

Caso práctico

El nuevo clúster ya está funcionando: 1024 nodos, cada uno con procesador de 32 núcleos y 64 GB de memoria RAM.

Su HDFS está ya listo para recibir los 2,5 petabytes que residen las distintas bases de datos y ficheros de texto plano con los que va a ser alimentado.

Hasta ahora nuestros protagonistas lo han vivido con la ilusión de unos niños que han recibido los regalos que más deseaban. Sin embargo tras comprobar que todo el hardware está preparado se ha hecho un silencio en la habitación.

—¿Y ahora? ¿Qué hacemos? —dice **Carlos**.

—Pues pasar la información al HDFS del clúster, ¿no? —responde **Sara**.

—Sí, estupendo. Y repito mi pregunta. ¿Qué hacemos?



[Brett Sayles](#) (Dominio público)

En esta unidad de trabajo vamos a tratar lo referente a la gestión de datos en entornos de Big Data.

Comenzamos viendo el proceso de ETL, el cual constituye la puerta de entrada más generalmente empleada a la hora de nutrir con datos nuestros sistemas. Aquí estudiaremos tanto sus 3 fases como dos de las herramientas más empleadas como parte del proceso.

A continuación descubriremos la integración de datos, poniendo énfasis en la distinción entre ésta y el mencionado proceso de ETL, ya que están en cierto modo relacionados.

Más adelante daremos un repaso a la normativa de tratamiento de datos, enumerando las distintas leyes vigentes y realizando un análisis algo más detallado del RGPD (a nivel europeo y de aplicación a partir de 2018).

Por último trataremos el concepto de Gobierno de Datos, el cual constituye una supervisión de toda la gestión realizada sobre los datos.



[Ministerio de Educación y Formación Profesional](#) (Dominio público)

Materiales formativos de FP Online propiedad del Ministerio de Educación y Formación Profesional.

[Aviso Legal](#)

Caso práctico



[Lukas](#) (Dominio público)

El ambiente en la reunión puede cortarse con un cuchillo.

El equipo de informáticos ha estado trabajando durante un mes para producir una serie de informes que los directivos de la compañía pudiesen usar hoy para tomar importantes decisiones de futuro, pero todo ha fallado estrepitosamente.

En primer lugar, los resultados para el informe iban a estar preparados por la mañana y sin

embargo ha habido que retrasar la reunión a la tarde esperando a que terminasen de calcularse.

En segundo lugar, con el informe en la mano cualquiera puede ver que los datos no se han interpretado bien porque los resultados son incoherentes con la realidad que muy bien conocen los directivos.

Y por último, no dejan de llegar quejas de clientes porque el tiempo de espera para sus servicios online durante toda la mañana ha sido excesivo.

Tras escuchar varias quejas destinadas a los informáticos, **Jorge**, el consultor tecnológico (al que, por cierto, nadie preguntó antes de comenzar a elaborar los informes) decide tomar la palabra.

—Señores, se ha hecho todo mal —dice Jorge.

—Eso ya lo hemos visto, Jorge —replica uno de los directivos más mayores—. ¿Sabes tú cuál es el problema? Ilumínanos, anda.

—Pues veréis. Se han hecho las analíticas contra las tablas del ERP, que es la parte transaccional de la empresa. Por eso se ha bloqueado y se ha dado un pobre servicio a los clientes esta mañana. Aparte, esas tablas son relacionales y por lo tanto no están pensadas para ser el origen de analíticas. Por eso las analíticas han tardado tanto. Y por último, los datos que hay en el ERP no se han interpretado correctamente y por eso los resultados son a todas luces incorrectos.

—¿Y entonces cómo lo hacemos?

—Tenéis que montar un almacén de datos que funcionará de forma independiente del ERP y será sólo para ejecutar analíticas, las cuales así se producirán más rápido porque esos almacenes de datos están específicamente diseñados para ello. Y al ser independiente de los sistemas transaccionales, la parte transaccional no se resentirá cada vez que haya que realizar analítica.

—Pero los resultados seguirán siendo incorrectos.

—Son incorrectos posiblemente porque el significado de algunos datos según están en el ERP no es el que los de la analítica suponían. Eso se soluciona haciendo un buen trabajo al cargar datos en el almacén. Ese proceso se llama ETL y la "T" es por "transformar", lo que incluye el tener mucho cuidado de poner los datos en el almacén según el significado que están esperando los de analítica.

ETL ("Extraer, Transformar, Cargar", del inglés "Extract, Transform, Load"), es el proceso que de forma necesaria es necesario realizar si tenemos datos provenientes de una o varias fuentes y queremos enviarlos de forma unificada a un almacenamiento de destino. Como veremos más adelante, primero es necesario adquirirlos desde sus fuentes (extraer), después generalmente es necesario transformarlos de algún modo (transformar) y por último hay que realizar su carga en el ya mencionado almacén de datos.

En muchas ocasiones ETL es confundido con otro proceso o mecanismo en cierto modo relacionado que recibe el nombre de *integración de datos*.

Veremos más adelante lo que significa integración de datos, pero vamos a aprovechar ahora para aclarar sus diferencias.

ETL	Integración de datos
Extraer desde fuentes, transformar y cargar datos.	Ofrecer una visión unificada de datos que residen en distintas fuentes.
Se produce una copia (modificada) de datos.	En su uso ideal no se realiza ninguna copia sino que los datos siguen residiendo en las fuentes originales.

Es importante destacar que las fuentes de origen a emplear en el proceso de ETL no tienen por qué contener únicamente datos recientes sino que en ocasiones también pueden ser históricos. Por ejemplo, los procesos ETL también se utilizan para integraciones con o migraciones desde sistemas heredados. Un posible caso sería migrar los datos de un ERP preexistente a uno nuevo.

Históricamente los destinos más típicos de los procesos ETL han sido los almacenes de datos, sobre todo para un posterior uso OLAP. De este modo se descargan los sistemas OLTP de toda carga analítica y ésta puede realizarse sobre una instantánea de los datos almacenada en un formato directamente preparado para el trabajo analítico.

Sin embargo, con la aparición de muchos otros usos de los datos y sistemas de almacenaje de los mismos, los procesos ETL han comenzado a tener otros destinos. Por ejemplo en ambientes Big Data es muy común que tal destino sea un sistema de ficheros distribuido, como HDFS o Amazon S3.

Autoevaluación

¿Cuál de las siguientes afirmaciones es correcta en relación a ETL e integración de datos?

- Son equivalentes en todo caso.
- El proceso es el mismo pero en integración de datos el destino es un almacén de datos.
- Integración de datos es uno de los pasos que se realizan dentro del proceso de ETL.
- No son equivalentes.

Incorrecto. No son equivalentes.

Incorrecto. La integración de datos no tiene como destino típico un almacén de datos.

Incorrecto. Integración de datos no forma parte de ETL.

Correcto.

Solución

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta

1.1.- Fases de ETL.

A continuación vamos a ver cuáles son las 3 fases involucradas en el proceso de ETL.

Veremos aquí un esquema/resumen para que puedas tener una vista general:

- ✓ **Extraer (*extract*):**
Fase en la que se extraen los datos desde diversas fuentes, incluyendo una validación previa de los mismos.
- ✓ **Transformar (*transform*):**
Fase en la que se realiza todo el proceso de transformación de los datos para dejarlos en el formato finalmente deseado. Entre otras cosas incluye limpiar datos y crear datos sintéticos como resultado de otros.
- ✓ **Cargar (*load*):**
Fase en la que finalmente se cargan los datos en el destino final, el cual como hemos dicho suele ser o bien un almacén de datos o un sistema de ficheros distribuido.

Autoevaluación

La fase de extracción del proceso ETL consiste en extraer los metadatos incluidos dentro de los ficheros.

Verdadero Falso

Falso

Es la fase en la que se extraen los datos desde diversas fuentes, incluyendo una validación previa de los mismos.

La fase de transformación es en la que se realiza todo el proceso de transformación de los datos para dejarlos en el formato finalmente deseado.

Verdadero Falso

Verdadero

Cierto.

En la fase de carga de ETL es cuando se cargan datos desde disco a memoria.

Verdadero Falso

Falso

Es la fase en la que finalmente se cargan los datos en el destino final.

1.1.1.- Extraer.

La primera parte del proceso de ETL es la extracción de datos desde las diversas fuentes en las que estos puedan encontrarse o incluso provenir en forma de flujo (*stream*).

Esta primera parte del proceso es fundamental, ya que de su correcta ejecución depende por completo el éxito del resultado final.

Incluye tanto el acceso a fuentes de datos de entrada como una primera fase de chequeo de los datos para darlos por válidos o en su defecto rechazarlos.

Tales fuentes de datos pueden ser de tipo muy variado y con formatos heterogéneos.

- ✓ Bases de datos relacionales.
- ✓ Almacenes de datos.
- ✓ Ficheros en diversos formatos.
- ✓ Transacciones que se van produciendo sobre la marcha.
- ✓ Logs de servidores web.
- ✓ Resultados obtenidos mediante [arañas web](#).
- ✓ Mediciones producidas por dispositivos IoT.
- ✓ ...

Una de las mayores dificultades de esta fase viene precisamente por la gran variedad de fuentes de datos que podemos encontrar, y la necesidad para integrarse con ellas empleando sus mismos formatos y protocolos. Para los tipos de fuentes más típicas lo común será que la herramienta ETL a utilizar ya sea compatible. Sin embargo en otras ocasiones habrá que realizar algún tipo de desarrollo específico para realizar tal integración. Algunas opciones para ello pueden ser:

- 1.- Conseguir que el sistema en el que está la fuente deje los datos en un formato al que podamos acceder (por ejemplo en ficheros o en una base de datos auxiliar a tal efecto).
- 2.- Programar un intermediario o conector capaz de adquirir los datos según el protocolo y formato en el que la fuente los entrega.

Hay que tener en cuenta que el proceso de extracción no debe afectar al funcionamiento de los sistemas de los cuales vienen los datos, sobre todo en el caso de los que son transaccionales.

Por ello será importante escoger bien la estrategia a la hora de recibir datos actualizados en el almacén de datos. Puede ser deseable que el almacén de datos esté actualizado a tiempo real según las últimas transacciones. Sin embargo, si eso va a ralentizar el funcionamiento de éstas entonces quizás sea mejor permitir un cierto desfase y que esos datos se obtengan no en tiempo real sino en momentos en los que el sistema transaccional esté menos ocupado (por ejemplo de noche).

Como resultado de esta fase, los datos quedan en un formato válido para que comience la subsiguiente fase de transformación.

Autoevaluación

¿En qué momento es mejor realizar la fase de extracción en un proceso de ETL?

- Cualquier momento es bueno.
- Siempre por la noche.
- Si va interferir con el correcto funcionamiento de los sistemas que contienen las fuentes se realizará en el momento en el que tales sistemas o fuentes se encuentren menos cargados.
- Siempre por la mañana.

Incorrecto. Puede haber momentos mejores que otros según las circunstancias.

Incorrecto. En muchas ocasiones puede ser por la noche pero no siempre.

Correcto.

Incorrecto. La mañana no suele ser el momento más indicado.

Solución

1. Incorrecto
2. Incorrecto
3. Opción correcta
4. Incorrecto

1.1.2.- Transformar.

En la fase de transformación se emplea una serie de reglas sobre los datos extraídos con la intención de prepararlos para ser finalmente cargados en el destino correspondiente.

Durante esta fase nos aseguramos de que al destino sólo van a llegar datos realmente válidos, y que estos estarán en el formato deseado.

Para ello puede ser necesario realizar diversos tipos de transformación, entre los cuales se encuentran los siguientes:

- ✓ Seleccionar sólo determinadas columnas o atributos de cada registro.
- ✓ Seleccionar sólo determinados registros según si cumplen o no determinada condición.
- ✓ Transponer filas en columnas o columnas en filas.
- ✓ Eliminar datos que resulten estar duplicados.
- ✓ Cuando un dato no esté presente, intentar darle valor si puede deducirse a partir del resto de los datos.
- ✓ Ordenar los datos según los valores de ciertas columnas si de ese modo se acelera su uso en destino.
- ✓ Traducir valores codificados según la nomenclatura o codificación del origen a la que tendrán en destino (por ejemplo traducir de "Válido" a 1).
- ✓ Crear nuevos valores derivados de otros (por ejemplo $area = lado1 * lado2$).
- ✓ Dividir un atributo en varios (por ejemplo una fecha en año, mes y día).
- ✓ Realizar unificaciones de datos que provienen de distintas fuentes.
- ✓ Calcular datos agregados (por ejemplo suma de totales o cálculo de medias).
- ✓ Comprobar que los registros referenciados por claves externas realmente existen, y quizás realizar algún tipo de denormalización.

La fase de transformación suele ser por lo general la que mayor esfuerzo humano requiere (de comunicación y comprensión), ya que se necesita comprender a la perfección el significado que tienen los datos en las fuentes de origen y el que deben tener en destino. Por ello los expertos encargados del proceso de ETL deben realizar en primer lugar un diseño detallado de las distintas reglas que se van a ir aplicando y en qué orden.

Autoevaluación

¿Cuál de las siguientes no es un posible tipo de transformación a realizar con los datos en un proceso de ETL?

- Realizar unificaciones de datos que vienen de distintas fuentes.
- Eliminar el contenido de una base de datos.
- Transponer filas en columnas o columnas en filas.
- Seleccionar sólo determinadas columnas o atributos de cada registro.

Incorrecto. Esa sí es una de las posibles transformaciones.

Correcto.

Incorrecto. Esa sí es una de las posibles transformaciones.

Incorrecto. Esa sí es una de las posibles transformaciones.

Solución

1. Incorrecto
2. Opción correcta
3. Incorrecto
4. Incorrecto

1.1.3.- Cargar.

La fase de carga del proceso de ETL es con diferencia la más sencilla y liviana, ya que es la única en la que ya podemos decir que la situación está desde un principio bajo control. Ello se debe a que gracias a la fase de transformación los datos ya están en el formato deseado y a que el almacén de destino para los datos es propio. Dado que es propio lo habremos escogido según las necesidades y no tendremos ningún problema para conectarnos a él como sí podía ocurrir con las fuentes en la fase de extracción.

Los destinos de datos pueden ser muy variados:

- ✓ Almacenes de datos para uso OLAP.
- ✓ Ficheros planos.
- ✓ Sistemas OLTP (por ejemplo si migramos de un ERP antiguo a otro nuevo).
- ✓ Sistemas de ficheros distribuidos como HDFS o Amazon S3.

Algo muy a tener en cuenta al cargar los datos en destino es decidir cómo interactúan los datos nuevos con los que ya haya almacenados, lo cual dependerá del uso que se les quiera dar y de los requisitos de gobernanza de datos acerca de datos históricos.

Algunas opciones en este sentido son las siguientes:

- 1.- Cargar los datos completos cada vez, eliminando previamente los datos antiguos.
- 2.- Cargar sólo datos nuevos que no afectan a los que ya hay cargados.
- 3.- Cargar datos de forma incremental, lo cual puede implicar la actualización de los que ya hay cargados.

Al margen de todo ello, y también según los requisitos de gobernanza de datos, el almacenaje de destino puede incluir algún mecanismo que le permita guardar un registro histórico de los estados por los que va pasando cada dato, lo cual puede ser muy útil a la hora de auditar la información.

Cuando hay un mecanismo de este tipo, tras actualizar un dato veremos la última versión del mismo, pero podremos acceder al histórico del mismo para comprobar los distintos valores que ha ido teniendo a lo largo del tiempo. Ello permite saber qué valores ha tenido el dato, en qué fecha y hora, e incluso con qué procedencia (usuario, fuente o proceso).

Autoevaluación

¿Qué les ocurre a los datos que ya hay en el almacenamiento destino de un ETL cuando llegan datos nuevos?

- Se eliminan.
- Se actualizan.
- Depende el uso. En ocasiones se eliminan, en ocasiones no quedan afectados y en ocasiones son modificados.
-

Depende el uso. En ocasiones se eliminan y en ocasiones son modificados.

Incorrecto. No siempre se eliminan.

Incorrecto. No siempre se actualizan.

Correcto.

Incorrecto. Puede haber ocasiones en las que ni se eliminen ni se modifiquen.

Solución

1. Incorrecto
2. Incorrecto
3. Opción correcta
4. Incorrecto

1.2.- Herramientas.

A continuación vamos a ver 2 de las herramientas de uso más común en procesos ETL cuando nos encontramos en ambientes de Big Data.

Veremos aquí un esquema/resumen para que puedas tener una vista general:

✓ **Apache Sqoop:** *SQL-to-Hadoop*.

Es una herramienta de línea de comandos que nos permite obtener datos desde bases de datos relacionales para transferirlos generalmente a sistemas de ficheros distribuidos.

✓ **Apache Flume:**

Es un software distribuido que permite obtener datos en *streaming* desde gran cantidad de fuentes no estructuradas o semiestructuradas para transferirlos generalmente a sistemas de ficheros distribuidos.

Ambas forman parte del ecosistema de Hadoop, por lo que fueron en un principio diseñadas para enviar datos a HDFS. Sin embargo, a día de hoy son compatibles con otros almacenamientos distribuidos, como por ejemplo Amazon S3.

Autoevaluación

Apache Sqoop es un software distribuido que permite obtener datos en streaming desde gran cantidad de fuentes no estructuradas o semiestructuradas.

Verdadero Falso

Falso

Ese es Apache Flume.

Apache Sqoop es una herramienta de línea de comandos que nos permite obtener datos desde bases de datos relacionales.

Verdadero Falso

Verdadero

Cierto.

Apache Sqoop permite enviar datos a HDFS.

Verdadero Falso

Verdadero

Cierto.

Apache Flume permite enviar datos a HDFS.

Verdadero Falso

Verdadero

Cierto.

1.2.1.- Apache Sqoop.

Apache Sqoop (cuyo nombre proviene de *SQL-to-Hadoop*) es una herramienta de línea de comandos que forma parte del ecosistema Hadoop y que está diseñada para permitirnos obtener datos desde bases de datos relacionales.

Siendo parte de Hadoop, fue en un principio diseñada para enviar los datos resultantes a HDFS (el sistema de ficheros distribuido de Hadoop). Sin embargo con el tiempo ha ido adquiriendo compatibilidad con otros sistemas muy utilizados en el mundo Big Data, como por ejemplo S3 (la solución en la nube de Amazon).

Sqoop es una herramienta ideal si queremos obtener datos que se encuentran en bases de datos relacionales (Oracle, SQL Server, MySQL, Teradata, Postgres,). Puede conectarse con cualquier tipo de base de datos que tenga conectividad [JDBC](#).

Algunas de las principales características de Apache Sqoop son las siguientes:

- ✓ Permite importaciones en masa (*bulk*), siendo capaz de obtener tablas individuales o incluso bases de datos completas.
- ✓ Paraleliza la transferencia de datos para conseguir un alto rendimiento de lectura desde las fuentes y un uso óptimo del sistema.
- ✓ Cuenta con mecanismos para evitar sobrecargar las fuentes.
- ✓ Permite realizar mapeados directos de bases de datos relacionales hacia otras herramientas del ecosistema Hadoop, como HBase o Hive.
- ✓ Cuenta con interfaz de línea de comandos.
- ✓ También permite acceso programático mediante JDBC.
- ✓ Aunque no era la intención inicial, también puede conectarse a bases de datos NoSQL como MongoDB o Cassandra.

Para saber más

Puedes ver más información acerca de Apache Sqoop en los siguientes enlaces.

- ✓ Empieza por aquí: [Sqoop en la Wikipedia](#).
- ✓ A continuación te recomendamos que veas esta [visión general de Sqoop](#) (en inglés).
- ✓ También puedes acceder a la [web oficial de Apache Sqoop](#) (en inglés), dentro de la cual puedes encontrar toda su documentación.
 - ➔ Dentro de la web oficial, te recomendamos que entres en la [documentación de Sqoop](#).
 - ➔ Para ver cómo trabajar con Sqoop desde línea de comandos entra en esta [demostración en 5 minutos](#).

Autoevaluación

¿Cuáles de las siguientes son características propias de Apache Sqoop?

- Se puede emplear mediante interfaz de línea de comandos.

- Obtiene datos desde fuentes no estructuradas.

- Tiene mecanismos para evitar sobrecargar las fuentes.

- Permite acceso programático mediante JDBC.

Mostrar retroalimentación

Solución

1. Correcto
2. Incorrecto
3. Correcto
4. Correcto

1.2.2.- Apache Flume.

Apache Flume es un software distribuido que permite obtener y agregar datos en streaming desde gran cantidad de fuentes no estructuradas o semiestructuradas para transferirlos generalmente a sistemas de ficheros distribuidos.

En un principio la idea era que ayudase fundamentalmente a encaminar hacia HDFS ficheros de log (por ejemplo de servidores web), ya que constituyen una fuente muy típica de entrada a Hadoop. Sin embargo en la actualidad permite conectar con muchos otros tipos de fuente, principalmente en *streaming* y basadas en eventos.

Al igual que ocurre con Sqoop, al ser parte de Hadoop fue en un principio diseñado para enviar los datos resultantes a HDFS (el sistema de ficheros distribuido de Hadoop). Sin embargo con el tiempo ha ido adquiriendo compatibilidad con otros sistemas muy utilizados en el mundo Big Data, como por ejemplo Amazon S3.

Se basa en una arquitectura flexible basada en el encaminamiento de flujos de datos a través de sus 3 tipos de componentes (*Source*, *Channel* y *Sink*).

Algunas de las principales características de Apache Flume son las siguientes:

- ✓ Está basado en eventos y se adapta a fuentes en *streaming*.
- ✓ Permite adquirir flujos de datos desde múltiples canales de entrada de forma simultánea.
- ✓ Permite crear topologías para tratar los flujos de datos hasta llegar al resultado final.
- ✓ Diseñado para alto ancho de banda y baja latencia.
- ✓ Es tolerante a fallos (por ejemplo a errores producidos en las fuentes) e incluye mecanismos de recuperación.
- ✓ Permite escalar de un modo casi lineal añadiendo componentes a la topología.

Para saber más

Puedes ver más información acerca de Apache Flume en los siguientes enlaces.

- ✓ Empieza por aquí: [Flume en la Wikipedia](#).
- ✓ También puedes acceder a la [web oficial de Apache Flume](#) (en inglés), dentro de la cual puedes encontrar toda su documentación.
 - ➔ Dentro de la web oficial, te recomendamos que entres en la [documentación de Flume](#).
 - ➔ En esa documentación es interesante que veas la [guía del usuario de Flume](#).

Autoevaluación

¿Cuáles de las siguientes son características propias de Apache Flume?

- Permite conectar con cualquier fuente JDBC.

- Alto ancho de banda y baja latencia.

- Tolerancia a fallos.

- Basado en eventos.

Mostrar retroalimentación

Solución

1. Incorrecto
2. Correcto
3. Correcto
4. Correcto

2.- Integración de Datos.

Caso práctico

En **Arendis Consultores**, la empresa de **Fernando, Mónica y Rodrigo**, se avecinan muchos cambios.

Crearon una empresa de la nada hace 26 años y fueron adaptándose a los nuevos tiempos sobre la marcha. Ahora sienten que quizás deberían haber realizado algún tipo de planificación más meditada, porque cuentan con diversas bases de datos (para diversas ubicaciones, usos y clientes), y cada vez les resulta más complicado saber cómo acceder a aquello que necesitan en cada momento.



[panumas nikhomkhaj](#) (Dominio público)

Además, muchas de esas bases de datos se crearon en su momento mediante subcontratación con terceras empresas de desarrollo informático, por lo que cada una utiliza tanto distinto motor como un distinto modo de entender el significado de cada dato.

Ellos lo que querrían es poder unificarlo todo en una única base de datos central, pero esas bases de datos preexistentes se encuentran en producción, dando servicio a programas de diversos clientes que se conectan a ellas constantemente. No es una opción pararlas durante el tiempo necesario para la unificación.

Hoy Fernando, Mónica y Rodrigo están de enhorabuena. Tras una larga investigación creen haber encontrado con la solución.

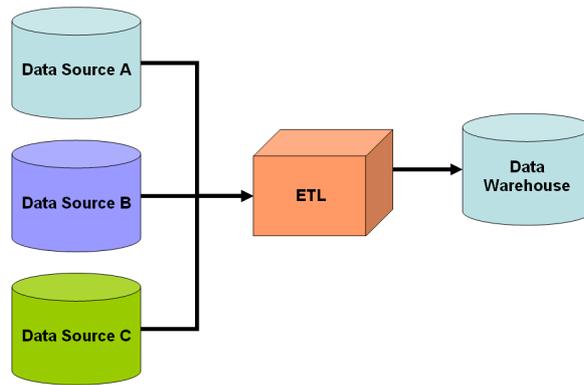
Y esa solución no es otra que una técnica conocida como **Integración de Datos**.

La Integración de Datos tiene objetivos similitudes a los que se persiguen en los proceso ETL, y por ello ambos términos suelen confundirse.

Por lo tanto, debemos dejar clara la diferencia desde un primer momento. Mientras que el proceso de ETL busca llevar los datos desde un origen a un destino, el procedimiento de integración busca entregar una visión unificada de los datos a aquellos usuarios o procesos que vayan a utilizarlos (lo que, como veremos, no sólo puede conseguirse mediante procedimiento de copiado).

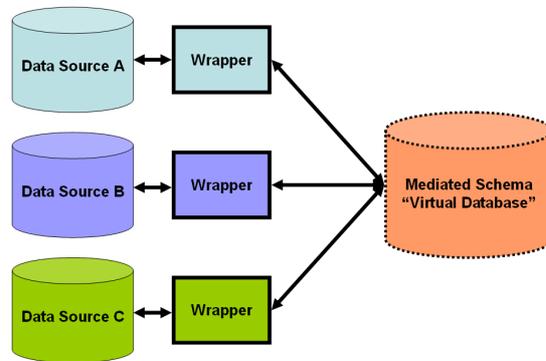
Podemos ilustrar esa diferencia con dos figuras, una correspondiente a un proceso ETL y otra correspondiente a una integración de datos:

Figura 1: Proceso de ETL



[Wikimedia Commons](#) (Dominio público)

Figura 2: Proceso de Integración de Datos



[Wikimedia Commons](#) (Dominio público)

Como podemos observar, en el caso del proceso de Integración de Datos, se diseña una base de datos virtual, la cual cuenta con un esquema mediado (aquel que define cómo queremos ver los datos).

Sin embargo, cuando realizamos una consulta el sistema emplea envoltorios de intermediación (*wrappers*) para acceder realmente a las bases de datos de origen, ya que es en ellas donde residen realmente los datos (siendo la base de datos virtual una mera construcción conceptual).

Para saber más

Puedes ver más información sobre Integración de datos en el siguiente enlace a artículo de la wikipedia. En el momento de crear estos contenidos no existía en castellano, de modo que te lo dejamos en inglés.

Autoevaluación

El proceso de Integración de Datos:

- Consiste en una replicación modificada de los datos.
- Busca entregar una visión unificada de los datos a aquellos usuarios o procesos que vayan a utilizarlos.
- Es un paso previo al proceso de ETL.
- Es una de las actividades englobadas dentro del proceso de ETL.

Incorrecto. Idealmente se realiza sin replicación de los datos.

Correcto.

Incorrecto. Son independientes.

Incorrecto. Existe de forma independiente a los procesos de ETL.

Solución

1. Incorrecto
2. Opción correcta
3. Incorrecto
4. Incorrecto

2.1.- Técnicas de integración de datos.

Hemos visto cómo debería ser una integración de datos en teoría.

Sin embargo, en la práctica se utilizan distintas técnicas para conseguir esos resultados (o similares), las cuales pueden enumerarse según distintos niveles de automatización de la integración (más automatizada cuanto más bajamos en el listado).

✓ Integración manual:

- Los usuarios operan con los datos accediendo directamente a los sistemas de origen.
- No existe una visión unificada de los datos.

✓ Integración basada en aplicación:

- Es una aplicación la que realiza toda la integración, accediendo a los sistemas de origen.
- Se producen resultados según lo que permita el interfaz de usuario de la aplicación y/o sus capacidades de volcado.

✓ Integración basada en middleware:

- La lógica de la integración en este caso no está en una aplicación sino en una capa de *middleware*, la cual facilita datos con algún tipo de transformación a las aplicaciones que a él se conectan.
- Por lo general ello implica que las aplicaciones aún tienen que participar en la integración. Por ejemplo si una aplicación recibe datos desde más de un *middleware* aún tendrá que fundirlos de algún modo para producir el resultado final.

✓ Integración virtual (o acceso uniforme a los datos):

- Deja los datos en los orígenes y permite acceder a ellos según una vista unificada de los mismos.
- Según tal visión unificada, accedemos a una base de datos virtual y las consultas se encaminan a los orígenes mediante envoltorios de intermediación (*wrappers*).
- Ventaja: no hay latencia en la vista final cuando un dato se añade o modifica en el origen.
- Inconveniente: carga los orígenes cada vez que se hace una consulta.
- Inconveniente: se pierde la capacidad de realizar algún tipo de gestión de histórico o de versiones de los datos al no emplear almacenamiento propio.

Autoevaluación

¿Cuál de los siguientes niveles de integración de datos es el que nos permite acceder a los datos mediante una visión unificada de los mismos?

- Integración basada en aplicación.
- Integración virtual.
- Integración basada en *middleware*.

Integración manual.

Incorrecto. En este caso no hay una verdadera visión unificada sino una aplicación intermediando.

Correcto.

Incorrecto. En este caso no hay una verdadera visión unificada sino un *middleware* intermediando.

Incorrecto. En este caso no existe ningún tipo de visión unificada de los datos.

Solución

1. Incorrecto
2. Opción correcta
3. Incorrecto
4. Incorrecto

3.- Normativa de tratamiento de datos.

Caso práctico



[stux](#) (Dominio público)

Hace tres años la vida de **Amalia** se convirtió en un pequeño infierno.

El historial médico que una clínica privada guardaba acerca de ella fue a parar a otras manos, poniendo a la vista diversas patologías que deberían haberse guardado en secreto.

Como resultado, comenzó a recibir diversas acciones comerciales tratando de venderle fármacos y tratamientos

relacionados con sus dolencias. En un principio Amalia pensó que debía de tratarse de una casualidad, pero pronto las casualidades comenzaron a parecerle demasiado grandes. Sencillamente no podía ser. Esas empresas necesariamente tenían que estar al corriente sobre sus problemas de salud. Era la única explicación posible.

Sin embargo lo peor no fue el acoso comercial sino que decidió hacerse un seguro de vida y las compañías aseguradoras o bien no querían asegurarla o le proponían una cuota desorbitada.

Hoy Amalia vuelve a sonreír. La ley está de su lado, y gracias a ella ha conseguido una indemnización por daños y perjuicios así como el borrado de sus datos sensibles en diversos ficheros.

Citas Para Pensar

1.- Se garantiza el derecho al honor, a la intimidad personal y familiar y a la propia imagen.

2.- El domicilio es inviolable. Ninguna entrada o registro podrá hacerse en él sin consentimiento del titular o resolución judicial, salvo en caso de flagrante delito.

3.- Se garantiza el secreto de las comunicaciones y, en especial, de las postales, telegráficas y telefónicas, salvo resolución judicial.

4.- La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos.

Artículo 18 de la [Constitución española de 1978](#).

Como podemos ver, ya en la [Constitución española de 1978](#) se recogían, en su artículo 18, derechos acerca del honor, la intimidad y la imagen personal, del mismo modo que se limitaba el uso de la informática para garantizar tales derechos.

Dada la gran importancia de regular estos derechos de un modo adaptado a las nuevas tecnologías, capaces de tratar datos de forma masiva, el 16 de mayo de 2016 se publicó el [RGPD](#) el cuál fue de aplicación a partir del 25 de mayo de 2018.

El RGPD era mucho más avanzado que la [LOPD](#) que funcionaba en España desde el año 1999, por lo que ésta tuvo que ser sustituida en 2018 por la [LOPD-GDD](#), la cual se escribió para estar acorde con el RGPD.

A continuación haremos un análisis de la normativa que se recoge en el RGPD, haciendo especial referencia a su impacto en relación al uso de datos para Big Data.

Para saber más

RGPD: Reglamento General de Protección de Datos

Puedes acceder al fichero PDF con la versión original del RGPD en castellano en el siguiente enlace:

[RGPD](#)

(2016)

El 23 de mayo de 2018 se publicó una corrección de errores. En el siguiente enlace puedes ver el texto final con la corrección de errores incorporada:

[RGPD](#)

(con corrección de errores de 2018)

También puedes acceder a información sobre el RGPD a través del siguiente enlace a la Wikipedia:

[Reglamento General de Protección de Datos](#)

(wikipedia)

Para saber más

LOPD-GDD: Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.

En el siguiente enlace puedes acceder al PDF de la LOPD-GDD de 2018.

[LOPD-GDD](#)

(2018)

3.1.- Definiciones.



[Matthias Zomer](#) (Dominio público)

El RGPD incluye un listado de definiciones para clarificar y delimitar diversos conceptos que en él se tratan.

A continuación enumeramos algunos de los más interesantes para el caso de Big Data:

✓ **Datos personales:**

Toda información sobre una persona física identificada o identificable (“el interesado”). Se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador online o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona (art. 4.1 RGPD).

✓ **Interesado:**

Una persona física identificada o identificable sobre la que los datos personales se están tratando (art. 4.1 RGPD).

✓ **Responsable de tratamiento (*controller*):**

La persona física o jurídica, autoridad pública, servicio u otro organismo que, solo o junto con otros, determine los fines y medios del tratamiento de datos personales (art. 4.7 RGPD).

✓ **Encargado del tratamiento (*processor*):**

La persona física o jurídica, autoridad pública, servicio u otro organismo que trate datos personales por cuenta del responsable del tratamiento (art. 4.8 RGPD).

✓ **Destinatario:**

La persona física o jurídica, autoridad pública, servicio u otro organismo al que se comuniquen datos personales, se trate o no de un tercero (art. 4.9 RGPD).

✓ **Tercero:**

La persona física o jurídica, autoridad pública, servicio u organismo distinto del interesado, del responsable de tratamiento, del encargado de tratamiento y de las personas autorizadas para tratar los datos personales bajo la autoridad directa del responsable o del encargado (art. 4.10 RGPD).

✓ **Delegado de protección de datos (*Data Protection Officer* o *DPO*):**

Constituye uno de los elementos claves del RGPD, y un garante del cumplimiento de la normativa de la protección de datos en las organizaciones, sin sustituir las funciones que desarrollan las Autoridades de Control.

✓ **Tratamiento:**

Cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, ya sea por procedimientos automatizados o no, como la

recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción (art. 4.2 RGPD).

✓ **Elaboración de perfiles:**

Toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física (art. 4.4 RGPD).

✓ **Consentimiento del interesado:**

Toda manifestación de voluntad libre, específica, informada e inequívoca por la que el interesado acepta, ya sea mediante una declaración o una clara acción afirmativa, el tratamiento de datos personales que le conciernen (art. 4.11 RGPD).

Autoevaluación

El interesado es la persona física o jurídica, autoridad pública, servicio u otro organismo al que se comuniquen datos personales, se trate o no de un tercero.

Verdadero Falso

Falso

Ese es el destinatario.

El responsable del tratamiento es la persona física o jurídica, autoridad pública, servicio u otro organismo que trate datos personales por cuenta del responsable del tratamiento.

Verdadero Falso

Falso

Ese es el encargado del tratamiento.

El delegado de protección de datos es la persona física o jurídica, autoridad pública, servicio u organismo distinto del interesado, del responsable de tratamiento, del encargado de tratamiento y de las personas autorizadas para tratar los datos personales bajo la autoridad directa del responsable o del encargado.

Verdadero Falso

Falso

Ese es el tercero.



3.2.- Ámbito de aplicación y bases legales.

El RGPD tiene aplicación sobre el tratamiento automatizado de datos personales que forman parte de un fichero o de un sistema.

Los artículos 2 y 3 del RGPD tienen en cuenta una doble visión en relación a su **ámbito de aplicación**:

✓ **Ámbito subjetivo:**

- Se aplica tanto a los responsables como a los encargados del tratamiento.
- No se limita a ciertas áreas de aplicación (como pudieran ser por ejemplo lo personal o lo doméstico).

✓ **Ámbito territorial:**

- Aplica si los responsables o encargados están establecidos en la UE.
- También aplica si los interesados se encuentran en el territorio de la UE, aunque no lo estén los responsable o encargados, cuando el tratamiento se destine a ofrecer bienes o servicios a los interesados o al control de su comportamiento.

En los artículos 6 al 8 del RGPD se establecen las **bases legales** para poder tratar datos personales.

La condición básica necesaria para ello es que el interesado **dé su consentimiento** para utilizar sus datos. En tal caso, el responsable debe ser capaz de demostrar que el consentimiento ha sido dado libremente por el interesado, y la solicitud de consentimiento debe ser claramente perceptible.

Existen, sin embargo, algunos casos excepcionales en los que pueden tratarse datos personales sin consentimiento explícito de los interesados. En concreto cuando tal tratamiento sea necesario para:

- ✓ Ejecutar o negociar un contrato con el interesado.
- ✓ Cumplir con una obligación legal.
- ✓ Proteger los intereses vitales del interesado o de otra persona cuando el interesado sea incapaz de dar su consentimiento.
- ✓ El cumplimiento de una misión realizada en interés público o en el ejercicio de poder público.
- ✓ La satisfacción de los intereses legítimos (pero sujetos a los derechos y libertades fundamentales).

También existen casos de "interés legítimo" en los que se considera que pueden tratarse datos personales (por ejemplo para prevención del fraude o cuestiones de seguridad de la red o de la información). En tales casos el responsable debe informar al interesado de que se está realizando un tratamiento en base a tal interés legítimo.

Autoevaluación

¿Cuál es la condición básica para poder tratar datos personales del interesado?

- No hay una condición necesaria.
- Que el interesado reciba el correcto pago una vez se han usado sus datos.
- Que el interesado dé su consentimiento.
- Que el interesado resida en el mismo país en el que se tratan sus datos.

Incorrecto. Sí que la hay.

Incorrecto. No es una cuestión de pago por uso.

Correcto.

Incorrecto. Puede residir en otro país.

Solución

1. Incorrecto
2. Incorrecto
3. Opción correcta
4. Incorrecto

3.3.- Derechos de los interesados.

Con la RGPD los interesados siguen manteniendo los derechos ARCO (acceso, rectificación, cancelación y oposición) que ya se contemplaban en anteriores normativas, pero ahora en una versión más completa.

En concreto, los interesados tienen los siguientes derechos en relación al tratamiento de sus datos:

- ✓ **Derecho de acceso:** los interesados tienen derecho a obtener copias de sus datos personales, junto con los detalles principales sobre cómo se tratan los datos.
- ✓ **Derecho de rectificación:** los interesados tienen el derecho a exigir la rectificación de sus datos personales, sin dilaciones indebidas y el derecho a completar los datos personales que sean incompletos.
- ✓ **Derecho al olvido:** los interesados tienen derecho a que sus datos personales sean suprimidos del fichero o sistema.
- ✓ **Derecho de limitación:** los interesados tienen derecho a impedir tratamientos adicionales.
- ✓ **Derecho a la portabilidad:** los interesados tienen derecho a exigir que sus datos sean proporcionados en un formato de “uso común y lectura por equipos y máquinas” para poder ser transmitidos a otro responsable.
- ✓ **Derecho a la notificación:** los interesados tienen derecho a ser notificados por el responsable ante cualquier rectificación, supresión y limitación salvo que le sea imposible o exija un esfuerzo desproporcionado.

Además, los responsables quedan obligados a ser más transparentes con los interesados.

En concreto, los interesados deben recibir información de cómo se tratan sus datos, incluyendo:

- ✓ La identidad del responsable e información de contacto.
- ✓ También la identidad de cualquier delegado de protección de datos.
- ✓ Las finalidades y bases legales para su tratamiento.
- ✓ Cualquier “interés legítimo” que sea la base del tratamiento.
- ✓ Cualquier transferencia internacional y garantías aplicables.
- ✓ El período de retención o los criterios para su determinación.
- ✓ El derecho a la portabilidad de datos y los derechos de oposición al tratamiento, de requerir la limitación y de retirar el consentimiento al tratamiento.
- ✓ El derecho a reclamar ante una autoridad de control.
- ✓ Cualquier requisito legal o contractual para proporcionar datos, así como las consecuencias de no proporcionarlos.

Autoevaluación

¿Cuáles de los siguientes son derechos de los interesados en relación al tratamiento de sus datos?

Derecho al olvido.

Derecho a la duplicidad.

Derecho de acceso.

Derecho de rectificación.

Mostrar retroalimentación

Solución

1. Correcto
2. Incorrecto
3. Correcto
4. Correcto

3.4.- Gobierno y rendición de cuentas.



[George Becker](#) (Dominio público)

A partir de la puesta en funcionamiento del RGPD, los responsables están obligados a:

- ✓ Garantizar su cumplimiento.
- ✓ Estar en condiciones de demostrarlo.

Aparte, los responsables están obligados a implementar la protección de datos:

- ✓ Por diseño (*privacy by design*).
- ✓ Por defecto (*privacy by default*).

En concreto deben tomar las siguiente **medidas**:

- ✓ Emplear políticas de protección de datos.
- ✓ Emplear códigos de conducta.
- ✓ Emplear mecanismos de certificación.

Usarán para ello las siguientes **técnicas** (entre otras):

- ✓ **Seudonimización de datos:** de modo que no se pueda reconocer la identidad de una persona sin utilizar información adicional.
- ✓ **Minimización de datos:** sólo tratar los datos personales que sean necesarios para la finalidad correspondiente.

Además, los responsables deben:

- ✓ Antes de comenzar con el tratamiento de los datos, realizar una evaluación del impacto de las actividades que supongan riesgos importantes para los derechos de los interesados.
- ✓ Designar un delegado de protección de datos (DPO) si:
 - ➔ son autoridades u organismos públicos
 - ➔ o realizan observación habitual y sistemática de interesados a gran escala (Big Data)
 - ➔ o tratan datos sensibles a gran escala.
- ✓ Mantener registros (documentación) de las actividades de tratamiento que contengan cierta información requerida (fines de tratamiento, descripción de las categorías de los

interesados, datos personales, destinatarios, las medidas técnicas y organizativas, y cualquier transferencia de datos a terceros países).

Autoevaluación

¿Qué es la seudonimización de datos?

- Un mecanismo para evitar que entre los datos acerca de un usuario figure su género.
- Un mecanismo para que todas las fechas de nacimiento de los interesados se eliminen automáticamente.
- Un mecanismo para que cada identificador único en las bases de datos se almacene en codificación hexadecimal.
- Un mecanismo para evitar que se pueda reconocer la identidad de una persona sin utilizar información adicional.

Incorrecto. No tiene que ver (al menos no únicamente) con el género de las personas.

Incorrecto. No tiene que ver con fechas de nacimiento.

Incorrecto. No tiene que ver con cómo se codifican los identificadores únicos.

Correcto.

Solución

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta



3.5.- Obligaciones de los encargados del tratamiento.

El RGPD impone una serie de **requisitos para los contratos** entre los responsables de tratamiento de datos y los encargados designados para ello:

- ✓ Un responsable sólo puede encargar un tratamiento de datos a quien cumpla con garantías técnicas y organizativas para cumplir con el RGPD.
- ✓ Debe existir un contrato escrito entre el responsable y el encargado.

Tales **contratos deben estipular** los siguiente:

- ✓ El encargado únicamente tratará los datos personales de acuerdo con las instrucciones del responsable.
- ✓ El encargado debe asegurar que su personal está sujeto a una obligación de confidencialidad.
- ✓ El encargado debe implementar las medidas técnicas y organizativas adecuadas para garantizar un nivel de seguridad de los datos personales apropiado al riesgo.
- ✓ El encargado no puede subcontratar el tratamiento de datos personales sin la autorización previa y por escrito del responsable.
- ✓ Cualquier contrato entre un encargado y un subencargado debe proporcionar las mismas obligaciones de protección de datos que las previstas en el contrato con el responsable.
- ✓ El encargado debe asistir al responsable en garantizar el cumplimiento de las obligaciones de seguridad, la evaluación de impacto de protección de datos y la consulta previa a la Autoridad de Protección de Datos para el tratamiento de datos de alto riesgo.
- ✓ El encargado debe suprimir o devolver los datos personales cuando el tratamiento se haya completado.
- ✓ El encargado debe proporcionar al responsable toda la información necesaria para demostrar el cumplimiento, así como permitir y contribuir a la realización de auditorías.

Autoevaluación

¿Un encargado de tratamiento de datos puede subcontratar a su vez el tratamiento de datos?

- No puede en ningún caso.
- Sí, pero únicamente contando con la autorización previa y por escrito del responsable y mediante contrato con el interesado de modo que éste sepa que la ley se está cumpliendo.
- Sí, en cualquier caso.
- Sí, pero únicamente contando con la autorización previa y por escrito del responsable y mediante contrato con el subcontratado de modo que

quede obligado a cumplir con sus mismas obligaciones.

Incorrecto. En algunos casos sí puede.

Incorrecto. No se hace un contrato con el interesado.

Incorrecto. Necesitará algún tipo de autorización.

Correcto.

Solución

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta

3.6.- Seguridad de los datos.

Como ya hemos visto, el RGPD obliga a los responsables y encargados a aplicar las medidas técnicas y organizativas de seguridad apropiadas para garantizar un nivel adecuado de protección de los datos personales.

Las **medidas de seguridad** deben incluir:

- ✓ La seudonimización de los datos personales.
- ✓ El cifrado de datos personales.
- ✓ La capacidad de restaurar los datos personales de forma rápida.
- ✓ La implementación de procesos de verificación y evaluación.

El RGPD también obliga a la **notificación de la violación de seguridad** de los datos personales:

- ✓ Los responsables deben notificar las violaciones de la seguridad de los datos personales a la autoridad de control pertinente sin dilación indebida (cuando sea posible, dentro de las 72 horas posteriores a la detección de la violación), a menos que sea improbable que dicha violación de la seguridad constituya un riesgo para los derechos y libertades de los interesados.
- ✓ Los responsables deben notificar a los interesados afectados por la violación de sus datos personales cuando suponga un alto riesgo para los derechos y libertades de los interesados.
- ✓ Los encargados deben informar sobre las violaciones de la seguridad de los datos personales a los responsables sin dilación indebida y en todos los casos.

Autoevaluación

¿Cuáles de las siguientes son medidas de seguridad que se deben cumplir al tratar datos personales?

- El cifrado de los datos personales.

- La ocultación de fechas de nacimiento.

- La replicación en distintos clusters de toda la información.

- La capacidad de restaurar los datos personales de forma rápida.

Mostrar retroalimentación

Solución

1. Correcto
2. Incorrecto
3. Incorrecto
4. Correcto

3.7.- Otros aspectos de la normativa.

El RGPD es muy extenso, por lo que no cubriremos aquí todo lo que incluye.

En el caso de que el alumno necesite realizar actividades con la seguridad de que cumple con la legislación, deberá atender a los siguientes documentos legales:

- ✓ [Reglamento General de Protección de Datos \(RGPD\)](#) (con corrección de errores de 2018).
- ✓ [Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales \(LOPD-GDD\)](#).

En concreto, no quedan aquí cubiertos pero pueden ser interesantes los artículos del RGPD:

- ✓ Artículos 40 a 43, acerca de los **códigos de conducta** y la **acreditación de certificaciones** para demostrar el cumplimiento de las normas.
- ✓ Artículos 44 a 50, acerca de las **transferencias internacionales de datos**.
- ✓ Artículos 51 a 76, acerca de la **supervisión por las autoridades de protección de datos**.
- ✓ Artículos 77 a 84, acerca de los posibles **recursos** así como las potenciales **responsabilidades y sanciones**.

Autoevaluación

¿Cuáles de los siguientes son aspectos que cubre el RGPD?

- Transferencias internacionales de datos.

- Supervisión por las autoridades de protección de datos.

- Responsabilidades y sanciones.

- Tipos de impuestos aplicables.

- Códigos de conducta.

Mostrar retroalimentación

Solución

1. Correcto
2. Correcto
3. Correcto
4. Incorrecto
5. Correcto

4.- Gobierno de Datos.

Caso práctico

—¿Dónde están los *datasets*?

—En la carpeta *datasets* que hay dentro de la carpeta del proyecto.

—Espera, ¿están copiados dentro del proyecto?

—Sí, en la carpeta de cada proyecto están copiados los *datasets* que ese proyecto use.

—¡Pero los *datasets* ocupan muchísimo espacio! ¿Y si hay dos proyectos que usan el mismo *dataset* de entrada?

—Pues entonces estará copiado dos veces.

—No me lo puedo creer.

—Ya, es una pérdida de espacio, pero se empezó a hacer así en su momento y hemos seguido haciéndolo igual.

—¿Y los resultados cómo se guardan?

—Eso depende del programador. Cada uno los mete en las carpetas que considera oportuno y con el formato que mejor le parezca.

—Esto es increíble. ¿Y todas estas carpetas tienen gestión de permisos o cualquier persona de la empresa puede entrar en ellas?

—Puede entrar cualquiera. ¿Ves? Yo desde aquí entro a toda las carpetas sin problema.

—¿Y si te da la gana puedes borrar lo que quieras?

—Claro.

—¿En serio?

—En serio.

—Esto es la mayor locura que he visto en mucho tiempo.



[Memed Nurrohmadi](#) (Dominio público)

—Lo sé. Por eso te hemos llamado. ¿Tú no eres experto en Gobierno de Datos?

Ya hemos visto la normativa relativa al tratamiento de datos, analizando en concreto el RGPD vigente a partir de 2018.

A continuación queda hablar del **Gobierno de Datos**, que abarca una gran variedad de aspectos relacionados con todo el ciclo de vida del dato, desde que es producido hasta que es eliminado del sistema.

Podemos definir el Gobierno de Datos como el ejercicio de control y autoridad y comunicación sobre la gestión realizada de los datos, con la finalidad de que asegurar que tal gestión es correcta de acuerdo con las políticas y las mejores prácticas.

En este sentido, el Gobierno de Datos incluye:

- ✓ Planificación.
- ✓ Ejecución.
- ✓ Seguimiento.

Dada la especial relación entre el Gobierno de Datos y la Gestión de Datos, merece la pena realizar aquí la **distinción entre ambos**:

- ✓ La **Gestión de Datos** se realiza para asegurar que la organización obtiene valor de los datos.
- ✓ El **Gobierno de Datos** se realiza para supervisar la Gestión de Datos, asegurándose de que la gestión es la correcta (cómo se toman las decisiones y cómo se comportan personas y procesos en relación con los datos).

Autoevaluación

¿Cuál es la relación existente entre Gestión de Datos y Gobierno de Datos?

- El Gobierno de Datos supervisa la Gestión de Datos.
- El Gobierno de Datos es supervisado por la Gestión de Datos.
- Son equivalentes.
- El Gobierno de Datos es una de las actividades contempladas dentro de la Gestión de Datos.

Correcto.

Incorrecto. La Gestión de Datos se realiza para asegurar que la organización obtiene valor de los datos.

Incorrecto. No son equivalentes.

Incorrecto. Una cosa no está incluida en la otra.

Solución

1. Opción correcta
2. Incorrecto
3. Incorrecto
4. Incorrecto

4.1.- Objetivos del Gobierno de Datos.

El Gobierno de Datos tiene los siguientes **objetivos generales**:

- ✓ Facilitar a las organizaciones la gestión de sus **datos** como los **activos** que son.
- ✓ Definir, implementar y comunicar, en relación a los datos:
 - Principios.
 - Políticas.
 - Procedimientos.
 - Métricas.
 - Herramientas.
 - Responsabilidades.
- ✓ Monitorizar y guiar el cumplimiento de las políticas definidas respecto a gestión de los datos.

También podemos considerar los siguientes objetivos, esta vez **enfocados por áreas**:

- ✓ Realizar una gestión general de riesgos respecto a posibles incidentes.
- ✓ Asegurar la seguridad de los datos.
- ✓ Asegurar la privacidad de los datos.
- ✓ Asegurar el cumplimiento de las normativas.
- ✓ Asegurar la calidad de los datos.
- ✓ Asegurar la correcta gestión de metadatos.
- ✓ Conseguir procesos eficaces.
- ✓ Asegurar que el ciclo de vida de los datos es claro y está controlado.

Autoevaluación

¿Cuáles de los siguientes están entre los objetivos del Gobierno de Datos?

- Asegurar la calidad de los datos.

- Evitar la necesidad de realizar procesamiento por lotes.

- Rebajar las necesidades de memoria RAM de los nodos del clúster.

- Asegurar la correcta gestión de metadatos.

- Asegurar que el ciclo de vida de los datos es claro y está controlado.

Mostrar retroalimentación

Solución

1. Correcto
2. Incorrecto
3. Incorrecto
4. Correcto
5. Correcto

4.2.- Marco de Referencia.



[David Puig y Juan Mañes](#) (Dominio público)

El Gobierno de Datos está presente en gran cantidad de actividades relacionadas con la gestión de los datos. Algunos autores llaman a estas actividades el marco de referencia del Gobierno de Datos:

- ✓ **Modelado y Diseño de Datos:**
Modelado lógico de datos y cómo se va a implementar en la organización.
- ✓ **Almacenamiento y Operación de Datos:**
Almacenamiento de datos, mecanismos de despliegue y administración de procesos de carga.
- ✓ **Seguridad de Datos:**
Diseño y desarrollo de políticas, estándares, auditoría de la seguridad y cumplimiento regulatorio.
- ✓ **Integración e Interoperabilidad de Datos:**
Diseño e implementación de arquitecturas y estándares de interoperabilidad e integración de datos.

- ✓ **Gestión de Documentos y Contenido:**
Políticas y actividades para la documentación de los datos a lo largo de su ciclo de vida.
- ✓ **Datos Maestros y de Referencia:**
Definición de requisitos y modelos de datos maestros críticos para la organización.
- ✓ **Data Warehousing & Business Intelligence:**
Definición de arquitecturas de Almacenes de Datos y sistemas de reportado para asegurar el uso correcto de los datos cuando se emplean para Inteligencia de Negocio.
- ✓ **Metadatos:**
Definición del modelo de metadatos, incluyendo tanto su descripción técnica como de negocio.
- ✓ **Calidad de Datos:**
Perfilado de datos, políticas, guías de calidad de datos y monitorización.
- ✓ **Arquitectura de Datos:**
Diseño de la estructura lógica y física de los sistemas que van a manejar los datos.

Autoevaluación

¿Cuáles de las siguientes son actividades en las que está presente el Gobierno de Datos?

- Definición del modelo de metadatos, incluyendo tanto su descripción técnica como de negocio.

- Optimización del desempeño de los microprocesadores.

- Políticas y actividades para la documentación de los datos a lo largo de su ciclo de vida.

- Almacenamiento de datos, mecanismos de despliegue y administración de procesos de carga.

- Control de la temperatura alcanzada por los componentes de los nodos.

- Diseño y desarrollo de políticas, estándares, auditoría de la seguridad y cumplimiento regulatorio.

- Políticas y actividades para la documentación de los datos a lo largo de su ciclo de vida.

- Monitorización de las funciones de red.

Mostrar retroalimentación

Solución

1. Correcto
2. Incorrecto
3. Correcto
4. Correcto
5. Incorrecto
6. Correcto
7. Correcto
8. Incorrecto

4.3.- Roles.

En relación a Gobierno de Dato existen diversos roles a desempeñar, siendo claves los siguientes:

Chief Data Officer (CDO):

- ✓ Perfil ejecutivo con función transversal (negocio e IT).
- ✓ Junto con el consejo de gobierno define la estrategia de datos y cómo será la gestión de los mismos.
- ✓ Responsable del modelo de datos.
- ✓ Asesora y supervisa.

Oficina de Gobierno del Dato:

- ✓ Ayuda y asiste al CDO en sus funciones (redacción de políticas, medidas de mejora y calidad, coordinación con IT, ...).

Data Owners (propietarios del dato):

- ✓ Perfil de negocio.
- ✓ Responsables de los departamentos que usan/producen los datos.
- ✓ Deben definir los objetivos de calidad de los datos.
- ✓ Suele haber un *data owner* por dominio o departamento.

Data Stewards (administradores del dato):

- ✓ Perfil de negocio.
- ✓ Responsables de implementar las políticas y procesos definidos.
- ✓ Identifican problemas y necesidades.
- ✓ Suele haber entre 1 y 3 *data steward* por dominio.

Data Custodians (custodios del dato):

- ✓ Perfil IT.
- ✓ Atienden las peticiones de tecnología de los propietarios de los datos.
- ✓ Son el punto de contacto del CDO con la tecnología.
- ✓ Suele haber un *data custodian* por sistema.

Autoevaluación

El *data owner* es aquel que ejecuta la gestión de los datos.

Verdadero Falso

Falso

Es un responsable de un departamento que usa/produce datos.

Los custodios del dato atienden las peticiones de tecnología de los propietarios de los datos.

Verdadero Falso

Verdadero

Cierto.

El CDO es responsable de implementar las políticas y procesos definidos.

Verdadero Falso

Falso

Es responsable del modelo de datos. Define la estrategia de datos y cómo será la gestión de los mismos.

Los *data stewards* son responsables de implementar las políticas y procesos definidos.

Verdadero Falso

Verdadero

Cierto.