Cloud Computing y *Big Data*, la próxima frontera de la innovación

Por Jordi Torres

UPC Barcelona Tech. Barcelona Supercomputing Center

Mateo Valero suele exponer una presentación en la que cuenta brevemente cómo comenzó la Ciencia. Empezó en el momento en que la Matemática, la teoría, permitió describir la experiencia. Éste fue un paso fundamental, pero ¿cuál es el siguiente paso fundamental? El siguiente paso es, o ha sido hasta ahora, la simulación. La simulación hecha por la supercomputación nos permite simular y crear escenarios que sin la supercomputación serían imposibles. Escenarios caros, peligrosos e imposibles. Primero fue la teoría, luego fue la simulación que nos ha permitido llegar hasta aquí y que se basa en muchas fórmulas, mucha matemática, y muchos cálculos. ¿Dónde se realizan estos cálculos? En este caso se hacen en Barcelona pero hay una red en España, la Red Española de Supercomputación, en la que los científicos españoles de diferentes áreas

La simulación hecha por la supercomputación permite crear escenarios que sin su intervención serían imposibles. Escenarios caros, peligrosos e imposibles.
Fuente: Prof. Mateo Valero, BSC-CNS 2010.

de investigación, no ingenieros informáticos sino precisamente de otras disciplinas, tienen una herramienta para desarrollar sus experimentos, escenarios caros, peligrosos o imposibles de crear.

Un supercomputador es una máquina de unas dimensiones y de unas características no normales para la mayoría de nosotros. En este caso, el MareNostrum, la máquina que tenemos en Barcelona y el nodo principal de esta red española de siete nodos, tiene 48.000 *cores*. Estas características implican unas dificultades importantes de gestión. Por ejemplo, existe un problema importante de infraestructura para disipar el calor, especialmente en latitudes como las nuestras, porque en Finlandia no tienen tantos problemas de refrigeración, y por tanto no tienen que asumir esos costes. Este supercomputador puede ser utilizado por muchos grupos españoles de investigación. Para ello existe un comité de expertos en diferentes materias que recibe propuestas de proyectos y que ordena y asigna los proyectos. Pero ¿qué pasa con el resto de grupos que no tienen acceso a un supercomputador? Por ejemplo, para ciertas empresas no es fácil entrar en esta red española de supercomputación, tienen que hacerlo a través de grupos de investigación, pero la investigación, por suerte, también se realiza en empresas. Por suerte también, a día de hoy el resto del mundo tiene el *cloud*. Amazon anunciaba hace un año que iba a contar con un supercomputador similar a los que tenemos en la red, que en su momento alcanzó el número 46 de una lista de 500 supercomputadores en el mundo.

La importancia del *Cloud Computing* para la Ciencia

El *Cloud Computing*, o computación en la Nube, es importante para la Ciencia porque, como servicio, ofrece lo que hasta ahora solo podían ofrecer ciertos centros muy especializados con unos costes muy elevados. Crear un centro de supercomputación es muy caro y hasta ahora los recursos tenían financiación pública, pero ya se

sabe cómo está la situación ahora mismo. Con lo cual, el *Cloud Computing* es algo que ya está aquí. WIRED, una revista técnica, publicaba hace un año un artículo titulado: "*Amazon builds world's fastest non existent supercomputer*" (Amazon construye el supercomputador, no existente, más rápido del mundo). Ahora todos tenemos acceso y capacidad para usar un supercomputador. *Cloud Computing* es, en el fondo, un gran número de máquinas en algún lugar del mundo, porque al final la computación y el almacenado sí existen, que se ubican en *data centers* (centros de datos) de los que, a día de hoy, hay decenas en el mundo. El de Amazon ocupa una superficie de 28.000 metros cuadrados, es decir, como cuatro campos de fútbol. El de Microsoft, por ejemplo, ocupa un 40% más, aunque su capacidad aumentó un 60%. La tecnología va avanzando.

Éstas son grandes factorías de información, grandes centrales de producción de información similares a las grandes centrales de producción eléctrica cuya existencia damos por descontado. Y sin embargo, algo similar pasó hace un siglo cuando las empresas dejaron de generar su propia electricidad y se conectaron a la red porque era más barato y les permitía centrarse en su negocio, dejando la producción de electricidad, que ya no era un elemento competitivo, a un profesional que, por economía de escala entre otras cosas, producía el mismo servicio más barato. Ahora está ocurriendo lo mismo en el ámbito de la computación y de los datos. Estos grandes centros de datos, por economía de escala y otros factores, generan el mismo producto, mi computación y mi almacenado, más barato. Así de simple. Y, además, se puede ubicar en Helsinki, donde el sistema de refrigeración es un 44% más eficiente que en Madrid, por ejemplo.

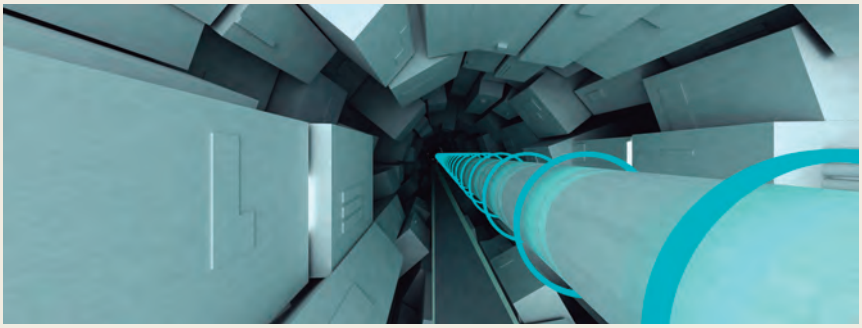
La idea es sencilla. La informática se convierte en un servicio. Un servicio que se paga por uso como la electricidad que pagamos en nuestras casas, donde si gastamos más, pagamos más, y viceversa, y donde puedo

utilizar puntas de energía si las necesito. Se elimina la posibilidad de un gasto innecesario, unas máquinas infrautilizadas o de un servicio insuficiente por no contar con suficientes máquinas. La idea es delegar la infraestructura y las necesidades de un supercomputador en un tercero. Por supuesto, no toda la Ciencia necesita supercomputación. La supercomputación es una parte de la computación que tiene unas características especiales, en lo que a tipo de *hardware* y de almacenado se refiere, que puede realizar un trabajo por partes y en paralelo. Es decir, si una empresa necesita 120 horas de computación para realizar una tarea, puede dividir la tarea por partes y utilizar un supercomputador para hacer todo el trabajo en una hora, porque es como si estuviese utilizando 120 máquinas. Y si lo hace en la Nube no necesita montar 120 máquinas con el coste que eso supone, además de los costes relacionados con el espacio, la refrigeración, la administración, los empleados, etc.

Para mí, éste es el *cloud* de verdad, el auténtico, el que supone la infraestructura como un servicio, una gran base para muchos grupos de investigación que necesitan hacer una simulación en un momento dado y pueden contratar este servicio. Una hora en Amazon cuesta 10 céntimos de euro. Es una oportunidad muy grande que tenemos todos los grupos de investigación y las empresas en general. Es algo muy simple, grandes centros de cálculo que son simples máquinas como las que vemos en casa, agrupadas, gestionadas por alguien y con acceso a través de Internet. Un centro de datos que no está en el piso -2 sino que está al otro lado del mundo y accesible por cables de fibra óptica en los que la latencia, el tiempo de acceso, puede ser pequeñísima si la red es del ancho de banda que requerimos.

En relación a los datos, el supercomputador produce un petabyte de información cada segundo. El *Big Data*, en sí mismo, es un reto muy grande que tenemos todos. Podemos definir el *Big Data* como aquellos problemas en los que los datos exceden los sistemas de almacenamiento que tenemos ahora. De hecho, una de las cuestiones es: ¿tenemos que almacenarlo todo o no? Y ¿cómo procesamos tanta información? La idea es que todo aquello que los sistemas convencionales hasta ahora existentes no soportan se denomina *Big Data*. Si el problema es asumible y tratable ahora, no es *Big Data*. Porque ¿qué es lo que hace que no podamos almacenar, gestionar estas cantidades de datos? No es solo una cuestión de volumen, sino también de la velocidad con la que se generan los datos que salen de sensores. Estamos ya en el mundo de la *Internet of things* (Internet de las cosas). Todo está o empieza a estar sensorizado y es una información de mucho valor que se debe usar, sobre todo en el mundo de la salud, que es uno de los temas más importantes de investigación. Es un *streaming* que va generando información constantemente. ¿Qué se hace con tanta información a esa velocidad? De repente llega el mundo sensorizado de las *Smart Cities*, otra de las áreas de investigación más relevantes, y aparece la posibilidad de añadir a su volumen de información factores como la contaminación, los recorridos de los au-

...”The LHC produces 1PetaByte of data every second, big data and lack of computing resources were becoming the European Organization for Nuclear Research’s biggest IT challenges...”



El “enorme” volumen de los datos es una de las variables que definen el fenómeno *Big Data*. El acelerador de partículas LHC produce 1 PetaByte (1 millón de GigaByte) de datos por segundo.

tobuses, los semáforos, para dar prioridad a los autobuses y que el tráfico sea más fluido. Estos datos no pueden ser procesados por modelos tradicionales de bases de datos estructuradas como hemos hecho hasta ahora.

Y finalmente, el último paso, y quizás el más importante, es cómo cambiamos la manera de analizar estos datos. Aplicamos algoritmos de minería de datos, de aprendizaje, etc., para extraer valor y conocimiento de los datos y muchos sistemas utilizan estos algoritmos para predecir escenarios a partir de los cuales nosotros podamos tomar decisiones. No obstante, estos algoritmos funcionan muy bien para miles de registros, miles de datos, pero no para millones de datos en tiempo real. La mayoría de los datos de la *Internet of things* no pueden ser almacenados, aunque los utilicemos en un momento dado. Bastante trabajo hay ya con los nuevos datos como para dedicar tiempo a los antiguos, con lo que el análisis se vuelve fundamental. En resumen, el mundo científico tiene cuatro retos fundamentales para poder aportar toda esta nueva tecnología que se llama *Big Data* al resto de grupos de investigación: almacenar, gestionar, procesar y analizar los datos. Todavía hay mucho por hacer, a pesar de las expectativas optimistas de mucha gente.

Los retos

Por ejemplo, ¿el almacenamiento de datos es viable económicamente? Claro que sí. Podemos conectarnos a Amazon y contratar dos terabytes por 82€, y esta capacidad de almacenamiento puede ser suficiente para muchas empresas que pueden almacenar el movimiento de una parte importante de su día. Es un gasto asumible, aunque hay que tener en cuenta que actualmente podemos leer discos a una velocidad de 100 Mb/s, por lo que necesitaríamos 5 horas para poder leer dos terabytes. Sin embargo, esto es un problema porque muchas empresas necesitan tomar decisiones empresariales con rapidez. ¿Qué hace Google? En mi opinión, nos ha hecho un flaco favor

La Nube es una oportunidad para acercar la supercomputación a todos aquellos grupos de investigación que hasta ahora no podían contar con ella

porque estamos acostumbrados a ir tecleando y que nos dé sugerencias de búsqueda relacionadas con búsquedas anteriores, ya que Google ha dotado a su buscador de una función de aprendizaje. Pero Google cuenta con 20.000 discos que, en paralelo, leen dos terabytes en un segundo. Lo mismo que se hace en computación, se hace también en almacenamiento. Aunque el primer reto implica cambiar el modelo de procesado. Existen iniciativas como Reduce, Storm o S4, pero el problema no ha sido resuelto todavía. Sobre todo en lo referente al tiempo real, como las decisiones que deben tomarse en una *Smart City* (ciudad inteligente), en donde hay situaciones que requieren encontrar una solución en menos de un segundo.

El almacenamiento es otro de los retos. Hasta ahora se utilizaba la RAM para aquello que se utilizaba mucho en nuestros cálculos y el disco para el resto de la información. La memoria es mil veces más rápida que el disco, pero también es cien veces más cara. En la actualidad, técnicamente, tampoco podemos contar con mucha memoria, lo que supone otro problema, y lo que se está desarrollando muy rápidamente es el denominado *storage class memory*. Hoy en día utilizamos en nuestros ordenadores discos sólidos, que son memorias que se han colocado donde antes había un disco. Son más rápidos, aunque más caros, y lo que se está investigando es el state *storage class memory*, que es colocar la memoria en su lugar. Cuando la comunidad científica haya solucionado este problema tendremos una capacidad de memoria equivalente a la capacidad de disco y con el tiempo esto tendrá un precio razonable. Este tipo de memoria es más económica en consumo porque no es un disco mecánico, sino que está compuesta de circuitos y, por tanto, consume menos energía, que es otra cuestión muy importante que debe tenerse en cuenta.

Las bases de datos relacionales que hasta ahora todos conocíamos y nos han explicado en las facultades ya no nos sirven para resolver grandes problemas. Están surgiendo nuevas propuestas de sistemas como los denominados "NO SQL". Podemos tener muchos datos, pero no sirven de nada porque no es información. Pero es que incluso la información no es conocimiento, y lo importante es lo que se denomina conocimiento accionable: algo que nos permite llevar a cabo una acción. Por ejemplo, no sirve que una aplicación nos informe del estado del tráfico en nuestro camino al trabajo porque el tráfico va cambiando en el tiempo que nosotros empleamos en desplazarnos: necesitamos una aplicación que haga una predicción a partir de datos actuales e históricos del tráfico, el tiempo, la hora etc., y que nos vaya indicando en tiempo real el camino para tardar el mínimo tiempo posible en llegar a nuestro destino. Los datos en sí no nos sirven. Necesitamos que generen conocimiento y esto no es trivial porque las técnicas de *machine learning* y de *data mining* sirven para miles de registros, pero no para millones, por el momento. Estamos trabajando en ello, pero actualmente no lo tenemos. Esto mismo es aplicable a otras ciencias y tenemos la suerte de contar con un centro multidisciplinar en el que colaboramos, por ejemplo, con investigadores de Ciencias de la Vida. Valorizar sus datos no es nada banal: tenemos problemas a todos los niveles, de almacenamiento, de gestión, de procesado, etc. La Nube es una oportunidad para acercar la supercomputación a todos aquellos grupos que hasta ahora no podían contar con ella.