Los datos en la factoría de información corporativa

PID_00270640

Juan Vidal Gil
Carles Llorach Rius

Tiempo mínimo de dedicación recomendado: 3 horas

Universitat Oberta de Catalunya

Juan Vidal Gil

Licenciado en Físicas por la Universidad Complutense de Madrid. Experiencia en soluciones tecnológicas de *Business Intelligence* y *Data Warehouse*, como jefe de proyectos en importantes compañías y como formador especializado en empresas del sector. Profesor colaborador de la UOC.

Carles Llorach Rius

Máster en Gestión de Empresas - MBA por la Universidad Rovira i Virgili e ingeniero en Informática por la Universidad Politécnica de Cataluña. Profesor colaborador en la Universitat Oberta de Catalunya.

La revisión de este recurso de aprendizaje UOC ha sido coordinada por la profesora: Àngels Rius Gavidia (2019)

Segunda edición: febrero 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

En otros módulos de esta asignatura hemos visto los almacenes de datos desde una perspectiva general de la organización. Dada la importancia que tienen los datos por sí mismos, en este módulo vamos a centrarnos en estudiarlos desde el punto de vista de la gestión/gobernanza de los datos.

Con ello, nos estamos refiriendo a la integración de los datos, a la gestión de los datos maestros, a la calidad de los datos, a la gestión de los metadatos y todos aquellos aspectos legales y éticos que debemos contemplar en el tratamiento de los datos.

Objetivos

En este módulo se pretende ofrecer una visión global de los procesos de gestión del dato en la organización, haciendo hincapié en aquellos estrechamente relacionados con los almacenes de datos.

Mediante el estudio, se conseguirán los objetivos siguientes:

1. Tomar conciencia del valor e importancia de los datos como activo de la organización.

2. Entender la función y el ámbito de las actividades de gobierno del dato.

3. Conocer la importancia y la gestión de los procesos de calidad del dato.

4. Profundizar en el concepto de metadato conociendo sus diferentes funciones y usos.

5. Conocer los aspectos legales y éticos para el adecuado tratamiento de los datos.

1. La importancia de los datos

1.1. Datos

Las organizaciones desarrollan sistemas informáticos donde residen sus datos: en el caso de la base de datos operacional, lo importante son los datos actuales, mientras que, en el caso del almacén de datos, la importancia está en los datos históricos.

En los dos entornos, el dato es muy importante. Aquellas organizaciones que así lo entienden y que actúan de acuerdo con esto suelen recibir el nombre de organizaciones orientadas al dato (en inglés, *data-driven*).

En este mismo contexto, surge el concepto de gobernanza de datos (en inglés, *data governance*, DG), que es una disciplina de control de calidad para la evaluación, la gestión, el uso, la mejora, la supervisión, el mantenimiento y la protección de información/datos de la organización.

Alguna de las actividades que típicamente se encuadran en la gobernanza de datos es la gestión de datos maestros (en inglés, *master data management*, MDM) y la limpieza de datos (en inglés, *data cleaning* o *data scrubbing*), entre otras.

La integridad de los datos es un problema importante en la mayoría de las organizaciones, y el desarrollo de un almacén de datos se utiliza con frecuencia como un vehículo para mejorar la calidad de los datos de manera significativa. La exactitud de los datos puede significar ahorros considerables en áreas como marketing, atención al cliente, etc. Existen estudios llevados a cabo por organizaciones tales como Gartner Group (una de las principales empresas de prospección de mercado) que estiman en un 4 % los ahorros obtenidos a partir la mejora de la integridad de los datos en las organizaciones.

Aparece así el concepto de *data warehouse governance*, que recoge aquellas prácticas centradas en cómo se crean los datos, cómo son recogidos, tratados y manipulados, almacenados, puestos a disposición para su uso o retirados.

Denominaremos programa al conjunto de prácticas que, pudiendo variar significativamente, dependiendo de su enfoque: en el cumplimiento (*compliance*), en la integración de datos, en la gestión de datos maestros (MDM), etc., están alineadas con las políticas corporativas: en un ámbito de lógica de negocio, estrategia tecnológica, seguridad, etc.

Las actividades de las organizaciones son generalmente horizontales y afectan a varios departamentos o funciones (comercial, tráfico, administración, etc.). La organización horizontal recibe el nombre también de «por actividades o procesos» y es totalmente contraria a la organización tradicional vertical, por departamentos o funciones. La organización «vertical» se visualiza como una agregación de departamentos independientes unos de otros y que funcionan autónomamente. Un buen despliegue de nuestros programas requiere una concepción amplia (horizontal) de nuestra organización.

Las organizaciones necesitan pasar del gobierno informal al gobierno de datos formal, cuando se da alguna de las siguientes situaciones:

- La organización llega a ser tan grande que la gestión tradicional no es capaz de entregar los datos relativos a actividades multifuncionales/transversales.

- Los sistemas de datos de la organización se hacen tan complicados que la gestión tradicional no es capaz de entregar los datos relativos a actividades multifuncionales/transversales.

- Los arquitectos de datos de la organización, los equipos de SOA (*service-oriented architecture*) u otros grupos enfocados horizontalmente, necesitan una visión corporativa (en lugar de fragmentada en silos) de las preocupaciones y las opciones relativas a los datos.

- La regulación: el cumplimiento legal o la existencia de requisitos contractuales que lo exigen.

Un *data warehouse* interactúa, por definición, con gran parte de la organización. Las políticas, procesos y procedimientos del programa deben ser claramente comunicados a todos los afectados para asegurar que el esfuerzo requerido generará beneficio.

La información proviene de fuentes internas (sistemas de producción) y externas (hasta un 20 %) y supone problemas como la saturación de información, la dificultad de acceso, no ser selectiva, etc. Todo esto deberá ser contemplado a la hora de diseñar nuestros programas.

2. Integración de datos

En el módulo «Construcción de la FIC» vimos la importancia de la integración de datos en los procesos de actualización de los almacenes de datos, concretamente en el componente de integración y transformación de datos. La correcta integración de datos en un almacén es una cuestión crítica para su correcta explotación. En este apartado vamos a ver las diferentes disciplinas asociadas a la integración de datos.

2.1. Disciplinas que intervienen en la integración de datos

Sabemos que el almacén de datos tiene un papel fundamental en lo relativo a consolidación e integración de la información: permite pasar de lo que llamamos telaraña de entorno operacional a un entorno centralizado e integrado. Sabemos también que la integración de los datos supone un auténtico reto si tenemos en cuenta la disparidad de orígenes, formatos, herramientas y sistemas que habitualmente tienen en las compañías.

La **integración de los datos** no es un problema exclusivo de los almacenes de datos, sino que se aplica a todos los sistemas que gestionan información y es una actividad que tiene todo un conjunto de disciplinas asociadas.

Algunas de estas disciplinas pueden ser: la calidad de los datos, la gestión de datos maestros, la definición de métricas homogéneas, el ciclo de vida del dato, etc. Estas disciplinas son solo parte de las disciplinas que intervienen en los procesos de gestión de datos en las compañías y pueden englobarse dentro de otra disciplina denominada gobierno del dato.

2.2. Gobierno del dato

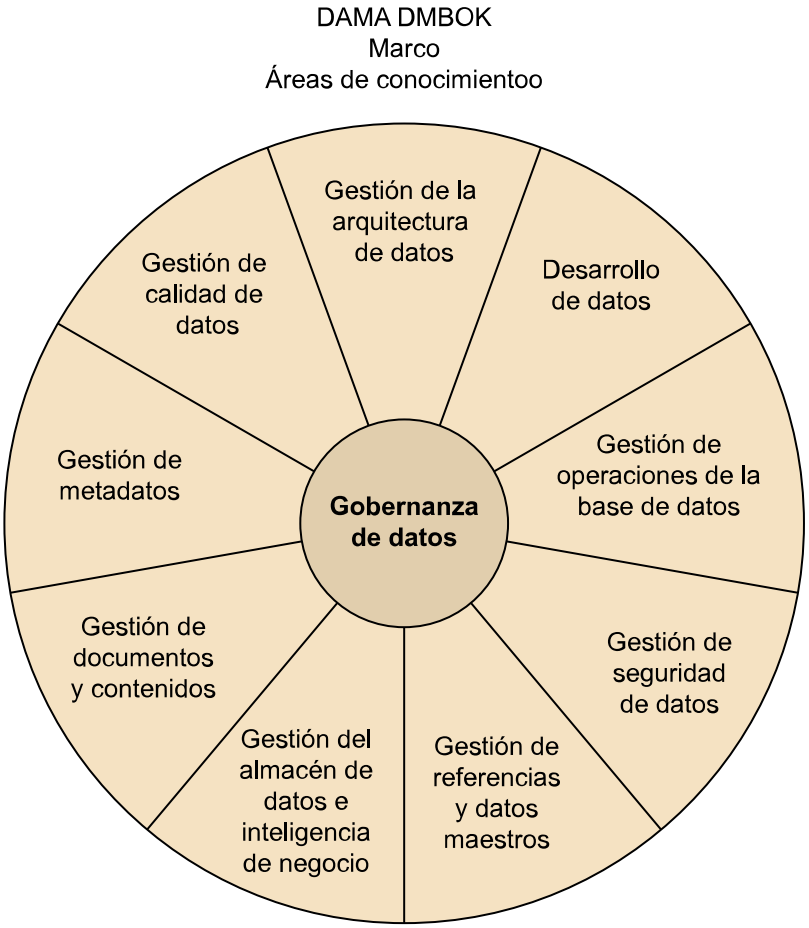
El **gobierno del dato** es una importante disciplina empresarial cuyo objetivo es proporcionar mayor control sobre la creación, manejo, mantenimiento, almacenamiento, uso e intercambio de información vital para el negocio.

Según el diccionario de la DAMA (Data Management Agency) el gobierno de los datos o la gobernanza de datos son los ejercicios de control y autoridad (planificación, monitorización y mejora) sobre la gestión de los datos.

La gobernanza de datos se compone del conjunto de áreas siguientes:

- Arquitectura de datos: análisis y diseño.
- Gestión de bases de datos.
- Gestión de seguridad del dato.
- Gestión de calidad del dato.
- Gestión de datos maestros.
- Gestión de sistemas de inteligencia de negocio y almacenamiento del dato.
- Gestión de documentos y contenidos.
- Gestión de metadatos.

Figura 1. Áreas que considera la gobernanza de datos



Fuente: www.dama.org.

En este módulo vamos a estudiar aquellas disciplinas del gobierno del dato más directamente relacionadas con los almacenes de datos, como pueden ser:

- Calidad del dato.
- Gestión de metadatos.

3. Calidad del dato

Tal y como ya se ha explicado en otros módulos, la calidad del dato es una cuestión crucial para los almacenes de datos y, en general, para la organización. La falta de calidad de los datos es uno de los principales problemas a los que se enfrentan los responsables de sistemas de información y las empresas, pues representa claramente uno de los problemas «ocultos» más graves y persistentes en cualquier organización.

La **gestión de datos** constituye un recurso estratégico en la organización y su calidad, un punto crucial en esta gestión.

Una correcta gestión de la calidad de los datos nos va a aportar los siguientes beneficios:

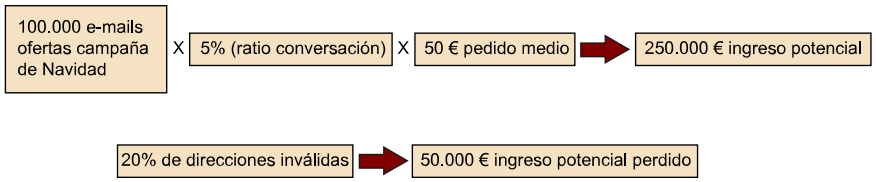
- Visión única del cliente-usuario/producto-servicio/proveedores.
- Mejora de la comunicación cliente-usuario/proveedores.
- Ahorro de tiempos de conciliación de la información.
- Garantía de la correcta unificación de bases de datos (fusiones de empresas).
- Confianza en el dato, mejora de procesos de *reporting* y analítica.

Ejemplo de procesos empresariales que se benefician de una correcta gestión de la calidad del dato

- Campañas de marketing.
- Análisis geomarketing.
- Cumplimiento de las normativas. Protección de datos (LOPD).
- Ahorro de costes de errores de facturación, envíos, comunicación con clientes.
- Procesos de detección del fraude.

En la figura 2 podemos ver el coste que puede tener una baja calidad del dato en una campaña de e-mail marketing.

Figura 2. Ejemplo del impacto económico de una baja calidad del dato



3.1. Objetivos de la calidad del dato

Con el fin de garantizar la calidad de los datos, los procesos de calidad de los mismos buscan los siguientes objetivos:

- **Precisión de los datos**: que cada dato sea fiel representante de lo que la función que se le atribuye requiere, haciéndolo de la forma establecida.

- **Confiabilidad de los datos**: que el dato que representa la información sea coherente y estable.

- **Completitud de los datos**: que garantice que ni en los propios datos, ni en los registros o tablas donde se almacenan falten campos o valores, que todo esté completo.

- **Conformidad de los datos**: que se respeten las condiciones de formato establecidas al dar de alta el dato.

- **Consistencia de los datos**: que, además de garantizar que el dato es correcto en cuanto a sus atributos, no vulnere ninguna regla de negocio.

- **Unicidad de los datos**: que no existan duplicidades.

El cumplimiento de estos objetivos garantiza una adecuada calidad en el dato, y para revisar su cumplimiento debemos establecer todo un conjunto de procesos y comprobaciones.

3.2. Etapas principales en la gestión de la calidad del dato

Conseguir los objetivos de calidad indicados en la sección anterior supone la realización de una serie de etapas bien definidas:

- **Perfilado del dato**: procesos encaminados a explorar las fuentes origen, obteniendo información estadística acerca de la fuente (rangos de valores, distribución, nulos, valores únicos, patrones).

- **Validación del dato**: batería de comprobaciones encaminadas a asegurar la corrección, consistencia, conformidad y completitud de los datos.

- **Limpieza del dato**: detección y corrección de errores en los datos (falta de completitud, inconsistencias, incorrecciones, etc.) bien sea modificando, o bien eliminando los registros afectados.

- **Enriquecimiento del dato**: procesos encaminados a mejorar, depurar, normalizar o completar la información de una fuente, utilizando otras fuentes complementarias.

3.2.1. Perfilado del dato

Los procesos de perfilado del dato nos permiten realizar una exploración previa para obtener información estadística de los datos que también nos pueden ayudar a conocer la calidad de la información origen. Hay procesos de perfilado de estructura y de contenido.

Ejemplos de operaciones en un perfilado de datos

Algunos ejemplos de operaciones de perfilado de datos que se pueden realizar sobre una entidad de datos son las siguientes:

a) En la tabla:

- Calcular el volumen de registros.
- Verificar el cumplimiento de reglas de negocio.
- Verificar la integridad referencial (padre-hijo). Detectar valores huérfanos.
- Detectar dependencias entre columnas (correlaciones).

b) En la columna:

- Obtener los valores únicos columna, duplicados, frecuencias.
- Calcular la columna: media, máximo, mínimo, desviación típica, varianza.
- Comprobar el tipo de dato, longitud, distribución de longitudes.
- Revisar el número de nulos, blancos.
- Comprobar la distribución de patrones (dd/mm/yuya, XX-XXX).
- Ajustar a patrones predefinidos (direcciones, código postal, e-mail, teléfono, etc.).
- Ajustar a patrones definidos por el usuario (expresiones regulares).

3.2.2. Validación del dato

Los procesos de validación del dato realizan las comprobaciones necesarias para asegurar la corrección, consistencia, conformidad y completitud de los datos. Existen dos tipos de validaciones:

a) Validaciones técnicas: todas las que nos garantizan la consistencia técnica de lo datos, evitando duplicidades, campos nulos, *outliers*, falta de integridad referencial, etc.

b) Validaciones de negocio: todas las que nos garantizan la consistencia de los datos en base a reglas de negocio.

Ejemplos de validaciones técnicas

- Duplicados: detección de duplicados por repetición clave exacta y mediante algoritmos de comparación de cadenas.
- Campos obligatorios: validar que estén informados todos los campos obligatorios.
- Tipo de datos: tipo de dato esperado (numérico, alfanumérico).

- Consistencia claves foráneas: validar integridad referencial entre las claves de tablas referenciadas (padre-hijo).
- Conciliación entre fuentes: conciliación de registros entre fuente origen y fuente destino (agregados, conteo de registros).

Ejemplos de validaciones de negocio

- Rango de valores: para determinadas variables podemos tener un rango de valores posible, ejemplo: la edad no puede ser negativa.
- Patrón del campo: el campo debe tener un patrón establecido (ejemplo: e-mail xxxx@yyyy.zzz).
- Codificación interna: variables que cumplen en su codificación interna (ejemplo: DNI).
- Cumplimiento de reglas de negocio: validaciones de negocio particulares de cada fuente y negocio concreto.

Una problemática muy habitual en los procesos de validación del dato es la relacionada con la deduplicación del dato.

La **deduplicación** es un proceso que persigue la identificación de duplicados por diferentes criterios. Los procesos de deduplicación son imprescindibles no solamente para eliminar registros y datos redundantes, sino también para proyectos de consolidación de fuentes de información y enriquecimiento de datos.

Existen diferentes técnicas para identificar registros duplicados. El caso más sencillo es cuando nuestros registros coinciden por clave; sin embargo, hay casos de duplicados en los que la clave no coincide exactamente, incluso aunque se trate del mismo registro. A menudo suelen ser casos en que hay algún campo de la clave con nulos, campos de tipo cadena de caracteres en los que hay errores tipográficos, campos con el mismo valor y diferente formato (por ejemplo, campos de tipo fecha). Para estos casos en los que no buscamos una clave exacta tenemos que realizar cruces y podemos hacerlo con distintos tipos de cruce o *matching*:

a) *Matching* determinístico: se comparan los diferentes atributos asociados a la entidad y se obtiene un resultado positivo o negativo. Antes de la comparación se suelen realizar transformaciones, normalizaciones, codificaciones y limpiezas previas a la comparación. En la figura 3 se puede ver un ejemplo de este caso.

b) *Matching* probabilístico: mediante algoritmos específicos se comparan diferentes atributos. Dichos algoritmos devuelven un porcentaje que indica el grado de similitud entre los atributos comparados. Los algoritmos de comparación deberán ser adecuados para el tipo de datos, puesto que no es lo mismo comparar una cadena de texto libre (como un nombre, razón social, descripción de producto, etc.) que un código (teléfono, CIF, código postal, núme-

ro de ref., etc.). Al igual que con el *matching* determinístico, es conveniente realizar transformaciones, codificaciones y limpiezas previas. Finalmente, se toman todos los porcentajes obtenidos de las diferentes comparaciones y se realiza una media ponderada. Ciertos atributos pueden tener mayor peso que otros, por ejemplo, al comparar empresas tendrá más peso la razón social que el teléfono. Un ejemplo sería el que se presenta en el caso b) de la figura 6.

Figura 3. Ejemplos de *matching* determinístico (a) y probabilístico (b)

a)

Razón social	CIF	Dirección	Cód. Postal	Teléfono	Municipio	Población
Laboratorios Roma S.A.	A11112222	Polígono industrial Los Sauces	30140	9611112233	Santomera	Murcia
✗	✗	✗	✓	✓	✗	✓
Razón social	**CIF**	**Dirección**	**Cód. Postal**	**Teléfono**	**Municipio**	**Población**
Roma S.A.		Polígono Los Sauces, C/Enebro, 10	30140	9611112233	Santmera	Murcia

b)

Razón social	CIF	Dirección	Cód. Postal	Teléfono	Municipio	Población
Laboratorios Roma S.A.	A11112222	Polígono industrial Los Sauces	30140	9611112233	Santomera	Murcia
40%		70%	100%	100%	90%	100%
Razón social	**CIF**	**Dirección**	**Cód. Postal**	**Teléfono**	**Municipio**	**Población**
Roma S.A.		Polígono Los Sauces, C/Enebro, 10	30140	9611112233	Santmera	Murcia

3.2.3. Limpieza del dato

Los procesos de limpieza del dato corrigen los errores detectados en los datos (falta de completitud, inconsistencias, incorrecciones, etc.) modificando o eliminando los registros afectados. Existen diferentes alternativas para corregir el dato:

- **Eliminación de registros**: puede venir obligada por posibles errores (registros duplicados).
- **Valor indeterminado**: en casos de valor no informado de una columna, una alternativa puede ser definir un código indeterminado para dicha columna.
- **Estimación del valor**: en casos de valor no informado, una alternativa puede ser estimar el valor (extrapolaciones, media de valores de columna).
- **Corrección de palabras**: corrección de errores ortográficos o tipográficos, eliminación de blancos innecesarios.

- **Normalización y estandarización de datos**: corrección de campos que deben seguir valores estándares (calles, municipios, ciudades, nombres de personas).

3.2.4. Enriquecimiento del dato

Los procesos de enriquecimiento del dato permiten mejorar los datos existentes complementando la información de las fuentes con otras fuentes, ya sean internas o externas. Es muy común el empleo de bases de datos estándares para completar información geográfica (ciudades, municipios, códigos postales, calles, coordenadas, etc.) o la información sociodemográfica de los clientes (edad, estado civil, número de hijos, etc.), bases de datos de empresas. En otros casos los campos informados serán modificados por su valor normalizado (calles, municipios, etc.).

3.3. Implementación de los procesos de gestión de la calidad del dato

Las etapas definidas en el apartado anterior dan como resultado la identificación de una serie de reglas de validación, corrección, limpieza y enriquecimiento del dato. Estas reglas deben ser implementadas en diferentes procesos de gestión del dato con objeto de garantizar la calidad del dato en la organización.

Algunos de los procesos más críticos son los siguientes:

- **Componente de transformación e integración de la CIF**: debe garantizar que la información que se consolida en los almacenes de datos esté depurada.

- ***Hub* MDM**: se implementarán estas reglas para garantizar la calidad en la integración de datos sobre los datos maestros.

- **Sistemas operacionales que traten información crítica de la compañía**: se consigue así que la información en origen sea lo más fiable posible.

- **Procesos de monitorización de la calidad del dato**: sobre algunas entidades se diseñarán y aplicarán procesos para medir la calidad de sus datos.

Respecto al último proceso mencionado, el de monitorización, se definirán una serie de métricas para medir la calidad de los datos de entidades determinadas. De los resultados obtenidos en las monitorizaciones podremos obtener una serie de indicadores o ratios sobre la calidad del dato Igualmente podemos crear alertas sobre estos ratios. El objetivo final será tener un cuadro de mando sobre calidad del dato.

3.4. Tendencias en los procesos de gestión de la calidad del dato

En los últimos años en los procesos de gestión de datos en entornos *Big Data* se ha desarrollado el concepto de *Data Curation*, el objetivo de esta técnica es automatizar todos los procesos de limpieza, estandarización y enriquecimiento del dato basándose en algoritmos de *machine learning* y sistemas expertos.

Estos procesos aplican diferentes técnicas analíticas para mejorar la calidad de los datos, relacionar diferentes fuentes y enriquecer los datos partiendo de un conjunto numeroso y heterogéneo de fuentes orígenes de datos, que pueden contener un volumen muy elevado de registros e información estructurada y no estructurada.

Dentro de las múltiples técnicas que se pueden aplicar, mencionamos algunas:

- Identificación de registros de diferentes fuentes que hacen referencia a la misma entidad de datos, lo que permite establecer relaciones (*clustering*, distancias, etc.).

- Identificación de registros duplicados (*clustering*, *matching*, etc.).

- Empleo de patrones para identificar determinadas entidades (nombres, teléfonos, direcciones, etc.).

- Revisión de columnas: obtención de relevancia de términos en colecciones o documentos, comparación de distribuciones para columnas numéricas.

- Minería de textos para identificar patrones, relaciones y enriquecer la información.

- Obtención de probabilidades en la identificación de entidades.

Los resultados obtenidos se relacionan con un nivel de confianza y son visualizados y presentados de forma que favorecen la intervención manual.

4. Gestión de metadatos

En otros módulos se ha hablado sobre la importancia de los metadatos en el contexto de la FIC. Se trataron diferentes tipos de metadatos de la FIC (los metadatos de fuentes de datos, los del almacén de datos y los del componente de integración y transformación), que nos ayudan a gestionar la FIC, hacerla evolucionar y, en algunos casos, son consultables por diferentes tipos de usuarios con el fin de conocer mejor la estructura de los modelos, cómo se relacionan las entidades o cómo se transforman los datos.

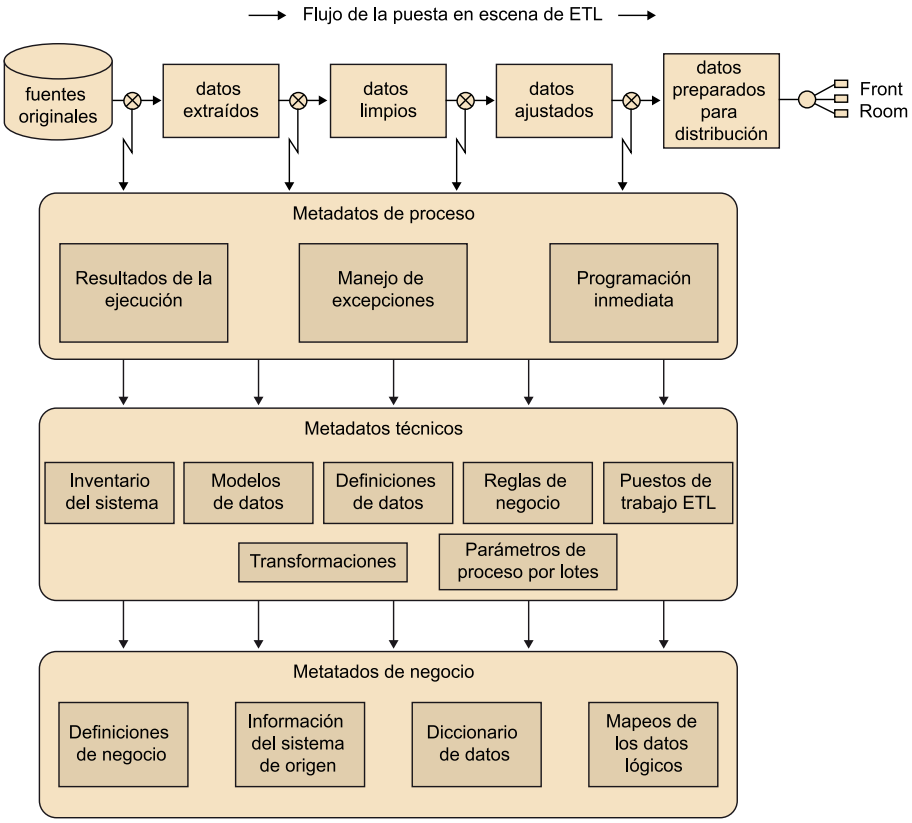
Los metadatos no son ámbito exclusivo de la FIC y son un elemento crítico en las actividades de gobierno del dato.

Los metadatos nos dan información sobre cómo almacenar los datos, cómo obtenerlos, cómo explotarlos y cómo se relacionan las entidades y los procesos.

4.1. Tipos de metadatos

Existen diferentes tipos de metadatos según su uso y diferentes posibles clasificaciones, una de las más extendidas es la que propone Ralph Kimball y que se muestra en la figura 4.

Figura 4. Tipos de metadatos



Fuente: www.kimballgroup.com.

Esta clasificación propone los siguientes tipos de metadatos:

- **Metadato de negocio**: describe los datos desde un punto de vista de negocio, y muestra un diccionario de datos que traduce el modelo de datos a términos de negocio.

- **Metadato técnico**: describe los aspectos técnicos tales como tipos de datos, longitudes, relaciones entre tablas, pasos de transformación, composición y dependencias de los *jobs*, etc.

- **Metadato de procesos**: muestra datos sobre las ejecuciones (registros leídos, escritos, rechazados, tiempos de proceso, errores, *logs* de procesos, etc.).

4.2. Retos en la gestión de los metadatos

El crecimiento exponencial de la información de los últimos años obliga a una mejora en los procesos de gobierno de datos que pasa por una gestión más óptima de los metadatos. Mostramos a continuación los principales retos que se deben afrontar en la gestión de los mismos:

4.2.1. Gestión integral de los metadatos

El concepto calidad de metadatos surge en grandes corporaciones que cuentan con miles de atributos e indicadores. Se trata de una problemática de integración y/o de herramientas de gestión de metadatos, no de calidad de datos en sí.

Existen soluciones que permiten integrar los metadatos generados en diferentes componentes con objeto de tener una visión común. Debemos integrar la gestión de metadatos que realizamos en el componente de transformación e integración de la FIC con la gestión de los metadatos que se lleva a cabo en el *hub* MDM y los orientados a la explotación del dato.

Esta gestión integral debe marcar un lenguaje de negocio común para unificar las definiciones y los criterios que hay que aplicar de los indicadores, atributos y cálculos comunes. Por ello son necesarios estándares de metadatos.

4.2.2. Estándares de metadatos

Para compartir los metadatos entre componentes, estos deben «hablar» el mismo idioma en este aspecto. Un estándar de definición de metadatos representa este idioma común.

En lo relativo a estándares tenemos el *common warehouse metadata* (CWM), que nos ayuda a definir y compartir metadatos entre componentes de nuestra arquitectura y soluciones de software.

Es necesario potenciar el uso de este tipo de estándares para mejorar la gestión de los metadatos, definirlos de forma eficiente, estándar y sencilla de compartir.

4.2.3. Información semiestructurada o no estructurada

Es un hecho que la información semiestructurada o no estructurada supone una fuente de información cada vez más relevante en las compañías. Aunque por su naturaleza no se trate de una información almacenable en estructuras bien definidas, sí es necesaria una capa de metadatos que será más ligera que en el caso de información estructurada, pero que nos ayudará a almacenarla y gestionarla.

5. Aspectos legales y éticos de los datos

5.1. Normativa legal de los datos

El primer aspecto que debemos considerar cuando trabajamos con datos es si están sujetos a alguna normativa legal vigente donde se lleva a cabo la actividad organizativa o contractual que hayamos podido suscribir con nuestros proveedores, clientes o socios.

Nos encontramos con dos principios generales del derecho que nos son de aplicación: la protección de datos, que está configurada en un ámbito europeo como un derecho fundamental de los ciudadanos; y la transparencia, regulando país a país el acceso a la información en poder de las administraciones públicas, premisa indispensable para la rendición de cuentas.

5.1.1. Protección de datos

La Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos y Garantía de los Derechos Digitales, despliega la principal norma de la Unión Europea en la materia, que es el Reglamento 2016/679, General de Protección de Datos (RGPD). Pero, con anterioridad, ya en 1981, el Consejo de Europa había adoptado el Convenio n.º 108, sobre la protección de las personas, que es el único instrumento internacional vinculante sobre protección de datos.

Nota

El Reglamento General de Protección de Datos entró en vigor el 25 de mayo de 2016.

Relacionamos, a continuación, y de manera muy resumida, los principales aspectos que hay que considerar:

- El interesado tiene el **derecho a ser informado** cuando sus datos personales deben ser recogidos.

- El responsable del tratamiento tiene que proporcionar su nombre y dirección, la finalidad del tratamiento, los destinatarios de los datos y toda otra información que sea necesaria para garantizar un tratamiento leal con el interesado (art. 10 y 11).

- **Tratamiento**: los datos pueden ser tratados solo bajo las siguientes circunstancias (art. 7):
  - Cuando el interesado ha dado su consentimiento.

  - Cuando es necesario para la ejecución de un contrato o por medidas precontractuales.

  - Cuando es necesario para el cumplimiento de una obligación jurídica.
  - Cuando es necesario para proteger los intereses vitales del interesado.
  - Cuando es necesario para el cumplimiento de una misión de interés público o inherente al ejercicio del poder público conferido al responsable del tratamiento.
  - Cuando es necesario para el propósito del interés legítimo perseguido por el responsable del tratamiento, siempre que no prevalezcan el interés o los derechos y libertades fundamentales del interesado.

- El interesado tiene el **derecho de acceso** a todos sus datos tratados. Incluso tiene el derecho de pedir la **rectificación**,**supresión** o **bloqueo** de datos que sean incompletos, inexactos o que no sean tratados de acuerdo con las disposiciones de la Directiva de protección de datos (art. 12).

- **Legitimidad**: los datos personales solo pueden ser recogidos para finalidades determinadas, explícitas y legítimas, y no pueden ser tratados posteriormente de manera incompatible con dichos fines (art. 6b).

- **Proporcionalidad**: los datos personales tratados solo pueden ser los adecuados, pertinentes y no excesivos en relación con la finalidad para la que fueron recogidos. Los datos deben ser exactos y, cuando sea necesario, actualizados; se deberán tomar las medidas razonables para suprimir o rectificar los datos inexactos o incompletos, con respecto a la finalidad con la que fueron recopilados. Los datos deben ser conservados de una manera que permita la identificación de los interesados durante un periodo no superior al necesario para la finalidad para la que fueron recopilados (art. 6).

- **Tratamientos especiales**: se aplican cuando los datos personales son sensibles: origen racial o étnico, opiniones políticas, convicciones religiosas o filosóficas, afiliación a sindicatos, salud o sexualidad (art. 8).

- El interesado puede **oponerse** en cualquier momento al tratamiento de datos personales, con la finalidad de prospección de mercado (art. 14).

5.1.2. Transparencia

Podríamos considerar bajo este campo todo aquello destinado a fortalecer la confianza de la sociedad en las instituciones públicas y organizaciones, a través del impulso del buen gobierno, la transparencia y la rendición de cuentas de sus actividades.

A diferencia de la protección de datos, no existe una norma europea o internacional que regule la transparencia de forma universal, y son los países los que legislan para garantizar el derecho a la información pública.

En la mayoría de los casos, el despliegue de estas leyes está soportado por un **portal de transparencia** que proveerá la navegación por el portal y la presentación de datos. Se deberá establecer la arquitectura tecnológica que mejor se ajuste a estos propósitos.

Con esta misma voluntad, hace años que han surgido iniciativas que fomentan la transparencia/apertura de las organizaciones a través de la publicación de sus datos. Los **datos abiertos** (*open data*) son aquellos que se consideran accesibles y reutilizables, sin exigencia de permisos específicos.

De este modo, estamos creando las condiciones para el desarrollo del mercado de la **reutilización de la información**, así como la **interoperabilidad** entre distintas organizaciones.

5.1.3. Otros principios

Además de la protección de datos y la transparencia, existen otras leyes que rigen el tratamiento de datos y debemos actuar de acuerdo con las mismas: la Ley de Servicios de la Sociedad de la Información (LSSI), la Ley Orgánica de Regulación del Tratamiento Automatizado de Datos (LORTAD) y la ley de *cookies* (BOE).

También existe un marco de actuación definido por un conjunto de principios relacionados con el uso de datos. Los más relevantes y que deberíamos considerar con atención son: privacidad, seguridad, identidad, confidencialidad, etc.

5.2. Ética de los datos

La ética en la informática es una disciplina nueva que pretende abrirse campo dentro de las éticas aplicadas, por lo que encontramos varias definiciones:

- Mario González Arencibia la define como «la disciplina que analiza los problemas éticos que son creados por la tecnología de los ordenadores o también los que son transformados o agravados por la misma». Es decir, por las personas que utilizan los avances de las tecnologías de la información.

- María Bolaño nos dice: «Es el análisis de la naturaleza y el impacto social de la tecnología informática y la correspondiente formulación y justificación de políticas para un uso ético de dicha tecnología». Esta definición

está relacionada con los problemas conceptuales y los vacíos en las regulaciones que ha ocasionado la tecnología de la información.

- Otros autores (J. B. Peña y E. A. Fernández) formulan la ética informática como «la disciplina que identifica y analiza los impactos de las tecnologías de la información en los valores humanos y sociales». Estos valores afectados son la salud, la riqueza, el trabajo, la libertad, la democracia, el conocimiento, la privacidad, la seguridad o la autorrealización personal.

La ética informática se plantea varios objetivos:

- Descubrir y articular dilemas éticos clave en informática.
- Determinar en qué medida son agravados, transformados o creados por la tecnología informática.
- Analizar y proponer un marco conceptual adecuado y formular principios de actuación para determinar qué hacer en las nuevas actividades ocasionadas por la informática en las que no se perciben con claridad líneas de actuación.
- Utilizar la teoría ética para clarificar los dilemas éticos y detectar errores en el razonamiento ético.
- Proponer un marco conceptual adecuado para entender los dilemas éticos que origina la informática y, además, establecer una guía cuando no existe reglamentación de dar uso a Internet.

La ética informática (y, por extensión, la de sus datos) debe estar, por lo menos, presente en las siguientes áreas:

- La utilización de la información.
- Lo informático como nueva forma de bien o propiedad.
- Lo informático como instrumento de actos potencialmente dañinos.
- Miedos y amenazas de la informática.
- Dimensiones sociales de la informática.

Así, el profesional que trabaja con datos sensibles (por ejemplo: sobre personas o grupos) destinados a **tomar decisiones** debe adoptar una forma de proceder que garantice:

- Responsabilidad.

- Confidencialidad.
- Calidad del producto.
- Juicio.
- Promover un enfoque ético en la gestión.
- Promover el conocimiento.
- Actualización permanente.

No es necesario establecer regulaciones sobre lo que se debe hacer con los datos. El objetivo debe ser ayudar a **tomar decisiones** efectivas en el ámbito de los negocios, a través de métodos y técnicas que faciliten discusiones internas, rigurosas y productivas. Estas discusiones pueden expresar posiciones coherentes y consistentes de la perspectiva de una organización sobre el uso de sus datos.

Si nos trasladamos al ámbito de las redes sociales y los grandes volúmenes de datos (*big data*) que se generan, se nos ocurre fácilmente que captar esos datos y hacer minería de datos (*data mining*) de ellos para vender información es lo que da valor a las redes sociales. Estos datos, una vez procesados y convertidos en inteligencia, son de un valor monetario incalculable.

Esto hace pensar que, una vez conseguida y procesada la información, podría ser usada para manipular y tratar de modificar el comportamiento humano. Esto lleva, como consecuencia lógica, a un problema ético: ¿quién y cómo va a controlar el uso de toda esta información? Se hace necesario mantener prácticas éticas, las cuales no se encuentran del todo definidas.

Un último aspecto que hay que considerar, relativo a la explotación de datos, es que la minería de datos tiene muchas aplicaciones útiles, pero también un enfoque meramente exploratorio que hace discutible la validez de ciertas deducciones. El uso de información personal con fines predictivos tiene consecuencias directas sobre la vida de las personas y exige, por lo tanto, actuar en un marco de responsabilidad. Se hace necesario, entonces, un código de ética.

No podemos finalizar este bloque sin comentar que es muy habitual que las organizaciones desarrollen instrumentos para proteger sus datos, pensando en el su uso fraudulento por parte de empleados, clientes o proveedores.

El objetivo deseado es garantizar la disponibilidad, integridad y confidencialidad de los datos que gestionamos, proporcionando los recursos y aplicando los controles necesarios para conseguirlo.

Para ello, encontramos: códigos de conducta, normas de uso de herramientas electrónicas, políticas de seguridad de la información, acuerdos de confidencialidad, derechos de propiedad intelectual, etc.

Ejemplo de contenidos comunes de estos instrumentos

- Partes afectadas.
- Responsabilidades.
- Definición de información confidencial.
- Deber de confidencialidad.
- Protección de datos: cuentas de usuario, contraseñas, etc.
- Uso de la informática y las comunicaciones.
- Control de acceso y privacidad.
- Propiedad intelectual e industrial.
- Etc.

Resumen

En este módulo hemos abordado la gestión de datos más allá de la FIC y hemos introducido diferentes actividades de gobierno del dato. Para aquellas que están más directamente relacionadas con la FIC, como la gestión de la calidad del dato, hemos entrado en más detalle.

Así mismo, se ha resaltado el concepto de metadato como aspecto crítico en la gestión de datos de la organización, sin olvidar el cumplimiento de la normativa vigente y la atención a la ética en el tratamiento de los datos.

Abreviaturas

API Siglas en inglés de interfaz de programación de aplicaciones: es el conjunto de subrutinas, funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.

BPEL Siglas en inglés de lenguaje de ejecución de procesos de negocio: lenguaje de orquestación de procesos de negocio basada en estándares, compuesto por servicios.

CDI Siglas en inglés de integración de datos de clientes. Disciplina de gestión de datos maestros centrada en la integración y normalización de datos de clientes.

DAMA Siglas de Data Management Agency: asociación Internacional dedicada al avance y definición de mejores prácticas en el entorno de la gestión de datos.

MDM Siglas en inglés de gestión de datos maestros: disciplina que ofrece una visión única de los datos buscando su estandarización y fiabilidad.

Bibliografía

Berson, A.; Dubov, L. (2010). *Master Data Management and Data Governance*. McGraw-Hill/ Osborne Media.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality*. Nueva York: John Wiley & Sons, Inc.

Fan, W.; Geerts, F. (2012). *Foundations of Data Quality Management*. Morgan & Claypool Publishers.