El datascientist

Author : paucabla

Categories : General, Sistemas de Información Empresarial

Date : 11 enero, 2019

Los temas propuestos para la reflexión en este tercer post son muy heterogéneos. Por ello, una de las pocas cosas que he podido sacar en común aparte de la disrupción digital es que todos producen una cantidad ingente de datos.

Pienso, a su vez, que actualmente vivimos en la época de la (des)información ya que es cierto que existe una cantidad ingente de datos generada cada día, pero, ¿cuántos de estos datos son aprovechables? ¿De verdad diferenciamos el valor de entre el ruido?

Por ello he decidido dedicar este post a explicar un poco cuáles deberían ser las funciones de uno de los perfiles profesionales más buscados en la actualidad, el data scientist o científico de datos. Además, tiene bastante que ver con el trabajo que estuve desarrollando durante mi proyecto fin de grado. Por tanto, trataré de definir a este profesional y al entorno al que se enfrenta desde mi experiencia en el ámbito.

Lo primero, por tanto, es definir el perfil en sí. Un Data Scientist es un profesional que traduce grandes volúmenes de información proveniente de distintas fuentes y las convierte en respuestas. Es decir, el que sabe discernir los datos que tienen valor y son aprovechables de los que no. Y, además, es capaz de sacarles partido a posteriori.

Teniendo esto en cuenta, aún nos faltaría abordar una de las partes de esta definición, los grandes volúmenes de datos conocidos comunmente como Big Data. Este termino hace referencia a las tareas relacionadas con ingentes cantidades de datos e información provenientes de distintas fuentes. Las principales características del Big Data son comúnmente conocidas como las 5V’s:

- **Volumen:** hace referencia al enorme volumen de datos e información que se maneja.
- **Variedad:** Hace referencia a la gran cantidad de fuentes de datos e información utilizadas, debido a ello surgen problemas con la estructura y formato de los datos.
- **Velocidad:** hace referencia a que se necesita tratar y procesar todos estos datos en el menor tiempo posible para que las respuestas o conocimientos sigan siendo válidos y no hayan quedado desfasados.
- **Veracidad:** hace referencia a la necesidad de que todos los datos e información utilizada sean veraces y se deben desechar los incorrectos.
- **Valor:** esta es la característica más importante ya que hace referencia al valor que tiene el conocimiento extraído de los datos para las personas encargadas de tomar las decisiones.

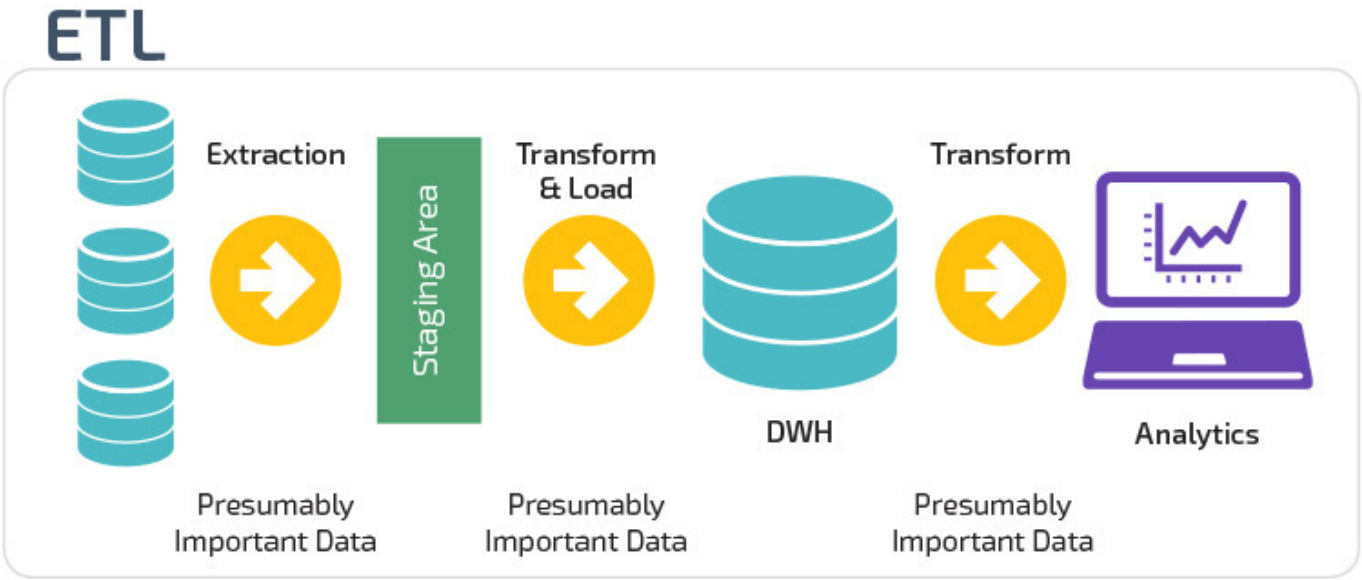
Este es el enfoque tradicional de las características del Big Data, pero actualmente se habla de 7, que son estas 5 añadiéndoles otras 2. Estas 2 son:

- **Viabilidad:** el Big Data es una herramienta fundamental para determinar la viabilidad de la empresa.
- **Visualización de los datos:** hace referencia a la presentación de los datos para ser fácilmente entendibles por las personas a quienes están dirigidos.

Una vez descrito en que consiste el Big Data, que es el ámbito en el que se desenvuelve el Data Scientist, podemos especificar las tareas necesarias para procesarlo y que son realizadas por este profesional.

- **Extracción de los datos:** la primera tarea a realizar será tanto la búsqueda de fuentes de datos e información, como la extracción de los datos desde estas fuentes.
- **Limpieza o curación de datos:** en esta tarea se deben de revisar los datos para evitar problemas con la estructura de estos, desechar datos incorrectos o sin relevancia y solucionar problemas que puedan causar los datos nulos.
- **Procesamiento de datos:** se utilizan métodos o algoritmos estadísticos para extraer información e incluso conocimiento de los datos que ya han sido curados.
- **Rediseñar estructura:** si fuese necesario, se volvería a editar o rediseñar la estructura de los datos, por ejemplo, se podrían añadir nuevos campos para dotarlos de mayor valor.

Estas tareas también pueden entenderse como un proceso de ETL que consiste en la extracción de datos, transformación de datos y su posterior carga. Además, al realizar el procesamiento, se llevaría a cabo la fase que se suele aplicar después del proceso ETL, que es el análisis.



El análisis descriptivo busca entender el perfil general del grupo de datos observado, una de las formas de llevarlo a cabo es mediante visualizaciones fácilmente comprensibles de conjuntos de datos complejos. Por ejemplo, mapas enriquecidos.

El análisis por inferencia busca obtener conclusiones para un grupo mayor a partir de una pequeña muestra analizada.

La principal diferencia de estos dos métodos es que el análisis descriptivo no busca hacer

generalizaciones mientras que el análisis por inferencia sí.

Una vez definido el perfil y sus funciones o tareas, según el proyecto que yo estuve desarrollando, me gustaría hablar un poco sobre la profesión en sí. Me parece que este perfil tiene unas oportunidades en el mercado actual y futuro inigualables por ningún otro perfil. Gracias a esto, estos perfiles tendrán la capacidad y libertad de elegir el lugar de trabajo que más les guste y no solo por la retribución.

Muchas empresas actualmente se lanzan a proyectos de analítica de datos o de tratamiento de datos por el mero hecho de usar estas tecnologías. Este enfoque tieneunas altísimas probabilidades de estar abocado al fracaso. Uno de los puntos más importantes a la hora de abordar un proyecto de estas características es tener muy claro el objetivo que se persigue y luego tratar de alcanzarlo. Por ello, se hace tan necesario este perfil profesional.

Por otra parte, me gustaría recalcar que este no es un perfil sencillo de encontrar, tiene que tener grandes conocimientos acerca de estadística y programación entre otras. Por esta dificultad debido a los grandes requisitos exigidos, lo más extendido es no tener a una única persona para cubrirlo, sino a distintas personas que tengan una base común pero estén especializadas en cada uno de los requisitos de este perfil ejerciéndolo como un equipo.

Dentro de poco se podrá ir viendo en el mercado laboral como se cubren estos perfiles y que skills son las que las empresas realmente demandan.