

OPINIÓN

El Big Data al servicio de la ciencia en el CERN

Helix Nebula quiere desarrollar una nube científica que forme un mercado abierto para la ciencia

ANDRZEJ NOWAK | 19 JUN 2013 - 11:24 CET

1

Archivado en: CERN i+d Opinión Suiza Política científica Tecnología Europa occidental Organizaciones internacionales Europa Relaciones exteriores Ciencia Sociedad

El [CERN](#) acoge el acelerador de partículas [Large Hadron Collider](#) (LHC), posiblemente la máquina más grande que se ha construido jamás. Este anillo superconductor de 27 kilómetros está situado en un túnel 100 metros bajo la frontera entre Francia y Suiza, cerca de Ginebra. Al colisionar protones a casi la velocidad de la luz, el LHC ofrece una visión inédita de la estructura de la materia y de la historia de nuestro Universo. Las observaciones del año pasado de una nueva partícula que encajaba con el tan deseado bosón de Higgs, que pasa por ser el responsable de la masa de todo lo que nos rodea, se han confirmado y se han estudiado con más profundidad gracias al análisis avanzado de decenas de petabytes de datos, lo cual podría bien llamarse Big Data.

Una empresa como ésta requiere una infraestructura científica y tecnológica única: desde detectores de partículas del tamaño de una catedral hasta equipos de primer orden para la adquisición y el procesamiento de datos.

No es posible observar directamente la interacción que tiene lugar en un detector, pero hay maneras de observar sus consecuencias. Después de una colisión, las partículas siguen descomponiéndose en otras hasta que abandonan el detector o éste las absorbe. Así, algunas partículas que son ya muy conocidas para la ciencia (y que pueden observarse) dejan rastros en diversos subdetectores, que tienen del orden de 100 millones de canales. Si tenemos en cuenta que las colisiones suceden 40 millones de veces por segundo, toda la operación puede compararse con tomar 40 millones de fotos por segundo con una cámara digital de 100 megapíxeles. Para añadir más dificultad, esta «cámara» se desenfoca constantemente y necesita calibrarse de manera regular. La precisión con la que funcionan los detectores puede compararse a la precisión que haría falta para poner un campo de fútbol boca abajo encima de otro y asegurarse de que todas las briznas de hierba están perfectamente alineadas.

Para reducir la carga de almacenamiento, los sistemas de adquisición de datos instalados en los componentes electrónicos y en granjas subterráneas de computación cerca de los detectores siguen buscando señales de nuevos elementos de la física. En primer lugar, la combinación de señales recibida debe entenderse, y esto supone un ejercicio de procesamiento paralelo en tiempo real. Después, debe compararse con decenas de parámetros que describen aspectos ya conocidos. Las muestras que representan elementos físicos ya conocidos se descartan inmediatamente. El uso de sistemas avanzados de filtración como estos implica que apenas llega a almacenarse una pequeña porción de los datos de entrada: aquellos que potencialmente constituyen un nuevo fenómeno. Al final, solo una colisión de cada 10.000.000.000.000 es realmente interesante, pero se mantienen almacenadas muchas más. Los datos escogidos se envían al Centro de datos principal del CERN a una velocidad de 10GB/s, y después se distribuyen por la Red mundial de Computación del LHC, o [Worldwide LHC Computing Grid](#) (WLCG), para procesarlos de nuevo, y se almacenan en cintas magnéticas para archivarlos. Cada año se añaden más de 25 petabytes.

La Red es una federación de más de 150 centros de datos repartidos por todo el mundo que

comparten recursos tales como procesadores y almacenamiento en disco o en cinta para formar una red enorme de unidades de computación interconectadas en la que todas trabajan conjuntamente. La WLCG se basa en la European Grid Initiative (EGI) de Europa y la Open Science Grid (OSG) de Estados Unidos, y por ello es parte de una infraestructura mayor que no está limitada a la física de alta energía sino que también trabaja para la investigación en astronomía, la biología o energía de fusión. En conjunto, la Red funciona sobre más de 350.000 núcleos de arquitectura Intel y ya almacena 0,25 exabytes (250 PB) de Big Data. Con esta infraestructura, un científico puede sentarse en su escritorio en cualquier lugar del mundo y ejecutar cómodamente tareas de simulación o análisis. Es entonces cuando nuestro gran montón de Big Data se convierte en información realmente útil.

Aunque el CERN y sus socios han conseguido liderar y explotar los nuevos recursos de computación (la [World Wide Web](#) nació en el CERN), todavía queda mucho trabajo por hacer. Se espera que el volumen de datos en crudo se multiplique por cien en los próximos años, así que el CERN afronta algunos retos tecnológicos considerables similares a los de los gigantes mundiales del procesamiento de datos y otros grandes laboratorios.

Bajo esas circunstancias, resulta natural unir fuerzas con otros para embarcarse en los grandes desafíos del procesamiento de datos del futuro, y por eso se creó el [CERN openlab](#). Es una investigación conjunta entre el CERN y algunas empresas destacadas de la industria tecnológica (HP, Huawei, Intel, Oracle y Siemens) que quiere desarrollar soluciones informáticas de última generación para el LHC. Desde su inicio en 2001, el CERN openlab ha colaborado en un gran número de proyectos de investigación, sobre todo en los campos de la eficiencia de los procesos, las bases de datos y las redes.

Otra iniciativa de futuro destacada es el proyecto [Helix Nebula](#), que quiere desarrollar una nube científica que forme un mercado abierto para la ciencia. Con el apoyo de un consorcio de 34 empresas y laboratorios de investigación, la iniciativa trabaja para establecer una Nube europea de ciencia de acceso fácil. Entre los participantes hay tres de los grandes nombres de la ciencia europea: el Laboratorio europeo de biología molecular (EMBL), la Agencia Espacial Europea (ESA) y el CERN. Los tres perciben el empuje del Big Data, ya sea en el ámbito de la secuenciación de genoma del futuro, en el de la predicción de terremotos o en el de la exploración de la materia.

En último término, lo más importante de todo es la dedicación de miles de científicos e ingenieros de 110 países que hace posible la captura y el análisis de grandes volúmenes de datos hasta en un momento en que los requisitos cada vez demandan más. Sus esfuerzos incansables son los que dirigen la tecnología para ayudar a convertir el Big Data en una gran ciencia mejor más rápida y beneficiar así a la sociedad en su conjunto.

Andrzej Nowak, ponente del 15º BDigital Global Congress, dirige el CERN-Openlab Platform Competence Center Leader (Suiza).

© 2013 CERN