



Dades massives (big data): la propera "gran cosa" en la gestió de la informació

[Versión castellana]

JULIO ALONSO ARÉVALO
Grup de recerca E-LECTRA
Universidad de Salamanca
alar@usal.es

MARTA VÁZQUEZ VÁZQUEZ
Biblioteca Abraham Zacut
Universidad de Salamanca
rulinha@usal.es

Paraules clau: Gestió de la informació, Dades massives

Keywords: Information management, Big Data

Rebut: 13/05/2016. Acceptat: 13/05/2016.

Segons la investigació del MGI i la McKinsey's Business Technology Office (Mannycia *et al*, 2011), la quantitat d'informació que genera qualsevol activitat pública o privada proporciona grans conjunts de dades, i la seva anàlisi s'ha convertit en una de les bases clau per a la competència en un futur immediat que sustentará les noves onades de creixement, productivitat, innovació i excedent del consumidor. Els líders de tots els sectors han de tenir en compte les conseqüències de la gestió adequada d'aquesta quantitat de dades ingent, no només orientada cap als que han de gestionar-les de manera directa com ara els mateixos administradors de dades, sinó també cap als que han de prendre decisions en les organitzacions. L'augment del volum i el detall de la informació capturada per les empreses, l'augment dels multimèdia, les xarxes socials i la "Internet de les coses" impulsaran un creixement exponencial de les dades en el futur previsible.

Els serveis com ara les xarxes socials, el web semàntic i intel·ligent i el comerç electrònic sovint han de manejar dades a una escala massa gran per a una base de dades tradicional. A mesura que augmenta l'escala i la demanda, també ho fa la complexitat. Afortunadament, l'escalabilitat i la simplicitat no són mútuament excloents —en lloc d'utilitzar una tecnologia de moda, cal un enfocament diferent—, ja que els sistemes que utilitzen grans quantitats de dades utilitzen moltes màquines que treballen paral·lelament per emmagatzemar i processar dades, fet que introdueix reptes fonamentals i desconeguts per a la majoria dels desenvolupadors. Les dades massives (*big data*) mostren com construir aquests sistemes utilitzant una arquitectura que aprofita els avantatges de maquinari agrupat juntament amb noves eines dissenyades específicament per capturar i analitzar dades a escala web. En aquestes dades massives es descriu l'escalabilitat, per entendre l'enfocament dels sistemes de grans volums de dades que es poden construir a partir d'un equip petit (Marz; Warren, 2012).

Gairebé tots els analistes consideren les dades massives com una de les tendències de futur que han de tenir en compte la majoria de les empreses i institucions. La societat de les tecnologies de la informació i la comunicació (TIC) propicia i requereix un diluvi universal de dades; processar-les, entendre-les i transformar-les en decisions de valor és el repte de l'anàlisi de les dades massives. Això és vital per a les empreses l'actiu de les quals és la informació. Segons estima la **International Data Corporation** (IDC) avui les dades s'incrementen un 50 % anualment, o sigui que es dupliquen cada dos anys. Un informe del Foro Económico Mundial va declarar que les dades constitueixen una nova classe d'actiu econòmic, com ara la moneda o l'or. Per la revista *Forbes*, les dades massives han estat la principal tendència tecnològica dels últims anys que es mantindrà els propers anys. Segons la IDC, fins a 2015 el seu creixement serà set vegades superior al de la mitjana de tot el sector de les TIC. En àrees tan variades com ara la ciència i els esports, la publicitat i la salut pública, s'ha produït un salt cap al descobriment i la presa de decisions a partir de les dades (López García, 2013). La tendència lligada a les dades massives també s'alimenta d'un millor accés a la informació, especialment d'ençà que la majoria de les empreses i les institucions han passat a la informàtica en núvol (*cloud computing*), la qual cosa hi facilita l'accés des de qualsevol temps i lloc mitjançant dispositius mòbils, i fins i tot que aquestes dades pugui utilitzar-les qualsevol objecte o dispositiu electrònic en l'anomenat "Internet de les coses".

Les dades massives deixen d'estar limitades al món de la tecnologia. Avui en dia es tracta d'una prioritat empresarial atesa la seva capacitat per influir profundament en el comerç d'una economia integrada a escala global. A més de proporcionar solucions a antics reptes empresarials, les dades massives inspiren noves formes de transformar processos, empreses, sectors sencers i, fins i tot, la mateixa societat. Així i tot, l'àmplia cobertura mediàtica que està rebent no ens permet distingir clarament el mite de la realitat: què està passant realment? (*Analytics*, 2014). Les empreses utilitzen les dades massives per obtenir resultats centrats en el client, aprofitar les dades internes i crear un millor ecosistema d'informació. L'anàlisi de totes les dades disponibles s'està convertint en un element de disrupció, així com en un factor de desintermediació que

afecta la cadena de valor. L'anàlisi d'informació en grans volums, de diverses fonts, a gran velocitat i amb una flexibilitat sense precedents pot suposar un factor diferencial per als que decideixin adoptar-lo (*Big Data*, 2013).

L'agència Gartner proporciona una descripció del terme en la frase següent: El *Big Data* es refereix al volum, varietat i velocitat de dades estructurades i no estructurades que s'aboquen a través de xarxes en els processadors i dispositius d'emmagatzematge, així com la conversió d'aquestes dades per a l'assessorament empresarial. Aquests elements es poden dividir en tres categories diferents: volum, varietat i velocitat.

- Volum (terabytes, petabytes i exabytes, eventualment). La quantitat cada vegada més gran de dades creada pels éssers humans i les màquines posa un repte als sistemes informàtics, que lluiten per emmagatzemar, protegir i posar a disposició tota la informació per utilitzar-la en un futur.
- Varietat. Les dades massives són també el nombre creixent de tipus de dades que han de manejar-se de manera diferent a partir de simples correus electrònics, registres de dades i de targetes de crèdit. Es reuneixen dades per a estudis científics, registres de salut, dades financeres i multimèdia: fotos, presentacions gràfiques, música, àudio i vídeo.
- Velocitat. Es tracta de la velocitat en què aquestes dades es mouen a partir de criteris de valoració en el processament i l'emmagatzematge.

La *ciència de dades* es refereix a una àrea emergent de treball que s'ocupa de la recollida, preparació, anàlisi, visualització, administració i conservació de grans col·leccions d'informació. Malgrat que el terme *dades científiques* sembla connectar més fortament amb àrees com ara bases de dades i informàtica, inclou molts tipus d'habilitats diferents —incloent-hi habilitats no matemàtiques.

Les dades massives, sense cap mena de dubte, són un dels camps més importants de treball per als professionals de les TIC. No hi ha cap àrea ni sector que no estigui afectat per les implicacions que aquest concepte incorpora; algunes eines canvien, es modifiquen estratègies d'anàlisi i patrons de mesura. Un dels reptes i oportunitats que tenen els professionals de la informació en aquest entorn és relacionat amb l'alfabetització sobre dades en biblioteques universitàries i de recerca. L'alfabetització informacional i digital a les biblioteques s'ha discutit i aplicat àmpliament en la literatura professional, però fins fa molt poc s'ha prestat poca atenció a l'alfabetització de dades. No obstant això, les noves iniciatives de govern electrònic i de dades que s'han obert en l'última dècada han creat dades públiques àmpliament disponibles que són molt interessants per a investigadors i estudiants. L'augment de la capacitat tecnològica per processar una gran quantitat de dades ofereix noves oportunitats tant per al profà com per a l'investigador. Conèixer i alfabetitzar sobre aquestes qüestions requereix un esforç de readaptació professional per fomentar una mentalitat sobre la importància d'aquestes dades i la cultura d'anàlisi, ja que es tracta de l'adopció de les noves tecnologies, fet que planteja reptes únics per als bibliotecaris. Com poden exercir les biblioteques el seu paper en aquest esforç de requalificació per desenvolupar una "mentalitat basada en dades"?

Entre les qüestions que es plantegen es troba la diferència entre què és l'alfabetització de dades i en què es diferencia de la seva contrapart, l'alfabetització informacional?, què pot aportar a un pla d'estudis l'alfabetització de dades en institucions d'educació superior? Aquest treball intenta abordar aquestes qüestions. A la part I s'examina críticament el concepte d'*alfabetització de dades* —en què es diferencia o s'assembla a altres tipus d'alfabetitzacions i per què és important. A la part II s'analiza l'alfabetització de dades en el món acadèmic, on es pot trobar una breu ressenya bibliogràfica de les pràctiques d'instrucció recents en aquest sentit. El document conclou amb un conjunt de pràctiques millors per als bibliotecaris que volen seguir l'alfabetització de dades en les seves institucions i recomanacions per a futures investigacions (*Big Data Now*, 2011).

El món científic, caracteritzat per haver de gestionar grans volums de dades, s'ha vist molt beneficiat per Big Data Analytics. Així trobem des d'aplicacions per a ciències naturals i del cosmos, com ara l'astronomia, la botànica i la geologia, fins a funcionalitats que permeten analitzar detalladament els casos i oferir tractaments més personalitzats en l'àmbit de les ciències de la salut, passant per les diferents ciències econòmiques i socials que obtenen més profit, en lucrar-se dels beneficis aportats per aquestes eines d'anàlisi de dades —estadística, economia o sociologia, entre d'altres.

Com es pregunta Mario Tascón (2013), "¿Va a ser Big Data una etiqueta más que añadir a las múltiples modas que hemos ido viendo a lo largo de los últimos años en el panorama de Internet y los desarrollos digitales o es una tendencia de fondo que está afectando en su totalidad a la evolución de la Web? ¿Se trata de un verdadero reto para las empresas en los próximos años o una nueva estrategia de marketing y *vaporware* de los proveedores tecnológicos? Big Data puede llegar a ser el activo más valioso de una organización o una de sus obligaciones más costosas, todo depende de las estrategias y soluciones que se pongan en marcha a corto plazo para afrontar el ingente crecimiento del volumen, la complejidad, la diversidad y la velocidad de los datos". Com veurem, és una tendència important per a les organitzacions i els seus processos de presa de decisions, però no afectarà de la mateixa manera, en absolut, totes les firmes i sectors. A Espanya, segons el Big Data d'IDC, prop d'un 5 % de les empreses ja utilitza aquesta tecnologia.

La *ciència de dades* es refereix a una àrea emergent de treball que s'ocupa de la recollida, preparació, anàlisi, visualització, administració i conservació de grans col·leccions d'informació. Tot i que el terme *dades científiques* sembla connectar més fortament amb àrees com ara bases de dades i informàtica, inclou molts tipus diferents d'habilitats —fins i tot les habilitats no matemàtiques. Per a alguns, el terme *ciència de dades* evoca imatges dels estadístics al laboratori que miren fixament com parpellegen les pantalles d'ordinador plenes de nombres en desplaçament. Res podria estar més lluny de la veritat. Moltes d'aquestes dades disponibles en el món tampoc són només numèriques i estructurades. En aquest context, "no estructurat" significa que les dades no estan disposades en files i columnes ordenades. Si bé és cert que les empreses, les escoles i els governs utilitzen una gran quantitat d'informació numèrica —vendes de productes, mitjana de qualificacions i avaluacions fiscals en són alguns exemples—, hi ha un munt d'altra informació en el món diferent de la que utilitzen els matemàtics i els estadístics. Malgrat que sempre és útil tenir grans habilitats matemàtiques, hi ha molt per fer en el món de la ciència de dades per als altres tipus de dades, com ara les que contenen paraules, llistes, fotografies, sons i altres tipus d'informació. A més, la ciència de dades va més enllà d'una simple anàlisi, ja que aquesta ciència ofereix una gamma de funcions i requereix una sèrie d'habilitats molt diferents (Stanton, 2013).

Ja existeix alguna aplicació com ara **Dataverse** (<http://dataverse.org/>), aplicació web de codi obert desenvolupada per la Harvard University, que permet compartir, preservar, citar, explorar i analitzar dades d'investigació. El programa facilita la presa de dades, les posa a disposició dels altres i permet replicar altres treballs d'investigació. Un repositori Dataverse allotja diversos *dataverses*. Cada *dataverse* conté un *dataset* o altres *dataverses*, i cada conjunt de dades conté metadades descriptives i arxius de dades (on figuren la documentació i el codi que acompanyen les dades). El Dataverse normalitza la citació dels conjunts de dades perquè sigui més fàcil per als investigadors publicar les dades i obtenir un millor reconeixement del treball. Quan es crea un conjunt de dades a Dataverse, es genera la citació i es presenta de manera automàtica com un marc únic de codi obert i repositori de dades d'investigació, i això fa que les dades científiques siguin tan accessibles, reutilitzables, i obertes com es pugui. Per això les dades massives també són una oportunitat, un repte i un desafiament professional per a documentalistes i bibliotecaris, que com a experts en la compilació, organització, gestió i difusió de la informació haurem d'estar alineats amb l'evolució d'aquesta proposta, que és,

sense cap dubte, una de les grans línies de desenvolupament professional. Per aquest motiu, haurem de conèixer i adquirir les destreses necessàries per saber com gestionar amb eficiència aquesta quantitat ingent de dades per dotar-les de valor en la societat de la informació (Torres i Viñals, 2012).

Bibliografia

Analytics: el uso de Big Data en el mundo real. Cómo las empresas más innovadoras extraen valor de datos inciertos (2014). IBM Institute for Business Value. <http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf (http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf) >. [Consulta: 30/12/2015].

Big Data: es hora de generar valor de negocio con los datos (2013). BBVA. <https://www.centrodeinnovacionbbva.com/sites/default/files/bigdata_spanish.pdf ([http://www.centrodeinnovacionbbva.com/sites/default/files/bigdata_spanish.pdf](https://www.centrodeinnovacionbbva.com/sites/default/files/bigdata_spanish.pdf)) >. [Consulta: 30/12/2015].

Big Data Now (2011). Cambridge: O'Reilly Media. <http://www.onmeedia.com/downloads/Big_Data_Now_Current_Perspectives_from_OReilly_Radar.pdf (http://www.onmeedia.com/downloads/Big_Data_Now_Current_Perspectives_from_OReilly_Radar.pdf) >. [Consulta: 30/12/2015].

International Data Corporation (IDC). <<http://www.idc.com/> (<http://www.idc.com/>) >. [Consulta: 30/12/2015].

López García, David (2013). *Análisis de las posibilidades de uso de Big Data en las organizaciones*. Santander: Universidad de Cantabria. <<http://repositorio.unican.es/xmlui/bitstream/handle/10902/4528/TFM%20-%20David%20L%C3%B3pez%20Garc%C3%ADaS.pdf?sequence=1> (<http://repositorio.unican.es/xmlui/bitstream/handle/10902/4528/TFM%20-%20David%20L%C3%B3pez%20Garc%C3%ADaS.pdf?sequence=1>) >. [Consulta: 30/12/2015].

Mannyica, James et al. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York: McKinsey & Company. <<http://lazowska.cs.washington.edu/escience/McKinsey.big.data.pdf> (<http://lazowska.cs.washington.edu/escience/McKinsey.big.data.pdf>) >. [Consulta: 30/12/2015].

Marz, Nathan; Warren, James (2012). *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. New York: Manning Publications. <http://www.manning.com/marz/BD_meap_ch01.pdf (http://www.manning.com/marz/BD_meap_ch01.pdf) >. [Consulta: 30/12/2015].

Stanton, Jeffrey M. (2013). *An Introduction to Data Science*. Syracuse: Syracuse University. <https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf>. [Consulta: 30/12/2015].

Tascón, Mario (2013). *Big Data*. Madrid: Fundación Telefónica. <http://www.fundaciontelefonica.com/arte_cultura/publicaciones-listado/pagina-item-publicaciones/?itempubli=264 (http://www.fundaciontelefonica.com/arte_cultura/publicaciones-listado/pagina-item-publicaciones/?itempubli=264) >. [Consulta: 30/12/2015].

Torres i Viñals, Jordi (2012). *Del Cloud Computing al Big Data: visión introductoria para jóvenes emprendedores*. Barcelona: UOC. <http://www.jorditorres.org/wp-content/uploads/2012/03/Del.Cloud_.Computing.al_.Big_.Data_.JordiTorres.ES_.pdf (http://www.jorditorres.org/wp-content/uploads/2012/03/Del.Cloud_.Computing.al_.Big_.Data_.JordiTorres.ES_.pdf) >. [Consulta: 30/12/2015].

Citació recomanada

Alonso Arévalo, Julio; Vázquez Vázquez, Marta (2016). "Dades massives (big data) : la propera «gran cosa» en gestió de la informació". *BiD: textos universitaris de biblioteconomia i documentació*, núm. 36 (juny) . <<http://bid.ub.edu/36/alonso.htm>>. DOI: <http://dx.doi.org/10.1344/BiD2016.36.1> [Consulta: 14-07-2016].

DOI: <http://dx.doi.org/10.1344/BiD2016.36.1> (<http://dx.doi.org/10.1344/BiD2016.36.1>)