Del Data Mining al Big Data

Luis Carlos Molina Félix

Director de Power Builders - Data Mining Solutions – México
Email: luiscarlos.molina@powerbuilders.com.mx

Resumen

Desde los 90's, en que Data Mining se consolidó como una tecnología de apoyo a la toma de decisiones, se han venido dando grandes avances, sobre todo en la parte de comunicaciones y manejo de grandes cantidades de información. Al día de hoy, nuevos retos surgen sobre todo en la parte de integrar la información de los dispositivos móviles a la ya gran Base de Datos. Dado las experiencias que se han tenido, se propone que se incorpore en los proyectos de Data Mining, un Modelo de Datos Analítico (MDA), que sin ser un Data Warehouse, pueda ser útil a los usuarios finales para una mejor explotación de lo que existe oculto en las bases de datos. Sin embargo, es importante adquirir las tecnologías y metodologías que Big Data ofrece para poder alcanzar esto. El objetivo es uno: Proveer al usuario de "toda" la información que existe, fácil de explotar, y que ayude a tomar decisiones más asertivas.

1. Los inicios de Data Mining

En el 2002, siendo Coordinador del Curso de Data Mining, en la Universitat Oberta de Catalunya (UOC) en Barcelona, España, me pidieron que escribiera un artículo relacionado a Data Mining (en español conocido como Minería de Datos). Lo titulé: "*Data Mining: Torturando los datos hasta que confiesen*"[i]. En aquella época fue un artículo muy citado, por la poca información al respecto que se tenía en español y por el gran potencial que esta tecnología ofrecía a través de varios ejemplos de diversos sectores. Lo que intentaba transmitir era en dejar claro que Data Mining, no era estadística, ni redes neuronales, ni visualización de datos, ni pronóstico, sino una tecnología orientada a los negocios y que mediante el análisis de grandes bases de datos iba en búsqueda de lo que se llama el conocimiento mediante la integración de un conjunto de técnicas.

Haciendo un poco de historia, para consolidar el término "Data Mining", se tuvo que pasar desde los 60's por *Data Arqueology*, *Data Dredging*, *Data Fishing*, *Data Snooping*, KDD (*Knowledge Discovery in Databases*), entre otros. A finales de los 90's era común ver el *Proceso de Fayyad* (ver Fig. 1), para usarlo como referencia para comenzar un proyecto de Data Mining, sin embargo, había problemas en las definiciones, por ejemplo para dimensionar el concepto "trabajar con grandes volúmenes de información", y acotar lo que era grande. También, cada investigador le daba una definición dependiendo del área de formación de la que provenía.

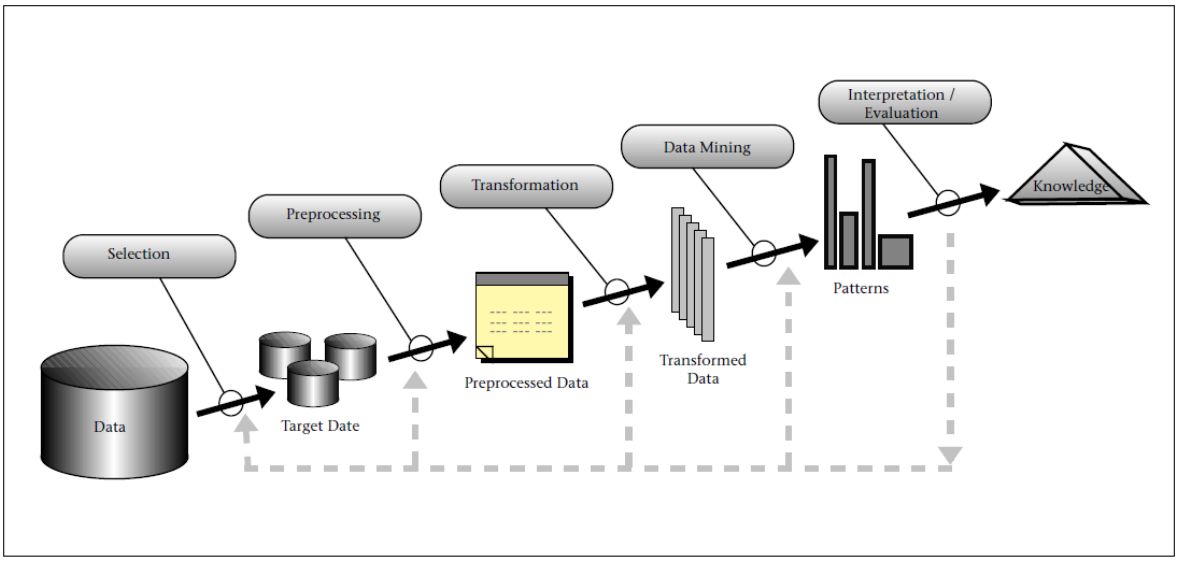

Figura 1. Pasos que contenían el Proceso KDD de Fayyad.

Desde mi punto de vista, en los 90's Data Mining tenía ciertos problemas, además de los tecnológicos:

- No había una metodología consolidada y completa.
- Se le vinculaba mucho como una etapa posterior y dependiente del Data Warehouse.
- Los proyectos tenía una fuerte dependencia del "gurú" que lo había desarrollado y solo él sabía lo que había hecho, ya que muchos proyectos no contaban con documentación detallada.
- Muchos de los proyectos realizados solo con estadística, pronóstico o inclusive Data Warehouse se vendía como de Data Mining.
- El especialista de Data Mining no tenía experiencia en mejoramiento de la calidad de datos.
- El especialista en Data Mining tenía poca experiencia en temas de negocio.
- Al momento de recibir los resultados, algunos directivos los hacían más como casos anecdóticos, que como verdaderos elementos de apoyo a la toma de decisiones.

Sin duda uno de los más importantes problemas de estos proyectos era la gran dependencia del "gurú", por lo que en muchas ocasiones cuando se necesitaba realizar nuevamente un estudio comparativo, muchas veces este no estaba disponible, y dado que casi no dejaba documentación sobre sus extracciones de SQL, causaba que no existiera continuidad en los estudios de Data Mining. Ante la falta de regularización de los procesos de Data Mining, un grupo de investigadores y empresas se reúnen y en 1999, aparece la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*)[ii] [iii] [iv], que se consolida en diferentes llamados del 2002[v], 2004[vi], and 2007[vii] como la más usada en un proyecto de inicio a fin, por lo tanto, estandariza, en gran medida, los criterios para establecer la estrategia de resolución a los proyectos de Data Mining.

Acompañado a esta metodología, diversos trabajos de Ron Kohavi, Tom Mitchel, Liu Huan, Hiroshi Motoda, Peter Shappiro, Heikki Mannila, John Dougherty, Jiawei Han, Micheline Kamber, Gregory Piatetsky-Shapiro [viii] [ix] [x] [xi] [xii] [xiii] [xiv] [xv], entre muchos otros, vinieron a demostrar los mejores métodos de muestreo, selección de atributos relevantes, algoritmos de clasificación, y de cálculo del error principalmente. Así que ya no había que inventar muchas cosas. La conclusión de todo esto es que son los propios datos, sus características y el

objetivo a alcanzar lo que nos va llevando a la técnica a usar, y que había ciertos métodos que tenían en lo general un mejor desempeño que otros que hacían lo mismo.

De ese entonces muchos eventos han sucedido: Mayor capacidad de almacenamiento (Cómputo en la Nube) y procesamiento; mejores herramientas analíticas con sorprendentes capacidades de visualización de datos; mejores herramientas para mejorar la calidad de datos; personal que realiza Data Mining con mejores habilidades de entender los negocios y; sobre todo, un cliente con cada vez mayores necesidades de analizar a profundidad sus datos para beneficiar al negocio.

2. La evolución de DM hacía el Modelo de Datos Analítico

Uno de los hechos relevantes es que Data Mining ya ha evolucionado a venderse más como concepto que como producto, por lo tanto, ahora se vende como mejora en la utilidad, propensión de fuga del cliente, perfilamiento del comportamiento de los defraudadores de tarjetas de crédito o mejora en el ajuste de parámetros en las herramientas de calificación de riesgo. Todo esto se resume, en que Data Mining se convierte en un concepto que incorpora en las diferentes áreas de organizaciones la práctica de la **Cultura Analítica**.

Para vender Cultura Analítica se debe de comenzar a diferenciar en una organización cuáles son sus procesos operativos y cuales sus analíticos. Los procesos operativos se refieren a todos ellos que trabajan en la continuidad del objeto del negocio, mientras que los segundos registran y miden el desempeño de ese objeto desde diversos aspectos. Una métrica simple es calcular en una organización las horas/hombre dedicadas a la operación y al análisis. En nuestras experiencias, encontramos casos de empresas mexicanas con una relación operación/análisis entre un 98%/2% y un 89%/11%, respectivamente. Desde mi punto de vista, y de acuerdo a los diversos tipos de negocios los rangos deben de estar entre un 80%/20% a un 60%/40%. El vender Cultura Analítica afecta de manera significativa toda la estructura de la empresa, por lo tanto, surge la importancia de tener una visión integral del negocio.

La Cultura Analítica dentro de unos de sus diagnósticos, estudia las actividades sin valor dentro de los procesos analíticos. Una experiencia sucedió en una institución bancaria, cuando la persona que analizaba casos de operaciones inusuales de depósitos bancarios, tenía que copiar las cuentas detectadas por un sistema y analizarlas en otro. Esto le consumía el 45% de su tiempo laboral en “copiar y pegar”. Al detectar esta actividad si valor, se desarrolló un programa que lo hacía en minutos, lo que le permitió, a la persona, realizar análisis de mayor profundidad e incorporar nuevas técnicas analíticas. Otro punto a incorporar, sin duda es la capacitación, sobre todo en estrategias para resolver los problemas analíticos y en el uso de herramientas de explotación de datos.

De igual manera que se ha aprendido a vender los proyectos, también ha habido ciertos aprendizajes citados a continuación:

- La Cultura Analítica debe de ser para el alcance de muchos. Se debe de apegar a metodologías analíticas bien documentadas, donde una persona con ciertos conocimientos técnicos, sin ser “gurú” las pueda entender. También los usuarios deberían de tener acceso a la explotación de la información con herramientas amigables con destacados componentes de visualización de datos.
- Un proyecto debe de hacerse inmune a quien lo diseña, por lo tanto, no debe de haber dependencia del “gurú”, de tal forma, que el proyecto analítico se pueda repetir a lo largo del tiempo.
- Se debe de tener habilidades que permitan mejorar la calidad de los datos.

- Se deben de dar resultados que impacten el núcleo del negocio, teniendo una visión lo más integral posible, por lo que el responsable del proyecto debe de involucrarse mucho en el negocio de la organización.
- El tener una certificación en el uso de herramientas de Data Mining no garantiza el éxito de un proyecto.

Adicionalmente, un elemento trascendental que ha surgido para cubrir las diversas necesidades actuales de las organizaciones, es el "**Modelo de Datos Analítico**" (MDA). Es un modelo generalmente bajo el esquema entidad-relación que tienen algunas diferencias respecto a un modelo de datos tradicional o un *Data Warehouse*. En un modelo tradicional, de forma general, se parte de determinar objetivos y alcances, entrevistas con los usuarios, mapeo de los procesos, definición de necesidades futuras, análisis de las fuentes de datos, así hasta diseñarlo, construirlo, probarlo, documentarlo, liberarlo, capacitar a los usuarios, dependiendo de la metodología usada. Un MDA, una vez definidos los objetivos y el alcance, se integran todos los elementos que permiten tomar de decisiones tanto a los directivos, como a los que están analizando la información. Se coloca como eje al usuario y sobre eso, se intenta proporcionarle la información que necesite mediante gráficas, tablas, reportes, indicadores, entre otros. Todo siguiendo las metodologías tradicionales, pero con insumos y estrategias diferentes (ver Fig. 2).

Por ejemplo, en un modelo tradicional se necesita definir en su contenido la dirección completa de un cliente, en cambio un MDA, solo necesitará de la Colonia, el Código Postal (CP) y variables hijas de este como CP2 y CP3, refiriéndose al Código Postal que contienen los 2 y 3 primeros dígitos, para garantizar que los algoritmos puedan consolidar por grandes grupos de localidades. También en un MDA, en la gran mayoría de los casos, el nombre del cliente no importará, sino únicamente algún identificador único. De igual forma no existe el concepto de hechos, ni dimensiones que tiene un *Data Warehouse*, aunque existen históricos, granularidad y metadatos.

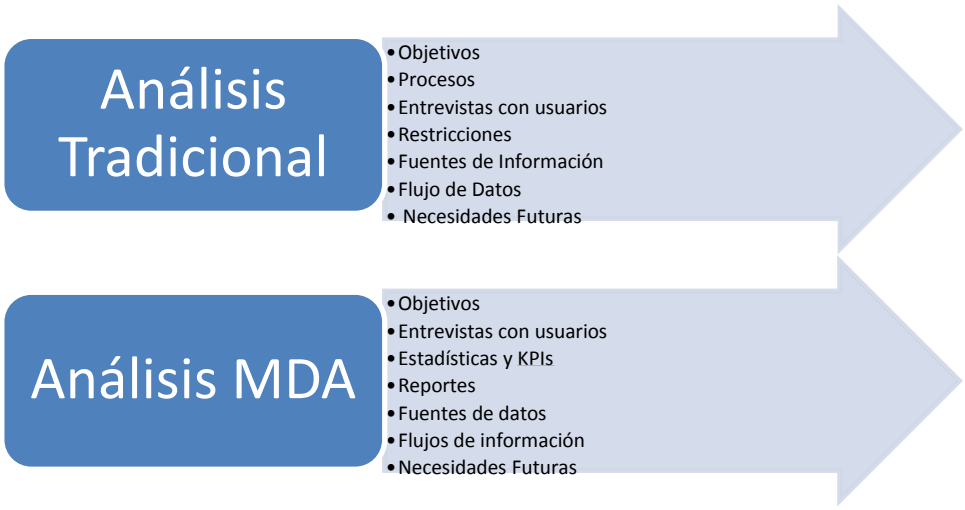

Figura 2. Comparación de la etapa de Análisis entre el Tradicional y el MDA.

Un aspecto importante dentro de un MDA es la política de nomenclatura de variables. Se deben de tener nombres que apoyen al usuario a entenderlas de manera intuitiva. Se deben de clasificar, así que esto permitirá saber cuántas variables pertenecen a catálogos numéricos, alfanuméricos, cuantas variables fecha se tienen, cuantas son indicadores, cuantas están relacionadas a montos, importes, etc. Existió un caso de

un *Data Warehouse* que tenía 14 formas de nombrar a la misma variable llave. Después de 2 años y medio de construirlo se decidió cancelar el proyecto.

Finalmente una vez construido el MDA, se debe de apoyar con poderosas herramientas gráficas de consulta de información para usuarios no informáticos. Esto permite que los usuarios de negocio puedan explotar la información al momento que la necesiten, sin tener dependencia de las áreas de TI. A su vez, una herramienta de Data Mining se conecta al MDA, lo que facilita aplicar las diversas técnicas, repetir los estudios a lo largo del tiempo para comparar los avances de la organización. Entre las técnicas más comunes se destaca árboles de decisión, reglas de asociación, clustering, métodos bayesianos, principalmente.

Cuando se implementó un MDA en una compañía telefónica, al principio se visitaron las diversas áreas de negocio para recopilar sus elementos de decisión y saber las necesidades de información que tenían. Se partió de poner al usuario en el centro para observar la propensión de abandono de los clientes, conocido en inglés como *churn*, y determinar qué elementos necesitaba para desempeñar su trabajo, además de la información de consumo que se le daba, era necesario saber cuántas campañas de promoción se le había hecho al cliente, cuantas quejas se tenían, como había evolucionado tecnológicamente en sus diversos teléfonos que había adquirido, si usaba su plan de datos para acceder a Facebook, o Twitter, cada cuando perdía su teléfono, entre otros. Cuando se presentó el proyecto al área de TI, se nos informó que lo que solicitábamos era imposible, pues cada información estaba en sistemas independientes. Afortunadamente, se pudo pasar sobre ese paradigma, gracias a los directivos que apoyaron el proyecto, de tal forma que sus primeros resultados incrementaron la respuesta usando las mismas campañas pasando de un 3% a un 30%, haciendo una mejor selección de clientes propensos a abandonar la compañía.

3. La informática no ha podido cumplir los requerimientos de los usuarios

Comenzaremos con la historia de importantes modelos de bases de datos tradicionales. En un banco, hace más de 10 años, se diseñó un modelo de datos con los productos tradicionales que ofrecía. Después comenzaron normativas regulatorias de los gobiernos, así que el modelo tuvo que sufrir modificaciones. Los ejecutivos pedían información nueva que no estaba considerada inicialmente en el modelo. El Director de Sistemas se fue a otro banco, y llegó uno proveniente de la industria farmacéutica. El banco sumó nuevos productos, y adquirió productos de otros bancos, por lo que hubo que integrarlos. Un nuevo Director de Sistemas fue incorporado. Hoy en día, el banco no tiene la visión global de su modelo de datos, faltan varios metadatos, y no tiene un control total de sus procesos. El usuario se encuentra ante el dilema, porque le gustaría saber cuantos productos tiene un cliente de los más de 40 que ofrece el banco y como los ha ido adquiriendo. Infelizmente varios productos tienen claves de cliente diferentes: las tarjetas de crédito con 16 dígitos, los seguros de vida con 12 dígitos, los créditos nómina de 8 dígitos. El usuario percibe que en algunas ocasiones el cliente incumple en algunos pagos a créditos y por otro lado le llegan campañas de promoción ofreciéndole nuevas líneas de crédito. También los saldos difieren en algunas ocasiones con lo real.

De igual forma al hacer un estudio de las actividades del usuario, se observa que el 20% de su tiempo la dedica a validar algunas soluciones parciales que le ha ofrecido el área de sistemas. El usuario le ha regresado los desarrollos informáticos en varias ocasiones debido a las inconsistencias de los saldos y ciertas reglas de negocio que han salido sobre la marcha, que no fueron consideradas.

El anterior relato es una historia real, y de igual forma aplica a muchas otras instituciones, donde el usuario no tiene la información suficiente para desempeñar sus actividades analíticas y solo se le dan pequeñas soluciones para responder a sus necesidades.

Este crecimiento de los modelos de datos de cierta forma poco planeado y más resolviendo el problema inmediato, ha originado que el tema de la calidad de datos tenga un papel importante. Aunque existen muchos conceptos para corregir esto tales como: higienización de datos, limpieza de datos, filtrado de datos, estandarización de datos, depuración de datos, hemos decidido llamarle procesos de mejora de calidad de datos, pero que sin duda es un aspecto fundamental antes de considerar hacer un proyecto de Data Mining.

La informática debe volcar sus esfuerzos en dar soluciones integrales analíticas, colocando como eje central al usuario y proveerlo de toda la información necesaria y fidedigna para cumplir con su actividad, lo que traería beneficios directos a las instituciones. Aquí surge el fundamento de lo que se le ha llamado *Big Data*[xvi].

4. El Big Data

A partir de una necesidad para proporcionar al usuario una visión analítica 360 grados sobre los clientes, los productos, los empleados, las transacciones que a su vez estén inter-ligados surge *Big Data*, conocido también por otros términos *Big Data Analytics*, *Value Data*, *Smart Data*, entre otros. Aunque existen muchas discusiones se hacen por el nombre, lo que es importante es atender una importante necesidad y que debe de contener varias palabras: almacenamiento y procesamiento masivo, heterogeneidad de datos, integración, fácil explotación de datos, análisis avanzado y data mining. En muchos textos al respecto se preocupan más por el tipo de almacenamiento de datos heterogéneos hablando de pentabytes, exabytes, zettabytes y yottabytes, sin embargo, eso no necesariamente responde a la necesidad del usuario.

En una compañía, el Director Nacional de Ventas me preguntó sobre un ejemplo del Big Data aplicado directamente hacía su área, yo les plantee un esquema del punto de vista *empleado* donde estuviera la fecha de ingreso de esa persona, su desempeño laboral, sus salarios, sus ausencias, la tecnología que ha ido manejando, sus clientes, sus ventas cerradas, su oportunidades de venta, su red social a la que pertenece en la empresa, su puesto en el organigrama, documentos donde aparece su nombre, su pronóstico de ventas y una clasificación del tipo de segmento de cliente que tiene mejor utilidad al momento de vender. Todo mostrado de forma integral y fácilmente consultable. De ahí mismo me puede llevar a la vista *cliente*, donde puedo obtener los diversos contratos que se le han hecho, ver la facturación, atrasos de pagos, quejas que ha hecho, utilidad, productos sugeridos para vender, entre otros. En varios casos, no se trata de adquirir nuevas tecnologías, sino de agregar e integrar la información de la institución.

Otro de los factores que pueden detonar el Big Data serían tomando en consideración una mejor explotación de la información que pueden enviar los sensores o dispositivos móviles, tales como información de camiones de transporte, de dispositivos de geolocalización, de tarjetas con antenas transmisora, estas últimas por ejemplo, pudiendo rastrear con la instalación de muchas antenas la navegación de los clientes en una tienda, o de sistemas más complicados como en la red de transporte público de la ciudad.

Junto con Big Data hay que proveer al usuario de herramientas para ejecutar acciones oportunas como respuesta del negocio. Aunque falta mucho camino para entender a los clientes, algo que se ha visto como un gran error que comente los ejecutivos es que si detectan algo en la red social, quieren atacar ese

mercado desde esa misma red social y eso no siempre es lo mejor. En su gran mayoría de los casos, las redes solo sirve para detectar patrones o grupos y deben de ser atraídos desde otros medios. En las campañas presidenciales de México en el 2012 existieron dos propuestas políticas (PRD-PT y PAN) que quisieron convencer a parte de ese electorado de su voto a favor desde las redes sociales, sin embargo, tuvieron un efecto contrario y acabaron llenándolos de infinidad de información, no toda fidedigna, y hasta con cierto grado de agresión no solo a los candidatos, sino a los cibernautas. En mi opinión concluí, en base a varias encuestas, que muchos electores indecisos acabaron no dando su voto a la propuesta política que enviaba mensajes agresivos en la red social en la que estaban. Cabe aclarar que muchos de estos mensajes no eran realizados por los estas las propuestas políticas, sino por cibernautas afines a ellas.

El equipo científico de Obama en las elecciones del 2012, dirigido por Rayid Ghani se dedicó a analizar los diferentes perfiles de electores en un lugar llamado "La Cueva". La situación al principio prácticamente era un empate técnico entre ambos candidatos, por lo que había que hacer cosas diferentes. Durante 18 meses unificaron todas las bases de datos que emplearon los equipos de campaña de Obama que le ganó a McCain en el 2008, en lo que podemos llamar el Big Data, combinando las redes sociales, listas de donantes, encuestas, las bases de datos del partido que determinaban sus preferencias políticas o la indecisión en cada estado de importancia. Entre las variables introducidas estaban: sexo, edad, raza, etnia, lugar de residencia, idioma, ingreso, tendencia política, historial de participación electoral, junto con aficiones, red de amigos, preferencias de consumo, la mayoría obtenidos del Facebook. Entre las cosas que encontraron los científicos fueron:

- El 20% de los que recibían un mensaje vía Facebook lo leían y lo mandaban a sus amigos. La acción fue diseñar una aplicación que transmitía mensajes muy bien estructurados animando a sus contactos a registrarse para algún evento donde Obama estaría presente.
- Se descubrió que en Florida era necesario convencer a las mujeres del condado de Dade de menos de 35 años, que les gustaban ciertos programas de televisión. La acción fue contratar publicidad en *Sons of Anarchy* y *The Walking Dead* que eran programas que la gran mayoría de ellas veía con frecuencia.
- Había un importante grupo de votantes indecisos en la red social Reddit. La acción fue que Obama se registró para interactuar junto con su equipo dentro de esa red.
- Se encontró un grupo de mujeres de la Costa Este estaban indecisas. La acción fue hacer un sorteo en esa región para ir a visitar, junto con Obama, a la actriz de *Sex and the City* Sarah Jessica Parker, nacida en 1965 y conocida como un referente en la moda.
- Se descubrió que las mujeres de entre 40 y 49 años de la Costa Oeste soñaban con tener una cena con George Clooney. La acción fue hacer un sorteo para cenar con Obama y el actor en Hollywood.

En la noche del escrutinio, Romney, vio cómo se iban sus estados como Ohio, Virginia, New Hampshire, Indiana, Colorado, Florida, Iowa a favor de Obama. Unas horas después la revista TIME[xvii] fue quien develó la existencia de "La Cueva". Actualmente el presidente Obama lanzó un proyecto en marzo del 2012 llamado "*The Big Data Research and Development Initiative*"[xviii]. La iniciativa está compuesta por 84 diferentes programas de Big Data distribuidos en seis dependencias.

5. El Reto Futuro

Las instituciones tanto públicas como privadas han hecho esfuerzos por conceptos como Cuenta Única, Cliente Único, Clave Única de Registro de Población (CURP), Registro Federal de Causantes (RFC), Documento Nacional de Identidad (DNI), entre otros. Sin embargo, por ejemplo, en los sistemas es común

hablar de la CURP16 o CURP18, para decir que la primera le faltan dígitos y en el segundo está completa. En un estudio de CURP de ciudadanos que asistían a escuelas del Distrito Federal, encontramos que un grupo importante de estos aparentemente habían nacido en el estado de Aguascalientes. Al buscar la fuente del problema, encontramos que el sistema de captura asumía el estado que aparece como primero y lo ponía por defecto, en caso de que no se hubiera llenado, en vez de haber puesto Distrito Federal por defecto. Por lo tanto debemos decir que el esfuerzo de cumplir con las claves de identificación únicas completas debe ser prioridad de las instituciones del gobierno, certificando que los sistemas puedan llenar de forma correcta y completa datos como la CURP.

De igual forma sucede con las direcciones, es necesario que cada predio tenga una dirección única, ya que resulta curioso, que la correspondencia que recibe un ciudadano de los distintos órganos del gobierno (agua, luz, predial, multas de tránsito, etc.) tienen direcciones diferentes, que varían principalmente en el nombre de la colonia y el código postal. La necesidad de que los municipios cuenten con catálogos de calles y colonias es primordial, así como establecer criterios para dar los nombres a las calles. Existe el caso de un municipio que decidió realizar una política de nomenclatura a sus calles, de tal forma que en lo referente a nombres de héroes o personajes distinguidos, se comienza por el nombre y termina por el apellido sin abreviaturas. Esto facilita bastante la forma de que cada predio tenga una dirección única y la correspondencia llegué realmente a donde tenga que llegar. Actualmente en México existen 2 organismos que indican como se debe de establecer una dirección, una es el Servicio Postal Mexicano (SEPOMEX) y el otro es el Instituto Nacional de Geografía, Estadística e Informática (INEGI). Sin embargo, en el tema de usar o no abreviaturas, por ejemplo, estas se contraponen, situación que se debe unificar.

Por otro lado, cada vez que se pide una factura fiscal, resulta que se debe de capturar todos los datos; siendo que si existieran sistemas más eficientes, el solo teclear el RFC, y solicitar a un centro del gobierno que llene los campos restantes en forma automática, eso ahorraría miles de horas/hombre anuales gastadas en esa actividad. Algunos me entenderán cuando piden una factura a un restaurante y esta llega a tardar más de 30 minutos y después de su revisión observamos que tiene errores.

Una estrategia para vender Big Data se refiere al plantearle situaciones de peligro a la compañía y ver si está preparada tecnológicamente para responder ante eso, por ejemplo, que información crucial de requiere al momento de que hubiera un accidente para definir acciones. Ante eso el posible cliente se da cuenta de que solo tiene información parcial ante un posible escenario y percibe inmediatamente la necesidad del Big Data, así como su necesidad de analizar toda la información en conjunto.

Para concluir, a lo largo de este artículo hemos hecho una revisión de como ha venido evolucionando las actividades analíticas a lo largo del tiempo, y como se han ido solventando problemas con la incorporación de metodologías, de una visión integral del negocio, del Modelo de Datos Analítico, y que esto a su vez está convirtiéndose en lo que se está llamando el Big Data. Hoy algunas de las empresas que vende está tecnología encabezan sus discursos de venta sobre almacenamiento masivo de datos heterogéneos, incorporando términos como *cómputo en la nube*, sin embargo, el verdadero poder de esta evolución tecnológica está en darle a los usuarios una información completa, integral y fidedigna, acompañada de elementos analíticos y de fácil explotación, que le ayuden a entender mejor el negocio, permitiendo beneficios como ahorros, identificación de nuevos perfiles, búsqueda de nuevos nichos de mercado, recomendación de nuevos productos, detección de fallas en los procesos, entre otros.

6. Bibliografía

[i] http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
[ii] http://www.crisp-dm.org
[iii] *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, P. Chapman, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, 2000.
[iv] *The CRISP-DM User Guide,* Pete Chapman, 1999.
[v] Gregory Piatetsky-Shapiro (2002); KDnuggets Methodology Poll.
http://www.kdnuggets.com/polls/2002/methodology.htm
[vi] Gregory Piatetsky-Shapiro (2004); KDnuggets Methodology Poll.
http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm
[vii] Gregory Piatetsky-Shapiro (2007); KDnuggets Methodology Poll.
http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm
[viii] *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, editors, AAAI/MIT Press, 1996.
[ix] *Machine Learning*, T. Mitchell, McGraw Hill, 1997.
[x] *Feature Selection for Knowledge Discovery and Data Mining*, Series: The Springer International Series in Engineering and Computer Science, Vol. 454, Huan Liu, Motoda, Hiroshi, 1998, XXIV, p. 214.
[xi] *Irrelevant Features and the Subset Selection Problem*, GH. John, R. Kohavi, K. Pfleger, Proceedings of the eleventh international conference on machine learning, p. 129 .
[xii] *Supervised and Unsupervised Discretization of Continuous Features*, J. Dougherty, R. Kohavi, M. Sahami, Machine Learning-International Workshop then Conference, p. 194.
[xiii] *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, R. Kohavi, International joint Conference on artificial intelligence 14, p. 1137.
[xiv] *Principles of Data Mining*, D. J. Hand, H. Mannila and P. Smyth, MIT Press, Fall 2000.
[xv] *Data Mining : Concepts and Techniques* , J. Han, M. Kamber, 2nd edition, Morgan Kaufmann, 2006.
[xvi] *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. J. Manyika, M. Chui, B. Brown, J.s Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, McKinsey Global Institute, May 2011.
http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
[xvii] http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/
[xviii] *Big Data Across the Federal Government*, Executive Office of the President, March 2012, White House. Retrieved 26 September 2012.

Luis Carlos Molina comenzó en Data Mining desde 1996. Sus estudios de maestría los realizó en la Universidad de Sao Paulo, en Brasil y de doctorado en la Universidad Politécnica de Barcelona, en España. Es autor del libro: “Data Mining – Una Introducción” FUOC 2000, y del artículo “Data Mining: Torturando los datos hasta que confiesen” UOC 2002. Ha trabajado de manera ininterrumpida en el tema con proyectos muy importantes que han sido presentados en Brasil, España, Japón, Portugal, Singapur y México. Las áreas principales que ha trabajado son Banca, Gobierno, Retail, y Telefonía Celular. Desde el 2005 es director de Power Builders, una empresa, localizada en México, que se dedica a dar soluciones analíticas y de limpieza de datos.