

# Tema 2.

## Fundamentos de la gestión de datos.

Gobernanza de datos y Big Data en los ayuntamientos del siglo XXI



**FVMP**  
Federació Valenciana  
de Municipis i Províncies

# Contenidos del tema



- Tipos de datos en el contexto de las entidades públicas locales
  - Datos estructurados, no estructurados y semi estructurados.
  - Ejemplos
- Ciclo de vida de los datos
  - Fases del ciclo de vida recolección, almacenamiento, procesamiento, análisis y disposición. Concepto de ETL.
  - Importancia de cada fase en la gestión eficaz del dato.

- Principios de gestión de datos
  - Principios de integridad, disponibilidad, confidencialidad y trazabilidad
  - Implementación en entornos de EELL





# Tipos de datos en el contexto de la ciencia de datos

- Categorías o clasificaciones que se pueden aplicar a los datos según sus características y propiedades. → Ayudan a comprender la naturaleza de los datos y a determinar qué métodos y técnicas son más apropiados para analizarlos y procesarlos.
- De **clasificación**: Numéricos, categóricos, textuales, binarios, ....
- De **estructura**: datos estructurados, semiestructurados y no estructurados
- De **calidad**: datos limpios, datos ruidosos, datos faltantes
- De **frecuencia**: datos estáticos vs. datos en tiempo real
- Entender las tipologías de datos es fundamental en la ciencia de datos porque influyen en las decisiones sobre qué técnicas y herramientas utilizar para el análisis, procesamiento y visualización de los datos, así como en la forma en que se interpretan los resultados obtenidos.





# Tipologías de clasificación

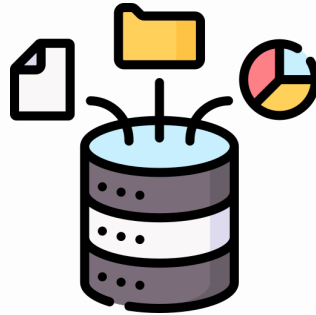
1. **Numérico:** Datos que representan valores numéricos. Pueden ser enteros o de coma flotante. Por ejemplo, la edad de una persona o la temperatura de un lugar.
2. **Categorico:** Datos que representan categorías o grupos discretos. Por ejemplo, el género de una persona o el tipo de producto.
3. **Ordinal:** Datos que tienen un orden inherente entre las categorías, pero la diferencia entre los valores no es significativa. Por ejemplo, las calificaciones de un producto en una escala del 1 al 5.
4. **Binario:** Datos que solo pueden tomar dos valores distintos, como verdadero/falso, sí/no, 0/1.
5. **Texto:** Datos que consisten en palabras, frases o texto completo. Por ejemplo, comentarios de clientes o descripciones de productos.
6. **Fecha/Hora:** Datos que representan fechas y/o horas. Por ejemplo, la fecha de nacimiento de una persona o la hora de registro de un evento.
7. **Imágenes:** Datos que representan imágenes digitales, como fotografías o imágenes médicas.
8. **Audio:** Datos que representan señales de audio, como grabaciones de voz o música.
9. **Video:** Datos que representan secuencias de imágenes en movimiento, como películas o videos de vigilancia.
10. **Geoespacial:** Datos que representan ubicaciones físicas en la Tierra, como coordenadas GPS.



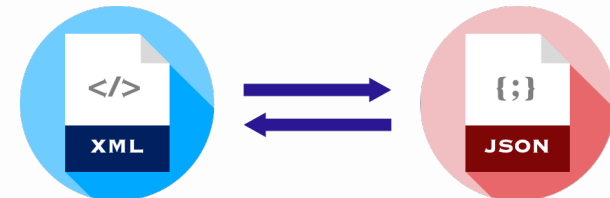


# Tipología de datos basados en su estructura

- **Datos estructurados:** Estructura clara y marcada en filas y columnas, con campos. Hojas **Excel**, **bases de datos** relacionales. Fáciles de consultar y de tratar.











- **Datos semiestructurados:** Son datos que no se ajustan perfectamente a un formato tabular, pero aún tienen cierto grado de estructura. A menudo, los datos semiestructurados se organizan en formatos como **JSON** (JavaScript Object Notation) o **XML** (eXtensible Markup Language), donde los datos están agrupados en campos con etiquetas o nombres de clave, pero no necesariamente todos los registros tienen los mismos campos. Esto es común en datos web, archivos de configuración y registros de eventos.





# Tipología de datos basados en su estructura

- **Datos no estructurados:** No tienen una estructura predefinida y no se pueden organizar fácilmente en un formato tabular o en cualquier otro formato estructurado. Pueden incluir texto sin formato, imágenes, archivos de audio y video, correos electrónicos, redes sociales y otros tipos de contenido generado por el usuario. Analizar y extraer información significativa de datos no estructurados puede ser más difícil que con datos estructurados o semiestructurados, pero es un área de investigación activa en la ciencia de datos, utilizando técnicas como el procesamiento del lenguaje natural (NLP), visión por computadora y aprendizaje automático

 Text files and documents	 Server, website and application logs	 Sensor data	 Images
 Video files	 Audio files	 Emails	 Social media data





# Calidad de los datos (1/2)

- **Datos limpios** (*Clean data*):
  - Son datos que están completos, precisos y sin errores. Estos datos cumplen con los estándares de calidad establecidos y son confiables para su uso en análisis y toma de decisiones.
- **Datos ruidosos** (*Noisy data*):
  - Son datos que contienen errores, inconsistencias o valores atípicos que pueden distorsionar los resultados del análisis. El ruido en los datos puede deberse a errores de entrada, errores de medición, corrupción de datos o problemas durante el proceso de recopilación.
- **Datos faltantes** (*Missing data*):
  - Son datos que faltan en una o más variables o atributos en un conjunto de datos. Esto puede deberse a diversos factores, como fallos en la recopilación de datos, problemas de calidad de los datos, o simplemente porque ciertos datos no están disponibles en el momento de la recopilación.





## Calidad de los datos (2/2)

- **Datos duplicados** (*Duplicate data*):
  - Son datos que aparecen más de una vez en un conjunto de datos, ya sea debido a errores en la recopilación de datos o a la duplicación intencional de registros. La presencia de datos duplicados puede distorsionar los resultados del análisis y afectar la precisión de las conclusiones.
- **Datos inconsistentes** (*Inconsistent data*):
  - Son datos que presentan discrepancias o contradicciones entre diferentes fuentes o dentro del mismo conjunto de datos. Estas inconsistencias pueden surgir debido a problemas en la integración de datos, errores humanos, cambios en los requisitos de datos o problemas en la calidad de los datos.
- **Datos desactualizados** (*Outdated data*):
  - Son datos que han perdido su relevancia o precisión debido a cambios en el contexto o a la obsolescencia de la información. Mantener datos actualizados es importante para garantizar la vigencia y la utilidad de los análisis y las decisiones empresariales basadas en datos.







# Tipologías según su frecuencia (1/2)

- **Datos estáticos** (*Static data*):
  - Son datos que no cambian con el tiempo o que cambian muy raramente. Estos datos se capturan en un momento específico y permanecen constantes a lo largo del tiempo. Por ejemplo, datos demográficos de una población en un censo, información de productos en un catálogo, o registros históricos de transacciones.
- **Datos en tiempo real** (*Real-time data*):
  - Son datos que se generan, capturan y procesan instantáneamente, sin demoras perceptibles. Estos datos reflejan eventos y actividades que ocurren en el momento en que suceden. Por ejemplo, datos de sensores IoT (Internet de las cosas), transacciones financieras en línea, interacciones en redes sociales en tiempo real, o datos de tráfico en una red de transporte.

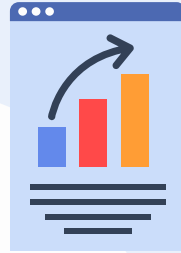




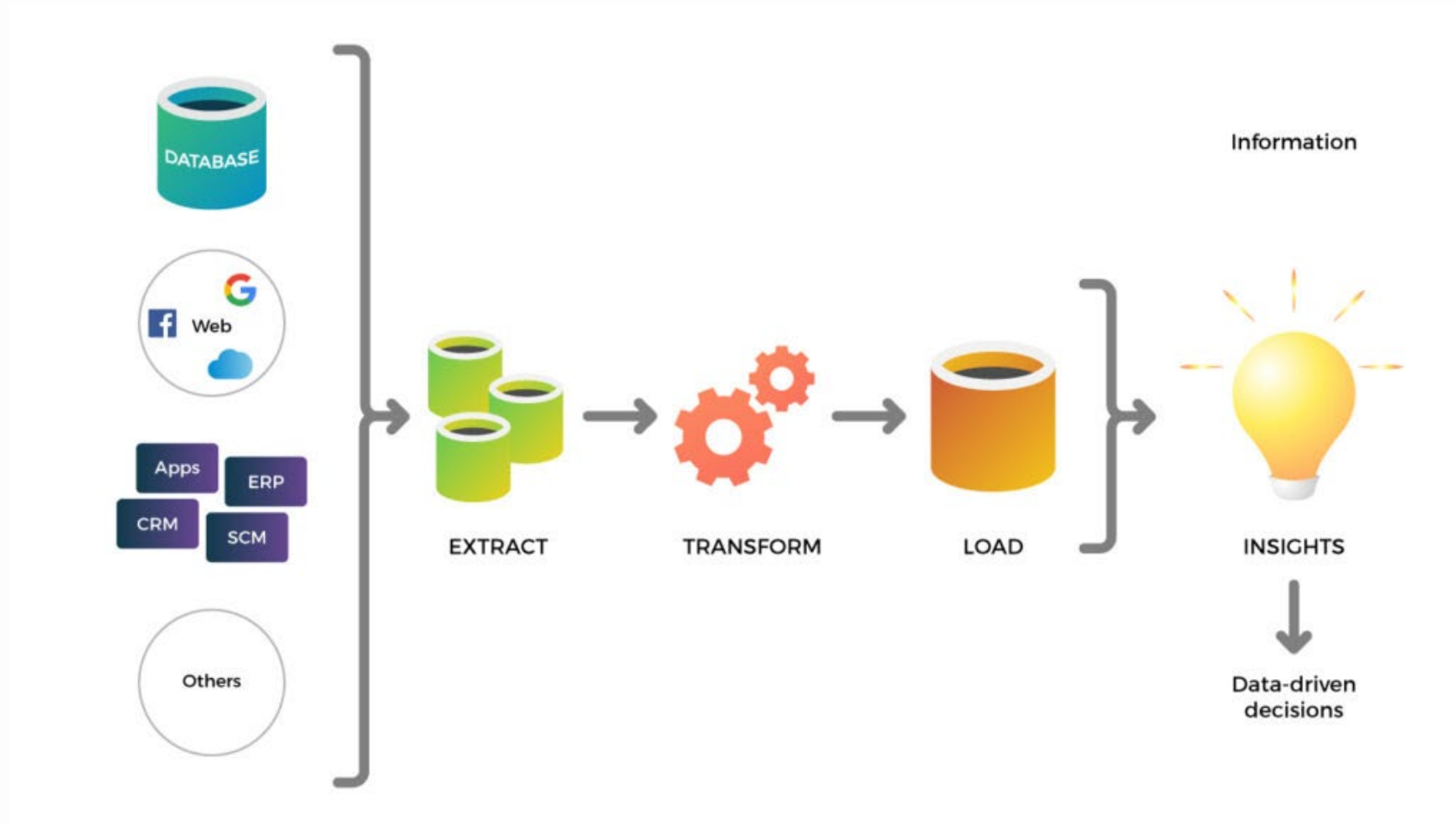
## Tipologías según su frecuencia (2/2)

- **Datos de transmisión** (*Streaming data*):
  - Son datos que se generan continuamente en tiempo real y se transmiten de manera continua a través de una red. Estos datos se procesan mientras están en movimiento y pueden incluir eventos de alta velocidad que deben ser analizados y respondidos en tiempo real. Ejemplos incluyen datos de sensores industriales, datos de clics en sitios web, o flujos de datos de redes sociales.
- **Datos históricos** (*Historical data*):
  - Son datos que se han recopilado y almacenado en el pasado y que pueden ser utilizados para análisis retrospectivos, modelado predictivo o toma de decisiones basada en el pasado. Estos datos proporcionan una visión de cómo han evolucionado las cosas con el tiempo y pueden ser útiles para identificar tendencias, patrones y correlaciones. Por ejemplo, datos de ventas pasadas, registros climáticos históricos, o datos de pacientes en registros médicos.



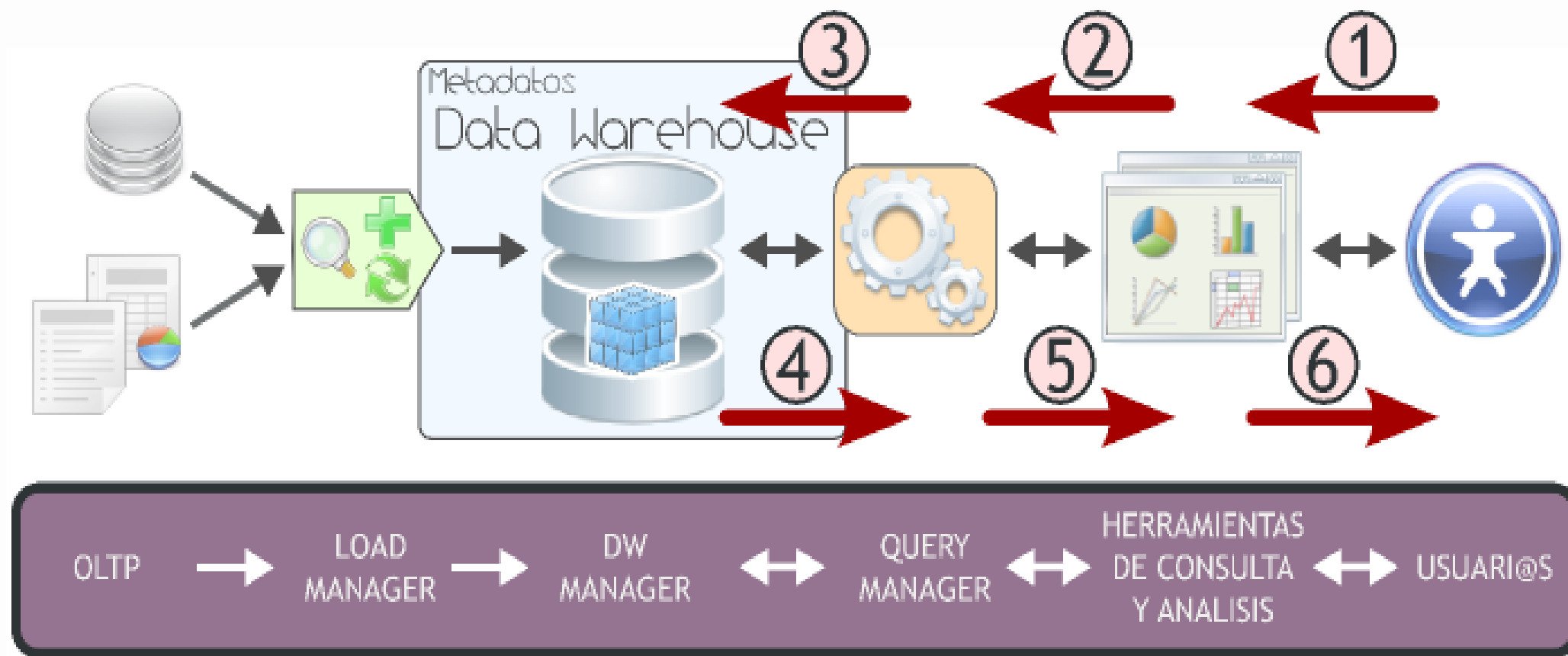


# Ciclo de vida de los datos. ETL





## Ciclo de vida de los datos. Post ETL





# La importancia del proceso ETL (1/2)

- El proceso ETL (*Extract, Transform, Load*) es fundamental en el ciclo de vida de los datos para garantizar que los datos se muevan de manera efectiva desde sus fuentes de origen hasta su destino final, ya sea una base de datos, un almacén de datos o un sistema de análisis.
  - **Extract – Extracción:** En esta fase, los datos se extraen de diversas fuentes de origen, como bases de datos, archivos planos, APIs web, sistemas de terceros, etc. Es crucial asegurarse de que los datos se extraigan de manera completa y precisa para evitar pérdidas de información y garantizar la integridad de los datos en el proceso.
  - **Transform – Transformación:** Durante esta fase, los datos extraídos se transforman en un formato que sea adecuado y útil para su análisis y uso posterior. Esto puede incluir limpieza de datos, normalización, conversión de formatos, agregación, enriquecimiento de datos, y otras operaciones. La transformación garantiza que los datos sean coherentes, completos y estén listos para su carga en el sistema de destino.





## La importancia del proceso ETL (2/2)

- **Load – Carga:** En esta fase, los datos transformados se cargan en el destino final, como un almacén de datos, un data lake o una base de datos para su almacenamiento y análisis posteriores. Es crucial garantizar que la carga de datos sea eficiente y segura, y que los datos estén disponibles para su acceso y consulta según las necesidades del negocio. Además, se debe tener en cuenta la integridad y la calidad de los datos durante la carga para evitar problemas posteriores en el análisis. “Vista minable”.
- Cada fase del proceso ETL es crítica para garantizar que los datos estén disponibles, limpios, consistentes y listos para su análisis y uso en la toma de decisiones empresariales. Un proceso ETL bien diseñado y ejecutado puede mejorar la eficiencia operativa, reducir los errores y garantizar la calidad de los datos en toda la organización.





# Principios de gestión de datos

- La gestión de datos es fundamental para garantizar que los datos de una organización sean
  - Confiables
  - Precisos,
  - Seguros
  - Que estén disponibles cuando sean necesarios.
- Aquí exploraremos algunos de los principios clave de la gestión de datos, centrándonos en la **integridad, disponibilidad, confidencialidad y trazabilidad**.





# Integridad de los datos

- La **integridad** de los datos se refiere a la **precisión, consistencia y fiabilidad** de la información almacenada. Los datos deben ser **exactos y completos**, sin errores ni discrepancias.
- Para garantizar la integridad de los datos, se pueden implementar controles como **validaciones de datos, restricciones de integridad en las bases de datos y auditorías regulares** para detectar y corregir problemas de calidad de datos.







# Disponibilidad de los datos

- La disponibilidad de los datos se refiere a asegurar que los datos estén **accesibles** cuando se necesiten. Esto implica tener sistemas y procedimientos en su lugar para garantizar que los datos estén para los usuarios autorizados en todo momento.
- Esto puede incluir la implementación de **redundancia** de datos, copias de seguridad regulares, sistemas de recuperación de desastres y monitoreo continuo de la infraestructura de datos.





# Confidencialidad de los datos

- La **confidencialidad** de los datos se refiere a proteger la **privacidad** y la confidencialidad de la información sensible y privada. Esto implica restringir el acceso a los datos a personas autorizadas y protegerlos contra accesos no autorizados o filtraciones de información.
- Para garantizar la confidencialidad de los datos, se pueden implementar **controles de acceso**, **cifrado** de datos, **políticas** de seguridad de la información y **capacitación** del personal sobre la importancia de proteger los datos confidenciales.





# Trazabilidad de los datos

- La **trazabilidad** de los datos implica mantener un **registro** de los cambios y las actividades que ocurren en torno a los datos a lo largo de su ciclo de vida. Esto permite rastrear quién accedió a los datos, cuándo y con qué propósito, así como también realizar un seguimiento de las modificaciones realizadas en los datos.
- Para garantizar la trazabilidad de los datos, se pueden implementar **registros de auditoría, registros de cambios en bases de datos y sistemas de gestión de versiones**.



# Comming up next...

Tema 3: **Infraestructura y tecnologías para Big Data** (Vicente Castelló) (Día 16 mayo)

- Arquitecturas de datos
- Tecnologías de almacenamiento y procesamiento (BDs, NoSQL, Hadoop, Spark..)
- Herramientas de análisis de datos (Power BI, Tableau..)

Tema 4. **Gobierno y regulaciones en la gestión de datos e IA.** Eduard Chaveli y Laura Vico. (20 mayo)

- Marco legal y regulatorio en la gestión de datos municipales
- Cumplimiento de normativas (GDPR, Ley de Protección de Datos)
- Ética en el uso de datos municipales e IA.

