Datos masivos y minería de datos sociales: conceptos y herramientas básicas

PID_00275695

Jordi Morales i Gras

Tiempo mínimo de dedicación recomendado: 3 horas

Universitat Oberta de Catalunya

Jordi Morales i Gras

Doctor en Sociología por la Universidad del País Vasco; profesor de Análisis de redes, Aprendizaje automático y Datos masivos, y socio director de Network Outsight, empresa especializada en el análisis sociológico de datos masivos.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Andrea Rosales

Primera edición: septiembre de 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Jordi Morales i Gras
Producción: FUOC

Índice

Introducción..... 5

1. El concepto de *big data*..... 7
1.1. Volumen 7
1.2. Variedad 8
1.3. Velocidad 10
1.4. Valor 11
1.5. Veracidad 12
1.6. Validez 13
1.7. Visualización 14
1.8. Virtualidad 15
1.9. Variabilidad o volatilidad 15
1.10. Complejidad 16

2. Minería de datos: explotando la cadena de valor del *big data*..... 17
2.1. Generación 17
2.2. Adquisición y limpieza 18
2.3. Almacenamiento 19
2.4. Análisis 21

3. Las herramientas del *big data*..... 24

4. Minería de datos de los *social media*..... 25
4.1. Generación 25
4.2. Adquisición 27
4.3. Almacenamiento 28
4.4. Análisis 29

Bibliografía..... 31

Introducción

Vivimos en un mundo cada vez más digitalizado. La aparición de internet, su expansión masiva a partir de la década de los noventa del siglo XX por todo el planeta —con diferencias sustanciales entre países y hemisferios— y su colonización posterior de la vida cotidiana mediante decenas de dispositivos conectados han facilitado la aparición de lo que denominamos *big data*. Pero, contrariamente a lo que se podría suponer, cuando hablamos de *big data*, o *datos masivos*, no estamos refiriéndonos meramente a una cuestión de medida o peso de los datos. En realidad, estamos caracterizando un paradigma comunicativo nuevo y propio del siglo XXI, y que comporta implicaciones de todo tipo: tecnológicas, sociales, culturales, legales y políticas.

El cambio más importante que ha comportado el proceso de digitalización del mundo ha sido el aumento de la trazabilidad de la actividad humana y no humana, lo cual ha resultado en un aumento de los sistemas de predicción automatizada. Actualmente, gran parte de aquello que hace o deja de hacer una persona, una comunidad, una máquina o un sistema complejo puede quedar registrado en una base de datos para su explotación posterior. De este modo, han proliferado y proliferan dispositivos *smart* que son capaces de registrar nuestra actividad, predecir nuestro comportamiento o nuestros deseos y hacernos todo tipo de propuestas para satisfacerlos con la máxima rapidez posible. Por ejemplo, hoy ya estamos acostumbrados a que el móvil nos proponga la mejor ruta para llegar al trabajo o que nuestro *marketplace* preferido nos diga el tipo de zapatos que compra la gente que ha adquirido los mismos pantalones que nosotros. También los edificios y las ciudades —y bien pronto los estados y las regiones— se han apuntado al paradigma *smart*, que sus analistas y gestores basan en algoritmos de inteligencia artificial cada día más complejos que los acompañan en la toma de decisiones.

Por todo esto, durante los últimos años, el volumen de datos accesibles que pueden ser procesados y analizados ha crecido exponencialmente, lo cual ha implicado una auténtica revolución para el conocimiento. De esta capacidad de acumular datos han derivado diferentes retos:

- **Retos tecnológicos y de ingeniería.** ¿Cómo pueden almacenarse volúmenes cada vez más grandes y diversos de datos? ¿Cómo pueden estructurarse y ordenarse? ¿Cómo pueden procesarse y disponerse para el análisis a una velocidad aceptable y, si es posible, en tiempo real?

- **Retos científicos e interpretativos.** ¿Cómo puede generarse conocimiento a partir de tantos datos y tan diversos? ¿Qué técnicas disponibles hay

que adaptar? ¿Qué técnicas nuevas hay que inventar? ¿Cuáles son las preguntas clave que hay que responder en el nuevo paradigma emergente?

- **Retos estratégicos y de mercado.** ¿Cómo puede traducirse el conocimiento en inteligencia competitiva? ¿Cómo puede añadirse valor a los datos? ¿Cómo pueden diseñarse sistemas de indicadores que faciliten la toma de buenas decisiones en periodos breves?

En esta asignatura nos centramos fundamentalmente en las necesidades y los retos del tercer tipo: los que nacen de los contextos de gestión de datos y de los problemas derivados de la generación de inteligencia o conocimiento aplicado. Para empezar, visitaremos y revisaremos el concepto de *big data*, que es el elemento central del paradigma comunicativo actual. A continuación, exploraremos la cadena de valor: lo haremos en términos generales en primer lugar y, finalmente, en el caso específico de los datos provenientes de los *social media*. Nos centraremos en este punto en los diferentes procesos de minería de datos sociales (*social data mining*) que pueden aportar valor a una estrategia de gestión de los *social media*.

1. El concepto de *big data*

Cuando hablamos de *big data* lo primero que hay que decir es que la medida importa, pero que no lo es todo. Es evidente que cuando el adjetivo que acompaña a *data* es *big* se está denotando cierta centralidad del aspecto del volumen, pero también es cierto que hay otras cuestiones que son tan importantes como el volumen a la hora de caracterizar este paradigma comunicativo.

Al inicio del siglo, el analista de datos Doug Laney definió las tres *v* del *big data*: volumen, velocidad y variedad (2001). A aquella definición inicial, varios analistas le han añadido nuevos conceptos que empiezan con *v* (tabla 1). Si atendemos estrictamente a las *v* que se han identificado en trabajos académicos —y dejamos de lado contribuciones elaboradas desde blogs o empresas de software como SAS u Oracle—, podemos destacar las cinco *v* y la *c* de Oguntimilehin y Ademola (2014), las siete *v* de Khan *et. al.* (2014) o las nueve *v* y la *c* de Patgiri y Ahmed (2016).

Tabla 1. Las *v* (y la *c*) del *big data* según varios autores

Laney (2001)	Oguntimilehin y Ademola (2014)	Khan *et. al.* (2014)	Patgiri y Ahmed (2016)
Volumen	Volumen	Volumen	Volumen
Variedad	Variedad	Variedad	Variedad
Velocidad	Velocidad	Velocidad	Velocidad
	Variabilidad	**Validez**	Valor
	Valor	**Veracidad**	Veracidad
	Complejidad	**Volatilidad**	Validez
		Valor	**Visualización**
			Virtualidad
			Variabilidad o Volatilidad
			Complejidad

Fuente: elaboración propia

Nota

Los conceptos en negrita son los que no se encuentran en columnas anteriores.

Es evidente que no todas las V y las C identificadas tienen el mismo valor, y también lo es que algunas apuntan a aspectos bastante similares. El «valor» es precisamente el componente que comparten las tres conceptualizaciones del *big data* hechas durante la segunda década del siglo XIX; esto es así porque el paradigma comunicativo del *big data* ha crecido como un paradigma fuertemente orientado al mercado y a la oportunidad de añadir valor a los datos.

1.1. Volumen

El uso intensivo de las nuevas tecnologías en todos los ámbitos de la vida y la gran capacidad a la hora de trazar y registrar la actividad humana y no humana tienen como consecuencia la generación de una gran cantidad de datos.

Nunca en toda la historia de la humanidad habíamos generado tantos datos como hasta ahora. Por la naturaleza descentralizada de internet, es imposible conocer con exactitud el volumen de datos que se producen cada día.

Reflexión

Aproximaciones llevadas a cabo el 2018 (Marr, 2018) apuntan al hecho de que diariamente se generan cerca de 2,5 quintillones de *bytes* o, lo que es lo mismo, cerca de 2.500 millones de *gigabytes*. Si usáramos discos de Blu-ray de una capa (25 GB por disco) para almacenar todos los datos y los apiláramos, cada día sumaríamos 1.200 metros de datos: 4 veces la Torre Eiffel de París. Y, por si no fuera suficiente, se estima que la cifra anterior se duplica cada dos años.

El paradigma de los datos masivos se caracteriza por un aumento exponencial y con carácter permanente en el volumen de los datos producidos. La recogida y el almacenamiento de los datos actualmente implican un reto de primer orden, como también la adquisición de las competencias necesarias para analizarlos. Precisamente por eso, es importante entender que la característica fundamental de los datos masivos no es el volumen, porque el volumen de los datos es relativo a la capacidad de computación de cada momento: lo que hoy consideramos una «gran» cantidad de datos es muy probable que se convierta en migajas dentro de unos cuantos años.

1.2. Variedad

La gran cantidad de dispositivos que son intensivamente utilizados para la producción de datos tiene como consecuencia directa una enorme variedad en la tipología, los formatos y la estructura de los datos que se generan. Además de datos cuantitativos (cifras) y corpus documentales (textos), el paradigma del *big data* se nutre de fotografías, vídeos, audios, coordenadas geográficas y un largo etcétera.

Según el modo de estructuración de los datos y la manera como se almacenan, es habitual clasificarlos en datos estructurados, no estructurados y semiestructurados (Gómez García y Conesa i Caralt, 2015).

Denominamos **datos estructurados** a los datos que se pueden integrar en una base de datos relacional (una base de datos MySQL, PostgreSQL o MS Access).

Una base de datos estructurada (tabla 2) tendrá una serie de campos o atributos predefinidos que serán equivalentes para cada fila o caso. Esto quiere decir que habremos planificado el formato del dato que almacenaremos antes de hacerlo: el tipo de dato, su posición, longitud, etc.

Tabla 2. Ejemplo de base de datos 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	Cartero	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

Fuente: elaboración propia

Denominamos **datos no estructurados** a los datos que no tienen una estructura fija, a pesar de que puedan tener una estructura implícita que hay que descubrir o una serie de propiedades que hay que identificar.

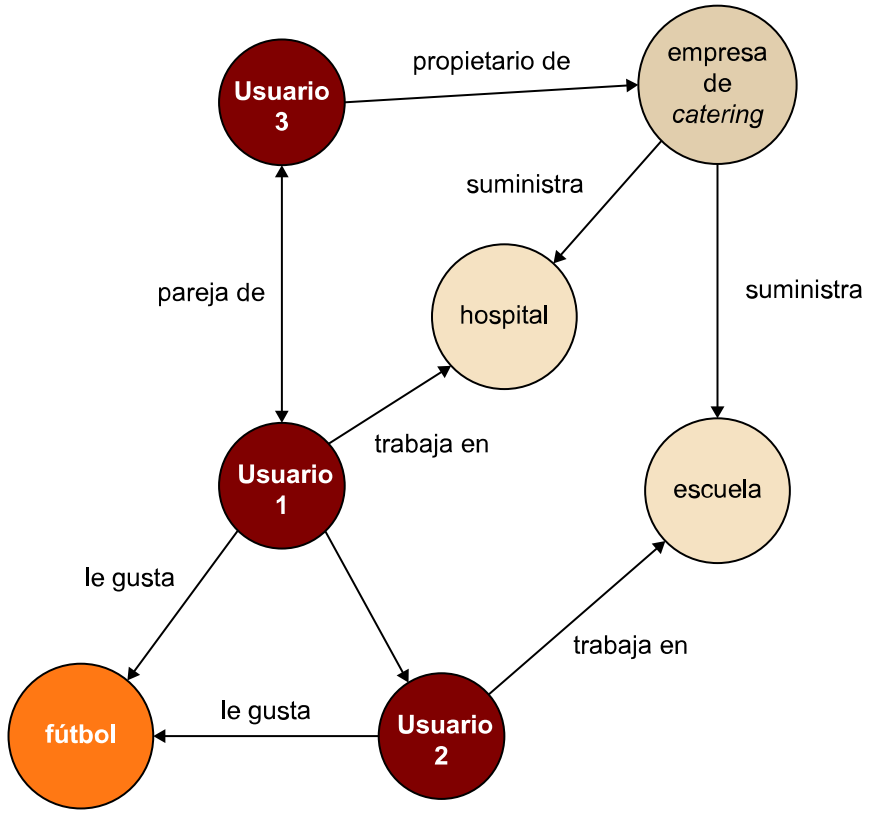
Se trata, en definitiva, de cualquier dato producido que no haya sido diseñado para encajar en un modelo de datos preestablecido o de los datos producidos espontáneamente para los cuales todavía no se ha creado un método de estructuración. Típicamente, estamos hablando de textos (por ejemplo, libros enteros, artículos, correos electrónicos, publicaciones en redes sociales y contenidos web) o de archivos multimedia (por ejemplo, fotos, vídeos y mú-

sica). Cualquier dato no estructurado puede ser estructurado, tanto por procedimientos manuales como automáticos. La gran mayoría de los datos de que disponen las empresas no son estructurados.

> Como resultado del proceso de estructuración de los datos no estructurados, nos encontramos también con la categoría de datos **semiestructurados**.

Aun así, la estructuración siempre es «semi» y nunca puede considerarse completa porque podríamos seguir aplicando más técnicas de estructuración sobre el dato no estructurado. Los datos provenientes de conversaciones en redes sociales pertenecen a esta categoría. Entre las diferentes técnicas de estructuración de los datos no estructurados destacan los lenguajes XML y JSON, muy utilizados en el paradigma del web semántico. También destacan las bases de datos NoSQL, como las bases de datos orientadas a redes o grafos (figura 1), que organizan la información en mapas conceptuales.

Figura 1. Ejemplo de base de datos orientada a grafo



Fuente: elaboración propia

1.3. Velocidad

Cuando hablamos de velocidad en *big data*, nos referimos a dos cosas diferentes:

- La rapidez en el crecimiento de las bases de datos, que, como ya hemos visto anteriormente, crecen de manera exponencial.
- La velocidad en la transferencia de datos para el análisis.

En términos generales, podemos afirmar que la velocidad de transferencia es muy importante en el paradigma de los datos masivos. Esto está estrechamente ligado con la cadena de valor de los modelos de negocio orientados a datos (*data-driven businesses*). Necesitamos poder almacenar, cargar, procesar y visualizar los datos a toda velocidad, para poder explotarlos y generar valor a partir de ellos, a veces en entornos de tiempo real.

Reflexión

La velocidad será un factor importante, por ejemplo: para gestionar la atención al cliente, para monitorizar el rendimiento de un servicio o producto, para identificar oportunidades en redes sociales o para generar contenidos periodísticos de actualidad. Y no lo será cuando haya que llevar a cabo un estudio de mercado.

Otras veces, la velocidad no será un factor tan importante o central. Cuanto más importante sea la velocidad, más rígido y previsible tendrá que ser el modelo de análisis de datos.

Para acabar, hay que considerar que la demanda de velocidad tiene efectos muy importantes en los costes de los servicios y los servidores que se utilicen.

1.4. Valor

Como ya hemos visto, la capacidad que hemos desarrollado las sociedades avanzadas a la hora de acumular información tiende al infinito. Por eso, es muy importante seleccionar adecuadamente la parte de los datos que tiene sentido guardar y prescindir de la parte de los datos que no es necesaria.

El de los datos masivos es un paradigma comunicativo fuertemente orientado al mercado. Esto quiere decir que, en la mayoría de casos, el sentido único de acumular, procesar y analizar datos es su explotación y la generación de valor. Por regla general —y dejando al margen algunas excepciones como pueden ser los usos más sociales o de investigación básica—, haremos bien de deshacernos de todo aquello que no aporte valor o que se considere que no lo puede aportar.

Además, hay que entender que la generación de valor económico o empresarial a partir de los datos es en una eventual interpretación que aporte un componente de innovación, de oportunidad de competitividad o de productividad. La pirámide informacional (figura 2) es un modelo teórico propuesto por Ponjuán Dante (1998) que pertenece a la teoría de la decisión y el aprendizaje organizativo. Según este modelo, los datos son la base para la toma de decisiones, pero son necesarias una serie de operaciones de refinación para transfor-

mar los datos en información, en conocimiento y, finalmente, en inteligencia. El valor añadido a los datos deriva del conocimiento aplicado o la inteligencia que un analista sea capaz de generar a partir de estos.

Figura 2. La pirámide informacional

Fuente: elaboración propia, a partir de Ponjuán Dante (1998)

1.5. Veracidad

Entre los riesgos más penetrantes que afectan la toma de decisiones basadas en datos nos encontramos con los datos de mala calidad, corruptos o inválidos. Así, el *big data* puede convertirse en el camino más rápido hacia el desastre más absoluto. Maximizar las condiciones de veracidad de los datos que se obtienen es una cuestión capital: es muy importante que los datos que se obtienen y procesan sean fiables y fieles a la realidad.

Son múltiples las causas que pueden comprometer la veracidad de los datos integrados en un sistema de *big data*. Las más habituales tienen que ver con errores informáticos, con el procesamiento incorrecto del ruido o las interferencias en un sistema (por ejemplo, con la estructuración incorrecta de datos no estructurados). Otras causas, menos habituales, pueden tener que ver con ataques malintencionados como, por ejemplo, cuando un pirata informático accede a una base de datos y manipula los datos para inducir un error en el sistema (por ejemplo, poner o sacar dinero de una cuenta bancaria).

No hay una fórmula mágica para garantizar la veracidad de los datos que se almacenan y procesan, puesto que esto depende en gran medida del tipo de dato que se esté procesando y de la metodología que se utilice. Siempre es adecuado evaluar los riesgos específicos de un sistema de *big data* y disponer de los mecanismos de protección, evaluación y corrección necesarios (por ejemplo, medidas de seguridad para evitar ataques y manipulaciones externas malintencionadas, sistemas de alertas para detectar anomalías). Así mismo, es im-

portante someter los métodos de tratamiento de datos a tests de estrés y hacer tantas comprobaciones manuales como sea posible antes de proceder a la automatización de los procesos.

1.6. Validez

Un aspecto relacionado con el anterior que puede comprometer seriamente la toma de decisiones basada en datos es la validez de la interpretación. Para estar seguros de que la interpretación de los datos con que contamos es válida para responder a la pregunta que queremos responder, tenemos que conocer muy bien las condiciones que han producido estos datos y tendremos que saber con certeza que estas condiciones son comparables a las actuales y que pueden conducirnos a interpretaciones igualmente válidas.

Es importante distinguir la validez de la veracidad. Mientras que la veracidad es una propiedad intrínseca del dato, la validez depende totalmente de la interpretación que quiera hacer el analista. Es perfectamente posible que un dato veraz sea inválido a la hora de establecer una interpretación.

> Imaginemos, por ejemplo, que queremos identificar cómo es de popular un usuario de Twitter o de Instagram por medio de su número de seguidores: el número de seguidores es un registro real y veraz, pero no es válido establecer una relación directa entre volumen de seguidores y popularidad, por una serie de razones, empezando por la existencia de usuarios falsos o con el comportamiento automatizado.

Otros aspectos clave que comprometen la validez de la interpretación son los siguientes:

1) Los sesgos que pueda haber en los propios datos (o sesgos estadísticos: diferencias sistemáticas entre los valores estimados y los reales).

Los sesgos estadísticos. Hay que ser conscientes de que, no por el hecho de ser masivos, los datos están ausentes de sesgos. Los datos masivos pueden estar sesgados por la manera como se han capturado, si no se han tenido en cuenta todos los tipos de sujetos de manera equilibrada y representativa (sesgo de selección). Otro tipo de sesgo se puede producir como consecuencia de problemas derivados del instrumento de captura de datos (sesgo de información).

2) Los sesgos en la mirada del analista (o sesgos cognitivos, culturales o políticos: creencias y subjetividades del analista que alteran el análisis).

En cuanto a los sesgos cognitivos, culturales o políticos, es importante que el analista entienda y gestione correctamente su relación personal con los datos que está analizando. Uno de los sesgos cognitivos más importantes en los *social media* es la tendencia a dar validez a la información que confirma las propias creencias y sacar la información que las contradice (sesgo de confirmación). Mediante esta predisposición natural humana proliferan noticias falsas y desinformación en redes.

3) Los sesgos en los algoritmos que reproducen los sesgos de su programador (sesgos algorítmicos que reflejan los valores y las creencias de su creador).

Los sesgos algorítmicos en realidad son una mezcla entre los dos anteriores: datos mal recogidos o mal procesados de manera sistemática como consecuencia de sesgos presentes en la mirada del analista que ha programado el algoritmo que recoge y procesa los datos. Veámoslo con un ejemplo:

Caso de ejemplo: Amazon

Amazon desarrolló el 2014 un algoritmo preparado para la selección de personal. Para entrenar el algoritmo, usaron una base de datos que contenía información sobre procesos de selección de personal durante diez años. La idea era utilizar la información de los últimos diez años, el perfil de los candidatos y su buen o mal rendimiento en la empresa, para crear un algoritmo predictivo capaz de proponer al empleador los cinco mejores perfiles entre una bolsa de candidatos.

El 2015 se dieron cuenta de que el algoritmo discriminaba a las mujeres, proponiendo de manera sistemática hombres para las diferentes posiciones, e incluyendo la variable «mujer» como un factor de penalización. La causa de esta discriminación no fue, en principio, una voluntad explícita y manifiesta de los programadores de discriminar las mujeres, sino una consecuencia derivada de la distribución de los datos que no fue correctamente centrada. Los datos que se habían usado para entrenar el algoritmo provenían de la industria tecnológica de los EE. UU., que está fuertemente masculinizada. De este modo, el algoritmo «entendió» que ser hombre constituía un factor de éxito en el modelo predictivo, y ser mujer un factor de fracaso.

Enlace de interés

Véase el artículo en que se explica el caso de Amazon en este enlace:
<https://www.bbc.com/news/technology-45809919>

A la hora de trabajar con datos, ya sea desde un punto de vista descriptivo o explicativo, o con la intención de generar un algoritmo predictivo, es muy importante que entendamos qué es lo que queremos conseguir y, en consecuencia, que podamos decidir si los datos de que disponemos son válidos para nuestro propósito.

Imaginemos, por ejemplo, que queremos entender el comportamiento de un cliente dentro de un comercio electrónico, o *e-commerce*. Probablemente, si trabajamos con series temporales largas habrá bastantes elementos externos al mismo dato que lo invaliden para tomar decisiones, como por ejemplo cambios en la estructura de la web o en los píxeles y etiquetas (*tags*) de monitorización de la actividad del usuario.

El factor tiempo suele ser un gran factor a la hora de «invalidar» datos, pero no es el único. En realidad, cualquier reestructuración del modelo de datos, tanto en su producción como en el almacenamiento posterior, puede invalidar un conjunto de datos, a pesar de que preserven intactos las condiciones de veracidad. Cómo en el caso de la veracidad, no hay ninguna fórmula mágica que garantice la validez de la interpretación: hay que disponer de los mecanismos y procesos adecuados para garantizarla.

1.7. Visualización

Una imagen vale más que mil palabras y probablemente un gráfico o diagrama vale más que cien mil matrices de datos. Visualizar correctamente los datos es de lo más útil para su análisis y para su comprensión.

Las propiedades de los datos que no se visualizan correctamente se convierten inevitablemente en invisibles o en tergiversaciones. Es, pues, muy importante invertir el tiempo necesario en optimizar los recursos de visualización/visibilidad de las propiedades relevantes de nuestros datos (las que aportan valor).

Respecto a las herramientas de visualización de datos, además de las herramientas habituales del análisis estadístico y demográfico[1] es importante tener en cuenta nuevas visualizaciones nacidas en los últimos años, que permiten representar ciertas propiedades y relaciones de los datos.[2] También hay que tener en cuenta que muchas de estas visualizaciones pueden presentarse en entornos de paneles de control interactivos (*dashboards*). En el último módulo, veremos todas estas cuestiones con más profundidad.

(1) Por ejemplo, histogramas, diagramas de caja y bigote, pirámides, gráficos de sectores, gráficos de dispersión, dendrogramas y mapas.

(2) Por ejemplo, diagramas de Sankey, gráficas de proyección solar, diagramas radiales o de cuerdas.

La visualización de datos es uno de los aspectos más atractivos e interesantes de la analítica de datos. Es muy importante tener siempre en cuenta que cualquier visualización de datos que se quiera implementar tiene que servir para responder a una pregunta clave y tiene que tener en cuenta la audiencia a la cual se dirige. Por eso, es importante no dejarse llevar solo por la belleza de ciertos modelos de visualización y asegurarnos de que cumplen su propósito con eficiencia.

1.8. Virtualidad

La gestión de los datos masivos no puede llevarse a cabo si no es en un entorno virtual. Resulta evidente que los antiguos archivadores analógicos no podrían servir en ninguno de los casos, pero en realidad, tampoco habría bastante con un ordenador personal convencional para almacenar, procesar y visualizar la cantidad de datos que pueden constituir un ecosistema de datos masivos.

Los servidores en red y los servicios de computación en nube (*cloud computing*) son herramientas que se vuelven necesarias en cuanto el almacenamiento de datos llega a cierto punto crítico. Además, los entornos de datos masivos suelen caracterizarse por la pluralidad de servicios que intervienen en ellos[3] y que es razonable mantener en servidores diferenciados. Por todo ello, la gestión de los datos tiene un elemento virtual inalienable.

(3) Por ejemplo, bases de datos, servicios de ETL, *dashboards* de analítica...

1.9. Variabilidad o volatilidad

Varios autores han identificado diferentes aspectos relativos a la naturaleza cambiante de los datos masivos. Mientras que el concepto de *volatilidad* se ha utilizado generalmente en un sentido negativo, señalando el cambio como un factor de incertidumbre y como posible resultado de una manipulación malintencionada, el concepto de *variabilidad* suele usarse en clave positiva, para indicar la vitalidad de los propios datos.

Las dos expresiones apuntan en una misma dirección: los datos masivos son dinámicos. Esto es así porque, por regla general, nos interesará monitorizar sistemas vivos y cambiantes para poder identificar estos cambios y encontrar oportunidades para la generación de valor.

1.10. Complejidad

Todos los factores anteriores hacen que los escenarios de datos masivos tiendan a una gran complejidad. La complejidad a que nos referimos se hace patente en los mismos protocolos de extracción y almacenamiento del dato y también en la analítica posterior.

Como hemos visto, los entornos de datos masivos nos plantean diferentes tipos de retos de diferente orden (tecnológicos, científicos y estratégicos), que se expresan por medio de todas las *v* del *big data*. Las soluciones disponibles para los retos derivados, sobre todo, del volumen y de la variedad de los datos son enormemente complejas: servidores conectados, bases de datos con varios grados de estructuración, visualizaciones complejas. Por todo esto, ha sido y es capital la configuración de nuevos perfiles profesionales preparados para afrontar —y eventualmente reducir— la complejidad del *big data*.

2. Minería de datos: explotando la cadena de valor del *big data*

Ya hemos visto que la generación de valor es un aspecto clave en el paradigma del *big data*. El valor que los datos tienen por ellos mismos suele ser muy bajo en términos económicos o empresariales, pero constituyen la materia prima a partir de la cual se puede añadir y generar más valor. El proceso a partir del cual se explota el dato como materia prima y se genera valor es conocido como *minería de datos*, y la operación consiste, en esencia, en capturar una serie de registros de información e interpretarlos para crear un patrón que nos aporte ideas accionables.

En esta sección del temario, introduciremos la cadena de valor del *big data* (García-Alsina, 2017) distribuida en cuatro fases fundamentales: generación, adquisición y limpieza, almacenamiento y análisis de datos.

2.1. Generación

Ya hemos visto que los datos masivos se caracterizan por una gran variedad en las fuentes de datos. En el contexto de la sociedad red (Castells, 2009), del mundo interconectado y de la digitalización de la vida cotidiana, buena parte de la actividad humana o no humana es susceptible de convertirse en generadora de datos: desplazamientos registrados por un GPS, rastros de navegación por internet, registros tomados por sensores urbanos (por ejemplo, *smart cities*) o industriales, el internet de las cosas (IdC) o la propia actividad en las redes sociales. Todo esto sumado a formatos y fuentes de datos también presentes en paradigmas comunicativos anteriores: información generada por las organizaciones o la administración pública, datos provenientes de estudios demoscópicos y encuestas, y un largo etcétera.

El proceso de generación de datos no consiste en «recogerlos» como si fueran alcachofas o petróleo. Los datos no son un recurso natural, sino que son un subproducto sociotecnológico que deriva de la interacción entre personas y sistemas digitales. El proceso de generación de datos está estrechamente ligado al proceso de estructuración o transformación de las actividades que registramos y que posteriormente denominamos *datos*, como también a la actividad de los usuarios. Como hemos visto antes, este proceso puede ser más o menos definitivo: mientras que los datos estructurados se presentan en un formato altamente estandarizado (filas y columnas conocidas y previsibles), los datos no estructurados o semiestructurados presentan un grado más bajo de homogeneidad.

Casi siempre, el proceso de generación de datos se lleva a cabo como proceso desvinculado respecto a las fases posteriores. Muchas veces, incluso son diferentes empresas las que se dedican a llevar a cabo los diferentes procesos de la cadena de valor del *big data*. En el caso de las empresas de redes sociales como Facebook, Twitter o LinkedIn, son las únicas encargadas de generar datos a partir de las interacciones de los usuarios con las plataformas (por ejemplo, conversaciones, relaciones de seguimiento, introducción de datos personales). El resto de agentes —corporativos o no— que quieran hacer uso de los datos que generan las redes tendrán que hacerlo desde su interfaz de programación de aplicaciones (API) y aceptando, por lo tanto, las condiciones de uso que imponga la plataforma. La información disponible por medio de las API tendrá que estar adaptada al marco legal vigente, y es la propia empresa de redes sociales la responsable principal de esta cuestión.

Una alternativa cada vez más extendida en el uso de las API es el «raspado web» o *web scraping*, que consiste en un conjunto de técnicas automatizadas que permiten extraer información de webs de manera sistemática. A pesar de que no se trata de técnicas ilegales *per se*, suelen ser técnicas no permitidas por la normativa de las plataformas generadoras de datos, y pueden plantear alguna violación de las leyes de protección de datos si no se implementan de manera responsable. Casos de raspado web muy intencionados pueden ser los comparadores de productos,[4] y casos de raspado malintencionado son la recuperación de correos electrónicos con finalidades de envío de correos basura (*spam*).

(4) Aplicaciones que disponen de robots que nos informan de ofertas u oportunidades en webs.

Tanto en un escenario de *web scraping* como de explotación de las API, hay que tener en cuenta algunas limitaciones, especialmente en cuanto a las categorías especiales de datos considerados en las diferentes legislaciones de protección de datos: legislaciones de estados y de entidades supraestatales como la Unión Europea. Típicamente, los datos no podrán utilizarse nunca con el objetivo de identificar categorías personales como la etnia, la ideología, el sentimiento religioso, la sexualidad, la salud u otros datos biomédicos o aspectos psicológicos. Es importante que el analista de datos conozca todos los límites y los respete escrupulosamente y, también, que conozca las atribuciones de su rol como responsable o como encargado del tratamiento de datos. Las diferencias entre países y continentes son importantes; por eso, será necesario tener en cuenta la legalidad vigente del territorio donde se opera.

2.2. Adquisición y limpieza

El caso de uso más habitual por la mayoría de analistas de datos será la explotación de datos generados por otra empresa y sobre la cual se tendrá un acceso limitado. Cualquiera que haga uso de la API de una aplicación o de una red social se encontrará en esta situación. En este caso, el proceso de adquisición del dato tendrá que tomar como materia prima la consulta (*query*) hecha sobre la misma API, que permitirá la recuperación de un volumen determinado de datos controlado por la empresa generadora. En *social media*, este es el caso de

las interacciones en una conversación de Twitter, las relaciones de seguimiento entre sus usuarios, las publicaciones y las interacciones en una página de Facebook o en un *hashtag* de Instagram.

La totalidad de los datos tal cual se generan (*raw data*) suele ser mucho mayor de lo que realmente se necesita para llevar a cabo cualquier análisis. El proceso de adquisición del dato es el segundo paso de la cadena de valor del dato y consiste precisamente en seleccionar los datos necesarios, transmitirlos a una base de datos propia y aplicar las operaciones de transformación necesarias. Estas operaciones serán diferentes en cada caso. Entre las más habituales encontramos las siguientes:

1) Integración de datos. El proceso consiste en combinar y unir datos provenientes de diferentes fuentes de datos y se suele llevar a cabo mediante herramientas de ETL (extracción, transformación y carga) por medio de las cuales se homogeneizan, estructuran e integran los datos, siguiendo un modelo predefinido. Recientemente, están proliferando herramientas de ELT, en las cuales se invierte el orden de los factores (extracción, carga y transformación), y que se presentan como potencialmente más escalables y versátiles. Como veremos más adelante, las herramientas de ETL integran datos en almacenes de datos (*data warehouses*), y las herramientas de ELT lo hacen en lagos de datos (*data lakes*). Una alternativa a los métodos de ETL y ELT, que consumen bastante espacio y recursos, es el método de federación de datos, que solo almacena metadatos por medio de los cuales se accede a bases de datos externas.

2) Limpieza de datos. Este proceso es clave para garantizar la calidad del dato. Se puede ejecutar de varias maneras, pero siempre con los mismos objetivos: detectar y corregir errores en los datos (inconsistencias, incoherencias y otros aspectos que pueden comprometer la veracidad del dato).

3) Eliminación de redundancias. Una parte fundamental de la limpieza de los datos es la eliminación de duplicaciones. Hay que eliminar duplicaciones y redundancias de manera sistemática para reducir costes de almacenamiento y para garantizar la veracidad de los datos.

2.3. Almacenamiento

Una vez obtenidos los datos, es el momento de almacenarlos. El objetivo principal de esta fase es poder disponer de datos cualificados con posterioridad, garantizando un acceso rápido y completo en la fase de análisis. A la hora de escoger el método óptimo de almacenamiento de datos hay que definir tres elementos: el lenguaje en que se quiere almacenar el dato, la usabilidad del dato almacenado y el tipo de visualizaciones que se tienen que derivar de este.

1) El **lenguaje de almacenamiento** del dato dependerá directamente de su grado y modo de estructuración. Cuando el tipo de dato que se quiere almacenar es estructurado, la mejor solución es una base de datos relacional de ti-

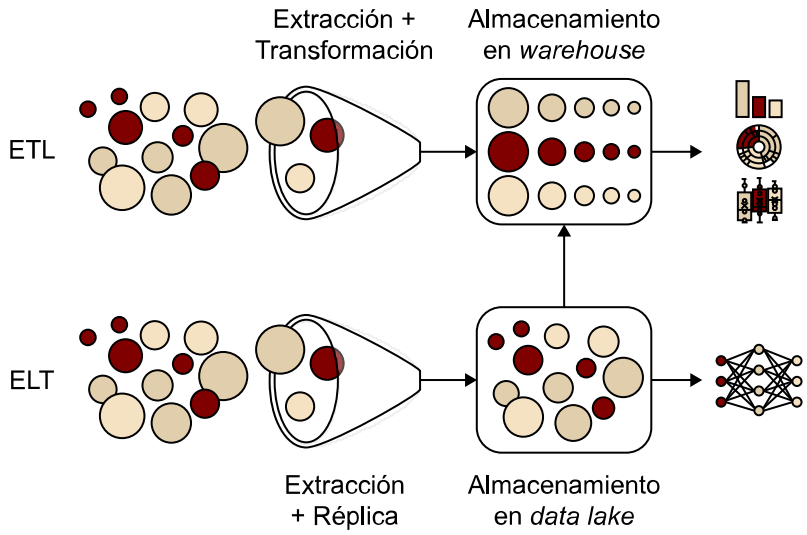
po SQL. Cuando el dato que se quiera almacenar presente una estructuración más baja (dato no estructurado o semiestructurado), será oportuno considerar lenguajes no relacionales, como JSON, o como las bases de datos orientadas a grafos.

2) Según la **usabilidad de los datos** almacenados, habrá que considerar el almacenamiento en tablas simples o en sistemas más complejos e integrados, como los almacenes de datos o los lagos de datos.

3) Finalmente, en función del **tipo de visualizaciones** que se quiera obtener de los datos, habrá que considerar los procesos y subprocesos de almacenamiento, para reducir la complejidad y permitir la carga rápida de los elementos necesarios para cada visualización.

A estas alturas, ya hemos podido ver como en *big data* todo está relacionado con todo. En este caso, el método óptimo de almacenamiento de datos depende completamente del método de integración llevado a cabo en la fase anterior de adquisición (figura 3). Si el método de integración es de ETL (extracción, transformación y carga), tendremos que almacenar datos preprocesados de manera estructurada o semiestructurada en un almacén de datos (*data warehouse* o DWH). En cambio, si el método de integración es de ELT (extracción, carga y transformación), almacenaremos los datos sin tratar (*raw data*) en su formato primitivo en un lago de datos (*data lake*). A su vez, la elección de uno de los métodos de almacenamiento tendrá consecuencias muy importantes en el análisis de los datos. Las exploramos a continuación.

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



Fuente: elaboración propia

2.4. Análisis

La fase de análisis es el cuarto y el último eslabón de la cadena de valor del *big data*. Aun así, tendrá que ser el primero en el razonamiento del analista: es en función del tipo de valor que se quiere extraer que hay que diseñar el resto de fases de la cadena.

Todos los procesos anteriores al análisis de datos tendrán que estar dispuestos de forma que optimicen un análisis orientado a responder las preguntas clave de cada proyecto.

A pesar de que las empresas siempre han dispuesto de datos con finalidades analíticas (por ejemplo, libros de contabilidad, ficheros de clientes), no ha sido hasta hace pocos años que los adelantos tecnológicos y la necesidad de inteligencia competitiva han hecho emerger los modelos de negocio orientados a datos (*data-driven business*). La disciplina que se encarga del análisis de los datos almacenados en sistemas de *big data* es la ciencia de datos (*data science*), y se nutre tanto de los adelantos recientes en ingeniería informática como de conocimientos estadísticos y de inteligencia artificial consolidados desde los siglos XIX y XX, respectivamente. Las siguientes son las herramientas principales de análisis de datos:

1) Estadística. La estadística es la ciencia matemática relacionada con la recopilación, análisis, interpretación y representación de datos. La estadística descriptiva sirve para resumir y presentar los conjuntos de datos: es, con mucha diferencia, la más utilizada, por su sencillez y facilidad de comprensión. La estadística inferencial sirve para extraer conclusiones a partir de datos disponibles. Su manera de construir conocimiento es refutando hipótesis, y es una disciplina fuertemente orientada a la explicación de las relaciones entre variables (formalización). En función de la naturaleza de estas variables,[5], la estadística dispone de diversas técnicas de validación hipotética, como la regresión, el análisis de varianza o el análisis de componentes principales, que se basan en la teoría de la probabilidad. El grueso de la estadística se fundamenta en el individualismo metodológico y se centra, por lo tanto, en el análisis de las propiedades y los atributos de los casos individuales que analiza.

(5) La diferencia más importante es entre variables cualitativas y cuantitativas.

2) Análisis de redes. Se trata de una técnica arraigada en una perspectiva estructural, diferente, por lo tanto, del individualismo metodológico, y orientada a la interpretación de las relaciones entre casos individuales y no tanto en sus propiedades o atributos. Matemáticamente, el análisis de redes se fundamenta en la teoría de grafos,[6] que ha disfrutado de menos popularidad durante los últimos siglos, pero que con la eclosión del mundo digital ha ganado prominencia. El análisis de redes se nutre simultáneamente de disciplinas como las matemáticas, la sociología o la biología, y sitúa las relaciones —y no los atributos— entre los nodos de una red como elemento central de análisis.

(6) La teoría matemática que estudia el concepto de *red*.

En el análisis de redes, se combinan una serie de elementos cuantitativos y cualitativos y hay una fuerte orientación hacia el componente visual mediante la representación de grafos o redes. El análisis de redes cuenta con algoritmos y métricas propias que permiten evaluar las propiedades de cada nodo, como también de la red en conjunto. En otra asignatura optativa del máster, profundizaremos en esta técnica de análisis.

3) Inteligencia artificial y aprendizaje automático. La inteligencia artificial (IA) es cualquier técnica que capacite un ordenador para llevar a cabo una o varias acciones que aparenten o emulen alguna de las dimensiones de la inteligencia humana. El aprendizaje automático (*machine learning*) es la subdisciplina de la IA que persigue la mejora del propio sistema mediante la experiencia. Hay varios tipos de algoritmos de aprendizaje automático. Veamos algunos:

a) El **aprendizaje supervisado** persigue la predicción de resultados futuros en función de series de datos conocidos. En contraste con la estadística inferencial, se pone el énfasis en la predicción (la estimación de valores futuros) y no en la formalización (el estudio de las relaciones entre variables). De hecho, es habitual que muchos algoritmos funcionen con «cajas negras» que no permitan estudiar las relaciones entre variables. Entre los algoritmos de aprendizaje supervisado más populares destacan la regresión (lineal, polinómica o logística), los árboles de decisión o el aprendizaje bayesiano.

b) El **aprendizaje no supervisado**, en cambio, busca clasificar los datos en función de sus propiedades intrínsecas, sin partir de un modelo predictivo preestablecido. Muchas veces se utilizan estos algoritmos con finalidades descriptivas (por ejemplo, para descubrir patrones de agrupación de casos) o en modelos de investigación inductiva (el desarrollo teórico nace de la observación y no del contraste hipótetico-deductivo). Entre los algoritmos más populares, destacan la agrupación *k-means* o el análisis de componentes principales, que también se usa en estadística inferencial.

c) Los **algoritmos de conjunto** combinan varios tipos de algoritmos para mejorar el poder predictivo de un modelo. Se trata de algoritmos preparados para resolver problemas de aprendizaje supervisado o no supervisado, pero que por su manera de proceder no forman parte de ninguno de los dos bloques anteriores. Los más populares son los algoritmos de impulso adaptativo (*adaptative boosting* o *AdaBoost*) y de bosques aleatorios (*random forests*), los dos basados en árboles de decisión.

d) Los **algoritmos de aprendizaje profundo** (*deep learning*) constituyen el campo de estudio más prometedor del aprendizaje automático desde el punto de vista de los resultados que obtienen, a pesar de que comporten ciertos problemas de interpretación que veremos más adelante. Se trata de algoritmos de

caja negra que establecen relaciones entre casos a varios niveles, y que podrían utilizarse para resolver problemas, sobre todo, de aprendizaje supervisado. Los algoritmos de redes neuronales son los más característicos.

El análisis de datos constituye, sin duda, el eslabón más importante de la cadena de valor del *big data*. En el proceso de diseño de la arquitectura de un sistema de datos masivos, es crucial evaluar con mucha precisión las necesidades analíticas específicas que se tienen, puesto que de estas necesidades derivarán las decisiones que tendrán que configurar la totalidad del sistema: la selección de las fuentes de datos, su tratamiento y el método de almacenamiento. La diferencia entre los sistemas de ETL y de ELT será crucial en este punto, puesto que el tipo de algoritmos a aplicar tendrá que ver con el grado de estructuración del dato y con su naturaleza bruta o preprocesada.

El análisis es también el área del *big data* que necesita ser más interdisciplinaria y transdisciplinaria. Es imposible que un equipo sin conocimientos sustantivos sobre el campo de estudio clave para el proyecto (por ejemplo, sobre comunicación, sobre sociología, sobre biología, sobre epidemiología, sobre literatura, sobre ciencias ambientales...) pueda generar valor a partir de la explotación de datos masivos, y también lo es que lo haga un equipo sin conocimientos técnicos, procedimentales y metodológicos (por ejemplo, informáticos, estadísticos, de ciencia de datos...). El *big data* constituye, así, un llamamiento a la comunidad científica en conjunto y a profesionales de todo tipo y no solamente a ingenieros informáticos y gestores de bases de datos.

3. Las herramientas del *big data*

A la hora de poner en marcha un ecosistema de datos masivos, el cómo es tan importante como el qué. Actualmente, hay varias empresas que ofrecen servicios integrales de *big data*. Las más importantes son **Amazon**, **Google**, **Microsoft**, **IBM** y **SAP**. Todas ofrecen varios servicios bajo varias modalidades de contratación (por ejemplo, *freemium*, coste por uso o licencias temporales) y que se ajustan en varios presupuestos. Por otro lado, también hay una serie de programas libres y de código abierto que cumplen varias funciones que hemos definido en la cadena de valor, a pesar de que ninguno no se puede considerar un servicio integral de *big data*:

- **Hadoop**, **Apache Spark** o **Red Hat** son marcos de trabajo o *frameworks* de computación en nube que permiten la computación distribuida: la utilización de varios ordenadores conectados en red para resolver problemas de computación masiva.

- **Spark SQL**, **Hive** o **Presto** son infraestructuras para el almacenamiento de datos relacionales basados en lenguaje SQL.

- **Talend**, **Pentaho** u **Oozie** son servicios de ETL que se integran fácilmente en *frameworks* de computación en nube.

- **Python**, **R** o **Scala** son lenguajes de programación con muy buenas capacidades para el análisis de datos masivos. También disponen de herramientas y complementos para la visualización de datos.

También hay un gran número de soluciones de software de propiedad que suplen varias necesidades no previstas por el software libre, como por ejemplo soluciones específicas por sectores (por ejemplo, publicidad, educación, inmobiliaria, etc.) o por departamentos específicos (por ejemplo, comercial, *marketing*, legal, etc.). Un punto muy sensible en el que el software de propiedad, actualmente, es mucho más fuerte que el libre es en la visualización de datos y las herramientas de inteligencia de negocios (*business intelligence*). Las plataformas principales de este mercado específico son Tableau, Looker, Microsoft Power BI y Qlik.

El paisaje de software y empresas de *big data* es enormemente variable y cambiante. Año tras año, surgen más y mejores sistemas y servicios, que obligan a todos los profesionales del sector a desarrollar un permanente estado de alerta. El analista de mercados Matt Truck publica cada año un *Data & AI Landscape* que puede ser muy útil a la hora de situarnos.

4. Minería de datos de los *social media*

Como ya hemos visto anteriormente, uno de los rasgos principales de lo que denominamos *big data* es la variedad en el formato y en las fuentes de datos. El paradigma de los datos masivos se nutre simultáneamente de datos que provienen de los servicios financieros, del comercio, del sector industrial, del sector de la salud, etc., y muy significativamente, se nutre del mundo de las telecomunicaciones y de las redes sociales de internet.

Enlace de interés

Internet y los *social media* son los productores principales de *big data*. La empresa DOMO publica regularmente la infografía *Data Never Sleeps*, en que consta todo lo que ocurre en Internet en un minuto: <https://www.domo.com/learn/data-never-sleeps-7>

En este apartado, veremos la cadena de valor de los datos que se producen en los *social media* y las redes sociales de internet. Veremos donde se generan los datos —y quienes los generan— y de qué manera son dispuestos por las empresas de redes sociales para su posterior adquisición, almacenamiento y análisis. Repasaremos, así, el proceso de minería de datos aplicado a los *social media* mediante cada una de las fases de la cadena de valor y viendo las características principales en el caso de los *social media*.

4.1. Generación

El primer paso en la cadena de valor del *big data* es la generación. Antes ya hemos visto que esta fase en la mayoría de ocasiones se lleva a cabo como proceso desvinculado de las fases posteriores. Esto quiere decir que, por regla general, los explotadores y analistas de datos querrán explotar y analizar datos generados fuera de sus ecosistemas o, incluso, integrar datos provenientes de una variedad de ecosistemas. En el caso de los *social media*, cada ecosistema podría ser una red social.

Desde un punto de vista conceptual —e incluso ético— también es importante entender que los datos no son creados por las empresas de *social media* de manera autónoma y autosuficiente, sino que son un subproducto tecnológico que integra la actividad de los usuarios y el diseño tecnológico de los sistemas, con sus implicaciones legales y éticas. Es la actividad y la interactividad de estos usuarios, es decir, sus publicaciones, sus relaciones, sus *likes*, sus retuits, sus *swipes*, etc., sumadas al modo de estructuración de los propios sistemas, lo que permite a las empresas registrar y empaquetar aquello que denominamos *datos*, tal como los conocemos y tal como son dispuestos en las API.

Muchas de las redes sociales disponen de API propias por medio de las cuales los explotadores y analistas de datos pueden acceder a ellas, cada una de las cuales mantiene su política de datos. Algunas redes apuestan por dar acceso a publicaciones y datos de usuarios (por ejemplo, Twitter e Instagram), otros solo sobre los usuarios (por ejemplo, LinkedIn), y otros hacen distinciones según perfiles y roles de usuario (por ejemplo, Facebook). El formato y el volumen de datos que las empresas generadoras deciden poner a disposición

de los explotadores y analistas es variable y sujeto a políticas empresariales y marcos legislativos en revisión permanente. Por todo esto, se trata siempre de un terreno complejo y de una fuente importante de incertidumbre para las empresas de datos masivos.

El tipo de datos que las diferentes API deciden poner o no poner a disposición de los explotadores y analistas pueden pertenecer a diferentes categorías. Como ya hemos visto, no todas las categorías están disponibles en todas las redes:

1) Perfiles de usuario. Se trata de información sobre el propietario de una cuenta. Algunas redes disponen de más datos que otras. Por ejemplo, mientras que en Facebook y LinkedIn esta información suele ser muy completa (género, aniversario, estado civil, estudios y un larguísimo etcétera), en Twitter o Instagram ni siquiera se conoce el género del usuario. Esto no quiere decir que Twitter o Instagram no dispongan internamente de algoritmos para identificar elementos como el género. Por regla general, cuanta más información potencialmente sensible tenga una empresa de *social media*, más cerrada será su API.

2) Conexiones. Un aspecto clave de las redes es el conjunto de relaciones que cada usuario establece. Estas relaciones tienen que ser necesariamente bidireccionales o unidireccionales:

- **Bidireccionales**: las amistades en Facebook y los contactos en LinkedIn.
- **Unidireccionales**: los seguidores y seguidos en Twitter e Instagram.

Como veremos más adelante, este elemento de direccionalidad tendrá mucha relevancia a la hora de analizar estas relaciones. Actualmente, solo Twitter proporciona este tipo de dato por medio de su API. Además de cuestiones más evidentes como es la identificación de grupos de usuarios, el análisis de las relaciones de un usuario es muy informativo respecto de sus gustos y aficiones, incluso de gustos y aficiones no declarados o sobre los cuales el usuario no habla.

3) Publicaciones. Los contenidos que publica cada usuario (tablas, tuits, fotografías, *stories*, vídeos) son también un dato clave de las redes sociales. Mediante el análisis de estas publicaciones, es posible generar una gran cantidad de conocimiento. Los múltiplos formados de los datos de este tipo también constituyen uno de los retos más importantes e interesantes que tendrá que afrontar cualquier analista de *big data*.

4) Interacciones. Las interacciones en redes a veces se pueden entender como un subtipo de publicación como, por ejemplo, cuando hay una mención o alusión incrustada en un texto. Las menciones en redes como Instagram o Twitter son el caso más evidente; otro caso serían los etiquetados a fotografías. Por otro lado, también hay interacciones que se establecen entre usuarios y contenidos: *likes*, *shares*, retuits, etc.

5) Grupos y listas. Muchas redes también disponen de elementos que agrupan usuarios en comunidades de intereses: los grupos profesionales en LinkedIn, los grupos o *fanpages* de Facebook o las listas de Twitter serían los ejemplos principales. Se trata también de un dato importante, y que requiere diferentes aproximaciones analíticas según la red social.

6) KPI. Parte de la actividad social anterior es sistematizada y organizada en indicadores clave de rendimiento (KPI) por las mismas empresas de *social media*. Muchas redes sociales disponen de *dashboards* de analítica básica que permiten explorar el rendimiento de estos indicadores. Por ejemplo:

- ***Engagement***: la suma de interacciones usuario-publicación que acumula una publicación.
- **Alcance**: la suma de usuarios que han visto una publicación.
- **Impresiones**: el número de veces que una publicación ha aparecido en una pantalla.

Cada red dispone de sus KPI y hay una gran variedad. Estos indicadores cobran especial importancia cuando se trata de analizar el rendimiento de una campaña publicitaria.

7) Metadatos. Finalmente, hay que tener en cuenta el grupo más voluminoso de datos: los datos que el usuario genera sin ser ni siquiera consciente. Aquí entran la marca y el modelo de dispositivo, la cámara que ha hecho la foto, la plataforma desde donde se ha publicado el contenido, e incluso el color del menú del usuario. Para hacernos una idea de la magnitud de esta cuestión, Twitter cuenta con más de cuatrocientas variables asociadas a cada tuit que son accesibles desde la API pública.

4.2. Adquisición

Los datos que se recuperan mediante las API de las redes sociales pertenecen al bloque de los datos semiestructurados. Se trata de datos que ya tienen cierto grado de estructuración o estandarización y que son devueltos por la API en un formato previsible, como por ejemplo los datos relativos a perfiles de usuarios, los KPI o los metadatos, pero que, por otro lado, presentan las típicas características de los datos pendientes de estructurar mediante algoritmos y otras técnicas clasificatorias, como por ejemplo las conexiones, las publicaciones, las interacciones y los grupos y listas.

Cómo hemos visto, destacan dos maneras diferentes de orientar el proceso de integración del dato, también la de redes sociales. La primera manera consiste en transformar el dato antes de almacenarlo, y la segunda consiste en todo lo contrario, primero almacenarlo y después transformarlo. El primero de los dos procesos es el denominado ETL, mientras que el segundo es el denominado ELT:

1) El ETL requiere más planificación y anticipación, y también es mucho más replicable y escalable. Siempre aplicaremos los mismos algoritmos y procesos al mismo tipo de dato y, por lo tanto, obtendremos un resultado más previsible que podremos integrar en una base de datos y analizarla con posterioridad. En el caso de los *social media*, gracias a un proceso de ETL podremos asociar un sentimiento a un fragmento de texto, podremos conocer los usuarios más importantes (teniendo en cuenta aspectos como el número de retuits o de «me gusta» en Facebook, u otros indicadores más complejos) en una conversación, o podremos generar una línea temporal para identificar los momentos clave de un debate.

2) El ELT es un proceso menos escalable y replicable pero que proporciona mucha más autonomía investigadora. En este caso, el analista podrá decidir qué tratamiento o qué técnicas aplica a los datos y responder a preguntas más específicas y específicas para cada cliente. En el caso de los *social media*, usaremos una lógica de proceso ETL cuando queramos identificar las subcomunidades dentro de una conversación o cuando queramos entrenar nuevos algoritmos.

Los datos de redes sociales se pueden adquirir de manera puntual o recurrente. Mientras que una adquisición puntual se asemeja al modelo de ELT[7], seguramente querremos optar por un proceso de ETL cuando la adquisición se tenga que efectuar de manera recurrente .

(7) Recuperaremos los datos, los guardaremos en el ordenador en el formato que proporciona la API, y después los procesaremos, transformaremos y analizaremos.

Muchas veces, sobre todo en fases de aprendizaje o de familiarización con los datos, podremos recurrir a herramientas de terceros, gratuitas o muy económicas, que nos permiten acceder a datos de redes sociales para llevar a cabo nuestros análisis. Estas herramientas típicamente recuperan y preformatean el dato, y finalmente lo guardan en nuestro ordenador (ejecutan un proceso de ETL); pero, desde el punto de vista del usuario, probablemente todavía habrá que ejecutar una serie de operaciones necesarias que aporten valor adicional a los datos.

4.3. Almacenamiento

Los datos de redes sociales se pueden almacenar en diferentes tipos de bases de datos, siempre en función del modelo de adquisición y de integración y del formato específico en que se quieran almacenar. Lo más habitual será hacerlo en una base de datos relacional y estructurada con sus atributos predefinidos, destinando algunas columnas a depositar datos en formatos no estructurados mediante lenguajes como por ejemplo JSON o XML. Por ejemplo, en la tabla siguiente (tabla 3) podemos ver como de cada ID de usuario dependen bloques de datos asimétricos, puesto que de cada usuario conocemos atributos diferentes.

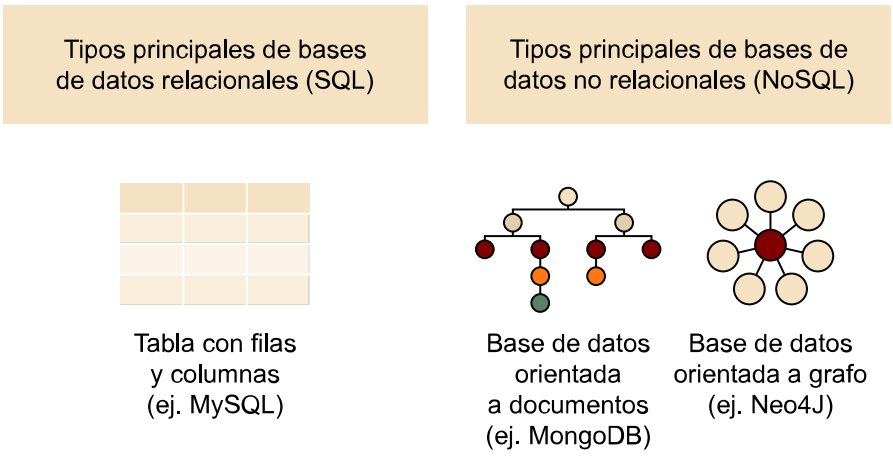
Tabla 3. Ejemplo de base de datos semiestructurada con datos incompletos

Id.	Atributos
001	{"Nombre": "Adrià", "Apellido": "Ramoneda", "Edad": "54" }
002	{"Nombre": "Anna", "Apellido": "Gutiérrez", "Ciutat": "Cerdanyola del Vallès", "Edad": "32" "Género": "mujer" }
003	{"Nombre": "Antonio", "Edad": "64", "Género": "hombre" }

Fuente: elaboración propia

En función de la magnitud del proyecto y de sus necesidades específicas, puede ser una buena idea almacenar los datos en bases de datos no relacionales. La alternativa más común a las bases de datos relacionales son las bases de datos orientadas a documentos como MongoDB, en las cuales se establecen conexiones entre documentos, generalmente escritos en lenguaje JSON, que admiten una gran diversidad de formatos y de campos. Una segunda alternativa, cada vez más popular, son las bases de datos orientadas a grafos, como Neo4j. En general, las bases de datos no relacionales (por ejemplo, figura 4) ofrecen una gran velocidad y escalabilidad, pero también es cierto que requieren una serie de conocimientos complejos y menos disponibles en el mercado que las relacionales.

Figura 4. Tipos principales de bases de datos



Fuente: elaboración propia

4.4. Análisis

A pesar de que el análisis de datos corresponde cronológicamente a la última fase de la cadena de valor de los *social media*, es muy importante que el analista la tenga en cuenta desde buen principio, de forma que se haga una implemen-

tación correcta de todas las fases anteriores. Esto es así porque, en cualquier escenario de datos masivos, y los datos provenientes de los *social media* no son una excepción, siempre hay disponible una pluralidad de estrategias de análisis de datos: la estadística inferencial, el análisis de redes y los algoritmos de aprendizaje automático son las más comunes. En todos el casos, siempre habrá que seleccionar las técnicas que aporten más valor, considerando también el esfuerzo necesario para aplicarlas.

La mayoría de operaciones analíticas, las más habituales y recurrentes tendrán que ver con la lectura de indicadores descriptivos relativamente sencillos. Las herramientas básicas de estadística descriptiva podrán servir a este propósito: medias, tablas de frecuencias, tablas de contingencia o de doble entrada, visualizaciones básicas, etc. Muchas redes sociales incorporan servicios básicos de analítica descriptiva para sus usuarios, como por ejemplo Twitter Analytics o Facebook Insights. Los datos que ofrecen gratuitamente las mismas plataformas siempre son muy limitados y con poco margen para la generación de valor; es por eso que suele ser necesario generar un ecosistema propio de datos masivos.

El análisis de redes sociales es una técnica de exploración empírica muy útil para analizar la actividad en *social media* por su naturaleza interconectada. Lo estudiaremos en profundidad en otra asignatura optativa del máster. Más adelante, en el segundo módulo, veremos algunas de las técnicas de aprendizaje automático más importantes y útiles para el análisis de datos provenientes de *social media*.

Bibliografía

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

García-Alsina, Montserrat (2017). *Big data. Gestión y explotación de grandes volúmenes de datos*. Barcelona: Editorial UOC («El Profesional de la Información», 36).

Gómez García, José Luis; Conesa i Caralt, Jordi (2015). *Introducción al big data*. Material docente. Barcelona: UOC.

Khan, M. Ali-ud-din; Uddin, Muhammad Fahim; Gupta, Navarun (2014). «Seven V's of Big Data Understanding Big Data to Extract Value». En: *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education* (págs. 1-5).

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

Marr, Bernard (2018, 21 de mayo). «How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read». *Forbes*.

Oguntimilehin, Abiodun; Ademola, Emmanuel Ojo (2014, junio). «A Review of Big Data Management, Benefits and Challenges». *Journal of Emerging Trends in Computing and Information Sciences* (vol. 5, n.º 6, págs. 433-438).

Patgiri, Ripon; Ahmed, Arif (2016). «Big Data: The V's of the Game Changer Paradigm». En: *2016 IEEE 18th International Conference on High Performance Computing and Communications* (págs. 17-24).

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