Del *cloud computing* al *big data*

Visión introductoria para jóvenes emprendedores

Jordi Torres i Viñals

PID_00194204

UOC
Universitat Oberta de Catalunya

Jordi Torres i Viñals

Jordi Torres i Viñals es profesor, investigador y divulgador de las nuevas tecnologías TIC. Catedrático en la UPC Barcelona Tech con más de veinte años de experiencia en I+D y dirige un grupo de investigación en el Barcelona Supercomputing Center, donde realiza proyectos con empresas. Actúa como experto para diferentes organizaciones, miembro del Consejo de Administración, imparte conferencias y colabora con diferentes medios de comunicación. Su interés actual se centra en conseguir un mayor y más sostenible provecho de los recursos TIC (*green computing*) en el escenario marcado por el *cloud computing* y la explosión de la información disponible (*big data*), tanto para el beneficio de las empresas como para el de la sociedad en general. Podéis encontrar más información sobre sus actividades en: http://www.JordiTorres.eu

El encargo y la creación de este material docente han sido coordinados por el profesor: José Antonio Morán (2012)

Primera edición: septiembre 2012

Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Eureca Media, SL

A todos mis alumnos, que hacen que
mi trabajo sea tan apasionante.

Este material no lo habría podido escribir si antes no hubiera escrito *Empresas en la nube*, por eso, nuevamente, vayan mis agradecimientos a todos y cada uno de los que me ayudasteis a hacerlo posible. En todo caso, me gustaría añadir explícitamente a los compañeros con los que hemos empezado a trabajar en este apasionante mundo del *big data* en el Barcelona Supercomputing Center y en la Universidad Politécnica de Cataluña, sin cuya colaboración tampoco habría podido escribir esta parte: David Carrera, Yolanda Becerra, Eduard Ayguadé, Juan Luis Pérez, Jordà Polo, Ricard Gavaldà y Mario Macías. Quiero agradecer igualmente a José Antonio Morán y Joan Antoni Pastor de la UOC por permitir que esta obra viera la luz. Y finalmente el mayor de los agradecimientos a mi familia, que han visto cómo les he recortado horas de estar con ellos los fines de semana para poder escribir estos contenidos.

Índice

Introducción

Para los estudiantes a los que va dirigido este material es un hecho natural que los servicios converjan y pasen del mundo físico al mundo digital, de modo que sean accesibles desde cualquier dispositivo electrónico. Forman parte de una generación que ha vivido toda su vida adulta en un mundo físico y digital a la vez y no entienden la tecnología solo como una herramienta de trabajo. Se levantan con la tecnología y se van a dormir con ella. No acceden a las redes sociales únicamente desde su ordenador, sino que es habitual que también lo hagan desde el móvil. Colgar fotos en un instante, enviar un *whatsapp* o compartir la ubicación con el resto de usuarios son hechos habituales.

Al mismo tiempo, el mundo de los negocios también se está transformando a gran velocidad. La desmaterialización ya ha revolucionado diferentes sectores de negocio y, en algunos, incluso ha acabado con un modelo que había estado funcionando durante años.

> Un ejemplo conocido por todos es el del mundo de la música, en el que, hasta hace poco, las discográficas controlaban toda la cadena de producción, desde el autor hasta el consumidor. Internet ha puesto fin a este negocio por el mero hecho de hacer prescindible el soporte físico del CD, lo que ha dejado sin sentido el rol que jugaban todos los intermediarios. No es simplemente que haya cambiado la manera de vender canciones sino que Internet ha convertido la música en un servicio más. Ahora, lo más importante ya no es poseer la música sino el acceso a las piezas que uno quiere escuchar, ¿no es cierto?

El ***cloud computing*** es lo que hace posible que la tecnología digital penetre en cada rincón de nuestra economía y nuestra sociedad, no solo permitiendo que los usuarios estén conectados a este nuevo mundo digital a través de sus dispositivos móviles, sino posibilitando que también lo esté en breve cualquier objeto o dispositivo, lo que se denomina ***Internet of things***. Todo ello provocará una marea de información digital que requerirá de una capacidad de almacenamiento y procesado de magnitudes hasta ahora nunca vistas, fenómeno que se empieza a denominar ***big data*** y que sin duda representa un océano de oportunidades profesionales para todos vosotros.

Este módulo hace un breve repaso a las tecnologías relacionadas con el *cloud computing* que marcarán nuestro futuro inmediato y, asimismo, intenta apuntar hacia dónde evolucionarán, para lo que incluye referencias a documentos que permiten profundizar en cada una de ellas y alcanzar así una formación más personalizada.

Por eso, además de describir el fenómeno del *cloud computing* que ya está aquí, intentaremos visualizar hacia dónde irá en los próximos años, con el objetivo de ayudar a los estudiantes, a los que va dirigido este material, a convertirse en partícipes de la profunda transformación que está provocando el *cloud computing* en la estructura económica, social y cultural, ya que incide en casi todos

los aspectos de nuestra vida: en el mundo laboral, mediante su actuación en la empresa, la economía, la sanidad, el diseño o la burocracia administrativa; y en el social y cultural, gracias a su presencia evidente en nuestras relaciones personales o en nuestros sistemas de información, educación, comunicación y ocio. Es indudable que ocupa y ocupará un papel central en nuestra vida, tanto personal como profesional, a pesar de que quizá no seamos plenamente conscientes de ello.

En definitiva, este libro quiere favorecer el descubrimiento de estas nuevas tecnologías a nuestros estudiantes, con el deseo de estimularlos para que se adentren algo más en este apasionante mundo de la tecnología y se animen a emprender.

1. *Cloud computing*

1.1. La red se ha convertido en el ordenador

A pesar de ser un módulo que quiere hablar de futuro, haremos una breve referencia a los orígenes del *cloud computing* para ayudar a comprender por qué se producen tan rápidamente los cambios tecnológicos: las tecnologías de la información y la comunicación[1] son muy jóvenes, y por eso todavía les queda mucho recorrido. Pensemos que los primeros ordenadores se remontan a no más allá de la mitad del siglo pasado, funcionaban con válvulas parecidas a bombillas y su uso estaba dedicado exclusivamente a los ámbitos científico y militar. La programación se hacía directamente en los propios circuitos de las máquinas. Poco después aparecieron los transistores y también los primeros ordenadores que permitían una cierta programación.

(1)A partir de ahora usaremos el acrónimo TIC.

Pero no fue hasta mediados de los años sesenta cuando la informática traspasó la frontera de la investigación para entrar progresivamente en el ámbito de los negocios. En aquellos momentos, la informática se basaba en grandes ordenadores ubicados en centros de proceso de datos de las empresas. Unos ordenadores muy costosos que eran compartidos por muchos usuarios y estaban gestionados por los nuevos departamentos de informática que iban apareciendo en las empresas. En los inicios se programaba con tarjetas perforadas por unas máquinas parecidas a las máquinas de escribir, que el ordenador leía con un lector de tarjetas parecido a los contadores de billetes que tienen los bancos. El resultado de la computación se imprimía normalmente sobre papel por medio de unas impresoras conectadas a estos ordenadores centrales. Más tarde llegaron los denominados terminales, con una pantalla y un teclado por cada puesto de trabajo, que estaban conectados a un ordenador central llamado *mainframe*. Varios usuarios ya podían interaccionar con el ordenador central a través del teclado y visualizar a la vez los resultados en pantalla. Todo un avance.

Fue hacia el año 1980 cuando llegó lo que se conoce por *PC*[2] u **ordenadores personales**. Los ordenadores dejaron de ser un producto extremadamente caro, lo que hizo posible que la informática estuviera al alcance de las pequeñas y medianas empresas, e incluso de los particulares. Ahora, las empresas podían disponer de tantos PC como hiciera falta para soportar las diversas aplicaciones, lo cual comportó que acabaran teniendo una serie de máquinas distribuidas por varios departamentos que gestionaban diferentes aplicaciones.

(2)Acrónimo en inglés de *personal computer*.

A pesar de que podríamos considerar que la infraestructura de Internet, tal como la conocemos hoy en día, nació a mediados de los años ochenta, no fue hasta mediados de los años noventa cuando empezaron las empresas a usarla realmente. Sin duda representó un cambio muy importante que permitió aumentar tanto el número de usuarios como el de aplicaciones disponibles.

A partir de aquí es cuando aparecieron webs como Yahoo, Google o Amazon, que daban servicio a un amplio abanico de usuarios que se conectaban desde cualquier lugar del mundo con el propósito de buscar información, comprar o simplemente navegar. En este período es cuando nace lo que se ha denominado **Web 2.0**, donde los usuarios ya no son solo consumidores de información sino también productores de una información que ahora puede ser consumida por el resto de las personas a través de la red Internet. Es el momento de los blogs, las wikis y el inicio de las redes sociales.

Todo ello hace aparecer la necesidad de potentes servidores que tienen que dar servicio a una infinidad de usuarios que se conectan desde cualquier tipo de dispositivo. Servidores que ya no se encuentran alojados en las empresas que han creado los programas informáticos o los servicios. Han desplazado toda la computación y almacenamiento de sus servidores hacia algún sitio externo desde el que puedan acceder fácilmente a través de Internet a grandes centros de proceso de datos[3]. La red se ha convertido en nuestro ordenador. Es lo que se conoce como ***cloud computing*** o informática **en la nube de Internet**. Aunque en los siguientes apartados profundizaremos algo más en este nuevo paradigma de computación, en la tabla 1 adelantamos un breve resumen esquemático que compara los tres escenarios de computación que hemos presentado en este punto.

[3]A partir de ahora usaremos el acrónimo CPD.

Tabla 1. Comparativa de los tres escenarios de computación y almacenamiento.

	Modelo de computación y almacenamiento	Características	Costes
Mainframe	Centralizado	Sistemas que procuraban aprovechar al máximo los recursos a causa del alto coste de estos.	Inversión inicial tanto para el hardware como para el software.
PC	Distribuido	PC y servidores distribuidos conectados en red (primero local y después Internet).	Inversión inicial para la compra de hardware y costes de licencias para el SO y las aplicaciones software que se usaban.
Cloud	Centralizado	Inmensos CPD con recursos TIC de bajo coste (*commodity*) que se optimizan aprovechando la economía de escala.	Se paga solo por lo que se gasta a medida que se va usando a lo largo del tiempo.

1.2. Definiendo el *cloud computing*

Parece que el nombre viene de los términos *cloud* (nube), que es el grafismo habitual que se usa para representar Internet, y *computing* (computación), que se podría considerar que reúne conceptos como informática y almacenamiento. No hay una única definición pero podríamos quedarnos con la que ha formulado el NIST (Instituto Nacional de Estándares y Tecnología estadounidense):

> "El *cloud computing* es un modelo que permite el acceso bajo demanda a través de la red a un conjunto compartido de recursos de computación configurables (como por ejemplo red, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente aprovisionados con el mínimo esfuerzo de gestión o interacción del proveedor del servicio".

Este instituto identifica varias características esenciales del *cloud computing*, indispensables para ver la potencia que puede tener. Lo primero que se debe subrayar es que ha de permitir que el consumidor se provea **unilateralmente** de los servicios que necesite sin la interacción de los recursos humanos del propio proveedor de servicios.

Actualmente se habla de diferentes modelos o niveles de servicio, de los que los más habituales son los tres siguientes:

- Infraestructura como servicio (IaaS[4])
- Plataforma como servicio (PaaS[5])
- Software como servicio (SaaS[6])

Lectura recomendada

Para definir los conceptos básicos relacionados con el *cloud computing* se suele utilizar como referencia el documento *The NIST definition of Cloud computing* del Instituto Nacional de Estándares y Tecnología (NIST) estadounidense. Lo podéis descargar en esta dirección:

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

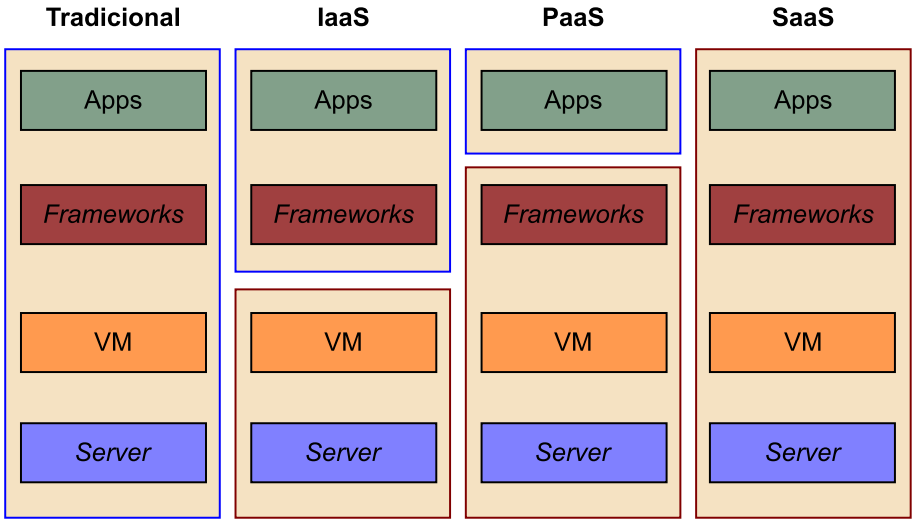
(documento en inglés)

(4)Acrónimo en inglés de *infraestructure as a service.*

(5)Acrónimo en inglés de *platform as a service.*

(6)Acrónimo en inglés de *software as a service.*

Figura 1. Esquema de los tres niveles de servicio del *cloud computing* habituales respecto al tradicional.



1.2.1. Infraestructura como servicio

La **infraestructura como servicio** hace referencia al hecho de ofrecer servicios de computación y almacenamiento, de tal manera que podamos disponer de recursos como ciclos de CPU, memoria, disco o equipamientos de red. El consumidor alquila los recursos de hardware en vez de comprarlos e instalarlos en su propio CPD, lo que le permite ir variando el consumo de los recursos en función de sus necesidades, lo que se conoce como elasticidad de la infraestructura.

Imaginemos que creamos una aplicación web que tiene que atender a diferentes cantidades de clientes en distintas franjas horarias. Si repentinamente se necesitan más recursos, estos se proveen puntualmente hasta que la necesidad desaparece.

Amazon Web Service

Un ejemplo muy referenciado para ilustrar este tipo de servicio es el de Amazon Web Service (AWS), con su *elastic cloud computing* (EC2). Además de proporcionar la flexibilidad necesaria para escoger fácilmente el número, medida y configuración de las instancias de servidores que necesita el usuario para su aplicación, Amazon EC2 proporciona a los clientes diferentes modelos que les permiten tener la flexibilidad necesaria para optimizar los costes. Por ejemplo, las instancias llamadas "bajo demanda" les permiten pagar una tarifa fija por hora sin ningún tipo de compromiso, mientras que las instancias llamadas "reservadas" ofrecen la posibilidad de abonar una tarifa de entrada y, a cambio, obtener un descuento por cada hora que se use.

Ejemplos europeos

Ejemplos cercanos, europeos, pueden serlo T-Systems, Flexiscale o CloudSigma, que proveen instancias virtuales y una interfaz para poner en marcha, parar y acceder a la configuración y control del servidor virtual del mismo modo que accedemos a cualquier servidor que tengamos en nuestro CPD.

Amazon

Amazon ofrece un kit de uso gratuito de AWS durante un año para poder iniciarse con el entorno. Se compone de 750 horas/mes de instancias de servidores micro, junto con varios recursos de almacenamiento y uso de red. Podéis encontrar la información en:

http://aws.amazon.com/es/free/

1.2.2. Plataforma como servicio

Ahora bien, lo que a veces hace falta es poder tener una máquina que ya disponga de un determinado software que facilite el desarrollo y la implantación de una nueva aplicación de manera rápida. Tal es el caso de lo que se conoce como "plataforma como servicio".

En la **plataforma como servicio** el proveedor ofrece algo más que la infraestructura, pues incluye todo lo necesario para construir nuevas aplicaciones y servicios, lo que facilita el ciclo completo de construcción y puesta en funcionamiento de las aplicaciones.

Este nivel está muy orientado a los equipos de desarrollo, a los que les permite reducir el tiempo de desarrollo e implementación de las aplicaciones, puesto que no se requiere instalar herramientas software de apoyo (lo que evita los problemas de versiones o licencias), y les facilita a la vez la compartición y la interacción de las aplicaciones.

Google App Engine

Hay varios ejemplos de plataforma como servicio, entre los que figura AZURE de Microsoft, pero el más popular y a vuestro alcance es probablemente Google App Engine, que ofrece servicios para crear aplicaciones dentro de toda la infraestructura de Google. La idea es que no se necesita ningún servidor, sencillamente se trata de cargar la aplicación en el espacio de Google App Engine para que sea accesible a vuestros usuarios o clientes, que así pueden acceder desde vuestro propio dominio. Google App Engine admite aplicaciones escritas en varios lenguajes de programación, como Python, Java, Javascript o Ruby. También sigue un modelo de pago por lo que se usa.

Google

Google ofrece un kit gratuito de su App Engine de hasta 500 MB de almacenamiento con los ciclos de CPU y ancho de banda de la red, necesarios para permitir un número muy elevado de visitas al mes. Podéis encontrar la información en:

https://developers.google.com/appengine/docs/whatisgoogleappengine?hl=se

1.2.3. Software como servicio

En el caso del **software como servicio**, el proveedor no solo ofrece la infraestructura hardware y los entornos de ejecución necesarios, sino también los productos software que interaccionamos con el usuario desde un determinado portal o interfaz a través de Internet.

Podríamos decir que es lo opuesto a un software como producto, ya que permite convertir el coste de uso en una variable más del negocio y evita las habituales inversiones iniciales en licencias. Por ejemplo, no es necesario comprar el software o sus licencias, ni ningún tipo de hardware, y deja a cargo del proveedor el mantenimiento y el apoyo tanto del software como del hardware. El precio por uso, en general, le resulta mucho más bajo a la empresa que si tuviera que adquirir el mismo servicio por sí sola, debido a las economías de escala que aplican los proveedores.

Ahora mismo existe un número muy grande de servicios de estas características con todo tipo de aplicaciones destinadas a diferentes áreas de la empresa, como la gestión de recursos humanos o la gestión financiera. También se puede encontrar un conjunto muy exitoso de servicios de herramientas de escritorio que incluyen las habituales hojas de cálculo o diversos procesadores de texto, entre otros.

Gmail y Hotmail

Algunos de los ejemplos más populares de software como servicio son Gmail y Hotmail que, en este caso, incluso son gratuitos para el usuario final. Este es probablemente el modelo de servicio más conocido por todos vosotros.

Ahora bien, además de estos tres tipos de servicios *del cloud computing* comentados, habría que mencionar que el mercado del *cloud computing* está madurando muy rápidamente y que en estos momentos ya se están ofertando otros productos "como servicio" y, sobre todo, se está preparando la nube para ofrecer muchos más.

Lectura complementaria

Un documento en el que se describen en detalle las diferentes modalidades de *cloud* y que constituye a la vez un buen punto de partida para conocer la estrategia europea sobre el *cloud computing* es el informe de la Comisión Europea *The Future of Cloud Computing*, que podéis descargar en la dirección siguiente:

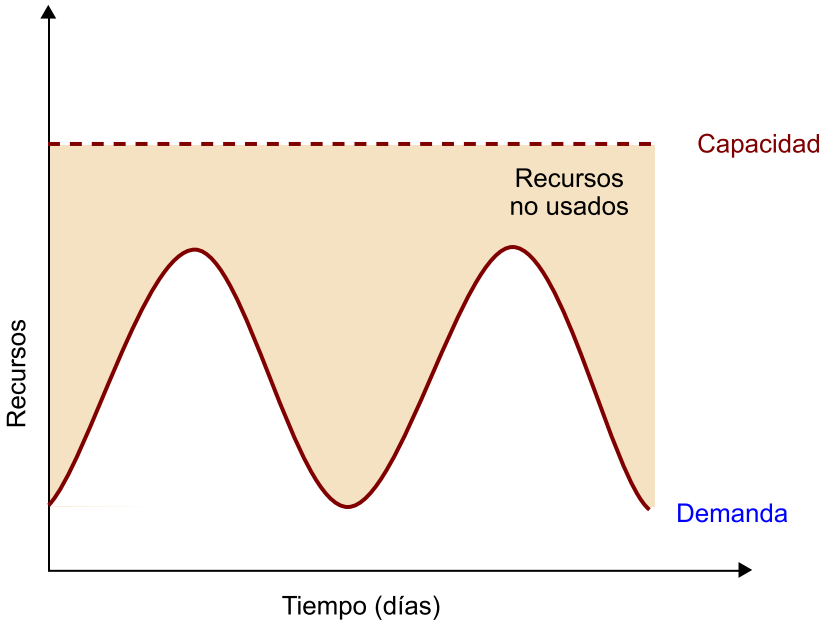
http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf (documento en inglés)

1.3. La flexibilidad del *cloud computing*

Con el *cloud computing* aparece un abanico de nuevos servicios a la vez que se ofrece la posibilidad de convertir gastos fijos en variables, lo que permite conocer mejor los costes reales de cada producto y minimizar a la vez los riesgos financieros en el lanzamiento de nuevos productos o servicios que dependan de las TIC. Una de las razones principales para su adopción tiene que ver con la **agilidad** con que podemos autoproveernos de los servicios ofrecidos a través del *cloud computing*. Ello comporta que se puedan reducir significativamente los plazos de implantación de una nueva necesidad TIC (la gestión dinámica de los servicios, de los recursos de cálculo, de almacenamiento o de acceso a la red), que pasan a ser de minutos u horas en vez de semanas o meses (que es lo que nos supondría en un CPD propio). Es decir, las infraestructuras o servicios TIC dejarán de ser un cuello de botella en la puesta en marcha de nuevos productos o servicios, lo que permitirá centrarse en la aplicación de negocio concreta que hay ante cualquier nuevo proyecto.

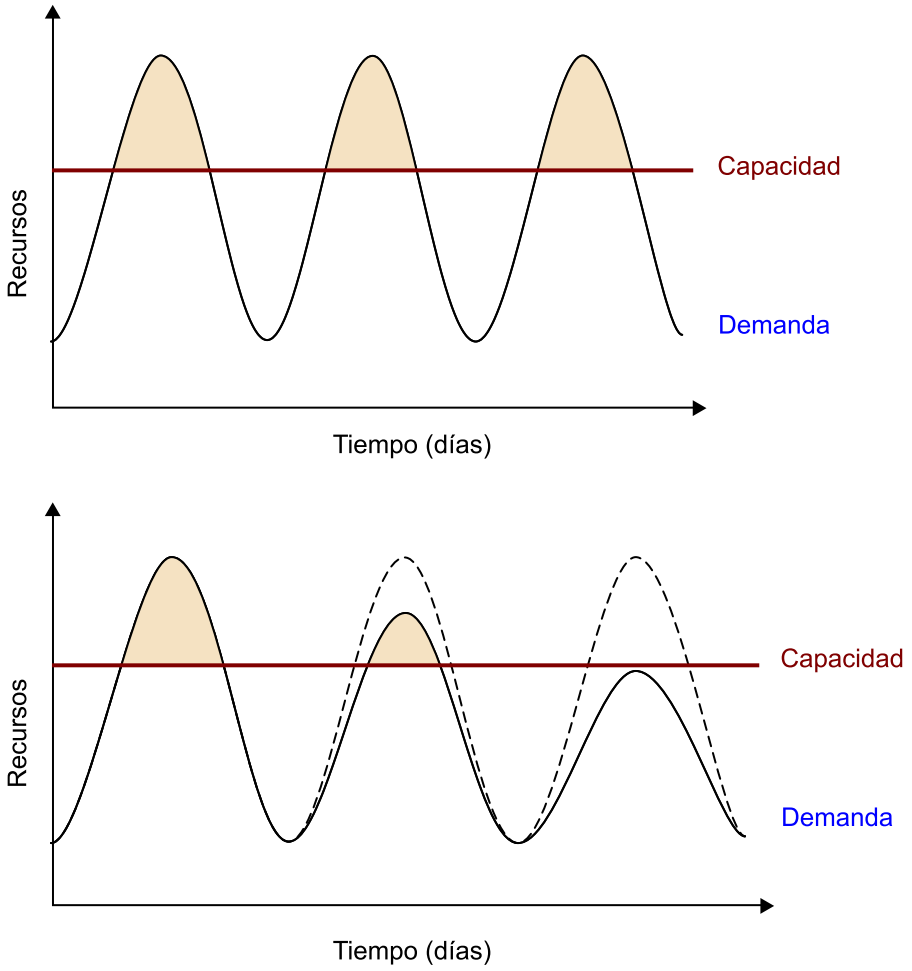
Otro aspecto relevante que presenta el *cloud computing* es lo que se denomina la ***escalabilidad***. Hasta ahora, si la empresa quería crecer en oferta, tenía que invertir primero en sus propios recursos informáticos (tanto hardware como software), pues de lo contrario podría llegar a colapsarse el sistema. Pero, por ejemplo, en el supuesto de que las ventas tengan un ciclo de temporada o basado en promociones especiales, lo que implica picos y valles enormes en el uso de los recursos TIC, cobra mucho interés poder disponer de un servicio flexible. Sin el *cloud computing,* la dificultad de prever la demanda motiva que se haga una provisión excesiva de sistemas TIC para garantizar que siempre se atenderán todas las solicitudes.

Figura 2. Esquema de demanda de carga variable cubierta por suficientes recursos para poder absorber las puntas de trabajo.



En la figura 2 se puede observar de manera esquemática la cantidad de recursos que una empresa necesita para soportar una determinada carga de clientes. La demanda de recursos es muy variable y proporcional a las peticiones de los clientes, y hacemos la suposición de que está relacionada con la hora del día en que dichas peticiones se producen. El gráfico representa esquemáticamente la llegada de peticiones al sistema (y el equivalente requerimiento de recursos) en función del momento en que se producen, siendo alta durante el día y baja durante la noche (una visión simplificada del comportamiento habitual en los portales web). Para cubrir una demanda variable se capacita al sistema con un número de recursos que permita absorber las puntas de trabajo, no previsibles fácilmente. Ello provoca una pérdida de recursos en el sentido de que no son usados, representada con la zona sombreada. Pensemos que los servidores, además del coste de compra, tienen un coste de consumo de energía por el solo hecho de haberlos puesto en marcha, aunque no estén atendiendo peticiones.

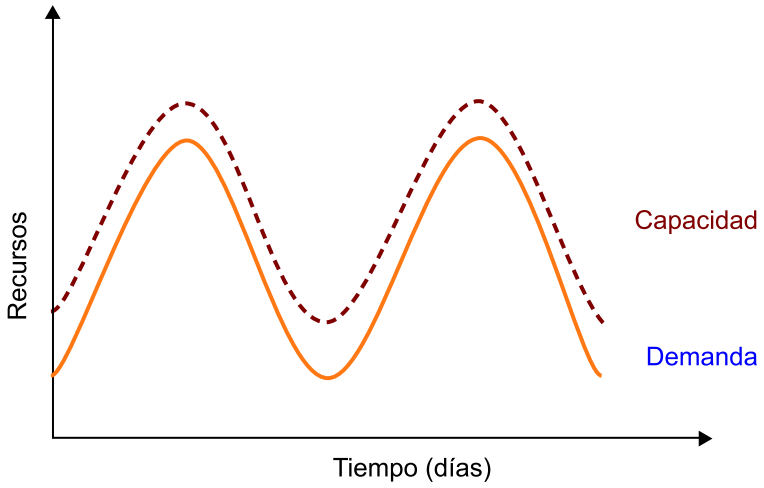
Figura 3. Comportamiento de los clientes ante una reducción de recursos disponibles para atenderlos.



Si se optara por reducir la provisión de recursos para minimizar este derroche, tal como se representa en la primera parte de la figura 3, entonces se estarían sacrificando ganancias potenciales provenientes de clientes no atendidos (todos los que no pueden ser atendidos por la falta de recursos en las puntas de trabajo). Se podría plantear que la pérdida por no atender a esta pequeña porción de clientes compensa la reducción del coste de derroche. Pero, habitualmente, los usuarios de servicios a través de Internet abandonan de manera permanente el sitio que no les ha prestado la debida atención. Es decir, la demanda se adapta rápidamente a la capacidad del sistema, reduciendo de manera definitiva nuestros clientes y por lo tanto, una parte del flujo de ingresos, como se muestra esquemáticamente en la segunda parte de la figura 3.

La elasticidad o escalabilidad que ofrece el *cloud computing* en la disponibilidad de recursos permite adaptarse a la demanda real, lo que reduce el riesgo de que estos queden infrautilizados durante los periodos menos productivos y a la vez permite que puedan soportar picos de trabajo imprevisibles, como se muestra esquemáticamente en la figura 4.

Figura 4. Esquema de cómo la elasticidad del *cloud computing* permite adaptarse a la demanda.



Por todo ello es por lo que la propiedad de elasticidad o escalabilidad del *cloud computing* es claramente una oportunidad para adaptarse a la demanda y eliminar el sobrecoste derivado de mantener los sistemas infrautilizados fuera de las puntas de demanda.

Lectura recomendada

Uno de los documentos técnicos más referenciados sobre el *cloud computing* es el informe *Above the Clouds: A Berkeley View of Cloud*, elaborado por un grupo de investigación de Berkeley, que, aunque es del 2009, continúa siendo de gran valor por sus didácticas explicaciones del nuevo paradigma de computación. El documento se puede descargar en la dirección siguiente:

http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

(documento en inglés)

1.4. Consumo energético y dependencia de la información

El verdadero corazón del *cloud computing* se encuentra en los grandes **centros de proceso de datos (CPD)**, donde la energía se ha convertido hoy en día en el principal coste para hacerlos funcionar. Afortunadamente, los avances en hardware y software permiten mejorar el consumo y gestión en estos CPD. Pero a pesar de las mejoras que han reducido la cantidad de electricidad que requiere su funcionamiento, consumirán cada vez más megavatios de potencia que en la actualidad. Pensemos que los ciudadanos van almacenando cada vez más información en esta nube, soportada por grandes CPD dispersos por todo el mundo, muchos de los cuales tienen el tamaño de varios campos de fútbol. ¿Os imagináis una nave industrial de esas dimensiones llena de hardware y software?

Internet no tiene fronteras como los países, lo que implica que nuestra información puede acabar siendo deslocalizada en la otra punta del planeta en uno de estos CPD –auténticas factorías de información del siglo XXI–, donde los costes de la energía pueden ser muy inferiores (pensemos que hay lugares donde el coste de la energía puede ser cinco veces más bajo que en Barcelona). Surgen entonces varias preguntas: ¿está garantizada la disponibilidad de nuestra información con esta deslocalización?, ¿podríamos encontrarnos algún día en la tesitura de no poder acceder a nuestros datos?

La respuesta la tenemos en el ejemplo del cierre por parte del Departamento de Justicia de EE. UU. y el FBI del popular servicio de intercambio de archivos Megaupload a principios del 2012, debido a la descarga de sus contenidos, una parte importante de los cuales pertenecían a usuarios que no eran los titulares

de los derechos. Pero no olvidemos tampoco que una parte de sus contenidos almacenados eran lícitos, de empresas y usuarios que tenían todos los derechos sobre sus datos.

¿Qué les pasó a las empresas que no pudieron acceder a sus contenidos de Megaupload? ¿Os imagináis lo que os pasaría si vuestras cuentas de Dropbox dejaran de ser accesibles?, ¿o si ya no os funcionara el servicio de Gmail? ¿Os habéis parado a pensar dónde están vuestros datos y a quiénes tendríais que dirigiros para poder recuperarlos?

La información es esencial en la sociedad actual, y por ello tenemos que evitar que se pueda perder por acontecimientos que se encuentran fuera de nuestro control. El motivo en el caso de Megaupload fue perseguir unos contenidos ilícitos de otros usuarios, pero en el escenario actual se puede perder el acceso a nuestra información por otras muchas causas. Para garantizar la accesibilidad de la información, tenemos que guardar nuestra parte de esta nube dentro de un territorio, donde se pueda tener un control sobre estas factorías de producción de información, que ya empiezan a ser un servicio público de interés general, lo mismo que el agua o la electricidad.

No obstante, estas fábricas de información requieren muchos miles de kilovatios de energía para funcionar, de forma que, si aproximadamente el 75% del sistema energético global en España es importado, nos encontraremos igualmente ante una situación de dependencia. Por eso, nuestra posición es que no nos queda más remedio que aumentar la producción de energía renovable (en realidad, la única que es autóctona) que se genera dentro del Estado para asegurar el mantenimiento de estas nuevas fábricas de información, de forma que no se dependa de la energía que viene de fuera y que nadie nos garantiza que siempre nos llegará.

Una posibilidad que se está valorando en el ámbito de la investigación es trasladar estas fábricas de información a los parques de energía renovable, de forma que estos parques no solo serán proveedores de energía, sino también proveedores de información, que no deja de ser la energía que necesita nuestra sociedad de la información en la que ya estamos inmersos. Esta propuesta es muy eficiente en varios aspectos, por ejemplo, la transferencia de datos a través de fibra óptica desde estos parques tiene un coste energético mínimo en comparación con las pérdidas que provoca el transporte de energía, que pueden llegar a ser del 10%. Aunque, debido a la discontinuidad de la energía renovable, por ahora se requiere energía de apoyo. Las investigaciones que se están llevando a cabo en estos momentos en centros de investigación, como el Barcelona Supercomputing Center (BSC) y la Universidad Politécnica de Cataluña (UPC) en colaboración con otros centros, nos hacen ser optimistas y pensar que la información se podrá mantener en gran parte solo con la energía renovable.

Por ejemplo, basándose en la predicción del tiempo, es posible ir moviendo cargas de computación entre diferentes parques, en función de la energía que cada parque tenga disponible en aquel momento, o retrasar automáticamente la ejecución de ciertos trabajos hasta que se vuelva a tener energía disponible.

Sin embargo, no podemos ignorar que una parte considerable de la energía consumida por las TIC corresponde en realidad al gran número de pequeños CPD existentes en pequeñas y medianas empresas. Estas instalaciones van desde una docena de servidores alojados en una sala de máquinas a varios centenares de servidores en el caso de empresas más grandes. Para este tipo de entornos, las investigaciones van hoy en día en la línea de construir contenedores (como los que se utilizan para el transporte marítimo de mercancías, pero de menor tamaño), que estarían equipados en las instalaciones con placas fotovoltaicas o pequeños aerogeneradores para proveerse de energía, a la vez que se conectarían a la red eléctrica como apoyo cuando la energía renovable no estuviera disponible.

El reto de la búsqueda actual de proveer los CPD con energía solar y eólica es que no siempre se encuentra esta disponible debido a su discontinuidad, si bien se pueden usar otras fuentes de energía, como la biomasa. Un tema al que todavía le queda mucho recorrido y que aún ofrece muchísimas oportunidades en investigación e innovación.

Lectura complementaria

Se pueden encontrar propuestas técnicas más detalladas desde nuestro grupo de investigación en este ámbito en el artículo "Towards Sustainable Solutions for European Cloud Computing" de la revista *UPGRADE,* publicada por la Asociación Europea de Profesionales Informáticos. El artículo se puede descargar en la dirección siguiente:

http://www.cepis.org/upgrade/media/le_2011_41.pdf (documento en inglés)

1.5. Los inhibidores del *cloud computing*

Como hemos visto en el subapartado anterior, Internet no tiene fronteras, y por eso nuestra información puede acabar deslocalizada en el otro extremo del mundo, provocando **problemas de disponibilidad**. Pero no solo es este uno de los temas que inquietan a los usuarios del *cloud*, también hay otros problemas que preocupan y que han frenado hasta ahora el desarrollo de la nube. Entre ellos, las dudas sobre **la seguridad**, **privacidad** e **integridad** de los datos en la nube constituyen una de las preocupaciones más frecuentes. También se alerta sobre la carencia de estándares que nos permitan un cambio fácil de proveedor. O sobre la madurez de los proveedores en este ámbito, que no siempre está suficientemente acreditada. Así como también sobre la legislación, todavía no homogénea del todo, en los diferentes países, a la vez que se requiere una mayor definición en aspectos como el control y la recuperación de los datos.

Lectura recomendada

El capítulo español de la Cloud Security Alliance publicó el informe “Cloud Compliance Report” en el que se tratan con una orientación práctica cuestiones como la protección de datos o los contratos relacionados con *el cloud*. El informe se puede descargar en:

http://cloudsecurityalliance.org/csa-news/csa-issues-first-cloud-compliance-report-for-spain/ (documento en inglés)

Para prevenirse contra estos problemas e, incluso, ir superando los propios miedos, no siempre fundamentados (como pensar que los datos están más seguros en un servidor en la propia empresa que en el servidor del proveedor), muchas empresas están optando por la nube híbrida, que combina el *cloud* (entendido como *cloud* público) y el *cloud* privado para la distribución de los recursos de una empresa.

Se entiende por ***cloud* privado** el configurado cuando se habilitan los propios CPD de la empresa con las últimas técnicas de gestión que permiten hacer un uso eficiente de sus recursos para ofrecer servicios a la propia empresa.

A pesar de que no se dispone de una escalabilidad casi infinita, de este modo se cuenta con altos niveles de seguridad y confidencialidad, lo que asegura el cumplimiento de la propia legislación (como puede ser la ley LOPD española).

Una solución óptima para no tener que sobredimensionar las máquinas (y recursos en general) ni provocar colapsos en el servicio es llevar al *cloud* únicamente los picos de actividad. Dejamos aquí este tema por motivos de espacio y de enfoque de este material, aunque es sin duda una materia muy extensa y de gran importancia para la implantación del *cloud*.

1.6. Un futuro móvil donde todo el mundo y todo estará conectado

El fenómeno *cloud* coincide en el tiempo con la masiva difusión de los móviles.

El ***cloud computing*** hace posible un futuro con una extrema movilidad en el que todo el mundo esté conectado.

Sin *cloud* no sería posible, por ejemplo, almacenar y manejar los datos de los más de 900 millones de usuarios de Facebook, de los que una parte importante se conectan a través de sus móviles.

Se estima que en el 2012 serán más de 15 millones los usuarios de redes sociales en España. A pesar de los problemas de seguridad y privacidad que conllevan, está clara la tendencia que marcan, y por eso las empresas ya consideran las redes sociales un medio eficaz para la comercialización y la comunicación de sus productos. No cabe ninguna duda de que a partir de ahora las empresas solo pueden aumentar su posicionamiento y visibilidad en ellas. Y aunque, debido a la diversidad de redes sociales, no sea fácil gestionar esa presencia, seguramente se impondrán nuevos sistemas y herramientas integradas que permitirán gestionar de manera más fácil la administración de los datos y procesos de una empresa en todo el amplio entramado de las redes sociales.

Téngase en cuenta que ya desde finales del 2010 se fabrican más dispositivos móviles que PC. El centro de gravedad se está desplazando rápidamente del PC a otros tipos de tecnologías móviles. Estos dispositivos, conectados a la nube, crearán una nueva relación entre los ciudadanos y sus dispositivos. Para mejorar la calidad de la interacción con el público se podrá utilizar información sobre el contexto y las preferencias del usuario, como servicios de localización, realidad aumentada en dispositivos móviles o comercio móvil. Las

empresas podrán anticiparse a las necesidades del usuario y, de forma proactiva, servir productos o servicios más apropiados y personalizados (por ejemplo, el marketing y la comercialización basados en la geolocalización para encontrar ofertas en tiempo real o apoyados en códigos promocionales, ofertas o descuentos, etc.).

Todo ello está comportando una mutación de los contenidos para llegar al público. Por ejemplo, a muchos usuarios les resultan más fáciles de entender las aplicaciones de los móviles que cualquier web, ya que aquellas les permiten acceder a una determinada funcionalidad de forma cómoda, rápida y fácil, una vez instalado el dispositivo. Por eso, el imparable crecimiento de estas *apps* representa un retroceso para el uso de la web desde los móviles, que como hemos visto son cada vez más numerosos. En consecuencia, el sitio web tradicional irá dejando de ser el núcleo de la estrategia de visibilidad y el eje de negocio de las empresas. El futuro pasa por apostar por el contenido digital con otros métodos y dispositivos. Además, más allá de la modernización de la web con nuevos estándares, como el HTML5 o el CSS3, la misma información que genera el usuario está adoptando otros formatos más dinámicos. Por ejemplo, cada minuto se generan 50 horas de contenido en YouTube, una muestra del crecimiento del tráfico *de streaming* frente a otros tipos de contenidos.

Pero todavía hay más, en un futuro no muy lejano incluso los móviles dejarán de ser el medio habitual de consumir la información contenida en el *cloud computing,* ya que serán sustituidos por otros dispositivos, como por ejemplo las gafas de realidad aumentada con conexión a Internet que Google está diseñando, según anunció la misma empresa hace unos meses. Unas gafas que a partir de la voz del usuario permiten hacer fotos, iniciar videochats, mostrar direcciones, pedir previsiones meteorológicas, estar conectado a la red social de la que se es usuario, entre otras decenas de opciones.

Dentro de poco, la nube podrá poner en contacto, no solo a los millones de personas que la utilizan a través de su ordenador o dispositivo móvil, sino también los miles de millones de objetos que todavía no están en ella actualmente. Cualquier objeto se podrá transformar en un elemento con capacidad de comunicarse en cualquier momento y en cualquier lugar. Es lo que se denomina ***Internet of things*** (Internet de las cosas), que implica que todo objeto pueda ser una fuente de datos y que cualquier cosa pueda ser monitorizada a través del tiempo y el espacio. Se trata de chips de bajo coste incrustados en casi cualquier objeto de consumo (artículos domésticos como la nevera o incluso los mismos coches). Todo ello generará nuevas áreas de negocio, entre las que se puede citar la de las *smart cities*. Pero sobre todo representará que los productos se conviertan en servicios. ¿Os imagináis que un fabricante de motores de aviación ponga sensores en sus equipos para poder alquilarlos por horas de vuelo, en lugar de venderlos? Todo un cambio de panorama para los servicios que empresas como Rolls Royce, fabricante de motores de aviación, ya están llevando a cabo.

Lectura complementaria

La Fundación de la Innovación Bankinter ha publicado el informe "El Internet de las cosas" donde se describe el estado del arte de esta tecnología emergente. El documento se puede descargar en la dirección siguiente:

http://www.fundacionbankinter.org/system/documents/8168/original/XV_FTF_El_internet_de_las_cosas.pdf (documento en castellano)

1.7. El *cloud computing* como motor de cambio

El *cloud* está aquí y ha venido para quedarse. Lo que veníamos intuyendo en los últimos años se ve confirmado con la aparición diaria de nuevos servicios en la nube o los datos de estudios, como los de CISCO, que nos dicen que dentro de un par de años todo el tráfico de la red tendrá su origen en el *cloud computing*.

Como fórmula eficiente para el suministro de servicios, el modelo *cloud computing* (especialmente su piedra angular, la infraestructura como servicio) ofrece un gran ahorro de costes, reduce drásticamente el espacio requerido en los CPD y ofrece una mayor simplicidad de las operaciones TIC. Lo que ha permitido el *cloud computing* ha sido básicamente la industrialización de las TIC, en un proceso similar al que se vivió con la generación eléctrica. A principios del siglo pasado, las empresas generaban su propia energía para autoabastecerse, hasta que surgieron las compañías eléctricas y las economías de escala del nuevo modelo de suministro superaron las ventajas de la producción propia.

A pesar de que los temas de seguridad o privacidad, entre otros muchos, todavía tienen que mejorar en *el cloud computing*, no cabe duda de que este abre un nuevo modelo de TIC flexibles que se adaptan a cada empresa y, algo muy importante, cuyos costes son predecibles. Entre las ventajas que ofrece, destacan la posibilidad de acceso desde cualquier lugar, la fórmula de pago por uso, el autoservicio a la carta o la rápida elasticidad.

Ahora bien, si una empresa se limita a adoptar el modelo *cloud computing* solo para ahorrar dinero, es muy probable que se equivoque de estrategia. La ventaja diferencial del *cloud computing* es que permite hacer realidad nuevos modelos de negocio, probando procesos de negocio y efectividad operacional en el momento en que se necesiten y con la potencia que requieran los proyectos. Pensemos, por ejemplo, que vale lo mismo alquilar una máquina 1.000 horas que 1.000 máquinas una hora. Esto abre un gran abanico de posibilidades a las empresas para probar y ensayar conceptos o procesos. La misma oportunidad se extiende a los emprendedores y las empresas más pequeñas, liberadas de la necesidad de realizar grandes inversiones de capital gracias al modelo de pago por uso descrito.

El *cloud computing* ha comportado además que los teléfonos inteligentes y otros dispositivos de consumo masivo hayan desplazado al PC del centro del mundo digital. El *cloud computing* está dando paso a una nueva era que ofrece a los usuarios el nivel de flexibilidad necesario con los dispositivos para la realización de las actividades diarias. Asimismo, las nuevas maneras de interactuar, como las que ofrece el reconocimiento táctil y gestual, junto con la identificación del habla y el conocimiento contextual, permiten una rica interacción entre el usuario y los dispositivos.

Todo ello acompañado de lo que se conoce como *Internet of things*, por obra de muchas empresas que ya lo están trabajando. Y es que casi cualquier cosa puede convertirse en digital, interconectando e integrando sus datos con todos los otros que capten los demás dispositivos, pudiendo, por lo tanto, hacer un tratamiento en tiempo real. No cabe duda de que, a medida que haya cada vez más objetos que tengan sensores incrustados y más posibilidades de comunicar, se generará una red de conexiones que logrará que haya ahora más información disponible que en toda la historia de la humanidad. El uso inteligente de esta información tendrá la capacidad de generar una transformación de la sociedad nunca vista hasta ahora, que, en el plano de los negocios, supondrá la creación de nuevos modelos, así como la mejora de los existentes, con una reducción de los costes y los riesgos de todos ellos.

Lectura complementaria

El libro que de forma divulgativa explica todo el fenómeno del *cloud computing* es: Jordi Torres i Viñals (2011). *Empresas en la nube, ventajas y retos del Cloud Computing*. Barcelona: Libros de Cabecera. Disponible en: http://www.librosdecabecera.com/empresas-en-la-nube

2. *Big Data*

2.1. Se acerca una marea de información digital

Ya hemos explicado que el *Internet of things* generará una ingente cantidad de datos desde cualquier tipo de dispositivo o sensor.

Para hacernos una idea de lo fácil que es producir datos, fijémonos en las redes sociales, por ejemplo Twitter, que gestiona más de 90 millones de tuits al día, lo que representa un total de 8 terabytes de datos cada día. O los datos transaccionales de las aplicaciones web, que han crecido a una gran velocidad. Hoy, Wal-Mart, la cadena minorista más grande del mundo, gestiona un millón de transacciones de los clientes por hora que alimentan una base de datos estimada en 2.5 petabytes. Y no olvidemos el mundo científico, del que es un buen ejemplo el colisionador de partículas del CERN, que puede llegar a generar 40 terabytes de datos por segundo durante los experimentos.

Boeing

Por ejemplo, los motores a reacción de Boeing que hemos comentado antes pueden producir 10 terabytes de información cada 30 minutos. Es decir, un Jumbo de 4 motores puede crear 640 terabytes de datos en un vuelo transoceánico.

Nota

1 terabyte (TB) = 1.000 gigabytes (GB)

1 petabyte (PB) = 1.000.000 gigabytes (GB)

1 exabyte (EB) = 1.000.000.000 gigabytes (GB)

1 zettabyte (ZB) = 1.000.000.000.000 gigabytes (GB)

Todo ello comporta un crecimiento exponencial de la información disponible y nos lleva a la siguiente gran tendencia que dominará la industria de las TIC en los próximos años, conocida como ***big data***. Según algunos estudios, en el 2015 circularán por el planeta 7.910 exabytes de información (en el 2005 eran 130 exabytes). En definitiva, es muy fácil generar terabytes de datos. Otra cosa es procesarlos y transformarlos en información de valor. El problema ya no radica en encontrar datos, sino en saber qué hacer con ellos. Por eso adelantamos a continuación los retos y oportunidades que el *big data* presenta.

Lectura complementaria

Un buen punto de partida para conocer el alcance del fenómeno *big data* lo podemos encontrar en el informe de McKinsey titulado "Big Data: The next frontier for innovation, competition, and productivity" que podéis descargar en la dirección siguiente:

http://www.mckinsey.com/insights/mgi/research/technology_and_Innovation/Big_data_The_next_frontier_for_innovation (documento en inglés)

2.2. Definiendo el *big data*

Igual que ocurría con el *cloud computing*, no hay una única definición de consenso para este nuevo fenómeno que denominamos *big data*.

En este material consideraremos como ***big data*** todo lo referente al hecho de que los datos se han vuelto tan grandes que no se pueden procesar, almacenar y analizar mediante métodos convencionales.

Una manera de caracterizar estos datos que se usan es recurriendo a lo que dicen las 3 *V* en referencia a volumen, variedad y velocidad:

- **Volumen**: el universo digital sigue expandiendo sus fronteras y se estima que ya hemos superado la barrera del zettabyte.

- **Velocidad**: la velocidad a la que generamos datos es muy elevada, y la proliferación de sensores es un buen ejemplo de ello. Además, los datos en tráfico –datos de vida efímera, pero con un alto valor para el negocio– crecen más deprisa que el resto del universo digital.

- **Variedad**: los datos no solo crecen sino que también cambian su patrón de crecimiento, a la vez que aumenta el contenido desestructurado.

A veces también se añade otra *V*, la de valor. Extraer valor de toda esta información marcará la próxima década. El valor lo podremos encontrar en diferentes formas: mejoras en el rendimiento del negocio, nuevas fuentes de segmentación de clientes, automatización de decisiones tácticas, etc.

Como ya hemos visto, el origen de los datos para una empresa puede ser diverso. Por ejemplo, le pueden llegar de sus propios sistemas de información de apoyo a las ventas o de interacción con sus clientes, así como estar generados por las máquinas o sensores incrustados en cualquier tipo de dispositivo o producto de la empresa. Y no olvidemos la información que circula por las redes sociales sobre una determinada empresa, que sin duda es muy valiosa para esta.

Pero hay otro origen muy importante de los datos, representado por las plataformas de información que varios gobiernos están abriendo. Estos datos públicos pueden ser informes, mapas, estadísticas, estudios, análisis, creados y gestionados por la administración en todos los ámbitos (sanidad, economía, educación, población, etc.), que son de gran interés público.

La filosofía de los ***open data*** (datos abiertos) ha dado lugar al movimiento Open Data, que persigue que las instituciones públicas expongan los datos públicos que están en su poder de forma reutilizable para que terceros puedan crear servicios derivados de dichos datos. Como consecuencia, los conjuntos de datos expuestos se ofrecen bajo licencias de propiedad abiertas, que permiten su redistribución, reutilización y aprovechamiento con fines comerciales.

El movimiento *open data* representa también el deseo de que las empresas liberen datos para permitir el desarrollo de nuevas aplicaciones y usos. Pero el papel protagonista en dicho movimiento es el de ciudadano. El usuario conectado a la red es una gran fuente de información. Al compartir lo que vemos (información de tráfico, opiniones, accidentes, etc.), el lugar donde estamos (geolocalización) o las imágenes que captamos, estamos construyendo, entre todos, una grande inteligencia colectiva.

Referencia web

Hay proyectos de *open data* que se están llevando a cabo en Euskadi, Cataluña, Asturias o Navarra, y desde ayuntamientos, como los de Barcelona, Zaragoza, Lleida, Badalona, Gijón o Córdoba. Recomiendo visitar:

http://datos.fundacionctic.org/sandbox/catalog/faceted

2.3. Hacen falta nuevas tecnologías de almacenamiento

El escenario que hemos tenido hasta ahora para el almacenamiento y acceso a datos de las aplicaciones es el de servidores que usan memoria RAM[7] para la información que necesitan a menudo y utilizan el almacenamiento en disco para el resto de los datos que la aplicación no está usando en ese instante. El hecho de no tenerlo todo en memoria se debe a que esta es cara, por lo que habitualmente hay poca. En cambio, las unidades de disco duro (HDD[8]) presentan una alta capacidad mucho más asequible, a pesar de que tienen un rendimiento mucho más bajo que la memoria. Por eso, habitualmente en un servidor solo hay memoria para contener la fracción de datos que el procesador necesita para trabajar. Además de ser cara, la memoria también es volátil, lo que significa que, si un servidor cae o se queda sin electricidad, la información almacenada en la memoria se pierde. Debido a este hecho, toda la información se mantiene en disco durante su vida útil, puesto que así se conserva aunque apaguemos el ordenador o lo desenchufemos de la corriente.

(7)Acrónimo en inglés de *random access memory*.

(8)Acrónimo en inglés de *hard disk drives*.

En estos momentos, el almacenamiento de información en HDD es cien veces más barato que en la memoria RAM, pero de acceso mil veces más lento. Para solucionarlo se están empezando a utilizar como sistemas de almacenamiento tecnologías de memoria no volátil, como las usadas hasta ahora en dispositivos móviles, que se denominan discos de estado sólido (SSD[9]).

(9)Acrónimo en inglés de *solid-state drive*.

SSD

Los **SSD** ya se encuentran en el mercado y son una opción para almacenar datos que requieren un tiempo de acceso muy rápido. Un buen punto de partida para conocer más detalles sobre ellos puede ser la Wikipedia:

http://en.wikipedia.org/wiki/solid-state_drive (documento en inglés)

Memoria no volátil

Un ejemplo son los chips de memoria flash, que, del mismo modo que los HDD, son no volátiles y mantienen la información cuando se quedan sin alimentación eléctrica.

Los **SSD** parecen una opción con futuro si tenemos en cuenta que el precio de la memoria flash está bajando y que, además, al ser básicamente un chip de semiconductores, tiene un consumo energético muy inferior al de los discos HDD, que usan componentes mecánicos para leer los discos. A pesar de ello, para ciertas aplicaciones que requieran reescribir muchas veces un mismo dato, hay que tener en cuenta que estos dispositivos todavía envejecen más deprisa que los HDD, y por lo tanto no son hoy por hoy una solución para cualquier tipo de aplicación.

La aparición de estas soluciones a la persistencia del almacenamiento de los datos por medio de tecnologías de memoria no volátil conducirá en breve a muchas mejoras de almacenamiento. Como memorias que son, se está inves-

tigando para conectarlas al bus interno de los ordenadores que conecta con el procesador. De este modo se conseguirá que el procesador pueda acceder a los datos persistentes a través de las instrucciones de memoria *load/store*, lo que permitirá un acceso simple y rápido que reducirá mucho más todavía el tiempo de acceso al almacenamiento persistente. Es la investigación conocida como *storage class memory*, que trata de diseñar un sistema de almacenamiento de capacidad y economía similar al de los HDD, pero con unas prestaciones de rendimiento equivalentes a las de la memoria. Actualmente, es una de las áreas de investigación más activas que sin duda tendrá un impacto muy importante en el mundo del *big data*, lo que permitirá acelerar la ejecución de las aplicaciones intensivas en datos.

2.4. Las bases de datos relacionales no sirven para todo

Hasta ahora, el escenario de las bases de datos ha estado marcado por lo que se conoce como *SQL*[10] (lenguaje de consulta estructurado). En el mercado encontramos muchas opciones consolidadas, como DB2, Oracle, MySQL, Informix, Microsoft SQL Server, PostgreSQL, etc. Estas bases de datos, que se suelen denominar *relacionales*, siguen las reglas ACID, hecho que les permite garantizar que un dato sea almacenado de manera inequívoca y con una relación definida sobre una estructura basada en tablas que contienen filas y columnas.

(10)Acrónimo en inglés de *structured query language*.

Reglas ACID

En el contexto de bases de datos, **ACID** (acrónimo inglés de *atomicity, consistency, isolation, durability*) son toda una serie de propiedades que tiene que cumplir todo sistema de gestión de bases de datos para garantizar que las transacciones sean fiables. Para entrar en algo más de detalle podéis consultar la Wikipedia:
http://es.wikipedia.org/wiki/ACID

Pero en el mundo del *big data* aparece el problema de que las bases de datos relacionales no pueden manejar el tamaño, ni la complejidad de los formatos, ni la velocidad de entrega de los datos que requieren algunas de las aplicaciones de hoy en día, por ejemplo, aplicaciones en línea con miles de usuarios concurrentes y millones de consultas al día.

En la figura 5 se presenta esquemáticamente el comportamiento de los sistemas de almacenamiento actuales. Podríamos resumir que tienen dos grandes limitaciones:

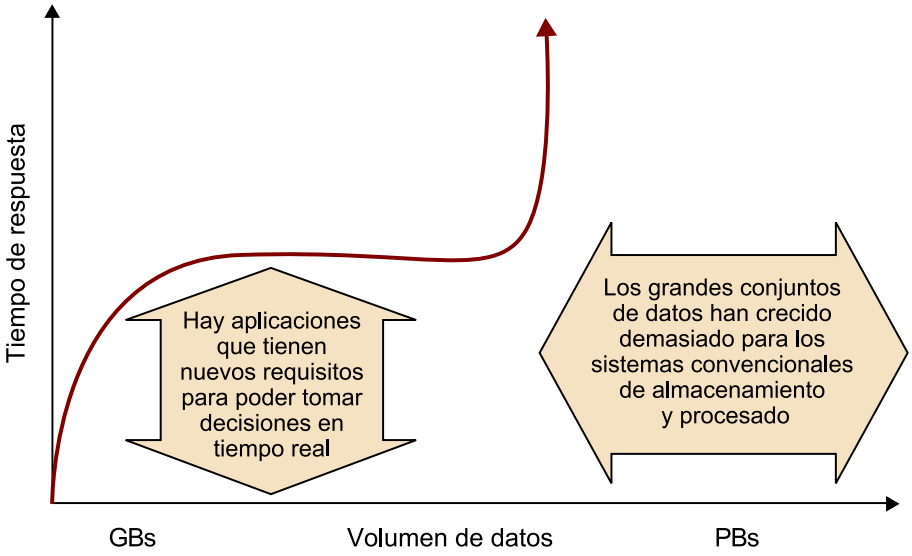
1) Por un lado, no pueden atender los requerimientos, como el de requerir un tiempo de respuesta muy bajo, que actualmente presentan algunas aplicaciones. En estos casos, hoy por hoy se están poniendo los datos en memoria (lo que se conoce por *in-memory*). Este tipo de aplicaciones se verán beneficiadas por las mejoras en almacenamiento que se prevén, comentadas en el apartado anterior.

2) Por otro lado, aparece el problema de que determinadas aplicaciones requieren un volumen de datos que ya no pueden ser almacenados y procesados usando bases de datos tradicionales.

Por eso han surgido nuevas variedades de bases de datos (llamadas NoSQL) que permiten resolver los problemas de escalabilidad y rendimiento que el *big data* presenta. NoSQL aglutina las diferentes soluciones de bases de datos centradas al ser no relacionales, distribuidas y escalables de forma horizontal. Hay nu-

merosos productos disponibles, muchos de ellos *open source*, como Cassandra, que pertenece al proyecto *open source* Hadoop. Este es un campo que durante el 2012 será muy activo, puesto que las principales empresas del sector ya han entrado en este nuevo mercado.

Figura 5. Comportamiento de los sistemas tradicionales de almacenamiento de la información.



Las bases de datos NoSQL no suponen que no se haya de usar el SQL, sino simplemente que hay soluciones mejores para determinados problemas y aplicaciones. Por eso, NoSQL también lo podemos leer como *not only SQL*. La idea es que, para determinados escenarios, hay que relajar muchas de las limitaciones inherentes a las bases de datos convencionales y la forma de acceder a ellas. Por ejemplo, podemos pensar en colecciones de documentos con campos definidos de manera laxa, que incluso pueden ir cambiando en el tiempo, en lugar de tablas con filas y columnas prefijadas. En cierto modo, incluso podríamos llegar a pensar que un sistema de este tipo no es ni siquiera una base de datos entendida como tal, sino un sistema de almacenamiento distribuido para gestionar datos dotados de una cierta estructura que puede ser enormemente flexible.

Referencia web

Si se tiene interés en conocer la larga lista de bases de datos NoSQL y averiguar más detalles, la página web siguiente es un buen punto de partida:
http://nosql-database.org/
(página web en inglés)

2.5. Se requieren nuevos modelos de programación

No cabe duda de que los datos aumentarán todavía más y muchas organizaciones querrán, no solo recolectarlos y almacenarlos, sino también utilizarlos, puesto que los datos solo les son útiles a las organizaciones si pueden hacer algo con ellos. Pero procesar estas grandes cantidades de datos almacenados distribuidamente de manera poco estructurada es un gran reto, y tampoco en este caso sirven los métodos tradicionales de procesado.

Para conseguir el objetivo de procesar grandes conjuntos de datos, Google creó **MapReduce**, motor que está actualmente detrás de los procesamientos de datos de Google. Pero ha sido el desarrollo **Hadoop MapReduce**, por parte

de Yahoo, lo que ha propiciado un ecosistema de herramientas *open source* de procesamiento de grandes volúmenes de datos que la mayoría de implementaciones de otras empresas están utilizando.

La innovación clave de MapReduce es la capacidad de hacer una consulta, dividiéndola y ejecutándola en paralelo a la vez, a través de muchos servidores sobre un conjunto de datos inmenso.

De este modo se resuelve el problema de los datos cuando son demasiado grandes para que quepan en una sola máquina. Este modelo tiene dos fases:

1) Fase Map, en la que los datos de entrada son procesados, uno a uno, y transformados en un conjunto intermedio de datos.

2) Fase Reduce, donde estos resultados intermedios se reducen a un conjunto de datos resumidos, que es el resultado final deseado. Hoy por hoy es un proceso tipo *batch,* que puede requerir de minutos u horas para completarlo.

El problema de MapReduce y el almacenamiento con sistemas NoSQL es que no sigue la manera de procesar y almacenar que desde hace años estamos empleando y enseñando en las universidades, y por eso se hace difícil "pensar" en este nuevo paradigma, que en cierto sentido, para entendernos, se tiene que "desaprender". Ante esto han surgido proyectos como **Hive**. De manera sencilla, podríamos decir que es un sistema *datawarehouse* basado en Hadoop que fue desarrollado por Facebook y ahora es un proyecto *open source* dentro del ecosistema de Hadoop. El interés principal de Hive es que ofrece la posibilidad de que los usuarios escriban consultas en SQL, que después se convierten en MapReduce de manera transparente para el programador. Ello permite a los programadores de SQL que no tienen experiencia en MapReduce usarlo e integrarlo en sus entornos habituales. Por otro lado, **Pig Latin** es un paquete también en el ecosistema Hadoop desarrollado originalmente por Yahoo, que ofrece un lenguaje de alto nivel para describir y ejecutar trabajos de MapReduce. Su objetivo es hacer también accesible Hadoop a los desarrolladores familiarizados con la manipulación de datos con SQL, para lo que les proporciona dos interfaces, una interactiva y otra para poder usarla desde Java.

Referencia web

Un buen punto de partida para entender más en detalle Hadoop y todo su entorno es su página en la Wikipedia, desde donde se puede ir profundizando a partir de sus enlaces:

http://es.wikipedia.org/wiki/hadoop (página web en castellano)

2.6. La información no es conocimiento

Parece que fue Albert Einstein quien dijo que "la información no es conocimiento". Cuánta razón tenía, los datos necesitan ser analizados para que se les pueda extraer el valor que contienen.

Es fundamental la habilidad para transformar los datos en información, y la información en conocimiento, de modo que con dicho conocimiento se puedan optimizar los procesos en los negocios.

Pensemos que esta **inteligencia de negocio** actúa como un factor estratégico para una empresa u organización, a la que le genera una potencial ventaja competitiva, que no consiste sino en proporcionar información privilegiada para responder a los problemas de negocio.

Pongamos por caso una de estas tecnologías, la del Machine Learning, que está detrás de muchos de estos análisis que comentábamos. Ciertamente ya están en un momento de mucha madurez. Por ejemplo, en febrero del año 2012, la computadora Whatson de IBM jugó durante tres noches consecutivas contra dos de los más brillantes campeones de Jeopardy[11]. La computadora era capaz de responder en menos de tres segundos ejecutando simultáneamente miles de tareas complejas que usaban Machine Learning. Ahora bien, Machine Learning es todavía un gran reto en *big data*, ya que la mayoría de los algoritmos simples se ejecutan bien en miles de registros, pero hoy por hoy son impracticables en miles de millones.

[11]Concurso cultural de la televisión estadounidense en el que los participantes deben responder a una pregunta con una serie de pistas.

Por eso, los métodos de análisis existentes hasta ahora todavía se tienen que adaptar, en general, al tamaño y la complejidad de los formatos del *big data*, lo mismo que las bases de datos que, al ser mayoritariamente relacionales, no permiten almacenar estas ingentes cantidades de datos, tal como hemos visto. Sin ningún tipo de duda, este será uno de los campos que más avanzará durante los próximos años, dada la importancia que tendrá para las empresas.

En esta línea, en el mundo *open source*, y dentro del ecosistema de Hadoop que antes hemos comentado, podemos encontrar iniciativas como **Mahout**. Su objetivo es producir una implementación libre de un paquete que incluya los principales algoritmos de Machine Learning escalables sobre la plataforma de Hadoop. El proyecto está muy activo, pero todavía quedan muchos algoritmos pendientes de incorporar. A pesar de ello, probablemente no es un reflejo de los avances que se están produciendo en este campo, puesto que seguramente se están desarrollando muchas investigaciones de manera silenciosa dentro de las grandes corporaciones, dada la importancia que pueden tener para sus negocios. Bien es verdad que esta es una opinión personal, pero estoy convencido de que es lo que está ocurriendo. Pongamos por caso a Google, que dispone de su sistema de almacenamiento GFS y MapReduce para el desarrollo de sus sistemas de Machine Learning a gran escala.

De Google, por ejemplo, sabemos que sus investigadores de Machine Learning desarrollan algoritmos de computación que mejoran el rendimiento de manera automática a través de la experiencia. Podemos encontrar que se aplica para resolver muchos problemas, como los de recomendar etiquetas para vídeos de YouTube o lograr la identificación de diferentes aspectos en las opiniones en línea de los usuarios.

Lectura complementaria

En el libro *Understanding Big-Data*, publicado por IBM, se describe el *big data* desde el punto de vista de negocio. El libro se puede descargar en la dirección siguiente:

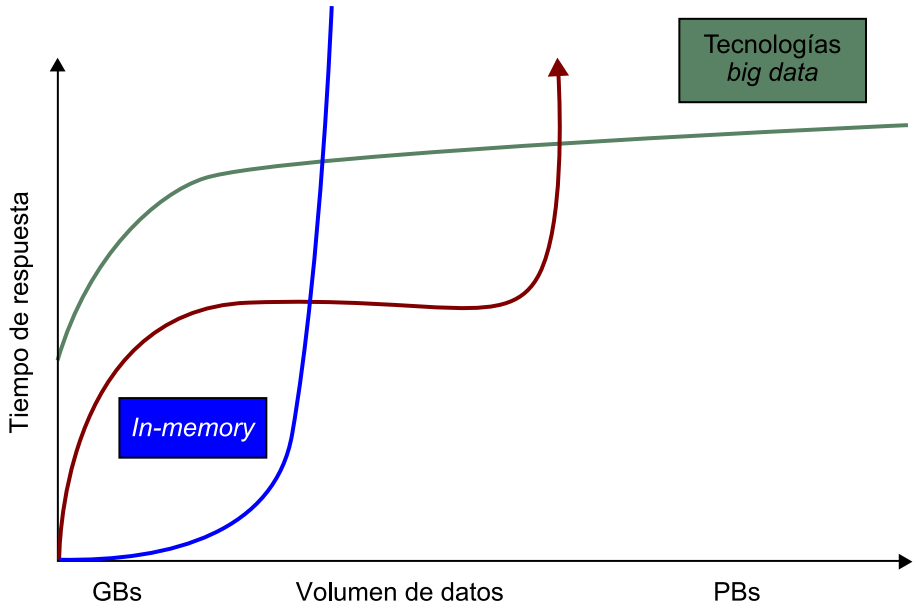
http://www-01.ibm.com/software/data/bigdata/ (libro en inglés)

2.7. Entrando en la nueva era del Data 2.0

El tamaño, la complejidad de los formatos y la velocidad de entrega superan las capacidades de las tecnologías de gestión de datos tradicionales. Por eso se requiere el uso de nuevas tecnologías que permitan la gestión de grandes volúmenes de datos. Surgirán nuevas tecnologías capaces de ejercer un gran impacto potencial, como por ejemplo las de gestores de bases de datos *in-memory* para atender los nuevos requerimientos de datos.

En todo caso, el futuro seguramente será híbrido, pues no contará con solo una tecnología de almacenamiento, sino que será una mezcla de todas ellas. Se ha de tener en cuenta que no todas las aplicaciones tienen su cuello de botella solo en los datos. Por otro lado, por más que se abaraten los costes de almacenamiento en memoria, no siempre resultará rentable el almacenamiento de unos determinados datos a unos costes más elevados. Por eso, uno de los temas de investigación importantes es considerar de manera global la gestión de recursos de almacenamiento en los CPD, considerando los requerimientos de las aplicaciones, a la vez que decidir dónde se ejecutan y dónde están sus datos. En la figura 6 se representa esquemáticamente el comportamiento de las nuevas propuestas frente a los sistemas tradicionales de almacenamiento de la información que hemos presentado antes.

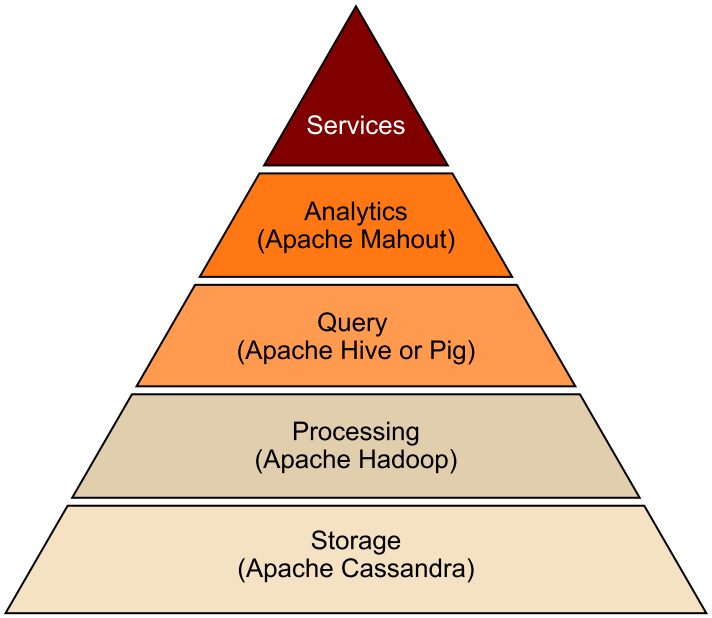
Figura 6. Comportamiento de las nuevas tecnologías para superar las limitaciones de la gestión de datos tradicional.



Un hecho muy importante es que estas nuevas tecnologías están al alcance de todo el mundo, no solo de las corporaciones que pueden hacer grandes inversiones. Del mismo modo que cuando maduró la web llegó la era en que

fue denominada Web 2.0, gracias a paquetes de software *open source* como los llamados LAMP (Linux, Apache, MySQL y PHP), ahora estamos entrando en una nueva era para los datos, la **Data2.0**, gracias a una nueva generación de tecnologías y arquitecturas, diseñadas para extraer valor a un gran número de datos de diversos tipos y en algunos casos en tiempo real. En la figura 7 se presentan algunas de las posibles tecnologías que están al alcance de todo el mundo para poder entrar en este mundo del *big data*.

Figura 7. Stack Data 2.0 de herramientas *big data open source*.



Para acabar, apuntar que las aplicaciones analíticas que se usan en el proceso de toma de decisiones en las empresas empezarán a recurrir al *cloud computing* usando como base las herramientas que hemos comentado en el apartado "*Cloud computing*". Para varios sectores (*smart cities*, *retail*, financiero, salud, etc.) aparecerán soluciones verticales, es decir, que integrarán todos los niveles para procesos específicos que permitirán lograr la diferenciación y la creación de ventajas competitivas.

Lectura complementaria

La organización O'Really ha publicado el libro *Biga Data Now, Current perspectives from O´Reilly Media*, que se puede descargar en la página web siguiente:

http://
shop.oreilly.com/
product/0636920022640.do
(libro en inglés)

Resumen

Conclusión: un mundo de oportunidades para todos vosotros

Todos los estudios apuntan a que el *cloud computing* transformará el mundo laboral, generará nuevos puestos de trabajo y, obviamente, hará obsoletos otros. Por eso no me cabe ninguna duda de que adquirir habilidades en el mundo del *cloud computing* es una apuesta segura para todos vosotros. Y lo mismo ocurre con el *big data*, ya que las compañías que quieran dar sentido a toda la información de que disponen habrán de contratar a personas con estos conocimientos.

Pero también podéis pensar en emprender con vuestra propia iniciativa. En este caso, recordad que el *cloud computing* es vuestro gran aliado. La informática, tanto en forma de un paquete de software como de un servidor hardware, ya se puede consumir como un servicio más, lo mismo que la electricidad o el agua. Solo hay que conectarse a la red de Internet y, de modo parecido a lo que ocurre con la electricidad, el consumidor paga según lo que consume, lo cual permite ajustar el consumo a las necesidades de cada momento y evita grandes inversiones iniciales en TIC. Además, recordad también que este nuevo paradigma ofrece una gran agilidad, lo que es muy importante a la hora de obtener nuevos recursos hardware o software, habida cuenta del tiempo que se requiere para crear una instalación informática propia, sobre todo si se piensa en la presión que les supone a los emprendedores tener que disponer de recursos con la rapidez y en la cantidad necesarias para poder innovar.

En cuanto al *big data*, ya hemos visto que hay a vuestra disposición todo un conjunto de paquetes *open source* que os pueden permitir grandes aventuras. Como ya se ha comentado, el conjunto de paquetes de software Linux, Apache, MySQL y PHP (LAMP) dio poder a los desarrolladores innovadores permitiéndolos programar potentes servidores web a partir de códigos *open source*. Todo ello desembocó en la creación de muchas nuevas empresas en el sector de las TIC y fue la base de lo que se conoce por Web 2.0. Estoy seguro de que a partir de ahora veremos nuevas empresas basadas en la pila de software Data 2.0 que hemos presentado anteriormente.

Recordad además que, para obtener datos a partir de los cuales generar valor, contáis con el movimiento Open Data que hemos presentado. Es una gran oportunidad, y querría hacer hincapié aquí en el hecho de que este movimiento brinda a los emprendedores la posibilidad de generar nuevas aplicaciones a partir de los datos que nos ofrece el *open data*. Y no olvidemos el mundo de los móviles y el *Internet of things*. Existe todo un filón en el desarrollo de

software más allá de las simples *apps*, que recolectan, almacenan, organizan, analizan e integran la información con otros programas o personas a través de sus dispositivos.

En resumen, mi visión es que el *cloud computing* y las nuevas tecnologías que lo acompañan ya están aquí y no hay marcha atrás. Estoy convencido de que los inhibidores del *cloud computing* que hemos comentado antes se irán solucionando. Para expresar el poder transformador que creo que tendrán el *cloud computing* y las tecnologías que lo acompañan, me permito usar la analogía con la aparición del automóvil a principios del siglo XX descrita por Rolf Harms y Michael Yamartino, de Microsoft, en su artículo "The Economics of Cloud":

> "Cuando los coches aparecieron a principios del siglo XX, se los llamó inicialmente "carruajes sin caballos". Es comprensible que las personas se mostraran escépticas entonces y vieran el invento a través de la lente del paradigma que había prevalecido durante siglos: el caballo y el carro. Por eso, los primeros coches buscaban una imagen muy similar a la del caballo y el carro, ya que ni siquiera los propios ingenieros llegaban a comprender las posibilidades del nuevo paradigma, como mayor velocidad o mayor seguridad. Increíblemente, los ingenieros mantuvieron la reproducción de una cabeza de caballo en la parte delantera de los primeros modelos de coche antes de darse cuenta de que eso no tenía ningún sentido. Inicialmente hubo una total incomprensión del nuevo paradigma, se afirmaba que el caballo está aquí para quedarse y que el automóvil es solo una novedad, una moda pasajera. Incluso los mismos pioneros del coche no acababan de comprender el impacto potencial que su trabajo podría tener en el mundo. Cuando Daimler, posiblemente el inventor del automóvil, intentó calcular el mercado potencial del automóvil a largo plazo, llegó a la conclusión de que nunca podría haber más de 1 millón de coches, debido a su alto coste y a la escasez de chóferes capacitados."
>
> Rolf Harms; Michael Yamartino (2010). "The Economics of Cloud". Microsoft. Disponible en: http://www.microsoft.com/presspass/presskits/cloud/docs/the-economics-of-the-cloud.pdf

Os animo a que avancéis y aprovechéis estas apasionantes oportunidades que tenéis delante, aunque no sepáis con seguridad qué os espera en cada paso, debido a la rapidez con que evoluciona la tecnología. Sobre todo confiad en vosotros mismos y perded el miedo a equivocaros, puesto que equivocarse es inseparable de actuar.

Espero que este material os haya ayudado al descubrimiento de estas nuevas tecnologías y que sobre todo os impulse a adentraros algo más en este apasionante mundo para comprender la transformación que está sufriendo la manera en que interactuamos y, particularmente, en que hacemos negocios. Y ¿por qué no?, estimularos a emprender aprovechando todas estas herramientas que están a vuestro alcance. ¡Suerte!

Referencia web

Como divulgador, mantengo un blog donde voy posteando los avances en estas tecnologías. Si creéis que os puede ser útil, podéis encontrarme en:

www.jorditorres.org/blog