"Se busca científico de datos para incorporar a nuestro equipo de Ciencia de Datos"

Este es el texto de un anuncio real de una empresa web norteamericana. Pero ¿qué hay detrás de la "Ciencia de Datos"?

Todos ya hemos oído, en mayor o menor medida, hablar del "dato grande" y tenemos una idea más o menos aproximada de lo que es, pero ¿qué se denomina como ciencia de los datos? La Ciencia de Datos es una disciplina de conocimiento que surge aparejada al fenómeno del Big Data.

MYRIAM CORRAL
Subdirección
de Inspección.
D.G. Ordenación
del Juego.
Ministerio de Hacienda
y Administraciones
Públcias

De los científicos de datos, también llamados superhéroes del Big Data por la cantidad de conocimientos que se les demanda, se espera oír mucho en los próximos años. Las expectativas del empleo alrededor del análisis de los datos se cifran entre 140.000 a 190.000 personas con experiencia en análisis de datos, sólo en Estados Unidos. La cifra asciende a 1,5 millones si incluimos a los gestores con los conocimientos necesarios para comprender y tomar decisiones basadas en el análisis de grandes volúmenes de datos. **(1)**

¿ Qué hay de nuevo?

A pesar de la novedad del término, no todo es nuevo alrededor de la Ciencia de Datos.

Hay personas ejerciendo como científicos de datos desde bastante antes de que apareciera el fenómeno Big Data, especialmente en áreas de conocimiento como la »

Ilustración 1. Anuncio de puesto de trabajo en Facebook

medicina o la astronomía.

¿Dónde está la novedad entonces y por qué se ha acuñado este nuevo término? Principalmente por las "nuevas tecnologías" para el tratamiento de grandes volúmenes de información liberadas por "empresas de datos" tipo Google, Yahoo, Facebook, Amazon, etc.

Para construir estos "productos" basados en datos, estas empresas han tenido que desarrollar toda una serie de tecnologías que permiten el tratamiento de grandes volúmenes de información a costes asequibles. No es que antes no se pudiera hacer: se podían mover Terabytes de datos pero la inversión en software y hardware hacía que solo estuviera al alcance de presupuestos millonarios. Hablamos incluso de centros de supercomputación como los que se dispusieron para la decodificación del genoma humano. E incluso así, había límites, porque no se trata de sistemas especialmente escalables, de modo que una vez superada la potencia de proceso y almacenamiento, habías tocado techo. Ahora estas tecnologías están orientadas a ser desplegadas en un hardware mucho más barato y más escalable y, por tanto, a un coste exponencialmente menor.

Así pues, el nacimiento del Big data se produce cuando estas empresas orientadas a los productos de datos liberan la tecnología que han desarrollado para tratar volúmenes grandes de información y se desarrollan las primeras implementaciones en la comunidad open source. A partir de ahí empieza a ser económicamente viable el análisis de volúmenes de información de tamaño creciente y se abre paso el término Big Data y la Ciencia de Datos para describir el trabajo de los que analizarán dicha información.

Un ejemplo paradigmático del objeto de la ciencia de los datos es el ocurrido durante la campaña electoral de Obama. En la fase más crucial, el analista Nate Silver recibió una enorme atención de la prensa por su precisión en la predicción de los resultados de cada uno de los estados del distrito de Columbia.

El propio Silver manifestó que no había realizado ningún cálculo muy sofisticado, y que la clave residía en la combinación y el análisis de un enorme volumen de datos. Para conseguir tal cosa, el responsable de campaña de Obama se había provisto de un sistema basado en Hadoop con el que conseguía la agregación de datos provenientes de las encuestas a nivel estatal y, además, era capaz de combinarlos con datos económicos y de encuestas previas. Una vez agregados se volcaron a una base datos columnar para permitir su análisis. Finalmente se dio acceso a un equipo de docenas de analistas que fueron extrayendo todos los resultados .

En resumen, la Ciencia de Datos aporta la combinación de las técnicas de análisis clásicas con nuevas tecnologías que hacen asequible la extracción de información de grandes volúmenes de datos con el mismo objetivo de siempre: extracción de información del dato. **(2)**

Definición del científico de datos

En este periodo han surgido definiciones del científico de datos con toques humorísticos como la del "estadístico de siempre, pero con casa en Silicon Valley". O aquella que dice que será la "profesión más sexy de este siglo", aparecida en un celebrado artículo del Harvard Business Review **(3)** . Pero, quizás, una de las definiciones más completas es la de Jeffery Stanton, de la Universidad de Siracusa **(4)** que se refiere a la Ciencia de Datos como un "área emergente de trabajo relacionada con la recolección, preparación, análisis, visualización, gestión y preservación de grandes cantidades de información". Esta definición da una idea aproximada de la variedad de conocimientos que comprende esta nueva disciplina:

Conocimientos informáticos como lenguajes de consulta, diseño de base de datos, minería y análisis interactivo de datos, scripting o algún lenguaje de programación, sistemas expertos y aprendizaje automático, etc.

Conocimientos analíticos basados en matemáticas, álgebra relacional, y sobre todo estadística, análisis predictivo y búsqueda de patrones, etc.

Conocimientos de técnicas de visualización de datos, siendo este un campo muy interesante. Uno de los principales problemas en esta área es cómo traducir el mar de datos a información para la decisión. El ojo humano es el principal canal de trasmisión y hay técnicas desarrolladas para trasmitir la información de la forma más efectiva posible.

Y, por supuesto, conocimientos del área de negocio en cuestión. El científico de datos es un especialista en la información que maneja y su cometido es explotar la información. Hay una diferencia por ejemplo con el especialista en Bussiness Intelligence que es muy ilustrativa. Este último, una vez desplegado el proyecto y transferido el conocimiento, tiende a ceder terreno a favor del usuario que explota la información, finalmente se irá y comenzará el despliegue de otro BI. El científico de datos se dedica a mover datos, combinarlos, analizarlos, sacar conclusiones, convencer de los resultados y rentabilizar todo ello para el negocio y por tanto se queda en el negocio.

La cantidad de conocimientos que se le presupone al científico de datos ha llevado a que con un poco de humor se le denomine el superhéroe del Big data.

Para hacernos idea aún mejor de cómo puede ser el trabajo de una persona en esta disciplina lo mejor es ilustrar una jornada de un científico de datos, Mr. Data.

Mr. Data llega por la mañana y hará un análisis de unos

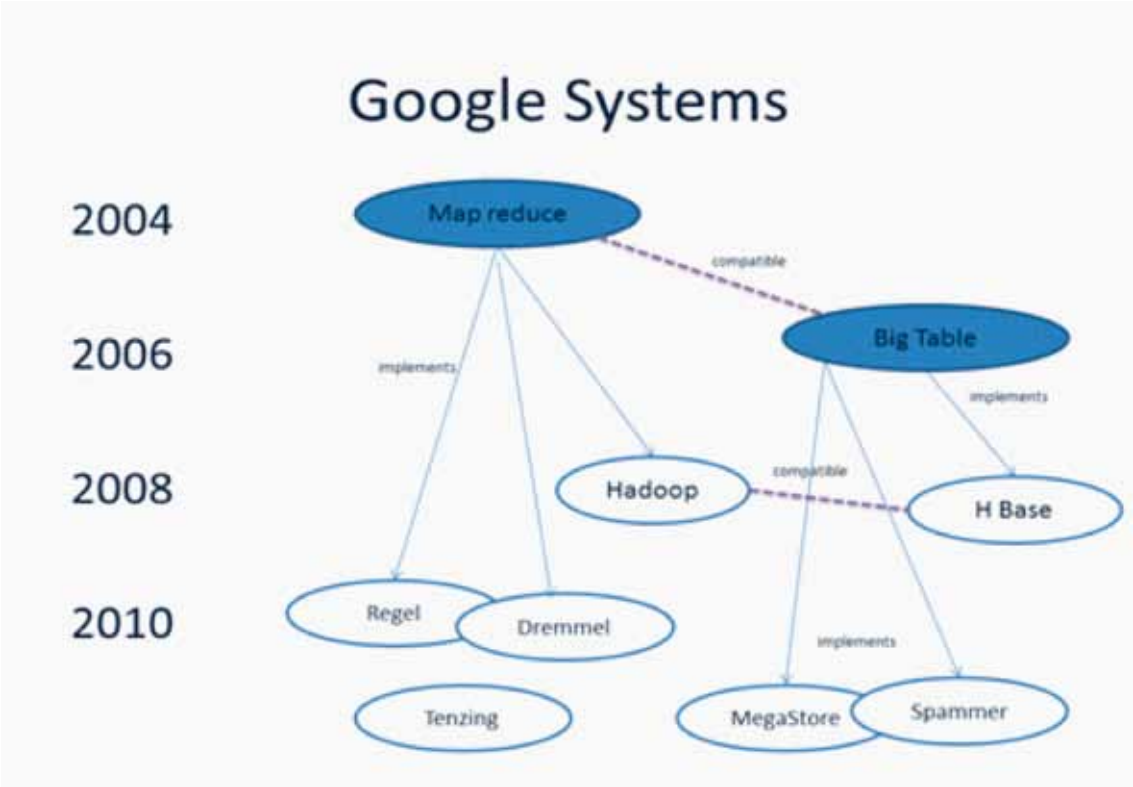

Ilustración 2. Evolución de las "nuevas tecnologías", para el tratamiento de grandes volúmenes

La Ciencia de Datos es una disciplina de conocimiento que surge aparejada al fenómeno del Big Data

ficheros planos de datos extraídos del web log de una empresa. Se trata de texto plano y sin ninguna descripción de su contenido, por lo que empleará una herramienta de visualización para hacerse una idea de qué representan. También empleará un poco de estadística descriptiva y de exploración de datos para completar esta fase previa.

A partir de ahí, se hará una idea de la información que puede extraer y lo conveniente de cruzarla con unos datos geoespaciales provenientes de una investigación previa. Para ello desarrollará un script de Python u otro lenguaje, con el que limpiará variables que le introducen ruido y prepara los dos datasets para el cruce. Como el volumen es considerable, volcará los datos procesados en la nube y realizará el cruce de información mediante algoritmos Map Reduce. Este cruce tardará varios días en realizarse. Cuando finalice elevará las conclusiones del trabajo.

Rol del científico de datos

Una vez desgranado los conocimientos y el tipo de trabajo asociado a esta profesión de nuevo cuño se distingue, con mayor claridad, que el científico de datos no es el técnico de Business Intelligence, aunque se le presuponen conocimientos de este campo. Y tampoco es el estadístico. Se trata de alguien con aptitudes que recorren el amplio espectro de materias que va desde las bases de datos relacionales y soluciones de Business Intelligence, hasta otro tipo de soluciones de alta escalabilidad, que se hacen imprescindibles cuando lo anterior no da para "mover" el dato (o se tardaría tanto que dejaría de tener sentido): los ecosistemas de soluciones surgidas de la liberación de Map Reduce por Google en 2004 y que dio origen a proyectos open source como Hadoop primero, H Base y lenguajes asociados como Hive, PIG, etc.

El trabajo de científicos de datos no difiere mucho de otros especialistas de datos en cuanto a las fases se refiere.

En la fase de diseño el científico de datos colaborará con el arquitecto de la solución para decidir el diseño y almacenamiento de los datos de modo que permita su análisis posterior. Por ejemplo, el científico de datos es el »

que aconsejará que el sistema se provea de una base de datos NO SQL en detrimento de otro tipo de bases de datos.

En la fase de obtención de los datos el científico se centrará en decidir cómo se van a recolectar, teniendo en cuenta cómo serán posteriormente analizados. Esta fase puede llegar a consumir gran parte del trabajo del científico de datos (casi un 80% del tiempo total). Se denomina de forma irónica "Data munging" o "Masticado de datos" y aunque aparentemente es sencilla, los que trabajan con datos reales saben que ahí comienza la pesadilla. Como ejemplo imaginad un log kilométrico de Twitter en formato json que hay que cruzar con datos de varias hojas Excel de formato.

La fase de análisis es la más parecida al trabajo que realizaban los analistas de datos antes de aparecer el concepto Big Data. Entran en juego capacidades de análisis y conocimientos de estadísticas para inferir datos.

La Fase de Visualización de los resultados supondrán una de las tareas más importantes del trabajo del científico de datos y su reto será convencer. Todas las fases previas no servirán de nada si no pueden trasladar los resultados, de forma efectiva, a las personas que toman las decisiones.

Fase de archivado de la información, de forma que sea altamente reusable. Una de las características asociadas al Big data es que los propósitos para los que se recoge la información son muy volátiles. Por ejemplo, los logs de búsqueda de Google jamás se habían pensado que pudieran ser de utilidad en la predicción de gripe.

Como convertirme en un científico de datos

Perfiles tan completos son difíciles de encontrar y éste es uno de los principales inhabilitadores del Big Data. Eso, y las enormes expectativas de trabajo puestas en esta área, hacen que muchos se pregunten pero ¿por dónde empezar?

Ante todo comentar que la capacitación en esta materia es un proceso, y no necesariamente rápido. Aunque hay cursos de todos los tipos y empieza a haber una oferta de estudios de posgrado en esta área, algunas de las ofertas más serias y completas cifran la duración de la formación como "Data Scientist" entre 2 y 3 años, dependiendo de la disponibilidad de tiempo y conocimientos previos. Esto nos da una idea de la complejidad del proceso.

Esbozando las líneas básicas, si la formación es cercana a la ingeniería se pueden refrescar o complementar algunas materias como álgebra y análisis numérico, ya que los algoritmos y los fundamentos matemáticos son a su vez base del data mining y del machine learning. Aquellos que no vengan de una formación en sistemas de información o informática deben conocer fundamentos de bases de datos y de estructuras de almacenamiento.

Hay personas ejerciendo como científicos de datos desde bastante antes de que apareciera el fenómeno Big Data, especialmente en áreas de conocimiento como la medicina o la astronomía.

También es fundamental conocer las bases de la computación distribuida. Cuando se mueve toda esta cantidad de información hay que saber distribuir, sí o sí.

Fundamental también es refrescar conocimientos de estadística. En la actualidad existe software que realiza mucho de los cálculos que antes se hacían a mano, no obstante hay que conocer los conceptos básicos para hacer aproximaciones con rigor. La buena noticia es que hay oferta abundante en esta materia para poder realizar al menos cursos introductorios o básicos. También es interesante la formación en cuestiones de optimización, minería de datos, inteligencia artificial, sistemas expertos y análisis predictivo comentado anteriormente.

Todo esto lo puedes hacer por tu cuenta o de forma ordenada siguiendo el catálogo de cursos y currículo para esta disciplina que aparece en universidades como MIT o Stanford, que ya lo tienen en su oferta. Otra opción interesante es seguir alguno de los MOOC **(5)** que se ofertan en plataformas tipo Coursera. **(6)** ✱

NOTAS

(1) *http://www.mckinsey.com/Features/Big_Data*
(2)*http://citoresearch.com/data-science/how-vertica-was-star-obama-campaign-and-other-revelations*
(3)*http://www.popsci.com/science/article/2012-09/harvard-business-review-data-scientist-sexiest-job-21st-century*
(4) *http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf*
(5)*http://en.wikipedia.org/wiki/Massive_open_online_course*
(6) *https://www.coursera.org/*