Auditoría Pública nº 75 (2020), pp. 9 - 13

Francisco Julián Chico Martínez
Auditor de la Sindicatura de Cuentas de Cataluña

"Los datos recopilados desde los albores de la humanidad hasta 2003 son el equivalente al volumen que ahora producimos cada dos días". (Eric Schmidt. Google)

El *big data* mató a la estrella del muestreo

RESUMEN/ABSTRACT:

Este artículo hace una aproximación al concepto de *big data* asociado a la auditoría vigente. Para ello, después de definir los principales elementos que lo componen, se compara con el actual esquema de *small data*, que tiene como base el muestreo estadístico y el método científico, se plantean algunos de los requisitos a los que tendrá que hacer frente la Administración para conseguir su implantación y se exponen las principales conclusiones.

Se analiza también el posible fin del muestreo ante la utilización de los algoritmos propios del *big data*.

This article takes a look at the concept of Big data associated with present-day auditing. To do this, after defining the main elements that compose Big data, it makes a comparison with the current small data scheme, based on statistical sampling and scientific method, it discusses some of the requirements that government will have to face in order to achieve its implementation, and it then presents its main conclusions.

The possible end of sampling is also analysed, faced with the use of Big data algorithms.

PALABRAS CLAVE/KEYWORDS:

BIG DATA, SMALL DATA, MACHINE LEARNING, MÉTODO MAPREDUCE, AUDITORÍA, CORRELACIÓN, CORRELACIÓN ESPURIA, METADATOS, MUESTREO ESTADÍSTICO, ALGORITMO, GRANULARIDAD, NUBE

BIG DATA, SMALL DATA, MACHINE-LEARNING, MAPREDUCE METHOD, AUDIT, AUDITING, CORRELATION, SPURIOUS CORRELATION, METADATA, STATISTICAL SAMPLING, ALGORITHM, GRANULARITY, CLOUD

INTRODUCCIÓN. (*TIEMPOS MODERNOS)*

El *Big data* promete ser un cambio equivalente al descubrimiento de la rueda o de la imprenta. Es un cambio que, nos guste o no, ya está en nuestras vidas, en nuestros teléfonos, en nuestras ciudades, en nuestras casas. A efectos del *Big data* las personas somos entes productores de datos.

En la actualidad ya no hablamos de bites ni Megabyte (10^6 bites) sino de Quintillón de bytes (QB= 10^{30} bytes) cifras que la mente humana apenas puede concebir. No sólo se trata de la cantidad de datos sino el soporte para contenerlos. Al respecto, Antonio Monleón-Getino[1] asegura que toda la música del mundo se puede introducir en un disco duro que cuesta unos 500 euros y que el 90 por ciento de los datos del mundo han sido creados en los últimos dos años.

A pesar de que se abren muchas puertas por la gran cantidad de datos que se generan y que se recopilan, éstos, por sí solos, no sirven de casi nada. Para que los datos se conviertan en información o conocimiento éstos se tienen que procesar para poder resolver respuestas a preguntas derivadas de éstos: es decir se tienen que convertir en útiles.

Por otra parte, el tener muchos datos supone adquirir mayor carga de trabajo ya que, si bien algunos problemas se eliminan mediante el *Big data*, otros no, y por tanto se pueden heredar los defectos de la Estadística a una escala mayor. "Hay muchos problemas asociados a la Estadística tradicional que también se dan con grandes cantidades de datos. No desaparecen porque estés analizando gran cantidad de información, se vuelven peores"[2]. El problema será cómo inferir lo que está ocurriendo y determinar las medidas para cambiarlo.

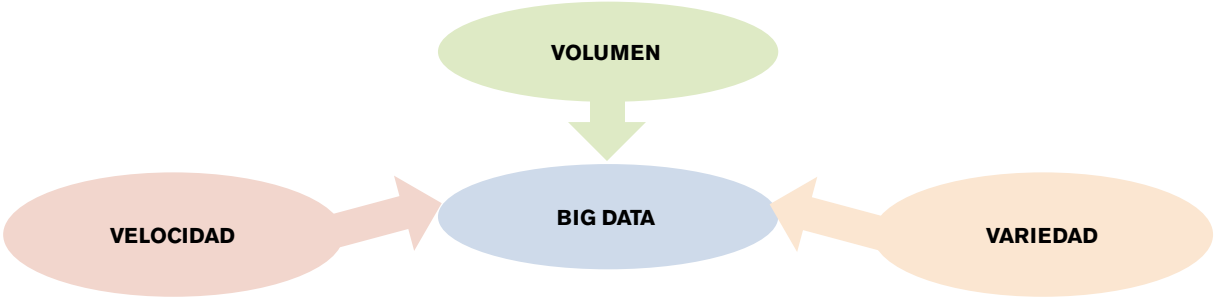
Pero ¿qué es el *Big data*?. Pocos tienen una idea clara de lo que significa, más allá de que se refiere a una gran cantidad de datos. El *Big data* es un concepto mucho más complejo que esa simple idea de volumen. Para una aproximación al concepto, el analista Douglas Laney's en su definición ya clásica del *Big data* lo relaciona con las tres uves (actualmente se agregan más "uves" a la definición, como veracidad, valor, etc.):

Volumen de datos que necesitan ser procesados y analizados a fin conseguir conocimiento útil y valioso

Velocidad en el procesamiento para conseguir unos tiempos de respuesta adecuados.

Variedad de estos datos.

Ilustración 1. Las uves del *big data*, según Douglas Laney's



Por su parte, la Wikipedia establece que los macrodatos (*Big data*) "es un término que hace referencia a conjuntos de datos tan grandes y complejos como para que hagan falta aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente. El concepto de *Big data* tiende a referirse al análisis del comportamiento del usuario, extrayendo valor de los datos almacenados, y formulando predicciones a través de los patrones observados".

EL BIG DATA NO TRATA SÓLO SOBRE EL VOLUMEN DE DATOS SINO SOBRE LA INFORMACIÓN QUE DE ELLOS PUEDE DESPRENDERSE E INTERPRETARSE

Hemos de tener en cuenta que para que los datos se conviertan en conocimiento útil y valioso cobra importancia el concepto de *Machine-Learning* (ML).

El ML lo definieron Kovahi y Provost[3] como "una disciplina proveniente de la ciencia y la ingeniería que se ocupa de la construcción y el estudio de algoritmos capaces de aprender a partir de datos. Es decir, se intenta realizar predicciones o toma de decisiones a partir de modelos construidos por los propios algoritmos en lugar de seguir instrucciones de manera explícita para lo que han sido programados[4]".

[1] El impacto del Big data en la Sociedad de la Información. Significado y utilidad. Antonio Monleón-Getino. Universidad de Barcelona. Noviembre de 2015.
[2] Big Data: *are we making a big mistake*. Hartford, Tim: Financial Times 28/3/2014.
[3] Ronny Kohavi y Foster Provost. *Applications of Data Mining to Electronic Commerce.* Kluwer Academic Publischers. 2001.
[4] Antonio Monleón-Getino Universidad de Barcelona. El impacto del Big data en la Sociedad de la información. Significado y utilidad. 2010.

De manera esquemática, el tratamiento de la información del *Big data* se realiza a través del método *MapReduce* que consta de dos fases:

1. Fase *Map* en donde se dividen y distribuyen en paralelo los datos en conjuntos más pequeños.

2. Fase *Reduce* en donde se combinan los resultados obtenidos en la fase *Map* para obtener un resultado mediante algoritmos.

SMALL-DATA FRENTE AL *BIG DATA*. (¡BIENVENIDO MISTER MARSHALL!)

En diversos foros especializados se da por sentado que la aparición de las técnicas del *Big data* y su aplicación en la auditoría supondrán la supresión inminente del muestreo, de manera equivalente a la película de los Inmortales en la que "sólo puede quedar uno".

El razonamiento, tan simple como equivocado, es que si con el *Big data* se pueden obtener todos los datos, no tiene sentido que se utilice el muestreo. Es decir, si N es la población y n es la muestra, con el *Big data*:

$$n \equiv N$$

Por tanto, según esta hipótesis, las técnicas estadísticas y de muestreo estarían de más.

Se aventura a decir, incluso, que con el *Big data* se encontraran correlaciones que permitirán pronosticar la reacción de cualquier variable a partir de criterios empíricos, sin conocer o pretender conocer las causas del fenómeno, o sea será el fin del método científico.

Quizás, en un futuro no muy lejano, los auditores públicos, gracias a algoritmos, podremos, como si de un film de Spielberg se tratara (*Minority Report*), detectar las salvedades y los fraudes antes de que se produzcan, pero de momento creo que eso está en el terreno de la ciencia ficción.

LA APARICIÓN E INTEGRACIÓN DEL BIG DATA EN LA FISCALIZACIÓN PÚBLICA HARÁ QUE LA UTILIZACIÓN DEL MUESTREO SE REALICE DE UNA MANERA DIFERENTE Y MÁS EFICIENTE, PERO NO DESAPARECERÁ

Lo que sí que parece evidente es que la aparición e integración del *Big data* en la fiscalización pública hará que la utilización del muestreo se realice de una manera diferente y más eficiente, pero no desaparecerá.

La fiscalización del futuro estará más enfocada a cómo se han introducido los datos, que a los datos en sí, es decir: se potenciará más el muestreo por atributos para revisar los controles internos que han servido de base para la introducción de datos a cambio de disminuir las pruebas substantivas al ser más susceptibles éstas de ser sustituidas por algoritmos. Veamos en la tabla 1 algunas de las principales diferencias entre el *Small data* y el *Big data*.

Tabla 1. Comparación *Small data* y *Big data*

Small data	***Big data***
Características de los datos	
En el muestreo los datos están estructurados según un esquema probabilístico soportado estadísticamente.	Los datos en el *Big data* se pueden encontrar estructurados, no estructurados, inconsistentes y con posible ruido o distorsión.
Esquema	
El esquema que se sigue es definir primero lo que queremos conseguir y posteriormente determinamos si lo obtenido se adecúa a lo establecido según el patrón estadístico.	El esquema que sigue el BD es, primero recopilar los datos y después se determinan las preguntas que contestan esos datos. Los datos se generan antes incluso de saber incluso qué tipo de información estadística puede extraerse de ellos
Recursos utilizados	
Se intenta optimizar los recursos haciendo la muestra lo más reducida posible de acuerdo con el modelo estadístico utilizado	Al disponer de la mayoría o todos los datos, es más probable que se detecten nuevas correlaciones entre variables. Hay que tener en cuenta que no todas las correlaciones serán útiles y reales para ello será necesario un proceso de filtrado para eliminar correlaciones espurias[5].
Eficiencia	
La idea es que modelos simples basados en una gran cantidad de datos pueden resultar más eficientes que modelos muy complejos basados en una cantidad relativamente pequeña de información. (Peter Norving, Google) Los datos obtenidos por muestreo se interpretan según los límites de los niveles de confianza exigidos.	Los datos más grandes no siempre son los mejores ya que debemos comprender sus propiedades y los límites y pueden dar lugar a malinterpretación de resultados, recopilación de datos basura, etc.
Utilización	
Se diseña para obtener solución a un problema, la utilización en casos diferentes a los que se ha diseñado no siempre es adecuada.	La información puede ser reutilizada para temáticas diferentes con diferentes niveles de desglose (granulidad de la información obtenida).
Correlaciones	
Las correlaciones obtenidas por muestreo permite encontrar un número limitado de patrones.	Las correlaciones mediante el *Big data* permite encontrar patrones que bajo un enfoque estadístico tradicional sería imposible detectar. El *Big data* puede contener respuesta a cuestiones que no estaban formuladas cuando se produjo la información.

Fuente: Elaboración propia

[5] La relación espuria, es la relación matemática que, por la existencia de un factor de confusión, presume la existencia de un vínculo apreciable entre dos factores o datos, cuando, en realidad, resulta inválido cuando se examina objetivamente.

REQUISITOS DE IMPLANTACIÓN EN LAS FISCALIZACIONES. (*EL VIOLINISTA EN EL TEJADO).*

No obstante, a pesar o además de todo lo expuesto anteriormente, todavía quedan muchos escollos por salvar y, sin ser exhaustivo, algunos de los requisitos que, como mínimo necesitaría la Administración para poder asumir e implementar la era *Big data* en la auditoría y suprimir así el muestreo son, entre otros, los siguientes:

- **Soporte informático adecuado**. - El *Big data* en auditoría requiere un soporte informático elevado, con maquinaria conectada masivamente en paralelo y que sea suficientemente potente para albergar y procesar todos los datos. Este gasto considerable ha de ser afrontado a nivel estado o de Unión Europea y requiere a su vez posibles convenios interadministrativos.
- **Integridad de los datos**. - Como segundo paso se necesita que todos los datos estén disponibles digitalmente. Es loable el esfuerzo de la Administración digital, pero estamos lejos de tener tabulados, por ejemplo, completamente los datos correspondientes a la contratación administrativa o de subvenciones, no sólo los establecidos por norma sino los derivados de posibles infracciones y de cualquier otro tipo.
- **Fiabilidad de los datos**. - Si se dispone de datos, pero no existe una verificación del control de su producción, éstos podrían ser totalmente inútiles, tendremos un incremento del ruido o basura en los datos.
- **Interconexión de diferentes bases de datos, públicas y privadas**. - Para la detección de irregularidades no solamente sería necesario los datos expedidos por la propia administración fiscalizada sino también datos de terceros. Por ejemplo, en contratación, datos de las empresas que se presentan a licitación, de su personal, datos correspondientes a las incompatibilidades, datos de las licitaciones en otras comunidades, etc. En este punto es posible que chocáramos con el derecho a la privacidad.
- **Algoritmos complejos**. - Para poder interpretar y poder sacar conclusiones de los datos se tienen que elaborar algoritmos que permitan correlacionar los datos a los fines perseguidos en la fiscalización pública. La elaboración de esos algoritmos requiere una gran especialización y un gran coste económico debido a la complejidad de trasladar los esquemas de fiscalización a unos parámetros informáticos.
- **Actualización de los datos**. - No sólo es necesario tener tabulados e interconectados los datos, sino que éstos han de estar actualizados, por ejemplo, con la normativa aplicable o con los requerimientos del momento que se fiscaliza. Un algoritmo que considere que un dato que puede ser relevante en un determinado momento, puede relevarse intrascendente ante un cambio normativo.
- **Preparación del personal**. - La fiscalización en la era *Big data* requiere un perfil de auditor que tenga suficientes conocimientos multidisciplinarios para formarse en cómo pensar para detectar la información que contienen los datos. Esto requiere un cambio de mentalidad del auditor y requerirá unas nuevas formaciones y nuevas capacitaciones para los nuevos técnicos.
- **Verificación del control interno**. – Como ya se ha apuntado, posiblemente el muestreo por variables se reducirá con el *Big data* por la no necesidad de realizar tantas pruebas sustantivas, pero se incrementará el muestreo por atributos por la necesidad de verificar los controles internos respecto a la introducción y

CONCLUSIONES. *(EL CIELO PUEDE ESPERAR).*

Entre las conclusiones más importantes se destacan las siguientes:

- El *Big data* al servicio de la Administración permitirá una mayor eficiencia en el análisis de datos y podrá responder a preguntas que todavía no se han formulado.
- Tener los datos no es lo importante, es saber cómo usarlos y qué hacer con ellos, es decir, dar valor a esos datos.
- El disponer y manejar muchos datos implicará heredar los problemas a una escala mayor. La cuestión será cómo inferir lo que está ocurriendo y averiguar cómo podemos intervenir para cambiarlo.
- El muestreo y la estadística en la fiscalización no desaparecerán, pero tendrán un enfoque diferente focalizado más en cómo se han introducido los datos y su análisis, y menos en los datos en sí. Posiblemente aumentará el muestreo por atributos en la revisión de pruebas del control interno y habrá una posible disminución del muestreo por variables.
- El paso del *Small data* al *Big data* requiere un esfuerzo económico, formativo y temporal muy grande que, de momento, está lejos de la administración actual.

- Por último, y no menos importante, no hay que olvidar las consideraciones éticas y de privacidad que supondrá el manejo de datos de carácter individual y privado a estas escalas.

Parece ser que la *Crónica de la muerte anunciada* respecto al muestreo se quedará, por el momento, en un *remake* de *El cielo puede esperar*.

BIBLIOGRAFIA

Rosana Ferrero. La estadística en la era del *Big data* https://www.maximaformacion.es/blog-dat/la-estadistica-en-la-era-del-Big data/

Universia Argentina. ¿Cuál es la relación entre *Big data* y estadística? https://noticias.universia.com.ar/educacion/noticia/2018/10/18/1162126/cual-relacion-Big data-estadistica.html

Reina García, Carmen. Manager Data Scientist .Estadística descriptiva, explorando el *Big data*. https://www.icemd.com/digital-knowledge/articulos/estadistica-descriptiva-explorando-Big data/

Toca Rey, Gonzalo. Estadística, la abuela del *Big data*. La Vanguardia. Historia y vida 601.https://www.lavanguardia.com/historiayvida/edad-moderna/20191222/472338388835/estadistica-Big data-historia.html

Martinez Vidal, Miguel Ángel. *Big data* y estadística oficial. Estadística actuarial número 40.Añoi 2017.https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/imagen_id.cmd?idImagen=1106865

Salgado, David. *Big data* en la estadística pública: retos ante los primeros pasos. Revista economía Industrial número 405 https://www.mincotur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/405/DAVID%20SALGADO.pdf