Married
Emma	Murphy	Female	21	Physicist	119587	Married
Tyler	Perkins	Male	23	Interior Designer	130840	Married
James	Thomas	Male	30	Photographer	126700	Single

Font: elaboració pròpia

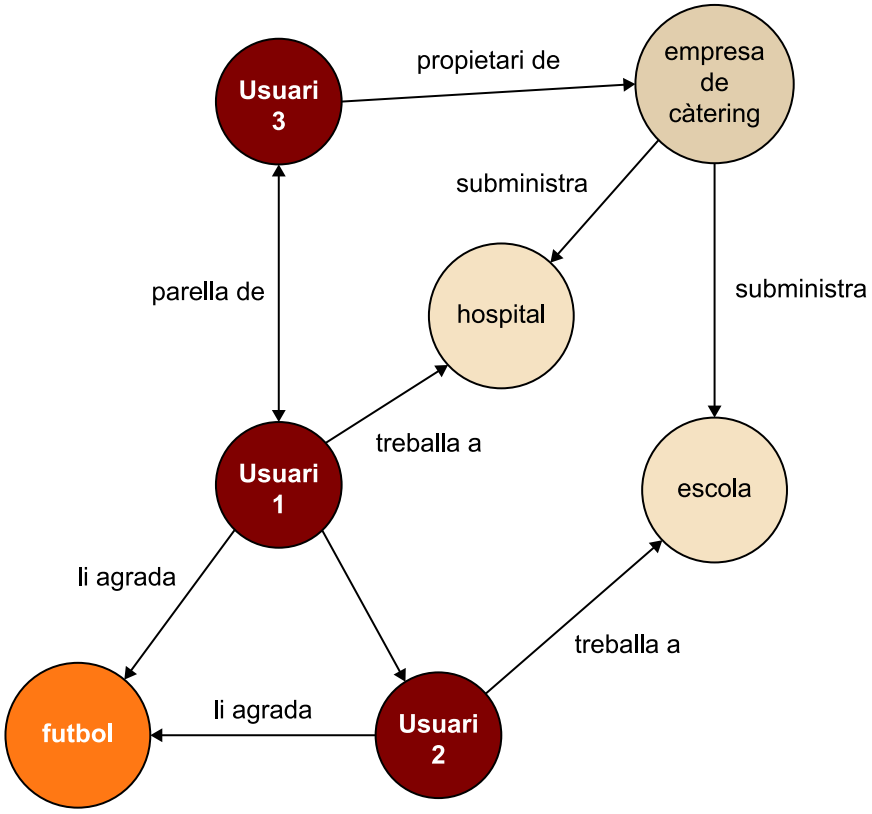
Anomenem **dades no estructurades** les dades que no tenen una estructura fixa, tot i que puguin tenir una estructura implícita que cal descobrir o una sèrie de propietats que cal identificar.

Es tracta, en definitiva, de qualsevol dada produïda que no hagi estat dissenyada per encaixar en un model de dades preestablert o de les dades produïdes espontàniament per a les quals encara no s'ha creat un mètode d'estructuració. Típicament, estem parlant de textos (per exemple, llibres sencers, articles, correus electrònics, publicacions en xarxes socials i continguts web) o d'arxius multimèdia (per exemple, fotos, vídeos i música). Qualsevol dada no estructurada pot ser estructurada, tant per procediments manuals com automàtics. La gran majoria de les dades de què disposen les empreses no són estructurades.

Com a resultat del procés d'estructuració de les dades no estructurades, ens trobem també amb la categoria de **dades semiestructurades**.

Tanmateix, l'estructuració sempre és «semi» i mai no es pot considerar completa perquè podríem seguir aplicant més tècniques d'estructuració sobre la dada no estructurada. Les dades provinents de converses en xarxes socials pertanyen a aquesta categoria. Entre les diferents tècniques d'estructuració de les dades no estructurades destaquen els llenguatges XML i JSON, molt utilitzats en el paradigma del web semàntic. També destaquen les bases de dades NoSQL com les bases de dades orientades a xarxes o grafs (figura 1), que organitzen la informació en mapes conceptuals.

Figura 1. Exemple de base de dades orientada a graf



Font: elaboració pròpia

1.3. Velocitat

Quan parlem de velocitat en *big data* ens referim a dues coses diferents:

- La rapidesa en el creixement de les bases de dades, que, com ja hem vist anteriorment, creixen de manera exponencial.
- La velocitat en la transferència de dades per a l'anàlisi.

En termes generals, podem afirmar que la velocitat de transferència és molt important en el paradigma de les dades massives. Això està estretament lligat amb la cadena de valor dels models de negoci orientats a dades (*data driven-businesses*). Necessitem poder emmagatzemar, carregar, processar i visualitzar les dades a tota velocitat, per tal de poder-les explotar i generar valor a partir d'elles, a vegades en entorns de temps real.

Reflexió

La velocitat serà un factor important, per exemple: per gestionar l'atenció al client, per monitorar el rendiment d'un servei o producte, per identificar oportunitats en xarxes socials o per generar continguts periodístics d'actualitat. I no ho serà, quan calgui dur a terme un estudi de mercat.

Altres vegades, la velocitat no serà un factor tan important o central. Com més important sigui la velocitat, més rígid i previsible haurà de ser el model d'anàlisi de dades.

Per acabar, cal considerar que la demanda de velocitat té efectes molt importants en els costos dels serveis i els servidors que s'utilitzin.

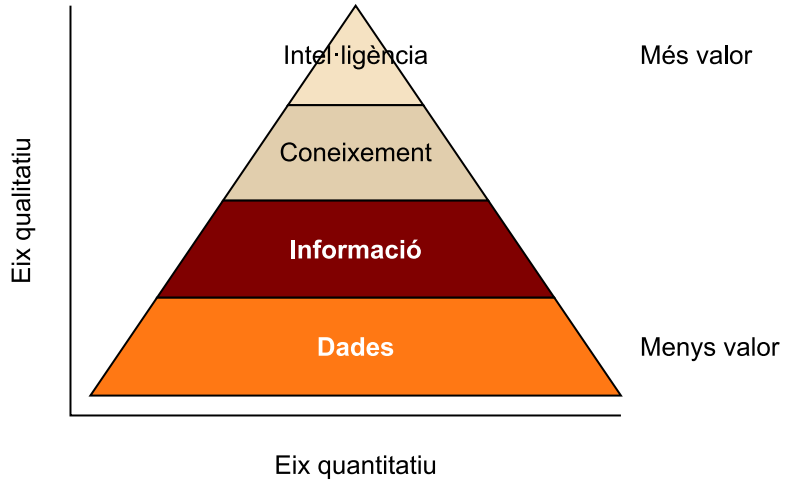
1.4. Valor

Com ja hem vist, la capacitat que hem desenvolupat les societats avançades a l'hora d'acumular informació tendeix a l'infinit. Per això, és molt important seleccionar adequadament la part de les dades que té sentit guardar i prescindir de la part de les dades que no és necessària.

El de les dades massives és un paradigma comunicatiu fortament orientat al mercat. Això vol dir que, en la majoria de casos, el sentit únic d'acumular, processar i analitzar dades és la seva explotació i la generació de valor. Per regla general —i deixant de banda algunes excepcions com poden ser els usos més socials o de recerca bàsica—, farem bé de desfer-nos de tot allò que no aporti valor o que es consideri que no en pot aportar.

A més a més, cal entendre que la generació de valor econòmic o empresarial a partir de les dades és en una eventual interpretació que aporti un component d'innovació, d'oportunitat de competitivitat o de productivitat. La piràmide informacional (figura 2) és un model teòric proposat per Gloria Ponjuán Dante (1998) que pertany a la teoria de la decisió i l'aprenentatge organitzatiu. Segons aquest model, les dades són la base per a la presa de decisions, però són necessàries una sèrie d'operacions de refinament per transformar les dades en informació, en coneixement i, finalment, en intel·ligència. El valor afegit a les dades deriva del coneixement aplicat o la intel·ligència que un analista sigui capaç de generar a partir d'aquestes.

Figura 2. La piràmide informacional



Font: elaboració pròpia a partir de Gloria Ponjuán Dante (1998)

1.5. Veracitat

Entre els riscos més punyents que afecten la presa de decisions basades en dades ens trobem amb les dades de mala qualitat, corruptes o invàlides. Així, el *big data* pot convertir-se en el camí més ràpid cap al desastre més absolut. Maximitzar les condicions de veracitat de les dades que s'obtenen és una qüestió cabdal: és molt important que les dades que s'obtenen i processen siguin fiables i fidels a la realitat.

Són múltiples les causes que poden comprometre la veracitat de les dades integrades en un sistema de *big data*. Les més habituals tenen a veure amb errors informàtics, amb el processament incorrecte del soroll o les interferències en un sistema (per exemple, amb l'estructuració incorrecta de dades no estructurades). Altres causes, menys habituals, poden tenir a veure amb atacs malintencionats com, per exemple, quan un pirata informàtic accedeix a una base de dades i manipula les dades per induir un error en el sistema (per exemple, posar o treure diners d'un compte bancari).

No hi ha una fórmula màgica per garantir la veracitat de les dades que s'emmagatzemen i processen, ja que això depèn en gran mesura del tipus de dada que s'estigui processant i de la metodologia que s'utilitzi. Sempre és adequat avaluar els riscos específics d'un sistema de *big data* i disposar dels mecanismes de protecció, avaluació i correcció necessaris (per exemple, mesures de seguretat per evitar atacs i manipulacions externes malintencionades, sistemes d'alertes per detectar anomalies). Així mateix, és important sotmetre els mètodes de tractament de dades a tests d'estrès i fer tantes comprovacions manuals com sigui possible abans de procedir a l'automatització dels processos.

1.6. Validesa

Un aspecte relacionat amb l'anterior que pot comprometre seriosament la presa de decisions basada en dades és la validesa de la interpretació. Per tal d'estar segurs que la interpretació de les dades amb què comptem es vàlida per respondre la pregunta que volem respondre, hem de conèixer molt bé les condicions que han produït aquestes dades i haurem de saber amb certesa que aquestes condicions són comparables a les actuals i que ens poden conduir a interpretacions igualment vàlides.

És important distingir la validesa de la veracitat. Mentre que la veracitat és una propietat intrínseca de la dada, la validesa depèn totalment de la interpretació que en vulgui fer l'analista. És perfectament possible que una dada veraç sigui invàlida a l'hora d'establir una interpretació.

> Imaginem, per exemple, que volem identificar com n'és de popular un usuari de Twitter o d'Instagram per mitjà del seu nombre de seguidors: el nombre de seguidors és un registre real i veraç, però no és vàlid establir una relació directa entre volum de seguidors i popularitat, per una sèrie de raons, començant per l'existència d'usuaris falsos o amb el comportament automatitzat.

Altres aspectes clau que comprometen la validesa de la interpretació són els següents:

1) Els biaixos que hi pugui haver en les mateixes dades (biaixos estadístics: diferències sistemàtiques entre els valors estimats i els reals).

Els biaixos estadístics. Cal ser conscients que, no pel fet de ser massives, les dades estan absentes de biaixos. Les dades massives poden estar esbiaixades per la manera com s'han capturat, si no s'han tingut en compte tots els tipus de subjectes de manera equilibrada i representativa (biaix de selecció). Un altre tipus de biaix es pot produir com a conseqüència de problemes derivats de l'instrument de captura de dades (biaix d'informació).

2) Els biaixos en la mirada de l'analista (biaixos cognitius, culturals o polítics: creences i subjectivitats de l'analista que alteren l'anàlisi).

Pel que fa als biaixos cognitius, culturals o polítics, és important que l'analista entengui i gestioni correctament la seva relació personal amb les dades que està analitzant. Un dels biaixos cognitius més importants en els *social media* és la tendència a donar validesa a la informació que confirma les pròpies creences i treure'n a la informació que les contradiu (biaix de confirmació). Mitjançant aquesta predisposició natural humana proliferen notícies falses i desinformació en xarxes.

3) Els biaixos en els algoritmes que reprodueixen els biaixos del seu programador (biaixos algorítmics que reflecteixen els valors i les creences del seu creador).

Els biaixos algorítmics en realitat són una mescla entre els dos anteriors: dades mal recollides o mal processades de manera sistemàtica com a conseqüència de biaixos presents en la mirada de l'analista que ha programat l'algoritme que recull i processa les dades. Vegem-ho amb un exemple:

Cas exemple: Amazon

Amazon va desenvolupar el 2014 un algoritme preparat per a la selecció de personal. Per entrenar l'algoritme, van fer servir una base de dades que contenia informació sobre processos de selecció de personal durant deu anys. La idea era utilitzar la informació dels últims deu anys, el perfil dels candidats i el seu bon o mal rendiment en l'empresa, per crear un algoritme predictiu capaç de proposar a l'ocupador els cinc millors perfils entre una borsa de candidats.

El 2015 es van adonar que l'algoritme discriminava les dones, proposant de manera sistemàtica homes per a les diferents posicions, i incloent la variable «dona» com un factor de penalització. La causa d'aquesta discriminació no fou, en principi, una voluntat explícita i manifesta dels programadors de discriminar les dones, sinó una conseqüència derivada de la distribució de les dades que no fou correctament centrada. Les dades que s'havien fet servir per entrenar l'algoritme provenien de la indústria tecnològica dels EUA, que està fortament masculinitzada. D'aquesta manera, l'algoritme va «entendre» que ser home constituïa un factor d'èxit en el model predictiu, i ser dona un factor de fracàs.

Enllaç d'interès

Vegeu l'article en què s'explica el cas d'Amazon en aquest enllaç:

<https://www.bbc.com/news/technology-45809919>

A l'hora de treballar amb dades, ja sigui des d'un punt de vista descriptiu o explicatiu, o amb la intenció de generar un algoritme predictiu, és molt important que entenguem què és el que volem aconseguir i, en conseqüència, que puguem decidir si les dades de què disposem són vàlides per al nostre propòsit.

Imaginem, per exemple, que volem entendre el comportament d'un client dins d'un comerç electrònic o *e-commerce*. Probablement, si treballem amb sèries temporals llargues hi haurà força elements externs a la mateixa dada que la invalidin per prendre decisions, com ara canvis en l'estructura web o en els píxels i etiquetes (*tags*) de monitoratge de l'activitat de l'usuari.

El factor temps sol ser un gran factor a l'hora d'«invalidar» dades, però no és l'únic. En realitat, qualsevol reestructuració del model de dades, tant en la seva producció com en l'emmagatzematge posterior, pot invalidar un conjunt de dades, malgrat que preservin intactes les condicions de veracitat. Com en el cas de la veracitat, no hi ha cap fórmula màgica que garanteixi la validesa de la interpretació: cal disposar dels mecanismes i processos adequats per garantir-la.

1.7. Visualització

Una imatge val més que mil paraules i probablement un gràfic o diagrama val més que cent mil matrius de dades. Visualitzar correctament les dades és d'allò més útil per a la seva anàlisi i per a la seva comprensió.

Les propietats de les dades que no es visualitzen correctament esdevenen inevitablement invisibles o tergiversacions. És, doncs, molt important invertir el temps necessari a optimitzar els recursos de visualització/visibilitat de les propietats rellevants de les nostres dades (les que aporten valor).

Respecte a les eines de visualització de dades, a més de les eines habituals de l'anàlisi estadística i demogràfica[1] és important tenir en compte noves visualitzacions nascudes els darrers anys, que permeten representar certes propietats i relacions de les dades[2]. També cal considerar que moltes d'aquestes visualitzacions poden presentar-se en entorns de quadres de comandament interactius (*dashboards*). En el darrer mòdul, veurem totes aquestes qüestions amb més profunditat.

(1)Per exemple histogrames, diagrames de caixa i bigoti, piràmides, gràfics de sectors, gràfics de dispersió, dendogrames, mapes.

(2)Per exemple diagrames de Shankey, gràfiques de projecció solar, diagrames radials o de cordes.

La visualització de dades és un dels aspectes més atractius i interessants de l'analítica de dades. És molt important tenir sempre en compte que qualsevol visualització de dades que es vulgui implementar ha de servir per respondre una pregunta clau i ha de tenir en compte l'audiència a la qual es dirigeix. Per això, és important no deixar-se portar només per la bellesa de certs models de visualització i assegurar-nos que compleixen el seu propòsit amb eficiència.

1.8. Virtualitat

La gestió de les dades massives no pot dur-se a terme si no és en un entorn virtual. Resulta evident que els antics arxivadors analògics no podrien servir en cap dels casos, però en realitat, tampoc no n'hi hauria prou amb un ordinador personal convencional per emmagatzemar, processar i visualitzar la quantitat de dades que poden constituir un ecosistema de dades massives.

Els servidors en xarxa i els serveis d'informàtica en núvol (*cloud computing*) són eines que esdevenen necessàries tan bon punt l'emmagatzematge de dades arriba a cert punt crític. A més, els entorns de dades massives se solen caracteritzar per la pluralitat de serveis que hi intervenen[3] i que és raonable mantenir en servidors diferenciats. Per tot això, la gestió de les dades té un element virtual inalienable.

(3)Per exemple bases de dades, serveis d'ETL, *dashboards* d'analítica...

1.9. Variabilitat/volatilitat

Diversos autors han identificat diferents aspectes relatius a la naturalesa canviant de les dades massives. Mentre que el concepte de *volatilitat* s'ha utilitzat generalment en un sentit negatiu, assenyalant el canvi com un factor d'incertesa i com a possible resultat d'una manipulació malintencionada, el concepte de *variabilitat* sol fer-se servir en clau positiva, per indicar la vitalitat de les pròpies dades.

Totes dues expressions apunten en una mateixa direcció: les dades massives són dinàmiques. Això és així perquè, per regla general, ens interessarà monitorar sistemes vius i canviants per poder identificar aquests canvis i trobar-hi oportunitats per a la generació de valor.

1.10. Complexitat

Tots els factors anteriors fan que els escenaris de dades massives tendeixin a una gran complexitat. La complexitat a què ens referim es fa palesa en els mateixos protocols d'extracció i emmagatzematge de la dada i també en l'analítica posterior.

Com hem vist, els entorns de dades massives ens plantegen diferents tipus de reptes de diferent ordre (tecnològics, científics i estratègics), que s'expressen per mitjà de totes les *v* del *big data*. Les solucions disponibles per als reptes derivats, sobretot, del volum i de la varietat de les dades són enormement complexes: servidors connectats, bases de dades amb diversos graus d'estructuració, visualitzacions complexes. Per tot això, ha estat i és cabdal la configuració de nous perfils professionals preparats per afrontar —i eventualment reduir— la complexitat del *big data*.

2. Mineria de dades: explotant la cadena de valor del *big data*

Ja hem vist que la generació de valor és un aspecte clau en el paradigma del *big data*. El valor que les dades tenen per elles mateixes sol ser molt baix en termes econòmics o empresarials, però constitueixen la matèria primera a partir de la qual es pot afegir i generar més valor. El procés a partir del qual s'explota la dada com a matèria primera i se'n genera valor és conegut com a *mineria de dades*, i la operació consisteix, en essència, a capturar una sèrie de registres d'informació i interpretar-los per crear un patró que ens aporti idees accionables.

En aquesta secció del temari introduirem la cadena de valor del *big data* (Montse García-Alsina, 2017), distribuïda en quatre fases fonamentals: generació, adquisició i neteja, emmagatzematge, i anàlisi de dades.

2.1. Generació (primer pas en l'explotació de les dades)

Ja hem vist que les dades massives es caracteritzen per una gran varietat en les fonts de dades. En el context de la societat-xarxa (Manuel Castells, 2009), del món interconnectat i de la digitalització de la vida quotidiana, bona part de l'activitat humana o no humana és susceptible de convertir-se en generadora de dades: desplaçaments registrats per un GPS, rastres de navegació per internet, registres presos per sensors urbans (per exemple, *smartcities*) o industrials, l'internet de les coses (IoT) o la mateixa activitat en xarxes socials. Tot això sumat a formats i fonts de dades també presents en paradigmes comunicatius anteriors: informació generada per les organitzacions o l'administració pública, dades provinents d'estudis demoscòpics i enquestes, i un llarg etcètera.

El procés de generació de dades no consisteix a «recollir-les» com si fossin carxofes o petroli. Les dades no són un recurs natural, sinó que són un subproducte sociotecnològic que deriva de la interacció entre persones i sistemes digitals. El procés de generació de dades està estretament lligat al procés d'estructuració o transformació de les activitats que registrem i que posteriorment anomenem *dades*, com també a l'activitat dels usuaris. Com hem vist abans, aquest procés pot ser més o menys definitiu: mentre que les dades estructurades es presenten en un format altament estandarditzat (files i columnes conegudes i previsibles), les dades no estructurades o semiestructurades presenten un grau més baix d'homogeneïtat.

Gairebé sempre, el procés de generació de dades es duu a terme com a procés desvinculat respecte les fases posteriors. Moltes vegades, fins i tot són diferents empreses les que es dediquen a dur a terme els diferents processos de la cadena de valor del *big data*. En el cas de les empreses de xarxes socials com Facebo-

ok, Twitter o Linkedin, són les úniques encarregades de generar dades a partir de les interaccions dels usuaris amb les plataformes (per exemple, converses, relacions de seguiment, introducció de dades personals). La resta d'agents —corporatius o no— que vulguin fer ús de les dades que generen les xarxes hauran de fer-ho des de la seva interfície de programació d'aplicacions (API) i acceptant, per tant, les condicions d'ús que imposi la plataforma. La informació disponible per mitjà de les API haurà d'estar adaptada al marc legal vigent, essent la mateixa empresa de xarxes socials la responsable principal d'aquesta qüestió.

Una alternativa cada cop més estesa a l'ús de les API és el «raspament web» o *webscraping*, que consisteix en un conjunt de tècniques automatitzades que permeten extreure informació de webs de manera sistemàtica. Malgrat que no es tracta de tècniques il·legals *per se*, solen ser tècniques no permeses per la normativa de les plataformes generadores de dades, i poden plantejar alguna violació de les lleis de protecció de dades si no s'implementen de manera responsable. Casos de raspament web ben intencionats poden ser els comparadors de productes[4], i casos de raspament mal intencionat són la recuperació de correus electrònics amb finalitats d'enviament de correus brossa (*spam*).

(4)Aplicacions que disposen de robots que ens informen d'ofertes o oportunitats en webs.

Tant en un escenari de *webscraping* com d'explotació de les API cal tenir en compte algunes limitacions, especialment pel que fa a les categories especials de dades considerades en les diferents legislacions de protecció de dades: legislacions d'estats i d'entitats supraestatals com la Unió Europea. Típicament, les dades no podran utilitzar-se mai amb l'objectiu d'identificar categories personals com l'ètnia, la ideologia, el sentiment religiós, la sexualitat, la salut o altres dades biomèdiques o aspectes psicològics. És important que l'analista de dades conegui tots els límits i els respecti escrupolosament i, també, que conegui les atribucions del seu rol com a responsable o com a encarregat del tractament de dades. Les diferències entre països i continents són importants; per això, serà necessari tenir en compte la legalitat vigent del territori on s'opera.

2.2. Adquisició i neteja de dades (segon pas en l'explotació de les dades)

El cas d'ús més habitual per la majoria d'analistes de dades serà l'explotació de dades generades per una altra empresa i sobre la qual es tindrà un accés limitat. Qualsevol que faci ús de l'API d'una aplicació o d'una xarxa social es trobarà en aquesta situació. En aquest cas, el procés d'adquisició de la dada haurà de prendre com a matèria primera la consulta (*query*) feta sobre la mateixa API, que permetrà la recuperació d'un volum determinat de dades controlat per l'empresa generadora. En *social media*, aquest és el cas de les interaccions en una conversa de Twitter, les relacions de seguiment entre els seus usuaris, les publicacions i les interaccions en una pàgina de Facebook o en un *hashtag* d'Instagram.

La totalitat de les dades tal qual es generen (*raw data*) solen ser moltes més de les que realment es necessiten per dur a terme qualsevol anàlisi. El procés d'adquisició de la dada és el segon pas de la cadena de valor de la dada i consisteix precisament a seleccionar les dades necessàries, transmetre-les a una base de dades pròpia i aplicar-hi les operacions de transformació necessàries. Aquestes operacions seran diferents en cada cas. Entre les més habituals trobem les següents:

1) Integració de dades. El procés consisteix a combinar i unir dades provinents de diferents fonts de dades i se sol dur a terme mitjançant eines d'ETL (extracció, transformació i càrrega) per mitjà de les quals s'homogeneïtzen, estructuren i integren les dades, seguint un model predefinit. Recentment, estan proliferant eines d'ELT, en les quals s'inverteix l'ordre dels factors (extracció, càrrega i transformació), i que es presenten com a potencialment més escalables i versàtils. Com veurem més endavant, les eines d'ETL integren dades en magatzems de dades (*data warehouses*) i les eines d'ELT ho fan en llacs de dades (*data lakes*). Una alternativa als mètodes d'ETL i ELT, que consumeixen força espai i recursos, és el mètode de federació de dades, que només emmagatzema metadades a través de les quals s'accedeix a bases de dades externes.

2) Neteja de dades. Aquest procés és clau per garantir la qualitat de la dada. Es pot executar de diverses maneres, però sempre amb els mateixos objectius: detectar i corregir errors en les dades (inconsistències, incoherències i altres aspectes que poden comprometre la veracitat de la dada).

3) Eliminació de redundàncies. Una part fonamental de la neteja de les dades és l'eliminació de duplicacions. Cal eliminar duplicacions i redundàncies de manera sistemàtica, per reduir costos d'emmagatzematge i per garantir la veracitat de les dades.

2.3. Emmagatzematge de dades

Un cop obtingudes les dades és el moment d'emmagatzemar-les. L'objectiu principal d'aquesta fase és poder disposar de dades qualificades amb posterioritat, garantint un accés ràpid i complet en la fase d'anàlisi. A l'hora d'escollir el mètode òptim d'emmagatzematge de dades cal definir tres elements: el llenguatge en què es vol emmagatzemar la dada, la usabilitat de la dada emmagatzemada i el tipus de visualitzacions que se n'han de derivar.

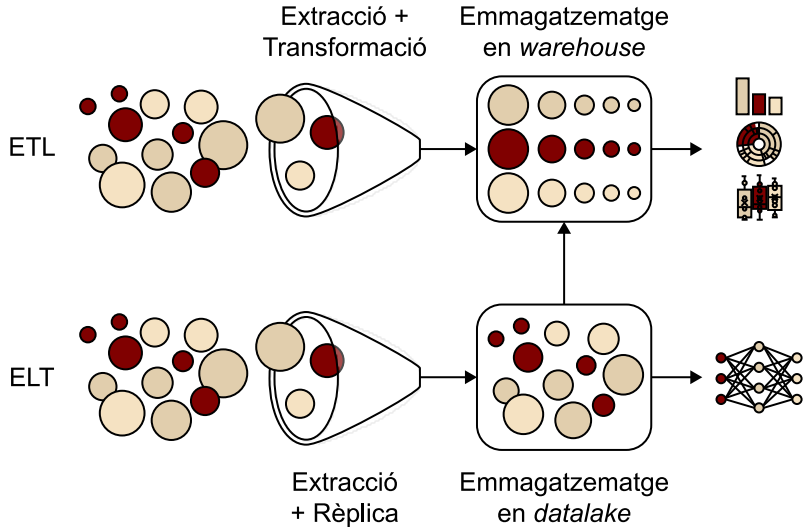
1) El **llenguatge d'emmagatzematge** de la dada dependrà directament del seu grau i mode d'estructuració. Quan el tipus de dada que es vol emmagatzemar és estructurat, la millor solució és una base de dades relacional de tipus SQL. Quan la dada que es vulgui emmagatzemar presenti una estructuració més baixa (dada no estructurada o semiestructurada), serà oportú considerar llenguatges no relacionals, com JSON, o com les bases de dades orientades a grafs.

2) Segons la **usabilitat de les dades** emmagatzemades, caldrà considerar-ne l'emmagatzematge en taules simples o en sistemes més complexos i integrats, com els magatzems de dades o els llacs de dades.

3) Finalment, en funció del **tipus de visualitzacions** que es vulgui obtenir de les dades, caldrà considerar els processos i subprocessos d'emmagatzematge, per a reduir-ne la complexitat i permetre la càrrega ràpida dels elements necessaris per a cada visualització.

A hores d'ara, ja hem pogut veure com en *big data* tot està relacionat amb tot. En aquest cas, el mètode òptim d'emmagatzematge de dades depèn completament del mètode d'integració dut a terme en la fase anterior d'adquisició (figura 3). Si el mètode d'integració és d'ETL (extracció, transformació i càrrega), haurem d'emmagatzemar dades preprocessades de manera estructurada o semiestructurada en un magatzem de dades (*data warehouse* o DWH). En canvi, si el mètode d'integració és d'ELT (extracció, càrrega i transformació), emmagatzemarem les dades sense tractar (*raw data*) en el seu format primitiu en un llac de dades (*data lake*). Al seu torn, l'elecció d'un dels mètodes d'emmagatzematge tindrà conseqüències molt importants en l'anàlisi de les dades.Les explorem a continuació.

Figura 3. Emmagatzematge en *data warehouse* o *data lake*



Font: elaboració pròpia

2.4. Anàlisi de dades

La fase d'anàlisi és la quarta i l'última passa de la cadena de valor del *big data*. Tanmateix, haurà de ser la primera en el raonament de l'analista: és en funció del tipus de valor que es vol extreure que cal dissenyar la resta de fases de la cadena.

Tots els processos anteriors a l'anàlisi de dades hauran d'estar disposats de manera que optimitzin una anàlisi orientada a respondre les preguntes clau de cada projecte.

Tot i que les empreses sempre han disposat de dades amb finalitats analítiques (per exemple, llibres de comptabilitat, fitxers de clients), no ha estat fins fa pocs anys que els avenços tecnològics i la necessitat d'intel·ligència competitiva han fet emergir els models de negoci orientats a dades (*data-drivenbusiness*). La disciplina que s'encarrega de l'anàlisi de les dades emmagatzemades en sistemes de *big data* és la ciència de dades (*data science*), i es nodreix tant dels avenços recents en enginyeria informàtica com de coneixements estadístics i d'intel·ligència artificial consolidats des dels segles XIX i XX, respectivament. Les següents són les eines principals d'anàlisi de dades:

1) Estadística. L'estadística és la ciència matemàtica relacionada amb la recopilació, anàlisi, interpretació i representació de dades. L'estadística descriptiva serveix per resumir i presentar els conjunts de dades: és, amb molta diferència, la més utilitzada, per la seva senzillesa i facilitat de comprensió. L'estadística inferencial serveix per extreure conclusions a partir de dades disponibles. La seva manera de construir coneixement és refutant hipòtesis, i és una disciplina fortament orientada a l'explicació de les relacions entre variables (formalització). En funció de la naturalesa d'aquestes variables[5] l'estadística disposa de diverses tècniques de validació hipotètica, com la regressió, l'anàlisi de variància o l'anàlisi de components principals, que es basen en la teoria de la probabilitat. El gruix de l'estadística es fonamenta en l'individualisme metodològic i se centra, per tant, en l'anàlisi de les propietats i els atributs dels casos individuals que analitza.

(5)La diferència més important és entre variables qualitatives i quantitatives.

2) Anàlisi de xarxes. Es tracta d'una tècnica arrelada en una perspectiva estructural, diferent, per tant, de l'individualisme metodològic, i orientada a la interpretació de les relacions entre casos individuals i no tant en les seves propietats o atributs. Matemàticament, l'anàlisi de xarxes es fonamenta en la teoria de grafs[6], que ha gaudit de menys popularitat durant els darrers segles, però que amb l'eclosió del món digital ha guanyat prominència. L'anàlisi de xarxes es nodreix simultàniament de disciplines com les matemàtiques, la sociologia o la biologia, i situa les relacions —i no els atributs— entre els nodes d'una xarxa com a element central d'anàlisi. En l'anàlisi de xarxes es combinen una sèrie d'elements quantitatius i qualitatius i hi ha una forta orientació cap al component visual, mitjançant la representació de grafs o xarxes. L'anàlisi de xarxes compta amb algoritmes i mètriques pròpies que permeten avaluar les propietats de cada node, com també de la xarxa en conjunt. En una altra assignatura optativa del màster aprofundirem en aquesta tècnica d'anàlisi.

(6)La teoria matemàtica que estudia el concepte de *xarxa*.

3) Intel·ligència artificial i aprenentatge automàtic. La intel·ligència artificial (IA) és qualsevol tècnica que capaciti un ordinador per dur a terme una o diverses accions que aparentin o emulin alguna de les dimensions de

la intel·ligència humana. L'aprenentatge automàtic (*machine learning*) és la subdisciplina de l'IA que persegueix la millora del propi sistema mitjançant l'experiència. Hi ha diversos tipus d'algoritmes d'aprenentatge automàtic. Vegem-ne alguns:

a) L'**aprenentatge supervisat** persegueix la predicció de resultats futurs en funció de sèries de dades conegudes. En contrast amb l'estadística inferencial, es posa l'èmfasi en la predicció (l'estimació de valors futurs) i no en la formalització (l'estudi de les relacions entre variables). De fet, és habitual que molts algoritmes funcionin amb «caixes negres» que no permetin estudiar les relacions entre variables. Entre els algoritmes d'aprenentatge supervisat més populars destaquen la regressió (lineal, polinòmica o logística), els arbres de decisió o l'aprenentatge bayesià.

b) L'**aprenentatge no supervisat**, en canvi, cerca classificar les dades en funció de les seves propietats intrínseques, sense partir d'un model predictiu preestablert. Moltes vegades s'utilitzen aquests algoritmes amb finalitats descriptives (per exemple, per descobrir patrons d'agrupació de casos) o en models d'investigació inductiva (el desenvolupament teòric neix de l'observació i no del contrast hipotèticodeductiu). Entre els algoritmes més populars, destaquen l'agrupació *k-means* o l'anàlisi de components principals, que també es fa servir en estadística inferencial.

c) Els **algoritmes de conjunt** combinen diversos tipus d'algoritmes per millorar el poder predictiu d'un model. Es tracta d'algoritmes preparats per resoldre problemes d'aprenentatge supervisat o no supervisat, però que per la seva manera de procedir no formen part de cap dels dos blocs anteriors. Els més populars són els algoritmes d'impuls adaptatiu (*adaptive boosting* o *AdaBoost*) i de boscos aleatoris (*random forests*), tots dos basats en arbres de decisió.

d) Els **algoritmes d'aprenentatge profund** (*deep learning*) constitueixen el camp d'estudi més prometedor de l'aprenentatge automàtic des del punt de vista dels resultats que obtenen, malgrat que comportin certs problemes d'interpretació que veurem més endavant. Es tracta d'algoritmes de caixa negra que estableixen relacions entre casos a diversos nivells, i que podrien utilitzar-se per resoldre problemes, sobretot, d'aprenentatge supervisat. Els algoritmes de xarxes neuronals són els més característics.

L'anàlisi de dades constitueix, sens dubte, la baula més important de la cadena de valor del *big data*. En el procés de disseny de l'arquitectura d'un sistema de dades massives, és crucial avaluar amb molta precisió les necessitats analítiques específiques que es tenen, ja que d'aquestes necessitats derivaran les decisions que hauran de configurar la totalitat del sistema: la selecció de les fonts de dades, el seu tractament i el mètode d'emmagatzematge. La diferèn-

cia entre els sistemes d'ETL i d'ELT serà crucial en aquest punt, ja que el tipus d'algoritmes a aplicar tindrà a veure amb el grau d'estructuració de la dada i amb la seva naturalesa bruta o preprocessada.

L'anàlisi és també l'àrea del *big data* que necessita ser més interdisciplinària i transdisciplinària. És impossible que un equip sense coneixements substantius sobre el camp d'estudi clau per al projecte (per exemple, sobre comunicació, sobre sociologia, sobre biologia, sobre epidemiologia, sobre literatura, sobre ciències ambientals...) pugui generar valor a partir de l'explotació de dades massives, i també ho és que ho faci un equip sense coneixements tècnics, procedimentals i metodològics (per exemple, informàtics, estadístics, de ciència de dades...). El *big data* constitueix, així, una crida a la comunitat científica en conjunt i a professionals de tota mena i no solament a enginyers informàtics i gestors de bases de dades.

3. Les eines del *big data*

A l'hora de posar en marxa un ecosistema de dades massives, el com és tan important com el què. Actualment, hi ha diverses empreses que ofereixen serveis integrals de *big data*. Les més importants són **Amazon**, **Google**, **Microsoft**, **IBM** i **SAP**. Totes ofereixen diversos serveis sota diverses modalitats de contractació (per exemple, *freemium*, cost per ús, llicències temporals) i que s'ajusten a diversos pressuposts. D'altra banda, també hi ha una sèrie de programes lliures i de codi obert que compleixen diverses funcions que hem definit en la cadena de valor, tot i que cap no es pot considerar un servei integral de *big data*:

- **Hadoop**, **ApacheSpark** o **Red Hat** són marcs de treball o *frameworks* d'informàtica en el núvol que permeten la computació distribuïda: la utilització de diversos ordinadors connectats en xarxa per resoldre problemes de computació massiva.

- **Spark SQL**, **Hive** o **Presto** són infraestructures per a l'emmagatzematge de dades relacionals basats en llenguatge SQL.

- **Talend**, **Pentaho** o **Oozie** són serveis d'ETL que s'integren fàcilment en *frameworks* d'informàtica en el núvol.

- **Python**, **R** o **Scala** són llenguatges de programació amb molt bones capacitats per a l'anàlisi de dades massives. També disposen d'eines i complements per a la visualització de dades.

També hi ha un gran nombre de solucions de programari privatiu que supleixen diverses necessitats no previstes pel programari lliure, com ara solucions específiques per sectors (per exemple, publicitat, educació, immobiliària, etc.) o per departaments específics (per exemple, comercial, màrqueting, legal, etc.). Un punt molt sensible en el que el programari privatiu, actualment, és molt més fort que el lliure és en la visualització de dades i les eines d'intel·ligència de negocis (*Business intelligence*). Les plataformes principals d'aquest mercat específic són Tableau, Looker, Microsoft PowerBi i Qlik.

El paisatge de programari i empreses de *big data* és enormement variable i canviant. Any rere any, sorgeixen més i millors sistemes i serveis, que obliguen a tots els professionals del sector a desenvolupar un permanent estat d'alerta. L'analista de mercats Matt Truck publica cada any un Data & AI Landscape que ens pot ser molt útil a l'hora de situar-nos.

4. Mineria de dades dels *social media*

Com ja hem vist anteriorment, un dels trets principals del que anomenem *big data* és la varietat en el format i en les fonts de dades. El paradigma de les dades massives es nodreix simultàniament de dades que provenen dels serveis financers, del comerç, del sector industrial, del sector de la salut, etc., i molt significativament, es nodreix del món de les telecomunicacions i de les xarxes socials d'internet.

Enllaç d'interès

Internet i els *social media* són els productors principals de *big data*. L'empresa DOMO publica regularment la infografia «Data Never Sleeps», en què consta tot allò que succeeix a internet en un minut: <https://www.domo.com/learn/data-never-sleeps-7>

En aquest apartat, veurem la cadena de valor de les dades que es produeixen en els *social media* i les xarxes socials d'internet. Veurem on es generen les dades —i qui les genera— i de quina manera són disposades per les empreses de xarxes socials per a la seva posterior adquisició, emmagatzematge i anàlisi. Repassarem, així, el procés de mineria de dades aplicat als *social media* mitjançant cadascuna de les fases de la cadena de valor i veient-ne les característiques principals en el cas dels *social media*.

4.1. Generació

El primer pas en la cadena de valor del *big data* és la generació. Abans ja hem vist que aquesta fase en la majoria d'ocasions es duu a terme com a procés desvinculat de les fases posteriors. Això vol dir que, per regla general, els explotadors i analistes de dades voldran explotar i analitzar dades generades fora dels seus ecosistemes o, fins i tot, integrar dades provinents d'una varietat d'ecosistemes. En el cas dels *social media*, cada ecosistema podria ser una xarxa social.

Des d'un punt de vista conceptual —i fins i tot ètic— també és important entendre que les dades no són creades per les empreses de *social media* de manera autònoma i autosuficient, sinó que són un subproducte tecnològic que integra l'activitat dels usuaris i el disseny tecnològic dels sistemes, amb les seves implicacions legals i ètiques. És l'activitat i la interactivitat d'aquests usuaris, és a dir, les seves publicacions, les seves relacions, els seus *likes*, els seus repiulades, els seus *swipes*, etc., sumades al mode d'estructuració dels propis sistemes, el que permet a les empreses registrar i empaquetar allò que anomenem *dades*, tal com les coneixem i tal com són disposades en les API.

Moltes de les xarxes socials disposen d'API pròpies per mitjà de les quals els explotadors i analistes de dades poden accedir-hi, cadascuna de les quals manté la seva política de dades. Algunes xarxes aposten per donar accés a publicacions i dades d'usuaris (per exemple, Twitter i Instagram), altres només sobre els usuaris (per exemple, Linkedin), i altres fan distincions segons perfils i rols d'usuari (per exemple, Facebook). El format i el volum de dades que les empreses generadores decideixen posar a disposició dels explotadors i analis-

tes és variable i subjecte a polítiques empresarials i marcs legislatius en revisió permanent. Per tot això, es tracta sempre d'un terreny complex i d'una font important d'incertesa per a les empreses de dades massives.

El tipus de dades que les diferents API decideixen posar o no posar a disposició dels explotadors i analistes poden pertànyer a diferents categories. Com ja hem vist, no totes les categories estan disponibles a totes les xarxes:

1) Perfils d'usuari. Es tracta d'informació sobre el propietari d'un compte. Algunes xarxes disposen de més dades que d'altres. Per exemple, mentre que a Facebook i Linkedin aquesta informació sol ser molt completa (gènere, aniversari, estat civil, estudis i un llarguíssim etcètera), a Twitter o Instagram ni tan sols es coneix el gènere de l'usuari. Això no vol dir que Twitter o Instagram no disposin internament d'algoritmes per identificar elements com el gènere. Per regla general, com més informació potencialment sensible tingui una empresa de *social media*, més tancada serà la seva API.

2) Connexions. Un aspecte clau de les xarxes és el conjunt de relacions que cada usuari estableix. Aquestes relacions han de ser necessàriament o bidireccionals o unidireccionals. Per exemple:

- **Bidireccionals**: les amistats a Facebook i els contactes a Linkedin.
- **Unidireccionals**: els seguidors i seguits a Twitter i Instagram.

Com veurem més endavant, aquest element de direccionalitat tindrà molta rellevància a l'hora d'analitzar aquestes relacions. Actualment, només Twitter proporciona aquest tipus de dada per mitjà de la seva API. A banda de qüestions més evidents com és la identificació de grups d'usuaris, l'anàlisi de les relacions d'un usuari és molt informativa respecte dels seus gustos i aficions, fins i tot de gustos i aficions no declarats o sobre els quals l'usuari no parla.

3) Publicacions. Els continguts que publica cada usuari (posts, piulades, fotografies, *stories*, vídeos) són també una dada clau de les xarxes socials. Mitjançant l'anàlisi d'aquestes publicacions, és possible generar una gran quantitat de coneixement. Els múltiples formats de les dades d'aquest tipus també constitueixen un dels reptes més importants i interessants que haurà d'afrontar qualsevol analista de *big data*.

4) Interaccions. Les interaccions en xarxes a vegades es poden entendre com un subtipus de publicació com, per exemple, quan hi ha una menció o al·lusió incrustada en un text. Les mencions a xarxes com Instagram o Twitter són el cas més evident; un altre cas serien els etiquetatges a fotografies. D'altra banda, també hi ha interaccions que s'estableixen entre usuaris i continguts: *likes*, *shares*, repiulades, etcètera.

5) Grups i llistes. Moltes xarxes també disposen d'elements que agrupen usuaris en comunitats d'interessos: els grups professionals a Linkedin, els grups o *fanpages* de Facebook o les llistes de Twitter en serien els exemples principals. Es tracta també d'una dada important, i que requereix diferents aproximacions analítiques segons la xarxa social.

6) KPI. Part de l'activitat social anterior és sistematitzada i organitzada en indicadors clau de rendiment (KPI) per les mateixes empreses de *social media*. Moltes xarxes socials disposen de *dashboards* d'analítica bàsica que permeten explorar el rendiment d'aquests indicadors. Per exemple:

- ***Engagement***: sol ser la suma d'interaccions usuari-publicació que acumula una publicació.
- **Abast**: la suma d'usuaris que han vist una publicació.
- **Impressions**: el nombre de cops que una publicació ha aparegut en una pantalla.

Cada xarxa disposa dels seus KPI i n'hi ha una gran varietat. Aquests indicadors cobren especial importància quan es tracta d'analitzar el rendiment d'una campanya publicitària.

7) Metadades. Finalment, cal tenir en compte el grup més voluminós de dades: les dades que l'usuari genera sense ser-ne ni tan sols conscient. Aquí entren la marca i el model de dispositiu, la càmera que ha fet la foto, la plataforma des d'on s'ha publicat el contingut, i fins i tot el color del menú de l'usuari. Per fer-nos una idea de la magnitud d'aquesta qüestió, Twitter compta amb més de quatre-centes variables associades a cada piulada que són accessibles des de l'API pública.

4.2. Adquisició

Les dades que es recuperen mitjançant les API de les xarxes socials pertanyen al bloc de les dades semiestructurades. Es tracta de dades que ja tenen cert grau d'estructuració o estandardització i que són retornades per l'API en un format previsible, com ara les dades relatives a perfils d'usuaris, els KPI o les metadades, però que, d'altra banda, presenten les típiques característiques de les dades pendents d'estructurar mitjançant algoritmes i altres tècniques classificatòries, com ara les connexions, les publicacions, les interaccions i els grups i llistes.

Com hem vist, destaquen dues maneres diferents d'orientar el procés d'integració de la dada; també la de xarxes socials. La primera manera consisteix a transformar la dada abans d'emmagatzemar-la, i la segona consisteix en tot el contrari, primer emmagatzemar-la i després transformar-la. El primer dels dos processos és el denominat ETL, mentre que el segon és el denominat ELT:

1) L'ETL requereix més planificació i anticipació, i també és molt més replicable i escalable. Sempre aplicarem els mateixos algoritmes i processos al mateix tipus de dada i, per tant, obtindrem un resultat més previsible que podrem integrar en una base de dades i analitzar-la amb posterioritat. En el cas dels *social media*, gràcies a un procés d'ETL podrem associar un sentiment a un fragment de text, podrem conèixer els usuaris més importants (tenint en compte aspectes com el nombre de repiulades o de «m'agrada» a Facebook, o altres indicadors més complexos) en una conversa, o podrem generar una línia temporal per a identificar els moments clau d'un debat.

2) L'ELT és un procés menys escalable i replicable però que proporciona molta més autonomia investigadora. En aquest cas, l'analista podrà decidir quin tractament o quines tècniques aplica a les dades i respondre a preguntes més específiques i específiques per a cada client. En el cas dels *social media*, farem servir una lògica de procés ETL quan vulguem identificar les subcomunitats dins d'una conversa o quan vulguem entrenar nous algoritmes.

Les dades de xarxes socials es poden adquirir de manera puntual o recurrent. Mentre que una adquisició puntual s'assembla al model d'ELT[7], segurament voldrem optar per un procés d'ETL quan l'adquisició s'hagi d'efectuar de manera recurrent .

(7)Recuperarem les dades, les desarem a l'ordinador en el format que proporciona l'API, i després les processarem, transformarem i analitzarem.

Moltes vegades, sobretot en fases d'aprenentatge o de familiarització amb les dades, podrem recórrer a eines de tercers, gratuïtes o molt econòmiques, que ens permeten accedir a dades de xarxes socials per dur a terme les nostres anàlisis. Aquestes eines típicament recuperen i preformaten la dada, i finalment la desen al nostre ordinador (executen un procés d'ETL); però, des del punt de vista de l'usuari, probablement encara caldrà executar una sèrie d'operacions necessàries que aportin valor addicional a les dades.

4.3. Emmagatzematge

Les dades de xarxes socials es poden emmagatzemar en diferents tipus de bases de dades, sempre en funció del model d'adquisició i d'integració i del format específic en què es vulguin emmagatzemar. El més habitual serà fer-ho en una base de dades relacional i estructurada amb els seus atributs predefinits, destinant algunes columnes a dipositar-hi dades en formats no estructurats mitjançant llenguatges com ara JSON o XML. Per exemple, en la taula següent (taula 3) hi podem veure com de cada ID d'usuari en depenen blocs de dades asimètrics, ja que de cada usuari en coneixem atributs diferents.

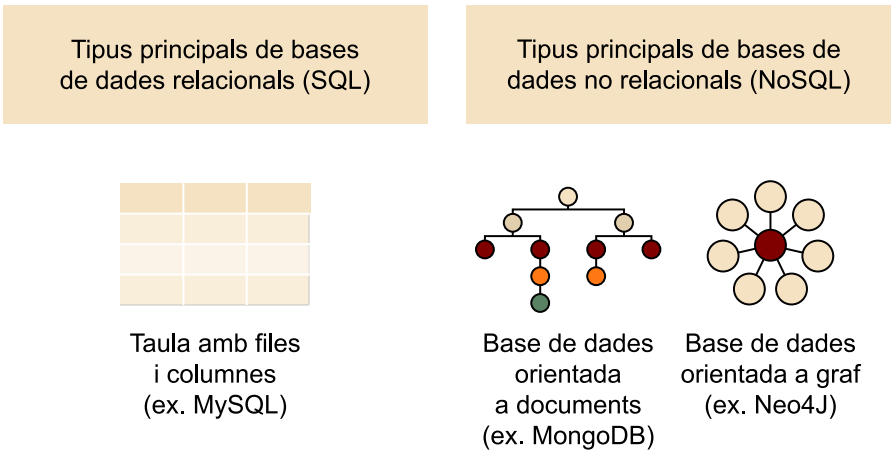
Taula 3. Exemple de base de dades semiestructurada amb dades incompletes

id	atributs
001	{"Nom":"Adrià", "Cognom":"Ramoneda", "Edat": "54" }
002	{"Nom":"Anna", "Cognom":"Gutiérrez", "Ciutat":"Cerdanyola del Vallès", "Edat": "32" "Gènere":"dona" }
003	{"Nom":"Antonio", "Edat": "64", "Gènere":"home" }

Font: elaboració pròpia

En funció de la magnitud del projecte i de les seves necessitats específiques, pot ser una bona idea emmagatzemar les dades en bases de dades no relacionals. La alternativa més comuna a les bases de dades relacionals són les bases de dades orientades a documents com MongoDB, en les quals s'estableixen connexions entre documents, generalment escrits en llenguatge JSON, que admeten una gran diversitat de formats i de camps. Una segona alternativa, cada cop més popular, són les bases de dades orientades a grafs com Neo4J. En general, les bases de dades no relacionals (per exemple,figura 4) ofereixen una gran velocitat i escalabilitat, però també és cert que requereixen una sèrie de coneixements complexos i menys disponibles en el mercat que les relacionals.

Figura 4. Tipus principals de bases de dades



Font: elaboració pròpia

4.4. Anàlisi

Malgrat que l'anàlisi de dades correspon cronològicament a la darrera fase de la cadena de valor dels *social media*, és molt important que l'analista la tingui en compte des de bon principi, de manera que es faci una implementació correcta de totes les fases anteriors. Això és així perquè, en qualsevol escenari de dades

massives, i les dades provinents dels *social media* no en són una excepció, sempre hi ha disponible una pluralitat d'estratègies d'anàlisi de dades: l'estadística inferencial, l'anàlisi de xarxes i els algoritmes d'aprenentatge automàtic són les més comunes. En tots el casos, sempre caldrà seleccionar les tècniques que aportin més valor, considerant també l'esforç necessari per aplicar-les.

La majoria d'operacions analítiques, les més habituals i recurrents tindran a veure amb la lectura d'indicadors descriptius relativament senzills. Les eines bàsiques d'estadística descriptiva podran servir aquest propòsit: mitjanes, taules de freqüències, taules de contingència o de doble entrada, visualitzacions bàsiques, etc. Moltes xarxes socials incorporen serveis bàsics d'analítica descriptiva per als seus usuaris, com per exemple Twitter Analytics o Facebook Insights. Les dades que ofereixen gratuïtament les mateixes plataformes sempre són molt limitades i amb poc marge per a la generació de valor; és per això que sol ser necessari generar un ecosistema propi de dades massives.

L'anàlisi de xarxes socials és una tècnica d'exploració empírica molt útil per analitzar l'activitat en *social media* per la seva naturalesa interconnectada. L'estudiarem en profunditat en una altra assignatura optativa del màster. Més endavant, en el segon mòdul, veurem algunes de les tècniques d'aprenentatge automàtic més importants i útils per a l'anàlisi de dades provinents de *social media*.

Bibliografia

Castells, Manuel (2009). *Comunicación y Poder*. Madrid: Alianza.

García-Alsina, Montse (2017). *Big data: gestión y explotación de grandes volúmenes de datos*. Barcelona: Editorial UOC / El Profesional de la Información.

Gomez García, José Luis; Conesa i Caralt, Jordi (2015). *Introducción al big data*. Barcelona: Oberta

Khan, M.Ali-ud-din; Uddin, Muhammad Fahim; Gupta, Navarun (2014). «Seven V's of Big Data understanding Big Data to extract value». *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education* (pàg. 1-5).

Laney, Doug (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. META Group.

Marr, Bernard (2018). «How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read». *Forbes*.

Oguntimilehin, Abiodun; Ademola, Emmanuel Ojo (2014). «A Review of Big Data Management, Benefits and Challenges». *Journal of Emerging Trends in Computing and Information Sciences* (núm. 5, pàg. 433-438).

Patgiri, Ripon; Ahmed, Arif (2016). «Big data: The v's of the game changer paradigm». *2016 IEEE 18th International Conference on High Performance Computing and Communications*(pàg.17-24).

Ponjuán Dante, Gloria (1998). *Gestión de información en las organizaciones: principios, conceptos y aplicaciones*. Santiago, CL: Universidad de Chile, Centro de Información en Capacitación.