Big Data: Desafíos para la investigación en Informática

Las Doctoras Verónica Gil Costa y Marcela Printista de la Universidad Nacional de San Luis, fueron las encargadas de dictar uno de los cursos de la Escuela de Informática en el Congreso Argentino de Ciencias de la Computación 2017. En este marco, brindaron su aporte acerca del Big Data.

En su opinión, ¿cuáles son los mayores desafíos que plantea Big Data?

Creo que uno de los mayores desafíos para Big Data es a partir de la gran cantidad y variedad de datos encontrar la manera de brindarles a las personas información relevante para cada uno de ellos en tiempo real, ya que esto permitirá una mejora considerable en la toma de decisiones. En otras palabras, permitir mejorar significativamente el acceso a la información más relevante, oportuna y precisa entre una amplia gama de fuentes. Sin embargo, los datos que circulan son mayores comparados con los volúmenes de información que se pueden digerir. Por lo tanto, no sólo es necesario agregar recursos físicos (computadores, hardware) sino que hay un condimento fundamental en lo que respecta a la explotación de datos y cómo sacar provecho de ellos, el factor humano.

¿Cómo el Big Data ha impactado a la informática?

Big Data tiene un gran impacto en la informática. Ha permitido desarrollar nuevas tecnologías, y nuevos modelos de procesamiento de datos. Es un área que requiere la interacción de diferentes especialidades como la matemática y estadística para hacer correlaciones de los datos, limpieza,de-duplicación y normalizado de información no estructurada. Presenta una oportunidad para formar "científicos de datos" que incluya una formación multidisciplinaria que está en continuo avance y actualización. Creo que actualmente el alcance e impacto de Big Data está limitado por los dispositivos (hardware), y a medida que estos dispositivos tengan mayor potencia de cómputo, pero también permitan aumentar nuestros sentidos, podremos ser capaces de expandir nuestra forma de adquirir conocimientos y experiencias.

¿Qué soluciones puede dar el ámbito de HPC a los problemas de Big Data?

El cómputo de alto rendimiento (HPC) se apoya en tecnologías computacionales como los clúster, supercomputadores o mediante el uso de la computación paralela. HPC permite realizar ejecuciones de simulaciones, aplicaciones y programas para el análisis de datos sobre recursos de gran escala. Para la comunidad de investigadores, la combinación de HPC y Big Data ha permitido el desarrollo de ambientes de software más eficientes, escalables y que tengan la flexibilidad y usabilidad de las herramientas de Big Data.

¿Existen suficientes herramientas actualmente?

Hoy en día existe una gran variedad de herramientas que pueden ser utilizadas para el procesamiento de datos masivos.

Un modelo muy popular es MapReduce, a partir del cual se han desarrollados diferentes herramientas como Hadoop de Apache, Hortonworks, Amazon Elastic, entre otros. Por otro lado, existen las herramientas desarrolladas para plataformas de stream processing como Storm, S4, Spark, Flink, etc. Cada una de estas herramientas provee diferentes enfoques de procesamiento de datos (como batch processing, stream processing, micro-batching) y son eficientes para resolver diferentes tipos de problemas. Continuamente se están desarrollando nuevas herramientas para abordar problemas de Big Data, pero cada una de estas está enfocada en un tipo particular de problema. Además, estas herramientas disponibles requieren de personal muy calificado para obtener información de calidad.

¿Hacia qué aspectos debería enfocarse principalmente la investigación en esta área?

Como ya comenté anteriormente, el área de Big Data involucra diferentes especializaciones y disciplinas, y para ello es importante implementar formatos que permitan compartir datos entre dichas disciplinas. Sin embargo, muchas veces es difícil combinar datos de distintas fuentes (Instagram, Facebook, Twitter, metadatos, etc.) con estructuras a veces incompatibles. Big Data puede ser utilizada para encontrar relaciones sutiles entre datos que a simple vista parecen no tener relación, realizar estimaciones en áreas como la economía, estudios en medicina, etc. Por otro lado, crowdsourcing (colaboración abierta distribuida) es una componente importante para Big Data. Hoy en día existen algunas plataformas como Tomnod en la cual los voluntarios participan en campañas para resolver problemas de etiquetado de imágenes. Por lo tanto, existen numerosas iniciativas que intentan combinar diferentes enfoques, tecnologías, datos; y Big Data es un área que podría lograr un puente entre cada una de ellas con el fin de obtener información relevante para mejorar el conocimiento de las personas •