Introducció al *big data*

José Luis Gómez García
Jordi Conesa i Caralt

PID_00234261

Universitat Oberta de Catalunya

Índex

Introducció

És difícil definir amb rigor què és el *big data* (dades massives), ja que és un concepte relativament nou que encara està en evolució. D'altra banda, com veurem més endavant, la definició més acceptada no s'implementa a partir del que és, sinó a partir de les característiques de les dades que vol analitzar.

En aquest mòdul introductori començarem descrivint els orígens del *big data* i justificarem per què el *big data* pot considerar-se com un nou paradigma a l'hora de prendre decisions i no solament una nova tecnologia relacionada amb la programació distribuïda. Finalment, es definirà *big data* i es mostrarà un exemple en què és aconsellable l'ús de tècniques *big data*.

1. Orígens

El terme *dades massives* apareix per primer cop en l'entorn de les ciències. En particular, en l'astronomia i la genètica, motivat per la gran explosió de disponibilitat de dades que van experimentar aquestes ciències durant la primera dècada del segle XXI. Alguns exemples d'això podrien ser el projecte d'exploració digital de l'espai denominat *Sloan Digital Sky Survey* o el projecte del genoma humà. El primer va generar més volum de dades durant els primers mesos de funcionament que el total de les dades acumulades en la història de l'astronomia fins aquell moment. D'altra banda, el projecte del genoma humà tenia com a objectiu trobar, seqüenciar i elaborar mapes genètics i físics de gran resolució de l'ADN humà. Cal tenir en compte que el genoma d'una persona és de l'ordre dels 100 gigabytes.

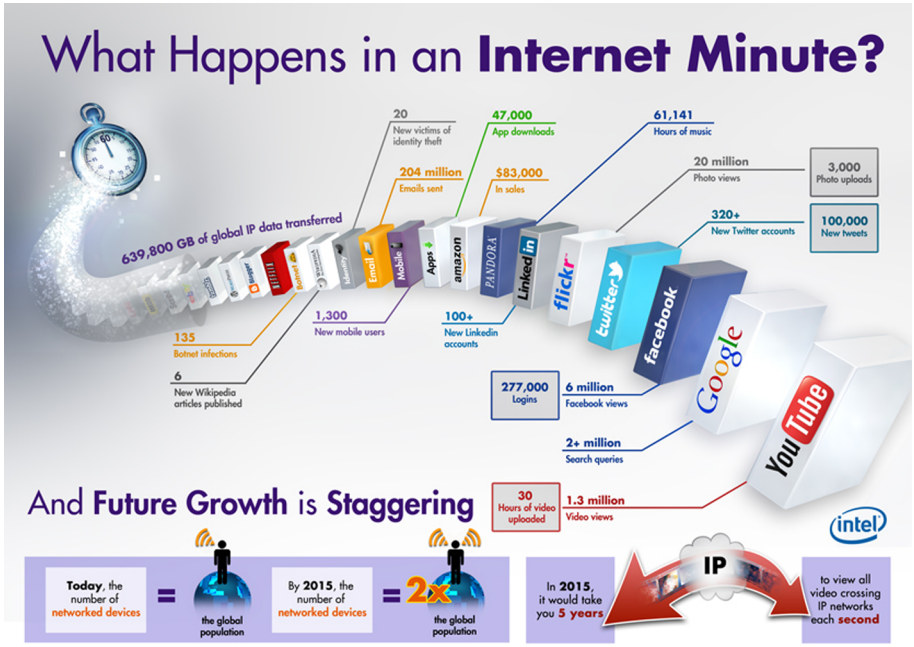
El projecte Sloan Digital Sky Survey

El projecte Sloan Digital Sky Survey té com a objectiu identificar i documentar els objectes observats a l'espai. Aquest és un dels estudis més ambiciosos i influents que s'han portat a terme mai en la història de l'astronomia. Mitjançant el processament d'imatges de gran part de l'espectre lluminós s'han obtingut llistes d'objectes observats, així com també diverses característiques i magnituds astronòmiques tals com la posició, la distància, la brillantor o l'edat. En poc més de vuit anys d'operacions s'han obtingut imatges que representen més de la quarta part del cel, fet que ha permès crear mapes en tres dimensions que contenen més de 930.000 galàxies i més de 120.000 quàsars. http://www.sdss.org/

Des d'un context més general, l'explosió de dades en aquests últims anys també ha estat una realitat. De fet, des de mitjan primera dècada del segle XXI, l'increment del nombre de dispositius amb connexió a internet, on també cal afegir l'auge de les xarxes socials, ha provocat una explosió en el volum de dades disponibles. Moltes d'aquestes dades són obertes i accessibles, cosa que permet que puguin ser explotades per qualsevol tipus d'agent, incloent-hi les empreses.

A tall d'exemple, la figura 1 mostra la quantitat de dades que es mouen a internet cada minut (dades del 2013).

Figura 1. Què passa en un minut a internet?



Font Intel: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

Les dades massives existeixen, però disposar d'una gran quantitat de dades no aporta valor en si mateix. El veritable valor de les dades es troba en llur anàlisi i interpretació, no en la seva generació. Per tant, l'aparició de les dades no solament respon a la seva disponibilitat, sinó que també respon a l'aparició de tecnologies que permetin processar-les, analitzar-les i interpretar-les.

A mesura que va anar augmentant el volum de les dades, es va fer més difícil allotjar-les a la memòria que els ordinadors feien servir per a processar-les. Això va motivar la modernització i l'evolució de les tècniques i tecnologies de processament de dades tradicionals. Una part molt important d'aquesta modernització va produir-se gràcies a les millores del maquinari dels ordinadors i al seu abaratiment, que va ajudar de manera decisiva l'entorn de les ciències, almenys pel que fa als primers projectes de dades massives. Això no obstant, amb més i millor maquinari no n'hi ha prou. Penseu, si no, com hauria de ser l'ordinador de Google per a indexar tots els continguts de la web. També han estat necessaris canvis en les tecnologies de programari per a processar una gran quantitat de dades eficientment.

L'evolució de la tecnologia basada en programari va sorgir en el si de grans empreses d'internet, com Google, Amazon i Yahoo! Aquestes empreses van adonar-se que les tècniques de processament de dades tradicionals no permetien tractar totes les dades que utilitzaven de manera eficient i van haver de crear les seves pròpies tecnologies per a poder continuar amb el model de negoci que ells mateixos havien creat. Les premisses que van seguir per al replantejament tecnològic van ser les següents:

- Hi ha una gran quantitat de dades que fa inviable el seu processament en un únic ordinador. Per tant, cal que s'utilitzi un processament distribuït

per a involucrar diversos ordinadors que treballin amb les dades de manera paral·lela. Així podran processar més dades en menys temps.

- Les dades són heterogènies i això requereix nous models de dades per a facilitar la inserció, la consulta i el processament de dades de qualsevol tipus i estructura. Aquests nous models de dades han originat noves bases de dades, anomenades *NoSQL*, que utilitzen estructures de dades diferents de les del model relacional i que permeten tractar més eficientment tipus de dades heterogènies o molt relacionades.

- Les dades han de processar-se amb rapidesa. Encara que s'hagin de processar moltes dades, el seu processament ha de ser ràpid. Per exemple, un cercador web no seria útil si la consulta que hem fet sortís un dia (o un minut) després d'haver-la fet.

Per exemple, en el cas del projecte del genoma humà, l'any 2012 l'empresa Life Technologies va presentar la seva eina The Ion Proton, la qual, seguint les premisses anteriors, era capaç de seqüenciar el genoma complet d'una persona en un dia. L'eina utilitzava tècniques de processament paral·lel i tècniques estadístiques de comparació, molt utilitzades en *big data*. De manera resumida, els passos que seguia aquesta eina per a processar el genoma humà d'una persona en un dia eren els següents:

1) Dividir el problema en subproblemes de menys volum i complexitat: seqüenciadors d'ADN digitalitzen el genoma per parts, petits fragments de la seqüència de l'ADN. Es distribueixen les parts a diferents ordinadors, que estan distribuïts de tal manera que poden processar la informació de forma paral·lela.

2) Compondre la solució final a partir de la integració de les solucions parcials dels subproblemes: a través de processament paral·lel s'ajusten totes les petites seqüències resultants de la resolució dels subproblemes per a formar la seqüència del genoma complet. En el processament s'executen diferents controls de qualitat que permeten, per exemple, arreglar possibles duplicitats i errors d'ajustament i aplicar tècniques de comparació amb els genomes d'altres individus per a detectar variacions i resoldre ambigüitats en la seqüència individual.

Aquesta tècnica de dividir un problema en problemes més petits i de menys complexitat perquè puguin tractar-se de forma paral·lela i combinar després els resultats finals respon al nom de *MapReduce* i és una de les tècniques de *big data* més utilitzades.

2. Canvi de paradigma del *big data*

Les dades massives imposen un nou paradigma en què la correlació «substitueix» la causalitat. Fins ara, els mètodes de recollida i processament de dades eren costosos i això provocava que quan es volia avaluar un fenomen no es poguessin recollir totes les dades que hi estaven relacionades. En aquests casos s'escollia una petita mostra aleatòria del fenomen, es definia un conjunt d'hipòtesis que calia comprovar i es feia una estimació d'una certa probabilitat que, per a la mostra escollida, aquestes hipòtesis fossin vàlides. Avui en dia el paradigma ha canviat, ja que és possible recollir dades de forma massiva, cosa que ens permet tenir informació sobre la mostra completa de dades (o gairebé) relacionada amb el fenomen que es vol avaluar, és a dir, tota la població. Per exemple, si una empresa vol analitzar les piulades que parlen d'ella, és perfectament factible recollir totes les que en fan menció i analitzar-les. Quan es troben correlacions entre diferents variables de la mostra (per exemple, els adults d'una regió geogràfica consumeixen més productes de l'empresa), podem explotar-les encara que no en sapiguem la causa. Trobar i provar la causa pot ser molt complicat i per al negoci no és gens necessari. Això implica un canvi de paradigma, on explicar la causalitat perd importància respecte a la correlació.

Tal com s'ha comentat, el canvi de paradigma mental provocat pel *big data* es basa en el següent:

- Ja no es tracta que la nostra experiència o intuïció ens indiqui si alguna cosa és plausible i, a posteriori, intentar confirmar-la a través de diferents enfocaments amb unes poques dades recollides per a tal efecte (la mostra).

- Ara es tracta d'ajuntar la informació disponible de tota la població en una diversitat de mitjans (xarxes socials, botigues, clients, investigació de mercats, vídeos, textos, sensors, etc.) i analitzar-la a través de diversos mètodes estadístics per a descobrir aquells fets que realment impacten en la nostra cerca, així com també les interrelacions entre aquests mateixos fets.

Aquest canvi de paradigma provoca que els sistemes analítics se centrin a trobar «quins» aspectes afecten la presa de decisions i no en «per què» l'afecten. De la mateixa manera que passa en els sistemes *business intelligence* (BI) tradicionals, es podrien respondre qüestions del tipus: «què va passar», «què està passant» i «què passaria si», però des d'un punt de vista estadístic, no causal, no es busca l'explicació del fenomen sinó només el descobriment del fenomen en si. En conseqüència, la causalitat entre els fets perd terreny en favor de l'associació (connexió, analogia, paral·lelisme i reciprocitat) d'aquests fets.

2.1. Analítica de negoci

L'objectiu principal de l'analítica de negoci és fer inferències, és a dir, fer prediccions o descobrir tendències sobre certes característiques d'una població, per a prendre decisions que repercuteixin de manera positiva en el negoci. Aquestes inferències es fan sobre la base de la informació continguda en una mostra de la població escollida de forma aleatòria. La condició d'aleatorietat és essencial per a assegurar-se que la mostra és representativa respecte de la població.

Quan es planteja una investigació estadística, el volum de la mostra és un factor crucial que cal tenir en compte. Si n'hi ha prou amb la representativitat, com més gran sigui la mostra, més exacta serà l'estimació resultant i la prova d'hipòtesi es durà a terme amb un millor criteri estadístic. Evidentment, si la mostra abraça tota la població, la generalització dels resultats obtinguts serà immediata i indiscutible.

En l'entorn *big data* podem arribar a utilitzar mostres que s'aproximen molt més al total de la població que les aproximacions tradicionals. Això és possible tant perquè som capaços de recollir més dades (observacions), com perquè som capaços de processar més quantitat de dades en menys temps.

Una altra característica, que es deu a la gran varietat de dades, és que fins i tot es poden analitzar dades que en principi no semblaven prou rellevants per a ser incloses en l'enquesta, o simplement les descartàvem per la impossibilitat de recollir-les o per la seva alta subjectivitat.

Aquests fets eleven l'anàlisi estadística, sobre dades massives, a nous nivells d'eficàcia. Això vol dir que, quan s'analitzen dades procedents d'una mostra més pròxima a la població real, podem descobrir més informació i fer-ho amb més fiabilitat. Alguns exemples que il·lustren aquest canvi de paradigma són els següents:

- Google és un dels més grans exponents a l'hora de recollir i correlacionar grans volums de dades. De fet, emmagatzema tots els criteris de cerca utilitzats pels usuaris, així com també les pàgines a les quals s'accedeix després de les seves cerques, a més de certa informació personal dels usuaris (com, per exemple, la data, l'hora, el tipus de navegador, l'idioma del navegador i l'adreça IP de cada consulta), les pàgines per les quals navega, etc.

- L'internet de les coses es basa en el fet que els objectes quotidians tinguin capacitat per a connectar-se a la xarxa, ja sia per a enviar informació sobre el seu funcionament o sobre el seu entorn (a través de sensors integrats) o per a rebre dades d'altres dispositius. L'aplicació d'aquesta filosofia de forma massiva augmentaria de manera significativa la informació que tenim sobre el món que ens envolta, ja que permetria digitalitzar i distribuir

L'adreça IP

L'adreça IP (IP és un acrònim d'*internet protocol*) és un nombre únic i irrepetible amb el qual s'identifica una computadora o dispositiu connectat a una xarxa. A internet, i en combinació amb les bases de dades de proveïdors d'accés a internet, serveix per a localitzar geogràficament de manera aproximada un dispositiu.

informació fins ara desconeguda i que podria originar correlacions fins ara insospitades.

- Analitzant les paraules clau i els enllaços seleccionats juntament amb l'adreça IP, Google ha estat capaç de predir, amb més anticipació que els organismes oficials, futures epidèmies, com, per exemple, les epidèmies de grip (Google Flu Trends, http://www.google.org/flutrends/intl/es/es/#ES). Tot això es duu a terme no pas perquè es coneguin els factors que produeixen la grip (causalitat), sinó perquè s'ha vist que una part d'una població geogràficament pròxima (localitzada a partir de la seva adreça IP) busca informació sobre símptomes o remeis per a combatre la grip (correlació). Amb aquesta informació, els mecanismes d'anàlisi de dades de Google són capaços de deduir que si molts veïns d'una determinada zona estan interessats en les causes o els remeis de la grip, és molt probable que existeixi un focus de grip en aquesta zona.

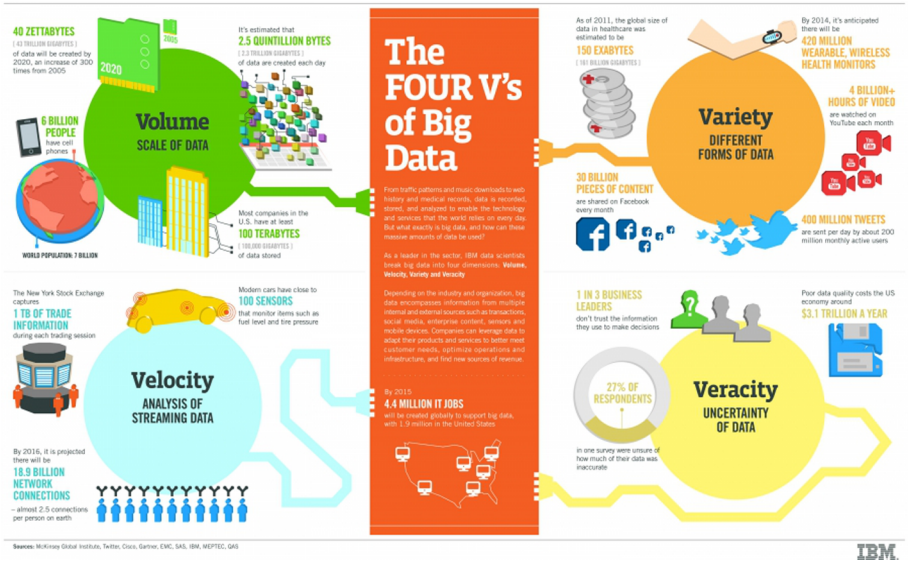
3. Definició de *big data*

El terme *big data* mira de descriure les tecnologies, tècniques i metodologies relacionades amb el processament de volums de dades grans i heterogenis.

El 2001 l'analista Doug Laney, del META Group (actualment Gartner), utilitzava i definia el terme *big data* com el conjunt de tècniques i tecnologies per al tractament de dades, en entorns de gran volum, varietat d'orígens i en els quals la velocitat de resposta és crítica. Aquesta definició, a partir de les característiques de l'entorn de les dades, es coneix com les 3 V del *big data*: volum, velocitat i varietat. Avui en dia s'accepta de manera generalitzada que la definició de les 3 V s'hagi ampliat a una quarta V, la veracitat.

L'esquema següent mostra com interactuen les 4 V del *big data* segons IBM: existeixen grans volums de dades (*volume*), d'una confiabilitat si més no discutible (*veracity*), procedents d'una gran varietat de fonts (*variety*) i que poden necessitar ser processades per a obtenir respostes ràpides (*velocity*) que ajudin a prendre més i millors decisions.

Figura 2. Les 4 V del *big data* segons IBM



Font Intel: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

A continuació es descriuen amb més detall les 4 V de la definició de *big data*.

3.1. Volum

Aquests últims anys hem viscut una gran explosió de dades. S'estima que el volum de dades existent actualment està per sobre del zettabyte i que creixerà de forma exponencial en el futur. A escala mundial, per posar un parell d'exemples il·lustratius, cada dia es creen 2,5 trilions de bytes de dades. A més, el 90% de les dades existents avui en dia s'han creat en els últims dos anys.

Zettabyte (ZB) = 1.000.000.000.000.000.000.000 bytes = 10^{21} bytes.

En l'entorn empresarial, els orígens de les dades tradicionals, com ERP, CRM o aplicacions d'RH, tenen uns requisits d'emmagatzematge molt controlats i acostumen a estar acotats a màxims de creixement d'uns pocs gigabytes diaris. Aquest és el límit de confort per a un *data warehouse* tradicional. Si després d'incloure nous orígens de dades, multipliquem el volum d'informació i sobrepassem aquest límit de confort, el rendiment del sistema podria veure's greument afectat i, per tant, caldria replantejar-se reestructurar el sistema de BI considerant un entorn de *big data*.

Enterprise resource planning

ERP (*enterprise resource planning*): sistemes informàtics de suport a la planificació de recursos empresarials. Típicament gestionen la producció, logística, distribució, inventariat, enviaments, factures i comptabilitat de la companyia de forma modular.

En la figura 3 podem veure els volums i la complexitat de dades generades pels orígens de dades més comuns en una empresa. Podem comprovar que la gran explosió de dades que origina el *big data* està relacionat amb:

Customer relationship management

CRM (*customer relationship management*): sistemes informàtics de suport a la gestió de relacions amb els clients, vendes i màrqueting.

1) l'aparició de nous orígens de dades, com podrien ser les xarxes socials, els vídeos o els sensors RFID;

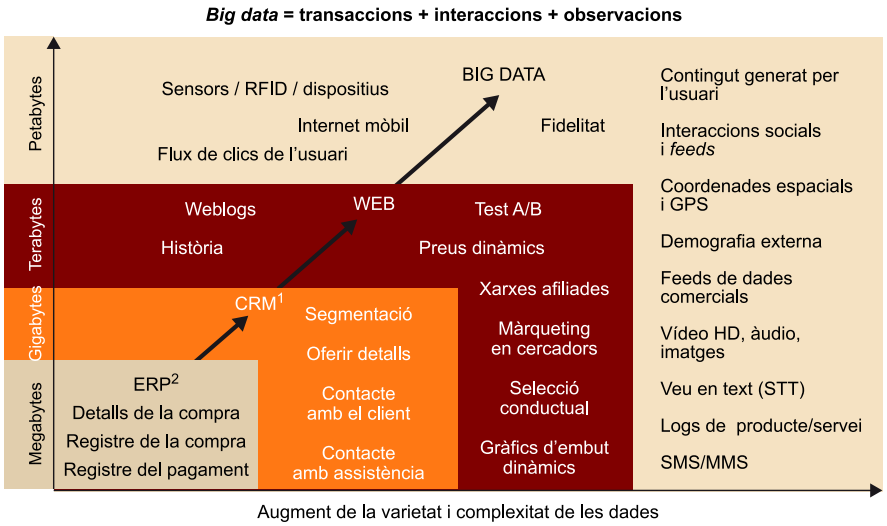
2) l'aplicació de processos analítics que fins ara no s'aplicaven de forma massiva, com, per exemple, analitzar els textos dels missatges dels clients per a valorar el seu sentiment/opinió sobre l'empresa, i

3) la recollida d'informació de dades que anteriorment es rebutjava, com, per exemple, els desplaçaments del ratolí en una pàgina web o els recorreguts dels cotxes recollits pels sistemes de geoposicionament (GPS).

RFID

Sigla de *radio frequency identification*, que en català vol dir 'identificació per radiofreqüència'. Aquesta tecnologia permet, entre altres coses, identificar, posicionar i traçar els moviments de qualsevol objecte marcat amb una etiqueta RFID.

Figura 3. Increments de volum per origen de dades



1 - Màrqueting relacional
2 - Sistema de planificació de recursos empresarials

Font: http://www.Hortonworks.com

Aquests volums addicionals podrien desbordar la capacitat d'emmagatzematge o de gestió dels *data warehouse* de l'empresa. Per exemple, si comparéssim el volum de dades de la informació d'un tiquet de compra amb el volum de dades que obtindríem si monitoritzéssim totes les operacions fetes per un caixer d'un supermercat (denominat generalment *captura d'instant* i *acció d'un TPV*), podríem conèixer, per exemple, errors freqüents, velocitat de registre de cada producte, velocitat de connexió segons el tipus de targeta de crèdit, etc. Però, d'altra banda, fàcilment estaríem multiplicant per diversos ordres de magnitud el volum de dades. És a dir, passaríem de volums mitjans diaris d'escala megabyte a escala gigabyte, per exemple.

3.2. Velocitat

En un entorn tan dinàmic com l'actual, moltes decisions han de prendre's amb gran rapidesa. El temps que triguem a reaccionar davant d'un problema és un factor tan crític com la decisió en si mateixa, i actuar tard pot ser catastròfic. Per exemple, un moviment social contra una empresa, provocat per una notícia mal interpretada, pot afectar molt negativament la fidelització dels seus clients si no es detecta a temps i s'actua amb celeritat. Les xarxes socials com Twitter poden propagar ràpidament una informació incorrecta, cosa que fa que el volum de dades s'incrementi en funció del temps que es deixi passar, per exemple.

Detectar i impedir que es porti a terme un accés indegut o un frau, recomanar un determinat producte segons la navegació d'un client, esbrinar si un client està a punt de donar-se de baixa o descobrir que una peça d'un motor està a prop del final de la seva vida útil són altres exemples que requereixen decisions que han de prendre's al moment en què es produeix el detonant que les

provoca; això significa pràcticament en temps real. Per tant, un objectiu del *big data* és tractar de proporcionar la informació necessària per a la presa de decisions en el menor temps possible.

Malgrat que es tracta d'un objectiu i d'una característica desitjable en el *big data*, tècnicament no sempre és possible treballar en temps real, ni tan sols en freqüències d'actualització pròximes.

Els sistemes de BI

Els sistemes de BI en temps real acostumen a estar emmarcats en entorns molt controlats, en què moltes vegades el sistema operacional, generador de les dades, està condicionat i dissenyat per a l'accés d'aquestes dades.

Existeixen certes barreres que cal que el *big data* superi per tal de millorar la velocitat respecte dels sistemes tradicionals. Ens referim a:

1) Velocitat de càrrega. Abans que una dada sigui analitzada ha de passar per diferents processos que la preparin, la interpretin i la integrin amb la resta de les dades. Es tracta dels processos d'extracció, transformació i càrrega (ETL), que permeten transformar, normalitzar i carregar les dades al *data warehouse*, fent que es mantinguin, a més, certs criteris de qualitat de dades. Aquests processos són costosos en temps i en recursos de maquinari i de programari.

ETL

ETL (*extract, transform and load*, que en català és 'extreure, transformar i carregar') és el procés que permet a les organitzacions moure dades des de múltiples fonts, reformatar-les, netejar-les, normalitzar-les i carregar-les en una altra base de dades, *data mart* o *data warehouse* per a analitzar i donar suport a un procés de negoci.

D'altra banda, garantir la qualitat de la dada des d'una multitud de perspectives possibles pot generar processos innecessaris o redundants segons el tipus d'anàlisi que es faci. Com més vulguem assegurar el grau de qualitat, més costosos es tornen.

Un altre aspecte que cal tenir en compte és la necessitat d'accelerar l'accés a les dades més freqüents; normalment, les dades més recents es reclamen amb més freqüència i s'exigeix una resposta més ràpida per a obtenir els informes que les contenen, davant, per exemple, de les dades de diversos anys d'antiguitat. Per tal que això sigui possible, normalment l'administrador de la base de dades mou les dades a les unitats més ràpides (o a la memòria RAM) i genera estructures per a accelerar les operacions més freqüents sobre aquestes dades. Lògicament, aquests processos d'optimització d'accés a dades, de la mateixa manera que succeeix amb els processos ETL o de qualitat, consumeixen temps i recursos en la creació de la dada, cosa que alenteix la disponibilitat inicial de la dada, això sí, en benefici d'un accés més ràpid posteriorment.

Tècniques d'optimització d'accés a bases de dades

Les tècniques d'optimització d'accés a bases de dades acostumen a estar basades en la generació i actualització d'índexs, recàlcul d'estadístiques d'accés, agregació i materialització de vistes.

Aquests processos d'ETL, qualitat i optimització de dades es duen a terme en sistemes amb un gran volum de dades, de manera que qualsevol problema que passi inadvertit en sistemes petits (segons de retard) pot convertir-se en minuts o hores en sistemes *big data*.

2) Velocitat de processament. Quan s'opera amb les dades, les funcions de consulta permeses en els sistemes gestors de bases de dades relacionals són bàsicament la selecció, la projecció, la combinació i l'agregació; a més, algunes bases de dades relacionals poden incloure altres funcions bàsiques aritmètiques i estadístiques.

Altres tipus de processament, com l'aplicació de funcions estadístiques avançades o tècniques d'intel·ligència artificial, acostumen a requerir implementacions a mida, que impliquen:

- Consulta per a l'extracció del conjunt de les dades d'interès.
- Emmagatzematge intermedi d'aquestes dades.
- Aplicació dels càlculs sobre el conjunt de les dades extretes.
- Explotació i emmagatzematge del resultat.

Pel fet que no s'executen de forma nativa en la base de dades, aquestes funcions no aprofiten al màxim els recursos del sistema, motiu pel qual es genera una pèrdua potencial de rendiment. A més, és freqüent que s'executin en un servidor diferent del de la base de dades, cosa que repercuteix negativament en la velocitat i que pot produir sobrecàrrega a la xarxa de dades, ja que les dades (recordem que són massives) han de moure's des del servidor de bases de dades al servidor responsable d'emmagatzemar i executar el programari especialitzat per al càlcul.

En aquests casos, pot ser recomanable l'ús d'un sistema de processament distribuït com *MapReduce* o un gestor de bases de dades del tipus NoSQL amb un model de dades més adequat a les necessitats concretes del sistema que s'ha d'implementar.

3.3. Varietat

La varietat es refereix als diferents formats i estructures en què es representen les dades. Definim estructura de dades com la forma com estan organitzades un conjunt de dades. Des de la perspectiva de BI, podem classificar els orígens de dades segons el seu nivell d'estructuració en:

a) Orígens de dades estructurades. La informació està representada per un conjunt de dades atòmiques elementals, que són dades de tipus simple, no compostes per altres estructures, o agrupacions d'aquestes. Es coneix amb antelació l'organització de les dades, l'estructura i el tipus de cada dada elemental, la seva posició i les possibles relacions entre elles. Les dades estructurades són de fàcil interpretació i manipulació.

Alguns exemples d'orígens de dades estructurades poden trobar-se en les bases de dades relacionals, en les aplicacions operacionals (ERP, CRM, aplicacions d'RH), o en fitxers amb una estructura fixa en forma de taula, com, per exemple, fitxers CSV o fulls de càlcul.

Fitxer CSV

Els fitxers CSV (de l'anglès *Comma-Separated Values*) són un tipus de document en format obert senzill per a representar dades en forma de taula, en què les columnes se separen per comes (o punt i coma quan la coma és el separador decimal: Espanya, França, Itàlia...) i les files, per salts de línia.

b) Orígens de dades no estructurades. Són aquells en què la informació no apareix representada per dades elementals, sinó per una composició cohesionada d'unitats estructurals de nivell superior. El valor informacional d'aquests orígens de dades tendeix a ser més gran que el dels estructurats, però la seva interpretació i manipulació resulta molt més complexa.

Alguns exemples d'orígens de dades no estructurades són textos, àudios, imatges o vídeos.

c) Orígens de dades semiestructurades. Les dades semiestructurades són aquelles que, tractant-se de dades elementals, no tenen una estructura fixa, malgrat que tenen algun tipus d'estructura implícita o autodefinida; o aquelles en les quals no totes les dades presenten una estructura elemental.

Un anunci de feina podria entendre's com una dada semiestructurada, ja que pot tenir alguns camps diferents en funció de l'anunci: una estructura lleugerament semblant en tots els anuncis i algun camp format per un text lliure o fotografia (no estructurats).

Les bases de dades tradicionals tenen limitacions a l'hora de processar dades no estructurades. Normalment permeten emmagatzemar textos, documents i arxius multimèdia, però no disposen de funcionalitats addicionals per a processar el seu contingut convenientment. Habitualment es necessiten aplicacions de tercers, o extensions de la base de dades, per a processar o visualitzar la informació no estructurada emmagatzemada a les bases de dades tradicionals. Per exemple, una base de dades pot emmagatzemar un document PDF en format binari. Consultant la base de dades es podria mostrar el conjunt de bits de què està compost, però de manera il·legible. Cal una tercera aplicació per tal que l'interpreti i el mostri de forma llegible o permeti buscar una paraula dins del document.

Algunes de les bases de dades relacionals més populars incorporen capacitats documentals que permeten un tractament de textos més eficient. Per exemple, incorporen funcions de cerca en textos i documents, així com també altres funcions de gestió documental i multimèdia, que permeten extreure metadades dels fitxers emmagatzemats, com, per exemple, el nom de la cançó i de l'autor emmagatzemats en un fitxer en format mp3. Tot i això, el seu tractament és limitat i es produeix de forma no nativa, de manera que ofereix una funcionalitat limitada i un rendiment millorable.

D'altra banda, els sistemes de BI tradicionals poden treballar amb un gran nombre d'orígens de dades diferents, però sempre assumeixen que aquests estan estructurats. Quan l'origen de dades no és estructurat, la solució més freqüent per a la seva inclusió en un sistema BI és intentar estructurar les seves dades a través d'un procés d'ETL.

Estructurar els orígens de dades no estructurades comporta una pèrdua d'informació, ja que només s'extreuen i s'emmagatzemen les qüestions que prèviament han estat considerades rellevants. Posem l'exemple d'una radiografia. Podríem crear programes que descobrissin fractures i taques i establissin la seva intensitat i posició. Les dades obtingudes serien dades estructurades i podrien emmagatzemar-se en un *data warehouse*. Això no obstant, si més endavant es descobrís, per exemple, que una determinada claredat és relativa a certa malaltia, aquesta característica no estaria recollida a la base de dades si no s'hagués emmagatzemat la radiografia original.

Estructurar els orígens de dades semiestructurades complica el procés de càrrega, ja que obliga a afegir tantes excepcions com possibilitats de variació tinguin les dades. Per exemple, a l'hora de carregar un CSV, coneixem la posició i el tipus de dada de cada columna, de manera que –llevat dels errors– podríem dur a terme una càrrega directa massiva en una base de dades relacional. Malgrat tot, a l'hora de carregar un fitxer XML, primer caldrà interpretar-lo. Això es deu al fet que en XML l'ordre de les dades no és important i que la mateixa informació es pot representar de maneres diferents. Òbviament, si algun dels camps de dades fos no estructurat, patiríem també una pèrdua d'informació en el procés d'estructuració d'aquest camp.

Fitxer XML

Fitxer XML: fitxer semiestructurat, compost per dades elementals però de definició no prèviament coneguda, sinó que inclou etiquetes per a descriure la seva pròpia definició.

3.4. Veracitat

La confiança en la veracitat de les dades és una característica que ha d'existir en qualsevol sistema de recolzament per a la presa de decisions. Prendre decisions a partir de dades errònies pot tenir conseqüències desastroses. En el *big data*, la gran quantitat i orígens de les dades provoca que la veracitat de la dada hagi de ser especialment considerada i s'hagi d'acceptar cert grau d'incertesa. A continuació descrivim en què consisteix aquest grau tolerat d'incertesa, que pot tenir origen en la veracitat (o exactitud) de la dada i en la fiabilitat del seu processament (exactitud del càlcul).

En un sistema de BI tradicional es pressuposa la veracitat de la informació, cosa que s'anomena *exactitud de la dada*. Per a satisfer aquest requisit, una gran part de la feina, tant dels desenvolupadors com dels usuaris, és assegurar la qualitat de les dades; per a tal efecte s'utilitzen tècniques i procediments com, per exemple, tècniques de neteja de dades, d'enriquiment, de mapatge, de control d'integritat, de gestió de dades mestres i de modelatge de dades. Tot això abans que les dades estiguin llestes per a l'anàlisi.

Neteja de dades

El *data cleansing, data scrubbing* o la neteja de dades és l'acte de descobrir, corregir o eliminar dades errònies d'una base de dades.

Moltes de les dades analitzades mitjançant el *big data* són intrínsecament dubtoses, relatives o amb un cert grau d'error inherent. Exemple d'això són les dades procedents de xarxes de sensors utilitzades per a fer prediccions de les condicions climàtiques. Són dades en què uns quants mesuraments es fan extensibles a zones i períodes més grans. En contrast amb els sistemes de BI tradicionals, als repositoris de *big data* rarament es porten a terme els processos de qualitat de dades, o si més no al començament no se'n fan, ja que podrien elevar el temps de càrrega o el cost del maquinari a un nivell no assumible.

En els sistemes d'ajuda per a la presa de decisions, les dades poden generar-se a partir de processos de modificació/anàlisi sobre les dades originals. En el cas dels BI tradicionals, aquests càlculs es basen en l'agregació sobre dades absolutes. L'agregació de dades és un procés determinista sense marge d'interpretació. Si les dades són veraces, la seva agrupació també ho serà. És el que es denomina *exactitud del càlcul*.

Això no obstant, una part molt important dels càlculs en *big data* es basen en mètodes analítics que permeten cert grau d'incertesa. És a dir, encara que les dades originals es considerin veraces (els comentaris d'usuaris de Facebook sobre una empresa), el resultat de la seva anàlisi pot no ser-ho (la informació sobre l'opinió dels usuaris respecte a l'empresa obtinguda automàticament a partir dels seus comentaris té una fiabilitat per sota del 100%).

La mineria de dades, el processament del llenguatge natural, la intel·ligència artificial o la mateixa estadística permeten calcular el grau de fiabilitat. Es tracta d'indicadors de la fiabilitat o l'exactitud de la predicció. Per exemple, l'error mostral o error d'estimació és l'error que es produeix quan s'observa una mostra en comptes de la població completa. No és el mateix prendre una decisió a partir d'una mostra del 50% de la població que d'una mostra de tan sols el 0,5%, motiu pel qual és un indicador que cal tenir en compte.

4. Escenari d'adopció de *big data*

A continuació es mostra un escenari a mode d'exemple amb l'objectiu de descriure una situació en la qual les tècniques i tecnologies del *big data* poden ser d'ajuda. En aquest cas no podem parlar de *big data* estrictament com a volum de dades, però sí que ho podem fer per la seva variació, veracitat i velocitat.

Suposeu una empresa en la qual, periòdicament, es fan anàlisis amb un full de càlcul que ocupa 1 GB. El PC amb el qual s'opera el full de càlcul és capaç de processar-lo, però triga gairebé una hora a obrir-lo i recalcular les noves dades.

L'empresa creix i s'hi incorporen altres divisions, cosa que provoca un gran creixement del volum de dades i, a final d'any, el full de càlcul arriba a ocupar més de 100 GB. Per aquest motiu, es decideix incorporar un PC més potent (8 nuclis de 64 bits, 128 GB de memòria, etc.) només per a l'execució del full de càlcul. Però ens trobem de nou amb la situació inicial: necessitem una hora per a carregar el full de càlcul.

A mesura que passa el temps, el volum de dades continua creixent, 1.000 GB, 10 TB, etc. Per aquest motiu, es decideix utilitzar un sistema clàssic de BI.

Conversió a bytes

Gigabyte = 10^9 = 1.000.000.000 bytes.

Terabyte = 10^{12} = 1.000.000.000.000 bytes.

Petabyte = 10^{15} = 1.000.000.000.000.000 bytes.

Exabyte = 10^{18} = 1.000.000.000.000.000.000 bytes.

Suposem que el nombre d'usuaris i que, sobretot, el volum de dades continua creixent: 100 TB, 1.000 TB, etc. Arriba un moment en què el sistema gestor de bases de dades comença a tenir problemes de rendiment. També creixen les necessitats informacionals i analítiques. Certs processos de segmentació i d'identificació de patrons requereixen diversos dies d'execució.

Un dia l'empresa es planteja llançar un nou producte. Per a analitzar la viabilitat d'aquest llançament diferents departaments porten a terme les següents tasques orientades a recollir dades sobre la potencial acceptació del nou producte:

- L'administrador de comunitats envia els missatges oportuns a les xarxes socials per a captar la impressió dels internautes sobre el nou producte.

- L'equip de màrqueting fa diferents tipus d'enquesta i dirigeix diferents *focus groups* o reunions de grup per a analitzar el llançament.

- S'utilitzen tecnologies RFID a les botigues de l'empresa amb l'objectiu de traçar els moviments dels seus clients.

- L'administrador web analitza l'activitat dels clients al portal de vendes i a la pàgina de Facebook de l'empresa.

Per tant, la informació que tindrà per a analitzar el llançament del nou producte serà la següent:

- Anàlisis tradicionals de vendes, comportament de compres, etc.

- Segmentació de clients.

- Dades facilitades per instituts d'estadística: demogràfiques, socials, econòmiques, etc.

- Els gustos dels fans a Facebook i al portal web.

- Opinions formulades pels clients o clients potencials (recollides a la web, a les xarxes socials i a les reunions de grup).

- Els desplaçaments que els clients fan per les botigues gràcies a les etiquetes RFID que duen els productes.

- Resultats d'enquestes i reunions de grup.

Aquest escenari presenta els problemes següents relacionats amb les 4 V, cosa que el fa ser un bon candidat per a aplicar les tècniques de *big data*:

1) Volum. Té molt de terreny abans no superi els límits físics de les bases de dades relacionals, tot i que se situa al límit aconsellat per al *data warehouse* de l'empresa. Utilitzar tècniques tradicionals de BI per a fer l'anàlisi podria requerir un canvi de maquinari, ja que al volum de l'actual *data warehouse* s'hi haurien d'afegir els nous orígens de dades obtingudes per a aquesta anàlisi.

2) Velocitat. Trobem processos estadístics que triguen diversos dies a executar-se. D'altra banda, la construcció dels processos ETL pot ser molt complexa a causa de la gran quantitat i varietat de dades. A més, el temps necessari per a executar els processos ETL pot ser molt elevat a causa de l'heterogeneïtat de les dades i dels seus orígens.

Reunions de grup

Tècnica qualitativa d'estudi de les opinions o actituds d'un públic utilitzada en ciències socials i en estudis comercials. Consisteix a reunir un grup de persones, normalment entre 6 i 12, amb un moderador, investigador o analista que s'encarrega de fer preguntes i dirigir la discussió. Normalment l'objectiu és avaluar el nivell d'acceptació o identificar les característiques buscades en un determinat producte o element publicitari.

Etiquetes RFID

Els *tags* o etiquetes RFID són la forma d'empaquetament més habitual dels dispositius RFID. Són autoadhesives i es caracteritzen per la seva flexibilitat, pel fet de ser «molt primes», per la capacitat de poder ser impreses amb codi humanament llegible a la seva cara frontal i per les capacitats de memòria, que dependran del circuit integrat que dugui incorporat.

3) Varietat. Existeixen diferents orígens de dades, alguns dels quals no estructurats o semiestructurats, com, per exemple, els comentaris a Facebook o els formularis de les reunions de grup, on trobem alguns camps de text lliure (els utilitzats per a recollir opinions i impressions, per a proposar millores, etc.).

4) Veracitat. Existeixen dades provinents de xarxes socials (amb faltes d'ortografia, abreviatures i interpretacions ambigües), d'enquestes (en què les respostes poden estar anotades en llocs equivocats o, de vegades, els entrevistats no hagin respost o hagin donat respostes incorrectes o inadequades) i de reunions de grup (que també poden presentar cert escepticisme, pel fet que es tracti d'una petita mostra de la població, per les inferències d'experiències prèvies o per la presència d'un líder molt marcat al grup). El fet de tractar amb aquestes dades provoca que el grau d'incertesa sigui elevat.

Les 4 V són els símptomes que indiquen la conveniència d'utilitzar un sistema de *big data* per a dur a terme una determinada anàlisi. L'anàlisi de *big data* difereix lleugerament de les anàlisis tradicionals, pel fet que s'analitzen totes les dades de les diferents fonts de dades de manera integrada. Aquest tipus d'anàlisi pot tenir algunes implicacions en els seus resultats. Seguidament comentarem els possibles avantatges de fer servir una anàlisi de tipus *big data* per a l'exemple que s'ha presentat.

En els sistemes tradicionals es tendeixen a fer diferents anàlisis a partir de cada àrea (vendes, xarxes socials, botigues, clients, investigació de mercats, etc.) a causa de la dificultat d'analitzar la combinació de totes les dades d'origen. Posteriorment, les conclusions obtingudes de les diferents anàlisis es combinen en una conclusió final. Un dels principals problemes d'aquesta manera de treballar és que les dades no es tracten en el seu conjunt, sinó que es fa des d'illes de coneixement. Això pot provocar una pèrdua d'informació pel que fa a les relacions que existeixen entre dades de diferents àrees, que poden ser rellevants i fins i tot decisives per al resultat final.

Contràriament a això, amb el *big data* es tendeixen a analitzar les dades combinades de totes les fonts d'informació. Quan es compta amb les dades combinades des del principi, es minimitza la pèrdua d'informació i s'incrementen les possibilitats de trobar noves correlacions no previstes.

A continuació veurem quines implicacions podria tenir portar a terme una anàlisi o una altra en el cas que ens ocupa.

En cas que utilitzéssim una anàlisi tradicional, s'analitzarien les dades de les reunions de grup i de les xarxes socials per separat i després s'integrarien les seves conclusions. Suposem que quan s'avaluen les reunions de grup, l'acceptació del producte ha estat majoritàriament afirmativa i que, segons l'anàlisi de les xarxes socials, el llançament del producte ha despertat un gran interès. Segons aquestes dades, els analistes podrien concloure que el llança-

ment ha de fer-se tal com s'havia proposat, ja que el producte ha agradat (conclusió de l'anàlisi del de la reunió de grup) i el seu llançament ha despertat interès (conclusió de l'anàlisi de les xarxes socials).

En cas que utilitzéssim una anàlisi més propera al *big data* s'analitzarien les dades de les reunions de grup i de les xarxes socials de forma conjunta. Suposem que a través de l'anàlisi es descobreixen dos fets:

1) Els grups als quals ha agradat menys el producte (segons les reunions de grup) coincideixen amb els grups en què hi ha hagut més interès pel que fa al llançament (segons les xarxes socials).

2) Els grups als quals ha agradat més el producte coincideixen amb els de menor penetració en xarxes socials, de manera que ni tan sols s'haurien tingut en compte en la combinació d'anàlisis. Segons aquests resultats, sembla coherent pensar que els analistes no serien tan optimistes com en el cas anterior i sol·licitarien alguns canvis en la campanya de llançament abans de treure el nou producte al mercat, ja que sembla que la campanya de llançament no ha estat enfocada al públic potencial del producte.

Un altre punt en què l'ús de *big data* pot aportar avantatges en el procés d'anàlisi és en la cerca d'informació històrica. Davant de qualsevol nova acció comercial, és habitual revisar la història de l'empresa per a trobar precedents semblants. Establir aquestes semblances permet tenir un punt de referència amb el qual poder mesurar-se i aplicar les lliçons apreses en experiències passades.

Condicionades pel tipus d'anàlisi fet tradicionalment, les bases d'informació de les empreses tendeixen a recollir dades dels resultats de les anàlisis dutes a terme en el passat, però no de les dades d'origen utilitzades en les anàlisis. Això dificulta la cerca de situacions passades similars a l'actual, ja que només es podran comparar anàlisis contra anàlisis o, en el cas més extrem, conclusions contra conclusions. Escollir experiències prèvies que no s'ajusten prou podria reforçar una idea d'èxit poc realista i forçar, per exemple, un llançament genèric a una escala desmesurada.

Al *big data* és més natural emmagatzemar totes les dades d'origen. Això facilita que es puguin fer cerques de precedents similars pel que fa a les dades, i no només a l'anàlisi o les conclusions. En el cas d'exemple, això permetria comparar les dades del llançament actual amb les dades dels llançaments anteriors. El fet de treballar amb les dades d'origen (no estructurades) permet una comparació més realista i contextualitzada.

Resum

Els canvis produïts en els últims anys ens han portat en un context en què les dades són més massives que mai, tant pel que fa al volum com al tipus d'informació que recullen i la velocitat en què es produeixen. Aquest gran volum de dades requereix noves tecnologies i noves filosofies per a emmagatzemar i processar les dades. Aquestes noves tecnologies permeten recollir i processar grans quantitats de dades i, gràcies a aquestes, dur a terme anàlisis més precises i generalitzables. A més, el gran volum de dades està potenciant l'ús de la correlació entre dades més que no construir models per a intentar explicar la causalitat.

El *big data* es defineix com el conjunt de tècniques i tecnologies per al tractament de dades, en entorns de gran volum, varietat d'orígens i en els quals la velocitat de resposta és crítica. A més, en els sistemes de BI també s'ha de tenir en compte un altre factor: la veracitat de les dades. Si prenem una decisió en funció de les dades, cal conèixer el marge d'error (o exactitud) que tenen, tant si es deu al seu origen com als processos utilitzats per a generar-les. Dit això, un problema susceptible de ser atacat mitjançant una aproximació de *big data* és un problema que satisfaci les quatre característiques descrites (volum, variabilitat, velocitat i veracitat).