Los datos en sistemas de *data warehouse*

Alberto Abelló Gamazo
Josep Curto Díaz
Carles Llorach Rius
José Samos Jiménez

PID_00236084

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Los sistemas de inteligencia de negocio (BI) son los encargados de transformar los datos en información completa, correcta y consistente, lo que resulta una ventaja competitiva que nos permitirá formular una sólida estrategia corporativa.

William Inmon presentó en 1998 lo que se denomina factoría de Información corporativa (en adelante, FIC). Se trata de un conjunto de componentes que interactúan para ayudar a gestionar todos los flujos de datos desde los sistemas operacionales de la empresa hacia los analistas. Su objetivo es transformar los datos de los sistemas operacionales (materias primas) en información de apoyo a los analistas (producto elaborado), para utilizarla en los procesos de toma de decisiones en la organización.

Una pieza esencial del sistema BI (o FIC) es el almacén de datos (en inglés, *data warehouse*, DW), cuya función es la de repositorio de información orientado a recopilar, resumir y tratar eficientemente el gran volumen de datos presente en las organizaciones, para facilitar el análisis y la toma de decisiones, añadiendo valor y generando beneficio para el negocio.

Definimos *data warehouse* como la colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo a la toma de decisiones.

Entenderemos por sistema de *data warehouse* el formado por dicho almacén y por el conjunto de herramientas y procedimientos que nos permiten explotarlo y gestionarlo a lo largo del tiempo.

Con la llegada del *big data*, los almacenes de datos evolucionan y surgen nuevos almacenes optimizados para soportar las características propias del mismo, que complementaran los almacenes de datos tradicionales, los cuales seguirán existiendo por las bondades que ya hemos explicado.

Para que la información que obtenemos del sistema responda a los propósitos anteriores, deberemos tratar las fuentes de datos para que los datos provenientes de estas y que alimentan el sistema cumplan las características exigidas para su procesamiento. De ahí la importancia de los datos y su adecuado tratamiento.

Objetivos

Los contenidos incluidos en este módulo se orientan a conseguir que el estudiante alcance los objetivos siguientes:

1. Tomar conciencia de la importancia de los datos de un almacén de datos por sí mismos y de la necesidad de su adecuada gestión.

2. Comprender la utilidad de los metadatos y los tipos existentes.

3. Conocer el marco legal existente y los principios que deben regir nuestra actuación en torno a los datos almacenados en los sistemas *data warehouse*.

4. Identificar apropiadamente las necesidades de los usuarios según las tipologías descritas en este material.

5. Comprender la importancia de la visualización de datos en los procesos de toma de decisiones de una organización, usando aquellas técnicas y herramientas idóneas para la explotación de datos requerida.

6. Conocer ampliamente las actividades que se deben llevar a cabo para poner en funcionamiento y, posteriormente, mantener un almacén de datos a lo largo de su vida útil, garantizando su accesibilidad, fiabilidad y adaptación a las nuevas necesidades que se generen.

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 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 son la gestión de datos maestros (en inglés, *master data management*, MDM) y la limpieza de datos (en inglés, *data cleaning* o *data scrubbing*), entre otros.

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 (es 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 genera 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 de ser contemplados a la hora de diseñar nuestros programas.

1.2. Los metadatos

Los metadatos no son un elemento específico de la FIC: aparecen en muchos contextos del mundo del software. La definición más frecuente que hay del concepto de metadato está basada en su etimología[1]: «Los metadatos son datos sobre datos». Los datos generalmente representan características de las entidades que modelan; en el caso de los metadatos, representan características

(1)*Meta*, en griego, significa 'sobre'.

de otros datos que facilitan su administración y uso. Es decir, lo que diferencia a un dato de un metadato, más que su estructura o contenido, es su propósito y uso.

Teniendo en cuenta un conjunto de datos, los metadatos sobre estos describen sus características (por ejemplo, formato, origen, uso, etc.). Estos metadatos son datos y a su vez podemos tener otros metadatos que describan sus características (metadatos sobre metadatos), y así de manera sucesiva.

En este apartado, empezamos revisando el uso de los metadatos en la FIC. A continuación, se presentan diferentes tipos de metadatos según el uso de los mismos. También se analiza la manera en que se crean los metadatos, así como los estándares definidos para permitir compartirlos entre distintos componentes. La sección acaba comentando la necesidad de utilizar diferentes versiones de metadatos en la FIC.

Los metadatos son el componente más importante de la FIC, puesto que cohesionan el resto de los componentes de los que también forman parte.

1.2.1. Uso y tipos de metadatos

La información ofrecida por los metadatos nos permite entender mejor la estructura, el funcionamiento y los resultados de los sistemas que describen. Es decir, los metadatos resultan interesantes para el equipo de desarrollo del sistema, los técnicos que hacen que el sistema funcione y los usuarios finales que lo utilizan. De este modo, los podemos clasificar según el papel de las personas que los utilizan.

Estos conjuntos de metadatos no son disjuntos, es decir, se utilizarán los mismos metadatos con objetivos diferentes.

Metadatos de construcción

Los equipos de desarrollo de los sistemas definen gran parte de los metadatos; posteriormente, estos se usarán con otros objetivos diferentes en la construcción de los sistemas. En el caso de la FIC, definen la estructura de las distintas fuentes de datos, de los almacenes de datos, las transformaciones que hay que hacer, la planificación, etc.

Los metadatos de construcción tienen gran importancia, puesto que hacen que los sistemas sobre los que se definen sean más flexibles y fáciles de evolucionar.

Los metadatos son tan importantes en este aspecto de los sistemas que a veces este es el único uso que se les reconoce.

Metadatos de gestión

Durante el funcionamiento del sistema, para gestionarlo se utilizan algunos de los metadatos definidos durante la construcción y también se definen otros nuevos. Todos estos forman los metadatos de gestión. En la FIC se define a los usuarios que utilizarán los diferentes almacenes de datos, se almacena información sobre el uso que hacen de los mismos, sobre el resultado de las extracciones y las transformaciones de datos, etc.

Los metadatos de gestión son utilizados por los técnicos que administran el sistema y hacen que este funcione.

Metadatos de uso

Los analistas generalmente no definen metadatos (tampoco datos), o no al menos directamente; sino que se limitan a hacer consultas sobre ellos. Además de consultar datos, también necesitan hacer consultas sobre los metadatos tanto de construcción como de gestión. No tendrán acceso a todos los metadatos definidos, sino solo a aquellos que los constructores del sistema hayan considerado de su interés según su perfil de usuario.

Ejemplo de metadatos de uso

Un analista necesita consultar los resultados de las ventas de una cadena de tiendas. Las ventas registradas en los sistemas operacionales de las tiendas se cargan cada día en el almacén de datos utilizado. El analista puede consultar los resultados propiamente dichos y el significado de un dato concreto (la fórmula utilizada para calcularlo: metadatos de construcción), y también puede hacer consultas sobre incidencias particulares que han tenido lugar para obtenerlos (si faltan los datos de alguna tienda: metadatos de gestión).

Generalmente, los usuarios de los sistemas operacionales solo necesitan trabajar con los datos de negocio almacenados en los sistemas. Los usuarios de los almacenes de datos, además de datos, necesitan metadatos.

Los metadatos, tanto de construcción como de gestión, tienen gran importancia para los usuarios de los almacenes de datos, ya que les suministran la información que necesitan sobre el significado o el estado de los datos que consultan.

1.2.2. El proceso de definición de los metadatos

Las ventajas de disponer de metadatos en cualquier sistema son indudables, puesto que proporcionan de manera explícita información que facilita la evolución, la gestión y el uso del sistema. Podemos definir los metadatos de manera manual o bien mediante alguna herramienta. Asimismo, se pueden definir antes, durante o después de la construcción del sistema al que están asociados.

La situación ideal es que dispongamos de una herramienta para construir el sistema y que la definición de los metadatos asociados forme parte del proceso de construcción y mantenimiento, de modo que el sistema y los metadatos asociados evolucionen de manera conjunta.

Si no forman parte necesaria del proceso de desarrollo del sistema, se corre el riesgo de que, por limitaciones en el tiempo de desarrollo, en el presupuesto o por otros motivos, los metadatos no se actualicen con el sistema al que están asociados y se produzca, de este modo, una discordancia entre los metadatos y el sistema que describen.

Ejemplo de metadatos obsoletos

Una editorial quiere gestionar los metadatos de sus libros. Los metadatos son fundamentales para la correcta distribución y venta, y su uso (y mal uso) afecta a toda la industria editorial.

La mayoría de los metadatos se crean en la editorial durante los procesos de edición y producción del libro, y en la toma de decisiones de marketing y ventas. El responsable de derechos, el de edición, el de producción, el de marketing y el de ventas crean y añaden al libro los metadatos correspondientes a su gestión.

El problema surge cuando cada uno de estos responsables o sus departamentos utilizan métodos de gestión manual de metadatos diferentes o independientes: cuando en la editorial no existe una gestión central de metadatos. Entonces, lo más probable es que los metadatos se dupliquen, se contradigan, pierdan calidad o queden obsoletos. Por no hablar del gasto innecesario en tiempo y dinero.

1.2.3. Estándares de metadatos

En sistemas complejos (como el caso de la FIC), cada componente dispone de sus metadatos. Para definirlos, se han podido utilizar diferentes herramientas de apoyo: herramientas CASE, herramientas del SGBD, herramientas del componente de integración y transformación, etc. Por lo tanto, cada componente tiene sus metadatos, almacenados según su criterio y formato particular.

Metadatos asociados al sistema

De este modo, como se presenta en Inmon, Imhoff y Sousa (1998), se consigue que los metadatos sean completos, no sean un elemento opcional, se actualicen de manera automática cada vez que se hagan modificaciones en el sistema y no requieran un esfuerzo adicional para mantenerlos.

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

A lo largo de la historia (es una historia relativamente corta, pues se inicia a principios de los noventa), se han definido diferentes estándares relacionados con metadatos. De manera particular, relacionados con la FIC, encontramos principalmente el *common warehouse metadata* (CWM). Su objetivo es definir un repositorio central que permita integrar los metadatos que hay definidos para las diferentes herramientas, de modo que se mantenga una sola versión de todos ellos. Con este fin, el CWM define un modelo de datos para el almacenamiento formado por submodelos específicos para cada área, y un conjunto de capas de acceso al repositorio que ofrecen distintos grados de funcionalidad.

Lectura complementaria

Podéis encontrar más detalles sobre los estándares en un anexo dedicado a este tema en **W. A. Giovinazzo** (2000). *Object Oriented Data Warehouse Design*. Nueva Jersey: Prentice Hall PTR.

1.2.4. Metadatos históricos

Una de las características de los almacenes de datos es que almacenan datos históricos, como hemos visto en el apartado «Características de un almacén de datos» del módulo «Introducción al almacenamiento de datos». A lo largo del tiempo, las estructuras y otras características de los componentes de la FIC, los datos de las fuentes de datos y de los almacenes de datos, las correspondencias entre estos y las transformaciones que se llevan a cabo han podido cambiar. Junto a estos componentes, también habrán cambiado los metadatos que los describen.

Por lo tanto, si en los almacenes de datos tenemos datos históricos con características diferentes, para cada conjunto de datos definido bajo las mismas características tendremos que almacenar la versión de los metadatos que los definen. De este modo, los analistas podrán saber para cada dato cómo está almacenado o en qué condiciones se obtuvo.

Se necesita mantener un control de versiones de los metadatos de la FIC.

1.2.5. Metadatos de usuario

El espaciamiento de los momentos de desarrollo de cada aplicación, las diferencias de presupuestos y requerimientos y la falta de planificación hacen que encontremos más heterogeneidades de las que querríamos entre los sistemas operacionales. Si los analistas quisieran acceder directamente a los datos de estas fuentes de información, lo primero que deberían hacer sería superar estas heterogeneidades entre las aplicaciones.

Podemos encontrar tanto heterogeneidades semánticas (el mismo tipo de información representado de maneras diferentes), como de sistemas (por ejemplo, hardware diferente, sistema operativo distinto o simplemente sistema de gestión de base de datos –SGBD– diferente).

Para resolver el primer tipo, nos encontramos con herramientas que nos permiten definir una capa semántica (también llamada lógica) de traducción entre la base de datos y la capa de presentación. Esta capa intermedia nos permite mapear columnas de la base de datos con objetos de negocio de manera que el usuario, para crear una consulta, no tiene por qué conocer nombres de tablas ni de columnas, que normalmente, por temas de normalización, suelen tener nombres bastante extraños.

Ejemplo de cambio de aplicación transaccional

Imaginemos que hemos utilizado en nuestros informes las entidades de negocio representadas por Clientes y Facturas. Estas ya las hemos «vinculado» a las tablas TR00_COS y TR00_INV respectivamente, donde sus datos residen. Si cambia la aplicación transaccional y la nueva tabla de facturas se llama ahora «FAC_CLIENTE», solo necesitamos vincular la nueva tabla, pero no será necesario cambiar todos los informes que haya podido crear hasta ese momento.

Serán los usuarios del sistema quienes definan y mantengan esta capa semántica, que formará parte del conjunto de metadatos de usuario, junto al resto de «metadatos» que un usuario genera en su interacción con el almacén de datos.

2. Aspectos legales y éticos de los datos

2.1. 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.

2.1.1. Protección de datos

La Ley Orgánica de Protección de Datos (LOPD), despliega la principal norma en la materia es la Directiva 95/46 de la Unión Europea. Pero con anterioridad, ya en 1981, el Consejo de Europa había adoptado el Convenio nº 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 ha entrado 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).

2.1.2. 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.

2.1.3. 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 de tratamiento automático 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.

2.2. Ética de los datos

La ética en la informática es una disciplina nueva que pretende abrirse campo dentro de las éticas aplicadas, por lo cual 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 EI «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 se 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 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 uso fraudulento de los mismos 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.

3. Usuarios del sistema

3.1. Usuarios en función de los datos

En primer lugar, en este apartado veremos los dos extremos de la cadena de obtención de información, es decir, quiénes son los usuarios de la FIC (qué queremos) y cuáles son las fuentes de información que han de satisfacer sus necesidades (qué tenemos). Esto nos hará pensar sobre qué debe haber en medio para que la información fluya del uno al otro.

3.1.1. Los usuarios

Los sistemas operacionales tienen muchos usuarios que acceden a muy pocos datos, mientras que, en lo que respecta a los sistemas de análisis, los utilizan muy pocos usuarios que quieren ver muchos datos. Recordemos que los sistemas operacionales se utilizan en el día a día de la empresa. Sirven para facilitar tareas rutinarias y repetitivas de los oficinistas. Cuando hablamos de tareas de análisis, las cosas se vuelven algo más complejas y podemos identificar diferentes tipos de usuarios, que podemos denominar: granjero, explorador y turista. Realmente, estos nombres no son muy importantes. Lo que sí importa son las características de cada uno y los requisitos que presentan.

Lectura recomendada

Podéis encontrar los tres tipos de usuarios extensamente explicados en la obra siguiente: **W. H. Inmon; C. Imhoff; R. Sousa** (1998). *Corporate Information Factory*. EE. UU.: John Wiley & Sons, Inc.

Los analistas tienen requerimientos diferentes de los que presentan los oficinistas. Además, podemos diferenciar distintos tipos de analistas con características muy diferentes, que la FIC debe tener en cuenta.

Granjero

Este primer tipo de usuario lleva a cabo accesos a la información absolutamente predecibles y repetitivos. De manera regular, encuentra cosas interesantes que ayudan a que la empresa funcione. En todo momento sabe qué quiere y cómo lo ha de obtener, porque, generalmente, repite las consultas de manera periódica. Podríamos decir que tiene su parcela de información y se dedica a cultivarla para extraer provecho de la misma regularmente. No accede a grandes cantidades de datos (puesto que nunca sale de su parcela) y los suele pedir resumidos, aunque le puede llegar a interesar ver diferentes niveles de detalle.

Este tipo de usuario suele utilizar herramientas OLAP[2]. Estas herramientas están pensadas para ser utilizadas por personal no informático. Son sencillas, comprensibles y ponen énfasis en la presentación de los resultados. Mediante

(2)Sigla de la expresión inglesa *on-line analytical processing*, 'procesamiento analítico en línea'.

el modelo multidimensional (muy cercano a la manera de entender el negocio de este tipo de usuarios), consiguen reflejar la complejidad que hay en las estructuras y relaciones de la vida real.

En este grupo tenemos a los empleados, los proveedores y los clientes a los que la organización proporciona servicios informacionales. Actualmente, la inteligencia de negocio operacional, que potencia el uso de estos sistemas en todas las capas de la organización, permite a los usuarios de negocio utilizar los datos y la información en los procesos de negocio de manera natural, sin tener que salir de sus aplicaciones.

Esto se debe al hecho de que la información se encuentra integrada y en cualquier momento es accesible a los procesos de negocio, de modo que los usuarios mismos muchas veces no son conscientes ni del hecho de que usan el almacén de datos.

Ejemplo de análisis en línea

Como ejemplo de granjero, podemos pensar en la persona encargada de hacer previsiones de *stock* para los almacenes. Esta persona seguramente querrá disponer de los datos de *stock* de cada producto durante los últimos años, y también de los pedidos pendientes de servir. Basándose en estos datos, tendrá que decidir qué hay que comprar y cuándo. Si compráramos demasiado o a deshora, la empresa podría perder mucho dinero. No se tiene que confundir este analista con la persona que simplemente registra las entradas y salidas del almacén, que no debe tomar ninguna decisión.

Explorador

Hay otros usuarios analistas que, al contrario que los granjeros, tienen unos accesos totalmente imprevisibles e irregulares. Pasan una gran parte del tiempo sin consultar los datos, planificando o preparando su estudio y, cuando lo tienen todo a punto, empiezan a explorar de repente una gran cantidad de datos tan detallados como sea posible. Realmente, no saben exactamente qué buscan hasta que lo encuentran, y los resultados en ningún caso están garantizados. Sin embargo, a veces encuentran algo realmente interesante que claramente mejora el negocio. Con frecuencia se conocen como usuarios exploradores (*power users*) de la organización.

Un usuario explorador suele ser informático y/o estadístico, experto en prospección de datos y por lo tanto, con dominio de herramientas de análisis estadístico. Estas herramientas tratan de extraer información oculta (no evidente) de un conjunto de datos. Generalmente, son semiautomáticas (como mínimo piden algunos parámetros o que los usuarios validen los resultados) y tienen que estar controladas por técnicos especializados.

En el contexto actual, como resultado de la problemática conocida como *big data*, la figura del explorador ha evolucionado hacia una nueva figura: el científico de datos (*data scientist*). Un científico de datos tiene que ser capaz de extraer información de grandes conjuntos de datos (en términos del problema de *big data*) de acuerdo con un objetivo claro de negocio, no aleatoriamente, y después presentarla de manera sencilla al resto de los usuarios no expertos de la organización. Por lo tanto, se trata de un perfil transversal con conocimientos de informática, matemáticas, estadística, minería de datos, diseño gráfico, visualización de datos y usabilidad.

Big data

Cuando hablamos de *big data*, nos referimos al crecimiento de los datos en volumetría, en velocidad de generación y en variabilidad de origen y forma.

Este perfil será clave para las organizaciones que quieren generar ventajas competitivas a partir de la información. En los próximos años, la demanda de este perfil se incrementará precisamente en aquellas organizaciones que ya tienen en consideración este tipo de necesidad y están desplegando iniciativas de analítica de negocio, es decir, en las organizaciones que ya han logrado un nivel de madurez alto en la explotación de datos y en la generación de información de valor.

Ejemplos de minería de datos

Podemos utilizar herramientas de minería de datos para reconocer patrones de comportamiento para detectar fraudes (facturas, hipotecas o llamadas telefónicas impagadas); generar reglas de manera automática para componer una cartera de valores invertidos en bolsa; encontrar factores de riesgo en un postoperatorio; o descubrir relaciones entre las compras de ciertos productos en el supermercado (por ejemplo, pañales y cerveza).

Turista

Tendríamos que entender este último tipo de usuario como un equipo formado por dos o más personas. Por un lado, tendríamos a la persona que posee una visión global de la empresa a la que se le ocurre la posibilidad de hacer un estudio sobre un cierto tema. Por otro lado, habría un experto en informática, conocedor de los sistemas de análisis de la empresa, encargado de averiguar si el estudio es factible con los datos y las herramientas disponibles o no.

Este equipo mirará datos sin seguir ningún patrón de acceso, y raramente observará dos veces los mismos datos. Por lo tanto, tampoco podemos conocer sus requerimientos *a priori*. Además de los datos, también estará especialmente interesado en consultar los metadatos. Las herramientas que utilizarán los turistas son navegadores o buscadores (tanto de datos como de metadatos), y el resultado de su trabajo serán proyectos que llevarán a cabo los granjeros o los exploradores.

Un usuario turista es, en definitiva, un usuario casual de la información.

Vamos a encontrar evoluciones a esta clasificación, como la que proponen Imhoff y Pettit (2004) con los siguientes tipos de usuarios: agricultores (*farmers*), turistas (*tourists*), exploradores (*explorers*), mineros (*miners*) y operadores (*operators*).

- Los **agricultores** vienen de la parte administrativa o de negocios de la empresa. Puede ser el analista financiero o el analista de ventas, y ve el mundo desde la perspectiva de productos, segmentos de mercado, campañas y canal de ventas. El detalle de los análisis efectuados irá hasta dos niveles hacia abajo, sin llegar al más preciso detalle.

- Los **turistas** vienen de la parte ejecutiva de la corporación, o de departamentos técnicos con mucho dominio de internet, y son los más críticos del sistema. En muchos casos, tienen una perspectiva muy amplia del negocio; en la mayoría de los casos, requieren de una interface muy consistente para buscar en múltiples bases de datos de una manera sencilla y poder identificar asuntos de interés. La arquitectura de los metadatos debe ayudar en esto.

- Los **operadores** son los usuarios más comunes del sistema, y normalmente solicitan información estandarizada en forma regular, para lo cual necesitan herramientas de búsqueda estandarizada. Normalmente provienen de la parte administrativa o del nivel administrativo intermedio, y requieren de información táctica e histórica de una manera rápida e integrada.

- Los **exploradores** son usuarios poco convencionales que llevan a cabo análisis específicos que, en muchos casos, ofrecen conocimiento muy relevante. Llevan a cabo búsquedas al azar, procedimientos poco convencionales y determinación de patrones y relaciones, y plantean sus propias hipótesis, que luego tratan de probar. Utilizan herramientas de OLAP, minería de datos y herramientas de visualización.

- Los **mineros** buscan en grandes bases de datos para encontrar algunos datos o patrones específicos, para lo cual requieren datos históricos y muy detallados, herramientas específicas de minería de datos y otras herramientas de búsqueda, las que utilizarán para hacer clasificaciones, estimaciones, predicciones, segmentaciones, y descripciones.

3.2. Usuarios según rol en la organización

A medida que las organizaciones y sus sistemas de inteligencia de negocio han ido evolucionando y aumentando en complejidad, aparecen nuevos roles para cubrir las necesidades que supone la gestión de los datos.

Parece que hay consenso en identificar los roles de propietarios (*owners*), los administradores (*stewards*) y los usuarios (*users*), que corresponden a los descritos en el punto anterior.

3.2.1. Propietario

La propiedad de los datos (y el acceso a ellos) es uno de los puntos más delicados de los sistemas de *data warehouse*. No siempre será posible encajar las necesidades de una organización aplicando la estrategia más simple: acceso total, departamental o por tema específico.

En primer lugar, identificamos dos niveles de propiedad de los datos: según quién los produce y según el cubo/tema donde van a residir. Si atendemos al nivel de producción de datos, está claro que implica que los usuarios que desean datos de los sistemas de producción tienen que pedir al propietario del sistema de producción permisos y derechos de acceso. El propietario del sistema de producción permite acceso al cubo de datos correspondiente y/o permite utilizar los datos de producción.

Cubo OLAP

Un cubo OLAP (*online analytical processing*), término acuñado por Edgar EF Codd, es una base de datos multidimensional, en la cual el almacenamiento físico de los datos se hace en un vector multidimensional.

La situación es más complicada cuando los cubos o temas están hechos a partir de varias fuentes de producción. En estos casos: ¿quién es el dueño del cubo? Generalmente, se determina que sea la persona que hizo la petición que inició la construcción del cubo. Así, el propietario del cubo o el tema es el segundo nivel de propiedad. Cualquiera que quiera utilizar datos de un cubo o tema debe preguntar y solicitar derechos de acceso al propietario del cubo. Este solicitará el acceso a los datos en los sistemas de origen, no los usuarios potenciales del cubo.

Hacer que la propiedad esté fragmentada en un ámbito de dimensiones y medidas es complicado, y requiere procedimientos complejos para gestionarlo.

¿Quién es dueño del almacén de datos? Si la titularidad tiene que pertenecer a un solo departamento, es una decisión un poco arriesgada. Por ejemplo, si establecemos como propietario el Departamento de IT (tecnologías de la información, en inglés *information tecnologies*), ellos solo son proveedores de servicios sin un conocimiento profundo y sin saber qué hacer exactamente con el contenido. Quizá la mejor opción es dar a la propiedad gestión estratégica o finanzas, para ser más precisos para el control. El Departamento de Control de Gestión sabe mejor qué tipo de productos se hará efectivo después del depósito de datos para los informes oficiales. Si esta propuesta no es suficientemente válida, quizá la mejor solución sería establecer un comité de expertos de todas las áreas de negocio, con la ventaja de controlar y dar la propiedad a este equipo.

Sin propiedad, no habrá desarrollo estratégico del sistema de almacén de datos. Solo funcionará para cubrir las necesidades actuales, sin ningún enfoque proactivo. La gestión de calidad de los datos será muy dudosa, y al final no habrá nadie para decidir cómo gestionar la gran cantidad de acciones necesarias.

Hay una fuerte necesidad de declarar oficialmente la propiedad de almacenamiento de datos y establecer los procedimientos de seguridad de datos tan pronto como sea posible.

3.2.2. Administrador

La administración de datos (*data stewardship*) es la gestión y supervisión de los activos de datos de una organización, para ayudar a proporcionar a los usuarios de negocio datos de alta calidad fácilmente accesibles y consistentes.

Dependiendo de la organización, el papel del administrador de datos puede estar definido formalmente o informalmente y reconocido por el negocio. Los administradores de datos suelen ser aquellas personas a las que todos se dirigen (dentro de su grupo de negocios) para tratar las consultas/problemas relacionados con los datos. Los compañeros de trabajo y los administradores de sistemas se acercan a los administradores de datos cuando lo que necesitan saber es qué datos se utilizan para responder a una pregunta de negocio, o si necesitan validar la precisión, integridad o validez de los datos dentro de un contexto de negocios.

En las organizaciones con programas de gestión de la información (*data governance*) bien definidos, la administración de datos puede ser un rol de trabajo formalmente establecido, pero en muchas organizaciones las tareas de administración de datos son una responsabilidad añadida para trabajadores de la información con otras funciones de trabajo.

De manera independiente de la formalidad de la función, los administradores de datos son los usuarios de negocio con los conocimientos técnicos de los procesos de negocio y cómo se utilizan los datos dentro de estos procesos.

Un administrador de datos es el principal responsable de:

- Identificación y adquisición de nuevas fuentes de datos.
- La creación y el mantenimiento de los datos de referencia consistentes y definiciones de datos maestros.
- La publicación de los datos pertinentes a los usuarios apropiados en una organización, y el seguimiento de las fuentes de datos publicados para la retroalimentación de uso/relevancia/calidad.

- Creación y gestión de metadatos comerciales para fuentes de datos publicados para asegurarse de que es fácilmente detectable, y significativa para trabajadores de la información.

- La resolución de problemas de integridad de datos a través de las partes interesadas.

- El análisis de los datos para los problemas de calidad de datos y conciliación.

Algunos autores distinguen las figura del guardián (*custodian*) de datos de forma separada y complementaria a la figura del administrador.

Su función principal es asegurar una buena continuidad y calidad de los datos. Mientras que el contenido es importante para ellos, su atención se centra en la infraestructura y las actividades necesarias para mantener intactos y a disposición de los usuarios los datos subyacentes. Ellos colaboran con los administradores de datos para implementar las transformaciones de datos, resolver los problemas de datos y colaborar en los cambios del sistema.

3.3. Usuarios según su relación con los datos

Antes de finalizar la caracterización de los usuarios de un sistema de *data warehouse*, encontramos la interesante clasificación y las definiciones que proponen W. W. Eckerson y C. Howson.

Nos proponen dividir a los usuarios en dos grandes grupos:

- Los productores de información: normalmente, se trata del 20 % de los usuarios y utilizan herramientas *desktop* para crear informes o modelos. Por lo general, se trata de estadísticos que utilizan herramientas de *data mining* o autores de informes que utilizan herramientas de diseño o de programación para crear informes específicos.
  Habitualmente, los autores de informes son técnicos de sistemas de información o usuarios de negocio avanzados que son capaces de entender la información y la informática. Los usuarios avanzados pueden crear o utilizar informes, por lo que están a medio camino entre los productores y los consumidores de información. Usualmente utilizan hojas de cálculo, herramientas de consulta e informes para acceder a la información y analizarla.

- Los consumidores de información: la mayoría de los consumidores de información son usuarios no habituales que regularmente consultan informes para la toma de decisiones, pero no acceden a los números o hacen análisis detallados diariamente. Los usuarios no habituales son directivos, gestores, responsables, colaboradores y usuarios externos.

Este numeroso grupo está bien servido con cuadros de mando con análisis guiados, informes interactivos (por ejemplo: OLAP, informes parametrizados, vinculados, etc.) e informes de gestión estandarizados. La mayoría de estas herramientas proveen ahora acceso vía web para promover el acceso desde cualquier lugar y facilitar el uso y minimizar los costes de administración y mantenimiento.

Es importante señalar que se diferencian los roles, no los individuos. Esto permite que un individuo tenga asignado más de un rol. Se deberán establecer las opciones que permitimos al usuario: usar los roles uno a uno; hacer uso simultáneo de todos los roles aplicando el criterio más restrictivo o el más permisivo.

4. Explotación y administración del sistema

4.1. Explotación de los datos

Una vez traspasados los datos al *data warehouse*, llega el momento de hacer uso de esta información para atender las necesidades estratégicas, tácticas y operativas que requiere nuestra organización.

Nos encontramos con una gran variedad de elementos de generación de valor en la inteligencia de negocio, que nos permiten presentar la información solicitada. Desde una perspectiva tradicional y consolidada a lo largo de las últimas décadas, destacamos:

- Informes y cuadros de mando.
- Análisis multidimensional (OLAP).
- Visualización de datos masivos.
- Análisis estadístico.
- Sistemas de información geográfica (GIS).
- Herramientas de minería de datos.
- Etc.

En estos últimos años, han aparecido una gran variedad de herramientas de visualización analítica que se caracterizan por:

- Permiten análisis interactivo, apoyándose en ágiles funcionalidades de visualización y gestión de datos, lo que facilita un análisis libre sobre el modelo de datos importado en la herramienta.
- La orientación de estas herramientas suele ser de autoservicio BI, lo que facilita la integración y el análisis de datos con poca intervención de IT.
- Las capacidades de visualización permiten, asimismo, llevar a cabo presentaciones claras y eficaces que ayuden en la toma de decisiones.
- Agilidad y rapidez en el manejo de datos, apoyándose en tecnologías *in-memory*.

- Estas herramientas se apoyan en una interfaz intuitiva que facilita la exploración de datos orientada tanto a perfiles TI como analistas de negocio.

Determinar cuál de estos instrumentos será el más adecuado para servir los datos es uno de los elementos más complicados que se deben atender. En el módulo «Explotación de datos» de esta asignatura, desarrollaremos ampliamente las posibilidades que se nos ofrecen y la idoneidad de todas ellas, según el propósito buscado.

4.2. Administración del sistema

Ya hemos comentado que un proyecto de almacén de datos supone el desarrollo y mantenimiento de un sistema informático por un largo periodo de tiempo, que acabará siendo un sistema crítico para la organización, con un coste económico importante que se deberá mantener bajo control.

Eckerson (2004) describe el ciclo de madurez de los *data warehouses* con las siguientes etapas: prenatal, infantil, niñez, adolescencia, adultez y madurez. En la prenatal, la empresa trabaja con reportes generados por la computadora central y que son codificados manualmente. En la etapa infantil, la empresa trabaja con hojas de cálculo que sirven para trabajar los datos y para almacenarlos y que son un *data mart* subrogado. En la etapa de la niñez la empresa trabaja con *data marts*, y luego, en la etapa de adolescencia, trabaja con un *data warehouse* que agrupa varios *data mart*. Ya en la etapa de adultez, trabaja con una *enterprise data warehouse*, que agrupa toda la información de la empresa en un solo punto, y utiliza *scorecards* para hacer un seguimiento del desempeño, y la empresa empieza a conseguir una gran cantidad de beneficios, incluyendo el retorno de la inversión.

Data mart

Es un subconjunto de los datos del *data warehouse* con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Al igual que en un *data warehouse*, los datos están estructurados en modelos de estrella o copo de nieve, y un *data mart* puede ser dependiente o independiente de un *data warehouse*.

En la última etapa, la de madurez, con los servicios de BI, la empresa se conecta con clientes y proveedores compartiendo los datos, sacando todas las ventajas, y además las conexiones se empiezan a hacer por Internet. En esta etapa también se utilizan los motores de decisiones, que automatizan muchas funciones.

4.2.1. La tecnología

El uso efectivo de los *data mart* en un entorno/contexto de *data warehousing* es un factor importante para la efectividad del almacén, y también puede ser determinante en el éxito del proyecto de desarrollo. Dado que los *data mart* son un factor crítico para el éxito del proyecto de *data warehousing* de mayor escala, también lo son su creación y mantenimiento.

Actualmente, las organizaciones se están convenciendo de que los *data warehouse* corporativos son complejos tanto de construir como de usar. Implementar un *data warehouse* requiere un considerable equipo de desarrolladores, hardware, software, tiempo y dinero.

Incorporaremos en este apartado el marco de referencia teórico que vamos a seguir al diseñar, implantar y gestionar nuestro almacén de datos. También aquellas infraestructuras y equipos necesarios para el funcionamiento del sistema. Y por último, aquellos programas que nos permiten manipular, almacenar y servir nuestros datos.

4.2.2. Entorno

Los sistemas de *data warehouse*, al igual que muchos otros sistemas TI, están fuertemente influenciados por el entorno en el cual están operativos.

Los principales elementos que les pueden afectar y que deberemos gestionar adecuadamente son:

- Los posibles cambios que nos hagan los usuarios (*demand*).
- Cambios externos en las fuentes de datos o marcos legislativos.
- Mantenimientos en la tecnología, en los tratamientos de datos, etc.
- Riesgos inherentes de todo proyecto tecnológico.
- Incidencias que puedan surgir y que deberán ser atendidas lo antes posible para no perder valor.

4.2.3. Arquitectura

Principalmente, encontramos tres enfoques en la arquitectura corporativa de un *data warehouse*:

- *Enterprise bus architecture* (o *data warehouse* virtual / federado): también conocido como MD (*multidimensional architecture*), consiste en una arquitectura basada en *data marts* independientes federados que pueden hacer uso

de una *staging area* en el caso de que sea necesario. Federados quiere decir que se hace uso de una herramienta EII (*enterprise information integration*) para llevar a cabo las consultas como si se tratara de un único *data warehouse*. Puede existir en el caso de ser necesario un ODS (*operational data store*).

- *Corporate information factory* (o *enterprise data warehouse*): consiste en una arquitectura en la que existe un *data warehouse* corporativo y unos *data marts* (o incluso cubos OLAP) dependientes del mismo. El acceso a datos se hace a los *data marts* o a la ODS en caso de existir, pero nunca al propio *data warehouse*. Puede existir en el caso de ser necesaria una *staging area*.

- *Enterprise data warehouse* 2.0: consiste en la revisión de la metodología de Bill Inmon para incluir toda la experiencia de los últimos veinte años. El punto diferencial es que se separa la información por la edad de la misma, y la clasifica por su uso. Se caracteriza por completar tanto la inclusión de información estructurada como no estructurada, y por focalizarse en tener el objetivo de responder a todas las necesidades actuales de negocio. Es una propuesta para evitar que la factoría de información crezca de manera desordenada.

Staging area

Es el sistema que permanece entre las fuentes de datos y el *data warehouse* con el objetivo de facilitar la extracción de datos, ser usado como caché de datos operacionales y mejorar la calidad de datos.

Operational data store

Es un tipo de almacén de datos que proporciona solo los últimos valores de los datos y no su historial, y generalmente es admisible un pequeño desfase o retraso sobre los datos operacionales.

4.2.4. Tareas administrativas

En esta sección, se proporciona información acerca de la administración y el mantenimiento óptimo del almacén de datos corporativo. En la misma, se enumeran las tareas más comunes que deberán llevarse a cabo.

Las tareas relacionadas con la gestión del almacén de datos se pueden dividir en las siguientes categorías:

a) Tareas de preparación

- Programar y gestionar procesos.

- Configurar autorizaciones.

- Configurar controles de consistencia de metadatos y datos.

b) Tareas llevadas a cabo regularmente

- Supervisar el rendimiento de los procesos.
- Supervisar el estado de los procesos y solicitudes.
- Supervisar el rendimiento de las aplicaciones.
- Mostrar las estadísticas de uso.
- Supervisar la carga de base de datos.
- Supervisar las cadenas de procesos periódicos.
- Monitorizar los archivos de registro.

c) Tareas a demanda

- Ajustar los procesos y objetos.
- Ajustar las cadenas de procesos.
- Configurar las autorizaciones.
- Modificar los agregados para reflejar los cambios en las jerarquías y atributos.
- Cambiar los parámetros del sistema.
- Ajustar parámetros relacionados con el rendimiento.
- Ajustar los parámetros de caché OLAP.

d) Análisis de errores

- Hacer comprobaciones de coherencia de datos y metadatos en el sistema.
- Prueba, ejecutar y administrar consultas y vistas de consulta.

Glosario

almacén de datos *m* Bases de datos orientadas a áreas de interés de la empresa que integran datos de distintas fuentes con información histórica y no volátil y que tienen como objetivo principal hacer de apoyo en la toma de decisiones. Puede ser corporativo o departamental.
en data warehouse

almacén de datos operacional *m* Conjunto de datos integrado y orientado al tema, pero sin datos históricos. Se suele utilizar como paso intermedio en la construcción del almacén de datos corporativo.
en operational data store

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

data governance *m* Véase gobierno del dato.

data mart *m* Subconjunto de los datos del *data warehouse* con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica.

data mining *m* Véase **minería de datos**.

data warehouse *m* Véase **almacén de datos**.

dato *m* Medida, observación hecha y almacenada en algún sistema (definición desde el punto de vista de los sistemas decisionales).

factoría de información corporativa *f* Conjunto de elementos de software y hardware que ayudan al análisis de datos para tomar decisiones.
sigla **FIC** (*Corporate Information Factory*)

FIC *f* Véase **factoría de información corporativa**.

gestión de datos maestros *f* Metodología que identifica la información más crítica de una organización y crea una única fuente fiable.
en master data management

gobierno del dato *m* Metodología que tiene por objeto asegurarse de que los datos son siempre fiables y válidos en cada contexto empresarial, de que su calidad se mantiene a lo largo del tiempo y de que existen mecanismos de control sobre quién puede hacer qué con los datos en cada momento.
en data governance

master data management *m* Véase **gestión de datos maestros**.

metadato *m* Datos sobre datos, que representan características de otros datos que facilitan su administración y uso.

minería de datos *f* consiste en extraer información de alto valor añadido (patrones ocultos, tendencias y correlaciones) a partir de datos en bruto.
en data mining

OLAP *f* Siglas que hacen referencia a las herramientas de análisis, normalmente multidimensional (*online analytical processing*).

OLTP *m* *Online transactional processing*

operational data store *m* Véase **almacén de datos operacional**.

SGBD *f* Véase **sistema de gestión de bases de datos**.

sistema de gestión de bases de datos *m* Software que gestiona y controla bases de datos. Sus funciones principales son las de facilitar su uso simultáneo a muchos usuarios de distintos tipos, independizar al usuario del mundo físico y mantener la integridad de los datos.
sigla **SGBD**

staging area *f* Sistema que permanece entre las fuentes de datos y el *data warehouse*, con el objetivo de facilitar la extracción de datos, ser usado como caché de datos operacionales y mejorar la calidad de datos.

Bibliografía

Inmon, W. H.; Strauss, D.; Neushloss, G. (2010). *DW 2.0:The Architecture for the Next Generation of Data Warehousing*. Morgan Kaufman Series in Data Management Systems.

Inmon, W. H.; Linstedt, D. (2014). *Data Architecture: a primer for the Data Scientist.* Burlington: Morgan Kaufman Series.

Kord, D.; Patterson, D. (2012). *Ethic of Big Data, Balancing Risk and Innovation.* O'Reilly Media.

Ladley, J. (2011). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*. The Morgan Kaufmann Series on Business Intelligence.