Calidad de datos

PID_00246838

David Cabanillas Barbacil

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Índice

Introducción

El potencial valor del dato no ha parado de crecer en los últimos años, como resultado de la progresiva transformación digital. Actualmente, las organizaciones tienen a su disposición múltiples estrategias para alcanzar dicho valor, como por ejemplo *big data**, *business analytics* o *business intelligence*. Sin embargo, tal y como apunta McKinsey, las organizaciones, en la gran mayoría de los sectores, aún no han logrado capturar el valor.

Lectura complementaria

Henke, N.; Bughin, J.; Chui, M.; Manyika, J.; Saleh, T.; Wiseman, B.; Sethupathy, G. (2016). *The age of analytics: Competing in a data-driven world*. McKinsey Global.

*** Más información en: https://goo.gl/7nHbyp**

Tal y como apuntan Ransbotham y Kiron, el gobierno del dato es clave para desbloquear la oportunidad que proporcionan los datos y los algoritmos.

Lectura complementaria

Ransbotham, S.; Kiron, D. (2017). *Analytics as a Source of Business Innovation*. MITSloan.

Los motivos detrás de la imposibilidad de capturar el valor de los datos son múltiples y diversos, e incluyen aspectos como la falta de talento, la existencia de silos de información o incluso el escepticismo por parte de la dirección.

Dentro de las áreas de aplicación del gobierno del dato (aunque puede encontrarse de forma independiente), destaca la **calidad del dato**. Sin ciertos niveles en la calidad de los datos, la toma de decisiones y la eficiencia de los procesos pueden verse afectadas por una pérdida de integridad, por falta de completitud o incluso por la aparición de inconsistencia. Lo que, en definitiva, muchas veces denominamos como *garbage in, garbage out*.

En este módulo estudiaremos la necesidad e importancia de la calidad del dato, en qué consiste, qué aporta, cómo implementarla, qué debemos tener en cuenta como mejores prácticas y qué tecnologías soportan la calidad del dato.

Objetivos

Este material didáctico está dirigido a:

1) Desarrolladores y consultores que quieren conocer qué significa calidad del dato o *data quality*.

2) Desarrolladores y consultores que quieren ayudar al desarrollo de estrategias de negocio que incluyan calidad de datos.

3) Gestores que están interesados en la transformación digital de su organización y en la inclusión de calidad del dato como uno de sus pilares fundamentales.

En los materiales didácticos de este módulo, encontraremos las herramientas indispensables para asimilar los siguientes objetivos:

1) Entender el concepto de *data quality*, las situaciones en las que es necesario desplegar una solución de este tipo y las ventajas que proporciona.

2) Conocer en qué consiste un programa de calidad de datos.

3) Enumerar y dar a conocer mejores prácticas de calidad de datos.

4) Conocer técnicas y tecnologías para la gestión de calidad de datos.

Si bien la obra es autocontenida en la medida de lo posible, los conocimientos previos necesarios son:

1) Conocimientos básicos sobre *business intelligence* y *big data*.

2) Conocimientos sobre estrategia y gestión de las tecnologías de la información (TI).

Se introducirán los conceptos necesarios para el seguimiento de este material.

1. Calidad de datos

1.1. La necesidad de la calidad de datos

Para que el dato pueda considerarse como un activo de valor y la organización pueda tomar mejores decisiones, mejorar procesos, reducir costes y crear nuevas fuentes de ingresos, es necesario que, a lo largo de su ciclo de vida, el dato no tenga problemas que afecten a su calidad.

Este tipo de problemas suponen graves costes asociados, derivados de tomar decisiones erróneas, incrementar los costes operativos, generar insatisfacción entre los clientes, deteriorar la imagen corporativa y una pérdida de confianza entre los clientes, empleados y proveedores. Sin la confianza de las personas que toman decisiones, se puede retrasar o incluso parar la explotación de los datos, lo que supone una barrera a convertirse en una empresa orientada al dato, o *data driven*.

Confianza

Confianza es la seguridad o esperanza firme que alguien tiene de otro individuo o de algo. También se trata de la presunción de uno mismo y del ánimo o vigor para obrar. En nuestro caso, es la confianza sobre los datos que tienen las personas en la organización.

Data driven

Data driven hace referencia a un proceso o una actividad que es guiada por los datos, en lugar de ser impulsada por la mera intuición o experiencia personal. En nuestro contexto, las decisiones se toman sobre los datos y no sobre especulaciones o sensaciones.

Los problemas de calidad de datos son sistémicos en las organizaciones y, de hecho, diferentes estudios llevan tiempo mostrando la magnitud del problema al que nos enfrentamos:

- En el 2002, la baja calidad en los datos sus clientes supuso pérdidas de 611 billones de dólares a las compañías de Estados Unidos*.
- En el 2004, se estimó que la calidad del dato suponía pérdidas de al menos el 10%, posiblemente estaba más cerca del 20%**.
- En el 2011, la baja calidad del dato se consideraba la principal razón por la que el 40% de las iniciativas de negocio fracasaban en conseguir los objetivos definidos***.
- En el 2014, el 59% de las empresas citaban la calidad de los datos como una barrera para la adopción de *business intelligence*****.
- En el 2016, las organizaciones perdieron en promedio 9,7 millones de dólares anualmente, debido a la mala calidad de datos******.

* Eckerson, W. (2002) *Data Quality and the Bottom line.* TDWI.

** Redman, T. C. (2004) *Data: An Unfolding Quality Disaster. DM Review Magazine.*

*** Friedman, T.; Smith, M. (2011) *Measuring the Business Value of Data Quality.* Gartner.

**** VV. AA. (2011) *2014 Analytics, BI, and Information Management Survey.* Information Week.

***** Duncan, A. D.; Selvage, M. Y.; Judah, S. (2016). *How a Chief Data Officer Should Drive a Data Quality Program.* Gartner.

El impacto de la calidad de datos va mucho más allá de la toma de decisiones, y es posible encontrarlo a lo largo de todos los procesos de una organización, por ejemplo:

- Datos de clientes erróneos inciden en la efectividad de las campañas de marketing.
- Direcciones incorrectas producen envíos fallidos de productos y un incremento en los costes operativos.
- Las mediciones incorrectas de productos pueden conducir a problemas significativos de fabricación y/o transporte; por ejemplo, que el producto no encaje en el camión que lo debe transportar.
- Los malos datos en la cadena de suministro de comestibles cuestan a la industria australiana más de mil millones de dólares australianos. IBM, junto con GS1 en Australia*, comparó los datos de los productos de supermercado de tres grandes superficies con los datos de los cuatro principales proveedores. Se reveló que los supermercados estaban trabajando con datos con inconsistencias de en torno al 80%. Desde errores o falta de dimensiones, hasta el número de palés que se pueden cargar por producto o las condiciones de almacenamiento.

  *** Más información en: https://goo.gl/nhqCVE**
- La recogida errónea de datos de presión y temperatura en la producción de piezas de plástico puede hacer que salgan al mercado vehículos que tengan problemas de incendio por una mala fabricación de sus componentes.

Por el contrario, una alta calidad de los datos puede mejorar la ventaja competitiva y capacidad de las organizaciones. Entre los impactos positivos, podemos destacar:

- Mayor efectividad en la adquisición y retención de clientes.
- Optimización en todas las áreas de la organización.
- Ejecución de procesos eficientes en la cadena de suministro y producción.
- Eliminación de costosos errores operativos.
- Penetración rápida en nuevos mercados.
- Toma de decisiones de negocios inteligente y oportuna.

Es necesario comentar que, en general, las organizaciones no inician proyectos de gobierno del dato, sino que identifican problemáticas específicas y, al final, las combinan bajo el mismo programa. De hecho, frecuentemente muchos

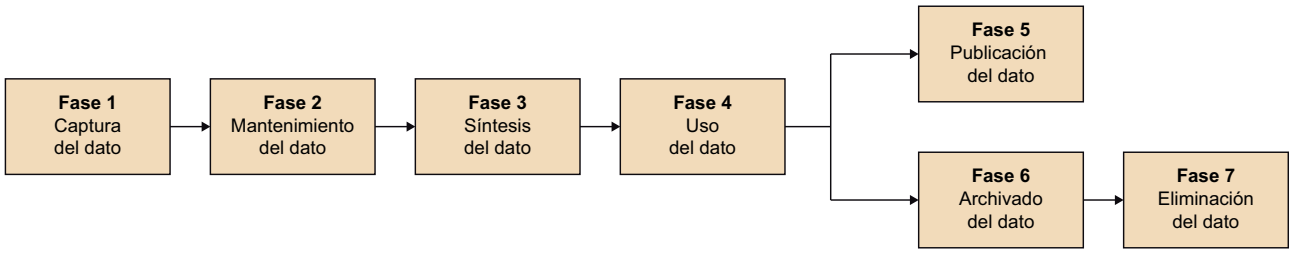
de los proyectos de gobierno del dato tienen sus orígenes en su intento de revolver problemáticas asociadas a la calidad de datos.

1.2. Motivos de la baja calidad de datos

La gestión de los datos con los que cuenta una organización ha ido ganando peso en los últimos tiempos, y hoy día se ha convertido en una operación clave para asegurar el futuro de cualquier organización. El éxito o el fracaso en la gestión de la calidad de datos está íntimamente vinculado con los riesgos que se asumen en este tipo de operaciones de gestión, y especialmente con el plan y las medidas adoptados para afrontarlos; sin embargo, y pese a la evidente importancia de que las organizaciones se doten de un correcto plan de gestión de calidad de datos, en muchas ocasiones o no se aplica ningún plan en la gestión, o solo se centran en la entrada de datos.

Es importante recordar que el dato tiene un ciclo de vida, como se ilustra en la figura 1, y que a lo largo del mismo podemos encontrar ejemplos de baja calidad, como se recoge en la tabla 1.

Figura 1. Fases *data governance*



Fuente: Marcos Pérez Rodríguez

Tabla 1. Fases del ciclo de vida del dato y ejemplo de baja calidad de datos

Fase	Ejemplo de baja calidad de dato
1 . . . 7	Cualquier intervención manual en el proceso de flujo de datos. Sistemas de información no sincronizados que deben compartir información.
2 . . . 3	Aplicar una fórmula o algoritmo sobre datos que es incorrecta.
5	Envío de informes desfasados.
6	Metadatos que describen los datos de manera incorrecta.

Metadato

Metadato hace referencia a datos sobre los datos. Por ejemplo: tiempo y creación del dato, creador/fuente del dato. . .

La captura del dato, **fase 1**, es el punto de entrada de los datos y el punto más frecuente de errores de la baja calidad de datos. Se trata de problemas como errores ortográficos, transposiciones de números, códigos incorrectos o partes del dato que no han sido incluidas, datos que han sido colocados en los campos incorrectos y nombres, apodos, abreviaturas, acrónimos irreconocibles, tipos de datos incoherentes o que la captura de los datos se hace de manera incorrecta (por ejemplo, sensores no calibrados correctamente). Estos tipos de errores están aumentando a medida que las empresas trasladan sus negocios al entorno digital y a la cuarta revolución, y se permite a los clientes/proveedores y terceros introducir o usar datos sobre ellos directamente en

los propios sistemas. Se pueden evitar muchos errores de entrada de datos mediante el uso de rutinas de validación que comprueban los datos a medida que se introducen en los sistemas comprobando sintaxis, formatos, datos extraños y/o estructuras no coincidentes.

En la **fase 2** entran en juego las herramientas ETL (*extract, transform, load*), que permiten extraer, transformar y cargar el dato. Pensemos en un pequeño ejemplo de ETL: supongamos que tenemos datos procedentes de dos sistemas diferentes, que deseamos almacenar en el mismo destino, pero podría haber algunas diferencias entre los dos. Por ejemplo, uno puede denotar el género como M y F, el otro lo designa como 0 y 1. Ahora, si se desea almacenar en un único fichero y el género se quiere almacenar como M y F, se debe transformar el 0 y 1 a M y F. Un proceso ETL debe traducir los 0 ⇒ M y los 1 ⇒ F.

ETL

ETL es el acrónimo de *extract, transform and load*, y hace referencia a los procesos que permiten extraer, transformar y cargar datos para habilitar su consumo eficiente.

En la **fase 3**, encontramos las reglas y los algoritmos aplicados al dato. Por ejemplo, se puede crear una regla que indique si dar o no crédito a un cliente. En este caso, la regla/proceso, a diferencia de la fase 2, es parte de la lógica de negocio de la organización. Una regla podría ser un campo booleano que indica si dar o no el crédito siguiendo la regla: si salario ≥ 1.000 y gastos ≤ 200, dar crédito, si no, no dar crédito. En los dos casos, si la regla o proceso fueran incorrectos, se introducirían errores en los datos.

En la **fase 5**, el tiempo en el que se utiliza el dato puede ser un factor de error. Por ejemplo, si estamos enviando datos de ventas correctos, pero indicamos que son de un periodo y en realidad son de otro.

Finalmente, en la **fase 6** un error puede venir si los metadatos que describen los datos almacenados son obsoletos o incorrectos.

Como es posible imaginar, los errores de entrada de datos se agravan cuando las organizaciones intentan integrar datos de múltiples sistemas y/o llevan a cabo cambios estructurales en los sistemas de origen. A veces estos cambios son deliberados, como cuando un administrador agrega un nuevo campo o valor de código y no se notifica al resto de la organización. En otros casos, se generan nuevos datos a partir de datos existentes (por ejemplo, el sexo se puede tratar de obtener a partir del nombre del usuario). Debido a la complejidad de los sistemas actuales, los cambios en los sistemas fuente se propagan de manera fácil y rápida a otros sistemas a través del ciclo de vida.

1.3. Motivos por los que no se invierte en la calidad de datos

Por lo comentado hasta el momento, debería ser prioritario para una organización invertir en la calidad del dato y, sin embargo, esto es algo que no forma parte de sus prioridades. Como apuntan los últimos informes de la Society for Information Management (SIM), llamados *IT Trends Study*, las principales

prioridades del departamento TI son la seguridad de activos digitales, la falta de talento, la alineación con el negocio, la credibilidad y la continuidad de negocio, por lo que esta tendencia está cambiando.

Las principales razones que justifican no invertir en la calidad de datos son:

1) Se considera que los datos son correctos: se trata del motivo más extendido. Sin embargo, la mayoría de los estudios defienden lo contrario. Esto se suele argumentar por dos razones principales:

a) Miedo a lo desconocido: dentro de la empresa, no se sabe muy bien cómo afrontar una iniciativa de calidad de datos, por desconocimiento de las técnicas existentes, de qué supone esta iniciativa, lo que implica, etc.

b) Miedo a lo que se puede encontrar: tradicionalmente, la calidad de los datos se había asignado a las áreas de TI, un departamento que no necesariamente debería conocer el significado de los datos. Por eso, es importante que la responsabilidad de un dato permanezca en el área de negocio que lo gestiona.

2) Descubrimiento de errores: con una mayor calidad de los datos, se produce incremento de la comprensión del rendimiento de la organización. Por ejemplo, se podría descubrir que se tienen menos clientes de los que se creía o que no se ha sido capaz de detectar fraudes (entre otros).

3) No se ve un valor directo en la inversión: la mayoría de los estudios defienden que disponer de datos de calidad evita que haya retrasos en los proyectos y que se entre en sobrecostes. Así pues, la inversión en este tipo de proyectos genera un gran retorno en el futuro.

4) No se percibe la necesidad: se trata de un motivo poco real ya que, ¿quién no necesita comprobar que sus informes son correctos? ¿Quién no quiere saber si sus decisiones son consistentes? ¿O quién no necesita detectar un fraude? Hay muchos motivos para apostar por implementar un proyecto de calidad de datos.

5) Excesivamente costoso: aunque pueda parecerlo, realmente es una percepción errónea. Actualmente, existen muchas herramientas asequibles o incluso gratuitas que permiten implementar soluciones, totales o parciales, de calidad de datos en tiempos cortos y con una gran efectividad.

1.4. ¿Qué es la calidad de datos?

Hemos estado hablando de la necesidad y de las barreras de la calidad del dato, pero no hemos introducido aún una definición formal. Es el momento de hacerlo, según la norma ISO 9000:2015*.

*** Más información en: https://goo.gl/kEwyR6**

Se entiende por **calidad de datos** el grado en el que los datos cumplen un conjunto de características y/o dimensiones.

ISO 9000

ISO 9000 es un conjunto de normas sobre calidad y gestión de calidad, establecidas por la Organización Internacional de Normalización (ISO), y especifica la manera en que una organización opera sus estándares de calidad, tiempos de entrega y niveles de servicio.

Para comprender esta definición, es necesario entrar en el detalle de las características y/o dimensiones que deben cumplir los datos. No existe un consenso en la industria sobre cuáles son estas dimensiones. Por ejemplo, según la OECD (Organization for Economic Co-operation and Development), son **relevancia**, **exactitud**, **credibilidad**, **oportunidad**, **accesibilidad**, **interpretabilidad** y **coherencia**. Según EUROSTAT (Statistical Office of the European Communities), cabe añadir: **puntualidad**, **transparencia**, **comparabilidad** y **exhaustividad** (que en el caso de la OECD, están incluidas en sus dimensiones).

En el presente material, y con el objetivo de ser lo más imparciales que resulte posible, consideraremos el enfoque de DAMA*. Para esta organización, las dimensiones de la calidad de datos son:

*** Más información en:** ***The six primary dimensions for data quality assessment.*** **Dama** ***UK Chapter,*** **2013.**

DAMA

DAMA es una asociación sin ánimo de lucro. Tiene como objetivo ayudar a los profesionales de los datos mediante la creación de taxonomías y documentos de referencia y su divulgación.

- **Completitud**, que consiste en la proporción de datos almacenados respecto al conjunto total.
- **Unicidad**, que consiste en que el dato debe guardarse de forma única para evitar inconsistencias.
- **Atemporalidad**, que consiste en el grado en que el dato representa la realidad en un momento temporal específico.
- **Validez**, que consiste en que el dato presenta conformidad (formato, tipo, rango) respecto a su definición.
- **Precisión/exactitud**, que consiste en el grado en el que el dato describe la realidad (con independencia del tiempo).
- **Consistencia**, que consiste en la ausencia de diferencias al comparar dos representaciones del mismo dato, evitando información contradictoria.

Otras dimensiones que pueden incluirse son usabilidad, duplicación, disponibilidad, confianza y/o valor.

En la práctica, la calidad de los datos es una preocupación por parte de los profesionales que participan en los sistemas de información, que van desde el almacenamiento de datos y la inteligencia empresarial a la gestión de la relación con los clientes y la gestión de la cadena de suministro. Es decir, cualquier dato que esté de algún modo relacionado con la empresa u organización y la calidad del dato. La preocupación de las empresas ha hecho incorporar la calidad de los datos como parte fundamental del *data governance*.

Para atacar el problema de la calidad de los datos, las organizaciones necesitan invertir en las personas, los procesos y las tecnologías necesarias para transformar datos defectuosos en información confiable y procesable, disponible para todas las partes en cualquier momento que la necesiten. Las mejores iniciativas de calidad de datos tienen estas cuatro características listadas en la tabla 2.

Tabla 2. Características y sus descripciones

Característica	Descripción
Colaborativo	Negocio y TI comparten la responsabilidad de la calidad de los datos, con funciones y tecnología claramente definidas y adaptadas a las habilidades y perspectivas únicas de los analistas de negocio, administradores de datos y desarrolladores y administradores de TI.
Proactivo	Negocio y TI reconocen que todas las organizaciones sufren algún grado de mala calidad de los datos y trabajan conjuntamente para identificar y corregir los problemas antes de que afecten al rendimiento del negocio.
Reutilizable	El perfil de datos y las reglas de negocio de limpieza pueden reutilizarse en cualquier número de aplicaciones, para agilizar y acelerar los procesos y ayudar a garantizar altos estándares de calidad.
Pervasivo	El entorno de calidad de datos se extenderá a todas las partes interesadas, dominios de datos, proyectos y aplicaciones, independientemente de dónde residan los datos.

Para que la calidad de los datos sea más efectiva, debe ser impulsada por una metodología que incorpore las características definidas anteriormente. Idealmente, la metodología será supervisada e implementada por un órgano del gobierno del dato.

1.5. ¿Cómo resolver los retos?

Para gestionar la calidad de los datos en una organización, se deben definir una serie de tareas y roles puesto que, como hemos comentado, la resolución de problemas en la entrada de datos no es suficiente:

- **Los roles técnicos y no técnicos o de negocios tienen que comunicarse**. La falta de colaboración entre estas dos partes de una organización es una de las principales razones por las que muchos proyectos de calidad de datos no cumplen sus expectativas iniciales. Tradicionalmente, negocio y TI se han basado en hojas de cálculo, documentos, correos electrónicos y otros mecanismos tediosos e imprecisos para comunicar los requisitos de calidad de los datos. Inevitablemente, bajo estas condiciones es difícil para los analistas de negocio y administradores de datos, es decir, las personas que deben seguir el plan de calidad de datos, esbozar requisitos de negocio de calidad de datos en términos claros para que la parte TI los pueda entender y ejecutar. La mala interpretación, los retrasos, los altos costes y los resultados no óptimos son comunes simplemente porque negocio y TI están hablando dos idiomas diferentes, sin un marco común. Los detalles críticos se pueden perder en la traducción. Por lo tanto, la colaboración entre negocio y TI es esencial para la calidad de los datos y las iniciativas de gestión de datos relacionadas.

- **Dotarse de un sistema de supervisión continua de los datos recientemente incorporados** (así como de los ya existentes). Un sistema que debe integrar herramientas de detección que permitan filtrar y categorizar los datos según su calidad, y detectar, antes de que sean requeridos por el sistema de gestión (que los transformará primero en información relevante y, después, en conocimiento sensible para la toma de decisiones), su grado de coherencia, oportunidad y fiabilidad, entre otros factores determinantes. Las empresas que están administrando la calidad de los datos están operando normalmente en un entorno por lotes o *batch*; es decir, ejecutan comprobaciones de datos, crean conjuntos de datos revisados periódicamente y los introducen en el sistema de datos. Si bien el procesamiento periódico es ciertamente útil, en algunos entornos comprobar la calidad de los datos tan pronto como se crea el dato, por ejemplo, cuando se introduce una nueva transacción o en el momento en que un nuevo conjunto de datos esté disponible, es obligatorio o, como poco, ventajoso.

En resumen, para llegar a obtener una calidad de los datos efectiva, debemos ser capaces de llevar a cabo las siguientes actividades:

- **Objetivos**: determinación de las mejores oportunidades para optimizar los datos.
- **Conceptualización**: definición de dimensiones en la organización.
- **Evaluación**: evaluación de los niveles de calidad actual de los datos.
- **Eliminación**: eliminación de las fuentes de problemas.
- **Automatización**: automatización del control de la calidad de datos.
- **Seguimiento**: seguimiento de la gestión de procesos y sus datos.
- **Medición**: medición de la calidad de datos.

Y la parte más importante, aunar todas estas tareas en un marco o programa entre los tres agentes que deben intervenir para asegurar una buena calidad de datos: **personas**, **procesos** y la **tecnología**.

2. Programa de calidad de datos

2.1. Introducción

Para gestionar la calidad de datos en una organización, debemos desarrollar lo que se conoce como programa de calidad de datos. Debemos, por lo tanto, introducir este concepto.

Se entiende por **programa de calidad de datos** la metodología estratégica y sistemática para la evaluación de la calidad de los datos dentro de una organización, la cual permite identificar atributos de calidad de datos, analizar dichos atributos en su contexto actual o futuro y proporcionar una guía para mejorar la calidad de los datos.

El programa permite orquestar personas, procesos y tecnología, con el objetivo de conseguir el mayor retorno de una iniciativa de calidad de datos.

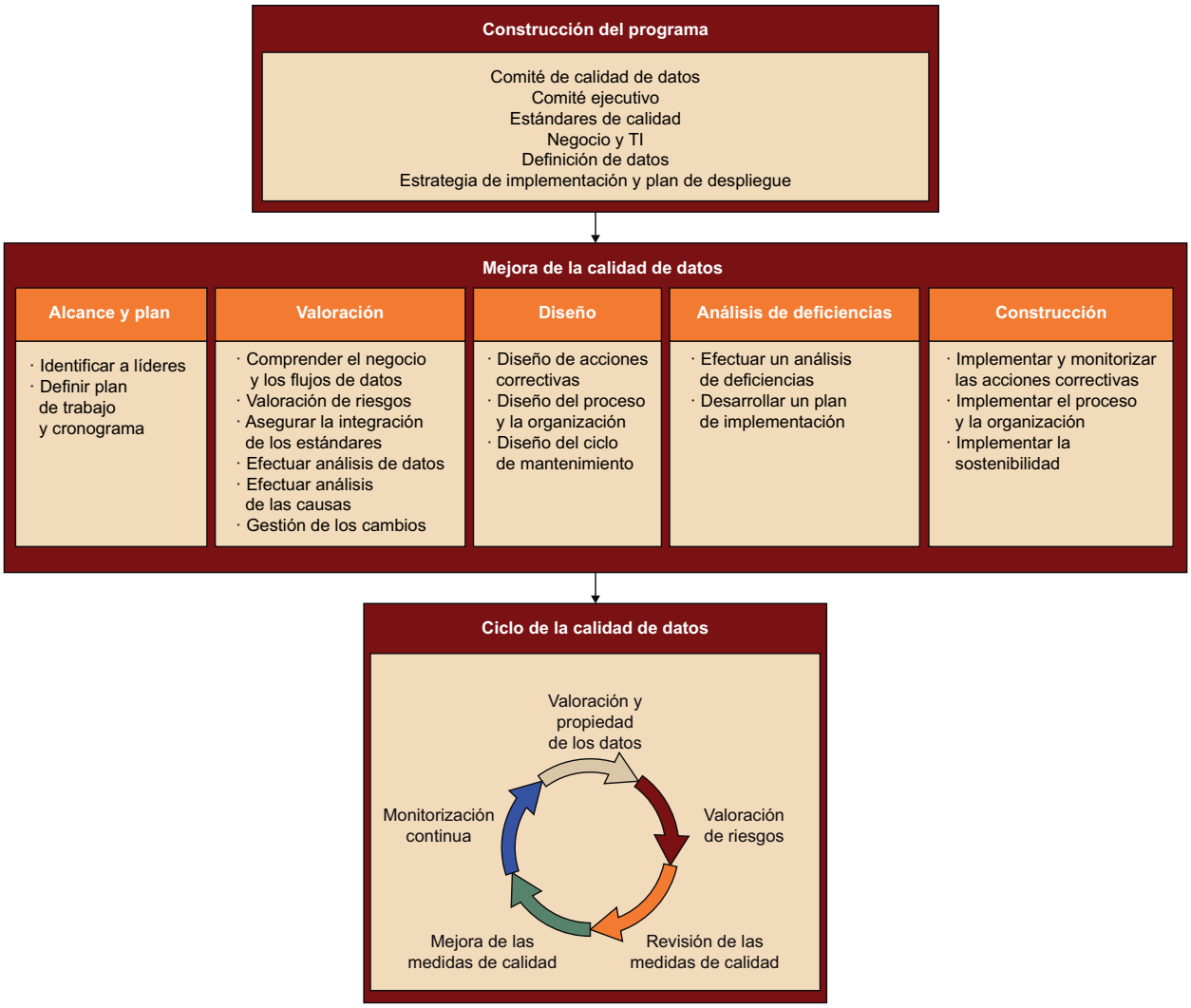
Si bien el valor de una metodología de calidad de datos puede parecer evidente, demasiadas organizaciones abordan iniciativas de calidad de datos sin planes o con planes mal definidos que introducen riesgos, pasan detalles por alto y llevan a cabo esfuerzos redundantes. Por el contrario, una metodología estratégica y sistemática permite evaluar adecuadamente su proyecto de calidad de datos, involucrar a las partes interesadas de negocio y TI, definiendo funciones y responsabilidades, y dotar de los procesos y las herramientas adecuadas para abordar el reto de la calidad de los datos. Como en toda tarea que hay que abordar, si desde el primer momento están bien definidas las metas, los roles y las actividades que hay que hacer, las probabilidades de éxito aumentan.

Instituir un programa de calidad de los datos dentro de una organización va mucho más allá de la adquisición de herramientas de limpieza de datos, la creación de un consejo de administración de datos o el hecho de documentar correctamente una serie de procesos. El programa de calidad de datos consiste en un **ciclo iterativo de evaluación**, **planificación**, **ejecución**, **gestión** y **revisión**.

Lo que significa que en el programa se determina el modo de actuar, se define objetivos, se designa a los responsables de cada iniciativa y las acciones que hay que tomar. Esto requiere procesos reproducibles, apalancados en los ins-

trumentos adecuados y en personas con formación y habilidades adecuadas, como se ilustra en la figura 2.

Figura 2. Proceso completo del programa de calidad de datos.



Fuente: David Cabanillas

El programa de calidad de datos consiste en un listado de elementos:

- **Madurez**: permite entender en qué estado se encuentra nuestra organización respecto a la calidad.

- **Expectativas**: permiten delimitar qué esperamos de la aplicación de un programa de calidad de datos. Es decir, se trata de prever/cuantificar beneficios.

- **Medición**: permite medir el estado de la organización antes y después de la aplicación del programa de calidad de datos.

- **Políticas**: definen la guía que hay que seguir sobre la calidad de los datos.

- **Procesos**: identifican y delimitan procesos de la organización que tienen contacto con los datos.

- **Gobierno**: define los roles dentro de la organización con respecto a la calidad de los datos.

- **Estándares**: establecen las normas sobre las que se debe de regir la calidad de los datos.

- **Metodologías**: se trata de las herramientas y técnicas utilizadas para conseguir una calidad de datos adecuada.

A continuación, revisaremos con detalle cada uno de estos elementos.

2.2. Madurez respecto a la calidad del dato

Para poder adoptar un programa de calidad de datos, es necesario comprender **en qué situación se encuentra nuestra organización**. Este asesoramiento se lleva a cabo a través de un modelo de madurez para la calidad del dato. El modelo de madurez de calidad de datos utilizado en este material se basa en el modelo de madurez de capacidades (CMM, que es el acrónimo de *capability maturity model*), desarrollado por el Instituto de Ingeniería de Software de la Carnegie Mellon University.

Lectura complementaria

Loshin, D. (2010). *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufmann.

Para adoptar un enfoque de gestión del rendimiento para la calidad de los datos, es necesario ir más allá de soluciones *ad hoc*, es útil visualizar cómo la gestión de la calidad de los datos encaja con todas las actividades dependientes de la información dentro de la organización. Las diferencias en la madurez de las organizaciones se miden por los procesos y las personas encargadas de la identificación de datos defectuosos.

Solución *ad hoc*

Una solución *ad hoc* es aquella que está específicamente elaborada para un problema o fin precisos y, por tanto, no es generalizable ni utilizable para otros propósitos.

El estado de madurez de las organizaciones se puede dividir en cinco estados (que van desde inicial, en el que no hay calidad del dato y se aplican soluciones puntuales, hasta optimizado, en el que existe un programa integrado en todos los procesos, como se muestra en la tabla 3).

Aunque los modelos de madurez fundamentados en CMM son los más extendidos en el ámbito de la calidad del dato, también es posible encontrar otros modelos fundamentados en cuatro etapas, como ilustra la tabla 4.

Buenas prácticas

Se refiere a toda experiencia que se guía por principios, objetivos y procedimientos apropiados que se adecuan a una determinada perspectiva normativa o a toda experiencia que ha arrojado resultados positivos.

2.3. Expectativas respecto a la mejora de los datos

Antes de aplicar un programa, es útil especificar de antemano sus expectativas con respecto a la calidad de los datos y los métodos que se utilizarán para eva-

Tabla 3. Estado de madurez de las organizaciones.

Estado	Soluciones	Colaboración
Inicial	Las soluciones para los problemas son *ad hoc*.	Las soluciones no son compartidas, lo que impide la replicación de la solución.
Repetible	Se identifican y evalúan las fuentes de la baja calidad del dato.	Se comparten las buenas prácticas entre los miembros de la organización.
Definido	Existe un entorno para monitorizar la calidad del dato.	El equipo de calidad de datos documenta problemas y soluciones.
Gestionado	Se evalúa y mide el impacto de la calidad de datos.	La información es compartida y se generan informes de impactos sobre posibles problemas.
Optimizado	Las mejoras estratégicas y la supervisión continua del proceso del ciclo de vida de los datos mediante paneles se aplican en toda la organización.	El entorno de calidad del dato incluye oportunidades de mejora.

Tabla 4. Estado de madurez de las organizaciones.

Estado	Importancia de la calidad del dato para la organización	Enfoque
Desconocido	Limitada	Se toman soluciones cuando la información suele ser inferior a la estándar.
Reactivo	Se comienza a reaccionar a los problemas de calidad de los datos, ya que impactan en el rendimiento del negocio.	Inversión en respuesta a un evento que ha causado problemas.
Proactivo	Se comienza a definir funciones y a crear figuras de gestión.	Se empieza a comprender el valor de los activos de datos más claramente, y a tener un proceso más estructurado para su análisis.
Optimizado	*Business as usual*, las decisiones se toman sobre los datos.	Vínculo entre la calidad de los datos y el rendimiento financiero.

luarlo. Por ejemplo: ¿se considerará aceptable una tasa de error en la variable x de menos del 1%?

En un ámbito de negocio, es posible hacerse preguntas como las siguientes:

- ¿Cómo ha disminuido el rendimiento debido a los errores?
- ¿Qué porcentaje de tiempo se gasta en la reelaboración de procesos fallidos?
- ¿Cuál es la pérdida de valor de las transacciones que fallaron debido a la falta de datos?
- ¿Con qué rapidez podemos responder a las oportunidades emergentes si disponemos de datos de calidad?

Sin embargo, en diferentes fases de madurez, se tendrán distintas expectativas. La tabla 5 relaciona expectativas con madurez de las organizaciones.

Es decir, las expectativas a la hora de mejorar la calidad de los datos no solo deben ser de un ámbito técnico; el valor añadido de la mejora de la calidad de los datos tiene que estar vinculado a la satisfacción de las expectativas de

Tabla 5. Expectativas para cada estado de madurez de las organizaciones

Estado	Caracterización
Inicial	La actividad de calidad de los datos es reactiva. No hay capacidad para identificar ni documentar las expectativas de calidad de los datos.
Repetible	Anticipación limitada de ciertos problemas de datos, y se identifican y notifican errores sencillos.
Definido	Las dimensiones de la calidad de los datos se identifican y documentan. Existe la capacidad de validar los datos utilizando reglas de calidad de datos definidas, así como métodos para evaluar el impacto en la parte de negocio.
Gestionado	La validez de los datos se inspecciona y supervisa. El análisis del impacto en la parte de negocio se hace de manera global. Y los resultados del análisis se han tenido en cuenta en la priorización de las expectativas.
Optimizado	Puntos de referencia de calidad de datos definidos. Las expectativas de calidad de los datos están vinculadas a objetivos de negocio. Es posible anticipar los niveles de calidad y fijar metas de mejora. Finalmente, se añaden controles para validación de datos integrados en los procesos de negocio.

negocio. Esto implica identificar los impactos empresariales, sus problemas y causas, y luego cuantificar los costes para eliminar todos los problemas relacionados con los datos y sus beneficios asociados.

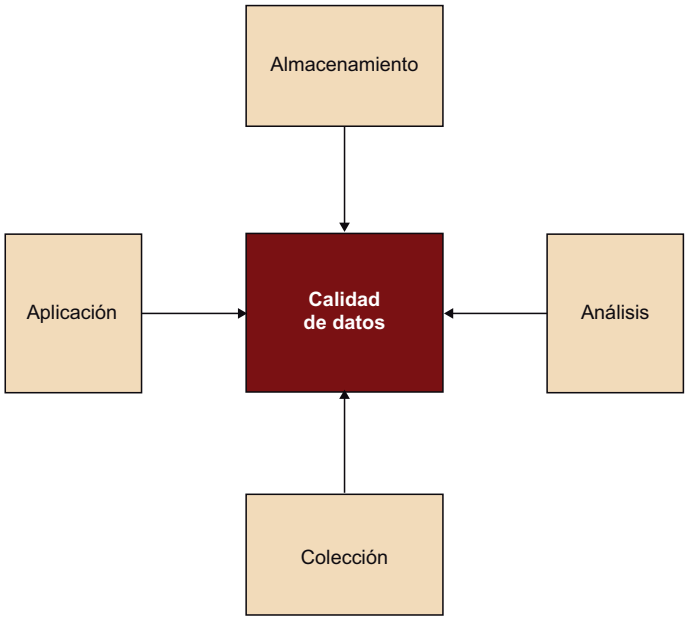
2.4. Medición

Encontramos cuatro grandes bloques que relacionan los datos con la organización, tal y como se ilustra en la figura 3*:

*** Más información en: https://goo.gl/q6RQ8u**

- **Aplicación**: el propósito para el que se recogen los datos.
- **Colección**: los procesos que acumulan los datos.
- **Almacenamiento**: procesos y sistemas utilizados para archivar los datos.
- **Análisis**: el proceso de comprender los datos para responder a la aplicación.

Figura 3. Programa de calidad de datos



Fuente: David Cabanillas

Como es posible imaginar, en cada uno de estos bloques la calidad se puede ver comprometida y es necesario controlarla y medirla.

¿Qué es importante medir? Debemos tener diferentes dimensiones para la calidad del dato*:

*** Más información en: https://goo.gl/Y22vhD**

- **Coherencia**: ¿se definen y entienden los elementos de datos de forma coherente?
- **Integridad**: ¿la estructura de datos y relaciones entre entidades y atributos se mantiene de manera consistente?
- **Completas**: ¿se presentan todos los datos necesarios?
- **Oportunas**: ¿se dispone de los datos cuando es necesario?
- **Accesibles**: ¿los datos son fácilmente accesibles, comprensibles y utilizables?
- **Validez**: ¿los valores de los datos están dentro de los rangos aceptables definidos por negocio?
- **Precisión:** ¿los datos representan con exactitud la realidad o una fuente verificable?

Los cinco primeros atributos pertenecen al contenido y la estructura de los datos, y cubren una multitud de aspectos que comúnmente se asocian con datos de mala calidad: errores de entrada de datos, reglas empresariales erróneas, registros duplicados y valores de datos que faltan o son incorrectos. Sin embargo, los datos sin defectos carecen de valor si no es posible entenderlos o acceder a los mismos. Por este motivo, las dos últimas dimensiones se evalúan mejor mediante entrevistas y encuestas a los usuarios de los datos, o por medio de métodos estadísticos.

Además de las medidas cuantitativas, también deben considerarse las medidas cualitativas. Algunos ejemplos incluyen:

- **Medidas de satisfacción del negocio**: miden el aumento/disminución en la satisfacción del negocio con la mejora en los datos.
- **Medidas de productividad**: consisten en el porcentaje de veces que el consejo de gobernanza de datos detectó y eliminó proyectos redundantes intra o interdepartamentales.
- **Oportunidad de negocio/medidas de riesgo**: miden el beneficio y el aumento de la competitividad vinculados a la calidad del dato.

- **Medidas de cumplimiento**: permiten entender el comportamiento de los usuarios respecto a sus niveles de acceso y actualización de datos.

Es muy importante establecer las medidas de calidad de datos más importantes para la organización. Las métricas pueden ser generales, como las que hemos discutido, o bien vinculadas a uno de los ámbitos del gobierno del dato. Esto es necesario para establecer una línea base para la calidad de sus datos y para monitorizar el progreso de las iniciativas en lo que a la calidad de datos se refiere.

Lectura complementaria

Pandey, R. K. (2014). *Data Quality in Data warehouse: problems and solution*. PhD Scholar, Surguja university (Chhattisgarh), India.

2.5. Políticas o reglas sobre los datos

¿Qué principios o reglas deben incluirse en su política de calidad de datos? El punto de partida es conocer qué es una política (en el contexto de una organización).

Se entiende por **política** un principio o regla para guiar las decisiones y lograr resultados.

Cabe comentar que el término no se utiliza normalmente para denotar lo que realmente se hace. Esto se conoce normalmente como procedimiento o protocolo.

La respuesta a la pregunta anterior es abierta, aunque tradicionalmente incluye responder a las siguientes preguntas:

- **Propósito**: ¿por qué necesitamos la política de calidad de datos? Sus propósitos deben de ser claros y directos.

- **Antecedentes**: ¿por qué es importante ahora el programa de calidad de datos? ¿Cómo se alinea con otras políticas y objetivos estratégicos en la organización? Esta sección es importante para proporcionar contexto a la política y dejar claro quién se beneficiará y por qué, además de ser útil para explicar la historia de la calidad de los datos en la organización.

- **Alcance**: ¿en qué circunstancias se debe habilitar la política? Por ejemplo, ¿se aplica la política a todos los datos de la organización? ¿Qué pasa con los datos de terceros?

- **Roles y responsabilidades**: ¿qué grupos (o roles individuales) serán responsables de asegurar que se ejecute la política? ¿La política será gobernada de manera centralizada o federada entre TI y negocio? ¿Qué se espera de cada rol?

- **Declaración**: ¿cómo se tratarán situaciones y conflictos específicos? Aquí es donde las medidas de la política de la calidad de los datos entran en juego.

- **Definiciones**: ¿qué se entiendo por calidad de datos? Es necesario incluir una sección de definiciones para que todos puedan entender cualquier acrónimo o términos poco comunes, y poner en común el argot entre TI y negocio.

- **Legislación**: ¿qué leyes y directivas hay que cumplir? Es una buena idea incluirlas en esta sección y proporcionar una mayor profundidad en los procedimientos y marcos individuales dictados por esos actos, pero tener supervisión ejecutiva de hasta qué punto son exactamente críticos los datos desde un punto de vista legal es siempre beneficioso.

- **Documentos de referencia**: ¿qué otras políticas y normas están vinculadas a esta política? ¿Se usaron otros documentos para formular esas políticas?

En la siguiente tabla 6, se relacionan las políticas con el estado de madurez de las organizaciones.

Tabla 6. Políticas para cada estado de madurez de las organizaciones

Estado	Caracterización
Inicial	Las políticas son informales y no están documentadas. Las acciones repetitivas son tomadas por diferentes miembros del personal sin coordinación.
Repetible	La organización intenta consolidar conjuntos de datos de fuente única de datos. Existen políticas iniciales.
Definido	Se establecen directrices personalizadas para establecer los objetivos de gestión de la calidad del dato, y las mejores prácticas están definidas.
Gestionado	Políticas establecidas y coordinadas en toda la empresa.
Optimizado	Notificación automatizada del incumplimiento de las políticas de calidad de datos.

2.6. Procesos o tareas sobre los datos

Todos los sistemas y procesos existentes para la recopilación, el registro, el análisis y el reporte de datos deben asegurar que sean exactos, válidos, fiables, oportunos, relevantes y completos dentro de la organización. Las cuatro principales familias de procesos son:

- Identificar el problema.

- Valorar el problema y su afectación e importancia.

- Acción que hay que llevar a cabo y personas involucradas.

- Evaluación de los resultados y medidas de mejoras.

Aplicar un conjunto de reglas de validación de datos a un conjunto de datos una sola vez proporciona información sobre el estado actual de los datos, pero no refleja necesariamente cómo las modificaciones y actualizaciones del sistema han mejorado la calidad general de los datos en la organización.

Sin embargo, el seguimiento de los niveles de calidad de los datos a lo largo del tiempo como parte de un proceso de monitorización continuo proporciona una visión histórica de cuándo y cuánto mejoró la calidad de los datos.

Los niveles de calidad de los datos pueden ser rastreados periódicamente (por ejemplo, a diario) para mostrar si el nivel medido en la calidad de los datos está dentro de un rango aceptable, comparado con los límites históricos de control.

¿Cómo controlar la evolución? A través de gráficos y herramientas de visualización de métricas.

- Los gráficos de control estadístico pueden ayudar a notificar a los administradores de datos cuándo un evento de excepción está afectando a la calidad de los datos y dónde buscar para rastrear el proceso de información incorrecta.

- Estas métricas se consolidan como herramientas de visualización de métricas, ya sea a través de un cuadro de mando (sistema de indicadores visual) y/o *scorecards* (sistema de métricas que busca mostrar el progreso hacia los objetivos).

- Estos sistemas evalúan el impacto empresarial de los defectos de datos, y determinan las dimensiones de la calidad de los datos que se pueden utilizar para definir las métricas de calidad de los datos en un formato visual, para que la organización tome las decisiones que considere oportunas.

En el programa de calidad de datos, se debe incluir:

- Plantillas de inspección de datos estandarizados.
- Calidad de los datos operacionales.
- Seguimiento de temas y soluciones.
- Intervención manual cuando sea necesario.
- Integridad del intercambio de datos.
- Planificación de contingencias.
- Validación de datos.

En la medida de lo posible, los procesos deben operar sobre una base de «la primera vez», en lugar de emplear la limpieza o manipulación de datos para obtener la información requerida. Es decir, se trata de poner el esfuerzo en el hecho de que los datos sean correctos desde el principio, para que causen el menor impacto negativo en la organización.

En los casos en los que no sea posible, cualquier ajuste de datos debe seguir un proceso claro y documentado, que pueda verificarse fácilmente. Hay que incorporar controles apropiados para reducir la probabilidad de error. Cuando se obtengan datos de terceros u otros departamentos, se acordará un protocolo para garantizar que estos datos cumplan los mismos niveles que se han definido en el programa.

2.7. Gobierno

El gobierno del dato y la calidad de datos están muy relacionados. Si los datos fueran agua:

- ***Data governance*** sería el grifo, encargado de que las personas tengan las herramientas y el conocimiento adecuados, así como de la distribución del agua en las cantidades adecuadas y a las personas correctas.

- ***Data quality*** sería la depuradora, encargada de que el agua sea buena, no esté contaminada y mantenga un nivel aceptable de calidad de forma continua.

En esta relación entre el gobierno y la calidad, es necesario recordar los roles que están involucrados en la gestión de calidad de datos:

- **Jefe de proyecto**: responsable de supervisar el programa de inteligencia de negocio o proyectos individuales y de administrar las actividades diarias basadas en el alcance, el presupuesto y las restricciones de horario. También el nivel de calidad de datos necesario; para ello, interactúa con los representantes de negocios para establecer los requisitos de calidad de los datos.

- **Administrador**: ayuda a la organización a comprender el valor y el impacto del entorno de *business intelligence*, así como a abordar los problemas que surgen. A menudo, los problemas de calidad de datos se detectan durante los proyectos de inteligencia de negocios, y el agente de cambio de organización puede desempeñar un papel instrumental: ayudar a la organización a entender la importancia de tratar con los problemas.

- **Analista de negocio y/o datos**: transmite los requisitos del negocio, y estos incluyen requisitos detallados de calidad de los datos. El analista de

datos refleja estos requisitos en el modelo de datos y en los requisitos para los procesos de adquisición y entrega de datos. Juntos, aseguran que los requisitos de calidad se definen, se reflejan en el diseño y se transmiten al equipo de desarrollo.

- **Administrador de datos**: el administrador de datos es responsable en última instancia de la gestión de datos como un activo corporativo.

2.8. Estándares

En el contexto de la calidad del dato, es importante contar con estándares o normas sobre los datos.

> Se entiende como **estándar** un proceso, protocolo o técnica utilizados para hacer una tarea concreta.

Estas normas están destinadas a ser utilizadas con flexibilidad para promover una mejor calidad de los datos, en lugar de constituir un conjunto rígido de requisitos. También pueden ser apropiados enfoques alternativos para lograr estos objetivos, siempre que consigan el resultado de obtener datos confiables que apoyen una toma de decisiones informada y que sea posible de llevar a cabo.

Las normas son esenciales para asegurar que:

- La recogida de datos es exacta y coherente en toda la organización.
- Los registros se completan y procesan con precisión.
- Los datos se mantienen seguros y confidenciales.
- Las salidas de datos pueden compararse interna y externamente.

En la tabla 7, se relacionan los estándares con el estado de madurez de las organizaciones.

Tabla 7. Estándares para cada estado de madurez de las organizaciones

Estado	Caracterización
Inicial	No se han definido estándares ni existen definiciones de los datos.
Repetible	Definiciones de elementos de datos y uso de metadatos.
Definido	Estándares de datos de negocio y gestión de metadatos.
Gestionado	Normas para el intercambio gestionadas a través del proceso de supervisión de estándares de datos.
Optimizado	Conformidad con estándares a través de una estructura orientada a las políticas.

2.9. Tecnología

La tecnología debe dar soporte a los puntos anteriores. Entre lo que se espera de la tecnología, listamos:

- Procedimientos estandarizados para el uso de herramientas de calidad de datos, para la evaluación y mejora de la calidad de los datos.
- Uso de técnicas basadas en reglas de negocio para la validación de datos.
- Corrección automática de datos guiada por políticas y reglas de negocio definidas. Cuando hablamos de automatización, puede fundamentarse en reglas sencillas (*if ... then ... else*) o incluso en patrones complejos fundamentados en *machine learning*.
- Análisis de impacto y escenarios hipotéticos compatibles con el panel de control y las herramientas de generación de informes.

En la siguiente tabla 8, se relaciona la tecnología con el estado de madurez de las organizaciones.

Tabla 8. Estándares para cada estado de madurez de las organizaciones

Estado	Caracterización
Inicial	Se desarrollan internamente rutinas *ad hoc*.
Repetible	Se dispone de herramientas para evaluar la calidad de los datos. Para el análisis de datos, la estandarización y la limpieza.
Definido	Procedimientos estandarizados para el uso de herramientas de calidad de datos, para la evaluación y mejora de la calidad de los datos.
Gestionado	Corrección automática de datos guiada por políticas de gobernanza y reglas de negocio.
Optimizado	Los usuarios no técnicos pueden definir y modificar dinámicamente las reglas y las dimensiones de la calidad de los datos.

2.10. Metodologías

¿Cómo detectar que no hay calidad de datos? Encontramos principalmente dos metodologías:

- De dentro hacia fuera.
- De fuera hacia dentro.

2.10.1. Metodología de dentro hacia fuera

El método de dentro hacia fuera comienza con el análisis de los datos. Se lleva a cabo un examen de los mismos (sobre las fuentes de datos existentes), utilizando la tecnología de perfilado de datos. Las imprecisiones de datos se revelan a partir del proceso, y luego se analizan juntas para generar un conjunto de cuestiones sobre datos, para su posterior resolución.

El análisis debe ser llevado a cabo por un analista de datos. La metodología comienza con un conjunto completo y correcto de reglas que definen la exactitud de los datos. Se trata de trabajar con los metadatos (es decir, datos sobre los datos).

Por ejemplo:

- El dato *temperatura* puede tener el metadato *rango* que, para datos en España, podría ser de -20 a 45 grados centígrados.

- Otro metadato en este contexto sería *temperatura_ciudad*, que relacionaría el dato *temperatura* con el dato *ciudad*.

El proceso de determinar los metadatos correctos implica inevitablemente relacionar TI y negocio. El analista debe detectar el comportamiento en los datos, y requerirá consulta para determinar por qué es así. Esto, con frecuencia, conduce a modificaciones en los metadatos. Estas consultas son siempre productivas, porque la pregunta siempre está respaldada por información de los datos.

2.10.2. Metodología de fuera hacia dentro

Este método busca problemas en el negocio, no en los datos. Identifica hechos que sugieren que los problemas de calidad de los datos están teniendo un impacto en el negocio.

Se buscan eventos como devoluciones de mercancías, reclamaciones de clientes, retrasos en la obtención de productos de información completados, altas cantidades de trabajo requeridas para obtener productos de información producidos, etc.

En este enfoque, se usan entrevistas de negocio que permiten determinar el nivel de confianza en la exactitud de los datos procedentes de los sistemas de información, así como su nivel de satisfacción respecto a conseguir todo lo que necesitan.

También puede incluir la búsqueda de decisiones tomadas por la corporación que fueron decisiones equivocadas. A continuación, los datos se examinan para determinar si tienen inexactitudes que contribuyen a los problemas, así como el alcance de la contribución. Generalmente, este examen apunta al problema específico y no es un ejercicio de perfiles de datos exhaustivo, aunque podría extenderse si hay la evidencia de un problema de calidad generalizado de datos.

2.10.3. Comparación de métodos

Ninguno de los enfoques es superior al otro: los dos aportan valor al proceso. El método de *dentro hacia fuera* es generalmente más fácil de lograr y necesita menos tiempo para su ejecución. Un solo analista puede analizar una gran cantidad de datos en poco tiempo. El enfoque *fuera hacia dentro* requiere dedicar mucho tiempo a entrevistar a personas de otros departamentos.

Veamos dos ejemplos:

- Imaginemos una empresa donde, en la generación de pedidos, hay un error en la inclusión del identificador del proveedor. Frecuentemente, en función del volumen de pedido, en ciertos sectores los proveedores acceden a descuentos (por fidelización). En este escenario, el proveedor puede estar perdiendo cantidades significativas cada año debido al error, y era completamente desconocedor de lo que estaba sucediendo. El enfoque de *dentro hacia fuera* es el adecuado en este caso.

- Lo opuesto también es cierto. Imaginemos que el sistema de información asigna identificadores erróneos a algunas piezas de un fabricante. Esto deriva en errores de envío y devoluciones. El análisis de *fuera hacia dentro* podría identificar la disparidad en los códigos.

2.11. La calidad del dato en el contexto del gobierno del dato

En el contexto de gobierno del dato, la calidad del dato es una función más para llevar a cabo. Como ya sabemos, cada función tiene distintas actividades (planificación, control, de desarrollo y operativas), cada una de las mismas llevada a cabo por el rol correspondiente.

Para la calidad de datos, estas actividades son:

- Desarrollar y promover la concienciación sobre la calidad de los datos (actividad operativa).

- Perfilar, analizar y evaluar la calidad de los datos (actividad de desarrollo).

- Definir requisitos de calidad de datos y reglas de negocio (actividad de desarrollo).

- Probar y validar los requisitos de calidad de los datos (actividad de desarrollo).

- Definir indicadores de calidad de datos y niveles de servicio (actividad de planificación).

- Medir y monitorizar la calidad de los datos (actividad de control).

- Gestionar problemas de calidad de datos (actividad de control).

- Corregir defectos de calidad de datos (actividad operativa).

- Diseñar e implementar procedimientos de calidad de datos operativos (actividad de desarrollo).

- Monitorizar los procedimientos operacionales y el rendimiento de la gestión de calidad de datos (actividad de control).

- Auditar la calidad de datos (actividad de control).

El programa de calidad de datos cubre estas funciones, que forman parte del marco más general.

3. Desarrollando un programa de calidad de datos

3.1. Desarrollo de un programa de calidad de datos

Como ya hemos comentado, un programa de calidad de datos proporciona una guía de buenas prácticas para mejorar y aprovechar mejor los datos. Implementar un programa de calidad de los datos en una organización exige superar una serie de desafíos que deben ser conocidos y superados. Las razones por las cuales las organizaciones no siguen una iniciativa formal y planificada de gestión de la calidad de los datos incluyen las siguientes:

- Ninguna de las unidades de negocio o departamento siente que es responsable del problema.
- Se precisa cooperación interdepartamental.
- Se requiere que la organización reconozca que tiene problemas significativos.
- Se necesita disciplina.
- Se requiere una inversión de recursos financieros y humanos.
- Se percibe que es un proceso costoso.
- El retorno de la inversión es, a menudo, difícil de cuantificar.

Dentro del desarrollo del programa, podemos identificar dos grandes tareas cruciales:

1) Evaluación. Es decir, desde dónde partimos y las medidas que hay que aplicar.

2) Seguimiento (de las soluciones aplicadas). Es decir, una vez detectados los errores y las oportunidades, y decididas las tareas y acciones que hay que llevar a cabo, valorar qué impacto han tenido estas en la organización, e insistir, si es necesario, en el proceso de mejora.

Estas dos grandes tareas se descomponen en diferentes pasos que se repiten de forma cíclica, como ilustra la figura 4:

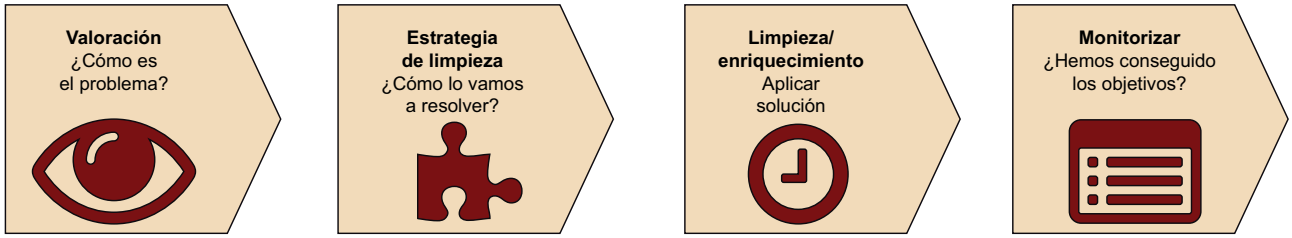
- **Valoración**: en este paso, se identifican los elementos que deben evaluarse para la calidad de los datos. Normalmente, estos serán elementos de datos considerados críticos para las operaciones empresariales y los informes de gestión asociados, y será necesario evaluar qué dimensiones de calidad de datos utilizar y su ponderación asociada.

- **Estrategia de limpieza**: para cada dimensión de calidad de datos, se definen los valores o rangos que representen datos de calidad buena y mala. Esto permite clasificar los datos y determinar cuáles necesitan ser tratados y qué medida aplicar.

- **Limpieza y enriquecimiento**: en este paso, se aplican los criterios de limpieza y de mejora de los datos y procesos para prevenir errores futuros.

- **Monitorizar**: finalmente, se revisan los resultados y se determina si la calidad de los datos es aceptable o no. Al introducir cambios en los modelos de negocio, es posible que aparezcan nuevos problemas de calidad de datos.

Figura 4. Pasos en el marco de la calidad de datos



Fuente: David Cabanillas, adaptado de Deloitte

Para controlar el flujo anterior y asegurar que se cumplan unos plazos de ejecución, es aconsejable llevar a cabo las siguientes acciones:

1) Determinar un calendario de actuación: la programación ha de estar expresada en términos claros, incluyendo un glosario, si se considera necesario; y debe ser compartida entre la parte de negocio y TI.

2) Establecer la frecuencia de evaluación: la periodicidad de las acciones de descubrimiento y evaluación de la calidad de los datos se determinará en función de la criticidad de la información contenida en los datos y la relevancia de las áreas a las que afecten.

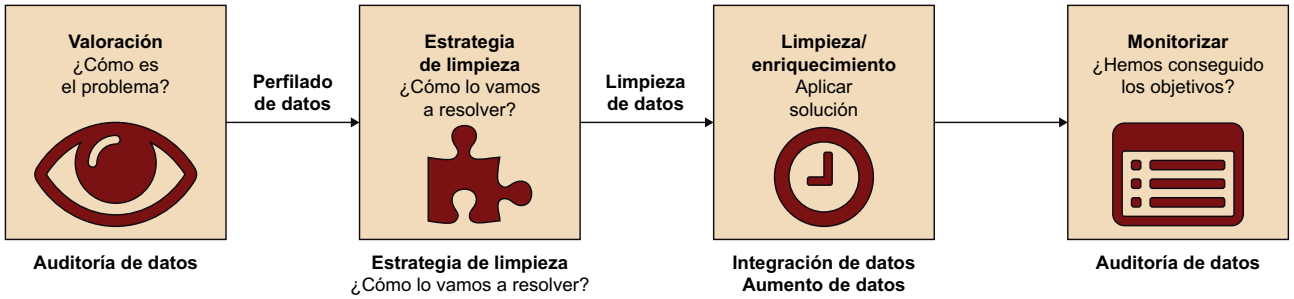
3) Designar a los responsables: la ausencia de propietarios de los datos es el principal problema de fondo en un elevado porcentaje de cuestiones relacionadas con la calidad de datos. Asignar responsabilidades y llegar a un consenso es la mejor prevención.

4) Definir los requisitos de *reporting:* la retroalimentación es imprescindible para fomentar el flujo de conocimiento y garantizar la actualización. Para ma-

ximizar la eficiencia de esta comunicación, han de establecerse los términos en los que se llevará a cabo, antes de comenzar el programa.

En los siguientes subapartados, se describen las tareas relacionadas con cada uno de los pasos del programa. La figura 5 relaciona cada una de estas tareas con el paso concreto del programa de calidad de datos en la que se encuentra.

Figura 5. Tareas y pasos en el marco de la calidad de datos



Fuente: David Cabanillas, adaptado de Deloitte

3.1.1. Perfilado de datos

Una de las tareas iniciales es el perfilado de datos. Dentro del programa de calidad de datos, el perfilado corresponde a las fases de **valoración** y **estrategia de limpieza**. Tal y como comenta Ralph Kimball:

> Se entiende como **perfilado de datos**, o *data profiling*, el análisis de contenido, estructura y anomalías de los datos.

El resultado de esta tarea son perfiles de datos, lo que proporciona un medio metódico, repetible, consistente y basado en métricas para evaluar sus datos. ¿En qué consiste el perfilado de datos? Incluye distintos tipos de análisis para clasificar el dato:

- **Análisis de completitud**. Da respuesta a la siguiente pregunta: ¿con qué frecuencia un atributo tiene valores, está vacío o es nulo?

- **Análisis de unicidad**. Da respuesta a las siguientes preguntas: ¿cuantos valores únicos encontramos en un atributo? ¿Hay duplicados? ¿Este fenómeno es esperado y normal?

- **Análisis de distribución**. Responde a la pregunta: ¿cuál es la distribución de frecuencias de los valores para un determinado atributo?

- **Análisis de rangos**. Responde a la pregunta: ¿cuáles son los valores mínimos, máximos, la media y el promedio para los valores de un determinado atributo?

- **Análisis de patrones**. Responde a las siguientes preguntas: ¿qué formatos encontramos para un atributo? ¿Cómo se distribuyen los valores en dichos formatos?

  Imaginemos que tenemos la base de datos de clientes, y uno de los atributos es el DNI (Documento Nacional de Identidad en España). Existen distintas reglas que definen y validan el hecho de que un valor en esta columna es, efectivamente, un DNI:

  - El valor debe tener nueve caracteres.
  - Los primeros ocho deben ser números.
  - El último debe ser una letra.
  - La letra se calcula tomando todas las letras, excepto la Ñ, la I y la O, porque pueden inducir a errores, en un orden concreto (que no es el orden alfabético lógico, sino este: TRWAGMYFPDXBNJZSQVHLCKET), y seleccionando la que coincide en la posición igual al resto de dividir el número del DNI entre 23.

Usar estas reglas corresponde en este caso a aplicar el análisis de patrones. En algunos casos, estas reglas son deducibles; en otros, la combinación de estos análisis ayuda a descubrir tanto los requisitos de calidad de datos como la forma de evaluar la calidad del dato.

Cabe comentar que los requisitos y reglas descubiertos no solo deben aplicarse sobre el conjunto de datos que se está analizando, sino que frecuentemente implican transformar las fuentes de origen, incluso la modificación de sistemas de información y de procesos de negocio.

3.1.2. Limpieza de datos

Una vez conocidos los problemas que tienen los datos, es necesario aplicar acciones para corregir dichos problemas.

Se entiende como **limpieza de datos**, o *data cleansing*, el proceso de arreglar o borrar datos que son incorrectos, incompletos, con un formato incorrecto o duplicados.

También es posible hacer referencia a este concepto como *data scrubbing* o *data correction*. Dentro del programa de calidad de datos, la limpieza de datos corresponde a las fases de **estrategia de limpieza** y **limpieza y enriquecimiento**.

El objetivo de esta tarea es mejorar la confianza de la organización en sus datos, pero podemos encontrarnos en diferentes escenarios. Por ejemplo, problemas relacionados con:

- **Datos procedentes de una única fuente**. Los errores:

– En un nivel de esquema, hacen referencia al diseño (falta de integridad de las restricciones, diseño ineficiente) y se observan por problemas de unicidad, integridad referencial, etc.

– En un nivel de instancia, hacen referencia a problemas de introducción de datos y se observan duplicados, valores contradictorios, errores ortográficos, etc.

- **Datos procedentes de distintas fuentes de origen**. Los errores:

– En un nivel de esquema, hacen referencia a diseños de datos y esquemas heterogéneos, lo que se traduce en conflictos de estructura, de nombres, etc.

– En un nivel de instancia, hacen referencia a inconsistencia en los datos, lo que se traduce en inconsistencias al agregar, combinar y cruzar datos.

Cuando se encuentran problemas de calidad de datos (por ejemplo, al importar datos en el almacén de datos), hay cuatro acciones viables que se pueden tomar:

- Rechazar el error.
- Aceptar el error.
- Corregir el error.
- Aplicar un valor predeterminado.

Cuando la precisión es más importante que la completitud, puede ser apropiado rechazar el error. Cuando se sabe que los datos contienen errores, pero están dentro del nivel de tolerancia, entonces puede ser apropiado aceptar el error. Cuando el valor se puede determinar, entonces el error se puede corregir. Por último, cuando no se puede determinar el valor correcto y la integridad es muy importante, entonces un valor predeterminado puede ser sustituido por los datos erróneos. En cualquier caso, es importante que los administradores de los datos entiendan las implicaciones de la solución elegida y esta esté consensuada con la parte de negocio.

¿Y qué opciones tenemos para la limpieza de datos? Disponemos de diferentes técnicas:

- **Duplicación**: permite detectar registros duplicados.

- **Búsqueda y reemplazo**: permite encontrar y reemplazar registros a partir de de un criterio.

- **División**: permite dividir un registro en varios valores según un criterio.

- **Diccionarios/referencias**: permiten validar registros respecto a un conjunto de valores que se aceptan como referente de calidad.

Estas técnicas suelen combinarse entre ellas para resolver los problemas de calidad, y pueden ser manuales o automáticas (en las que el sistema identifica y propone soluciones).

3.1.3. Auditoría de datos

El objetivo de la auditoría es comprender el grado en el que los problemas de calidad de los datos existen en nuestra organización, es decir, la extensión y la gravedad de los defectos de los datos. Este proceso implica revisar las métricas clave, aparte de los valores, para crear conclusiones. Los informes de auditoría pueden crearse para medir el progreso en el logro de los objetivos de calidad de los datos y cumplir con los acuerdos en la primera fase del programa de calidad de datos*.

*** Ejemplo de plantilla para informe: https://goo.gl/4TvXtj**

De este modo, el informe de auditoría determina:

- Declaraciones concretas sobre el *status quo* de la calidad de los datos.

- Recomendaciones de orientación, es decir, soluciones para optimizar la calidad de los datos a corto plazo y mantener el nivel alcanzado a largo plazo.

- Identificar posibles oportunidades de optimización de los procesos.

- Visión general del posible potencial de ahorro o beneficios conseguidos y los costes asociados a aplicar las soluciones.

Imaginemos que tenemos datos geolocalizados en nuestra organización. El proceso de auditoría seguiría este proceso:

- Chequeo de bases de datos internas.

- Chequeo de bases de datos externas.

- Revisar datos fuera de rango mediante una referencia GIS (*geographic information system*).

- Revisar datos fuera de rango con herramientas estadísticas.

- Informar de la calidad de los datos geolocalizados.

3.1.4. Integración de datos

Como es posible deducir, en los procesos de calidad de datos debemos transformar los datos con el objetivo de aumentar su calidad. En este sentido, no es necesario reinventar la rueda: dentro de la calidad de datos, una de las tareas más relevantes es la integración de datos.

Lectura complementaria

Curto, J. (2017). *Introducción al Business Intelligence*. Editorial UOC.

Se entiende por **integración de datos**, o *data integration*, el conjunto de aplicaciones, productos, técnicas y tecnologías que permiten una visión única y consistente de nuestros datos de negocio.

Es decir, la integración permite acceder a los datos presentes en diferentes fuentes, recuperarlos, combinarlos, transformarlos (siguiendo las reglas de calidad de datos) y aplicar los cambios en las fuentes de destino.

Imaginemos que tenemos un listado con dos archivos de productos. Una empresa puede haber vendido los mismos productos en diferentes sucursales, pero los productos pueden venderse bajo nombres distintos: la marca y los patrones de descripción en cada archivo se basaron en el personal de entrada de datos. El primer reto en la integración de datos es reconocer que el mismo cliente existe en cada una de las dos fuentes, y el segundo desafío es combinar los datos en una sola vista del producto (consolidación). Con los datos del cliente, a menudo encontramos un campo común, por ejemplo, el DNI, que se pueden utilizar para identificarlo e integrarlo.

3.1.5. Aumento de datos

El aumento de datos es el último paso. En este paso, se busca aumentar el valor del conjunto de datos inicial a través de la incorporación de datos externos.

Por ejemplo, este proceso puede ser usado para añadir las coordenadas geográficas a las direcciones de la base de datos de clientes.

Tenemos diferentes opciones respecto a las fuentes de datos:

- **De pago**: son fuentes de datos preparadas por terceros para su consumo. Pueden ser accesibles a través de ficheros independientes o una API (*application programming interface*). La calidad del dato la gestiona el proveedor, y el pago puede tener diferentes modalidades (pago único, suscripción, etc.). Entre estas fuentes, podemos considerar Crunchbase* o Bloomberg**.

*** Más información en: http://www.crunchbase.com**

**** Más información en: http://www.bloomberg.com**

- ***Open Data***: son fuentes de datos de terceros a las que se puede acceder libre y fácilmente. Como en el caso anterior, el proveedor asegura el nivel de calidad del dato. Algunos ayuntamientos ofrecen sus datos en esta modalidad. Por ejemplo, el Ayuntamiento de Barcelona***.

***** Más información en: http://opendata.bcn.cat**

- **Públicas**: son fuentes de datos disponibles de forma pública, pero cuyo acceso no está preparado para el consumo. Frecuentemente, estos datos pueden recuperarse de manera automática mediante técnicas de *web scrapping*. Un ejemplo de ello es recuperar información de Wikipedia.

- ***Crowdsourcing***: son fuentes de datos que se generan a partir de la colaboración de consumidores. Por ejemplo, pedir ayuda a nuestra comunidad para la traducción de nuestro *software*.

Open data

Los datos abiertos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, a lo sumo, al requerimiento de atribución y de compartirse de la misma manera en la que aparecen.

En la combinación de los datos propios y los de terceros, las técnicas de integración de datos juegan un papel fundamental.

3.2. Mejores prácticas

Como hemos aprendido durante este apartado, un programa de calidad de datos pasa por diferentes fases y tiene distintas tareas. Llevar a buen puerto este tipo de proyectos no es sencillo, puesto que requiere orquestar procesos, personas y tecnología. En este apartado, nos centraremos en revisar las mejores prácticas que pueden ayudar al despliegue de este tipo de proyectos.

Mejores prácticas

Las mejores prácticas son un conjunto de acciones que han rendido un excelente servicio en un determinado contexto y que se espera que, en contextos similares, rindan similares resultados.

La mayoría de las organizaciones dejan la calidad de los datos a los administradores de bases de datos (con la suposición de que si estos son dueños de los datos, son también responsables de su calidad), pero este enfoque no es siempre el más sensato. Las organizaciones deben adherirse a un programa de calidad de datos para asegurar que sus datos son de calidad. En el programa, se marcan los objetivos y hasta dónde se quiere y se debe llegar en la calidad de los datos, se describen los pasos y, finalmente, se evalúan los resultados. Como ya hemos comentado, para aplicar y seguir el programa, tanto negocio como TI deben coordinarse.

¡Recordad!

La prevención es mejor que la cura. Hay que tratar de evitar introducir datos de baja calidad en nuestros sistemas.

¿Qué debemos tener en cuenta como mejores prácticas? Existen una serie de principios clave que hay que considerar en el establecimiento de un programa eficaz:

- **Empezar de forma acotada**: el objetivo final es diseñar una visión completa del estado de calidad de datos en la organización. Sin embargo, esto puede ser una tarea inabordable directamente. Una buena hoja de ruta de implementación debe comenzar con un área acotada (por ejemplo, área de clientes) y construir sobre la misma el programa a lo largo del tiempo.

- **La calidad de los datos es un proceso continuo**: un programa no es un evento único, sino un proceso continuo que evolucionará con el tiempo. A menudo, se trata de un cambio cultural en una organización y, por lo tanto, su implementación lleva tiempo. El programa debe ser reevaluado periódicamente, y modificado cuando sea necesario.

- **Enfoque en los objetivos de calidad de los datos**: en este tipo de iniciativas, no se trata de solucionar rápidamente problemas de calidad de datos sin las raíces del problema. Es necesario establecer las medidas de calidad de datos importantes para la organización, entender los niveles de calidad de datos requeridos por el negocio y establecer los objetivos de calidad de datos apropiados.

- **Combinar calidad de datos en el desarrollo/implementación de software**: a menudo, la calidad de datos se considera un proceso posterior al desarrollo/implementación de software. Este enfoque reactivo supone un mayor esfuerzo para la organización. Es necesario establecer un enfoque proactivo que considere ya desde el principio la calidad en todos los procesos de negocio.

- **Empezar con los datos de alto retorno**: una forma de convencer a la organización de la necesidad de tener un programa global de calidad de datos consiste en centrarse inicialmente en las áreas específicas de datos que proporcionarán el retorno de inversión más alto si se aplican estas medidas.

- **Maximizar la participación multidepartamental en el programa**: la colaboración es crucial en el diseño e implementación de medidas tanto en un ámbito de negocio (usuarios de datos) como de TI (administradores de datos).

- **Hay más de una solución del programa**: los datos suelen fluir a través de una organización desde muchas fuentes/aplicaciones, a través de diferentes servicios de integración de datos y en muchos repositorios. La calidad de los datos se puede gestionar en cualquier punto del flujo de datos, dependiendo de las políticas de gestión de datos, los requisitos empresariales, la propiedad de los datos, la arquitectura del sistema, etc. Comprender las medidas de calidad de datos, el ciclo de calidad de datos y mantener los principios son aspectos clave para asegurar una base sólida del programa.

Flujo de datos

El flujo de datos es el movimiento que tienen los datos en un sistema determinado.

- Elementos que hay que tener en cuenta: aplicar la calidad del dato transforma la organización, de forma que es necesario documentar los estados antes y después. Esto significa que es conveniente tener en cuenta los siguientes elementos:

– Lista de conjuntos de datos y elementos que hay que tratar.
– Lista de tipos de datos y categorías.
– Catálogo, esquema o mapa de dónde residen los datos.
– Discusión de soluciones de limpieza por categoría de datos.
– Diagramas de flujo de datos existentes.

- Diagramas de flujo de trabajo existentes.
- Plan para decidir cuándo y dónde se accede a los datos para limpiarlos.
- Análisis de cómo cambiará el flujo de datos después de la implementación del proyecto.
- Discusión de cómo el flujo de trabajo cambiará después de la implementación del proyecto.
- Lista de actores afectados por el proyecto.
- Plan para educar a las partes interesadas en cuanto a los beneficios del proyecto.
- Plan para la formación de operadores y usuarios.
- Lista de medidas de calidad de datos y métricas para monitorizar.
- Plan de cuándo y dónde monitorizar.
- Plan para la limpieza inicial y, luego, regular.

Flujo de trabajo

Cómo se estructuran las tareas, cómo se llevan a cabo, cuál es su orden correlativo, cómo se sincronizan, cómo fluye la información que soporta las tareas y cómo se le hace seguimiento al cumplimiento de las tareas.

3.3. Impacto

Para que la parte de negocio y TI vayan de la mano en la aplicación del programa de calidad de datos, es necesario valorar su impacto antes de su ejecución. Son varios los puntos que hay que tener en cuenta en esta valoración:

- **Eficiencia operativa**: tiempo y costes de limpieza de datos o correcciones de procesamiento.
- **Métricas**: medidas de rendimiento inexactas para los empleados.
- **Riesgo/cumplimiento**: los datos incorrectos conducen a un riesgo *X*, violaciones de normas.
- **Ingresos**: coste de oportunidad perdida, identificación de oportunidades.
- **Productividad**: disminución de la capacidad de procesamiento directo mediante la automatización servicios.
- **Eficiencia de la adquisición**: mayor facilidad de uso para el personal (ventas, centro de llamadas, etc.), mayor facilidad de interacción para los usuarios.
- **Reducción**: reducción del tiempo desde el pedido hasta la entrega.
- **Rendimiento**: deterioro en la toma de decisiones.

Una buena metodología que hay que seguir es la utilización de una plantilla/*template* sobre los diferentes impactos que producirá el programa de calidad de datos. Un ejemplo es la estructura que se muestra en la figura 6:

- **ID problema:** identificador asignado para el problema.

- **Problema del dato:** descripción del problema.

- **Impacto:** descripción del impacto comercial atribuible a la emisión de datos; puede haber más de un impacto para cada problema de datos.

- **Medida:** medición en el grado de impacto.

- **Importancia:** una estimación de la cuantificación de los impactos en un ámbito monetario/temporal.

Figura 6. Template

Id problema	Problema del dato	Impacto	Medida	Importancia

Fuente: David Cabanillas

4. Técnicas y tecnología para la calidad del dato

El programa de calidad de datos se sustenta en técnicas y tecnología que permiten automatizar las principales tareas que hemos discutido en anteriores apartados.

4.1. Técnicas

Cuando hablamos de técnicas en el contexto de calidad de datos, hacemos referencia a dos tipos:

- Aquellas que permiten el análisis visual de los datos, con el objetivo de detectar anomalías.
- Aquellas que permiten automatizar el proceso de identificar problemas de calidad y su tratamiento.

4.1.1. Técnicas visuales

En el momento de identificar problemas de calidad del dato, una primera técnica es la visualización. Esta técnica puede apoyar el descubrimiento de patrones en los datos. La manipulación visual de datos (usando agregaciones, agrupamientos, clasificación, escalas de colores, diferentes tipos de gráficos, etc.) permite, a veces, identificar los problemas siguientes:

- Valores que faltan.
- Valores fuera de rango.
- Resultados de negocio que no encajan.

En este sentido, herramientas analíticas visuales como Tableau*, QlikSense**, GGobi*** o Improvise**** encapsulan estas técnicas y permiten a los analistas construir vistas multidimensionales de los datos que ayudan a evaluar los problemas de calidad de los datos.

* http://www.tableau.com
** http://www.qlik.com
*** http://www.ggobi.org
**** https://goo.gl/pMG6LS

Además, una vez creadas las métricas de control, el *scorecard* permite hacer un seguimiento del programa de calidad, como ilustra la figura 7.

Aunque las técnicas visuales son interesantes, no siempre resultan suficientes para el análisis de la calidad del dato, y por ello es necesario recurrir a técnicas de automatización.

Figura 7. Ejemplo de *scorecard*

DATA ELEMENT LEVEL

TABLE XYZ

Expected fields to be populated 100%

#	Table Column Name	8/4/2012	8/11/2012	8/18/2012	8/25/2012	9/1/2012	9/8/2012	Trend
1	Key Field 1	100%	100%	100%	100%	100%	100%	
2	Key Field 2	100%	100%	100%	100%	100%	100%	
3	Key Field 3	100%	100%	100%	100%	100%	100%	
4	Key Field 4	100%	100%	100%	100%	100%	100%	
5	Field 05	91%	72%	67%	70%	70%	70%	
6	Field 06	72%	78%	80%	81%	81%	83%	
7	Field 07	94%	96%	96%	100%	100%	98%	
8	Field 08	88%	74%	72%	72%	72%	70%	
9	Field 09	81%	74%	65%	70%	67%	64%	
10	Field 10	84%	70%	63%	72%	70%	66%	
11	Field 11	88%	74%	70%	74%	72%	70%	
12	Field 12	84%	74%	72%	74%	72%	72%	
13	Field 13	84%	74%	70%	74%	72%	68%	
14	Field 14	94%	98%	98%	95%	95%	98%	
15	Field 15	66%	74%	72%	70%	77%	79%	
16	Field 16	78%	80%	85%	88%	84%	83%	
17	Field 17	47%	52%	46%	44%	44%	47%	
18	Field 18	19%	15%	17%	16%	16%	17%	
19	Field 19	100%	100%	100%	100%	100%	100%	
	Average Score	83%	79%	77%	79%	79%	78%	
		B	C	C	C	C	C	

Fuente: Midior Consulting

4.1.2. Técnicas para la automatización

Como sucede en otros ámbitos, ya no es posible tratar de forma manual la calidad del dato. De hecho, la gran mayoría de las tareas dentro de esta disciplina se fundamentan en el análisis descriptivo, la minería de datos y la minería de textos.

1) El **análisis descriptivo** permite diferenciar los valores incompletos, vacíos, atípicos y rangos de valores, así como conocer la distribución de frecuencias, valores mínimos/máximos, etc. Este análisis se hará para cada columna del conjunto de datos analizados.

2) La **minería de textos** permite analizar atributos en formato de texto para identificar valores atípicos.

3) La **minería de datos** permite:

- Encontrar agrupaciones de datos y patrones. El análisis de patrones se utiliza para determinar si los valores de datos en un campo o campos coinciden con el formato o estructura esperados. Las técnicas de *clustering* se utilizan para detectar una variedad de errores relativos a una métrica de distancia elegida.

Clustering

El *clustering* es un procedimiento de agrupación de una serie de elementos de acuerdo con un criterio.

– La distancia euclidiana es útil para la detección de emisiones numéricas y de unidades de medición coherentes.

– Las distancias basadas en caracteres (distancia de *Levenshtein*), *token-based* (*atomic strings*) y en la fonética (*soundex*) son útiles para detectar inconsistencias en el texto, tales como errores ortográficos, diferentes ordenamientos de términos y palabras fonéticamente similares.

- Identificar anomalías. Mediante el análisis de los valores atípicos (ya sea a través del análisis estadístico, o técnicas más avanzadas como el análisis de series temporales).

- Completar y/o combinar datos. Los datos se pueden vincular implícitamente definiendo criterios de unión en valores similares, usando un valor único generado o códigos de coincidencia basados en algoritmos de lógica difusa *fuzzy logic*. También es posible usar otras técnicas de extrapolación. Por ejemplo, usando funciones heurísticas para extrapolar el salario actual de un cliente, a partir de su salario de hace cinco años.

Funciones heurísticas

Las funciones heurísticas tienen conocimiento de información de proximidad.

- Analizar e identificar causas de errores. La técnica de *root cause analysis* permite comparar escenarios con y sin errores para identificar los factores que los han motivado.

¿Cómo se usan estas técnicas?

El análisis de columnas proporciona una inspección de datos en la que un analista pasará a través de todos los registros, o un subconjunto de registros dentro de una tabla, e iniciará la identificación de varias estadísticas sobre los datos. Los tipos de análisis incluyen el número total de registros, su tipo de datos, cuántos registros contienen valores nulos o faltantes, cardinalidad o unicidad, valores mínimo y máximo, media, desviación estándar, duplicados y mucho más. Tener una idea de cómo se ven los datos ayudará a determinar cuánto trabajo se necesitará para solucionar o corregir las imprecisiones o inconsistencias en los datos.

Lectura complementaria

Elmagarmid, P. G.; Verykios, V. S. (2007). «Duplicate record detection: A survey». *IEEE Transactions on Knowledge and Data Engineering* (vol. 19, núm. 1, págs. 1-16).

Esto no termina aquí, puesto que el analista puede aplicar el análisis de dominios, y focalizarse en valores y rangos de datos esperados o aceptados. Es decir, decidir si un valor de datos específico es aceptable o cae dentro de un rango

aceptable de valores. Un ejemplo de estos criterios podría ser un campo de género donde los únicos valores aceptables son *Male* M o *Female* F. Otro ejemplo podrían ser las 50 abreviaturas de dos estados de caracteres en el formato apropiado (sin espacios, mayúsculas, sin periodos, etc.). Un informe de análisis de dominio produciría un gráfico que indicaría porcentajes de registros que caían dentro o fuera del valor aceptable.

Otro paso consistiría en identificar valores atípicos, es decir, valores fuera de rango. Los valores extremos pueden ser valores atípicos univariados estándar, o específicos del tipo. Por ejemplo, los valores atípicos de la serie temporal toman generalmente dos formas: un *outlier* aditivo es un movimiento inesperado y transitorio en un valor medido a lo largo del tiempo, mientras que un *outlier* inesperado es un movimiento inesperado que persiste en el tiempo.

Lectura complementaria

V. Chandola; A. Banerjee; V. Kumar (2009). «Anomaly detection: A survey». *ACM Comput. Surv* (núm. 41, vol. 3, art. 15).

Outlier

Es una observación numéricamente distante del resto de los datos, es decir, un valor atípico.

Y está claro que es necesario continuar con las potenciales dependencias con otros conjuntos de datos o sistemas.

Una vez hecho el análisis y conocido el estado, se usarían las técnicas de transformación para aplicar las medidas pertinentes, que van, como ya hemos visto, desde rechazar el error hasta transformar el dato.

Lectura complementaria

Y. Huhtala; J. Kärkkäinen; P. Porkka; H. Toivonen (1999). «Tane: An efficient algorithm for discovering functional and approximate dependencies». *The Computer Journal* (núm. 2, vol. 42).

4.2. Tecnología

Desde la perspectiva tecnológica, en la calidad de datos tenemos diferentes tipos de herramientas. El proceso de calidad de datos se ha llevado a cabo de forma manual, mediante *scripts* (SQL, librerías de R como dplyr*, librerías para Python como Pandas**, etc.) y mediante programas informáticos propietarios y *open source*.

*** Más información en: https://goo.gl/v3RbFJ**

**** Más información en: http://pandas.pydata.org**

Aquí nos centraremos en los programas. Dependiendo del fabricante, las herramientas incluyen más o menos características. Existen, principalmente, tres tipos:

1) Especializadas en la calidad de datos.

2) Que incluyen funcionalidad de calidad de datos, sin que sea su foco. Por ejemplo, las herramientas de integración de datos.

3) Que complementan las herramientas de calidad de datos. Por ejemplo, las herramientas de gestión del proyecto de calidad de datos, o para hacer pruebas.

4.2.1. Herramientas de calidad de datos

Este tipo de herramientas están especializadas en la calidad de datos. A veces, forman parte de una cartera más grande de productos (como una plataforma de integración de datos o una solución para el gobierno del dato), y otras son simplemente un producto independiente.

Esto significa que incluyen las siguientes características:

1) Análisis exploratorio de datos.

2) Perfilado de datos.

3) Capacidad de crear reglas de negocio.

4) Transformaciones de datos y *workflow* asociadas a calidad de datos.

5) Gestión de excepciones.

6) Gestión y acceso a tablas de referencia.

7) Identificador de duplicados e identidades.

8) Dominios de datos de negocio y capacidades de autodescubrimiento de dominios.

9) Gestión de metadatos.

Destacamos soluciones como las de Informatica*, Tamr**, DataCleaner***, Trillium Enterprise Data Quality**** o Datiris Profiler*****, aunque cabe comentar que los principales fabricantes del mercado disponen de solucione propias.

*** http://www.informatica.com**
**** http://www.tamr.com**
***** https://goo.gl/N9o3Gu**
****** https://goo.gl/KahJFS**
******* http://www.datiris.com**

La figura 8 ilustra la plataforma de calidad de datos. La plataforma se conecta con aquellos sistemas de información (internos y externos) que tienen datos relevantes. Esta comunicación, como ya sabemos, es bidireccional para poder propagar los cambios.

Dentro de esta categoría, podemos encontrar un nuevo tipo de herramientas híbridas que combinan la manipulación rápida de datos, la calidad de datos y la integración de datos, en una nueva categoría denominada *data wrangling*, cuyo principal cliente es el analista de datos y el científico del dato. Destacamos herramientas como Trifacta* u Open Refine**. Además, las propias herramientas de *data science* como Dataiku*** o Tibco Spotfire***** incluyen también estas capacidades. Cabe destacar, igualmente, que Pentaho Data Integration***** incluye múltiples *plugins* que extienden su funcionalidad en el ámbito de la calidad del dato.

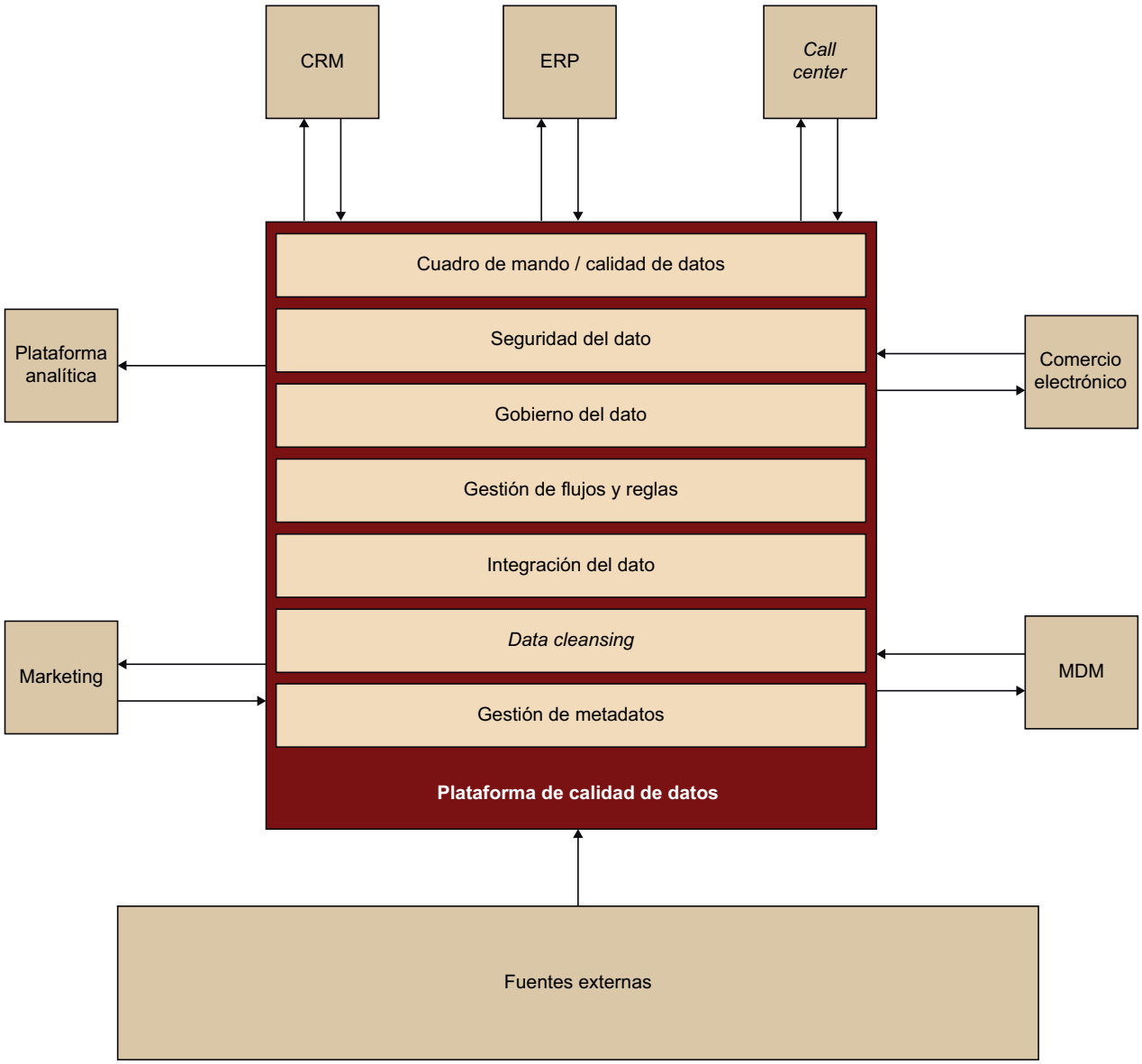
*** http://www.trifacta.com**
**** http://openrefine.org**
***** http://www.dataiku.com**
******* http://spotfire.tibco.com**
****** http://www.pentaho.com**

En el contexto de *big data*, han aparecido empresas especializadas como Zaloni*.

*** Más información en: http://www.zaloni.com**

Las soluciones actuales para la calidad de datos suelen incluir capacidades de monitorización y creación de *scorecards*.

Figura 8. Plataforma de calidad de datos



Fuente: David Cabanillas

4.2.2. Herramientas de integración de datos

Dentro de las herramientas de integración de datos, destacan las herramientas ETL, que se utilizan para:

- Extraer datos de fuentes de datos homogéneas o heterogéneas.
- Transformar los datos para almacenarlos en formato o estructura apropiados para el propósito de consulta y análisis.

- Cargar los datos en el destino final (base de datos, más específicamente, almacén de datos operativos, *data mart* y almacén de datos).

Estas herramientas han empezado a incluir pasos de transformación vinculados a la calidad de datos, como por ejemplo validar el formato de un campo (correo electrónico, cuenta bancaria, etc.) o incluso aplicar *data profiling*. Herramientas como Pentaho Data Integration* o Talend** caen en esta categoría.

* http://www.pentaho.com
** http://www.talend.com

4.2.3. Herramientas para la gestión del programa (de calidad de datos)

La aplicación del programa en un ámbito de organización se puede ejecutar como si de un proyecto se tratara. Es decir, es posible aplicar las herramientas de gestión de proyectos tradicionales para la implantación del plan de calidad de datos. Herramientas como Basecamp*, Trello** o Asana*** pueden ser de gran utilidad para coordinar el programa de calidad de datos.

* http://basecamp.com/
** http://trello.com
*** http://asana.com

4.2.4. Pruebas de calidad

Encontramos distintos tipos de pruebas que pueden aplicarse al *data warehouse* y bases de datos cuando se quiere garantizar los procesos de la organización en términos de calidad total. Algunas de las más interesantes son las siguientes:

1) Pruebas unitarias: consisten en validar cada uno de los componentes de una solución, aunque este tipo de test ha de llevarse a cabo durante la etapa de desarrollo, nunca después. Los elementos más críticos y que deben someterse a este tipo de prueba son, al menos, la lógica ETL, reglas de negocio y cálculos implementados en la capa de OLAP (*online analytical processing*) y la lógica de indicadores clave o KPI (*key performance indicator*). Este tipo de pruebas se hacen en varias ocasiones a lo largo del curso de un proyecto, y pueden automatizarse.

2) Pruebas del sistema de integración: dependen del éxito obtenido en las pruebas unitarias, y deben lograr dos metas principales:

a) Garantizar que se puede construir y desplegar con éxito: para lo que es necesario llevar a cabo pruebas de acumulación del sistema.

b) Asegurar que no surgen problemas durante la ejecución del trabajo: con este objetivo, una vez implementados y configurados, todos los trabajos deben ser ejecutados y los datos, procesados.

La adopción de este tipo de pruebas en el ciclo de desarrollo del *data warehouse* y bases de datos es un paso que sirve para confirmar que el sistema actúa del modo esperado, una vez que las partes constituyentes de la solución se conjuntan.

3) Pruebas de validación de datos: mediante este proceso, se someten a test los datos dentro de un *data warehouse*. Una forma habitual de llevar a cabo esta prueba consiste en el uso de una herramienta de consulta *ad hoc* (por ejemplo, Excel) que permita recuperar datos en un formato similar a los informes operativos existentes. Cuando se detecta la existencia de un vínculo entre el *data warehouse* y el informe operacional, se demuestra que los datos son válidos (a menos que, por supuesto, el informe original sea defectuoso). Esta prueba ha de ser llevada a cabo por un representante del negocio, ya que este perfil es el que mejor conoce los datos y puede validarlos con mayores garantías de éxito.

4) Pruebas de aceptación de usuario: su objetivo es asegurar que los datos que se proporcionan al usuario final cumplen con sus expectativas, y que lo mismo sucede con las herramientas que se ponen a su disposición.

5) Pruebas de rendimiento: se ocupan de validar adecuadamente el rendimiento de la solución en condiciones de trabajo reales. Para ello, en el *testing* hay que considerar factores como la arquitectura de datos, la configuración del *hardware*, la escalabilidad del sistema o la complejidad de las consultas.

6) Pruebas de regresión: este tipo de test es el proceso de volver a probar la funcionalidad para garantizar que el desarrollo del *data warehouse* y bases de datos no ha causado desperfectos en otras funciones y aplicaciones. Cada una de las distintas categorías de pruebas definidas anteriormente debe quedar sujeta a pruebas de regresión.

Resumen

En este módulo didáctico, hemos presentado el concepto de calidad del dato, que tiene el objetivo último de aumentar la confianza en el uso del dato para una toma de decisiones eficiente, óptima y rápida en la organización.

Primero, hemos discutido los motivos para la baja calidad de datos en las organizaciones, por qué no se invierte en ello y su necesidad, lo que nos ha llevado a definir el concepto.

A continuación, hemos revisado en qué consiste un programa de calidad de datos y sus componentes, desde la madurez (en qué estado se encuentra la organización respecto a la calidad del dato) hasta las diferentes metodologías existentes (desde el dato al negocio, o a la inversa). Todo esto con foco puesto en las personas, los procesos y los datos.

Para poder desarrollar el programa, hemos discutido las diferentes fases que lo componen, así como las mejores prácticas que debemos tener en cuenta y cómo medir el impacto en la organización.

Por último, se han revisado las técnicas y la tecnología que forman parte de lo que se conoce actualmente como calidad de datos.

Glosario

análisis y estandarización *m* y *f* Descomposición de los campos en partes y el formato de los valores para incorporar diseños basados en estándares industriales, estándares locales, reglas de negocio definidas por el usuario y bases de conocimiento de valores y patrones.

big data *m* Conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos.

business intelligence *m* Conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información, que permite tomar mejores decisiones a los usuarios de una organización.

ciclo de vida de un activo *m* Diferentes etapas por las que pasa un activo, desde su nacimiento hasta el fin.

data quality *f* Técnicas para la identificación, el control, el incremento y el mantenimiento de la calidad de datos en una organización.

data quality management **(DQM)** *m* La gestión de la calidad de los datos es un tipo de administración que incorpora el establecimiento de funciones, su despliegue, las políticas, las responsabilidades y los procesos con respecto a la adquisición, el mantenimiento, la disposición y la distribución de datos. Para que una iniciativa de gestión de la calidad de los datos tenga éxito, se requiere una sólida asociación entre los grupos tecnológicos y el negocio.

data warehouse *m* Repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo vayan a ser utilizados posteriormente por los consumidores o usuarios, con las propiedades siguientes: estable, coherente, fiable y con información histórica.

enriquecimiento *m* Mejorar el valor de los datos internos almacenados, añadiendo atributos relacionados de fuentes internas o externas.

ETL *m* Procesos que permiten la extracción, transformación y carga de datos desde fuentes de origen hasta el destino para su correcto consumo.

limpieza *f* Modificación de valores de datos para cumplir con las restricciones de dominio, restricciones de integridad u otras reglas de negocio que definen cuándo la calidad de los datos es suficiente para la organización.

matching *m* Identificar, vincular o combinar entradas relacionadas dentro o entre conjuntos de datos.

monitorización *f* Despliegue de controles que aseguran que los datos siguen cumpliendo con las reglas de negocio que definen la calidad de los datos para la organización.

perfilado *m* Captura de estadísticas (metadatos) que proporcionan información sobre la calidad de los datos y ayudan a identificar problemas de calidad.

programa de calidad de datos *m* Herramienta para la evaluación de la calidad de los datos dentro de una organización, y que permite identificar atributos de calidad de datos, analizar los atributos de calidad de datos en su contexto actual o futuro y proporcionar una guía para mejorar la calidad de los datos.

Bibliografía

Brackett, M.; Earley, P. S. (2009). *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*. Nueva York: DAMA.

Curto, J. (2017). *Introducción al Business Intelligence (nueva edición ampliada y revisada)*. Barcelona: Editorial UOC.

Hoberman, S. (2015). *Data Model Scorecard: Applying the Industry Standard on Data Model Quality*. Nueva York: Technics Publications.

Jugulum, R. (2014). *Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality*. Nueva York: John Wiley & Sons.

Loshin, R. (2010). *The Practitioner's Guide to Data Quality Improvement*. Nueva York: Morgan Kaufmann.

Mosley, M (2009). *DAMA-DMBOK functional framework*. Nueva York: DAMA.

McGilvray, D. (2008). *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM)*. Nueva York: Elsevier Science.

Sebastian-Coleman, L. (2013). *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework* . Nueva York: Morgan Kaufmann.