Gestión de datos en un *Data Warehouse*

Juan Vidal Gil

PID_00236071

Universitat Oberta de Catalunya

Índice

Introducción

En otros módulos de esta asignatura hemos visto los almacenes de datos desde una perspectiva general de la organización. En este módulo vamos a centrarnos en estudiarlos desde el punto de vista de la gestión 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 y a la gestión de los metadatos.

Trataremos temas como la gobernanza de los datos bajo el que se enmarcan entre otros la gestión de de datos maestros o la gestión de la calidad de los datos. Igualmente veremos el papel fundamental de los metadatos en todos los procesos relacionados con la gestión del dato.

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

2. Conocer la necesidad y función de la gestión de los datos maestros.

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.

1. 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.

1.1. Disciplinas que intervienen en la integración de datos

En el módulo «Construcción de la FIC», hemos visto 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.

1.2. Gobierno del dato

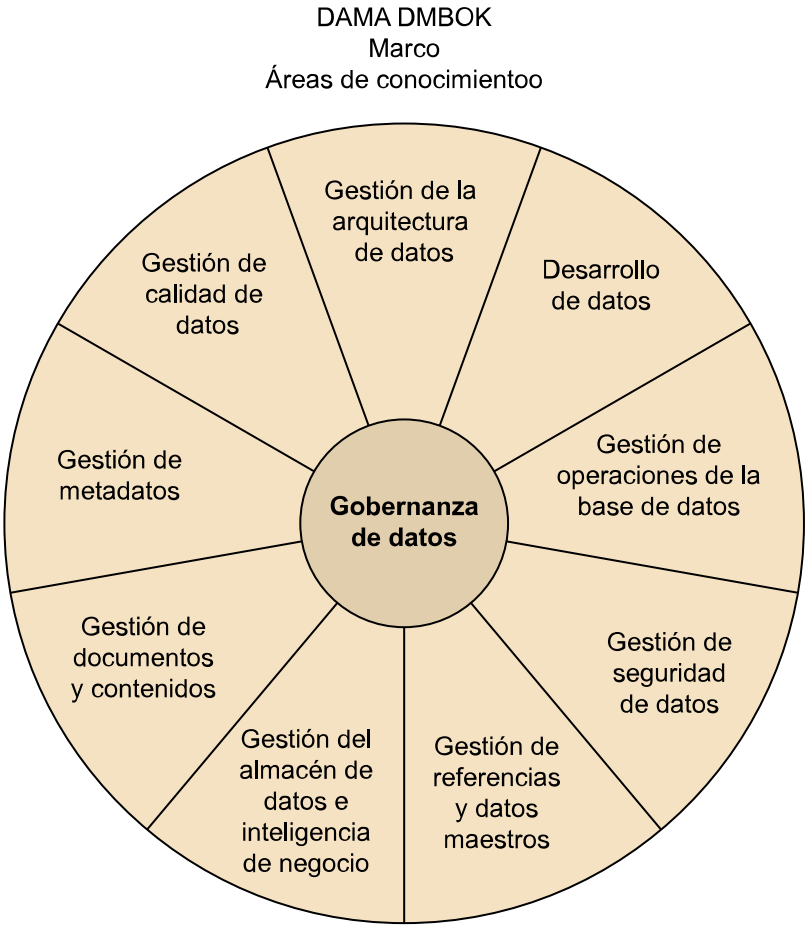
El **gobierno del dato** es una importante disciplina empresarial cuyo objetivo es proporcionar mayor control sobre la creación, manejo, mantenimiento, almacenamiento, uso e intercambio de información vital para el negocio.

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:

- Gestión de datos maestros.
- Calidad del dato.
- Gestión de metadatos.

2. Gestión de datos maestros

Una de las áreas críticas en el gobierno del dato o la gobernanza de datos es la gestión de datos maestros (MDM; *Master Data Management*) que surge de la necesidad de tener una versión única, fiable, compartida y actualizada de la información más crítica de la organización.

La **gestión de datos maestros** consiste en un conjunto de procesos y herramientas que define y gestiona de forma consistente las entidades que representan datos críticos dentro de la organización.

Por tanto, incluye procesos y herramientas para buscar, recopilar, agregar, identificar, asegurar la calidad, la persistencia y la distribución de los datos de forma uniforme.

Ejemplos de entidades críticas que son gestionadas por MDM

Los datos de clientes, de productos, de proveedores o de cuentas pueden ser ejemplos de entidades críticas en una organización.

Sobre estas entidades definiremos los datos maestros.

Un **dato maestro** es un registro único que sirve de referencia para toda la empresa.

Los datos maestros podemos decir que son o contienen datos de referencia dentro de la organización.

Ejemplos de datos maestros que forman parte de las entidades críticas

El nombre de un cliente, su código de contrato, el código de un producto o un número de cuenta son datos de referencia que forman parte de entidades críticas si consideramos el ejemplo anterior.

Veamos a continuación cuáles son los objetivos, etapas y diferentes arquitecturas de implementación de la gestión de datos o MDM.

2.1. Objetivos de la gestión de datos maestros

Los objetivos que persigue la gestión de los datos maestros son los siguientes:

1) Creación de repositorios de datos maestros centralizados. Se trata de evitar que existan diferentes versiones de los datos y/o heterogeneidad entre versiones de una misma entidad en diferentes sistemas.

2) Repositorios actualizados. Un repositorio central debe estar actualizado y ello se consigue mediante los procesos de captura, transformación, integración y actualización necesarios. El proceso de integración es crítico ya que partimos de fuentes origen que pueden ser heterogéneas y la ágil captura y actualización de cambios.

3) Gestión de la calidad de los datos maestros. Sobre un repositorio centralizado se activarán los procesos necesarios para garantizar la calidad del dato, tanto en el proceso de actualización del dato como en procesos de monitorización periódicos que medirán la calidad del dato, activando alertas si fuera necesario.

4) Distribución de los maestros a lo largo de la organización. El repositorio central deberá distribuir los datos de la entidad, para asegurar la visión única en toda la organización. Se implementarán los mecanismos de sincronización necesarios.

2.2. Etapas en el proceso de gestión de datos maestros

La consecución de los objetivos descritos en la anterior sección exige la implementación de una estrategia para llevar a cabo la gestión de datos maestros. Con esta finalidad el proceso de gestión de datos maestros se compone de las etapas siguientes:

- Identificar las fuentes origen de datos.
- Identificar los productores y consumidores de datos maestros.
- Recopilar y analizar metadatos sobre los datos maestros recopilados.
- Determinar los responsables (administradores) de los datos maestros.
- Implementar un programa de gobierno de datos (y por lo tanto, tener un grupo responsable de este programa).
- Desarrollar el modelo de metadatos maestros.
- Diseñar procesos para garantizar la calidad del dato.
- Diseñar la infraestructura necesaria.
- Generar y testear los datos maestros.

- Modificar los sistemas consumidores y productores de información.

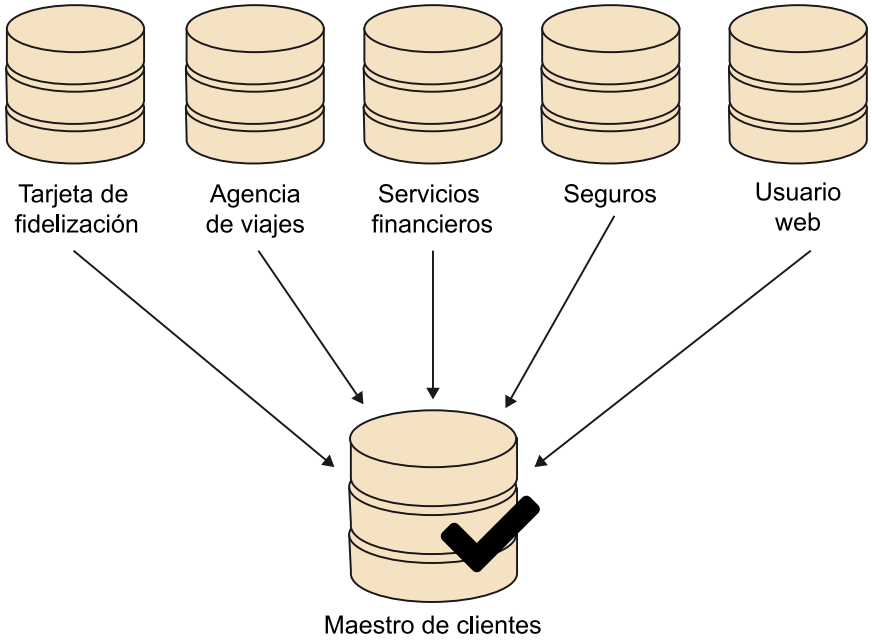
- Implementar un proceso de mantenimiento.

Una de las tareas que hay que abordar en la estrategia MDM es el despliegue de un programa de gobernanza de datos. Este programa une a personas, procesos y tecnología para cambiar la manera en que los datos son adquiridos, gestionados, mantenidos, transformados en información, compartidos en el contexto de la organización como conocimiento común y, de manera sistemática, obtenidos para mejorar la rentabilidad de la empresa. Es decir, hablamos de una disciplina en la que convergen conceptos como calidad de datos, gestión de datos, gestión de procesos de negocio y gestión del riesgo. Con frecuencia, estas iniciativas están motivadas por el cumplimiento de regulaciones que buscan mitigar el riesgo en lo que respecta tanto al ámbito español como al europeo (por ejemplo, Sarbanes-Oxley, Basilea I y II, Health InsurancePortability and Accountability Act –HIPAA–). Las entidades financieras y aseguradoras tienen experiencia en las mismas. También en este contexto confluyen prácticas de despliegue basadas en marcos de referencia en gestión de TI y calidad, tales como COBIT7 (*Control Objectives for Information and Related Technology*), ISO/IEC (*International Organization for Standardization and the International Electrotechnical Commission*), etc.

Ejemplo de aplicación de gestión de datos maestros: el maestro de clientes

Un caso común de MDM es la creación de un maestro de clientes. La visión de una empresa sobre sus clientes puede depender de los diferentes canales de comunicación y las diferentes áreas de la compañía que interaccionan con él. Así pues, dependiendo del canal de comunicación o del sistema de información con el que cada área gestiona sus clientes, podemos tener diferentes registros como representación de los datos que describen un mismo cliente. Teniendo en cuenta que partimos de múltiples y variadas visiones de cada cliente, no resulta trivial identificar y cruzar los datos de un cliente provenientes de diferentes orígenes de datos. Es posible, incluso, que no encontremos una clave que identifique unívocamente al cliente, ya que puede no haber campos comunes a todos los orígenes de datos (por ejemplo, DNI, número de teléfono, contrato, etc.).Dado que interesará guardar la historia de cada cliente, deben identificarse las diferentes versiones de datos de un cliente, las cuales reflejan los distintos estados por los que pasa y las enlaza convenientemente. Por ello hay que identificar valores origen de los campos clave, que permitan enlazar con los distintos cambios de valor de la clave que pueda tener un mismo cliente. Por ejemplo, un cliente puede cambiar de contrato porque han cambiado sus datos de facturación (cambio del titular que paga el servicio) o porque cambia el tipo de servicio que tiene contratado con la compañía. Un sistema de MDM debe permitir enlazar los datos de ese cliente a pesar de estos cambios de forma que pueda realizarle propuestas comerciales adecuadas en función de su evolución.

Figura 2. Maestro de clientes



Una vez que tengamos identificado el cliente, trataremos de obtener el *golden record*, que es un registro normalizado y enriquecido con información de varios orígenes de datos que permite guardar una visión única, en este caso, de los datos del cliente.

Una cuestión crucial en MDM es la generación del ***golden record*** o registro normalizado y enriquecido que debe guardar la visión única del dato.

La obtención de este registro único se realiza partiendo de información de diferentes orígenes en los cuales existen múltiples versiones. Veamos cómo se realiza con un ejemplo.

Ejemplo de obtención *golden record* sobre maestro clientes

Supongamos que tenemos múltiples versiones de los datos de un cliente (ver la figura 3).

Figura 3. Diferentes visiones de un cliente

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

Un primer paso es realizar las comprobaciones necesarias para identificar los distintos registros como registros de un mismo cliente. Luego normalizaremos sus campos.

En la figura 3 vemos que el campo CIF no nos sirve como clave de cruce, ya que encontramos sistemas origen en los que no se ha registrado el CIF (ver segundo registro). Por tanto, habrá que realizar cruces adicionales para identificar clientes que, aunque sean el mismo, no cruzan por CIF. En estos cruces adicionales pueden emplearse múltiples campos (población, municipio, teléfono, código postal, dirección). Es posible que para algunos de los campos seleccionados no sea posible realizar un cruce por valor exacto, como el campo dirección (no es idéntico en los tres registros). Para este tipo de casos, habitualmente sucede con campos de tipo cadena, se utilizarán algoritmos de comparación de cadenas que retornan un valor indicador del grado de similitud entre las cadenas.

Ya identificados los registros que corresponden al mismo cliente, podemos crear el *golden record* que tiene en determinados campos un valor normalizado, que puede no corresponderse con el valor de este campo en las diferentes versiones. En el caso del ejemplo, los campos relativos a la localización (población, municipio, código postal, dirección) pueden normalizarse y completarse utilizando bases de datos de referencia con información geográfica o bases de datos de direcciones postales.

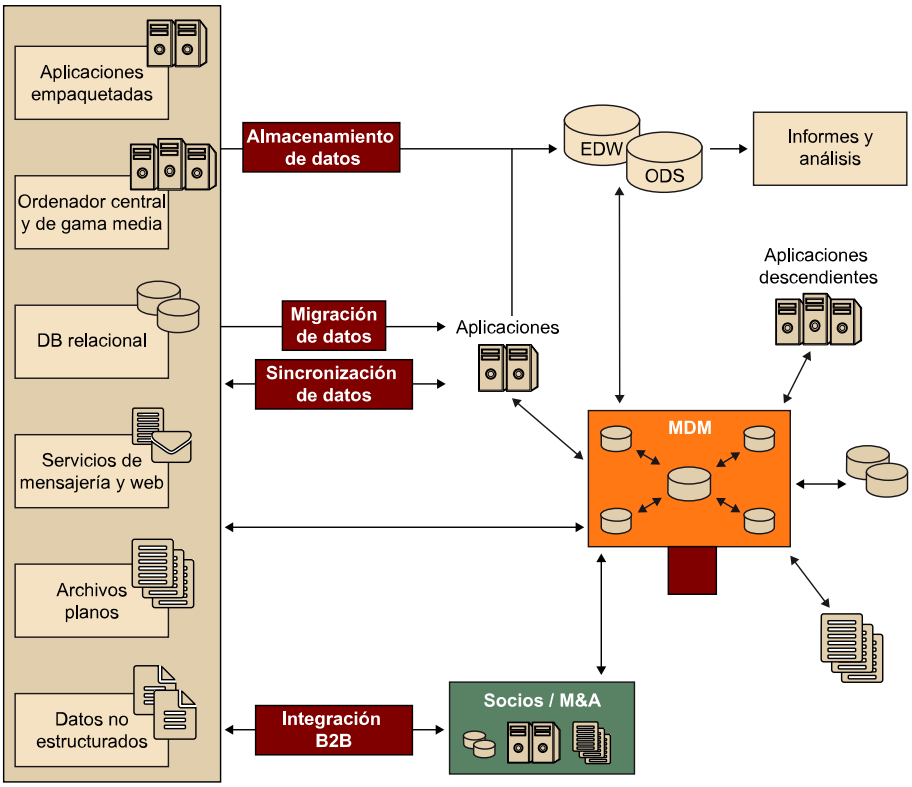
Una vez generado el *golden record*, ya podremos crear claves internas que permitan enlazar este registro con las diferentes versiones en los sistemas origen.

2.3. Arquitecturas de implementación del *hub* MDM

Hemos visto que para la gestión de los datos maestros es necesario realizar un conjunto de tareas y que para ello es necesario un repositorio centralizado. A este repositorio le llamamos *hub* y por referirse a datos maestros, *hub* MDM.

El ***hub*** es un repositorio con una base de datos centralizada de datos maestros y los procesos para sincronizarse con los sistemas origen y destino.

Figura 4. Conexión del *hub* MDM con los sistemas de la organización



Existen distintas arquitecturas de implementación del *hub*:

1) *Hub* MDM transaccional (también llamado centralizado): utiliza aplicaciones transaccionales para mostrar los datos unificados, almacenando los registros en su propia estructura de datos. El dato se encuentra tanto en los sistemas orígenes como en el almacén MDM, y requiere una capa ETL para la sincronización con los sistemas. También posee un tratamiento de los datos

sincronizados, dado que en este tipo de *hub* están incluidas las capas de calidad y estandarización que posteriormente serán replicadas a los sistemas orígenes. Por otro lado, adoptar este tipo de *hub* supone que el dato consolidado en el mismo es tratado como dato maestro dentro de la compañía. Este tipo de arquitectura es más lenta de implantar y más intrusiva con los sistemas origen, pero aporta una única fuente y la calidad de los datos está controlada desde un único punto.

2) *Hub* MDM registro (también llamado federado): los datos se integran virtualmente mediante aplicaciones. Esto es, a partir de los índices que relacionan los datos reales contenidos en los diferentes sistemas transaccionales de la compañía (claves principales del registro), crean sus metadatos permitiendo el acceso a los mismos en modo consulta. Por ejemplo, si hay registros de un cliente en el CRM, en el sistema de facturación y en el de servicio al cliente, el *hub* MDM tiene los tres registros relacionados bajo una clave común, pero no hay un registro único (*golden record*) almacenado.

Este tipo de *hub* es utilizado cuando existe un gran número de sistemas de origen de datos. Su sincronización es por medio de la lectura de los datos y un tratamiento en espacios de trabajo virtuales hasta que consiguen categorizar e indexar las relaciones de manera interna. Los datos nunca vuelven a incorporarse a los orígenes, queda todo en el *hub*, con lo que no existe una compleja capa ETL en esta arquitectura. Una de las cosas que hay que tener en cuenta en estos sistemas es la latencia de sincronización del dato. Un factor negativo de este tipo de aplicaciones es la alta frecuencia de lectura de los orígenes para traerse los datos de los registros que afectaría al rendimiento de los mismos.

3) *Hub* MDM híbrido: similar al estilo de las aplicaciones *hub* de registro pero implementado con una capa ETL que permite la sincronización de los datos y evita los problemas de latencia en las lecturas con los sistemas origen. Los registros son almacenados de forma virtual, pero la consolidación de los registros o registro único (*golden record*) sí que es almacenada en la estructura de metadatos de la aplicación, al menos en lo que respecta a sus principales atributos. Finalmente, para volver a replicar los datos consolidados a los orígenes, se suelen utilizar estrategias de sincronización «no invasivas» utilizando bien las API de los propios sistemas o bien otro tipo de soluciones de flujos de trabajo BPEL (*Business Process Execution Language* es una forma de orquestar procesos de negocio basada en estándares, compuestos por servicios).

Como resumen, en la tabla 1 se presenta una comparación de los distintos tipos de *hub* según sus principales características:

Tabla 1. Comparativa de diferentes tipos de *Hub*

	Transaccional	**Registro**	**Híbrido**
Intrusivo en sistemas origen	Sí	No	Parcialmente
Movimiento de datos	Sí	No	Parte de los datos

	Transaccional	Registro	Híbrido
Normalización, calidad y enriquecimiento	Sí	No	Parte de los datos
Sincronización sistemas origen	Sí	No	Estrategias no invasivas

En función de la arquitectura de *hub* escogida para el entorno MDM, quedará definida la importancia que le daremos al dato unificado obtenido. De esta forma un modelo transaccional o híbrido nos obligará a diseñar una estrategia de sincronización con una capa de integración.

Si la necesidad de disponer del dato unificado es crítica para nuestra organización, la sincronización requerida puede impactar de manera negativa en nuestros sistemas orígenes. Hemos de considerar que una arquitectura que replicase los datos en los sistemas orígenes podría ser lenta y costosa de implementar, por lo tanto debemos plantearnos si realmente la necesidad del dato unificado es crítica para nuestra organización. Por ejemplo, si pensamos en escenarios de compañías financieras o aseguradoras, la posibilidad de tener el dato correcto en tiempo real puede suponer una solución de alto coste, lo cual podría volverse un factor en contra del proyecto. Pero si la necesidad de tener el dato unificado no es crítica, existen opciones que evitan un desarrollo de un entorno complejo en cuanto a integración. Y aunque hay modelos que también exigen un importante coste de desarrollo en las aplicaciones (sistemas orígenes), quizás optar por un modelo híbrido resulte lo más eficiente.

Una cuestión crítica en el diseño del *hub* es la frecuencia de sincronización con los sistemas origen o destino y el modo de acceso (intrusivo o no intrusivo según penalice o no los sistemas).

En caso de un modelo transaccional o híbrido, en el que es necesaria una capa de integración para los datos, tenemos que consolidar primero nuestro *hub*, una vez consolidados los datos en nuestro entorno, se revertirían los datos ya depurados a los sistemas orígenes. Debemos utilizar cargas *batch* o latencias de refresco ajustadas para evitar colapsar los sistemas orígenes. El cómo de intrusivo o no queramos nuestro modelo marcará la diferencia a la hora de trabajar con los datos en nuestros sistemas transaccionales y, por ende, en nuestro entorno MDM.

En el *hub* MDM es útil añadir los campos adicionales de metadatos necesarios para auditar todos los cambios realizados. Por razones técnicas y legales es necesario conocer para cada dato maestro en una modificación:

- cuál era el valor original,
- cuál es el valor actual,
- quién lo cambió,
- motivo del cambio y

- fecha del cambio.

Actualmente, existen en el mercado diferentes soluciones para implementar procesos de MDM. Hay dos tipos principales:

a) Especializadas por dominios: clientes, productos, cuentas, proveedores, empleados. Un caso habitual es una solución CDI (*Customer Data Integration*).

b) Multidominio: gestión integral de todos los datos maestros de la compañía. No se centra en una entidad concreta, sino que abarca todas las entidades que se gestionan con MDM (clientes, productos, proveedores, cuentas, etc.).

2.4. Diferencias entre almacenes de datos y datos maestros

Aunque en algunos aspectos los almacenes de datos guardan similitudes con los datos maestros su función no es exactamente la misma.

Si bien en ambos casos se unifican los datos procedentes de los sistemas fuente, los almacenes de datos no han sido concebidos para devolver los datos modificados a dichos sistemas fuente.

Un **almacén de datos** emplea un proceso monodireccional para cargar los datos en él, mientras que el MDM necesita de un proceso bidireccional que garantice la sincronización de los datos entre el repositorio y los sistemas origen y destino asociados.

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:

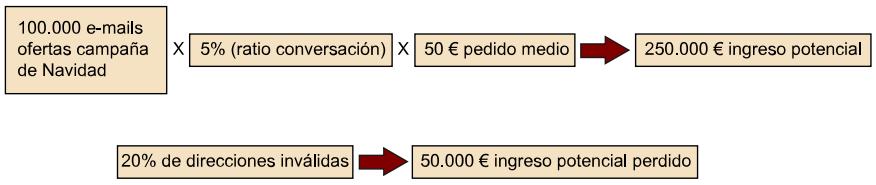
- Visión única del cliente-usuario/producto-servicio/proveedores.
- Mejora de la comunicación cliente-usuario/proveedores.
- Ahorro de tiempos de conciliación de la información.
- Garantía de la correcta unificación de bases de datos (fusiones de empresas).
- Confianza en el dato, mejora de procesos 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 5 podemos ver el coste que puede tener una baja calidad del dato en una campaña de e-mail marketing.

Figura 5. 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 6 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, número 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 6. 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.

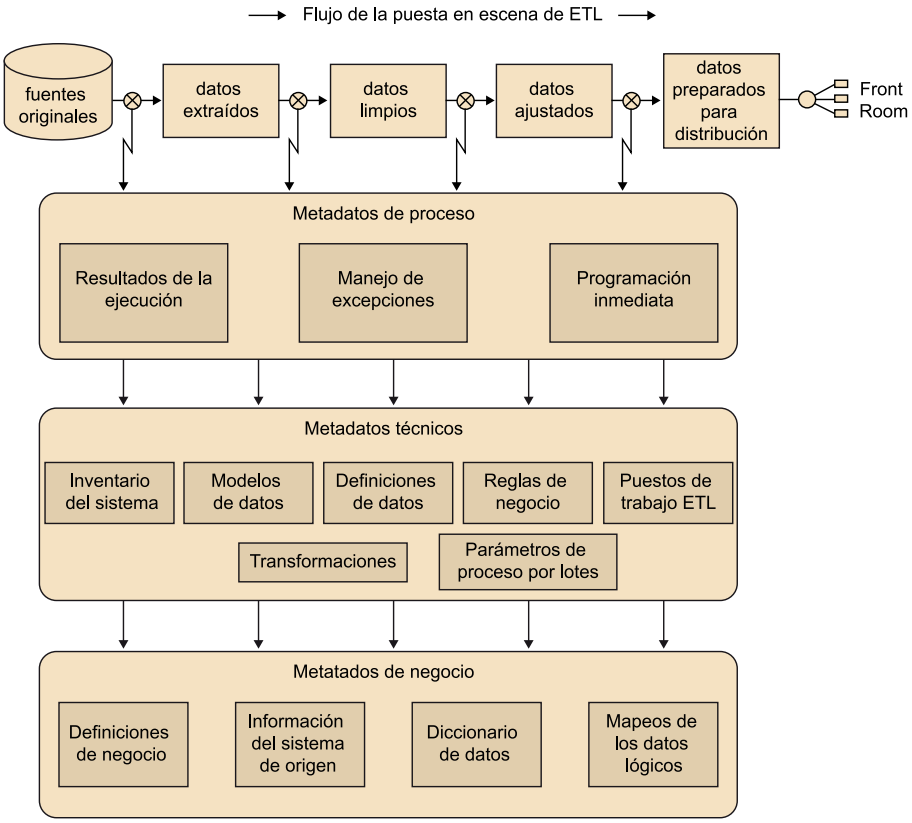
Los metadatos no son ámbito exclusivo de la FIC y son un elemento crítico en las actividades de gobierno del dato.

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 7.

Figura 7. 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.

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 los datos maestros y 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.

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.