

Jesús Marco de Lucas\*

# UNA TENDENCIA, DATA SCIENCE, Y TRES CLAVES: BIG DATA, SUPERCOMPUTACIÓN, CLOUD

*En este artículo se proporciona una visión general de la ciencia de los datos, o data science, un enfoque bajo el que se unen a la actual avalancha de datos digitales, muchos de ellos en abierto, los nuevos métodos matemáticos, técnicas e infraestructuras de computación, componentes todos ellos necesarios para abordar nuevos retos en áreas muy diversas, que están propiciando un cambio disruptivo no solo en la forma de hacer ciencia, sino también en innovación y las aplicaciones en la industria.*

**Palabras clave:** big data, estructura digital, ciencia de los datos, la nube digital, servicios de software.

**Clasificación JEL:** H54, O31, O33.

## 1. Introducción

Si quisiéramos resumir los avances en estos últimos años en ciencia y en tecnología desde la perspectiva de las infraestructuras de computación, como claves de un nuevo cambio digital, podríamos decir simplemente: *data science*<sup>1</sup>.

¿Pero qué es *data science*? Para muchos científicos, una etiqueta para vender mejor algo que llevan haciendo muchos años: entender y explotar los datos (en ciencia) que son la base de sus experimentos, especialmente cuando son muy complejos y/o extensos. Esta visión está ligada al desarrollo de los métodos estadísticos, y sus extensiones, como la minería de

datos, o las técnicas de aprendizaje automático (*machine learning*).

Si preguntamos a muchos profesionales TIC, pueden decir que *data science* es una forma más de renombrar la explotación de grandes bases de datos, que tradicionalmente se han operado en el ámbito comercial como *business analytics*, y especialmente incorporar al análisis datos no estructurados, como los ligados a la actividad de los usuarios en las redes, y hacerlo con nuevas herramientas (por ejemplo, bases de datos NoSQL<sup>2</sup>), para descubrir nuevas oportunidades de negocio.

Si preguntamos a los responsables de los centros de datos, nos dirán que es todo lo que se requiere

---

\* Profesor de investigación del Consejo Superior de Investigaciones Científicas (CSIC) en el Instituto de Física de Cantabria (IFCA). Coordinador del proyecto *DEEP Hybrid DataCloud*, en el European Open Science Cloud, participa en el WLCG de procesamiento de datos del CERN y en la Red Española de Supercomputación.

---

<sup>1</sup> A lo largo del texto se utilizan palabras en inglés, dado que este es el idioma empleado por la comunidad tanto científica como técnica.

<sup>2</sup> El término NoSQL identifica un tipo de sistemas de gestión de bases de datos diferentes de los sistemas clásicos relacionales que emplean el lenguaje SQL (*Structured Query Language*, un lenguaje de consulta en bases de datos estandarizado, muy extendido en el ámbito profesional). En <https://nosql-database.org>

para gestionar millones de *gigas* (es decir, *petabytes*) de información, transmitirlos, procesarlos en un tiempo razonable, o almacenarlos durante muchos años de forma que se puedan reutilizar y analizar para dar soporte a la toma de decisiones estratégicas.

Todos ellos tienen razón, pero el desafío realmente apasionante, que va un paso más allá, y que implica un nuevo cambio digital que está ya impactando en nuestras vidas, y va a hacerlo especialmente en nuestra economía, es la capacidad de contar con sistemas inteligentes que se alimentan de estos datos utilizando las nuevas técnicas, para realizar tareas «humanas» de modo mucho más eficiente.

Se trata de la integración entre un área de conocimiento como es la inteligencia artificial, con una base técnica muy amplia de soluciones, que globalmente han ido convergiendo en lo que se denomina problemas de *big data*, gracias a una e-infraestructura muy potente, que se apoya en los supercomputadores y las plataformas de computación *cloud*.

Por ello, en lo que sigue intentaré describir en primer lugar los problemas de *big data*, explicar cómo han evolucionado las plataformas de computación, en particular los supercomputadores y los sistemas *cloud*, para proporcionar soluciones en esta área, y terminar planteando la relación con los nuevos desafíos en *data science*, y cómo están relacionados con la naturaleza disruptiva del cambio digital.

En primer lugar se analizan los retos que han surgido asociados a problemas de *big data* y la destacada evolución de las técnicas matemático-estadísticas para abordarlos, cada vez más complejas, pero también más potentes, que perfilan un nuevo currículum profesional.

Se presentan a continuación dos infraestructuras de computación clave en *data science*. Se describe cómo los supercomputadores continúan creciendo en tamaño y capacidad, y en particular cómo ha aumentado enormemente el interés desde la industria en los últimos años, con China encabezando esta evolución. Se analiza el modelo *cloud* de computación, y especialmente su flexibilidad como entorno de *data science*, y la

oferta cada vez mayor de servicios HPC Cloud de las grandes multinacionales, atrayendo hacia este entorno los servicios de *big data*.

Por último, se destaca el carácter disruptivo de las nuevas técnicas, como es el caso del *Deep Learning*, y la oportunidad que ofrece la nueva iniciativa European Open Science Cloud (EOSC), y se proponen varias acciones para aprovecharla en el contexto nacional: formación, colaboración empresa-investigación pública, y especialización de cara a desarrollar y explotar estas nuevas técnicas a todos los niveles.

## 2. *Big data*: problemas y oportunidades

La denominación *big data* se empieza a aplicar hace más de 20 años para diferenciar colecciones de datos que son tan grandes o tan complejos que no se pueden procesar y analizar de la forma en que tradicionalmente se ha estado haciendo, con aplicaciones informáticas clásicas.

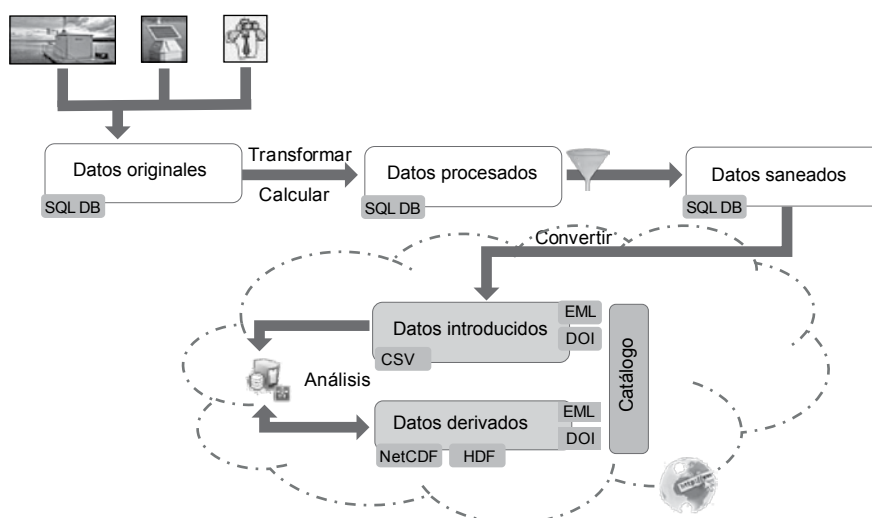
La consultora Gartner destacó, muy acertadamente, en un análisis presentado en 2011<sup>3</sup>, que si bien el volumen de la información digital, a nivel mundial, crece a un ritmo anual superior al 60 por 100, y gestionar este volumen de datos es en sí un desafío, es necesario además tener en cuenta otros dos parámetros: variedad y velocidad. En este análisis se precisaban algunos aspectos clave en relación a los tres parámetros indicados: el problema no es solo la gestión del incremento en tamaño global de los datos clásicos, como los asociados a los registros de las bases de datos, sino de nuevos tipos de datos, de diferentes fuentes, que siguen diferentes modelos de datos, incluyendo información multimedia o de actividad en Internet, y que se genera y debe ser procesada a gran velocidad.

Este análisis en base a estas 3V originales, volumen, variedad y velocidad, ha logrado un cierto consenso en

<sup>3</sup> GARTNER INC. (2011). *Gartner Says Solving «Big Data» Challenge Involves More Than Just Managing Volumes of Data*. Junio. En <http://www.gartner.com/it/page.jsp?id=1731916>.

FIGURA 1

### CICLO DE VIDA DE LOS DATOS EN EL CASO DE USO DE *ALGAE BLOOM* DEL PROYECTO INDIGO-DATACLOUD



FUENTE: INDIGO-DataCloud project, 2016.

la comunidad, aunque también se ha propuesto añadir otras «V», como veracidad o valor, que suponen evidentemente un nuevo reto a considerar por las tecnologías necesarias para resolver los problemas de *big data*.

Desde una perspectiva actual, muy en línea con un enfoque como *data science*, abordar un problema de *big data* supone considerar desde el principio el «ciclo de vida» de dichos datos: desde la creación de los mismos, su captura y transferencia, almacenamiento, filtrado o preprocesado, en resumen, su «ingestión», hasta su análisis, visualización, publicación, preservación, y la forma de compartir dichos datos (Figura 1).

En paralelo a este ciclo de vida de los datos, está su versión dual: el *software* asociado a estas acciones sucesivas de transformación de los datos, que permite su evolución desde la información hasta el conocimiento.

Ambas capas, datos y *software*, se deben implementar sobre una infraestructura de sistemas de información, que para cubrir todo el ciclo de vida va a necesitar integrar sistemas muy diversos, y además garantizar cuestiones clave transversales, como la seguridad y privacidad.

Para entender mejor la base de estos desafíos en *big data* se comentan a continuación brevemente algunos ejemplos y los factores que han influido para ello.

### La avalancha de datos

El aumento imparable del volumen de datos que registramos digitalmente y en muy diversas áreas ha sido posible por la mejora en el diseño y fabricación de los sistemas electrónicos asociados, integrando sensores, procesadores y red, aumentando sus

prestaciones y reduciendo a la vez su consumo, tamaño y coste.

El ejemplo más cercano a todos es el de los teléfonos móviles, que pueden integrar sensores que eran relativamente sofisticados hace unos años, como cámaras fotográficas de alta resolución o receptores GPS, y también otros más sencillos, de aceleración, o el propio micrófono, y además son capaces de transmitir esta información digitalizada a la Red de diversas formas (4G, wifi). Esta evolución de la microelectrónica, y más recientemente de la nanotecnología, permite contar hoy en día con todo tipo de sensores para medidas físicas y químicas, y cada vez más también biológicas, con buena resolución y fiabilidad, y a un coste cada vez menor. Sensores ideados en un principio para una función específica, como los sensores de gases que incorporan los motores de los coches para mejorar su rendimiento, o los sensores de consumo eléctrico, pueden integrar su información fácilmente y casi en tiempo real en la Red.

En la idea inicial de un *Internet of Things* todos estos sensores ubicuos se autoidentificarían a la Red y proporcionarían de forma autónoma un gran volumen de datos, útil para entender diferentes modelos físicos, sociales o económicos, para sistemas complejos de modelar. En iniciativas como las *smart cities* se integran este tipo de sensores y otros similares como los sensores de paso, de presencia, o estaciones medioambientales, con otros más complejos como redes de cámaras, con el propósito de optimizar un modelo de uso de recursos urbanos.

El primer reto viene dado por la integración de esta información: la agencia norteamericana de investigación de mercados ABI Research estima que en 2020 podría haber 30.000 millones de dispositivos conectados. La utilidad que tiene asignar una identificación única de todos los objetos digitales es poder hacer una planificación de recursos en la industria de la manera más precisa posible, controlando las materias primas, seleccionando los «productos» a demanda de los clientes y gestionando los residuos de la mejor manera.

Por otra parte, y especialmente en el ámbito de la investigación, continúa el desarrollo de detectores y sensores cada vez más complejos para lograr medidas de gran precisión. Detectores como los empleados por el experimento CMS del Large Hadron Collider (LHC) del CERN<sup>4</sup> integran más de 75.000.000 de canales de lectura, y registran más de 1.000 colisiones por segundo, lo que resulta en un volumen de datos muy elevado (los detectores de LHC han almacenado cerca de 200 *petabytes* en los años pasados, y esta cifra se incrementará un orden de magnitud en la nueva fase de «alta luminosidad»).

Una evolución similar ocurre en Astronomía, con los enormes flujos de datos de los nuevos telescopios, y también en Observación de la Tierra, con el despliegue creciente de satélites, como la reciente generación *Sentinel*, y de aviones de observación dotados de nuevos sistemas de teledetección de todo tipo cada vez con mayor resolución y cobertura espacial y temporal.

Pero quizás la mayor avalancha de datos va a venir unida a nosotros mismos. Dejando a un lado la «traza digital» de nuestra actividad en Internet, en nuestro móvil y en nuestras tarjetas de pago, que pueden considerarse sistemas «heterodoxos y no suficientemente controlados» de adquisición de datos, otra clave puede ser el desarrollo de la medicina personalizada, gracias al desarrollo de los nuevos chips genómicos, el uso cada vez más extendido y sistemático de equipos de adquisición y análisis de imagen médica, y la incorporación a la vida diaria de los sistemas de teleasistencia. Los sistemas de adquisición para este caso son muy diferentes. Los equipos de imagen médica, normalmente manejados en hospitales y por personal especializado, son cada vez más precisos y complejos, aportando en un solo examen una gran cantidad de información (por ejemplo, en un TAC). Los sistemas de secuenciación genómica, hasta ahora relativamente complejos, están en plena

<sup>4</sup> CERN (2011). En cuanto a la transferencia de datos en tiempo real, Google Earth tiene una imagen muy representativa de la actividad del LHC del CERN: *running jobs*, 246.190; *Active CPU cores*, 542.066 y *transfer rate* 26.01 GiB/sec.

evolución: empiezan a estar disponibles chips de bajo coste conectables directamente a un sistema sencillo de adquisición. De la misma forma, los sistemas de teleasistencia pueden incorporar no solo cámaras y micrófonos de conversación del paciente con el personal sanitario, sino también conexión para sensores básicos (tensión, temperatura, pulsaciones, ritmo cardíaco) o no tan básicos (como cámaras para detección de melanos). Aunque se trata de medidas sencillas, la monitorización de cientos de miles o millones de pacientes, y el interés de analizar correlaciones y contar con sistemas de alarma ante epidemias, hace de este campo uno de los mayores retos en el campo de *big data*. En conjunto el diseño, operación, explotación, protección y preservación de datos personales es un área muy compleja por la importancia de la privacidad de dichos datos.

Por último, destacar la importancia creciente, en un contexto no solo comercial, de las redes sociales, reflejada por ejemplo en los cientos de millones de mensajes anuales en Twitter o los cientos de *terabytes* diarios procesados por Facebook: son canales que proporcionan igualmente información sobre nuestra actividad y entorno.

Desde el punto de vista nacional, el informe *Análisis de la estrategia Big Data en España*<sup>5</sup> recoge además iniciativas clave a nivel europeo como la Big Data Value Association (BDVA)<sup>6</sup>.

### Técnicas estadísticas y matemáticas para *big data*

Sin entrar en detalle, dada la amplitud del tema, puede decirse que muchas de las técnicas que se aplican en los problemas de *big data* se corresponden con una evolución, muy modulada tecnológicamente, de métodos estadísticos más o menos clásicos, integrándose más recientemente con métodos fuertemente computacionales cuyo ejemplo más destacado son las redes neuronales. Como ejemplo, se incluye en el Cuadro 1 el listado de

técnicas y métodos básicos que se abordan en un máster de *Data Science*, y cuyos algoritmos correspondientes se encuentran implementados en diferentes lenguajes de programación, y se utilizan de forma más o menos transparente en las diferentes aplicaciones disponibles, tanto de investigación como comerciales.

Un punto muy importante es que la implementación y el rendimiento de estas aplicaciones, como veremos, depende fuertemente del tipo de infraestructura de computación disponible.

### 3. E-infraestructuras para *big data*

En términos generales, ¿es posible implementar y utilizar las técnicas citadas para abordar estos retos de *big data*, desde el punto de vista de los recursos de computación necesarios?

La respuesta, también en términos generales, es sí. El mismo desarrollo tecnológico que ha posibilitado la avalancha de datos, también ha llevado a un crecimiento exponencial de los recursos informáticos, a todos los niveles: computación, almacenamiento, red.

Por poner un ejemplo concreto, hoy en día un sistema de almacenamiento con capacidad de 1 *petabyte*<sup>7</sup> cabe en un armario (*rack*) estándar, cuesta menos de 100.000 euros, su gestión no es especialmente complicada, y su rendimiento permite acceder a los datos almacenados a gran velocidad. La razón hay que buscarla en la evolución de la capacidad y precio de los discos magnéticos, la incorporación de los discos de estado sólido de acceso indexado mucho más rápido, la mejora de la interconexión interna de estos discos, y el uso de nuevas soluciones de acceso a los datos en paralelo, tanto a ficheros como a objetos digitales.

Por otra parte, el análisis o procesamiento de estos grandes volúmenes de datos se realiza normalmente en *clusters* de servidores de alto rendimiento, interconectados entre sí y a los sistemas de almacenamiento

<sup>5</sup> MIRÓN, F. *et al.* (2017).

<sup>6</sup> En <http://www.bdva.eu>

<sup>7</sup> 1 *petabyte* (1 Pb) es aproximadamente 1.000.000 de *gigabytes* (Gb), o 100.000.000 de fotos de alta resolución.

## CUADRO 1

TÉCNICAS ESTADÍSTICAS Y MATEMÁTICAS PARA *BIG DATA*

| Técnicas Estadísticas                                     | Minería de datos                                                              |
|-----------------------------------------------------------|-------------------------------------------------------------------------------|
| Estadística descriptiva.                                  | Problemas de asociación, segmentación, clasificación, y predicción.           |
| Muestreo y Monte Carlo.                                   | Aprendizaje no supervisado y supervisado.                                     |
| Inferencia estadística.                                   | Sobreajuste, validación cruzada ( <i>k-fold</i> ).                            |
| Contrastes paramétricos y no paramétricos.                | Técnicas de vecinos cercanos, distancias, núcleos y funciones de base radial. |
| Técnicas de remuestreo ( <i>bootstrap</i> ).              | Segmentación jerárquica, <i>k-medias</i> , SOM.                               |
| Modelos de regresión.                                     | Árboles de clasificación y regresión.                                         |
| Estimación de máxima verosimilitud.                       | Modelos lineales y aditivos generalizados.                                    |
| Regularización. Regresión contraída.                      | Aprendizaje por conjuntos: <i>boosting</i> y <i>bagging</i> .                 |
| Redes neuronales                                          | Aprendizaje estadístico                                                       |
| Redes de topología multicapa y recurrente.                | Márgenes y vectores soporte. Máquinas de vector soporte (SVM).                |
| Algoritmos iterativos de aprendizaje ( <i>backprop</i> ). | Métodos basados en núcleos.                                                   |
| Reservorios y técnicas de proyección aleatoria.           | Variables latentes y método EM.                                               |
| <i>Extreme Learning Machines</i> .                        | Modelos de Markov ocultos (HMM).                                              |
| <i>Deep learning</i> . <i>Autoencoders</i> y convolución. | Aprendizaje bayesiano. Redes probabilísticas.                                 |
| <b>FUENTE:</b> Elaboración propia.                        |                                                                               |

descritos mediante redes de alta velocidad y baja latencia. Cada servidor, con varios procesadores y múltiples núcleos por procesador, puede superar fácilmente el *teraflop*<sup>8</sup> (un billón de operaciones por segundo). Interconectados entre sí y con los sistemas de almacenamiento, se denominan genéricamente sistemas HPC (High Performance Computing), y constituyen actualmente la base de muchos entornos de *big data*. A partir de este sistema básico hay dos formas de «escalar» para abordar los retos más complejos.

### Supercomputadores y *big data*

Muchos de los supercomputadores actuales, a diferencia de la mayoría de los existentes hasta hace 20

años, se construyen empleando los mismos procesadores que se utilizan en los sistemas HPC: la diferencia es sobre todo de escala.

Por ejemplo, el supercomputador Cray XC40, denominado Trinity (2017), instalado en Los Álamos (EE UU), utiliza procesadores Intel Xeon de 16 *cores*, eso sí, masivamente: más de 300.000 *cores*.

Evidentemente, este escalado conlleva la necesidad de contar con una red de conexión interna especial, y con una instalación física adecuada: el supercomputador requiere una potencia eléctrica de más de 4 MW, y el correspondiente sistema de refrigeración. Pero permite alcanzar una potencia de más de 10.000 *teraflops*, que permite resolver problemas extraordinariamente complejos en un tiempo reducido.

Antes de continuar hablando de supercomputadores, conviene discutir su evolución, y su papel en diferentes áreas, y especialmente en relación a los problemas de *big data* y más en general en *data science*.

<sup>8</sup> Recientemente, en junio de 2017, Intel ha presentado el primer procesador con 18 núcleos que supera el billón de operaciones por segundo o *teraflop*: *Core i9 Extreme Edition*. Su coste está por debajo de los 2.000 dólares.

Tradicionalmente la aplicación de los supercomputadores se ha centrado sobre todo en problemas de simulación<sup>9</sup>. Hoy en día la simulación se aplica a todas las áreas de la ciencia y la tecnología, y especialmente en el campo industrial. Se simulan sistemas a todas las escalas y en todos los campos: se simulan colisiones de partículas elementales, propiedades de estructuras atómicas y moleculares, nuevos materiales, reacciones químicas, fluidos, la aerodinámica de un coche o de un avión, la hidrodinámica de un canal o de una bahía completa, la evolución de la atmósfera, o incluso la formación de galaxias. Por poner dos caras de la misma moneda, se simulan explosiones nucleares, y reactores de fusión.

En muchos de estos campos la simulación requiere una gran capacidad de cálculo, pero no conlleva en general un procesamiento de datos especialmente complejo. Las técnicas más importantes son las que permiten la «paralelización»<sup>10</sup> de los cálculos: cómo repartir los mismos entre miles o cientos de miles de procesos, de modo que el tiempo necesario para completar un cálculo complejo se reduzca proporcionalmente. Obtener el máximo rendimiento de un supercomputador es un problema extraordinariamente delicado, y con soluciones en muchos casos muy específicas<sup>11</sup>.

Sin embargo, en los últimos años la importancia del procesamiento de datos en los problemas relevantes para las comunidades que usan la supercomputación crece de forma continua. En primer lugar, está el posprocesado de los datos producidos por las propias simulaciones. Los investigadores normalmente aplican esta fase para producir un nuevo conjunto de datos que a su vez exploran en detalle en sistemas más pequeños,

pero de uso más flexible, que permiten por ejemplo un análisis interactivo. Debe tenerse en cuenta que el modo clásico de uso de un supercomputador se basa en la asignación de tiempo de computación a cada grupo durante un cierto período, y el envío de los trabajos de computación para su ejecución en el supercomputador de acuerdo a las prioridades asignadas.

Pero a medida que las simulaciones son más realistas, y permiten la comparación con los datos reales, los investigadores quieren aplicar las mismas técnicas a ambos conjuntos de datos, reales y simulados, y en los últimos años esto significa aplicar técnicas de *big data*, que también requieren del uso masivo de procesadores pues se benefician del paralelismo de algoritmos y datos. Por tanto, los investigadores empiezan a plantearse realizar no solo un posprocesado sino todo, o al menos una parte relevante, del análisis de los datos producidos en la simulación en el propio supercomputador. E inversamente, se plantean desplazar al supercomputador también el análisis de los datos reales, ya que los algoritmos a utilizar son los mismos. En resumen, el supercomputador pasa de ser una infraestructura muy especializada, orientada sobre todo a simulaciones que escalan masivamente, a convertirse en una máquina de propósito más general en la que el sistema de acceso a los datos y el procesamiento de los mismos puede ser crítico.

Un ejemplo muy claro lo tenemos en el uso de supercomputadores para el procesamiento masivo de secuencias genéticas. Los algoritmos muy complejos están en muchos casos paralelizados y permiten utilizar de forma eficiente miles de *cores* a la vez: el supercomputador permite reducir el tiempo de cálculo de meses a días.

El uso de un supercomputador como servicio<sup>12</sup> es actualmente una de las opciones más interesantes a considerar en el área: problemas importantes de flexibilidad y seguridad, como por ejemplo la necesidad de instalar

<sup>9</sup> En los problemas de simulación se utiliza un modelo matemático para describir los procesos del sistema considerado, y se aplican métodos numéricos mediante programas (de simulación) ejecutados por computadores.

<sup>10</sup> En la «paralelización», o computación en paralelo, se ejecutan simultáneamente múltiples instrucciones de un programa usando a la vez (en paralelo) varios procesadores o *cores*.

<sup>11</sup> Y ha anticipado también muchas de las soluciones técnicas para los problemas de *big data*.

<sup>12</sup> Se entiende aquí por «servicio» el uso de un *software* de interfaz para acceder remotamente desde otro sistema a las aplicaciones mediante un protocolo de comunicación, a través de la Red.

y soportar las aplicaciones de los usuarios, se van resolviendo con soluciones técnicas recientes como el uso de «contenedores»<sup>13</sup>, que sin reducir el rendimiento permiten que el usuario utilice sus aplicaciones optimizadas transparentemente en diferentes plataformas.

Evidentemente hay un compromiso entre el coste de un supercomputador y la explotación del mismo para tareas que no requieren toda su potencia.

Por ello conviene considerar a continuación la alternativa actualmente más importante de servicios de computación de alto rendimiento: HPC Cloud Computing.

### **Cloud computing para big data**

Siguiendo la definición del NIST de *cloud computing*<sup>14</sup>, podemos decir que se trata de un modelo para permitir el acceso bajo demanda y a través de la Red (Internet) a unos servicios que utilizan un conjunto de recursos computacionales, como por ejemplo servidores, sistemas de almacenamiento, y conexiones de red, de modo flexible.

*Cloud computing* se basa pues en un modelo de servicios, es decir un «usuario accede a un servicio que ha contratado con un proveedor». Los servicios son a tres niveles diferentes (Figura 2):

— **Infraestructura como Servicio (IaaS)**: se facilita directamente al usuario el equivalente a un computador virtual, con capacidad de computación, almacenamiento y conexión de red, y el usuario puede instalar desde el sistema operativo hasta las aplicaciones, pero no gestiona ni controla la infraestructura subyacente, que se encuentra típicamente en un gran centro de datos remoto.

— **Plataforma como Servicio (PaaS)**: permite desplegar directamente aplicaciones del usuario, incluyendo en particular las que puede crear mediante una serie de herramientas que se encuentran ya instaladas

en el sistema y permiten aprovechar el carácter flexible del entorno *cloud*. Por ejemplo, puede poner en marcha y gestionar un *cluster* virtual con 100 procesadores, pero no se tiene que ocupar de instalar y configurar el sistema operativo.

— **Software como Servicio (SaaS)**: el usuario accede directamente a aplicaciones que suministra el proveedor, y que se ejecutan en una infraestructura Cloud, del propio proveedor del servicio o de terceros, por lo que disfruta de las ventajas propias (escalabilidad, flexibilidad de uso, etc.). Un ejemplo típico es por ejemplo acceder a una base de datos escalable.

En general los servicios Cloud son accesibles a través de una interfaz tipo web, y desde cualquier sitio y de modo múltiple. Usualmente el usuario paga por consumo, es decir solamente cuando el servicio se utiliza. Aunque muchas aplicaciones exigen que algunos servicios estén disponibles de forma continua, si están bien configuradas solo consumirán los recursos necesarios. De esta forma el proveedor de infraestructura Cloud puede compartirla entre los diferentes usuarios y reducir el coste a repercutir. El pago de los servicios normalmente es por uso, y el precio puede ser fijo o en subasta de acuerdo a la prioridad requerida. En el caso de los servicios IaaS este precio depende críticamente de la dimensión de los recursos solicitados: por ejemplo, una instancia de computación con 2 núcleos y 4 Gb de memoria cuesta 0,03 euros/h mientras que una instancia con 64 núcleos y 128 Gb de memoria puede costar 100 veces más, 3 euros/h.

Evidentemente cuando pensamos en recursos orientados a problemas de *big data*, puede parecer-nos que son este último tipo de instancias las que necesitamos, pero en el ciclo de vida de los datos hay muchas etapas en las que las necesidades son mucho menores, y por ello el uso de servicios Cloud puede ser especialmente adecuado.

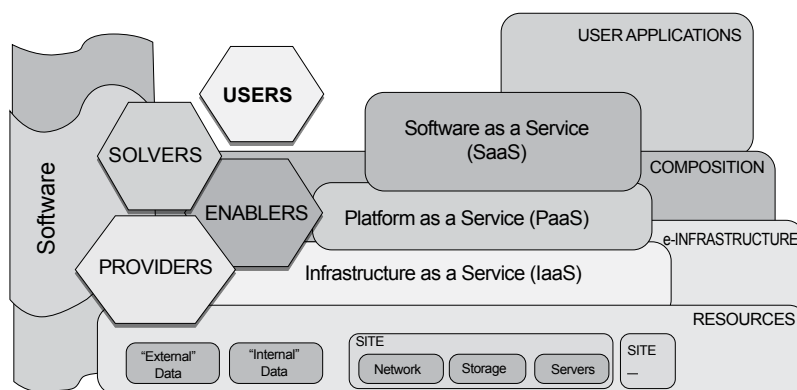
Los servicios Cloud se implementan de diferentes formas según la propiedad de los recursos y el tipo de acceso a los mismos. Muchas empresas y centros de investigación cuentan con un sistema de servicios

<sup>13</sup> Como por ejemplo la solución *-u docker*, desarrollada en el proyecto europeo INDIGO-DataCloud, que permite ejecutar aplicaciones en paralelo sobre InfiniBand en contenedores sin necesidad de privilegios de administración.

<sup>14</sup> NIST (2011).

FIGURA 2

## ESQUEMA DE DESARROLLO DE APLICACIONES SOBRE LAS CAPAS DE SERVICIOS CLOUD



NOTA: *Adapting LW Layered Architecture.*

FUENTE: NIST.

Cloud privados, que no ofertan a otros usuarios. Los sistemas denominados «públicos» ofrecen el acceso a sus servicios a terceros, al público en general. Ejemplos típicos son Amazon Web Services, Google Cloud Compute, IBM Bluemix o Microsoft Azure.

En un servicio Cloud híbrido se combinan los servicios de diferentes proveedores Cloud; típicamente se permite que servicios en un Cloud privado puedan escalar accediendo a recursos de un Cloud público cuando sea necesario.

En este caso, y también en general para evitar la dependencia de una solución ligada a un solo proveedor, es especialmente importante la existencia de estándares abiertos, que permitan la interoperabilidad de los servicios, además de resolver los problemas de autenticación y autorización. Un ejemplo es el uso de estándares como OCCI<sup>15</sup> o CDMI<sup>16</sup> para el acceso a los servicios IaaS de computación y de almacenamiento, o de

TOSCA<sup>17</sup> como lenguaje de orquestación de recursos en servicios PaaS, aunque la competencia con las soluciones propietarias es muy difícil por el peso *de facto* de iniciativas en el área como *Open Stack*<sup>18</sup>.

Como ya hemos dicho, la filosofía de servicios se adapta muy bien a soportar el ciclo de vida de los datos. Es cierto que hay un problema de coste evidente en cuanto el volumen de datos es muy elevado, y cuando se pretende además preservar los datos durante mucho tiempo, pero existen soluciones adaptadas a casi todos los casos. Una cuestión diferente es si realmente se puede disponer de recursos HPC como servicios Cloud, y la respuesta empieza a ser también afirmativa, aunque el coste sigue siendo relativamente elevado.

Los grandes centros de infraestructura Cloud son cada vez mayores, buscando optimizar sus costes de operación, y también el coste de los equipos instalados.

<sup>17</sup> OASIS *Topology and Orchestration Specification for Cloud Applications* (TOSCA). En [https://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=tosca](https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca)

<sup>18</sup> *Open source software for creating private and public clouds*. En <https://www.openstack.org/>

<sup>15</sup> *Open Cloud Computing Interface* (OCCI). En <http://occi-wg.org/>

<sup>16</sup> *Cloud Data Management Interface* (CDMI). En <https://www.snia.org/cdmi>

Aunque esto hace difícil en principio que compitan con el *hardware* de los supercomputadores, como ya hemos comentado, en muchos casos los procesadores son similares, y la mejora de los componentes disponibles masivamente de forma comercial pero de alto rendimiento, como discos o tarjetas de red, y también de coprocesadores especializados, como veremos más adelante, junto al hecho de ser compradores masivos, hace posible que los proveedores de servicios Cloud empiecen a competir también en esta área.

Además, existe un problema, o una oportunidad, que es la necesidad de un servicio de desarrollo o adaptación de aplicaciones HPC. El entorno *cloud* típico está orientado a la gestión de soluciones informáticas en muchos casos de tipo comercial, y la implementación de aplicaciones que deben escalar a cientos o miles de procesos no es en general sencilla. El soporte y sobre todo el conocimiento que se ofrece desde los centros de supercomputación, es un punto clave a tener en cuenta en este sentido.

### Aplicaciones de la supercomputación y de los servicios Cloud HPC

Tanto los supercomputadores como los servicios Cloud HPC cuentan con múltiples aplicaciones tanto en investigación como en la industria y cada vez más en el ámbito comercial.

Entre los dominios que siguen reservados a los supercomputadores destacan especialmente las simulaciones que requieren un uso masivo de procesadores y de memoria. Si recorremos la lista de aplicaciones en que se utilizan los supercomputadores más potentes del mundo, encontramos aplicaciones prácticas, como la predicción y simulación de tsunamis y terremotos, o de las condiciones meteorológicas y climáticas, junto a otras de interés puramente científico, como la simulación de la interacción fuerte de los *quarks*, o la evolución de la estructura del universo a diferentes escalas.

¿Y las empresas, están interesadas en el uso de supercomputadores?

La mayoría de los supercomputadores de centros de investigación públicos y de universidades mantienen contratos para el desarrollo de diversos proyectos con empresas. Pero además, entre los 500 supercomputadores de la lista publicada en 2017<sup>19</sup> exactamente la mitad figuran como de uso industrial, y la mitad de ellos están en China, una cuarta parte en EE UU, y el resto repartido entre Europa (30), Corea del Sur (5), Arabia Saudí<sup>20</sup> (4) y Japón (2), Australia, Nueva Zelanda, Canadá y Brasil. El supercomputador de uso industrial más potente que aparece en la lista es Pangea, de la empresa TOTAL, instalado en Pau en Francia, y cuya misión fundamental es el estudio del subsuelo. No es casualidad que el siguiente sea el de una empresa petrolera americana, y uno de los siguientes tenga un uso similar en Italia. Igualmente aparecen múltiples sistemas de proveedores de servicios, muchos de ellos ofreciendo por tanto computación como servicio, de empresas de *software*, de telecomunicación, varios de empresas de automóviles, y uno de Airbus.

Sí que es una sorpresa encontrar en los puestos 31 y 32 del *ranking* mundial dos supercomputadores con procesadores especializados, uno de Facebook y otro de la empresa que fabrica dichos procesadores, NVIDIA, pero de ello hablamos un poco más adelante.

También es una «anécdota» interesante la presencia en la lista de un supercomputador de Amazon que entró en la misma en 2013 en el puesto 64, para demostrar simplemente que sus servicios Cloud HPC podían estar al nivel de los mejores supercomputadores del mundo.

Las pymes también pueden beneficiarse del acceso a supercomputadores: la iniciativa europea PRACE<sup>21</sup> de supercomputación ha promovido el desarrollo y uso de aplicaciones para pymes en temas tan diversos como el diseño de turbinas eólicas, la optimización de soldaduras, o la simulación de veleros de competición.

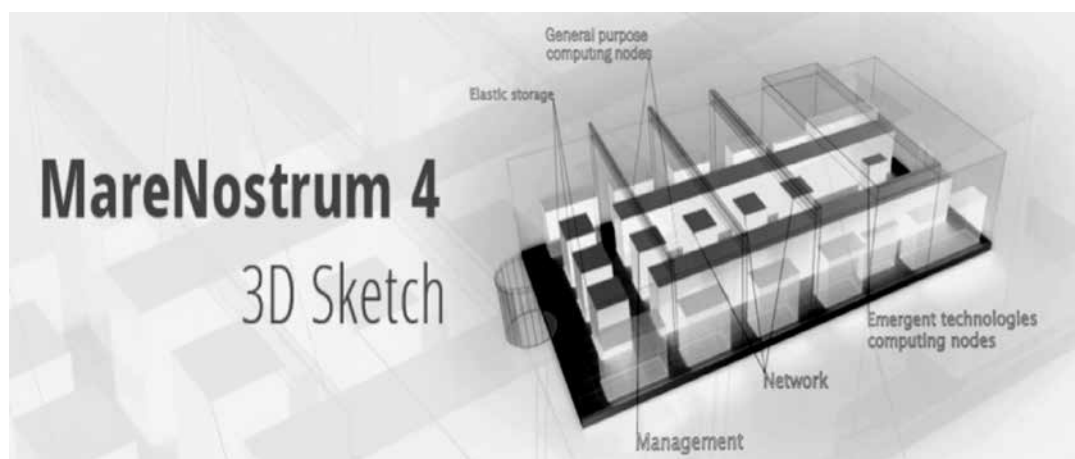
<sup>19</sup> TOP500 LIST SITE (2017). Consultado en junio en <https://www.top500.org/list/2017/06/>

<sup>20</sup> Todos ellos de la misma compañía petrolera.

<sup>21</sup> PRACE, ver <http://www.prace-ri.eu/>

FIGURA 3

ESQUEMA 3D DEL SUPERCOMPUTADOR MARENOSTRUM-4, INSTALADO EN EL BSC



FUENTE: BSC (2017).

### Situación a nivel mundial, en Europa y en España

La mayoría de las empresas y centros de investigación abordan los problemas de *big data* con ayuda de recursos propios, de tamaño medio, usualmente *clusters* HPC, que tienen a su disposición alojados en centros de datos propios o externos (en modalidad de *hosting*).

A medida que los problemas de *big data* se tornan más complejos, tanto empresas como centros de investigación y universidades buscan recursos externos.

Como ya se ha comentado, la lista pública «oficial» que incluye los 500 supercomputadores más potentes, que se actualiza dos veces al año, se denomina Top500, y puede consultarse en <https://www.top500.org/>

Destaca en la lista el crecimiento de estos sistemas en China, con los dos supercomputadores más potentes del mundo, y el mayor número de supercomputadores instalados, especialmente en el campo industrial y en el área de servicios. Aunque ha desplazado a EE UU, este país sigue teniendo el mayor número de supercomputadores dedicados a la investigación,

si bien Europa globalmente iguala en esta área, y además cuenta actualmente con el número 3 de la lista, instalado en Suiza.

España cuenta únicamente con un supercomputador en la lista (Figura 3), aunque destaca por dos motivos: *i)* con una potencia que excede los 10.000 Tflops, Mare Nostrum-4 es el segundo sistema en Europa, y *ii)* está emplazado en el BSC<sup>22</sup>, Barcelona Supercomputing Center/Centro Nacional de Supercomputación, lo que asegura la interacción con los mejores especialistas a nivel europeo tanto en la arquitectura del sistema como en las aplicaciones en diferentes áreas. A nivel nacional España cuenta además con la Red Española de Supercomputación (RES)<sup>23</sup>, coordinada por el BSC y en la que se integran los sistemas emplazados en las diferentes comunidades autónomas, universidades y centros de investigación. Esto permite llegar a los investigadores y también a las empresas en todo el territorio

<sup>22</sup> BSC, ver <http://www.bsc.es>

<sup>23</sup> RES. <http://www.bsc.es/RES>

nacional, ya que existe una oferta a nivel global de cerca de 500.000.000 de horas al año, abierta a toda la comunidad, como ICTS<sup>24</sup>.

Esta oferta se extiende a nivel europeo con la iniciativa PRACE, que permite a investigadores de toda Europa cubrir, de forma competitiva, las necesidades para proyectos muy grandes, que necesitan de grandes supercomputadores y millones de horas de ejecución.

En el área Cloud HPC la situación es más compleja. Como ya se ha indicado, la mayoría de los proveedores de servicios Cloud HPC son grandes multinacionales, y el coste de estos servicios es relativamente elevado. La mayor iniciativa hasta ahora en computación distribuida en Europa, tanto en volumen de datos (más de 200 *petabytes*) como capacidad de procesamiento (más de 600.000 *cores*), es la que coordina la iniciativa paneuropea EGI.eu<sup>25</sup>, y que coordina recursos de todos los países europeos, y da soporte a grupos de investigación en todas las áreas científicas, aunque destaca especialmente el procesamiento de datos del LHC del CERN. Sin embargo, esta infraestructura ha operado hasta el momento utilizando la tecnología GRID, y la evolución a un modelo Cloud, mediante la iniciativa FedCloud, está aún pendiente. A nivel nacional el IFCA coordina la participación española en EGI.eu, y también ha liderado diversos desarrollos en FedCloud, y en particular su uso por la ESFRI<sup>26</sup> LifeWatch<sup>27</sup>.

En paralelo, varios proyectos europeos han explorado la integración de proveedores de recursos Cloud HPC, destacando por ejemplo el desarrollo de soluciones en el proyecto INDIGO-DataCloud que permiten

que los grupos de investigación integren sus recursos locales en un Cloud híbrido. Este modelo puede ofrecer una excelente oportunidad de colaboración entre grupos de investigación y empresas proveedoras, que actualmente en Europa son muy limitadas, y en España no están claramente identificadas.

El modelo Cloud híbrido puede además ser una excelente solución para implementar la e-infraestructura requerida por el European Open Science Cloud.

#### 4. ¿La disrupción? del EOSC al *deep learning*

La iniciativa actual más relevante a nivel europeo en cuanto a e-infraestructura para *data science* es el European Open Science Cloud. Con un presupuesto inicial de más de 200.000.000 de euros, el EOSC se plantea como una infraestructura de datos que unirá los recursos existentes a nivel nacional y también europeo, en particular las ESFRI y las e-infraestructuras. Se implementará como un sistema de servicios, muy posiblemente siguiendo un modelo Cloud. La idea del EOSC es apoyar el ciclo de vida de los datos para impulsar la investigación, integrando los servicios computacionales necesarios para ello.

Entre estos servicios destaca la iniciativa EuroHPC para adquirir y desplegar una infraestructura paneuropea de supercomputación *exascale*, es decir al menos diez veces más potente que el conjunto de recursos existentes actualmente, y especialmente orientada a la explotación intensiva de esos datos.

La ambiciosa iniciativa del EOSC tenía inicialmente tres pilares: ciencia en abierto (es decir, lograr que todos los resultados de los proyectos financiados públicamente en Europa sean públicos), implicación de la industria TIC europea, e impacto en la innovación en Europa.

Cuando pensamos en un modelo de innovación basado en *data science*, esta conjunción parece una muy buena idea. Pero hay que analizar muy cuidadosamente qué va a significar su implementación. En primer lugar, es cierto que el acceso en abierto a los datos es la mejor forma de permitir su reuso, y esto es

<sup>24</sup> Instalación Científico-Técnica Singular del Ministerio de Economía, Industria y Competitividad.

<sup>25</sup> EGI.eu, <https://www.egi.eu/>

<sup>26</sup> El término ESFRI hace referencia a las infraestructuras de investigación consideradas en el «European Strategy Forum on Research Infrastructures», acrónimo ESFRI, y que cuentan por ello con un apoyo a nivel paneuropeo.

<sup>27</sup> LifeWatch es una ESFRI (infraestructura de investigación europea) orientada al estudio de la biodiversidad y los ecosistemas, liderada por España, y con sede en Sevilla. <http://www.lifewatch.eu>

crítico en los proyectos *de Data Science* multidisciplinarios. Pero para que esos datos sean reusados por los investigadores y la industria europea, es esencial que dispongan de los recursos y herramientas necesarios a la vez. De lo contrario podrán repetirse ejemplos como la actual experiencia con los datos en abierto de la misión Copernicus: es Google quien es capaz de gestionar y ofrecer esos datos a los investigadores de forma más efectiva, a cambio de que usen una plataforma en la que depositan su conocimiento.

En segundo lugar, está la cuestión de cómo reforzar el papel de la industria TIC europea. La experiencia hasta el momento es que los incentivos para competir con grandes multinacionales en esta área no parecen suficientes, y ello implica que van a tener una participación reducida en los posibles desarrollos tecnológicos y nuevos estándares.

En tercer lugar, está la apuesta por un sistema *exascale* como solución de computación. Esta idea tiene un interés estratégico claro: para construir dicho sistema, que requiere millones de procesadores de bajo consumo, Europa podría desarrollar su propia tecnología. Pero, por otro lado, ni los problemas usuales de *big data* que se van a abordar con mayor impacto en la innovación en la industria, desde medicina personalizada al diseño de nuevos productos, requieren sistemas de tan gran escala, ni el entorno de supercomputación es tan flexible como el de los servicios Cloud HPC.

Sin embargo, varios factores pueden hacer cambiar esta perspectiva del EOSC, en principio no muy favorable a la innovación ni a la participación de las empresas.

Y el factor clave es el impacto de los nuevos avances en inteligencia artificial que, explotando técnicas de *data science*, están revolucionando muchas áreas de aplicación.

Desde finales de los noventa, y gracias a la mejora de la potencia de cálculo de los sistemas, esta combinación empieza a ser aplicada en áreas como el diagnóstico médico o la logística. En 2010 una aplicación de aprendizaje automático de IBM fue capaz de ganar

en un concurso de TV de preguntas de cultura general, integrando técnicas diversas que incluían el procesamiento de lenguaje natural, pero también el acceso a grandes bases de datos de conocimiento, derivadas de... Wikipedia. La aplicación de este sistema, denominado Watson<sup>28</sup>, a áreas específicas, tales como medicina o derecho, permite contar con asistentes virtuales que facilitan la labor de los profesionales<sup>29</sup>. Una parte de la tecnología de esta aplicación se basa en el uso de la semántica, un área clásica en inteligencia artificial.

Ya en 1997 otra aplicación de IBM, *Deep Blue*<sup>30</sup>, había derrotado al campeón mundial de ajedrez Gary Kasparov, gracias a su gran capacidad de exploración de las jugadas futuras empleando su capacidad de cálculo. En contraste, 20 años después, en 2017, una aplicación de Google consiguió derrotar al campeón mundial de un juego oriental mucho más complicado, el Go, pero la técnica empleada es totalmente diferente: *deep learning*.

En los últimos años, el *deep learning*<sup>31</sup> (Figura 4) se ha establecido como uno de los candidatos más prometedores para alcanzar el objetivo a largo plazo de construir una solución de inteligencia artificial multipropósito. Ya ha revolucionado las aplicaciones en varios campos que van desde la visión por computador al reconocimiento de voz. Un marco flexible y multiuso ha permitido extender su aplicación a muchos campos de la ciencia. Aunque las herramientas básicas de esta tecnología se conocen desde hace mucho tiempo, la mayor revolución en este campo se debe a la evolución de técnicas especializadas, como las *Convolutional Neural Networks* (CNN), y al aumento de la potencia de cálculo empleando coprocesadores especiales especialmente adaptados

<sup>28</sup> Las publicaciones relativas al sistema Watson de IBM se encuentran disponibles online en [http://researcher.watson.ibm.com/researcher/view\\_group\\_pubs.php?grp=2099](http://researcher.watson.ibm.com/researcher/view_group_pubs.php?grp=2099)

<sup>29</sup> Un ejemplo reciente es el uso de Watson para la «optimización» de las declaraciones fiscales.

<sup>30</sup> Más información en <https://www.research.ibm.com/deepblue/>

<sup>31</sup> LECUN, Y.; BENGIO, Y. y HINTON, G. (2015).

FIGURA 4

CATÁLOGO DE *DEEP LEARNING APPLICATIONS* DEL IFCA:  
UNA TÉCNICA, VARIAS APLICACIONES



Deep Learning Applications <https://deep.ifca.es>  
Applications developed at IFCA using Deep Learning techniques



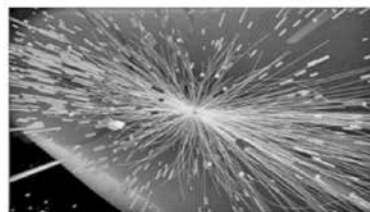
**PLANTS**

**Contact:** Ignacio Heredia  
Classification of a plant image  
among 6.000 plant species  
(mainly from Western Europe).  
Smartphone App. available

**Image Recognition**

**CONUS**

**Contact:** Lara Lloret  
Classification of a conus image  
among 68 conus species.  
Smartphone App. available



**PARTICLE PHYSICS  
COLLISIONS**

**Contact:** Celia Fernández  
Performance of convolutional nets at  
classifying the image of a collision in  
particle physics among several possible  
physical processes.

**FUENTE:** Instituto de Física de Cantabria.

a dichas técnicas, como son las GPU<sup>32</sup> (inicialmente unidades de proceso gráfico).

Estas redes neuronales convolucionales, inspiradas por modelos biológicos, están diseñadas específicamente para el reconocimiento de imágenes y señales, aprenden utilizando grandes colecciones de datos que toman como referencia, y combinan la extracción de características relevantes con el proceso de aprendizaje. Las GPU permiten realizar operaciones matriciales de una manera muy eficiente en comparación con una CPU, que son claves en el entrenamiento.

<sup>32</sup> Mientras que la CPU (*Central Processing Unit*) es un procesador de propósito general que se encarga de controlar y realizar la mayoría de las tareas en un computador, una GPU (*Graphics Processor Unit*) es un coprocesador especializado que realiza de forma muy eficiente operaciones dedicadas al procesamiento de gráficos.

El avance en esta línea requiere grandes colecciones de datos «etiquetadas», uno de los objetivos del European Open Science Cloud, y el *hardware* necesario para entrenar las redes sobre estas grandes colecciones.

Y la buena noticia es que muchos de los equipos actuales de supercomputación ya incorporan GPUs: en la lista Top500 hay más de 70 supercomputadores con chips del fabricante NVIDIA, incluyendo los dos supercomputadores citados anteriormente, uno de Facebook y otro de la propia empresa NVIDIA, que ocupan los lugares 3 y 4 entre los supercomputadores de uso industrial. En realidad, la incorporación masiva de estos chips a los supercomputadores tiene mucho que ver con su aplicación también en procesos de simulación, especialmente de fluidos, pero es de esperar que el número de supercomputadores para

aplicaciones de *deep learning* crezca rápidamente para abordar problemas en áreas muy diversas.

Del mismo modo, en el entorno de HPC Cloud varias compañías ya ofrecen GPUs como IaaS y hay proyectos que están desarrollando *deep as a service*, es decir la posibilidad para el usuario de definir la arquitectura de una aplicación de *deep learning* y ejecutarla en una infraestructura Cloud con GPU, sin preocuparse de los detalles de *hardware*, acceso a los datos de entrenamiento y de escalabilidad.

¿Podemos pues ser optimistas y pensar que en el contexto del European Open Science Cloud, quizás en España, vamos a ser capaces de explotar colecciones de datos en abierto con estas técnicas sobre sistemas que integren procesadores específicos desarrollados en el contexto EuroHPC con el objetivo de alcanzar la *exascale*? Y ¿serán nuestras empresas las que se beneficien de estas iniciativas tanto para desarrollar los sistemas y proveer servicios Cloud como para desarrollar las múltiples aplicaciones de las que esperamos que se acabe beneficiando la sociedad, pero que van a suponer el reemplazo de muchas técnicas, y profesiones, actuales?

Para ello, se proponen varias acciones:

— Formación, tanto académica como profesional, en una nueva disciplina transversal como es *Data Science*, que debe consolidar un currículo común<sup>33</sup>, y que además requiere una formación previa en un dominio de especialización (sea técnico, como matemáticas, física, informática o ingenierías, o no, por ejemplo derecho, economía, medicina).

— Las alianzas entre centros de investigación y proveedores de servicios Cloud HPC en España, para desarrollar soluciones híbridas que permitan por un lado consolidar una base de test que incluya tanto acceso a datos en abierto como a recursos HPC incluyendo GPUs, y por otro escalar cuando sea necesario y se

cuenta con financiación para abordar proyectos ambiciosos.

— Expandir la base de acceso a los recursos de supercomputación, y planificar la conexión con los recursos de datos, siguiendo de cerca la iniciativa EDI (*European Data Infrastructure*) del European Open Science Cloud, y el desarrollo de procesadores específicos en EuroHPC.

— Aprovechar el potencial de las técnicas generales de *machine learning* para aplicarlas en nuevos campos, como la medicina personalizada, o la agricultura de precisión, integrando grandes colecciones de datos (médicos, genéticos, de uso del suelo, climáticos, etc.).

— Promover la colaboración de pymes TIC innovadoras en España con grupos profesionales en áreas como medicina, derecho, medio ambiente, o gestión de recursos locales, para desarrollar y desplegar una nueva generación de «asistentes digitales» que, usando técnicas de *deep learning* sobre recursos HPC, permita aliviar la carga de trabajo en estos sectores y a la vez mejorar enormemente el tiempo de respuesta a los ciudadanos.

## Referencias bibliográficas

- [1] AGUILAR, F. *et al.* (2016). «Initial Specifications of Data Ingestion in INDIGO» – Deliverable D2.11 INDIGO-DataCloud project, <https://www.indigo-datacloud.eu/documents/initial-specifications-data-ingestion-indigo-d211>
- [2] BEYER, M (2016). «Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data». GARTNER, <http://www.gartner.com/it/page.jsp?id=1731916>
- [3] CERN (2011). *The CMS Experiment at LHC*, CERN. En <https://cms.cern/>
- [4] LECUN, Y.; BENGIO, Y. y HINTON, G. (2015). «Deep Learning». *Nature*, vol. 521, nº 7.553, pp. 436-444, mayo.
- [5] MELL, P.; GRANCE, T. (2011). *The NIST definition of Cloud computing*. Sp 800-145 En <https://csrc.nist.gov/publications/detail/sp/800-145/final>
- [6] MIRÓN, F. *et al.* (2017). *Análisis de la estrategia big data en España*. Disponible en <http://planetica.es/content/whitepaper-iniciativa-big-data>
- [7] TRINITY CRAY XC40 (2017). En <https://www.top500.org/system/178610>

<sup>33</sup> El proyecto europeo EDISON ha propuesto un currículo estándar, que ya ha sido adoptado por diferentes grados y másteres en Europa, incluyendo uno oficial en España, el máster UIMP/UC en *Data Science*.

## ***Jorge Juan y la Ciencia ilustrada***



La presente publicación pretende poner al día la figura de Jorge Juan y reivindicar la importancia de su trayectoria en el intento ilustrado que, a lo largo del siglo XVIII, trató de equiparar la actividad científica en nuestro país con la que se desarrollaba en Francia e Inglaterra. Aunque existe una amplia bibliografía sobre el personaje, los estudios que se integran en este volumen, elaborados por los mejores expertos en la biografía y el trabajo científico de Jorge Juan, permiten una visión de conjunto de la contribución del marino alicantino a la ciencia de su tiempo. Se destaca el recorrido de esta personalidad polifacética y visionaria, ofreciendo una aproximación multidisciplinar a su figura y sus logros.

**Número de páginas:** 94

**Precio papel:** 8,00 €

**Pdf:** 4,00 €

**ePub:** 6,00 €

(IVA incluido)