Capítulo 3
Preparar sus datos para sacarles todo su potencial

1. La calidad de los datos: un recordatorio

La calidad de los datos es un elemento fundamental que hay que tener en cuenta antes de abordar técnicas de limpieza y procesamiento de datos. Para cualquier organización que quiera tomar decisiones con conocimiento de causa y sacar el máximo partido de su información, este aspecto no puede pasarse por alto. Todos los procedimientos que examinaremos en este capítulo tienen un único objetivo: proporcionar a los distintos equipos datos fiables.

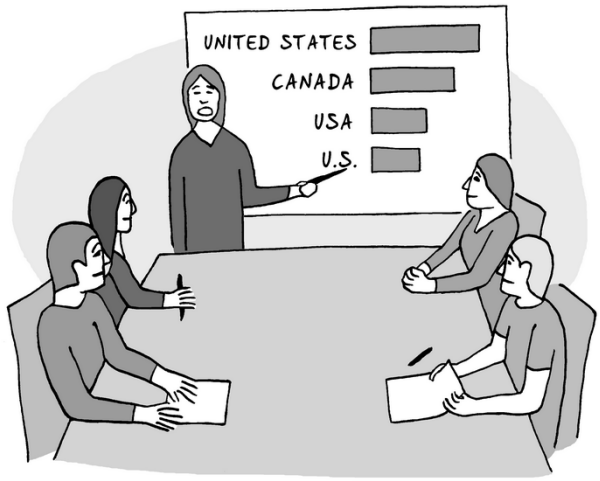
1.1 ¿Qué es la calidad de los datos?

La calidad de los datos refleja la capacidad de una organización para mantener la exactitud y sostenibilidad de su información a lo largo del tiempo. Como expertos en la materia, debemos ofrecer datos irreprochables, basados en indicadores claros y fácilmente interpretables. Empezaremos por examinar en detalle los seis criterios que definen las dimensiones de calidad de los datos (DQD).

Esta noción engloba tanto las características intrínsecas de los datos como los métodos aplicados para garantizarlas. En esencia, la calidad de los datos se define por su capacidad para servir a los fines previstos.

Una iniciativa de calidad de datos es un proceso a largo plazo, integrado en todo el ciclo de vida de los datos. Requiere un cambio cultural en la forma en que la organización gestiona sus datos. Es un enfoque global que afecta a toda la empresa y a sus prácticas cotidianas.

Es importante señalar que la introducción de datos erróneos en un proceso producirá inevitablemente datos inexactos en la salida. Por consiguiente, una estrategia basada en datos de mala calidad dará lugar a decisiones ineficaces, con consecuencias directas en el retorno de la inversión.



Créditos: https://dataedo.com/

1.2 ¿Por qué es importante la DQD?

La calidad de los datos suele verse comprometida por diversos factores. Entre ellos se encuentra el error humano en el momento de su introducción inicial. Errores tipográficos, diferentes convenciones de nomenclatura entre fuentes de datos o abreviaturas incorrectas son una fuente frecuente de problemas. Además, una información inicialmente exacta puede quedar obsoleta con el tiempo, al cambiar el contexto.

Si la calidad de los datos se ve comprometida, habrá consecuencias costosas para las empresas. He aquí algunos ejemplos concretos de problemas frecuentes y sus repercusiones:

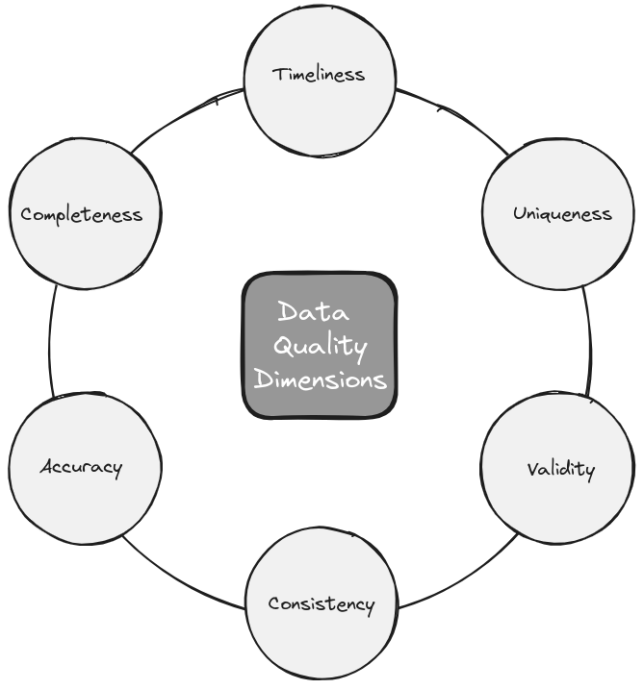
– Errores de introducción de datos: en 2018, un empleado de Samsung Securities cometió un monumental error de introducción de datos al distribuir dividendos, emitiendo inadvertidamente 2.800 millones de acciones «fantasmas», unas treinta veces el número total de acciones existentes. Este error provocó una perturbación masiva del mercado y una pérdida de confianza en la gestión de la empresa, lo que obligó a adoptar costosas medidas correctoras. [Fuente: https://www.sirfull.com/en/blog/poor-data-quality-impacts/]
– Incoherencias en los datos: en 2022, Unity Technologies se enfrentó a un grave problema con su herramienta de publicidad dirigida Pinpointer. Los datos inexactos de un cliente importante corrompieron los modelos de IA, lo que provocó una pérdida de ingresos de 110 millones de dólares y una caída del 37 % en el valor de las acciones de la empresa. [Fuente: https://zeenea.com/what-are-the-most-common-data-quality-issues-and-how-can-you-solve-them/]

– Datos obsoletos: Tesco, el gigante británico de la distribución, tuvo problemas de inexactitud en sus datos de existencias, lo que provocó frecuentes desabastecimientos en sus tiendas. Estas imprecisiones provocaron pérdidas de ventas, insatisfacción de los clientes y un impacto negativo directo en el volumen de negocios de la empresa. [Fuente: https://www.sirfull.com/en/blog/poor-data-quality-impacts/]
– Errores críticos en los datos: entre marzo y julio de 2017, Equifax generó puntuaciones crediticias incorrectas para millones de consumidores debido a datos erróneos. La empresa se enfrentó a multas reglamentarias, demandas judiciales y una pérdida de credibilidad, lo que puso en peligro las decisiones de concesión de préstamos y erosionó la confianza del público. [Fuente en francés: https://www.datagalaxy.com/fr/blog/big-data-attention-aux-donnees-de-mauvaise-qualite/]
– Datos de localización inexactos: Royal Dutch Shell tuvo errores en los datos de localización de sus pozos petrolíferos, lo que provocó una perforación ineficaz. Estos errores supusieron millones de dólares en costes adicionales y una considerable pérdida de tiempo debido a una perforación incorrecta. [Fuente en francés: https://solutions-business-intelligence.fr/data-quality-enjeux-et-bonnes-pratiques/]

Cada uno de estos problemas de calidad de datos tuvo un impacto significativo en las operaciones, la reputación y los resultados financieros de las empresas implicadas. Estos ejemplos ponen de relieve la importancia de invertir en procesos sólidos de gestión de la calidad de los datos para evitar estos costosos escollos.

1.3 Los principales criterios de la DQD

Veamos los principales criterios de la DQD, núcleo de cualquier estrategia eficaz de gestión de datos.



1.3.1 Precisión (accuracy)

La precisión es el grado en que los datos almacenados se ajustan a los valores reales que deben representar. Mide lo cerca que está el valor almacenado del valor real correcto o aceptado.

En el sector bancario, imaginemos un error en el cálculo de los intereses de un préstamo hipotecario. Si el tipo de interés se registra incorrectamente como 3,5 % en lugar de 3,05 %, esto podría dar lugar a que se cobrara de más a miles de clientes. No solo podría acarrear importantes pérdidas financieras para el banco en caso de reembolso, sino que también podría dañar gravemente su reputación y dar lugar a posibles sanciones.

Para garantizar la exactitud, las empresas pueden poner en marcha procesos de validación de datos, comprobaciones cruzadas automatizadas y auditorías periódicas. El uso de inteligencia artificial para detectar anomalías también puede ser eficaz.

1.3.2 Integridad (completeness)

La integridad es el grado de presencia de todos los elementos de datos necesarios en un conjunto de datos concreto. Evalúa el grado en que se incluyen todos los valores requeridos y están presentes todos los registros que deberían estarlo.

En el campo de la investigación médica, imaginemos un estudio sobre la eficacia de un nuevo tratamiento contra el cáncer. Si los datos de seguimiento posteriores al tratamiento están incompletos para un grupo significativo de pacientes, las conclusiones del estudio podrían estar distorsionadas. Las decisiones cruciales sobre si aprobar o rechazar el tratamiento podrían basarse en información parcial, lo que podría afectar a la vida de muchos pacientes.

El uso de campos obligatorios en los formularios de introducción de datos, la creación de alertas para los datos que faltan y la aplicación de procesos sistemáticos de recopilación de datos pueden mejorar la exhaustividad.

1.3.3 Coherencia (consistency)

La coherencia se refiere a la ausencia de contradicciones en los datos dentro de un conjunto de datos o entre distintos conjuntos de datos. Garantiza que los datos sean uniformes y lógicamente compatibles en todos los sistemas, aplicaciones y procesos de la organización.

En una empresa multinacional, imagine que el departamento de recursos humanos y el de finanzas utilizan sistemas diferentes para gestionar la información de los empleados. Si un empleado cambia de trabajo y esta información solo se actualiza en el sistema de recursos humanos, podrían producirse errores en las nóminas, las prestaciones e incluso en la planificación estratégica de recursos humanos.

El uso de un sistema de información integrado, la aplicación de procesos automáticos de sincronización entre distintos sistemas y el establecimiento de normas coherentes de gestión de datos en toda la empresa pueden mejorar la coherencia.

1.3.4 Vigencia (timeliness)

La vigencia mide el grado en que los datos están actualizados y disponibles en el plazo requerido para su uso previsto. Evalúa la frescura de los datos en relación con el momento en que se crearon o actualizaron por última vez, así como su disponibilidad en el momento en que se necesitan para los procesos empresariales.

En el trading de alta frecuencia, donde las decisiones de compra y venta se toman en milisegundos, la vigencia de los datos es fundamental. Un retraso de solo unos segundos en la actualización de los precios de las acciones puede acarrear considerables pérdidas financieras. Por ejemplo, si un acontecimiento importante afecta a la cotización de una acción, pero esta información no se refleja inmediatamente en los datos utilizados por los algoritmos de negociación, podrían tomarse decisiones de inversión desastrosas.

El uso de sistemas de procesamiento en tiempo real, la implantación de procesos automáticos de actualización de datos y la optimización de los flujos de datos pueden mejorar la puntualidad.

1.3.5 Validez (validity)

La validez es la medida en que los datos se ajustan a las normas empresariales definidas, los formatos especificados y las restricciones del dominio. Garantiza que los valores de los datos cumplen los criterios sintácticos y semánticos establecidos para el tipo de datos en cuestión.

En el sector de la aviación, imaginemos un sistema de reservas que acepte una fecha de vuelo anterior a la actual. Esto podría acarrear graves problemas en la planificación de vuelos, la gestión de tripulaciones y, potencialmente, comprometer la seguridad si estos datos no válidos se utilizan en otros sistemas críticos.

La aplicación de controles de validación estrictos en las interfaces de entrada, el uso de restricciones a nivel de base de datos y el establecimiento de procesos periódicos de limpieza de datos pueden mejorar la validez.

1.3.6 Singularidad (uniqueness)

La singularidad es la propiedad de que cada entidad distinta del mundo real se represente una y solo una vez en el conjunto de datos. Garantiza que no haya duplicados involuntarios ni redundancias en los registros de datos.

En el sector sanitario, imagine un paciente con varios historiales médicos debido a duplicados en la base de datos. Esto podría dar lugar a graves errores en el tratamiento si un médico no tiene acceso al historial médico completo del paciente. Por ejemplo, podrían pasarse por alto alergias o interacciones entre medicamentos, poniendo en riesgo la salud del paciente.

El uso de identificadores únicos, la aplicación de procesos de desduplicación y la implantación de controles estrictos a la hora de crear nuevos registros pueden mejorar la singularidad de los datos.

2. Depuración de datos

2.1 Primeros pasos con la biblioteca pandas

pandas es una potente y versátil biblioteca de Python diseñada para la manipulación y el análisis de datos. Fue desarrollada por Wes McKinney, un investigador que empezó a construir lo que se convertiría en pandas. El nombre «pandas» deriva de «Panel Data», un término econométrico para conjuntos de datos que incluyen observaciones a lo largo de varios periodos.

pandas es particularmente adecuada para trabajar con datos tabulares, similares a una hoja de cálculo de Excel o una tabla SQL. Las principales estructuras de datos gestionadas por esta biblioteca son series, que almacenan datos a lo largo de una dimensión, y DataFrames, que lo hacen a lo largo de dos dimensiones (filas y columnas). Estas estructuras de datos facilitan su manipulación, así como su limpieza, preprocesamiento, análisis y visualización.

pandas es muy utilizada en el análisis de datos. A menudo se presenta como la herramienta ideal para manipular datos que pueden organizarse en filas y columnas. Es más, dominar pandas es una habilidad muy buscada por los reclutadores, ya que muchas empresas de todos los sectores recurren cada vez más a la ciencia de datos.

Existen varias alternativas a pandas, entre ellas polars, dask y cudf. Cada una de estas soluciones tiene sus ventajas, en particular la velocidad de procesamiento en comparación con pandas. No las trataremos en este libro, ya que pandas sigue siendo la librería más utilizada para el análisis de datos en Python. Su riqueza y versatilidad, así como su amplia adopción en la comunidad de análisis de datos, la convierten en una herramienta esencial.

2.2 Presentación de nuestro conjunto de datos

En este capítulo, trabajaremos con un conjunto de datos que está disponible gratuitamente en la plataforma Kaggle.

El comercio electrónico se ha convertido en un nuevo canal de apoyo al desarrollo empresarial. A través del comercio electrónico, las empresas pueden acceder a un mercado más amplio y establecer una mayor presencia proporcionando canales de distribución más baratos y eficientes para sus productos o servicios. El comercio electrónico también ha cambiado la forma en que las personas compran y consumen productos y servicios. Muchas personas recurren a sus ordenadores o dispositivos inteligentes para encargar productos, que pueden recibir fácilmente en sus domicilios.

Se trata de un conjunto de datos de un año de transacciones de ventas en el Reino Unido para el comercio electrónico (venta minorista en línea). Esta tienda londinense vende regalos y artículos para el hogar para adultos y niños en su sitio web desde 2007. Sus clientes proceden de todo el mundo y suelen hacer compras directas para ellos mismos. También hay pequeñas empresas que compran al por mayor y venden a otros clientes a través de canales minoristas.