Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)

Sistemas de Big Data

UD03. Gestión de Datos.
Resumen.

JUAN ANTONIO GARCIA MUELAS

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.

Los procesos ELT **no son equivalentes a un proceso de Integración de datos**.

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.

Las fuentes de origen pueden contener datos recientes, pero también históricos.

Las **3 fases de un proceso ELT**:

- ✓ **Extraer (extract):** Fase en la que se extraen los datos desde diversas fuentes, incluyendo una validación previa de los mismos. De su correcta ejecución depende el éxito del resultado final. Es importante elegir bien la estrategia para recibir datos para no afectar al funcionamiento de los sistemas desde los que vienen los datos.
- ✓ **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.
  Se emplea una serie de reglas sobre los datos extraídos con la intención de prepararlos para ser finalmente cargados en el destino correspondiente, asegurando que son válidos y tienen el formato deseado. Suele ser el proceso que mayor esfuerzo humano requiere.
  - ➢ Seleccionar sólo determinadas columnas o atributos de cada registro.
  - ➢ Seleccionar sólo determinados registros según si cumplen o no determinada condición.
  - ➢ Transponer filas en columas 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.
- ✓ **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. Es la fase más sencilla y liviana.
  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.

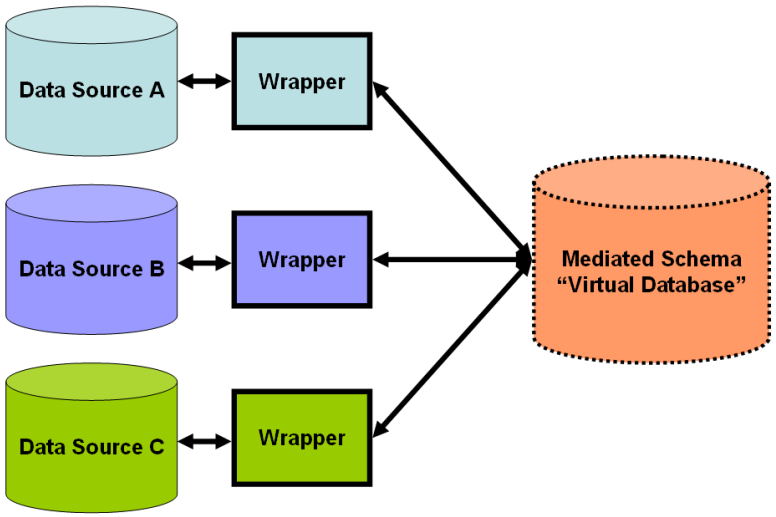
Si enviamos datos de un sistema **OLTP a uno OLAP** con intención de emplear el segundo de ellos para realizar la analítica **descargamos al OLTP de la carga de la analítica**.

Las dos **herramientas más comunes** dentro del ecosistema Hadoop (diseñadas para envíos de datos a HDFS, pero compatibles con otros como Amazon S3) **para procesos ETL** son:

- ✓ **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. Puede conectarse con cualquier base de datos con conexión 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.
- ✓ **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.
  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.

INTEGRACIÓN DE DATOS.

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).



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 usario 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.

NORMATIVA DE TRATAMIENTO DE DATOS.

En el Art.18 de la Constitución se recogen los derechos acerca del honor, intimidad e imagen personal, así como la limitación a estos efectos del uso de la informática.

En 2016 se publica el RGPD, de aplicación desde 2018.

A continuación, enumeramos algunos de los conceptos 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).

El interesado debe dar su consentimiento para utilizar sus datos, aunque hay 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).

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 maquinas" 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.

Para implementarlo se usarán 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.

Se podrán subcontratar los servicios de tratamiento de datos 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.

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.

GOBIERNO DE DATOS

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).

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.

El marco de referencia (actividades relacionadas con la gestión de datos) 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.

ROLES.

El **data Owner** es un responsable de un departamento que usa/produce datos.

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

El **CDO** es responsable del modelo de datos, definiendo estrategia y gestión de los mismos.

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

ENLACES DE INTERÉS.

https://sqoop.apache.org/docs/1.99.7/user/Sqoop5MinutesDemo.html

https://sqoop.apache.org/docs/1.99.7/index.html

https://flume.apache.org/documentation.html

https://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html

https://www.boe.es/doue/2016/119/L00001-00088.pdf

https://eur-lex.europa.eu/legal-content/ES/TXT/?qid=1532348683434&uri=CELEX%3A02016R0679-20160504

https://es.wikipedia.org/wiki/Reglamento_General_de_Protecci%C3%B3n_de_Datos

https://www.boe.es/buscar/pdf/2018/BOE-A-2018-16673-consolidado.pdf