Big data
Procesando los datos en la sociedad digital

Francisco Herrera
Departamento de Ciencias de la Computación
e Inteligencia Artificial
de la Universidad de Granada

Estamos inmersos en la era de la información, donde diariamente se registran grandes cantidades de datos, del orden de petabytes. Vivimos en una sociedad digital donde el desarrollo de las tecnologías de la información y las comunicaciones y su popularidad ayudan a eliminar fronteras y crear multitud de servicios donde los datos transmitidos son un eje central de su funcionamiento.

La gran cantidad de datos disponibles en la actualidad junto con las herramientas necesarias para su procesamiento conforman lo que hoy en día conocemos como *big data* (careciendo todavía de un término concreto en nuestro idioma, utilizándose con frecuencia el término de "datos masivos").

Internet es una galaxia de información y datos, un mundo de conexiones que constantemente genera enormes cantidades de información sobre toda clase de fenómenos y actividades, incluyendo resultados de búsquedas, comentarios en blogs y redes sociales, etc. Pero los datos no sólo se encuentran ligados a internet, sino que son parte fundamental de numerosas aplicaciones, tales como los instrumentos científicos, las redes de sensores, las transacciones comerciales, los sistemas de información empresarial, etc. Igualmente los encontramos en multitud de áreas científicas como genómica, biomedicina y astronomía. Un reciente informe calcula que se generan unos 2,5 trillones de bytes de datos al día, y se espera que el número de dispositivos en red duplique a la población mundial en 2015.

El 90 % de los datos hoy disponibles se ha creado en los últimos años. Objetos cotidianos como coches, relojes o gafas comienzan a conectarse a Internet para alimentar nuevos servicios que requieren un constante intercambio de información. Los ayuntamientos siembran las calles con sensores de recogida de datos para facilitar la vida de los ciudadanos. El progreso y la innovación no se ven actualmente obstaculizados por la capacidad de recopilar datos sino por la capacidad de gestionar, analizar, sintetizar, visualizar y descubrir conocimiento en los datos recopilados. Éste es el reto de *big data;* de nada sirve tener acumulada tanta información si no se puede usar. Hay que extraer el conocimiento útil y valioso, y ello requiere herramientas capaces de procesar los datos.

El término *big data* no sólo se refiere al tamaño de los conjuntos de datos a manejar, sino que también implica la necesidad de su análisis en tiempo real cuando recibimos flujos continuos e ingentes de datos, y la posibilidad de encontrarnos con una gran variedad de estructuras, datos numéricos, textuales, enlaces, etc. Hay que tratar con datos cuyo volumen, diversidad y complejidad requieren el uso de nuevas arquitecturas, técnicas, algoritmos y análisis para gestionar y extraer el valor y conocimiento oculto en ellos... por ello se habla de las 3 V (figura 1).

Fig. 1. Las 3 V que definen *big data.*

En esta sección, personas notables, no necesariamente físicos, son invitadas a hablar (a través de una entrevista) o a escribir acerca de aspectos de su profesión o de su actividad, o en relación con sus experiencias que pueden interesar a los físicos. Animamos al lector a debatir temas que aquí se presentan enviando sus comentarios para la sección "Pulsos e impulsos".

La tecnología en torno a *big data* y el análisis inteligente de datos han dado lugar recientemente al término "Ciencia de Datos" (*Data Science*) que se refiere al área emergente de trabajo relacionada con la preparación, análisis, visualización, gestión y mantenimiento de grandes colecciones de datos para la obtención de conocimiento que genere ventajas de negocio. Debido al impacto que estas temáticas están llegando a alcanzar, ha aparecido el término profesional: el «científico de datos».

El alto potencial de estas nuevas técnicas ha sido reconocido de inmediato debido a su influencia sobre problemas de diversos campos de conocimiento. Entender la economía global, obtener una mejor planificación de servicios públicos, desarrollar investigaciones científicas o buscar nuevas oportunidades de negocio son algunas de las grandes aplicaciones relacionadas con estos repositorios de datos. Se identifica *big data* como uno de los grandes impulsores de la era digital.

Existen diversas herramientas *big data* diseñadas por los gigantes tecnológicos. La aproximación más popular es el sistema de procesamiento distribuido *MapReduce,* presentado por Google en 2003. A partir de ahí, se han desarrollado aplicaciones y tecnologías siguiendo su estela o explorando otras estrategias de procesamiento. Yahoo! desarrolló *Hadoop,* implementación de software libre de MapReduce. Posteriormente, con otras grandes firmas, creó la Fundación Apache que desarrolla más de 150 proyectos de software libre para *big data*. Recientemente, ha apadrinado el proyecto *Spark,* desarrollado inicialmente en la Universidad de California en Berkeley, que aspira a ser el nuevo referente para el diseño de algoritmos para el procesamiento de datos masivos.

¿Por qué surge *big data?*

A continuación vamos a mostrar un caso de estudio que introduce de forma natural la necesidad de más y más capacidad de cálculo, lo que condujo al desarrollo de la primera tecnología específica para *big data*.

Supongamos que tenemos que procesar un conjunto de datos del tamaño de 1 terabyte en un ordenador que es capaz de procesar 25 megabytes por segundo. El tiempo total de procesamiento sería aproximadamente de 46 días. Si disponemos de un clúster de 1.000 ordenadores y podemos paralelizar mediante el procesamiento distribuido clásico, el tiempo de cálculo se reduciría a 66 minutos, aceptable para una empresa que necesite dar respuestas en el día. ¿Qué ocurre si tenemos un conjunto de datos 100 veces mayor? Necesitaríamos 4,6 días, inaceptable para la empresa. Y la estrategia no puede consistir en aumentar los recursos de hardware, el tamaño del centro de proceso de datos con más y más ordenadores, pues se choca con restricciones de espacio, consumo eléctrico, costes... Este sencillo planteamiento de un problema difícil llevó al paradigma en análisis de datos a gran escala: MapReduce.

Google necesitaba manejar en 2003 la tabla de índices invertidos para procesar las consultas en su buscador, y se contabilizaban más de 20 petabytes de información diaria creciendo sin parar. En agosto consiguieron la primera versión estable de MapReduce, desarrollando más de 10.000 aplicaciones en los 4 siguientes años, logrando convertirse así en la tecnología básica para el procesamiento de sus datos.

El paradigma MapReduce

El nuevo paradigma de programación para el procesamiento de datos masivos sigue la estrategia "divide y vencerás". Descompone el problema planteado, de gran tamaño, en un gran conjunto de problemas pequeños (bloques o fragmentos de datos) que se procesa en una cola virtual de tareas que se va ejecutando conforme las disponibilidades de ordenadores (función Map), y las salidas asociadas a los bloques de datos se fusionan para proporcionar una solución al problema (función Reduce). Esta función requiere diseñar algoritmos que integren los modelos o salidas del procesamiento de los bloques en un modelo o solución final para el problema.

Un ejemplo es el cálculo de las tendencias de consultas, el famoso *Google Trend*. Se tiene un sistema de tareas para procesar fragmentos de datos que se ponen en cola, se van procesando según las disponibilidades de cálculo y se fusionan las salidas de cada bloque para dar el listado de consultas más populares. Actualmente Google utiliza múltiples aplicaciones desarrolladas sobre MapReduce y sus extensiones para la indexación de las páginas web a partir del popular algoritmo *PageRank,* el análisis de las bases de datos de búsquedas en Google, procesamiento de imágenes de satélite, sistemas de traducción automática, etc.

Hadoop: implementación en código abierto

Los competidores de Google comenzaron enseguida la carrera para disponer de tecnología similar. Yahoo! creó un equipo con el objetivo de diseñar software similar de código abierto en JAVA. Presentaron el resultado en 2008 con el nombre de Hadoop (el nombre del elefante de peluche del hijo de Doug Cutting, ingeniero jefe del equipo).

En julio de 2008, un algoritmo desarrollado por Yahoo! sobre Hadoop ganó la competición anual de ordenación de un terabyte de datos. Los ordenó en 209 s, cuando el record estaba en 297 s. En 2013, otro algoritmo sobre Hadoop ganó la competición de ordenación de información en un minuto, ordenando 102,5 terabytes en 4.328 s. Hadoop, que funciona sobre cualquier clúster o red y no depende del número de ordenadores disponible, mostraba así su capacidad como plataforma para el desarrollo de potentes aplicaciones de análisis de datos masivos.

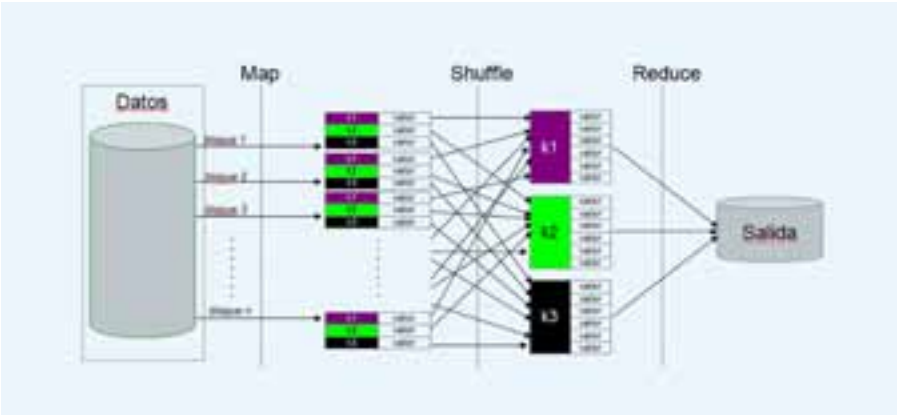

Fig. 2. Flujo de datos en MapReduce (los 3 colores corresponden a 3 trabajos diferentes de análisis de los datos que se procesan en paralelo).

¿Cómo funciona Hadoop? Se creó primero la solución más popular de la Fundación Apache, el sistema distribuido de ficheros HDFS (Hadoop Distributed File System), que permite el procesamiento de cantidades ingentes de datos en una red o nube de ordenadores conectados. Una instalación de este tipo se compone de un nodo máster (*Namenode)* y un gran número de nodos de almacenamiento (*Datanodes).* Los ficheros de datos a procesar se dividen en bloques de 64 megabytes que se distribuyen por el conjunto de ordenadores a utilizar en el análisis. Se almacenan en los Datanodes, asignándoles una etiqueta de identificación única de 64 bits en el Namenode. Cuando una aplicación pretende leer un archivo, contacta con el Namenode para determinar dónde está. Una característica importante del diseño es que los datos nunca se mueven a través del Namenode; toda la transferencia de datos se produce directamente entre clientes y los Datanodes. Las comunicaciones con el Namenode sólo implica la transferencia de metadatos. Conseguir niveles adecuados de fiabilidad, disponibilidad y rendimiento requiere redundancia, por lo que se almacena cada bloque en al menos tres servidores del clúster, siendo el número de copias un parámetro del sistema. HDFS es resistente a los fallos de un sistema distribuido. El fallo de un ordenador no afecta al sistema dado que tiene las réplicas de los bloques y dispone de un sistema automático de control de fallos.

Sobre esta instalación del sistema de ficheros HDFS trabajan las aplicaciones para el procesamiento de datos, distinguiéndose dos fases principales, Map y Reduce; hay una fase intermedia, Shuffle, de emparejamiento de ficheros. La idea subyacente parte del hecho de que muchas de las tareas de análisis tienen una estructura similar, es decir, se aplica el mismo cálculo sobre un gran número de registros (fase Map) y, a continuación, los resultados intermedios han de fusionarse de alguna manera (fase Reduce) para resolver el problema.

El programador debe especificar las funciones Map y Reduce dentro de cada trabajo. Entonces, se divide el conjunto de datos de entrada en bloques independientes que se procesan en paralelo por las tareas Map y Reduce, teniendo asignada cada una de ellas un número determinado de ordenadores. MapReduce ordena las diferentes salidas de las tareas Map en la etapa Shuffle, para ser procesadas por la tarea Reduce de acuerdo a una clave de salida que agrupa las tareas del mismo trabajo, puesto que se pueden realizar diferentes trabajos sobre el mismo bloque de datos. Los componentes principales de este modelo de programación se muestran en la figura 2.

Como ilustración, supongamos que interesa calcular el coste medio por año de una larga lista de productos a partir de los registros de coste para cada producto. Cada registro puede estar compuesto por una variedad de datos que, al menos, incluye los años y el coste asociado. La base de datos se divide en bloques de entrada. La función Map extrae, para cada bloque de registros, todos los pares <año,coste>, y los transmite como salida. La fase Shuffle agrupa los pares por su correspondiente año, creando la lista de costes por año <año,lista(coste)>. Finalmente, la fase Reduce calcula los costes medios contenidos en la lista de cada año.

¿Qué arquitectura hardware conviene?

Una opción es un clúster con un número de procesadores variable en función de las disponibilidades. Por ejemplo, el cluster de ordenadores del grupo de investigación «Soft Computing y Sistemas de Información Inteligentes» de la UGR, ATLAS, está formado por 4 servidores con 8 microprocesadores de 6 núcleos, 192 núcleos. Permite ejecutar aplicaciones *big data* con el número de bloques que se desee, haciendo una asignación previa del número de núcleos para las funciones Map y Reduce, y teniendo en cuenta que Hadoop distribuye las tareas Map y Reduce en los núcleos que van quedando disponibles. Para la instalación de Hadoop se puede acceder a *Cloudera* (www.cloudera.com), que proporciona soporte no comercial de los productos de la Fundación Apache. Se puede instalar incluso en un ordenador portátil en modo local para depuración de fallos.

Una tendencia en *big data* es el cálculo en la "nube" (*cloud computing*). La alta escalabilidad y flexibilidad que ofrecen estas plataformas permiten contar con servicios a la medida de las necesidades y pagar sólo por el uso que realmente se hace de esta tecnología. Sin necesidad de invertir en centros de procesamiento de datos, pueden alquilarse máquinas y servicios o comprarse almacenamiento siempre de modo acorde a las necesidades. Es una opción, hoy ofrecida por proveedores como Windows Azure y Amazon Elastic Cloud, que se potenciará en los próximos años. Por otra parte, la Fundación Apache proporciona aplicaciones de análisis de datos, siendo muy popular la biblioteca *Mahout* con un amplio con-

junto de algoritmos para clasificación, *clustering,* extracción de asociaciones entre variables y sistemas de recomendaciones.

Es destacable que Hadoop va más allá de una escalabilidad lineal en el procesamiento de los datos atendiendo a la arquitectura distribuida y al número de ordenadores disponibles, pues influyen diferentes aspectos asociados al tiempo de análisis de cada bloque, como el rendimiento en función de la memoria disponible o la escalabilidad de los algoritmos utilizados en las tareas Map. Como ejemplo, hemos ejecutado el algoritmo de clasificación *Random Forest* sobre un conjunto de datos con 2,4 millones de instancias (41 variables y 2 clases). El tiempo de ejecución en un ordenador I7 fue de 49.134 s, y en ATLAS, utilizando sólo 20 bloques, fue de 221 s. Claramente la escalabilidad está muy por encima de un valor lineal. El mismo problema con 4,8 millones de instancias, que no puede hacerse en un I7, requiere 236 s en ATLAS. Esto es, duplicando el tamaño del problema sólo se necesitan 15 s más de cálculo, debido a que en el primer caso no se alcanza la capacidad máxima de cálculo para cada núcleo.

Ecosistema Hadoop

Se llama así al conjunto de software, aplicaciones de código libre sobre Hadoop, actualmente más de 150 proyectos, desarrollado por la Fundación Apache para análisis, programación, almacenamiento de datos, etc. El portal www.kdnuggets.com destacaba recientemente 18 herramientas esenciales para trabajar con el entorno Hadoop, entre las que incluye la plataforma Apache Pig para ejecución de código sobre datos, el paquete Apache Ambari para el manejo de los clústers Hadoop, el conjunto de herramientas GIS Tools para manejar componentes geográficas en los datos, la base de datos HBase, y un nuevo sistema para el desarrollo de algoritmos que utiliza el procesamiento en memoria más rápido que el procesamiento en Hadoop, Apache Spark.

Frente a las bases de datos relacionales que se utilizan para el almacenamiento de datos, surge la nueva tecnología *NoSQL,* sistemas de bases de datos no relacionales para albergar y procesar grandes cantidades de datos complejos. Las bases de datos NoSQL no son compatibles con las actualizaciones y eliminaciones. Dos bases de datos para estas arquitecturas son HBase y Cassandra.

Limitaciones de MapReduce

MapReduce no es, a pesar de sus bondades, la panacea. Hay escenarios en los que este modelo de programación funcional no aporta el rendimiento esperado, y han de buscarse soluciones alternativas. Se pueden enumerar las siguientes limitaciones:

- Hay algoritmos que no pueden formularse eficientemente en términos de las funciones Map y Reduce. Por ejemplo los algoritmos iterativos, presentes en el ámbito del análisis inteligente de datos.
- MapReduce tiene problemas en el procesamiento de datos en red, es decir, con la estructura de grafo. Por ejemplo, los enlaces en la web que procesa Google para dar prioridad a las páginas con su buscador.
- El procesamiento de gran cantidad de archivos pequeños y cálculos intensivos con una pequeña cantidad de datos es otra debilidad de MapReduce.
- MapReduce es poco flexible a la hora de crear flujos de datos acíclicos: solamente permite usar un esquema "Map ▸ Shuffle ▸ Reduce". A veces sería interesante hacer un "Map ▸ Shuffle ▸ Reduce ▸ Shuffle ▸ Reduce" pero con MapReduce se tiene que transformar en un flujo "trabajo 1 (Map ▸ Shuffle ▸ Reduce); trabajo 2 (Map ▸ Shuffle ▸ Reduce)" (obligando a tener una fase Map "identidad" que carece de utilidad).

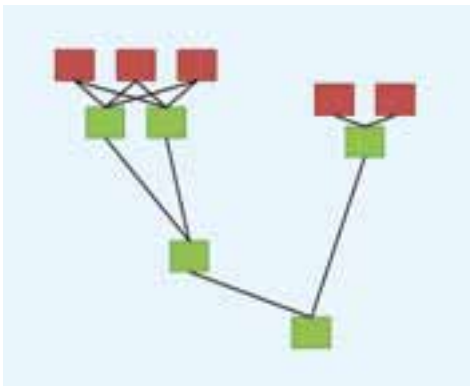

Fig. 3. Flujos acíclicos de procesamiento de datos.

Dependiendo de las características del problema, hay alternativas a MapReduce. Google desarrolló *Pregel* para el procesamiento iterativo de la información presentada como grafos y procesada con su algoritmo PageRank. En el ecosistema Hadoop, está Giraph, que tiene las mismas prestaciones.

Spark: la nueva tecnología emergente

Como hemos mencionado, Apache Spark es un nuevo sistema para el procesamiento de datos masivos, centrado en la velocidad por el procesamiento en memoria, facilidad de uso y un análisis sofisticado para el diseño de algoritmos. Aborda muchas de las limitaciones de MapReduce desde su propia concepción, y es hoy considerado la tecnología emergente del ecosistema Hadoop. Originalmente desarrollado en 2009 en el laboratorio AMPLab de la Universidad de California en Berkeley, presentado como código abierto en 2010, Spark se ha convertido en un proyecto de desarrollo de software de la Fundación Apache en febrero de 2014.

Spark, instalado sobre el sistema de archivos distribuido HDFS, no está ligado a MapReduce y permite realizar trabajos paralelizados totalmente en memoria, lo cual permite un rendimiento de hasta 100 veces superior para ciertas aplicaciones, sobre todo si se trata de procesos iterativos. Si hay datos que no caben en la memoria, Spark sigue trabajando y usa el disco duro para volcar los datos que no se necesitan en este momento. Spark resuelve la limitación de MapReduce para los flujos de trabajo basados en grafos acíclicos

reduciendo así aún más los tiempos de procesamiento y permite procesos como el mostrado en la figura 3.

Es posible así escribir una aplicación completa en Python, Java o Scala compilarla y ejecutarla en el clúster, y ofrece aplicaciones para la realización de diferentes tipos de análisis: explotar los datos con un lenguaje SQL (Spark SQL), procesamiento de flujo continuo de datos (Spark Streaming), una librería de aprendizaje automático (MLlib) y una aplicación para el procesamiento de grafos (GraphX). Spark no es todavía un producto estable para su producción, pero actualmente es una de las herramientas más prometedoras desarrolladas por la Fundación Apache; de hecho, los desarrolladores de Mahout han anunciado que su siguiente versión se implementará sobre Spark.

Big data y física

En el ámbito de la Física, encontramos muchas áreas involucradas. Por ejemplo, en relación con el lanzamiento del telescopio Solar Dynamics Observatory por la NASA en 2010, la sonda espacial está proporcionando alrededor de un petabyte de datos científicos al año (aportados por los instrumentos Atmospheric Imaging Assembly y Helioseismic and Magnetic Imager). Los retos en astronomía solar serán todavía mayores con el telescopio Daniel K. Inouye Solar Telescope, actualmente en construcción en Hawai, que se espera que genere entre 3 y 5 petabytes al año.

En 2009 *Science* publicó "Distilling Free-From Natural Laws from Experimental Data", donde se muestra cómo a partir de datos del movimiento de diferentes sistemas físicos es posible aprender las leyes de conservación utilizando algoritmos de aprendizaje de expresiones basados en Programación Genética. Dados los datos sobre oscilaciones de un péndulo doble, el algoritmo fue capaz de obtener el hamiltoniano del sistema. Para sistemas más simples, el algoritmo identificó otras leyes, tales como la función de Lagrange y la ecuación de movimiento. ¿Qué podrá llegar a comprenderse a partir de bases de datos masivos de sistemas físicos utilizando tecnologías *big data?*

El CERN es referente y pionero en el uso de *big data*. El estudio de las interacciones y partículas fundamentales que componen la materia genera un gigantesco volumen de información. El LHC provoca un choque de 2 protones a una velocidad cercana a la luz, mientras millones de sensores captan información sobre la colisión y ocurren millones de colisiones por cada segundo, de modo que la cantidad de datos generada por segundo es enorme. De hecho, el centro de procesamiento de datos del CERN ha recogido más de 100 petabytes en los últimos años, actualmente recoge más de 35 petabytes al año, que se doblarán con la actualización del LHC, y los investigadores desean guardar sus datos al menos durante 20 años. No hay duda de que van a ser necesarias nuevas tecnologías *big data*.

Concluyendo

La sociedad digital requiere almacenar, mover y procesar una enorme cantidad de datos, lo que ha llevado al desarrollo del conjunto de tecnologías *big data.* Ha habido importantes desarrollos en los últimos años, y habrá muchos más en el próximo futuro, pues el potencial es inmenso. *Big data* será imprescindible en el día a día de las empresas. International Data Corporation, principal proveedor mundial de tecnologías de inteligencia de mercado, predice que un 89 % de crecimiento de la industria de Tecnologías de la Información será en 2020 *big data*. Este entorno moverá 132.000 millones de dólares en 2015, calcula la consultora Gartner.

Recientemente se publicó en el periódico El País el artículo titulado «10 maneras de mejorar nuestra calidad de vida usando *big data*» (N. Palomino, 25 de Abril de 2014), donde se describen 10 ejemplos de cómo la sociedad puede beneficiarse de la actual revolución de los datos, con iniciativas concretas que se están desarrollando en esos ámbitos. Estos 10 problemas muestran un interesante panorama del uso de *big data:* (1) comprender los procesos demográficos y migratorios, (2) identificar hábitos y problemas sociales, (3) mejorar sistemas de alerta de desastres, (4) comprender tendencias económicas, (5) detectar riesgo de pandemias en tiempo real, (6) descubrir cambios topográficos, patrones de tráfico y emisiones, (7) entender el cambio climático, 8) mejorar servicios públicos, (9) organizar la ayuda humanitaria en desastres, y (10) fortalecer lazos comunitarios.

La demanda de profesionales es enorme. Gartner estima que en *big data* se crearán más de 4,4 millones de empleos entre 2012 y 2015 en el mundo, y España necesitará para 2015 más de 60.000 profesionales según el Subdirector General de Tecnologías.

Sin duda, *big data* jugará un papel esencial en nuestra sociedad como catalizador de la evolución hacia la sociedad digital basada en el conocimiento. Nuestra sociedad necesita estas tecnologías para transformar la información en conocimiento.

Lecturas complementarias

[1] V. Mayer-Schönberger y K. Cukier, *Big Data. La revolución de los datos masivos* (Turner, 2013).

[2] B. Schmarzo Big Data, *El poder de los datos* (Anaya, 2014).

[3] R. Schutt y C. O'Neil, *Doing Data Science* (O'Reilly, 2013).