El Ingeniero de datos

Miguel Toro, Arantza Illarramendi, Francisco Ruiz

El cada vez más dominante mundo de los datos

Hoy estamos inundados de una avalancha de datos. En una gran cantidad de diferentes áreas, los datos se recopilan a una escala sin precedentes. Las decisiones que anteriormente se basaban en conjeturas o en modelos de realidad minuciosamente construidos, ahora pueden también basarse en una mayor cantidad de datos. El análisis de esos datos impulsa muchos aspectos de nuestra sociedad moderna, incluidos los servicios móviles, el comercio minorista, la fabricación, los servicios financieros, las ciencias de la vida y las ciencias físicas. Es lo que popularmente se conoce como Big Data (datos a gran escala, inteligencia de datos). Es el mundo de los datos y su importancia es cada vez mayor.

La investigación científica ha sido revolucionada por el Big Data. El campo de la Astronomía se está transformando de tomar fotografías del cielo a encontrar objetos y fenómenos interesantes en una base de datos que contiene imágenes de las galaxias. En las ciencias biológicas, ahora existe una tradición bien establecida de depositar datos científicos en un repositorio público, y también de crear bases de datos públicas para el uso de otros científicos. A medida que avanza la tecnología, particularmente con la llegada de la secuenciación de próxima generación, el tamaño y la cantidad de conjuntos de datos experimentales disponibles aumenta de forma exponencial. Ocurre de forma similar en otras áreas de investigación.

El Big Data tiene el potencial de revolucionar no solo la investigación, sino también la educación. Podemos imaginar un mundo en el que tengamos acceso a una gran base de datos donde guardemos medidas detalladas del rendimiento académico de cada estudiante. Esta información podría usarse para diseñar los enfoques

más efectivos para la educación, comenzando por la lectura, la escritura y las matemáticas, hasta los cursos avanzados de nivel universitario. Estamos lejos de tener acceso a esos datos, pero hay tendencias poderosas en esta dirección.

El almacenamiento y uso intensivo de los datos puede reducir el costo de la atención médica y mejorar su calidad, al hacer la atención más preventiva y personalizada y basarla en un seguimiento continuo y extenso de las actividades y síntomas de las personas.

El uso intensivo de los datos se ha mostrado interesante para la planificación urbana (mediante la fusión de datos geográficos de alta fidelidad), el transporte inteligente (a través del análisis y la visualización de datos vivos y detallados de la red de carreteras), modelos medioambientales (a través de redes de sensores que recopilan datos de forma ubicua), ahorro de energía (mediante el descubrimiento de patrones de uso), análisis de riesgo sistémico financiero (a través del análisis integrado de una red de contratos para encontrar dependencias entre entidades financieras), seguridad nacional (a través del análisis de redes sociales y transacciones financieras de posibles terroristas), seguridad informática (a través del análisis de la información registrada), y así sucesivamente.

Sin embargo, los problemas aparecen de inmediato durante la adquisición de datos, cuando el tsunami de datos nos obliga a tomar decisiones sobre qué datos conservar y cuales descartar, cómo almacenarlos de manera confiable. Los datos actuales son de tipologías muy diversas: los tweets y blogs son fragmentos de texto débilmente estructurados, mientras que las imágenes y los videos están preparados, en un primer momento, para su almacenamiento y visualización, pero no tanto para su búsqueda y análisis. Transformar ese contenido en un formato adecuado para su posterior análisis es un desafío importante. El valor de los datos aumenta considerablemente cuando pueden vincularse con otros datos, por lo que la integración de datos es otro desafío relevante. Como la mayoría de los datos se generan hoy directamente en formato digital, tenemos la oportunidad de influir en la creación de

los datos para facilitar el enlace posterior y vincular automáticamente los datos creados previamente.

Como ya sabemos, incluso para análisis simples que dependen de un solo conjunto de datos, una cuestión importante es la elección y el diseño de la base de datos o del depósito de datos (en un sentido más amplio) adecuado. Por lo general, habrá muchas formas alternativas de almacenar los datos. Ciertos diseños tendrán ventajas sobre otros para ciertos propósitos, y posiblemente inconvenientes para otros. Véase, por ejemplo, la gran variedad en la estructura de los depósitos de datos bioinformáticos con información sobre entidades sustancialmente similares, como los genes. El adecuado diseño de los depósitos de datos es hoy en día un reto muy importante en este mundo de los datos.

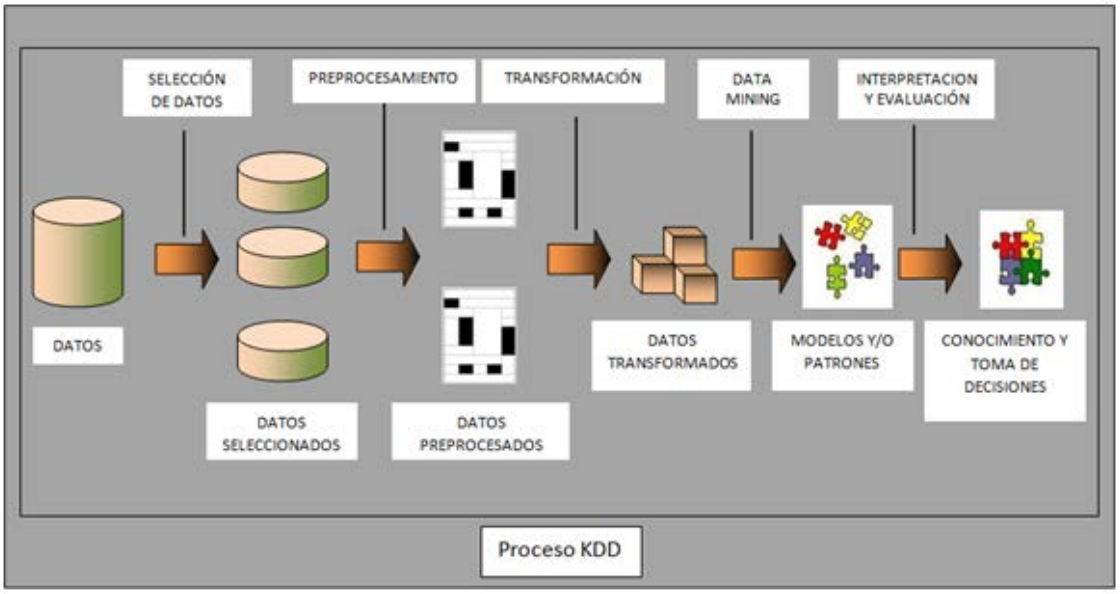
Los métodos para consultar y extraer conocimiento de los datos actuales son fundamentalmente diferentes de los análisis estadísticos tradicionales en muestras pequeñas. Los datos, en el mundo del Big Data, son distribuidos, tienen ruido, son dinámicos, heterogéneos, interrelacionados y en muchos casos poco fiables. Sin embargo, incluso los datos con mucho ruido podrían ser más valiosos que las muestras pequeñas porque los patrones obtenidos suelen dominar las fluctuaciones individuales y, a menudo, revelan patrones y conocimientos ocultos más confiables. Además, interconectando grandes redes de información heterogénea, se puede explorar la redundancia para compensar los datos que faltan, verificar casos conflictivos, validar relaciones y descubrir nuevas relaciones y modelos ocultos.

El análisis de datos es un cuello de botella en muchos casos, debido a la falta de escalabilidad de algunos algoritmos subyacentes y a la complejidad de los datos que deben analizarse. Finalmente, la presentación de los resultados y su interpretación por expertos en el dominio en el que se generan los datos, es crucial para extraer conocimiento relevante.

Los profesionales del mundo de los datos

Aun partiendo del hecho de que la Informática es una Ingeniería madura, desde la Informática se han hecho esfuerzos de acercamiento a las Ingenierías clásicas con el fin de aprender de su experiencia, surgen nuevos debates en la relación de la Informática con la Ingeniería. Uno de ellos está en el campo de los datos y las profesiones que están surgiendo alrededor. Se están ofertando Másteres para formar a científicos de datos. Se habla que el mundo del Big Data necesita Científicos de Datos. Pero se habla poco de Ingenieros de Datos. Claramente un Ingeniero, en general, debe conocer la ciencia disponible. La Ciencia estudia el mundo como es. Su pasión es descubrir la verdad. La Ingeniería crea el mundo que aún no existe. Su pasión es crear soluciones. En particular en el caso del Ingeniero de Datos debe conocer los algoritmos disponibles para analizar los datos, agruparlos o hacer predicciones. Pero un Ingeniero necesita muchas más capacidades: gestión de equipos, mantenimiento de la infraestructura, tratar con los clientes, evaluar y gestionar los costes, estar al día de las tecnologías y saber elegir, y adquirir, las más adecuadas en cada caso, etc. Son habilidades que debe tener el Ingeniero de Datos. En definitiva, debe saber elegir y mantener la infraestructura que da soporte a los datos a un coste adecuado y gestionar los recursos humanos necesarios. En el mundo del Big Data se observa la necesidad de formar Ingenieros de Datos y no solamente Científicos de Datos.

Con el fin de situar las características de los dos perfiles: Ingeniero de Datos y Científico de Datos vamos a utilizar el esquema ya definido en la literatura especializada que refleja los pasos necesarios para obtener conocimiento a partir de un conjunto de datos (KDD, Knowledge Discovery in Databases).



El perfil del Ingeniero de Datos y sus especificidades

De una forma resumida, el Ingeniero de Datos centrará fundamentalmente su labor en las etapas de selección, pre procesamiento y transformación y el Científico de Datos en las de Análisis de datos y Evaluación. Puede ocurrir que en organizaciones no tan grandes las competencias de ambos perfiles las desarrolle la misma persona.

Para precisar los dos enfoques veamos dos tipos de conocimiento que se complementan entre sí: el conocimiento científico, el que persiguen los científicos, y el conocimiento tecnológico, el que buscan los ingenieros. Tienen propósitos diferentes: el primero trata de ampliar y profundizar el conocimiento de la realidad; el segundo, de proporcionar medios y procedimientos para satisfacer necesidades. El conocimiento científico pretende saber, el conocimiento tecnológico saber hacer. Han sido muy estudiados en la literatura los criterios que indican la validez de la teoría científica. Sin embargo, el criterio de validez de una tecnología no es tanto que sea verdadera, sino que funcione en la práctica y sea útil. El conocimiento tecnológico, por otra parte, está orientado hacia la resolución de problemas complejos y la toma de decisiones en cuestiones que inciden muy directamente en el desarrollo económico de la sociedad.

Desde el punto de vista anterior un Científico de Datos busca nuevos algoritmos y usa los ya existentes para analizar datos, para encontrar patrones en los datos. Un Ingeniero de datos gestiona los datos, diseña los procedimientos para capturar los datos en crudo, realiza un pre procesamiento sobre dichos datos, transforma e interrelaciona los datos, escoge la infraestructura adecuada para el presupuesto disponible, y genera el conjunto de datos elaborados sobre los que se aplicarán técnicas de análisis. El Ingeniero de Datos se guía por un problema de datos concreto a resolver en una empresa o institución y busca soluciones al mismo. Para ello parte de un presupuesto y usa los algoritmos y tecnologías disponibles. El Científico de Datos se guía por la búsqueda de algoritmos para encontrar patrones en datos ya previamente organizados. Son dos perfiles distintos y complementarios.

El ingeniero de datos es alguien que desarrolla, construye, prueba y mantiene arquitecturas, bases de datos y sistemas de procesamiento a gran escala. Los Ingenieros de Datos tendrán que implementar nuevas formas de mejorar la fiabilidad de los datos, la eficiencia y la calidad de los mismos. Para ello, necesitarán emplear una variedad de lenguajes y herramientas que permitan agregar datos diversos y distribuidos o buscar oportunidades para adquirir nuevos datos no conseguidos hasta ahora. Relacionado con lo anterior está el hecho de que los Ingenieros de Datos deberán asegurar que la arquitectura elegida soporte adecuadamente los requisitos necesarios para el análisis de los datos y los especificados por los grupos de interés del negocio.

Un aspecto muy importante que deberá tener en cuenta el Ingeniero de Datos es la seguridad y la confidencialidad de los datos. Son aspectos presentes en cualquier aplicación, pero en el mundo de las aplicaciones intensivas en datos la granularidad que se exige en el control y el acceso al conocimiento y a la información (individual y agregada) cobra singular importancia. Especialmente, cuando se requiere garantizar la conformidad respecto de las normativas europeas, nacionales y organizacionales.

Existe una clara superposición en los conjuntos de habilidades de los Ingenieros de Datos y los Científicos de Datos, pero los dos perfiles se están volviendo cada vez más claros: mientras que el ingeniero de datos trabajará con sistemas de bases de datos, API (Application Programming Interface) de datos, herramientas para extraer, transformar y cargar los datos, estará involucrado en el modelado de datos elaborados y la creación de soluciones para almacenar y consultar los datos; el científico de datos necesitará saber sobre estadística, matemáticas y aprendizaje automático para construir nuevos algoritmos para el descubrimiento de patrones en los datos. El primero es un ingeniero, el segundo un científico. Un perfil se complementa con el otro.

Habilidades del Ingeniero de Datos

Un Ingeniero de Datos es un informático, con conocimientos avanzados de Ingeniería del Software, que conoce las características de los datos, el tipo de consultas frecuentes que son interesantes para la entidad correspondiente y los aspectos en los que la entidad está interesada en mejorar a través de la gestión intensiva de los datos. Diseñar y mantener la infraestructura que soporta los datos es el papel del Ingeniero de Datos.

Tiene que saber responder a preguntas como: ¿Qué infraestructura es necesaria para almacenar los datos relevantes? ¿Cuál es la mejor forma de almacenarlos para que se puedan llevar a cabo de forma eficiente las consultas o aplicar los algoritmos adecuados de extracción de patrones? ¿Cuál es la mejor política de mantenimiento de esa infraestructura? ¿Cuál es la política de seguridad adecuada para esos datos?

Para responder a esas preguntas debe conocer el software y el hardware disponibles y sus posibilidades. Debe conocer técnicas de almacenamiento eficiente, procesamiento de datos en arquitecturas avanzadas y distribuidas y técnicas de ingeniería del software para el pre procesamiento eficiente de los datos. Debe conocer la legislación y las normativas europeas, nacionales y organizacionales

referidas a la seguridad y la confidencialidad de los datos. Debe tener habilidades de comunicación adecuadas para interaccionar con los grupos de interés.

También debe conocer el estado del arte de la Ciencia de los Datos, el conjunto de algoritmos disponibles para extraer patrones de los datos y las posibilidades de visualizar los datos o los patrones extraídos.