# Anatomía, Fisiología y Patología del Big Data
## Reflexiones informático-clínicas en tres tomos

» **Guillermo Leale**
*Doctor en Ingeniería en Sistemas de Información, Investigador en Minería de Datos y Big Data, Secretario académico UAI - Facultad de TI Sede Rosario*

### Introducción

El presente artículo tiene como objetivo hacer un análisis sobre los principios, funcionamiento y desafíos futuros del Big Data, desde la perspectiva de la Minería de Datos, haciendo una descripción informático-clínica en tres tomos. El Tomo I versará sobre la anatomía del Big Data, en tren de buscar una forma de diseccionarlo en sus conceptos fundamentales. Por su parte, el Tomo II será una descripción de su fisiología, para poder descubrir su funcionamiento a la luz de las tareas de la Minería de Datos. Finalmente, el Tomo III pondrá la lupa sobre las patologías del Big Data, entendiéndolas como nuevos desafíos a ser superados con las nuevas Tecnologías de la Información.

### Propedéutico

¿Para qué un propedéutico? Una buena respuesta recursiva es "para saber lo que hay que saber, antes de lo que hay que saber". Vayan aquí entonces un par de conceptos necesarios que aclaren nociones básicas antes de hablar del Big Data.

### Minería de Datos

En los últimos años, mucho se ha hablado acerca de la Minería de Datos. Este concepto surge como una excelente metáfora que pone al minero en el centro de la escena, con la cualidad de poder extraer, con el uso de sus herramientas y su conocimiento, las escasas pepitas de oro presentes en grandes montañas de piedra. De la misma manera, el "minero de datos", trata de encontrar información valiosa que a priori se encuentra "oculta" —no aparente a simple vista— en la "montaña" de datos, utilizando sus "herramientas" de análisis y su experiencia.

Es interesante notar que, a partir de la definición misma de la Minería de Datos, no se soslaya la figura de la "montaña de datos" en la cual buscar, lo cual indica desde un principio que estamos pensando en volúmenes grandes de datos. Esta idea, desde luego, es un principio fundamental para poder pensar en contextos de Big Data, como veremos más adelante.

Una definición formal de la Minería de Datos se presenta a continuación. Se trata de un "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos" (Fayyad, Piatetsky-Shapiro & Smyth, 1996: 37). Una manera simple e intuitiva de pensar la Minería de Datos es concebirla como un proceso para enfrentar problemas que involucran datos, y que no pueden ser resueltas a mano, o utilizando código simple de programación.

## Bola de cristal

La intuición presentada anteriormente añade un segundo principio fundamental para tener en cuenta, cual es el de pensar en algoritmos, en términos de códigos de programación que ejecutan tareas determinadas. Dentro de diversas clasificaciones, una útil es pensar en algoritmos predictivos y descriptivos.

Los algoritmos predictivos, a la manera de una "bola de cristal" informática, predicen un valor determinado a partir de sucesos que ocurrieron anteriormente. Pueden encontrarse ejemplos exitosos de algoritmos predictivos en la detección de fraude en tarjetas de crédito; en recomendaciones de ítems que pueden llegar a gustarnos para ver, escuchar y comprar en los servicios de Spotify, Netflix y Amazon respectivamente; y en las Redes Neuronales artificiales, que pueden aprender desde cómo se comportan las tendencias de la bolsa hasta qué diagnóstico tentativo posee un paciente a partir de una radiografía. Existen variadas aplicaciones, todas con el denominador común de predecir valores de una forma rápida, eficiente y "casi humana", con una tasa de error cada vez más pequeña. Quizá ciertamente mejor que la predicción humana hecha a partir de, precisamente, una bola de cristal.

## Buscando telarañas

Las telarañas no son siempre visibles a simple vista. A veces los hilos que las componen son muy finos. No obstante ello, pueden formar estructuras fuertemente unidas y resistentes. Tomemos esta analogía para pensar en grupos de personas, con relaciones no siempre aparentes, como por ejemplo con respecto a sus opiniones sobre determinado tema. ¿Qué hilos tejen las relaciones entre las personas en este sentido? ¿Son estas relaciones, siguiendo la analogía, apenas visibles pero al mismo tiempo resistentes? Los algoritmos descriptivos buscan la visibilización de estas relaciones que a priori parecen ocultas para el observador. Cuando se toman en conjunto muchas variables que definen a los individuos, pueden establecerse relaciones complejas que dejan ver los hilos subyacentes. Los algoritmos que exploran estas relaciones en particular se denominan de "clustering" o agrupamiento. De esta manera pueden encontrarse grupos ("clusters") de usuarios con intereses u opiniones comunes en las redes sociales; grupos demográficos de interés con características singulares; o bien grupos de clientes con características particulares y potencialmente valiosas para una empresa. De esta manera podemos descubrir el tejido real de la telaraña de relaciones subyacente.


*Revista Abierta de Informática Aplicada*/vol 2 n1 (2018): 39-46
ISSN 2591-5320

## ¡Puede fallar!

Hablemos de dos personas en concreto. Al final de este apartado escribiré sus nombres. Por ahora, sepamos que ambos nacieron en 1948; ambos son (inmensamente) ricos; ambos no tienen jefe, o sea que no son empleados; ambos residen en Londres; ambos comparten los gustos de viajar, tener (y disfrutar de) muchos perros y conducir caros autos deportivos; ambos tienen hijos y además se casaron más de una vez. ¿Muchas coincidencias hasta ahora? Debo agregar que ambos son llamados "príncipes". Uno es Carlos, el Príncipe de Gales. El otro es Ozzy Osbourne, el "Príncipe de la Oscuridad".[1] ¿Cómo es esto posible? ¿Cómo pueden ser tan similares y a la vez tan distintos? Una respuesta sensata es que las singularidades ocurren, y además los algoritmos no son permeables a ellas. Quizá un algoritmo descriptivo que tomara en consideración esos aspectos personales concluiría que ambos príncipes formarían parte de un grupo con características similares. Esto puede ocurrir, y es incorrecto, pero no resta eficacia al algoritmo. ¿Por qué?

Es necesario entender que los algoritmos tienen una muy fuerte base matemática y estadística. Desde esta base, los resultados deben pensarse en términos de promedios y tasas de error. De esta manera, puede ser que un resultado puntual sea erróneo, pero en el total de datos, el promedio de errores será lo más pequeño posible. Por esto es importante remarcar que el resultado buscado debe ser estadísticamente válido y útil para los objetivos planteados, y no necesariamente en los (a veces naturalmente extraños) casos particulares.

## Tomo I. Anatomía

En este apartado hablaremos de Anatomía. Diseccionaremos al Big Data y trataremos de entender sus piezas. ¿Qué es en realidad el Big Data? ¿Quién lo hace, y cómo? Un buen punto de partida es comenzar planteando tres negaciones: no es "grandes datos"; no es sólo "grandes"; no es sólo "datos". Estas ideas son intuitivas, ya que si el Big Data simplemente fuera alguno de esos tres conceptos, bastaría con traducir las palabras al castellano. No obstante, el Big Data engloba una serie de características particulares que comentaremos a continuación.

### *El asado*

Voy a contar en este apartado una sabrosa historia, en parte cierta y en parte ajustada al objetivo de este artículo. Un soleado domingo de verano, invitamos a mi familia política en pleno a comer un asado en mi casa. Son unas veinte personas, con lo que tuve que comprar varios kilos de carne. Pero eso no es todo. Cada uno tiene un gusto particular para comer. Mi suegro, por ejemplo, es un purista del asado. No puede comer solamente carne, debe agregar además embutidos y achuras. Por su parte, mi cuñado está haciendo una dieta especial que excluye las carnes rojas, por lo que para él asamos un pollo. Mi cuñada es vegetariana, lo cual hizo sustituir para ella las carnes por una buena cantidad de verduras asadas. Con estos detalles, además era necesario que todos comiéramos a una hora razonable, y preferiblemente que comenzáramos todos al mismo tiempo, o sea, no dejar a ningún comensal esperando su plato.

En este punto es apropiado comentar que lo que se estaba cocinando en la parrilla para esta historia no era un simple asado de domingo. En particular, tenía que prepararse para muchas

1 Tomado de https://www.bbc.com/news/technology-37307829.

personas, y cada una de las piezas llevaba distinta preparación, así como diferente tiempo de cocción. Luego, era necesario que se preparara rápido, o al menos en un tiempo razonable. Finalmente, y por sobre todo: debía tener "gusto a asado", esto es, en el paladar debía seguir teniendo el mismo gusto que -ahora sí- un simple asado de domingo.

¿Qué tiene que ver el asado con el Big Data? Pensemos en la situación de la casi ficticia historia contada más arriba. Preparar este asado es un tipo de problema diferente a preparar un asado común. Lo mismo ocurre justamente con el Big Data. Podemos esbozar una definición enunciando que es un *tipo de problemas* que provee un *marco de referencia* con ciertas características particulares. Hacer un asado para mi familia política en pleno es un tipo de problemas con un marco de referencia distinto a preparar un simple asado para dos personas. De lo que concluimos un principio importante: el Big Data no refiere a un objeto, ni a una tecnología, sino a un tipo de problemática puntual, diferente a una situación cotidiana común. Hablaremos entonces de una *problemática* o un *contexto* de Big Data.

### *V de Big*

Todo contexto de Big Data tiene una serie de cualidades que mnemotécnicamente se denominan "las V's". De acuerdo a la bibliografía, hay hasta 7 conceptos que comienzan con esa letra, y que describen distintos aspectos. Los autores teóricos más sólidos (Gantz & Reinsel, 2011: 1-12; Hashem, Yaqoob, Anuar, Mokhtar, Gani & Khan, 2005: 98-115) señalan solamente 4 de estos conceptos, a saber: Volumen, Variedad, Velocidad y Veracidad. El volumen, desde luego, es inherente a toda problemática de Big Data, dado que cuando la cantidad de datos es demasiado grande surge la necesidad de considerar la opción de trabajar en otra escala, de forma diferente a la convencional. Luego, si al volumen agregamos la variedad, significa que nos enfrentaremos con datos complejos y no estructurados. Ejemplos de estos datos son los textos, las imágenes, los videos y el *streaming* (transmisión de multimedia en tiempo real), o cualquier combinación de estos con los datos estructurados tradicionales. A estas dos V's debemos tratarlas con velocidad, esto es, en un tiempo razonable. Idealmente es deseable que el tiempo sea similar a los problemas tecnológicos comunes. Finalmente, una buena metáfora para la veracidad es el ya mencionado "gusto a asado": los resultados del trabajo en un contexto de Big Data deben ser tan confiables como los que estamos acostumbrados a obtener en los contextos tradicionales.

Como comentario adicional, las otras tres V's que mencionan los otros autores bien pueden enmarcarse en este último concepto de veracidad: *valor* agregado para los resultados obtenidos; *viabilidad* del proyecto Big Data en términos de posibilidad cierta de aplicación; y *visualización*, en el sentido de comprensibilidad de los resultados para los interesados.[2]

Una última definición que continúa la idea inicial planteada en el apartado de Minería de Datos es la siguiente. Si este proceso resuelve problemas que no pueden atacarse con simple programación, entonces el Big Data, por su parte, resuelve situaciones con características particulares que escapan a las técnicas tradicionales de Minería de Datos, y por lo tanto no pueden ser resueltas por éstas.

### *Mejor con un ejemplo, primera parte*

Pensemos en la red social Twitter. Tomemos por caso el ingenioso primer tuit de la CIA en su cuenta oficial, publicado en 2014: "no podemos confirmar ni denegar que este es nuestro primer

2 https://impact.com/marketing-intelligence/7-vs-big-data/.

tuit",[3] haciendo un juego de palabras con el cliché de los guiones de las innumerables películas donde la entidad norteamericana tiene un papel protagónico. Este tuit tiene, a la fecha de este manuscrito, más de 246.000 "me gusta" y más de 310.000 retuits. Tiene además cerca de 26.000 respuestas. Estos números, que parecen muy grandes, son no obstante muy comunes para la vorágine de información que circula actualmente en Twitter en particular, y en las redes sociales en general. Poder procesar los tuits y el contenido multimedia, interpretar los textos, encontrar grupos de personas con intereses comunes o detectar noticias falsas —con estos volúmenes de datos y estas características del contexto— son todos desafíos que claramente constituyen problemáticas de Big Data.

## Tomo II. Fisiología

En este apartado hablaremos de la fisiología del Big Data. Entendamos a la fisiología como la comprensión del comportamiento y la dinámica del Big Data. Tengamos en cuenta entonces que el tiempo es un factor fundamental bajo esta definición. El tiempo, además, aquí también es dinero, como lo indica la conocida frase. No solo por la ganancia o pérdida potencial para el interesado, sino por el costo que involucra una mala decisión en contextos de Big Data. Veamos un ejemplo a continuación.

### *Mejor con un ejemplo, segunda parte*

Quiero compartir en este apartado algunas ideas sobre un trabajo de investigación que estoy realizando a partir de un conjunto de datos del sitio Web Quora.[4] En este sitio, un miembro de la comunidad de usuarios puede realizar preguntas y recibir respuestas a partir de expertos dentro de la misma comunidad. Nuestro conjunto de datos consta de cerca de 400.000 pares de preguntas en idioma inglés, y un marcador por cada par que indica si ambas preguntas significan lo mismo, o no. El objetivo de la investigación es lograr un algoritmo que identifique eficazmente preguntas similares, de tal forma de acortar el tiempo que un usuario tarda en encontrar la respuesta deseada. Si es posible identificar una pregunta similar a la que formuló un usuario, entonces puede presentarse al usuario la respuesta correspondiente, mientras este espera que un experto responda a su pregunta. De esta forma, la respuesta encontrada tiene buenas chances de ser justamente la que el usuario estaba buscando, resolviendo así el problema de la espera.

Uno de los puntos de esta investigación es poder cuantificar de alguna manera la similitud entre todos los pares de preguntas. Existen métodos conocidos para medir similitud entre textos, que asignan un valor entre 0 y 1, siendo 1 un valor que indica la máxima similitud y 0 la máxima diferencia. Suponiendo que trabajamos con todas las preguntas del conjunto de datos de prueba (800.000 preguntas individuales, dado que originalmente contábamos con 400.000 preguntas en conjuntos de a dos), la cantidad de valores de similitud posibles entre todas las preguntas es cerca de 320.000.400.000 (trescientos veinte mil millones cuatrocientos mil). Si este número genera vértigo, aún mucho más impresionante es el procesamiento que esto implica. En una computadora actual de última generación, hicimos la estimación del tiempo necesario para calcular esta cantidad de valores, y el resultado es del orden de los… ¡tres años! Esto muestra a las claras que es imposible atacar problemas como este con métodos comunes. Aquí es donde hace su entrada triunfal el Big Data como problemática y como marco de trabajo.

3 https://twitter.com/cia/status/474971393852182528.
4 www.quora.com.

### *Pensar diferente*

De acuerdo a lo que venimos planteando, atacar un problema como el mencionado debe generar un cambio de enfoque. Pensemos que si tuviéramos la mejor computadora que el dinero puede comprar, aún podríamos enfrentar tiempos de trabajo extraordinariamente largos. Por lo tanto, pensar solamente en tecnología no resuelve el problema. ¿Entonces qué hacemos?

Hay que pensar diferente. Esto implica considerar nuevas estrategias para las tareas de captura, procesamiento, almacenamiento y análisis de los datos. En términos de proceso, debemos correr el centro de gravedad de los resultados a las hipótesis, y entender el problema lo antes posible al comienzo del proyecto. Una buena forma de ver esto es pensar en un gurú imaginario que podrá responder correctamente una pregunta si está bien formulada, y si no lo está recibiremos una respuesta carente de valor, y perderemos la oportunidad de obtener la sabiduría. El método elegido, la estrategia planteada, la forma de encarar el problema son las herramientas que simbolizamos con esa pregunta que debemos preocuparnos en formular correctamente.

Por ejemplo, si tomáramos el caso de Quora mencionado más arriba, invertir en una computadora muy cara podría ser una mala estrategia. Es posible prevenir esa situación haciendo estimaciones de cuánto puede llevar resolver el problema en términos de tiempo y costo. Por otra parte, una alternativa viable puede ser la computación paralela, poniendo a trabajar decenas o cientos de computadoras de bajo costo en vez de una sola computadora con mucho poder de cálculo. Esta última alternativa puede ser incluso mucho menos costosa que la primera, y puede llevarnos por el buen camino que conduce a la respuesta correcta. En suma, hay que pensar primero en términos de entender el problema, y luego considerar la tecnología.

## Tomo III. Patología

La patología es, en términos del concepto de enfermedad, una situación de desequilibrio de la que se debe salir. En términos de Big Data, podemos pensar a la patología como una circunstancia particular que plantea un desafío que se debe superar. Veremos aquí entonces las aplicaciones prácticas y los desafíos futuros de las problemáticas de Big Data.

### *Aplicación*

Presentaremos aquí una ilación de conceptos que constituyen los sistemas tradicionales y su enclave dentro del contexto de Big Data. En primer lugar, los *sistemas transaccionales* generan datos, reportes y constituyen el entorno operativo en el que las empresas desarrollan su trabajo de todos los días. Por otra parte, la *inteligencia de negocios* provee un reacomodamiento e integración de los datos transaccionales para que sirvan a los objetivos de los niveles gerenciales. Decimos que aquí contamos con información obtenida a partir de los datos. Finalmente, la Minería de Datos y el Big Data proveen algoritmos avanzados sobre grandes volúmenes de datos complejos. De esta forma, la información obtenida con la inteligencia de negocios se transforma en conocimiento para los niveles estratégicos de la organización. Sobre estos tres pilares —datos, información y conocimiento— es que se toman las decisiones importantes que condicionan y definen las acciones de la empresa.

### *Desafíos: tres ideas finales*

Como cierre de este artículo, a modo de resumen, quiero dejar tres ideas interesantes para tener en cuenta como desafíos planteados por el concepto Big Data. Cada una de ellas es parte de un momento del proceso de resolución de problemas en este contexto.

En el inicio del proceso, es bueno tener "cimientos sólidos". Es necesario considerar nuevas estrategias para la resolución, que abarquen todas las tareas tecnológicas tradicionales y las moldeen a la forma del marco de trabajo de Big Data. Es importante recordar que el entendimiento del problema y el establecimiento de bases claras para la resolución del mismo son muy importantes.

Una vez iniciado el proceso, puede ser ventajoso considerar la opción de un "ejército de máquinas". En este sentido, es conveniente echar mano a nuevas tecnologías de bajo costo para poder atacar estos problemas eficazmente en un tiempo razonable. Ejemplos de estas tecnologías son los clusters de máquinas y la computación en la nube.

Finalmente, en la culminación del proceso, es importante tener el concepto del "ojo de agua". Un ojo de agua es un área reducida de mucha profundidad dentro de una gran superficie de agua con baja profundidad, como un lago. Los lagos son entonces anchos y poco profundos, y esto es lo que normalmente provee la aplicación de la Minería de Datos: ocuparnos de pocos detalles en la vastedad de los datos. Por su parte, el ojo de agua representa la profundidad, en términos de particularidad y características singulares, de cada uno de los datos dentro del gran volumen original considerado. Estos dos conceptos están en constante tensión, para lo cual es necesario repensar los criterios de análisis de tal forma de lograr procesar inteligentemente cada vez más datos, pero procurar en todo momento no perder las individualidades y características particulares de cada uno.

## Conclusiones

En este trabajo se pretendió mostrar el concepto, funcionamiento y desafíos del Big Data a la luz de la Minería de Datos. Como palabras finales es menester reafirmar la idea de que es un concepto que representa un gran avance y tiene un futuro promisorio. No obstante ello, es un campo nuevo que representa un abanico de posibilidades de aplicación en datos que ya están disponibles. Por esta razón, debemos ser cuidadosos y tener presente permanentemente la precaución de poder controlar los alcances, de tal manera de aprovechar el avance del Big Data en forma beneficiosa para la sociedad

## Bibliografía

» Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). "From data mining to knowledge discovery in databases". In *AI magazine*, 17(3), 37.

» Gantz, J. & Reinsel, D. (2011). "Extracting value from chaos". In *IDC iview*, 1142(2011), 1-12.

» Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. & Khan, S. U. (2015). "The rise of 'big data' on cloud computing: Review and open research issues". In *Information Systems*, 47, 98-115.