



Humanidades digitales, diálogo de saberes y prácticas colaborativas en red

Cátedra UNESCO de comunicación



NUEVAS ESTRATEGIAS, MÉTODOS E INSTRUMENTOS PARA LA INVESTIGACIÓN SOCIAL

Las nuevas plataformas tecnológicas de supercomputación y su disponibilidad para la investigación en las ciencias sociales

Alexander Herrera¹

La presente ponencia tratará el tema de las plataformas tecnológicas de supercomputación y su función en la investigación. En primera instancia, es necesario iniciar aclarando que las plataformas tecnológicas de supercomputación pueden estar enfocadas en cualquier tipo de investigación y en cualquier área de producción de conocimiento. Para efectos de orden de la ponencia, empezaré definiendo qué es *supercomputación*.

Normalmente, cuando se habla de *supercomputación* se piensa en computadores más avanzados a la máquina que se utiliza en el escritorio de las oficinas. Sin embargo, dicha idea resulta relativa. Por ejemplo, los celulares y tabletas que utilizamos hoy día son cerca de 100 000 veces más poderosos que el

¹ Ingeniero de Sistemas y magíster en *Software* libre, Ciencias de la Computación. Universitat Oberta de Catalunya (España). Ha sido profesor e investigador en las áreas de Información y en el desarrollo de simulaciones computacionales a problemas complejos. Es el coordinador de computación de alto rendimiento del Centro de Alto Rendimiento Computacional Javeriano (ZINE), de la Universidad Javeriana. Experto en HPC, Grid y Cloud Computing. Además imparte algunas asignaturas en el Departamento de Ingeniería de Sistemas, Facultad de Ingeniería de la Pontificia Universidad Javeriana. alexander.herrera@javeriana.edu.co

computador que llevaba el Apolo 11 a la Luna, y muy probablemente los celulares que utilizarán en 30 años serán 1 000 000 veces más poderosos que los celulares y tabletas que usamos hoy.

Entonces, resulta necesario contextualizar el término supercomputación, a partir de la noción de ciencia. La ciencia básicamente entrega una herramienta para poder mejorar, estudiar y aportar unos métodos científicos que superen los métodos de la ciencia tradicional que la precede. La nueva ciencia que surgió entre finales del siglo pasado y principios del presente, junto con sus prácticas científicas e investigativas, empezó a denominarse *e-ciencia*. En cuanto a sus definiciones, podemos encontrar miles. Aquí me centraré en la definición de la eScience Conference Series (IEEE, 2014); esta definición, que es una de las más aceptadas, dice que la *e-ciencia* estudia, promulga y mejora el proceso continuo de innovación en los métodos de investigación computacionalmente intensivos o intensivos de datos. Normalmente, esto se lleva a cabo en colaboración, a menudo utilizando infraestructura de cómputo distribuida, de esta forma se cambian elementos en el proceso investigativo tradicional: se centra en cómo se desarrolla la investigación, cómo se exponen y entregan los datos, cómo se recuperan esos datos, cómo se analizan y cómo se obtienen los resultados a partir de ellos.

Para aterrizar la anterior definición de e-ciencia, expondré un pequeño ejemplo. Primero, es importante aclarar que para todo procesamiento de datos se necesitan los programas o *softwares*, las máquinas que los ejecutan y los servidores. Para recolectar los datos y transferirlos a las plataformas tecnológicas hay un sinnúmero de técnicas, como las que expuso Rafael González en la ponencia anterior. Ahora bien, proseguiré con el ejemplo: hay un acelerador de partículas que realiza simulaciones en física. Este gran acelerador es gestionado por 174 instituciones, en 38 países diferentes. En el momento se están ejecutando unos 25 proyectos, cada dato producto del experimento lo recogen aproximadamente 150 millones de sensores que almacenan 27 *terabytes* de datos por día, se procesan dos millones de *jobs* por día (un *job* es trabajo de cómputo), y aun así quedan, todavía, 140 *petabytes* de datos importantes sin procesar. Para que se hagan una idea, un *petabyte* es aproximadamente 13 años de video continuo en alta definición.

¿Cómo hacer, entonces, o qué hacer para poder analizar esos datos? Hay muchos de esos datos, de hecho, que el ser humano jamás va a poder ver físicamente; de ahí la importancia de las nuevas tecnologías y plataformas para procesar dicha información. Un ejemplo de dichas plataformas tecnológicas es la computación distribuida en Grilla o en Malla denominada en inglés GRID. Se trata, básicamente, de una tecnología que permite entregar los resultados de los datos procesados a los investigadores a escala global.

Pero, para volver al ejemplo del acelerador de partículas, en este experimento trabajan 8 000 físicos de todo el mundo, consultando datos. Como parte de este equipo, hay estudiantes de maestría y pregrado que están trabajando en esos datos; es decir, buscan patrones o elementos significativos entre ese mar de información; todo ello, a partir de interfaces y *softwares* avanzados.

Centrémonos un poco en el caso concreto de la plataforma GRID. Este proyecto surgió con la idea de crear una unidad de recursos en investigación científica dispersos geográficamente. Para crear una unidad no solo en recursos económicos, sino, también, en recursos informáticos, tecnológicos —como unidades de almacenamiento—, humanos; es decir, todo lo referente a los recursos, de forma que contribuyesen a crear un entorno de investigación enfocado. El siguiente objetivo al que le apuntó GRID fue al uso de protocolos e interfaces de estándares abiertos. Antes de la creación de este tipo de plataformas, cada tecnología tenía su propio estándar, protegido por derechos de propiedad intelectual; de este modo, si se quería utilizar algún estándar, era obligatorio comprarle al fabricante las licencias de uso. Hoy en día, con estos protocolos abiertos, cualquier persona puede tomar dicho manual de referencia, implementar y

adaptar la interfaz estándar, con el fin de que sea compatible con las organizaciones con las que se espera trabajar.

Una vez obtenidos los protocolos e interfaces de estándares abiertos, GRID empezó a ofrecer características de servicio no triviales. Esto quiere decir, a grandes rasgos, que la plataforma presta servicios como los de la computación en la nube, pero enfocados en la ciencia; servicios como la transferencia de archivos, por medio de varios canales simultáneos de alta velocidad, de manera gratuita para la comunidad académica de un determinado proyecto; o el procesamiento paralelo para acelerar una simulación, que en un computador normal puede durar años en terminar; o servicios de recuperación de información que otros científicos pueden usar en nuevos descubrimientos.

En la actualidad, con el fin de organizar todos los servicios de la plataforma de la mejor manera, se crearon las organizaciones virtuales. Las organizaciones virtuales tienen varias capas; la primera es una capa de *red*, en la que están todos los dispositivos de red que realizan la conexión de los recursos físicos, como sensores para las GRID de sensores —que se usan, por ejemplo, para los volcanes—, que también interconectan los computadores, servidores y almacenamientos necesarios. Existe una segunda capa de *recursos computacionales*, en la que están los servidores, bases de datos, almacenamientos, que soportan todo el trabajo de las capas siguientes. La tercera capa es la de *software de gestión*, que permite interconectar de forma transparente todos esos recursos con los *entornos virtuales de investigación y usuarios de la organización virtual*, que son las capas restantes.

Aunque el ejemplo de la plataforma de GRID nos vislumbra un panorama general de las características base de la supercomputación y sus funciones, dejémoslo a un lado y hagamos la pregunta: ¿qué más encontramos en supercomputación? Se encuentran las máquinas y el *software*. De acuerdo con el último *ranking* mundial de los supercomputadores, entre los más potentes del mundo se encuentra el supercomputador de China, que tiene unos 10 600 000 procesadores; en Sudamérica, tenemos solo un país en el *ranking* de quienes poseen los 500 computadores más poderosos del mundo: Brasil, que tiene cuatro máquinas. En cuanto al *software* —elemento principal para que una plataforma pueda funcionar—, se destaca un tipo específico, el sistema operativo GNU²/Linux. Este sistema operativo viene de la línea del *software* libre; es decir, posee una licencia libre. La filosofía del *software* libre es compartir conocimiento, compartir ideas, compartir soluciones y generar nuevo conocimiento. Así mismo, esta forma de compartir conocimiento supone una mejora continua del *software*, en la medida en que todos tienen conocimiento de cómo es su configuración y todos pueden contribuir a mejorarlo. A partir de esta idea se han generado proyectos inter- y multidisciplinarios bastante interesantes, y aplicaciones para compartir y procesar datos e información de investigaciones en todas las áreas del conocimiento.

Entonces, la invitación, para finalizar, es que si tienen proyectos de este tipo, con una gran cantidad de datos, los incluyan y hagan parte de estas plataformas tecnológicas, que actualmente ya están al alcance de todos los investigadores.

Referencias

IEEE. (20-24 de octubre de 2014). *The 10th IEEE International Conference on e-Science*. Recuperado de <http://escience.ime.usp.br/events/ieee-escience-2014>

² Véase <https://www.gnu.org/home.es.html>.