Introducció al *big data*

PID_00250683

Jordi Casas Roma

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Índex

Introducció

Iniciarem aquest mòdul amb una introducció al concepte de les dades massives (*big data*) i ens centrarem en el canvi de paradigma que comporta l'arribada d'aquest tipus de dades.

A continuació, veurem una de les primeres definicions de dades massives, relacionada amb les magnituds de la dada. A partir d'aquesta primera definició n'han sorgit moltes més que, en general, amplien l'original. Així doncs, a partir d'aquesta definició inicial presentarem la que utilitzarem en aquest text i veurem alguns estàndards importants en relació amb les dades i la interconnexió de sistemes.

Finalment, introduirem un petit exemple que servirà per a il·lustrar un escenari d'ús de tecnologies de dades massives, en aquest cas concret, en una ciutat intel·ligent.

Objectius

Als materials didàctics d'aquest mòdul trobarem les eines indispensables per a assolir els objectius següents:

1. Conèixer els antecedents que han portat a l'aparició de les dades massives.
2. Comprendre el canvi de paradigma associat a les dades massives.
3. Descobrir els factors que poden fer que un problema analític pugui ser resolt emprant metodologies i eines de dades massives.

1. Antecedents i contextualització

El terme *big data* —terminologia anglosaxona àmpliament utilitzada que se sol traduir per *dades massives*— va aparèixer a principis del segle XXI en l'àmbit de les ciències, en particular de l'astronomia i de la genètica, ja que tots dos camps van experimentar una gran explosió pel que fa a la disponibilitat de dades. Per exemple, en el camp de l'astronomia, el projecte d'exploració digital de l'espai anomenat Sloan Digital Sky Survey va generar més volum de dades durant els seus primers mesos de funcionament que el total de dades acumulades en la història de l'astronomia fins al moment. En el camp de la genètica, un exemple rellevant seria el projecte del genoma humà. Aquest projecte té com a objectiu trobar, seqüenciar i elaborar mapes genètics i físics de gran resolució de l'ADN humà i genera una quantitat de dades de prop de 100 gigabytes per persona.

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. Podeu accedir a la seva pàgina web des de l'enllaç següent: http://www.sdss.org.

En aquests últims anys l'explosió de dades s'ha generalitzat en molts dels camps que envolten la nostra vida quotidiana. Entre d'altres, l'increment del nombre de dispositius amb connexió a internet i l'auge de les xarxes socials i de l'internet de les coses (IoT) han provocat una explosió en el volum de dades disponibles. A més de la gran quantitat de dades, és important destacar que moltes d'aquestes són obertes i accessibles, cosa que permet que puguin ser explotades per usuaris o institucions d'arreu del món.

IoT

L'internet de les coses o IoT (*Internet of Things*, en anglès) és un concepte que es refereix a la interconnexió digital d'objectes quotidians amb internet.

No obstant això, el simple fet de disposar d'una gran quantitat de dades no aporta valor. El veritable valor de les dades rau en la seva anàlisi i interpretació.

L'aparició de noves tècniques i tecnologies de processament de dades va sorgir a causa de la impossibilitat de processar l'enorme quantitat de dades que es generaven amb les tècniques tradicionals. Tot i que la millora i l'abaratiment del maquinari dels ordinadors permet, amb les tècniques tradicionals, carregar i processar més dades, l'augment en la quantitat de dades és de diversos ordres de magnitud superiors. Per tant, encara que puguem adquirir més maquinari i de més bona qualitat, això és absolutament insuficient per afrontar l'augment massiu de dades. Per exemple, imaginem com hauria de ser l'ordinador de Google per indexar tots els continguts del web. Per aquest motiu també va resultar necessària l'evolució de la tecnologia basada en el programari.

Les grans empreses d'internet, com Google, Amazon i Yahoo!, es van trobar amb diversos problemes importants a l'hora de continuar exercint les seves

tasques quotidianes. En primer lloc, la gran quantitat de dades que estaven acumulant feia inviable el seu processament en un únic ordinador. Així doncs, s'havia de fer servir un processament distribuït per involucrar diferents ordinadors que treballessin amb les dades de manera paral·lela i, per tant, poder processar més dades en menys temps. En segon lloc, l'heterogeneïtat de les dades va requerir nous models de dades que facilitessin la inserció, la consulta i el processament de dades de qualsevol tipus i estructura. I en tercer lloc, les empreses es van trobar amb la necessitat que les dades es processessin ràpidament, encara que n'hi hagués moltes. Per exemple, un cercador web no seria útil si retornés els resultats de la nostra cerca després d'un temps massa llarg, com pot ser després d'una espera de més d'una hora. En conseqüència, aquestes grans empreses que gestionaven grans volums de dades es van adonar que les tècniques de processament de dades tradicionals no permetien tractar de manera eficient totes les dades que utilitzaven i van haver de crear les seves pròpies tecnologies per poder continuar amb el model de negoci que ells mateixos havien creat.

Moltes de les tècniques que es van desenvolupar per donar resposta als problemes plantejats per les dades massives utilitzen un plantejament basat en el processament paral·lel. Aquest paradigma es basa en dos passos principals:

1) En primer lloc, es divideix el problema en subproblemes més petits i de menor complexitat. D'aquesta manera, es poden distribuir els diferents subproblemes en diferents ordinadors perquè cadascun s'encarregui d'un subproblema de manera independent.

2) En segon lloc, la solució final del problema es compon a partir de les solucions parcials dels subproblemes. Un cop resolt cada subproblema independentment, s'acoblen totes les petites solucions resultants per tal de crear la solució global del problema inicial.

2. El nou paradigma de les dades massives

Fins ara, si un agent o una institució volia avaluar un fenomen, generalment no podia recollir totes les dades que hi estiguessin relacionades. El motiu era que els mètodes de recollida i processament de dades eren molt costosos en temps i en diners. En aquests casos, s'escollia una petita mostra aleatòria del fenomen, es definien un conjunt d'hipòtesis que calia comprovar i s'estimava amb una certa probabilitat que, per a la mostra escollida, aquestes hipòtesis eren vàlides. Aquest és el paradigma de la **causalitat**, en el qual s'intenta establir una relació de causa-efecte entre el fenomen que s'analitza i les dades relacionades.

Lectura complementària

Viktor Mayer-Schönberger; Kenneth Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray.

Avui dia, en canvi, la recollida de dades massives ha permès obtenir informació sobre la mostra completa (o gairebé) de dades relacionada amb el fenomen que cal avaluar, és a dir, sobre tota la població. Per exemple, si una institució vol analitzar els tuits que tracten sobre un tema d'interès públic, és perfectament factible que pugui recollir tots els que parlen del tema i analitzar-los. En aquest cas, l'anàlisi no pretén confirmar o invalidar una hipòtesi, sinó establir **correlacions** entre diferents variables de la mostra. Per exemple, suposem que hi ha una forta correlació entre el lloc de residència dels veïns d'una ciutat i la seva opinió davant d'un problema determinat de la ciutat. En aquest cas, podem explotar la relació que hi ha entre les dues variables encara que no sapiguem la causa que fa que l'una porti a l'altra.

Tuits

Els tuits són petits missatges de text, limitats a 280 caràcters (originàriament 140), que permeten als usuaris de Twitter expressar el seu estat actual, informar sobre notícies o mantenir petites converses.

> Les dades massives imposen un nou paradigma en què la correlació «substitueix» la causalitat. Determinar la causalitat d'un fenomen perd importància i, en contraposició, «descobrir» les correlacions entre les variables es converteix en un dels objectius principals de l'anàlisi.

Aquest canvi de paradigma provoca que els sistemes de dades massives se centrin a trobar «quins» aspectes estan relacionats entre ells i no pas a saber «per què» ho estan. Aquests sistemes pretenen respondre qüestions del tipus: què va passar?, què està passant?, què passaria si...?, però des d'un punt de vista basat en les correlacions, ja que no es busca l'explicació del fenomen, sinó només el descobriment del fenomen en si. En conseqüència, la causalitat perd terreny a favor de l'associació entre fets.

2.1. Primera definició de dades massives

L'any 2001, l'analista Doug Laney* de META Group (ara Gartner) utilitzava i definia el terme *dades massives* com el conjunt de tècniques i tecnologies per al tractament de dades en entorns de gran volum amb una varietat d'orígens i en els quals la velocitat de resposta és crítica.

* http://gtnr.it/1bKflKH

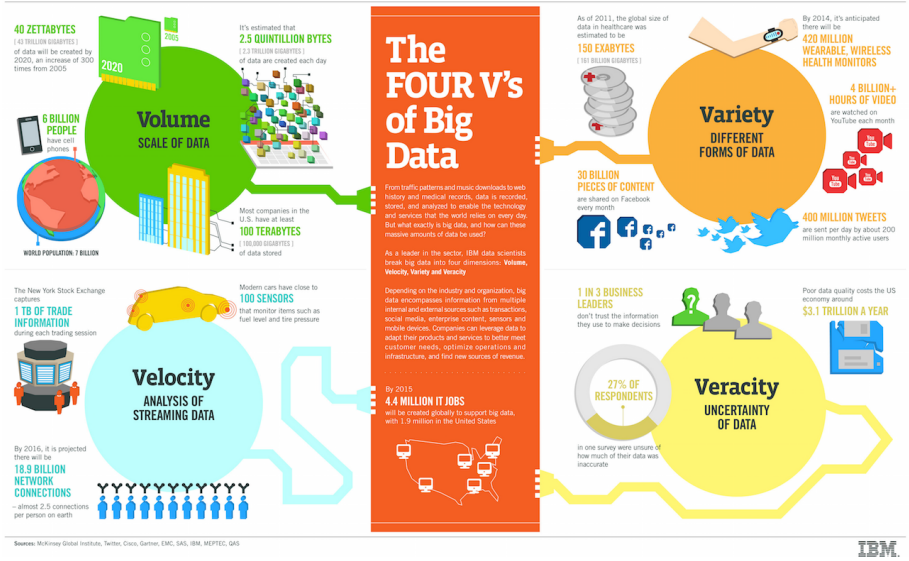
Lectura complementària

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

Aquesta definició es coneix com les tres V del *big data*: volum, velocitat i varietat. Avui dia està comunament acceptat que la definició de les tres V hagi estat ampliada amb una quarta V, la veracitat.

La figura 1 mostra la interacció de les quatre V de dades massives segons IBM: hi ha grans volums de dades (*volume*), procedents d'una gran varietat de fonts (*variety*) d'un cert grau d'incertesa (*veracity*) que pot ser que calgui processar per tal d'obtenir respostes ràpides (*velocity*).

Figura 1. Interacció de les quatre V de les dades massives segons IBM



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

A continuació veurem amb més detall les quatre V de la definició de dades massives.

2.1.1. Volum

Bases de dades relacionals

Les bases de dades relacionals emmagatzemen les dades en taules i permeten les interconnexions o relacions entre les dades de diferents taules. El llenguatge utilitzat per a consultes i manteniment es diu *Structured Query Language* (SQL).

S'estima que el volum de dades existent actualment està per sobre del zettabyte i que creixerà de manera exponencial en el futur.

Els magatzems de dades tradicionals, basats en bases de dades relacionals, tenen uns requisits d'emmagatzematge molt controlats i solen estar limitats per màxims de creixement de pocs gigabytes diaris. Si multipliquéssim el volum d'informació i sobrepasséssim aquest límit de confort, el rendiment del siste-

ma es podria veure greument afectat i, per tant, caldria replantejar-se reestructurar el sistema d'emmagatzematge i considerar un entorn de dades massives.

Zettabyte (ZB) = 10^{21} bytes = 100000000000000000000 bytes.

2.1.2. Velocitat

En un entorn dinàmic, el temps que es triga a obtenir la informació o el coneixement enfront de determinats successos és un factor tan crític com la informació en ella mateixa. En alguns casos, la informació extreta de les dades és útil mentre les dades estan «fresques», però perd valor quan les dades deixen de reflectir la realitat. La gestió del trànsit és un exemple que requereix decisions que s'han de prendre en un espai de temps breu; és a dir, pràcticament en temps real. Per tant, en alguns casos les dades massives han d'intentar proporcionar la informació necessària en el menor temps possible. Tot i que es tracta d'un objectiu i d'una característica desitjable en dades massives, tècnicament no sempre és possible treballar en temps real.

Hi ha dos tipus de velocitats que juntes condicionaran la velocitat final de resposta davant noves dades. Aquestes velocitats són:

- **Velocitat de càrrega**. Les noves dades han de ser preparades, interpretades i integrades amb la resta de dades abans de poder ser processades. Aquests processos inclouen l'extracció, la transformació i la càrrega (ETL), que són costosos en temps i en recursos de maquinari i de programari.

ETL

Extreure, transformar i carregar (en anglès, *extract, transform and load* (ETL)) és el procés que permet moure dades des de múltiples fonts, reformatejar-les, netejar-les, normalitzar-les i emmagatzemar-les al sistema utilitzat per a analitzar-les.

- **Velocitat de processament**. Un cop les noves dades estan integrades i a punt per ser analitzades, hem de considerar altres tipus de processament, com l'aplicació de funcions estadístiques avançades o tècniques d'intel·ligència artificial. Aquest processament sol implicar consultes per a l'extracció del conjunt de les dades d'interès, l'emmagatzematge intermedi d'aquestes dades o l'aplicació dels càlculs sobre el conjunt de les dades extretes, per exemple.

2.1.3. Varietat

L'estructura de dades es defineix com la manera en què es troba organitzat un conjunt de dades. La varietat es refereix als diferents formats i estructures en què es representen les dades. Segons el seu nivell d'estructuració, podem classificar els orígens de dades de la manera següent:

Fitxer CSV

Un fitxer CSV (de l'anglès *comma-separated values*) és un tipus de document en format obert que permet representar dades en forma de taula, en què les columnes se separen per comes i les files per salts de línia.

- **Orígens de dades estructurades**. La informació ve representada per un conjunt o una agrupació de dades atòmiques elementals, és a dir, dades simples que no estan compostes per altres estructures. Es coneix per endavant l'organització de les dades, la seva estructura i el tipus de cada dada elemental, la seva posició i les possibles relacions entre les dades. Les dades estructurades són de fàcil interpretació i manipulació.

Els fitxers amb una estructura fixa en forma de taula, com els fitxers CSV o els fulls de càlcul, són clars exemples d'orígens de dades estructurades.

- **Orígens de dades semiestructurades**. La informació ve representada per un conjunt de dades elementals, però a diferència de les dades estructurades no tenen una estructura fixa, tot i que tenen algun tipus d'estructura implícita o autodefinida.

  Exemples d'aquest tipus de dades són els documents XML o les pàgines web. En tots dos casos els documents segueixen certes pautes comunes, però sense arribar a un nivell d'estructuració fixa.

Fitxer XML

Un fitxer XML (de l'anglès *eXtensible Markup Language*) és un tipus de document semiestructurat, compost per dades elementals però de definició no prèviament coneguda, que inclou etiquetes per a descriure la seva pròpia definició.

- **Orígens de dades no estructurades**. La informació no apareix representada per dades elementals, sinó per una composició cohesionada d'unitats estructurals de nivell superior. La interpretació i la manipulació d'aquests orígens de dades resulten molt més complexes que les de les estructurades o semiestructurades.

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

2.1.4. Veracitat

La gran quantitat de dades i els seus orígens en les dades massives provoquen que la veracitat de cada dada s'hagi de considerar especialment i s'hagi d'acceptar un cert grau d'incertesa. Aquest grau tolerat d'incertesa pot tenir origen en l'exactitud de la dada i en la fiabilitat del seu processament (exactitud del càlcul):

- **Exactitud de la dada**. Moltes de les dades analitzades mitjançant dades massives són intrínsecament dubtoses, relatives o tenen un cert grau d'error inherent. Per exemple, les dades procedents de xarxes de sensors utilitzats per a mesurar la temperatura ambiental poden incloure un cert grau d'incertesa, atès que generalment unes poques mesures es fan extensibles a zones i períodes més grans.

- **Exactitud del càlcul**. Una part molt important dels càlculs en dades massives estan basats en mètodes analítics que permeten un cert grau d'incertesa. La mineria de dades, el processament del llenguatge natural, la intel·ligència artificial o l'estadística mateixa permeten calcular-ne el grau de fiabilitat. Es tracta d'indicadors de la fiabilitat o l'exactitud de la predicció, que pot ser inferior al 100% encara que les dades originals es considerin veraces.

  Per exemple, tot i que els comentaris dels usuaris de Facebook sobre una empresa són veraços, el resultat de la seva anàlisi mitjançant tècniques automàtiques de processament del llenguatge natural pot tenir una fiabilitat per sota del 100%.

2.2. La nostra definició de dades massives

A partir de la definició anterior, han aparegut altres definicions alternatives que han anat afegint, progressivament, més V a les definicions anteriors. Conceptes com ara la variabilitat, la validesa o la volatilitat s'han incorporat a aquesta definició de dades massives segons la proposta d'alguns autors.

El fet que hi hagi múltiples definicions complica la comprensió i la identificació d'escenaris. La gran majoria d'aquestes inclouen el que es coneix com les tres V del *big data* que hem comentat anteriorment i que són magnituds físiques de la dada.

Per tant, per tenir un enfocament pragmàtic, en aquest text farem servir la definició següent de dades massives.

Entenem per **dades massives (*big data*)** el conjunt d'estratègies, tecnologies i sistemes per a l'emmagatzematge, el processament, l'anàlisi i la visualització de conjunts de dades complexos.

A més, entendrem per conjunts de dades complexos els que, atès el seu volum, la seva velocitat o la seva varietat no poden ser tractats de manera eficient en un sistema tradicional d'anàlisi de dades.

2.3. Classificació de NIST

D'acord amb el NIST,* i en particular dins del seu grup de treball de dades massives, hi ha tres tipologies d'escenaris que requereixen l'ús d'aquests tipus de processos. Els tipus disponibles es resumeixen a continuació:

* https://www.nist.gov/

NIST

NIST és l'acrònim de National Institute of Standards and Technology, institució americana que estudia, defineix i promou estàndards tecnològics.

- **Tipus 1**, en què una estructura de dades no relacional és necessària per a l'anàlisi de dades.
- **Tipus 2**, en què cal aplicar estratègies d'escalabilitat horitzontal per a processar i analitzar de manera eficient les dades.
- **Tipus 3**, en què cal processar una estructura de dades no relacional mitjançant estratègies d'escalabilitat horitzontal per a processar i analitzar de manera eficient les dades.

Per tant, per a una determinada necessitat analítica, és possible identificar si estem en un escenari de dades massives o no, i si és necessari aquest tipus de tecnologies, fet que cada vegada més s'erigeix com un punt rellevant i de partida per a la implementació d'aquest tipus de projectes.

2.4. Estàndards en dades massives

A mesura que les dades massives han adquirit més importància per a les organitzacions i aquestes s'han començat a preocupar per com dur a terme un projecte d'aquest tipus, ha quedat palès que cal interconnectar múltiples sistemes i tecnologies. Aquesta integració i interoperabilitat de sistemes requereix estàndards de mercat.

Per exemple, dins el context de la intel·ligència de negoci i l'analítica ja existeixen estàndards com UIMA* (*Unstructured Information Management Architecture*), OWL** (*Web Ontology Language*), PMML*** (*Predictive Model Markup Language*), RIF**** (*Rule Interchange Format*) i XBRL***** (*eXtensible Business Reporting Language*) que permeten la interoperabilitat de l'analítica de dades en informació no estructurada, ontologies de models de dades, models predictius, l'intercanvi de dades entre organitzacions i regles i informes financers, respectivament.

*** https://uima.apache.org/**
**** http://bit.ly/2jdglBH**
***** http://dmg.org/**
****** http://bit.ly/2Bp6dO8**
******* https://www.xbrl.org/**

Des del 2012, diversos grups de treball de la comunitat internacional han començat a treballar en la creació d'estàndards, com ara NIST, TMForum, Cloud Security Alliance, ITU, Open Data Platform initiative (ODPi) i Common Criteria Portal.

En el cas d'ODPi, la seva recerca d'estàndards es fonamenta a proposar una configuració mínima d'Apache Hadoop que, segons el seu criteri, inclou només quatre components: HDFS, YARN, MapReduce i Ambari.

La gran majoria d'aquests grups dona suport a l'adopció efectiva de tecnologia de dades massives per mitjà del consens en definicions, taxonomies, arquitectures de referència, casos d'ús i *roadmap* tecnològics.

Taxonomia

Quan parlem de taxonomia fem referència a una classificació o ordenació en grups de coses que tenen unes característiques comunes.

3. Exemple d'escenari de dades massives

A continuació, es mostra el context d'una ciutat intel·ligent com una situació en què les tècniques i tecnologies de dades massives poden ser necessàries per a un correcte processament i anàlisi de les dades.

Suposem que la ciutat en qüestió recull les dades següents:

- Les càmeres de trànsit recullen imatges contínuament, tant les graven en format vídeo com les utilitzen per identificar a partir de la matrícula els vehicles que circulen en cada via.

- Els sensors de les zones d'aparcament exteriors proporcionen informació contínua sobre la seva ocupació i indiquen en cada moment si cadascuna de les places de la ciutat està buida o ocupada.

- Una gran quantitat de sensors repartits per tota la ciutat analitzen la qualitat de l'aire en períodes curts de temps i produeixen una anàlisi contínua de les diferents zones de la ciutat. En cada anàlisi s'inclouen molts factors relacionats amb la contaminació i els agents tòxics de l'aire.

- Els punts de transport urbà basat en l'ús compartit de bicicletes informa en cada moment sobre la seva disponibilitat en cada punt de recollida i devolució de la ciutat.

Però, a més, l'Ajuntament d'aquesta ciutat ha decidit complementar la informació que recull mitjançant els diferents sensors de la ciutat amb informació obtinguda a internet i a les xarxes socials. Entre altres, l'Ajuntament es planteja:

- Monitorar les accions dels usuaris que visiten alguna de les pàgines web municipals, de manera que registrin informació com ara les pàgines a les quals s'ha accedit, el dispositiu amb el qual s'hi ha accedit o la seva ubicació, o quan està disponible. Addicionalment, també es volen analitzar i registrar els comentaris dels usuaris en els fòrums municipals, on es discuteix qualsevol tema d'interès per a la ciutat, i de les enquestes en línia que l'Ajuntament realitza periòdicament.

- Recopilar informació de la xarxa social Twitter referent a l'estat del trànsit en qualsevol moment i punt de la ciutat. Per fer-ho, obtenen i emmagatzemen tots els tuits amb informació sobre algun dels principals carrers, vies o passejos de la ciutat.

- Emmagatzemar la informació de les interaccions dels usuaris a la pàgina municipal de Facebook, com ara el nombre de clics a m'agrada o els comentaris dels usuaris.

Aquest escenari presenta els problemes següents relacionats amb les quatre V, cosa que el fa un bon candidat per aplicar tècniques de dades massives:

1) Volum. El volum de dades generat diàriament en una gran ciutat pot superar els límits físics de les bases de dades i eines d'anàlisi tradicionals.

2) Velocitat. Les anàlisis de dades relacionades amb el trànsit han de tenir respostes ràpides que permetin detectar i corregir problemes, en la mesura del possible, de manera gairebé immediata. Una resposta tardana als problemes de trànsit és sinònim de no reaccionar.

3) Varietat. Hi ha diferents orígens de dades, algunes d'elles no estructurades o semiestructurades, com ara els comentaris al Facebook o els fòrums municipals, on trobem alguns camps de text lliure per recollir opinions i impressions.

4) Veracitat. Hi ha dades provinents de xarxes socials, d'enquestes en línia i de fòrums que poden contenir faltes d'ortografia, abreviatures i interpretacions ambigües. El fet de treballar amb aquestes dades provoca que el grau d'incertesa sigui elevat.

Les quatre V són els símptomes que indiquen la conveniència d'utilitzar un sistema de dades massives per realitzar una anàlisi determinada. L'anàlisi de dades massives difereix lleugerament de les anàlisis tradicionals, ja que s'analitzen totes les dades de les diferents fonts de dades de manera integrada. A l'hora de comptar amb les dades combinades d'arrel, es minimitza la pèrdua d'informació i s'incrementen les possibilitats de trobar noves correlacions no previstes.

Resum

En aquest mòdul didàctic hem presentat els antecedents històrics que han portat a l'aparició de les dades massives. A partir dels problemes per a la gestió de conjunts de dades complexes apareix la necessitat d'emmagatzemar i processar aquest tipus d'informació de manera més eficient.

És en aquest punt en què algunes empreses i institucions de tot el món comencen a treballar en solucions que permetin donar resposta a anàlisis de grans conjunts de dades, anàlisis en temps real (o gairebé) i anàlisis de conjunts de dades semiestructurades o no estructurades.

Aquestes iniciatives són les que han acabat definint el que avui dia coneixem com a dades massives i, encara que no tinguem una definició única i consensuada, sí que intuïtivament hi ha una definició col·lectiva més o menys general o acceptada, tot i que amb alguns matisos notables.

Després de presentar la definició inicial de dades massives i les seves implicacions, hem presentat la definició que farem servir en aquest text, a més d'algunes indicacions sobre els estàndards existents i una possible classificació de problemes que requereixen l'ús de dades massives.

Glossari

bases de dades relacionals *f pl* Bases de dades que emmagatzemen les dades en taules i permeten les interconnexions o relacions entre les dades de diferents taules. El llenguatge utilitzat per a consultes i manteniment es diu SQL (*Structured Query Language*).

dades estructurades *m pl* Dades que segueixen un patró igual per a tots els elements i que a més és conegut *a priori*. Per exemple, les dades d'un full de càlcul presenten els mateixos atributs per a cada fila.

dades no estructurades *m pl* Dades que no segueixen cap tipus de patró conegut *a priori*. Per exemple, dos documents de text o imatges.

dades semiestructurades *m pl* Forma de dades que no conté una estructura fixa predefinida *a priori*, però que conté etiquetes o altres marcadors per a separar els elements semàntics i fer complir jerarquies de registres i camps de les dades. Per exemple, els documents JSON o HTML.

Extreure, transformar i carregar *v* Procés que permet moure dades des de múltiples fonts, reformatejar-les, netejar-les, normalitzar-les i emmagatzemar-les al sistema utilitzat per a analitzar-les.
en Extract, transform and load
sigla **ETL**

fitxer CSV *m* Tipus de document en format obert que permet representar dades en forma de taula, on les columnes se separen per comes i les files per salts de línia.
en Comma-Separated Values

fitxer XML Tipus de document semiestructurat, compost per dades elementals però de definició no prèviament coneguda, que inclou etiquetes per a descriure la seva pròpia definició.
en eXtensible Markup Language

internet de les coses *m* Concepte que es refereix a la interconnexió digital d'objectes quotidians amb internet.
en Internet of Things
sigla **IoT**

taxonomia *f* Classificació o ordenació en grups de coses que tenen unes característiques comunes.

tuit *m* Petit missatge de text, limitat a 280 caràcters (originàriament a 140), que permet als usuaris de Twitter expressar el seu estat actual, comunicar notícies o mantenir petites converses.

Bibliografia

Mayer-Schönberger, Víktor; Cukier, Kenneth (2013). *Big data. La revolución de los datos masivos*. Madrid: Turner Publicaciones.

White, Tom (2015). *Hadoop: The Definitive Guide, 4th Edition*. O'Reilly Media.