Avance y Perspectiva

Revista de divulgación del CINVESTAV

Sobre el análisis inteligente de datos en el Gran Colisionador de Hadrones

Karina Galache · Thursday, May 30th, 2019

Categorías: Ciencias Exactas, Zona Abierta

Big Data es un concepto que involucra todo lo referente al hecho de que una determinada cantidad de datos es tan grande que no se puede procesar, almacenar y analizar mediante métodos convencionales y en un tiempo razonable. Se debe contar entonces con hardware y software especializado, como los sistemas computacionales distribuidos.

Científicos de todo el mundo en diversas áreas como la genética, física, astronomía, medicina, ciencias sociales, etcétera,. se enfrentan al desafío de analizar enormes volúmenes de datos para extraer valiosas conclusiones en sus investigaciones. En este artículo se describe cómo los experimentos de física de altas energías, en particular los llevados a cabo en el Gran Colisionador de Hadrones (LHC) del CERN, el acelerador de partículas más energético y grande del mundo, constituyen un proyecto que involucra desafíos tecnológicos e impulsa a los físicos a incorporar técnicas avanzadas de análisis de datos.

Introducción

La física de partículas, dedicada al estudio de los constituyentes fundamentales de la materia y las interacciones entre ellos, es también conocida como física de altas energías. Esta sinonimia no constituye una mera coincidencia trivial, sino que comprende un concepto sustancial de la física: cuanto más profundo se quiera ahondar en el conocimiento de los componentes elementales de la materia, mayor energía se precisará. Además, debido a la equivalencia masa-energía, altas energías permiten la creación de partículas pesadas que son inestables, es decir, no existen como forma estable de la materia.

De esta manera, los aceleradores modernos de partículas, como lo es el Gran Colisionador de Hadrones (LHC, por sus siglas en inglés), en el CERN (Centro Europeo para la Investigación Nuclear), constituyen una herramienta fundamental para los físicos de partículas experimentales.

Las partículas masivas que se espera crear en los aceleradores (como el Bosón de Higgs, cuya existencia pudo probarse en el LHC en 2012 [1]) al colisionar haces de partículas, tienen una muy corta vida media. Esto significa que muy rápidamente devienen, o decaen en otras partículas más

ligeras ya conocidas. Después de que se produce una colisión, cientos de esas partículas ligeras como electrones, muones y fotones, pero también protones, neutrones y otras, vuelan a través de detectores con velocidades próximas a la de la luz. La detección de estas partículas ligeras se utiliza para deducir, si las hay, la breve existencia de las nuevas y más pesadas partículas producidas.

Luego de un proceso de selección de los datos de interés mediante sofisticados sistemas, los experimentos del Gran Colisionador de Hadrones producen anualmente alrededor de 50 millones de GigaBytes (o 50 PB) de datos almacenados.

Para acumular y procesar enormes volúmenes de datos en un tiempo razonable se necesita una tecnología acorde, como son los sistemas distribuidos con sus softwares específicos. Existen varios paradigmas de sistemas distribuidos, como los *clusters* o la computación GRID, entre otros. En cualquier caso, lo que se busca es utilizar múltiples sistemas para abordar un único problema de gran escala.

Los *clusters* son conjuntos de computadoras conectadas a través de una red de área local (LAN) construidos mediante el uso de hardware común y que se comportan como si fuera una única computadora. El término GRID frecuentemente se usa para indicar una infraestructura de gestión de recursos distribuidos que se centra en el acceso coordinado a recursos informáticos remotos. Estos recursos de cómputo son reunidos desde múltiples localizaciones para alcanzar una meta común.

La diferencia fundamental entre estos dos paradigmas de computación distribuida es que un sistema GRID tiende a ser más heterogéneo y disperso geográficamente. Los *clusters* en general son homogéneos en el sentido de que las computadoras que lo forman comparten las mismas características de hardware. En un sistema GRID esto puede ser muy diferente.

El LHC

El LHC es el acelerador de partículas más grande y energético del mundo. Está situado al noroeste de Ginebra, en la frontera franco–suiza y utiliza la estructura de un acelerador precedente, el Gran Colisionador de Electrones y Positrones (LEP, por sus siglas en inglés), que operó entre 1989 y 2000.

Consiste en un túnel de 27 kilómetros de circunferencia situado a una profundidad media de 100 metros bajo tierra. En este túnel, heredado del acelerador LEP, se ubica un anillo de imanes superconductores que, en conjunto con varias estructuras de aceleración, permite aumentar la energía de las partículas a lo largo de su recorrido y controlar su dirección.

Los protones son hadrones, es decir, son partículas compuestas por otras partículas elementales denominadas *quarks* que se mantienen unidas mediante la interacción fuerte, que es una de las cuatro interacciones fundamentales.

En el LHC se aceleran a una velocidad muy cercana a la de la luz dos haces de protones (o de iones pesados que son también hadrones) que se hacen circular en sentido opuesto y viajan en cavidades separadas en las cuales existen condiciones de altísimo vacío. Esto último impide la colisión de las partículas aceleradas con moléculas de gas residual, que provocarían (y lo hacen a menudo

inevitablemente, a pesar de que la presión dentro de los tubos es incluso menor a la que existe en el espacio exterior) dispersión del haz y pérdidas de energía de las partículas aceleradas.

Figura 1: El Gran Colisionador de Hadrones con sus cuatro detectores principales.

Una vez acelerados, los haces de partículas se cruzan para producir colisiones en cuatro puntos alrededor del anillo, que se corresponden con las posiciones donde se ubican cuatro detectores de partículas: ATLAS, CMS, ALICE y LHCb.

Al producirse las colisiones, se generan cascadas de partículas secundarias y el objetivo de estos detectores es registrar las trayectorias de estas partículas, identificarlas, medir sus propiedades y consecuentemente hacer inferencias sobre el proceso físico que se llevó a cabo. ATLAS y CMS son los detectores más grandes con los que cuenta el acelerador.

Además de estos cuatro detectores, el acelerador cuenta con otros tres que se utilizan para fines más específicos: el TOTEM, LHCf y MoEDAL.

Los protones se obtienen a partir de Hidrógeno gaseoso y atraviesan una serie de aceleradores pequeños antes de ser derivados al gran anillo en donde alcanzan su energía final de 7 TeV [2]. En estas condiciones completan más de 11 mil vueltas al anillo por cada segundo.

Los haces de partículas se conforman por alrededor de 2500 paquetes (o *bunches*) con cien mil millones de protones cada uno que se encuentran separados entre sí espacialmente por 7,5 metros y temporalmente por 25 nanosegundos. Cada *bunch* es colimado hasta alcanzar una sección transversal de 16 micrones cuadrados cuando llegan a los puntos de interacción donde tienen lugar las colisiones. El espaciamiento temporal de 25 nanosegundos entre los *bunches* corresponde a una frecuencia de 40 MHz, lo cual implica que los *bunches* atraviesan cada punto de colisión en el LHC 40 millones de veces por segundo. Sin embargo, debido a varias razones técnicas, la frecuencia de cruce promedio resulta ser de unos 35 MHz.

A pesar de la gran cantidad de protones que contiene cada *bunch*, cuando dos de ellos que circulan en sentidos opuestos se cruzan en los puntos donde se encuentran los detectores, la probabilidad de que exista una colisión entre dos protones, siendo el tamaño de estos tan pequeño, es muy baja. Así, ocurren unas 55 colisiones efectivas (que permiten el buen registro de sus partículas secundarias) por cada vez que se cruzan 200 mil millones de protones (100 mil millones por cada *bunch*).

Sin embargo, como los *bunches* se cruzan en promedio 35 millones de veces por segundo, se producen algo menos de 2000 millones de colisiones entre protones por segundo.

En cada vuelta que se recorre, un cierto número de protones se pierden debido a diversos factores como son: la eficacia limitada de los sistemas magnéticos que controlan los haces, la interacción de los protones con las moléculas del gas residual en los tubos de vacío o la interacción electrostática entre los protones en las zonas de colisión o entre los mismos protones que forman los *bunches*.

Estos procesos pueden producir efectos negativos sobre la maquinaria del acelerador, por lo cual se

hace necesario vaciar los tubos luego de 10 horas de circulación a la máxima energía. Esto es conocido como *beam lifetime* o tiempo de vida del haz. Finalmente, se inicia un nuevo ciclo volviendo a inyectar protones que circularán otras 10 horas.

Detección de partículas

Los detectores rodean los puntos de colisión de los protones y están constituidos por capas de subdetectores que cumplen funciones específicas. De manera general, la capa más interna del detector (más cercana al punto de colisión), contiene un detector de trazas cuya misión es determinar la trayectoria de las partículas cargadas resultantes de la colisión. Estas trayectorias se separan de acuerdo a la carga de las partículas secundarias mediante potentes imanes superconductores.

Luego se encuentran los calorímetros electromagnéticos y hadrónicos, que se utilizan para determinar la energía de las partículas absorbiéndolas. El calorímetro electromagnético se encarga de determinar la energía de fotones y electrones, mientras que en el calorímetro hadrónico (que se sitúa más externamente) se mide la energía de los hadrones que se han formado durante la colisión.

Finalmente, en la parte más externa del detector, se encuentran las llamadas cámaras de muones, que sirven para registrar los muones más penetrantes.

Existen algunas partículas, como los neutrinos, que no dejan registro en los subdetectores. Su presencia es deducida a partir de desequilibrios de magnitudes físicas (como la energía) una vez que se reconstruye completamente el suceso. Es por esto que el detector debe ser hermético, evitando que partículas producto de la colisión puedan perderse sin ser detectadas.

Un gran volumen de datos

Se ha visto que el LHC produce alrededor de 2000 millones de colisiones protón-protón por cada segundo en los distintos detectores. Cada colisión produce una cascada de partículas secundarias que son detectadas. Tomando en cuenta que la cantidad de datos promedio que se generan en cada evento (un evento es una colisión y todas las partículas secundarias que produce) es de más 1 MB, se tiene entonces que se generan alrededor de 2 PB por segundo.

Si se consideran DVD’s de capa simple de unos 5 GB, se precisarían entonces algo menos de medio millón de DVD’s por segundo para almacenar estos datos, o bien, más de 60.000 pendrives de 32 GB por segundo.

No es posible manejar esta desmesurada cantidad de datos con el sistema de adquisición de datos de ningún detector actual. Aún si así fuera, restaría almacenarlos de forma permanente y distribuirlos a miles de físicos dispersos por el mundo que están preparados para analizarlos.

Por este motivo, los detectores del LHC cuentan con un sistema de análisis de datos en tiempo real, o sistemas de *trigger*, diseñados para rechazar los eventos que no resultan interesantes desde el punto de vista físico. Estos sistemas seleccionan algunos de los datos recolectados por los

detectores para un análisis posterior y descartan de forma permanente el resto.

Los sistemas de *trigger* se componen de varios niveles. En cada nivel se toma una decisión sobre los datos que se reciben y estos se transmiten al siguiente nivel.

Cuanto más profundo es el nivel, mayor cantidad de tiempo se dispone para tomar la decisión y mayor es la información con la que se cuenta para tomarla. Es decir, cada nivel refina la decisión que se toma en el nivel anterior.

Cada uno de los detectores del LHC tiene un sistema de *trigger* particular. El detector ATLAS durante el run I (período de tiempo de adquisición de datos del LHC comprendido entre 2009 y 2013), tuvo un sistema de *trigger* basado en tres niveles. Actualmente, durante el run II del LHC que comenzó 2015, el sistema de *trigger* de este detector pasó a tener dos niveles.

El primer nivel del detector ATLAS, o *Level-1*, acepta los datos a la frecuencia de colisión del LHC (35 MHz) y toma una decisión del tipo sí/no cada 10 microsegundos en promedio (aunque tiene un tiempo de latencia mayor y cercano a 1 milisegundo) teniendo en cuenta los datos que recibe de la cámara de muones y de los calorímetros. Esta decisión se basa en comparaciones de umbrales para desechar canales no válidos y delimitaciones de regiones de interés en el detector.

De acuerdo con la capacidad de manipulación de datos con la que se cuenta, la frecuencia de salida de datos hacia el segundo nivel es de alrededor de 100 KHz. Si la decisión que se toma sobre el evento resulta negativa, entonces los datos se perderán de manera permanente.

El segundo nivel, o HLT por *High Level Trigger*, trabaja con información de todos los subdetectores y reduce la frecuencia de salida de datos a 1 KHz. Recibe los eventos que no se desecharon en el primer nivel y utiliza sofisticados algoritmos de selección para tomar la decisión en un tiempo medio de 1 milisegundo y una latencia de hasta 2 segundos. Los eventos que logran pasar las decisiones negativas del sistema de *trigger* se almacenan de manera permanente para su análisis posterior (offline).

Considerando lo anterior, el detector ATLAS produce alrededor de 1 GB por segundo de datos. Al igual que ATLAS, el detector CMS también origina cerca de 1 GB por segundo de datos. LHCb produce algo más de medio GB por segundo y ALICE varios GB por segundo cuando se producen colisiones de iones pesados. En total, anualmente son producidos para almacenar de manera permanente alrededor de 50 millones de GB (o 50 PB).

La World Wide Web y la GRID

En 1989 mientras Tim Berners-Lee trabajaba en el CERN, escribió una propuesta que permitía, a través de la conexión de computadoras a internet, el intercambio de información entre científicos (más precisamente físicos de partículas) que se encontraban en distintas universidades e institutos alrededor del mundo. Un año más tarde, tras refinar esta propuesta con la ayuda de Robert Cailliau, nacía la *World Wide Web*, un servicio que hoy es utilizado por usuarios en todo el mundo y permite una comunicación global a una escala sin precedentes.

Años más tarde, los científicos del LHC se enfrentaron a un problema similar: el gran volumen de datos extraídos por los distintos detectores del acelerador debía ser almacenado, procesado,

analizado y compartido con toda la comunidad de física de altas energías que utiliza el LHC. Esto fue resuelto implementando la Worldwide LHC Computing Grid (WLCG), una estructura basada en la tecnología GRID.

La GRID es un sistema de computación distribuido que admite compartir recursos de una forma no centralizada geográficamente permitiendo resolver problemas a gran escala. Es decir, así como la World Wide Web permite el intercambio de información entre computadoras a través de internet, la tecnología GRID va más allá de este concepto, posibilitando a computadoras y otros instrumentos que se encuentren conectados en red compartir no sólo información, sino también poder de cómputo y recursos como almacenamiento en discos, bases de datos y aplicaciones de software.

Por sí sólo el CERN no cuenta con los servicios informáticos o financieros para procesar la enorme cantidad de datos del LHC. Por lo tanto, a través del sistema WLHC comparte esta carga vinculando miles de computadoras y sistemas de almacenamiento en más de 170 centros de computación dispersos en 41 países.

Estos centros de cómputo están organizados en cuatro niveles (o *tiers*) y juntos sirven a una comunidad de más de 8000 físicos con acceso prácticamente en tiempo real a los datos del LHC. El sistema permite almacenar, distribuir y analizar los aproximadamente 50 PB de datos que se producen anualmente en el LHC.

tier-0: El primer nivel (tier-0) lo constituye el CERN, contando con una granja de alrededor de 11.000 servidores que se disponen en 1450 metros cuadrados. Además, este centro es apoyado por el Centro Wigner de Investigación en Física de Budapest (Hungría), el cual actúa como extensión del mismo, añadiendo unos 20.000 núcleos de procesamiento adicionales y 5.5 PB de almacenamiento en disco. Estos dos centros se conectan por una red que permite una transferencia de 100 GB por segundo.

Todo esto resulta en sólo un 20% de la capacidad de procesamiento de la WLCG. Luego de que superan el HLT del *trigger*, los datos llegan al tier-0.

Este nivel almacena una primera copia de los datos crudos que se obtienen desde los detectores y empieza el proceso de reconstrucción, es decir, de determinación a partir de los datos crudos, objetos físicos del proceso de colisión, como los distintos tipos de partículas con su energía, trayectorias, etcétera. Todo esto es transferido al siguiente nivel de la red GRID.

tier-1: Formado por 13 centros de cómputo suficientemente grandes como para albergar la información que reciben mediante enlaces de fibra óptica de 10 GB/s. Funcionan como una extensión del *tier-0*, recibiendo los datos en bruto y reconstruidos. Aquí empieza el proceso de análisis de datos y es distribuido a los *tier-2*.

tier-2: Corresponde a universidades y otros centros de investigación científica capaces de almacenar suficiente información y proveer potencia computacional para determinados experimentos y análisis. Existen alrededor de 160 centros *tier-2* en todo el mundo. Estos centros reciben los datos del *tier-1* y completan en análisis de los mismos mediante simulaciones computacionales.

tier-3: Lo conforman los usuarios individuales o *clusters* alojados en instituciones específicas de universidades que colaboran en los distintos experimentos. No hay un compromiso formal entre la

WLCG y los recursos de este nivel. Aquí se realizan simulaciones puntuales que se necesiten para análisis particulares.

Hacia un análisis inteligente de los datos

Los datos por sí mismos no tienen valor, son sólo hechos o cifras sin procesar, sin ninguna interpretación y ningún análisis añadido. De los datos se deriva la información, que son los datos procesados a fin de que adquieran sentido mediante su contextualización, categorización, cálculo, corrección y condensación.

Finalmente, de la información se deriva el conocimiento, que es una mezcla de la información adquirida contextualizada y la experiencia que sirve como marco para la incorporación de nuevas experiencias e información. La información se convierte en conocimiento al ser analizada, comparando los resultados con modelos, buscando conexiones o patrones y determinando sus consecuencias.

Esta cadena Datos – Información – Conocimiento se denomina “Jerarquía de la Información” y es la base de la inteligencia de datos, un área de la informática que trata de generar o adquirir conocimiento a partir de estos datos.

Figura 3: Worldwide LHC Computing GRID. Imagen adaptada de http://wlcg-public.web.cern.ch/tier-centres.

En el LHC, los datos se obtienen en los detectores. La información comienza a generarse a partir de los datos en los sistemas *trigger* de los detectores y en el *tier-0* del sistema GRID, donde comienza el proceso de reconstrucción. La obtención del conocimiento a partir de esta información comienza en el *tier-1* con el análisis de los datos y se continúa en los *tier-2* y *tier-3*.

Los físicos que colaboran en alguno de los experimentos del LHC, analizan la información que reciben con el fin de corroborar predicciones del Modelo Estándar de partículas y sus interacciones, la teoría que mejor describe la naturaleza a nivel fundamental, pero que se sabe, es una teoría incompleta, o con el fin de encontrar nueva física. Uno de los grandes desafíos consiste en extraer señales extremadamente raras, si las hay, que se encuentran mezcladas entre otras señales que no resultan de interés (el fondo) y que surgen de procesos físicos ya conocidos. El uso de técnicas de análisis avanzadas es crucial para lograr este objetivo.

Existen dos métodos que tienen gran importancia en el análisis de datos dentro de la física de altas energías: la estadística bayesiana y los métodos estadísticos multivariables. En el enfoque bayesiano, la interpretación del concepto de probabilidad no es la frecuencia de resultados en experimentos repetibles, sino el grado de confianza de que un resultado ocurra. De esta manera, se puede asociar la probabilidad con una hipótesis y así responder directamente preguntas que no pueden abordarse fácilmente con los métodos frecuentistas tradicionales. No obstante, con excepción de casos particulares relativos a, por ejemplo, el análisis de la eficiencia de *triggers*, suele utilizarse la estadística frecuentista.

En la estadística multivariante se determina la contribución de múltiples variables simultáneas en un resultado, tratando de explotar tanta información como sea posible de las características que se

han medido. Esto resulta un tratamiento natural para abordar el problema en el cual se intenta distinguir dos tipos de eventos, la señal de interés y el fondo, basándonos en características que se han medido para cada evento.

En los últimos años, la utilización de técnicas de análisis multivariantes mediante algoritmos de aprendizaje automático se ha convertido en una parte integral del análisis de datos en la física de altas energías.

De manera general, puede decirse que el aprendizaje automático – o *machine learning* – es una rama de la inteligencia artificial en la que se diseñan mecanismos para dotar a los sistemas computacionales de capacidad de aprendizaje. En este contexto, la interpretación semántica de "aprendizaje" corresponde a la habilidad de reconocer patrones dentro de un conjunto de datos que se analizan.

Entre los algoritmos más usuales de *machine learning* que se implementan al realizar un análisis multivariable en física de altas energías se encuentran las redes neuronales artificiales, los árboles de decisión y las máquinas de soporte vectorial. Con este objetivo el CERN ha desarrollado ROOT, un *framework*, o entorno de trabajo, orientado a objetos para el análisis de datos a gran escala y que provee métodos estadísticos, de visualización y librerías específicas para el análisis de datos en física de altas energías, aunque es también utilizado en otros campos.

En particular, la herramienta TMVA de ROOT dedicada al análisis multivariable (TMVA, por sus siglas en inglés *Toolkit for Multivariate Analysis*), permite implementar los algoritmos de *machine learning* ya mencionados y algunos otros.

Conclusiones

La física y la tecnología parecen seguir caminos opuestos. A medida que la física avanza y se estudian estructuras más pequeñas mediante cada vez más altas energías, los principios físicos se vuelven cada vez más simples. Esta simpleza no implica que la matemática se vuelva cada vez más fácil o que los procesos sean cada vez menos complejos, sino que las reglas que describen la naturaleza son más coherentes, universales y unificadas.

Por otro lado, a medida que todo esto se desarrolla, los progresos tecnológicos parecen hacerse más complejos y más diversos. Cada generación de experimentos de física de altas energías es más grande, potente, complejo y exigente en términos del manejo y análisis de datos que el anterior. Esto requiere un desarrollo tecnológico acorde y de técnicas de análisis de datos más sofisticadas.

Se espera que en los próximos años la cantidad de datos producidos en los experimentos de altas energías se incrementen, proporcionando nuevos desafíos tecnológicos e impulsando a los físicos a incorporar nuevas técnicas y herramientas para el análisis de datos. El LHC constituye, en este sentido, un perfecto paradigma del análisis inteligente de datos en un entorno Big Data.

Agradecimientos

Sinceros agradecimientos a la Dra. M. T. Dova y los Dres. C. A. García Canal, F. G Monticelli, J.

A. Olivas Varela y S. J. Sciutto, por la lectura cuidadosa de este manuscrito, correcciones y sugerencias.

Referencias

[1] https://home.cern/. Accessed May 28, 2018.

[2] P. C. Bhat, *Advanced Analysis Methods in Particle Physics,* AIP Conf. Proc., 583 (2002), pp. 22–30.

Para saber más

G. Cowan, *Statistics for Searches at the LHC, in Proceedings, 69th Scottish Universities Summer School in Physics: LHC Phenomenology (SUSSP69)*: St. Andrews, Scotland, August 19- September 1, 2012, 2013, pp. 321–355.

G. Cowan, *Topics in statistical data analysis for high energy physics, in High energy physics. Proceedings, 17th European School*, ESHEP 2009, Bautzen, Germany, June 14-27, 2009, 2013, pp. 197–218.

D. Froidevaux, P. Sphicas, *General purpose detectors for the Large Hadron Collider*, Ann. Rev. Nucl. Part. Sci., 56 (2006), pp. 375– 440.

G. Kane, A. Pierce, *Perspectives on LHC Physics*, World Scientific, 2008.

S. V. Gleyzer, L. Moneta, O. A. Zapata, *Development of machine learning tools in root*, Journal of Physics: Conference Series, 762 (2016), p. 012043.

B. Kahanwal, T. P. Singh, *The distributed computing paradigms: P2p, grid, cluster, cloud, and jungle*, (2013).

A. R. Martz, *The Run-2 ATLAS Trigger System*, J. Phys. Conf. Ser., 762 (2016), p. 012003.

H. Voss, *Successes, challenges and future outlook of multivariate analysis in hep, Journal of Physics: Conference Series*, 608 (2015), p. 012058.

Para conocer detalles del descubrimiento del Bosón de Higgs y una introducción a la física de partículas, se recomienda la lectura de: M. T. Dova, *Qué es el bosón de Higgs*, Editorial Paidós, 2015.

El electronvoltio (eV) no es más que una unidad de energía adecuada para las mediciones en un acelerador y la física de altas energías. Un Teraelectronvoltio (TeV) corresponde a 1000000000000 electronvoltios.

Laura Calcagni
Departamento de Física, Universidad Nacional de La Plata

This entry was posted on Thursday, May 30th, 2019 at 7:56 am and is filed under Ciencias Exactas, Zona Abierta
You can follow any responses to this entry through the Comments (RSS) feed. Both comments and pings are currently closed.