

Data Cleansing: La información sí puede ser fiable



Para asegurar un análisis certero de la información, ésta debe estar libre de errores. La fiabilidad en la toma de una decisión está sujeta a la calidad de la información que la sustenta y ésta no debe ser objeto de duda.

[David Álvarez Rodríguez .Gerente de Business Intelligence en Single Consulting]

¿Cuántas veces nos hemos encontrado con el caso de ser objeto de campañas de marketing por triplicado (Estimado Sr. David / Estimado Sr. Álvarez / Estimado Sr. David Álvarez)? Para la empresa que ha lanzado esta campaña de marketing el impacto es claro; se han triplicado los costes de la iniciativa para contactar al mismo prospecto. Pero, ¿y los costes “subjetivos”? ¿Se puede cuantificar el coste de un cliente molesto por haberle llegado 3 veces el mismo comunicado?

Muchos otros ejemplos ocurren también a diario dentro de las organizaciones a la hora de realizar análisis de la información. En entornos datawarehouse, con fuentes de datos diversas -a veces incluso externas a la organización- cruzar información o consultarla de forma consolidada en ocasiones se convierte en una tarea complicada. Datos relacionados con productos y/o clientes representados de múltiples formas y totalmente “desnormalizados”, con diferencias en mayúsculas/minúsculas, abreviaturas, errores tipográficos, guiones y caracteres especiales, etc., hacen de esta labor, una misión imposible...

En estos casos, la solución obvia sería tratar de realizar una limpieza previa de los datos objeto del análisis. La cuestión radica en “como” llevarla a cabo. Un proceso “manual” en grandes conjuntos de datos quedaría inmediatamente descartado, no sólo por el propio coste en horas/hombre, sino también por la propia naturaleza del proceso, propenso a nuevos errores. Una herramienta que automatice en mayor o menor medida esta limpieza de datos ayudaría a obtener un nivel razonable en la calidad de la información de forma eficaz.



Data Cleansing

La limpieza de datos (*data cleansing*) debe considerarse algo mucho más complejo que una simple tarea de actualizar registros con información correcta. Una limpieza de datos exhaustiva requerirá una descomposición, análisis y posterior montaje del conjunto de datos. De hecho, esta tarea se define como un proceso completo y no como acciones individuales o puntuales. De forma general, se pueden definir 3 fases diferentes en todo proceso de limpieza de datos:

- Detección y definición de la tipología de errores.
- Búsqueda e identificación de los casos de error.
- Corrección de estos casos de error.

Cada una de estas tres fases constituye un problema complejo en sí mismo, aunque quizá son las dos primeras las que conllevan un mayor nivel de dificultad.

La mayoría de soluciones de *Data Cleansing* se centran exclusivamente en el análisis de la integridad de los datos para detectar errores. Esta tipología de análisis -enfocada a bases de datos relacionales- es la operativa más sencilla en una tarea de limpieza de datos. Para un conjunto de datos (base de datos), el análisis de integridad incluiría más

opciones: integridad relacional, referencias, relación entre entidades, integridad por columna, etc. y se podría obtener con consultas SQL directas contra dicha base de datos.

La función de análisis de integridad de datos permite destapar un gran número de errores, si bien no es capaz de identificar errores más complejos. Errores que involucran relaciones entre uno o varios campos son, a menudo, más complicados de encontrar. Esta tipología de errores en datos requiere un análisis más profundo basado en métodos más complejos.

Digamos que un gran porcentaje (99,5%) de los datos se comportan de forma similar, entonces podríamos decir que el resto (0,5%) podrían ser candidatos a ser erróneos. Estos datos se consideran *outliners*. El proceso para llegar a este conjunto de datos se compone de dos partes: por un lado la identificación de las tendencias de “normalidad” de los datos y por otro, la de los *outliners* o variaciones extrañas.

En el mundo real, para llegar a determinar una tendencia de normalidad de los datos rara vez basta con un único modelo de distribución. Este proceso suele basarse en varios métodos diferentes:

- **Modelo Estadístico:** identifica los valores erróneos en base a medias, desviaciones estándar, rangos, etc. (basado en el teorema de Chebyshev).
- **Modelo de Clustering:** modelos de Minería de Datos (*Datamining*) que permiten agrupar conjuntos de datos con patrones comunes, determinados también por el propio algoritmo.
- **Modelo Basado en Patrones:** Búsqueda de valores que no conforman un patrón específico,

Nombre: Ricardo García
DNI: 123456789
Email: R.Garcia@HTOMAIL.COM
Dirección: C/ Serrano 8. 5º. Madrid

Entrada datos manual

ASNEF RAI : Registro de Aceptaciones Impagadas de Asociación Nacional de Entidades de Financiación

Validaciones con terceros



Data Cleansing:
 Normalización de dirección, DNI, teléfono, fechas, cta. bancaria
 Errores en campos (email)

Data Enhancement (España):
 Sexo, Latitud/Longitud, Edad Media de la zona, etc.

Matching:
 De-duplicación de datos
 Complimentación extra de información

<p>Nombre: Ricardo García DNI: 123456789 Email: R.Garcia@HTOMAIL.COM Dirección: C/ Serrano 8. 5º. Madrid</p>	<p>Nombre: Ricardo Apellido: García DNI: 123456789 – B Email: r.garcia@hotmail.com Dirección: Tipo: Calle Calle: Serrano Número: 8 Piso: 5 Población: Madrid C.P.: 28001</p>
<p>Teléfono: 666555555 F. Nacimiento: 08/11/75 Nº Cta.: 21002254180200213776</p>	<p>Teléfono: +34-666-55-55-55 F. Nacimiento: 08/11/1975 Nº Cuenta: Entidad: 2100 Sucursal: 2254 D.C.: 18 Cta.: 0200213776</p>
<p>(marketing) (marketing) (marketing) (marketing)</p>	<p>Edad Media de la Zona: 34 Sexo: H Longitud: 40.42194 Latitud: -3.68847</p>

Email: r.garcia@hotmail.com
Telf: 666555555
F. Nacimiento: 08/11/75

Reg. existentes en BBDD

Nombre: R. García
Nº Cta: 21002254180200213776

Cleansing + Enhancement + Matching

bien manual o bien obtenido como combinación de técnicas matemáticas (particionado, clasificación y clustering). El patrón se define como el grupo de registros que cumplirían el mismo “comportamiento” según un x% de confianza definido por el usuario.

- **Modelo de Reglas de Asociación:** reglas de asociación con altos intervalos de confianza que definen diferentes tipos de patrones. Como en el modelo anterior, los registros que no sigan estos patrones serán considerados *outliners*. Este modelo se recomienda cuando se trata con datos de diferentes tipos. Habitualmente se utilizan reglas de asociación, ya definidas, como modelos estándar -similar al modelo basado en patrones- pero podría extenderse a otros tipos de asociación como, por ejemplo, correlaciones estadísticas.

En el mundo SAP en concreto, para cubrir estas necesidades descritas anteriormente, se dispone de la herramienta Business Objects Data Quality Management como solución estándar. Esta consta de las siguientes funcionalidades, que cubren notablemente cualquier nivel de integración de datos que fuese necesaria:

- **Data Analysis & Measurement.** Este componente es el encargado de reconocer, dentro del conjunto de datos, aquellos considerados como *outliner*. Una vez identificados los errores, el sistema provee de herramientas para conocer exactamente la criticidad de los errores en el conjunto de datos. Entre otras funcionalidades, permite establecer unos parámetros de control que ejecuten alertas cuando los resultados de los análisis superen un umbral determinado. Además, cuenta con un cuadro de mando que, de forma muy visual, permite identificar los problemas mediante diagramas de Benn, distribuciones de frecuencia, informes de integridad referencial, etc.
- **Data Cleansing.** Este componente es capaz de estandarizar la información en base a patrones definidos, utilizando para ello estándares internacionales de hasta 190 países para normalizar diferentes tipos de datos: teléfonos, direcciones de email, etc., incluso si la información es semi-estructurada (reconocimiento automático de “Pso. de la Castellana”, “Castellana” ó “Paseo Castellana”).
- **Data Enhancement.** *Data Enhancement* permite opciones de enriquecimiento de los

datos para conocer y acceder con mayor fiabilidad a clientes o prospectos. Quizá la más interesante y extendida sea el *Geocoding* (información geo-demográfica) para campañas de marketing orientadas a núcleos poblacionales.

- **Data Matching & Consolidation.** Esta herramienta es capaz de identificar y corregir duplicidades en los datos. Además, no sólo encuentra los patrones de duplicidad (**matching**), sino que es capaz de consolidar la información de un registro, cumplimentándola con la de sus duplicados.

Mantener una base de datos fiable que garantice una gestión eficaz de los clientes se está convirtiendo en una necesidad cada vez más relevante para las empresas. Single Consulting, como consultora experta en el asesoramiento y soporte a entidades inmersas en procesos complejos de transformación, entiende esta necesidad y la integra como parte de su apuesta a futuro dentro del mundo de *Business Intelligence*, ayudando a sus clientes en la definición y elección de los procesos necesarios para asegurar el éxito de estos proyectos. □