

Data Manipulation

Índice

- 01 ¿Qué es Data Manipulation?
- 02 Proceso ETL
- 03 Casos prácticos



01

¿Qué es Data Manipulation?

Data Analytics utiliza datos que comúnmente fueron recolectados con otro fin, lo que podrá atentar contra la **calidad** de nuestros datos.



La mala calidad en
nuestros datos nos
puede llevar a una
mala **calidad** en los
resultados de
nuestros análisis.



Data Manipulation

Es un proceso continuo donde buscamos asegurar la **calidad** de los datos. De esta forma asegura los mejores resultados, por lo que podremos volver a cada instancia para hacer los ajustes necesarios.

Las principales tareas que haremos en este proceso son:

- **Extracción:** Distintas bases o tablas que nos darán los datos necesarios para analizar.
- **Observación:** Entendimiento de cada variable para planificar la transformación.
- **Transformación:** Se aplicarán distintas técnicas que nos permitirán “limpiar” nuestros datos, dejando de lado aquellos que contienen “ruido” (que no reflejan la realidad o pueden llevarnos a conclusiones incorrectas).
- **Carga:** Dejamos el dataset listo para aplicar su análisis.
- **Control:** Auditar que los datos que necesitamos para el análisis sean de calidad para ser analizados.

02

Proceso ETL

Proceso ETL

Es el proceso en el cual consultamos los datos que necesitamos, si lo requieren podemos transformarlos y, por último, los cargamos en un conjunto de datos (dataset). Ahora ya están listos para aplicar a un modelo de Data Analytics,

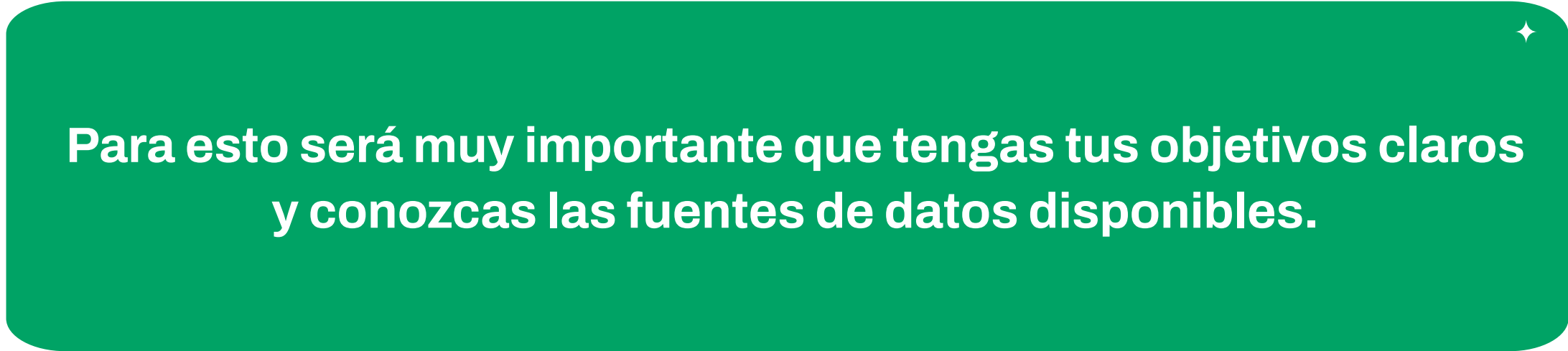
Los pasos del proceso son:

- ➔ **E:** Extract / Extracción
- ➔ **T:** Transform / Transformación
- ➔ **L:** Load / Carga

1. Extraer

Proceso en el cual consultamos los datos de las fuentes de datos origen.

El principal objetivo de esta instancia es tomar únicamente aquellos datos que vamos a necesitar para nuestro análisis.



Para esto será muy importante que tengas tus objetivos claros y conozcas las fuentes de datos disponibles.

2. Transformar

Una vez que seleccionamos y observamos los datos, llega el momento de **transformarlos**, es decir aplicar toda adaptación o modificación para tener los datos listos para su análisis.

Algunos ejemplos de transformación son:

- **Limpieza:** Completar datos faltantes o valores outliers.
- **Transformación:** Agregar datos sumariados (ej. Promedio país).
- **Integración:** Combinar datos de distintas fuentes.
- **Reducción:** Minimizar el volumen de datos, manteniendo el mismo sentido analítico.
- **Discretización:** Similar a la reducción, pero sobre datos numéricos.

3. Cargar

Momento en el que cargamos los datos, creando un dataset integrado para ser analizado. Es la última etapa antes de pasar a la aplicación de nuestro modelo de análisis de datos.



Es sumamente importante que hagamos un control para comprobar que los datos integrados y transformados no tengan ningún error.

03

Casos prácticos

1. Extraer: ¿Qué nos puede pasar a la hora de consultar datos?

No todos los datos van a ser fáciles de consultar ya que pueden provenir de múltiples **bases de datos**, con distintas **plataformas** y **protocolos**. A su vez tendrán distintos lenguajes y juegos de caracteres. Estas complicaciones y muchas más pueden surgir a la hora de extraer datos.

Por este motivo será muy importante **conocer las fuentes, asegurarnos que sean confiables** y luego trabajar en la **calidad de los datos** de cada una, entendiendo qué tipo de datos son y qué transformaciones serán necesarias.

Por ejemplo, si queremos analizar los sueldos de nuestros empleados, vamos a necesitar acceder a las bases internas de cada empleado y cruzarlo con bases macroeconómicas de cada mercado. Cada una de estas bases tiene distintos orígenes y han sido creadas con distintos fines, por lo tanto es muy probable que requieran transformaciones para ser integradas y analizadas.



1. Extraer: Consejos prácticos



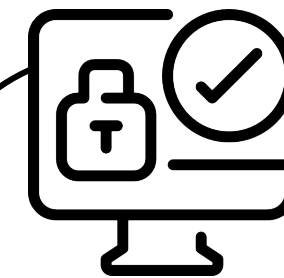
Dueños de las bases

Siempre resulta útil tomarse un tiempo para conocer a los dueños de las bases, de esta manera podremos contextualizar mejor los datos .



Cómo será la integración

Debemos tener presente cómo vamos a integrar las distintas bases de datos, validar qué variable será el ID de cada tabla.



Conocer los tipos de datos

Como hemos visto, saber qué tipos de datos estamos viendo contribuye a una mejor contextualización y planificación del proyecto.

Un valor **outlier** es un dato extremo que carece de sentido en nuestra base. Veamos un ejemplo...

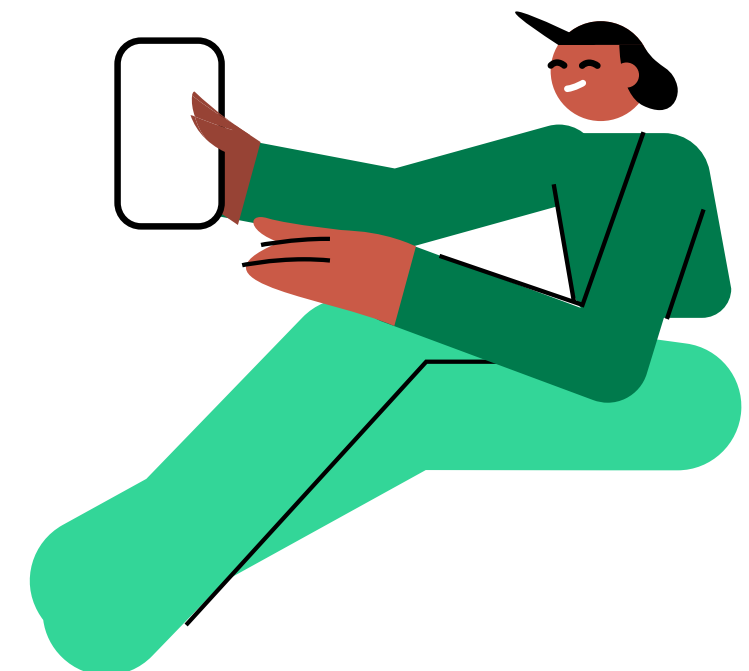


2. Transformar: ¿Cómo se ve un valor outlier?



2. Transformar: ¿Cuáles son las transformaciones más usuales?

- **Cambiar tipos de datos:** Muchas veces los datos numéricos están expresados en texto o tienen un separador de miles distinto al nuestro.
- **Trasponer:** Sucede cuando tenemos los datos expresados en una matriz, usualmente es mejor pasarlo a una tabla plana.
- **Valores nulos:** Completar valores vacíos.
- **Reducción:** Minimizar el volumen de datos, manteniendo el mismo sentido analítico.
- **Discretización:** Similar a la reducción, pero sobre datos numéricos.



3. Cargar: Momentos finales antes de pasar al análisis de datos

En esta instancia vamos a crear el dataset que será el input de nuestro modelo de datos. Por eso es muy importante que en esta etapa controlemos el proceso de Data Manipulation y si todo se ejecutó correctamente pasar a la carga de los datos.

Pensemos:

1. ¿Las transformaciones fueron ejecutadas correctamente?
2. ¿Existen valores faltantes o outliers luego de la transformación?
3. ¿La base de datos está completa o alguna variable perdió registros?
4. Luego de la carga: ¿Están todos los datos que necesitamos?



