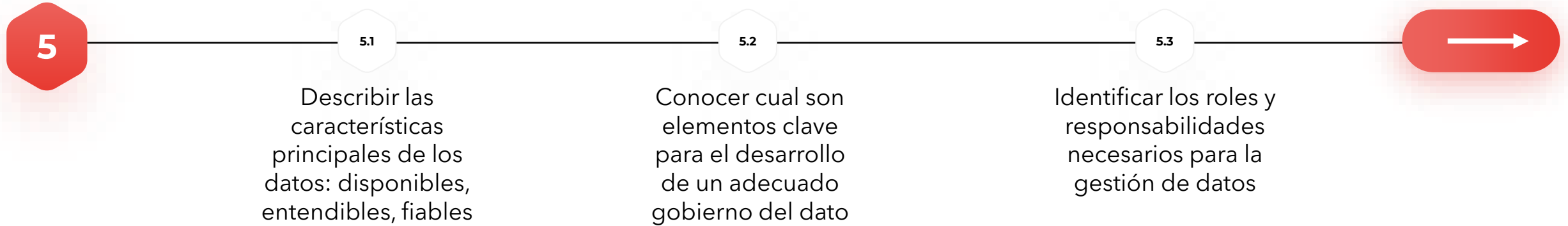




BIG DATA Y BUSINESS INTELLIGENCE

Identificación de los elementos clave para el gobierno y gestión de los datos

ÍNDICE



5.1

Describir las características principales de los datos: disponibles, entendibles, fiables

COMENZAR

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="UTF-8">
5     <title>Title goes here</title>
6   </head>
7   <body>
8
9   </body>
10 </html>
```





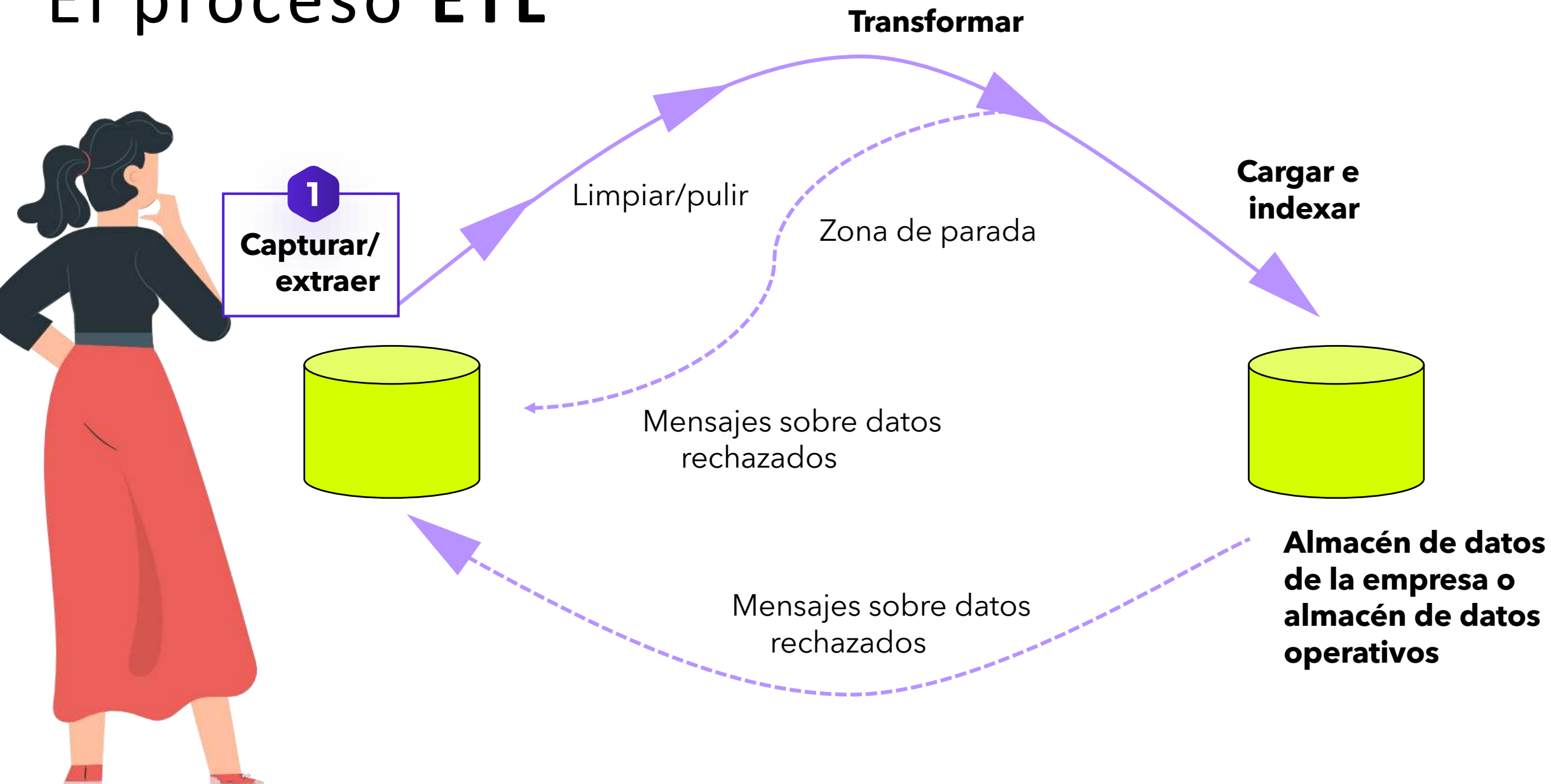
¿Qué es el **proceso ETL**?

El proceso ETL es un conjunto de técnicas y herramientas que se utilizan para extraer (Extract), transformar (Transform) y cargar (Load) datos desde diferentes fuentes de datos en un sistema de destino.

En este proceso, los datos se extraen de múltiples fuentes, se transforman para ajustarlos al formato del sistema de destino y se cargan en el sistema de destino.

El objetivo del proceso ETL es garantizar la calidad de los datos y su disponibilidad para su uso en la toma de decisiones empresariales.

El proceso ETL



1

Extracción

- Este paso cubre la recolección de datos del sistema fuente y los hace accesibles para su posterior procesamiento.
- El **objetivo principal** de la etapa de extracción es recuperar todos los datos necesarios del sistema fuente con los menores recursos posibles.

Pueden ser extraídos:

- Lógicamente
- Físicamente

2

3



Extracción lógica



Extracción completa

La extracción completa se utiliza cuando los datos deben ser extraídos y cargados por primera vez. En la extracción completa, los datos de la fuente se extraen por completo. Esta extracción refleja los datos actuales disponibles en el sistema fuente.

Extracción incremental

En la extracción incremental, los cambios en los datos de la fuente necesitan ser rastreados desde la última extracción exitosa. Sólo estos cambios en los datos serán extraídos y luego cargados. Estos cambios pueden detectarse a partir de los datos de origen que tienen la última marca de tiempo modificada. También se puede crear una tabla de cambios en el sistema de origen, que mantiene un seguimiento de los cambios en los datos de origen.

***Otro método** para obtener los cambios incrementales es **extraer los datos fuente completos y luego hacer una diferencia** (operación negativa) entre la extracción actual y la última extracción. Este método provoca un problema de rendimiento.

1

2

3

Extracción física

Extracción en línea

Los datos se extraen directamente del sistema fuente. El proceso de extracción se conecta al sistema fuente y extrae los datos de la fuente. Aquí los datos se extraen directamente de la Fuente para ser procesados en el área de preparación, por eso se llama extracción en línea. Durante la extracción nos conectamos directamente al sistema fuente y accedemos a las tablas de origen. No es necesario ningún área de preparación externa.

Extracción fuera de línea

Los datos del sistema fuente se vuelcan fuera del sistema fuente en un archivo plano. Este archivo plano se utiliza para extraer los datos. El archivo plano puede ser creado por un proceso rutinario diario. Aquí los datos no se extraen directamente de la fuente, sino que se toman de otra área externa que guarda la copia de la fuente.

El área externa puede ser archivos planos, o algunos archivos de volcado en un formato específico. Así, cuando necesitemos procesar los datos, podemos obtener los registros de la fuente externa en lugar de la fuente real.

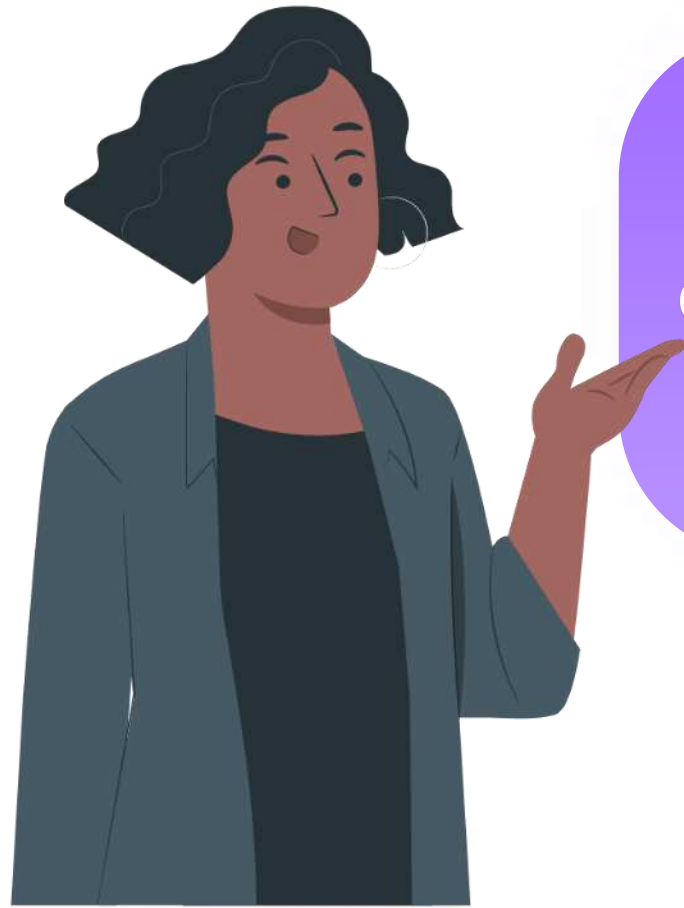


1

2

3

¿Qué datos se deben extraer?



La fase de
descubrimiento de
datos

La fase de
detección de
anomalías

1

2

3

La fase de descubrimiento de datos


- El **criterio clave** para el éxito en el almacenamiento de datos es la limpieza y la cohesión de los datos que contiene.
- Una vez que se entiende cómo debe ser el objetivo, es cuando pasamos a **identificar y examinar** las fuentes de datos.



1

2

3

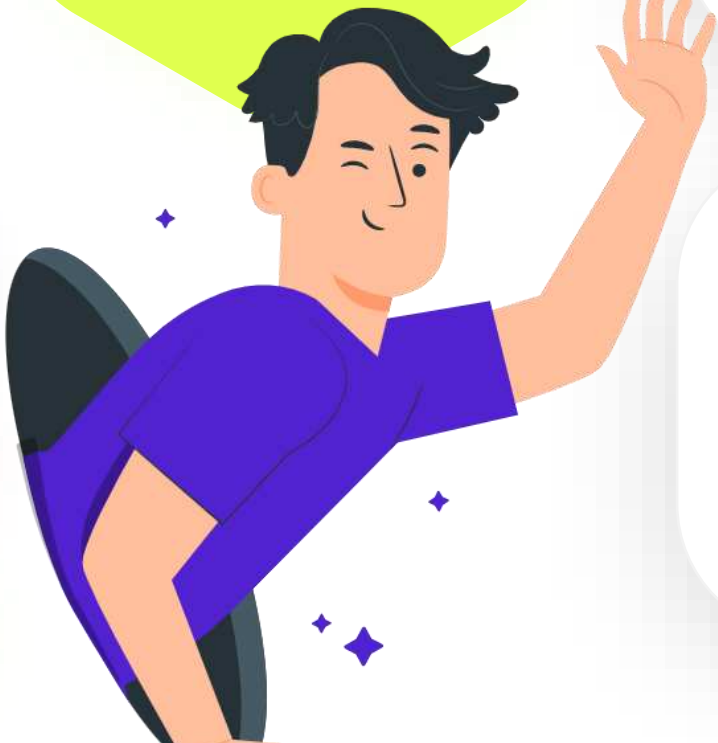


La fase de descubrimiento de datos

Corresponde al equipo de ETL profundizar en los requisitos de datos para determinar todos y cada uno de los sistemas de origen, tablas y atributos necesarios para cargar el almacén de datos o warehouse.

Los pasos a seguir son:

- Recogida y documentación de los sistemas fuente.
- Seguimiento de los sistemas fuente.
- Determinación del sistema de registro .
- Punto de origen de los datos.
- La definición del sistema de registro es importante porque en la mayoría de las empresas los datos se almacenan de forma redundante en muchos sistemas diferentes.
- Las empresas hacen esto para que los sistemas no integrados compartan datos. Es muy común que la misma pieza de datos sea copiada, movida, manipulada, transformada, alterada, limpiada o corrompida en toda la empresa, dando como resultado diferentes versiones de los mismos datos.



La fase de detección de anomalías

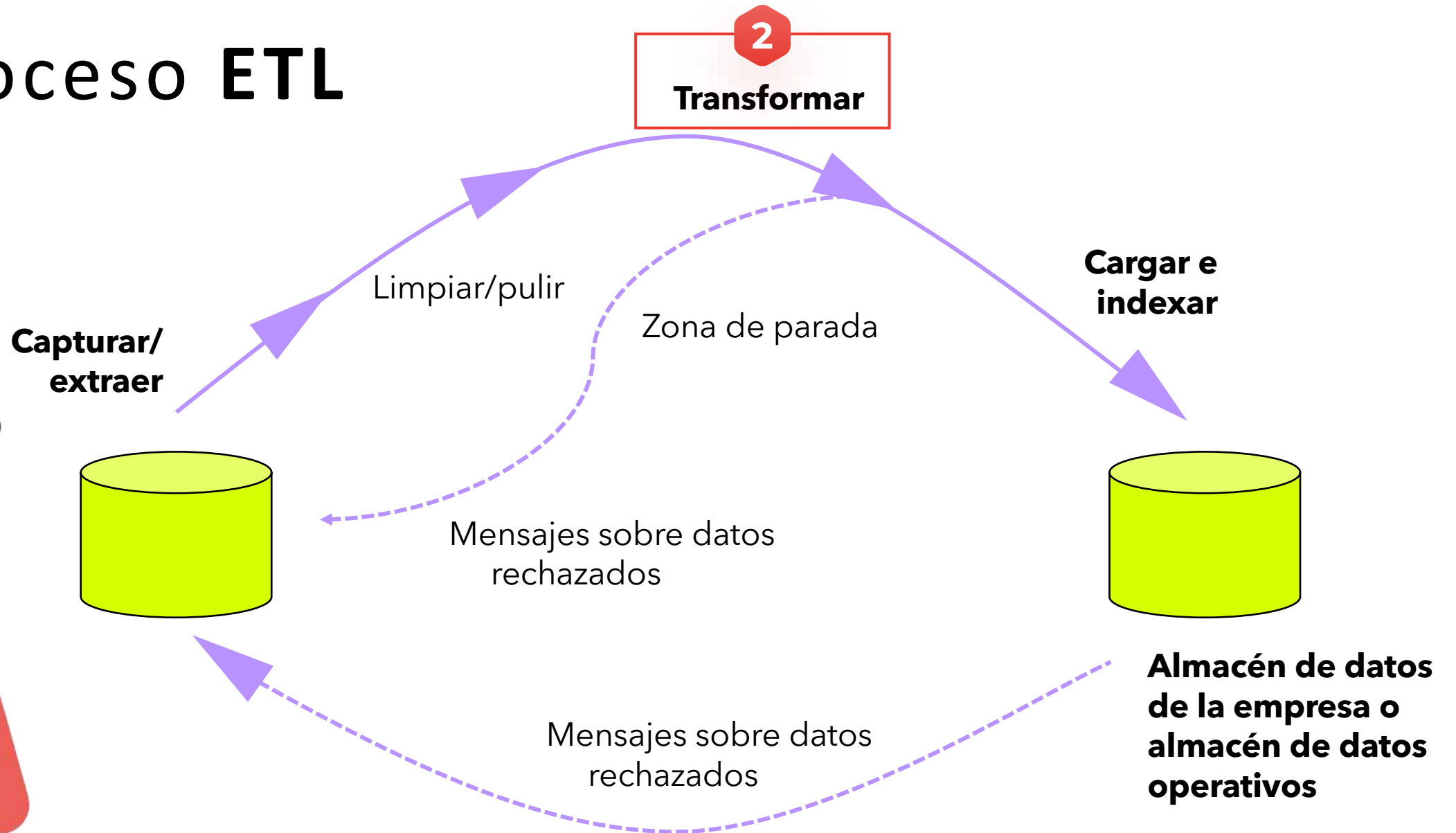
Datos con valores NULL

Un valor NULL no manejado correctamente puede destruir cualquier proceso ETL. Los valores NULL suponen un gran riesgo cuando se encuentran en columnas de clave foránea. Si se unen dos o más tablas basándose en una columna que contiene valores NULL, se perderán datos. Recuerda que en una base de datos relacional tipo MySQL, NULL no es igual a NULL. Por eso esas uniones fallan. Comprueba si hay valores NULL en todas las claves foráneas de la base de datos de origen. Si los valores NULL están presentes, debe realizar una unión manual de las tablas

Fechas en campos que no son de fecha

Las fechas son elementos muy peculiares porque son los únicos elementos lógicos que pueden venir en varios formatos, conteniendo literalmente diferentes valores y teniendo exactamente el mismo significado. Afortunadamente, la mayoría de los sistemas de bases de datos admiten la mayoría de los distintos formatos para su visualización, pero los almacenan en un único formato estándar

El proceso ETL



1

2

Transformación de datos

- La transformación es el componente de la conciliación de datos que convierte los datos del formato de los sistemas operativos de origen al formato del almacén de datos de la empresa.
- La transformación de datos consiste en una **variedad de funciones** diferentes:



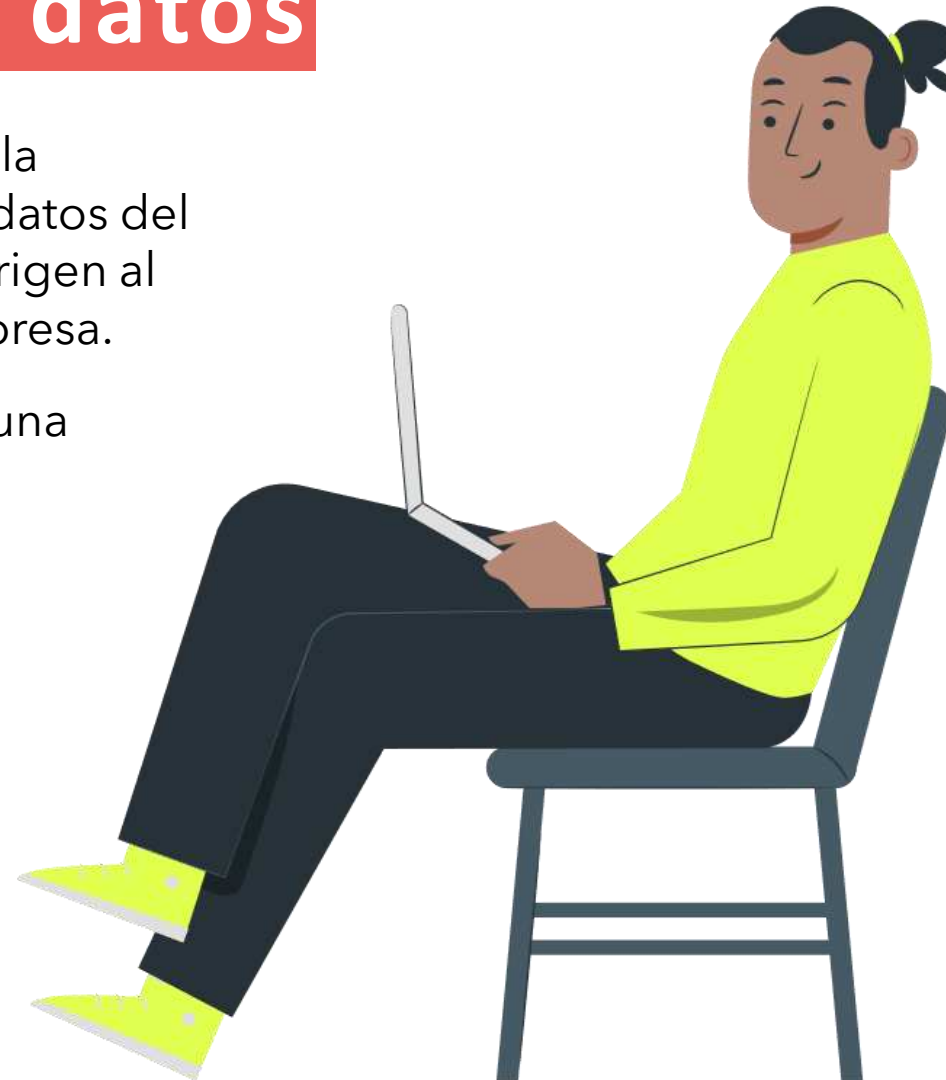
Funciones a nivel de registro.



Funciones a nivel de campo.



Transformaciones más complejas.



3

Transformación de datos

Funciones a nivel de registro

- **Selección:** partición de datos
- **Unión:** combinación de datos
- **Normalización**
- **Agregación:** resumen de datos, agregados

Funciones a nivel de campo

- **Transformación de campo único:** de un campo a otro campo
- **Transformación de múltiples campos:** de muchos campos a uno, o de un campo a muchos

1

2

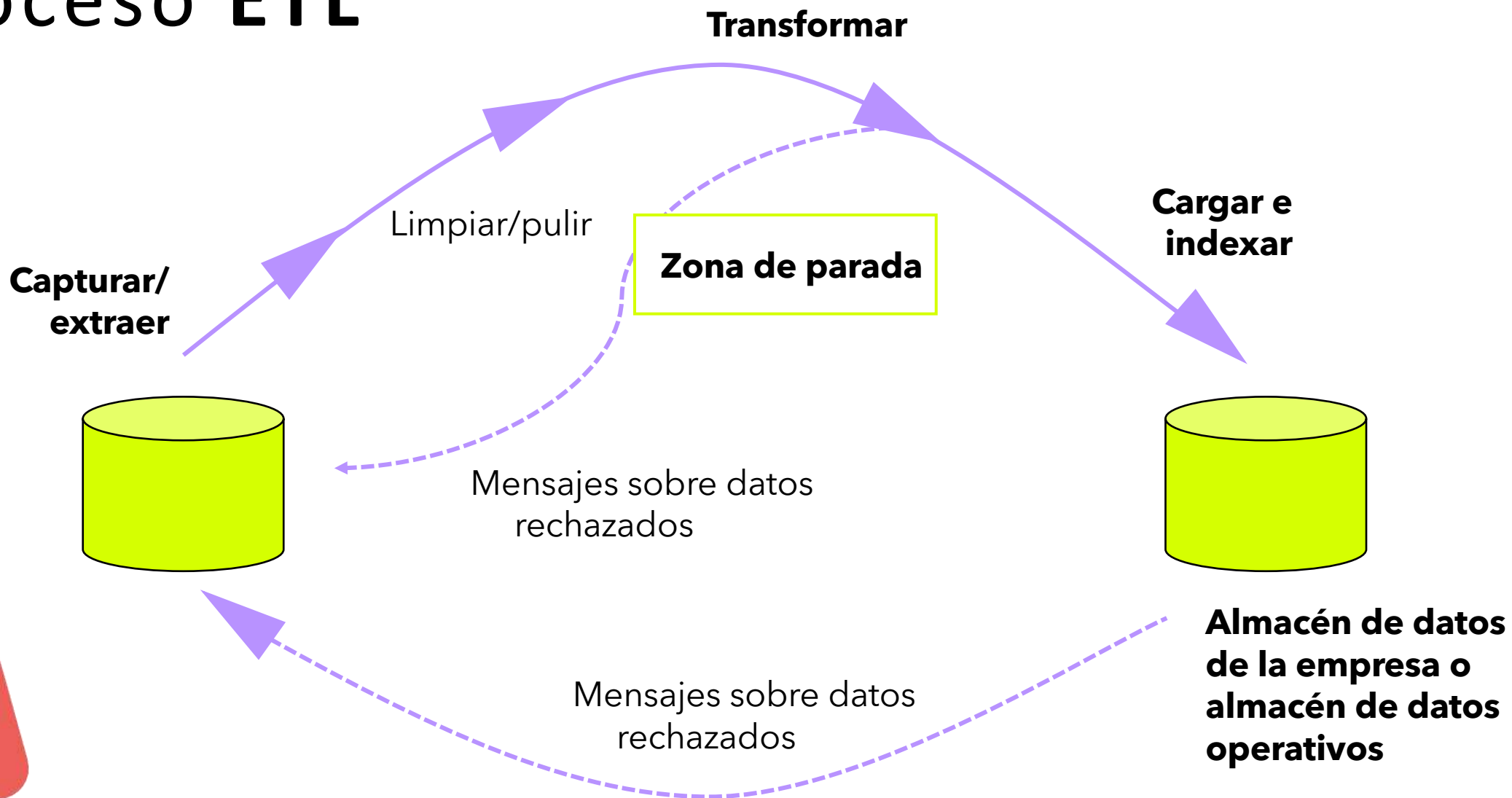
3

Transformación de datos

- **Es el paso principal donde el ETL añade valor.**
- En realidad, modifica los datos y proporciona orientación sobre si los datos pueden utilizarse para los fines previstos.
- Se realiza en el área de preparación.



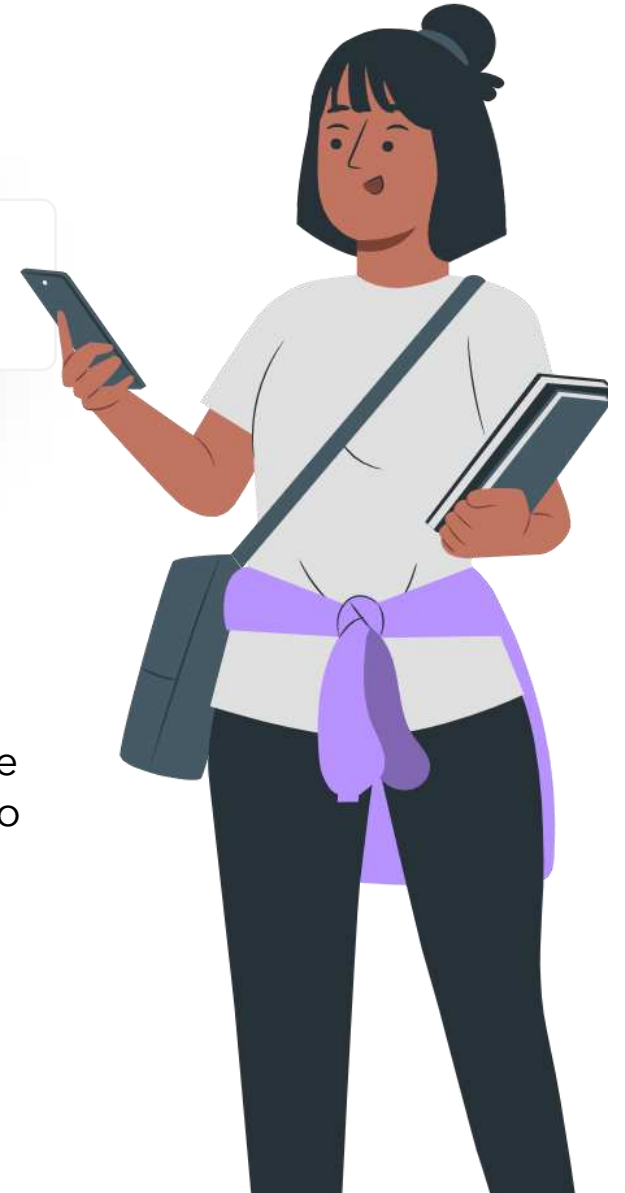
El proceso ETL



Staging o zona de parada

Significa que los datos se vuelcan simplemente a la ubicación para que luego puedan ser leídos tal como estaban por el siguiente procesamiento fase.

- La zona de parada se utiliza durante el proceso ETL **para almacenar los resultados intermedios** del procesamiento.
- **Permite reiniciar** al menos, **algunas de las fases** independientemente de las demás. Por ejemplo, si la fase de transformación falla, no debería ser necesario reiniciar el paso de Extracción.
- **Sólo el proceso ETL de carga puede acceder al área de staging.** Nunca debe estar disponible para nadie más, especialmente a los usuarios finales, ya que no está pensada para la presentación de datos al usuario final, ya que puede contener datos incompletos o en medio del proceso.



1

2

3

Transformación

Paradigma de la **calidad de los datos**:

CORRECTO

**SIN
AMBIÜEDADES
/ CLARO**

COHERENTE

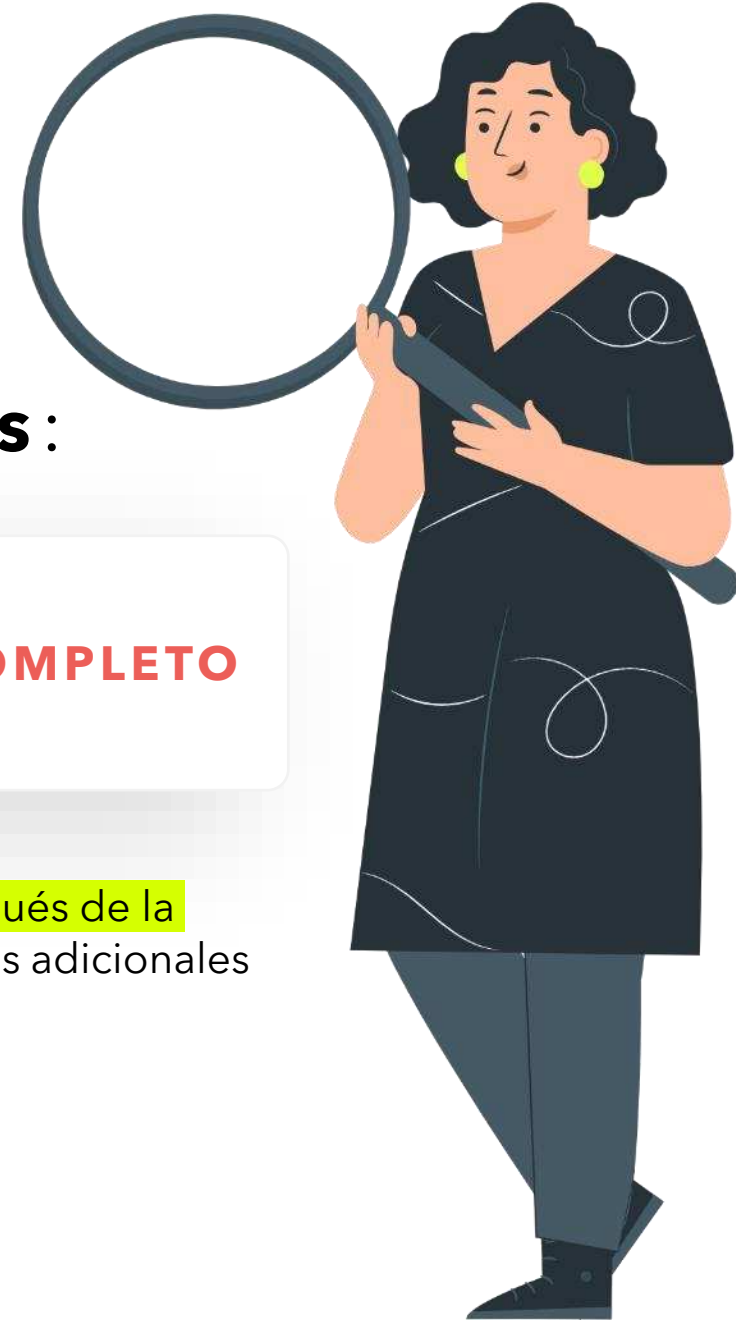
COMPLETO

Las comprobaciones de la calidad de los datos se realizan en dos lugares: **después de la extracción y después de la limpieza**, y en este punto se realizan comprobaciones adicionales de confirmación.

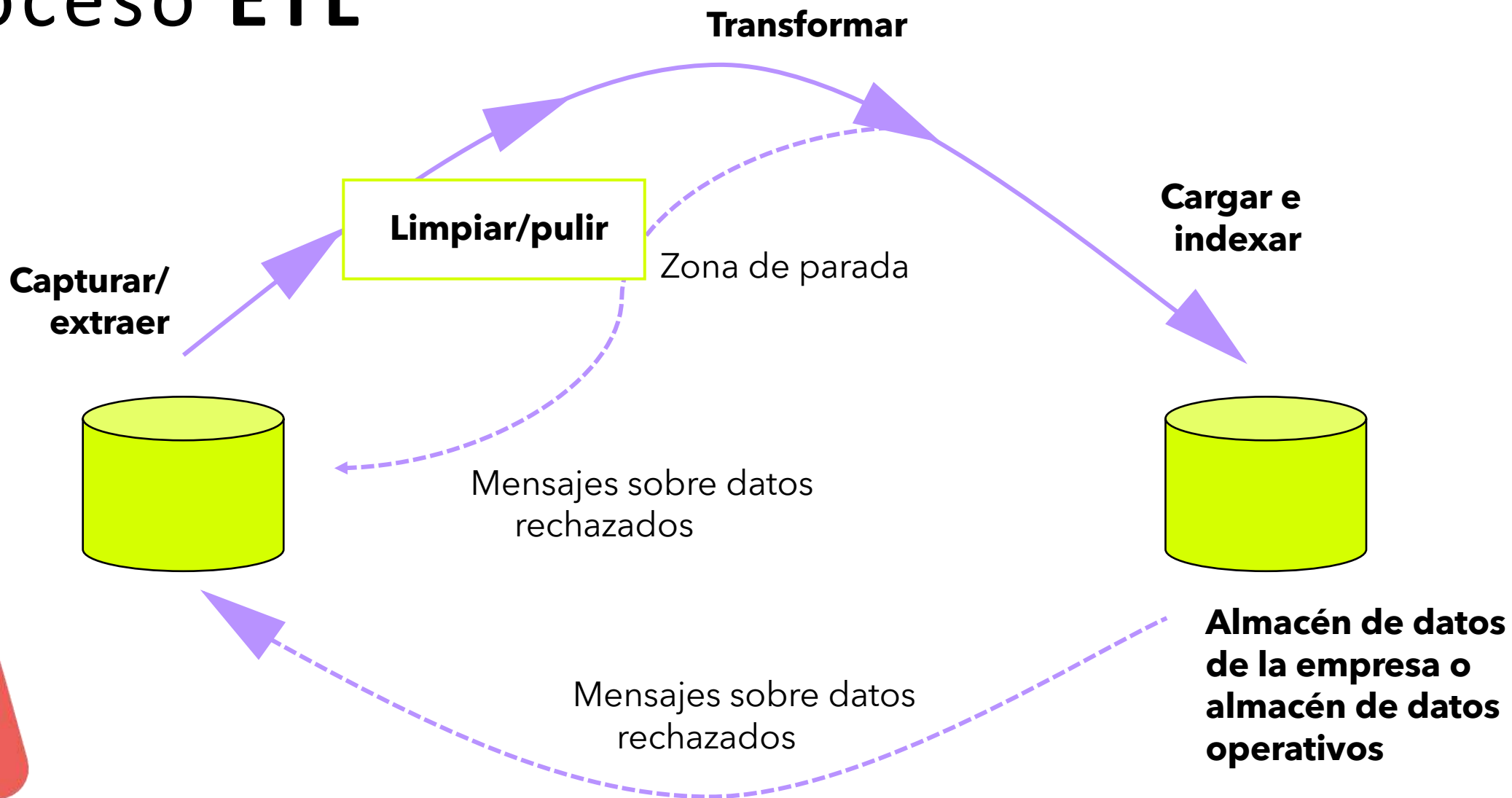
1

2

3



El proceso ETL



Transformación **Limpieza** de datos

Detección de anomalías

- **Muestreo de datos** - recuento(*) de las filas de una columna de departamento

Aplicación de la propiedad de la columna

- **Valores nulos** en las columnas.
- **Valores numéricos** que caen **fuera de los máximos y mínimos** esperados.
- Columnas cuya **longitud es excepcionalmente corta/larga**.
- **Columnas con determinados valores** fuera de los conjuntos de valores válidos discretos.



1

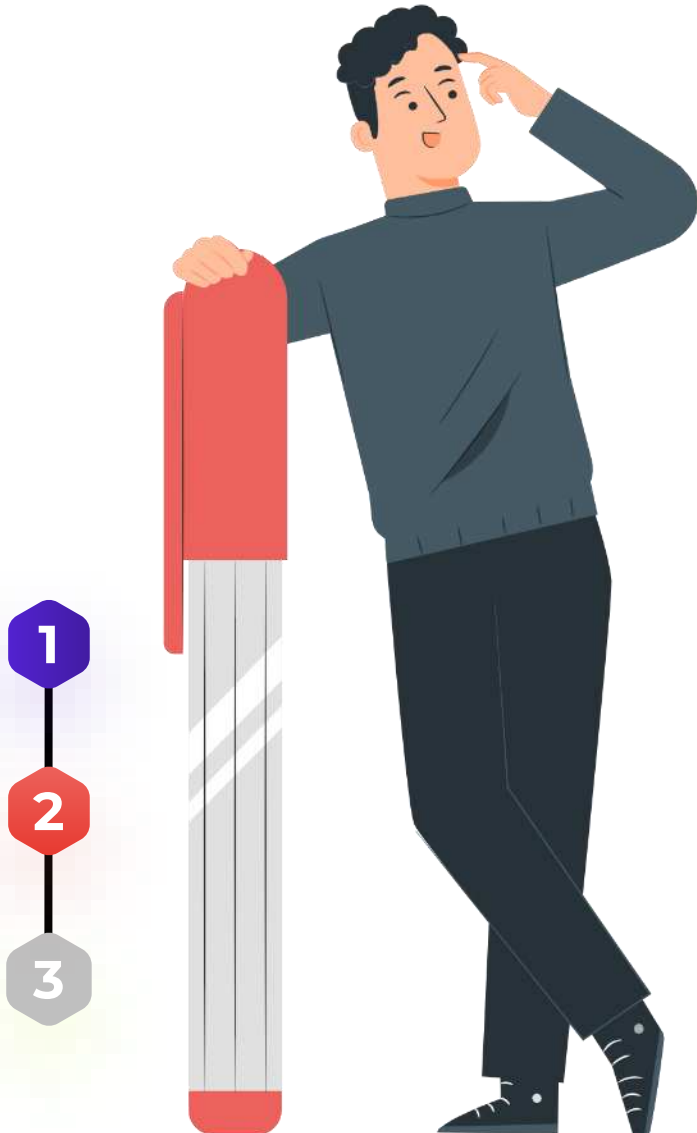
2

3

Limpieza de datos

La etapa de limpieza es una de las más importantes, ya que garantiza la calidad de los datos antes de transferirse al almacén de datos. La limpieza debe realizar reglas básicas de unificación de datos, como por ejemplo:

- ❧ **Hacer que los identificadores sean únicos** (las categorías de sexo Masculino/Femenino/Desconocido, M/F/nulo, Hombre/Mujer/No disponible se traducen al estándar Masculino/Femenino/Desconocido)
- ❧ **Convertir los valores nulos en un valor estandarizado** No disponible/No proporcionado
- ❧ **Convertir números de teléfono, códigos postales a una forma estandarizada.**
- ❧ **Validar los campos de dirección, convertirlos en una nomenclatura adecuada,** por ejemplo, Calle/St./Str.
- ❧ **Validar los campos de dirección entre sí** (Estado/País, Ciudad/Estado, Ciudad/Código postal, Ciudad/Calle).



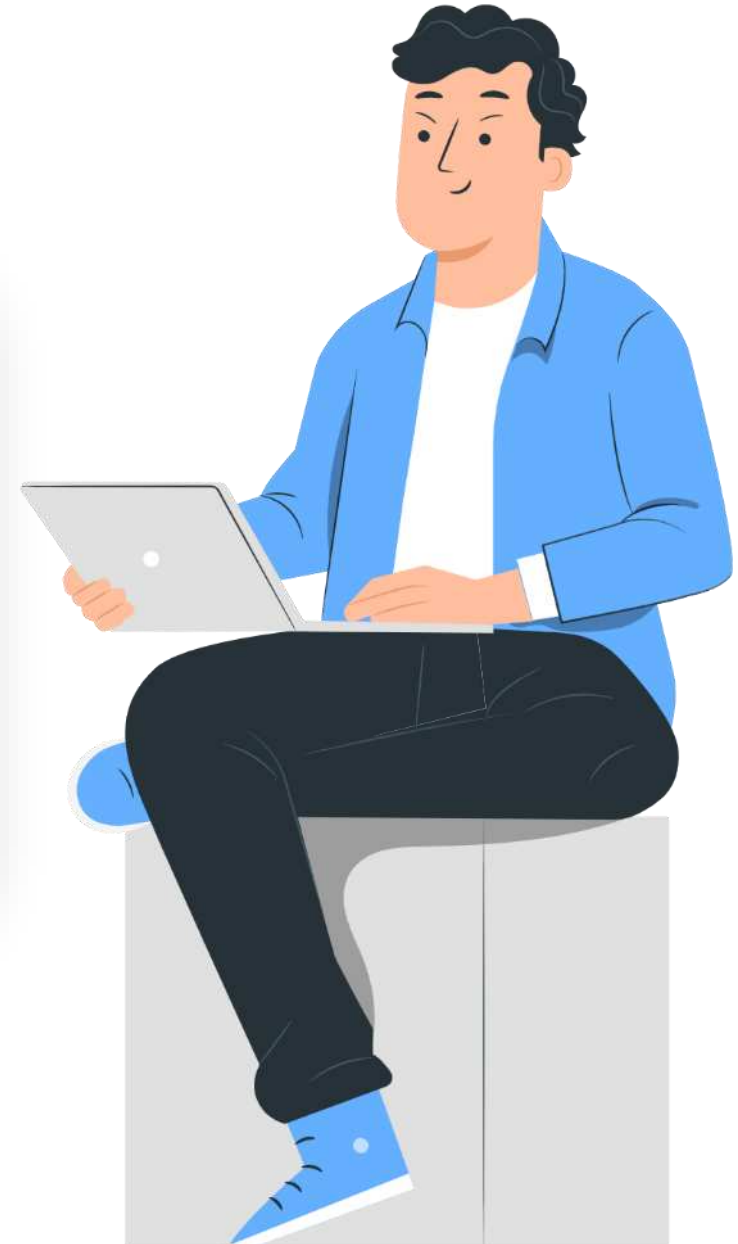
Limpieza de datos

Corrección de errores:

Faltas de ortografía, fechas erróneas, uso incorrecto de los campos, direcciones no coincidentes, datos que faltan, datos duplicados, incoherencias

También:

Descodificación, reformato, sellado de tiempo, conversión, generación de claves, fusión, detección/registro de errores, localización de datos perdidos



1

2

3

La depuración lleva al proceso ETL



Transformación- Confirmación

✓ Aplicación de la estructura

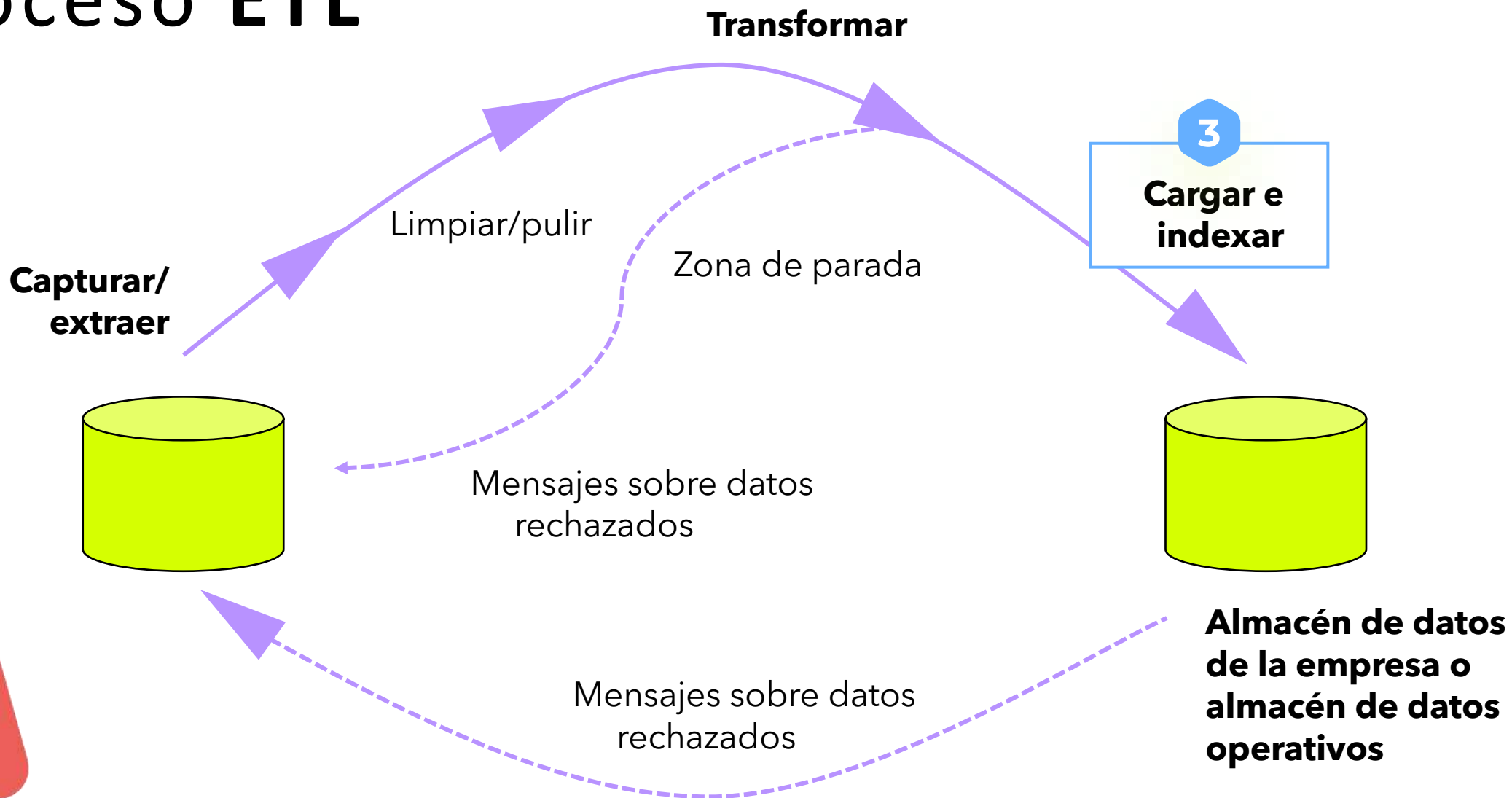
- Las tablas tienen claves primarias y foráneas adecuadas
- Obedecer la integridad referencial

✓ Cumplimiento de los valores de los datos y las reglas

- Reglas de negocio (business rules) simples
- Comprobaciones lógicas de datos



El proceso ETL



1

2

3

Carga de datos

El objetivo del proceso de carga suele ser una base de datos. Para que el proceso de carga sea eficiente, es útil deshabilitar las restricciones e índices antes de la carga y volver a habilitarlos sólo después de que se complete la carga. La integridad referencial debe ser mantenida por la herramienta ETL para asegurar la consistencia.

COMPLETA

Es la carga de todo el volcado de datos que tiene lugar la primera vez. En este caso damos la última fecha de extracción como vacía para que se carguen todos los datos.

INCREMENTAL

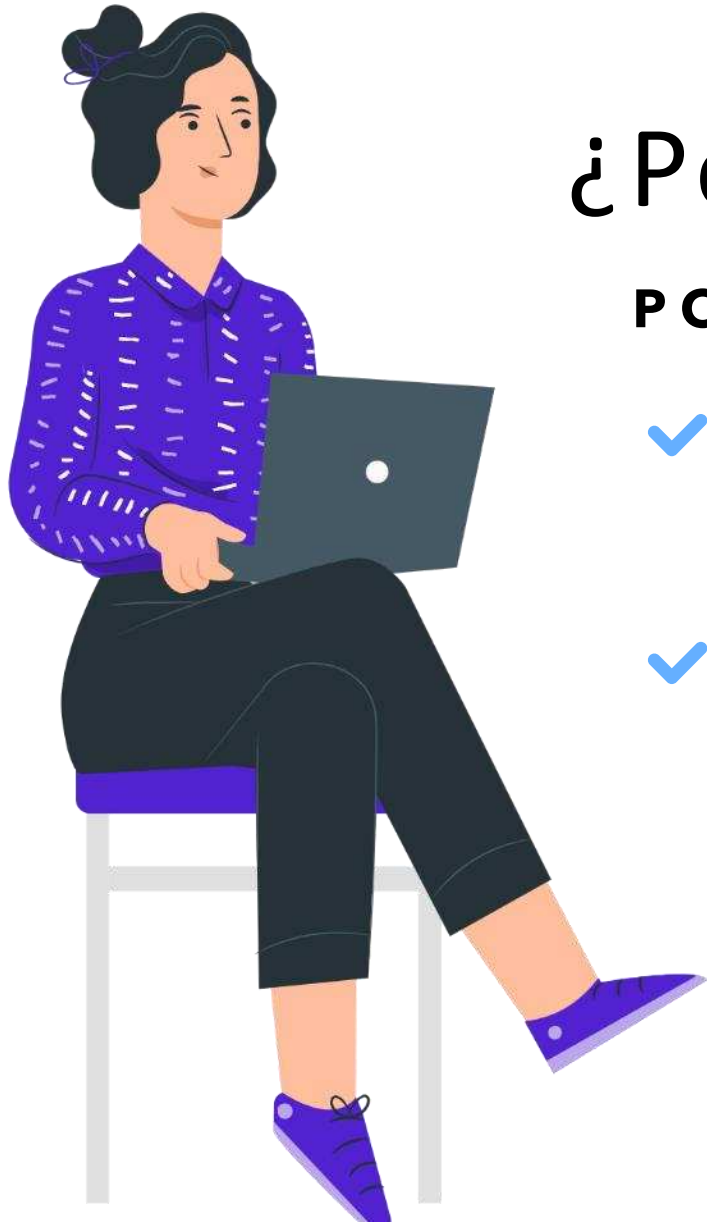
Donde la diferencia entre los datos de destino y de origen se vuelca a intervalos regulares. Aquí damos la última fecha de extracción para que sólo se carguen los registros posteriores a esta fecha.



¿Por qué **incremental**?

POR LA **VELOCIDAD**

- ✓ Optar por realizar una carga completa en conjuntos de datos de gran tamaño **requerirá una gran cantidad de tiempo** y otros recursos del servidor.
- ✓ Lo ideal es que **todas las cargas de datos se realicen durante la noche** con la expectativa de completarlas antes de que los usuarios puedan ver los datos al día siguiente. La ventana nocturna puede no ser suficiente para completar la carga completa.



1

2

3

Carga completa **vs** Carga incremental

CARGA COMPLETA

Trunca todas las filas y carga desde cero

Requiere más tiempo

Se puede garantizar fácilmente

Puede perderse

CARGA INCREMENTAL

Registros nuevos y actualizados se cargan

Requieren menos tiempo

Difícil. El ETL debe comprobar las filas nuevas/ actualizadas

Se mantiene

1

2

3

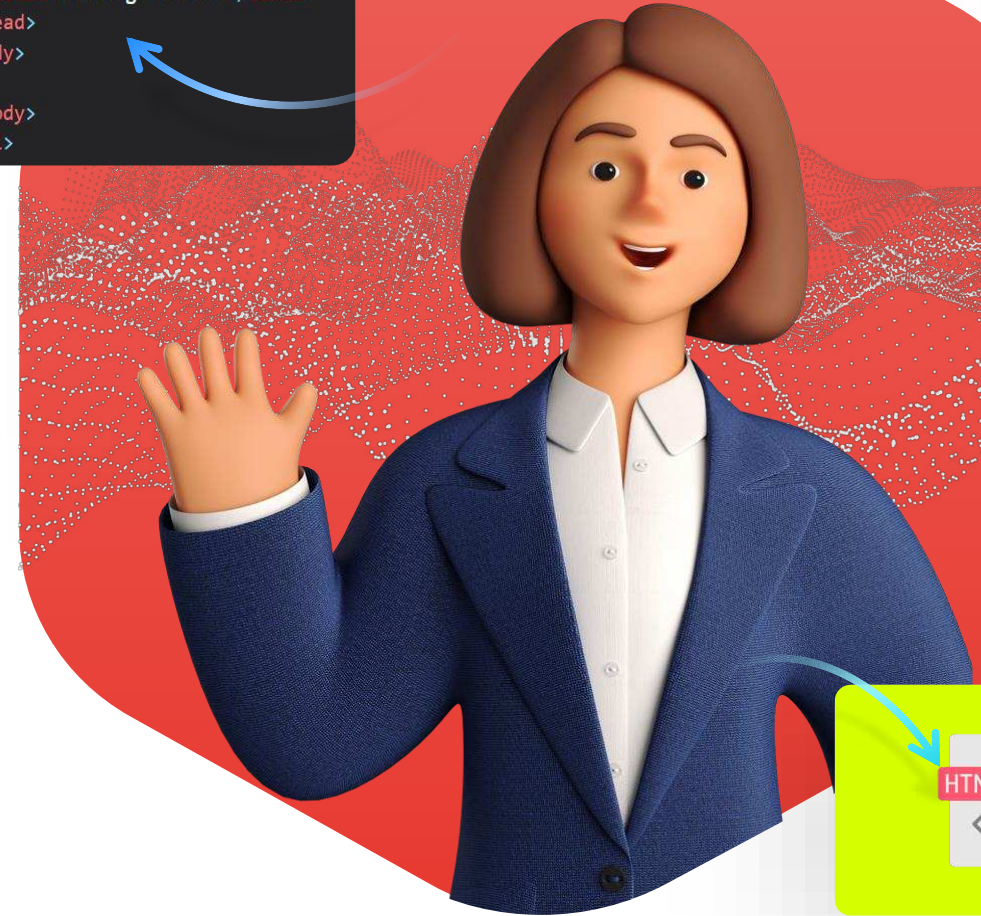


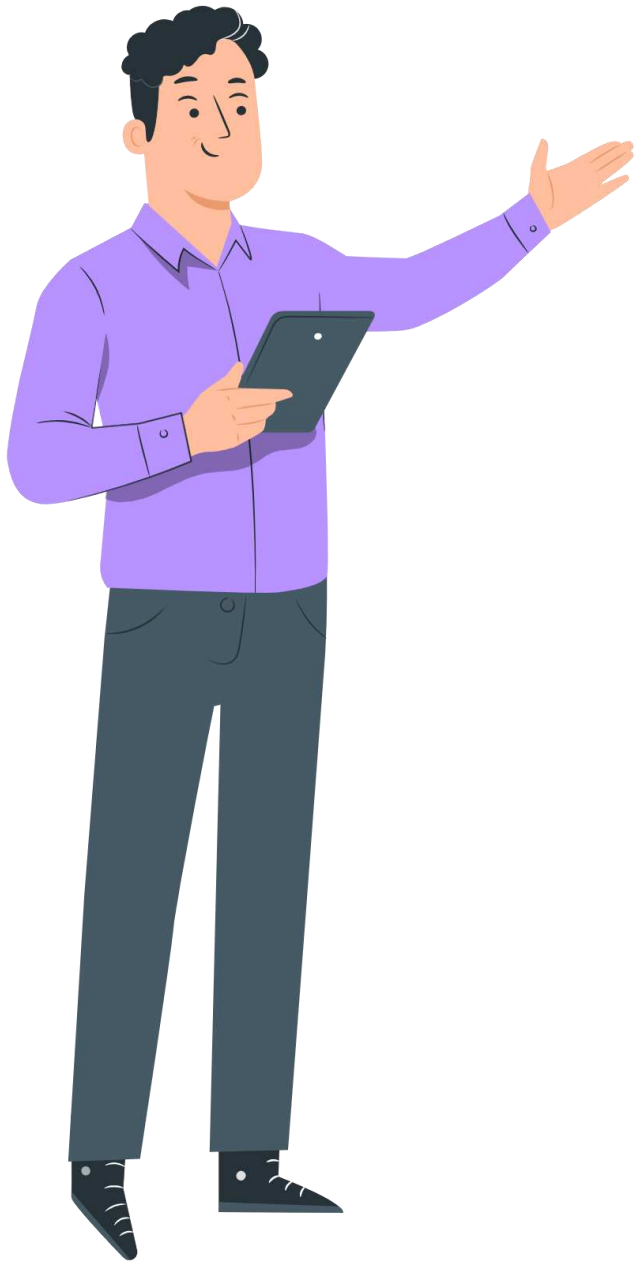
5.2

Conocer cual son
elementos clave para
el desarrollo de un
adecuado gobierno
del dato

COMENZAR

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="UTF-8">
5     <title>Title goes here</title>
6   </head>
7   <body>
8
9   </body>
10 </html>
```

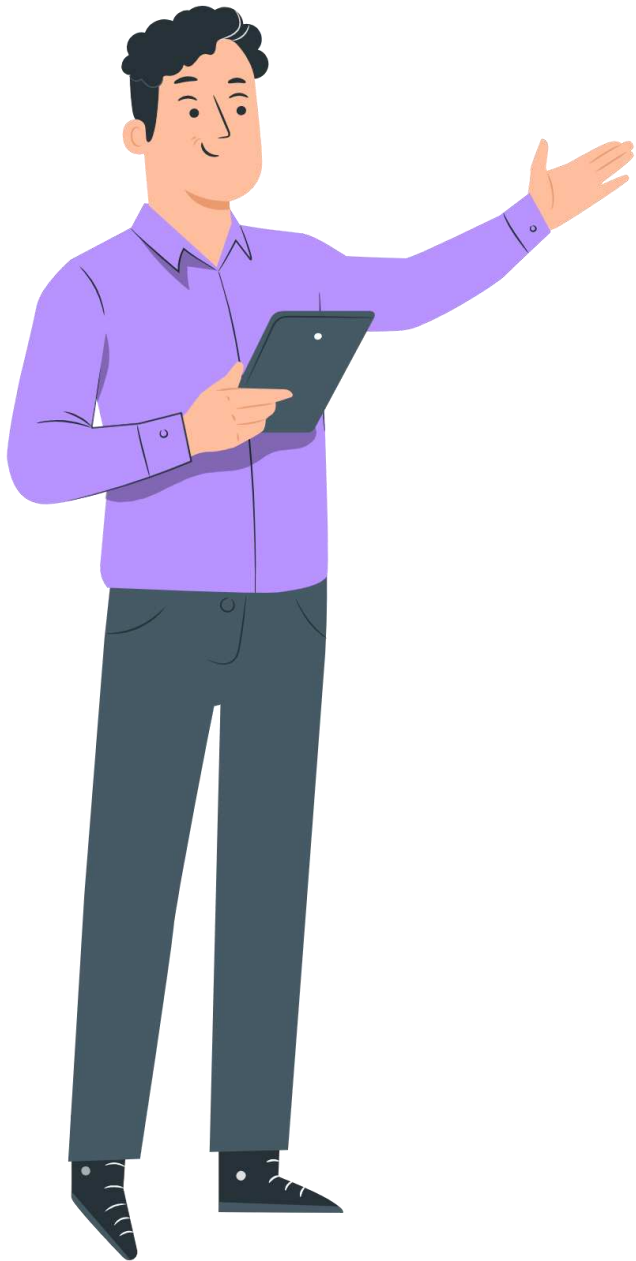




¿Qué es **inteligencia de negocios?**

En una sociedad en donde estamos inundados de datos, las compañías precisan de sistemas que les permitan analizar la ingente cantidad de información y transformarla en algo útil y de valor.

La inteligencia de negocios o Business Intelligence responde, precisamente a esta necesidad, pretendiendo ofrecer modelos predictivos en base a hechos históricos que brinde a las empresas un mejor posicionamiento competitivo y una mayor satisfacción de sus clientes.



Este concepto crítico no es nuevo y desde su definición en los años 60, su evolución ha sido constante.

BI no se puede definir como una tecnología o como una herramienta. Su definición es más amplia.

Se entiende por Business Intelligence el conjunto de tecnologías, metodologías y estrategias enfocadas a tratar los datos que tiene una empresa y convertirlos en conocimiento útil que permita tomar mejores decisiones y mejorar su eficacia y competitividad.

Introducción

03:29



Algunas de las tecnologías que
forman parte de

**Business
Intelligence** son:



- ORIGEN DE LOS DATOS
- CARGA DE DATOS EN UN REPOSITORIO CENTRAL
- TRANSFORMACIÓN DE DATOS
- EXTRACCIÓN DE DATOS
- DATAWARE HOUSE
- MINERIA DE DATOS Y MODELADO
- OLAP - OLTP
- PRESENTACIÓN DE LA INFORMACIÓN
- INFORMES
- CUADROS DE MANDO

NECESIDAD DE LA Inteligencia de Negocios

Estas son algunas situaciones en las que la implantación de un sistema de Business Intelligence es necesaria:

- 1 Toma de decisiones sin fundamento.
- 2 Problemas de comunicación dentro o fuera de la empresa.
- 3 Uso masivo y descentralizado de tablas Excel.
- 4 Enfoque de regresión múltiple para el análisis de la varianza.



5

La información no fluye entre departamentos o llega duplicada.

6

Hay silos de información.

7

Tareas de ventas y marketing ineficaces.

8

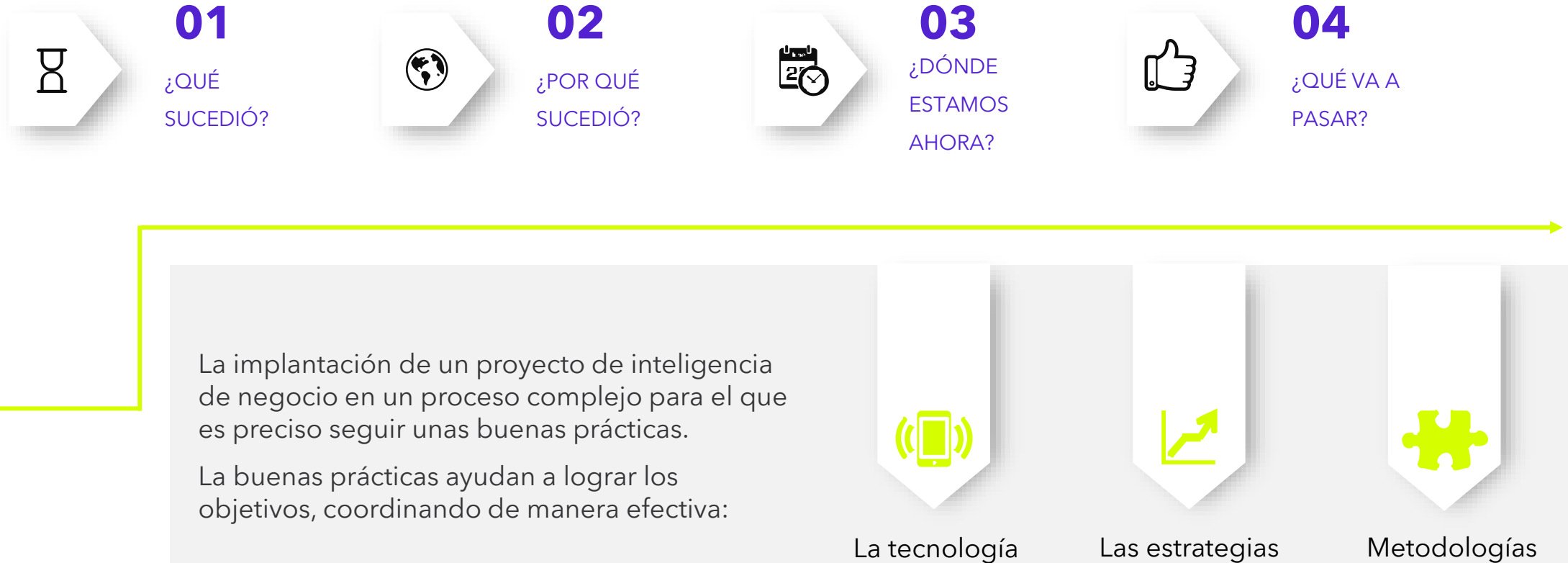
Volumen excesivo de información que la hace inmanejable.

9

Procesos manuales.



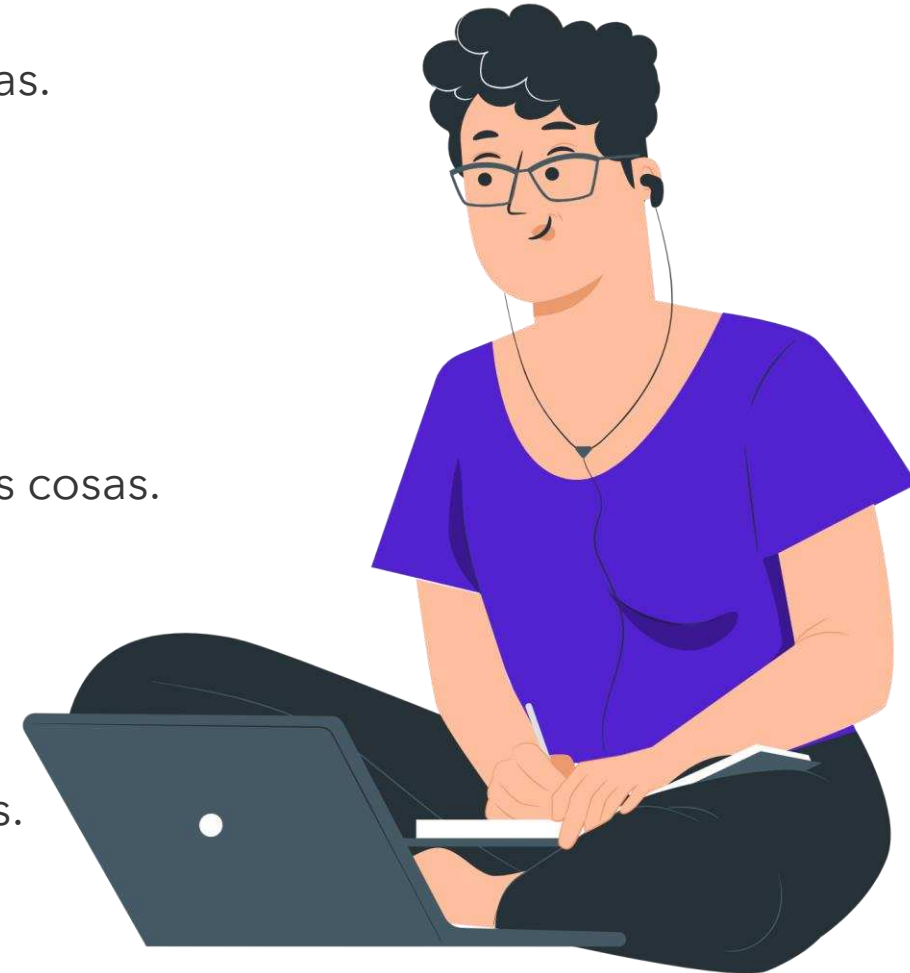
La implantación de un sistema de **Business Intelligence** responde a preguntas como:



¿Cuándo no existe una estrategia de BI?

Es posible detectar que no existe una estrategia cuando detectamos algunos de los siguientes puntos dentro de una organización:

- ✓ El departamento de informática es la cusa de todos los problemas.
- ✓ El BI no es crítico para la dirección.
- ✓ BI se considera lo mismo que reporting.
- ✓ Siguen existiendo información descentralizada.
- ✓ No es posible medir resultados ni saber el porqué sucedieron las cosas.
- ✓ DataMining es lo mismo que BI.
- ✓ No se invierte ni en tecnología ni en RRHH expertos en BI.
- ✓ Se piensa que la BI es sólo para directivos o mandos intermedios.



La implantación de una estrategia de inteligencia de negocio

es un proceso a largo plazo y complejo que implica múltiples departamentos, procesos y recursos entre los que es conveniente destacar:



**Medir los resultados de
aplicaciones analíticas**

03

Mediante herramientas de
datamining o similares que
faciliten la toma de decisiones.

Revisando casos de estudio y
consultando a las empresas
del sector para determinar
qué ha funcionado y qué no.

04

**Aprender de los
éxitos y fracasos**

**Evangelizar la
organización**

05

Concienciar a todo el
personal, desde la dirección
hasta el último empleado.

Este tipo de proyectos suponen una transformación cultural de toda la organización en donde es necesario tener respuesta a alguna de las siguientes preguntas:



¿Qué problemas o necesidades de negocio se busca resolver?

¿A qué en particular se debe dar respuesta y con qué prioridad?

¿De qué manera obtenemos respuesta actualmente?

¿Qué fuentes de datos y desde qué departamentos son necesarias? (marketing, operaciones, recursos humanos, etc.)?

¿Cómo es la calidad de los datos?

¿Qué cantidad de datos debe ser guardada como histórico?

¿Con qué frecuencia deben estar actualizadas los datos?

¿QUÉ ES EL EDA?

El **análisis exploratorio de datos** es un enfoque de análisis de datos para revelar las características importantes de un conjunto de datos, principalmente a través de la visualización.

Hay que conocer bien los datos:

- ✓ Distribuciones (simétrica, normal, sesgada).
- ✓ Problemas de calidad de los datos.
- ✓ Valores atípicos.
- ✓ Correlaciones e interrelaciones.
- ✓ Relaciones funcionales.
- ✓ Atributos derivados, claves como la primaria, claves foráneas, etc.
- ✓ Atributos estáticos, atributos dinámicos, etc.



Introducción al Análisis Exploratorio de Datos (EDA)



1

El análisis exploratorio de datos se refiere al proceso de realizar investigaciones iniciales sobre la naturaleza de los datos para identificar patrones, interceptar anomalías, evaluar hipótesis, y chequear asunciones con la ayuda de estadísticos y herramientas de representación gráfica.

2

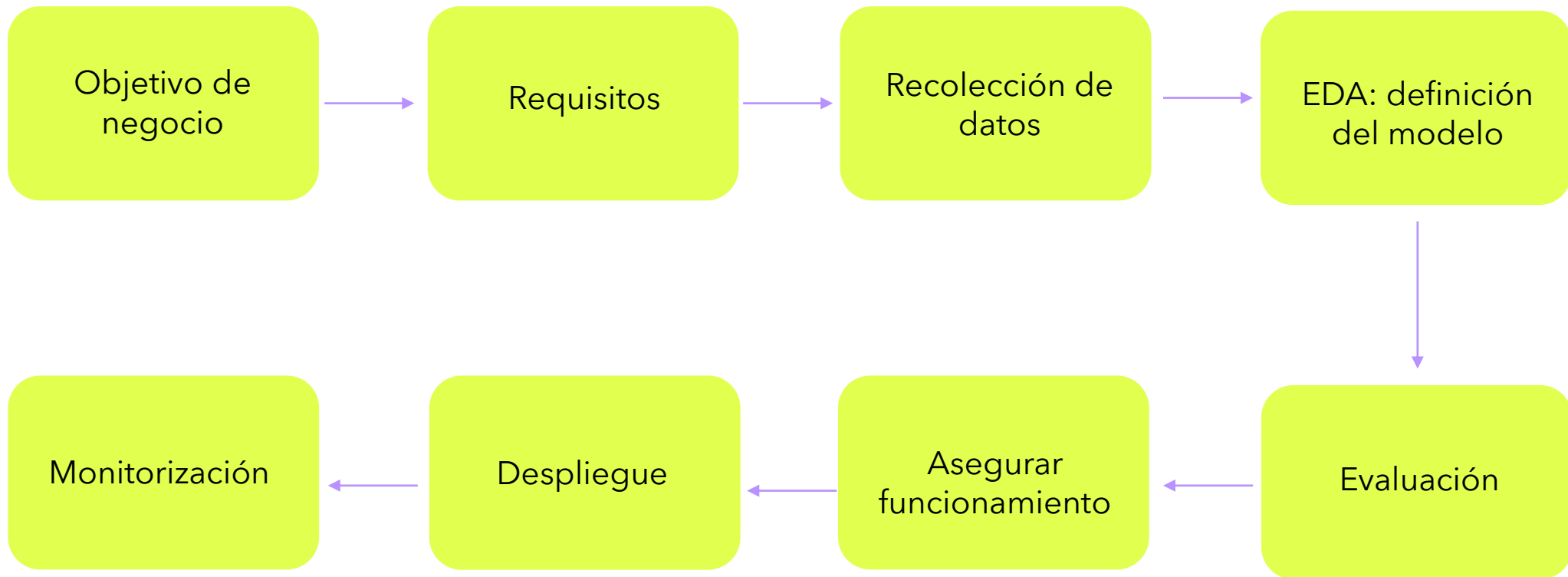
Es una práctica muy recomendada para comprender la naturaleza de los datos y tratar de extraer tanta información como se pueda de inicio.



Importancia del EDA

- ✓ Encontrar las características y variables más relevantes del dataset de datos.
- ✓ Testear hipótesis o probar asunciones sobre las características del conjunto de datos.
- ✓ Comprobar la calidad de los datos y la información para el procesamiento posterior.
- ✓ Proporcionar insights o conclusiones relevantes sobre los datos a los interesados.
- ✓ Verificar las relaciones existentes en los datos.
- ✓ Buscar patrones atípicos en la estructura de los datos.

Proceso del **Análisis Exploratorio de Datos (EDA)**



Dos categorías **de datos**



**Datos estructurados: archivo CSV,
archivo Excel, Base de Datos, etc.**



**Datos no estructurados: vídeo, imagen,
sonido, etc.**



Obtén una idea general de los datos

1

Asegúrate de que tu primera visualización esté basada en datos (sin modelos).

2

Piensa en lo interactivo y lo visual:

- **Los humanos son los mejores reconocedores de patrones**
- Utiliza tantas dimensiones como le permitan sus datos 2, 3
 - x,y,z, espacio, color, tiempo....

3

La visualización es útil en las primeras etapas de la minería de datos:

- detectar valores atípicos (por ejemplo, evaluar la calidad de los datos)
- comprobar los supuestos (por ejemplo, ¿distribuciones normales o sesgadas?)
- identificar datos brutos útiles y transformaciones (por ejemplo, $\log(x)$)

Conclusión: ¡siempre merece la pena examinar los datos!

Fundamentos de la visualización efectiva de datos

Estos son los cuatro pilares a la hora de realizar una visualización efectiva de datos:

Cuestiones de calidad de datos.

Una buena comprensión de las teorías estadísticas.

Cómo mover volúmenes de datos.

¿Debo utilizar el aprendizaje automático?

Problemas de la calidad de datos

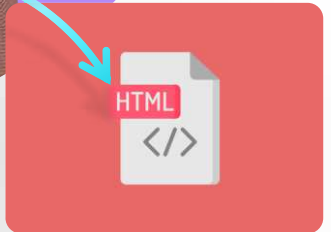


5.3

Identificar los roles y responsabilidades necesarios para la gestión de datos

COMENZAR

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="UTF-8">
5     <title>Title goes here</title>
6   </head>
7   <body>
8
9   </body>
10 </html>
```



El ingeniero de datos es el perfil que recibe los datos desde el exterior. Datos normalmente no trabajados y desordenados sin formato, y su principal tarea consiste en trabajar estos datos introduciéndolos en la base de datos correspondiente para dejarla preparada de cara al científico de datos y el analista de datos.

datos

El científico de datos, gracias a la arquitectura tecnológica provista por el arquitecto de Big Data y gracias a los datos provistos por el ingeniero de datos, trabajará mediante técnicas basadas en I+D. Principalmente su función está establecida en torno a una estrategia a más largo plazo de obtención de insights o conclusiones que no son necesarias a nivel de negocio en el corto plazo.

CIENTÍFICO
DE DATOS

ARQUITECTO DE
BIG DATA

ANALISTA
DE DATOS

ESPECIALISTA
EN IA

El analista de datos es el perfil más business, y es quien procesa los datos procedentes del ingeniero de datos y genera mediante todas las herramientas de visualización las gráficas, así como todas las métricas KPIs que van a ser utilizadas por el personal directivo para la toma de decisiones.

insights
decisiones

El especialista en IA cuya principal función es la generación de herramientas predictivas para que a través de los insights procedentes del científico de datos y a partir de los datos procedentes del ingeniero de datos que obtiene por medio del científico de datos, pueda generar estructuras algorítmicas predictivas que ayuden en la toma de decisiones que el analista de datos ha de proporcionar a los directivos de la empresa en la toma de decisiones correspondiente.

hemos
terminado

¡EXCELENTE TRABAJO!

