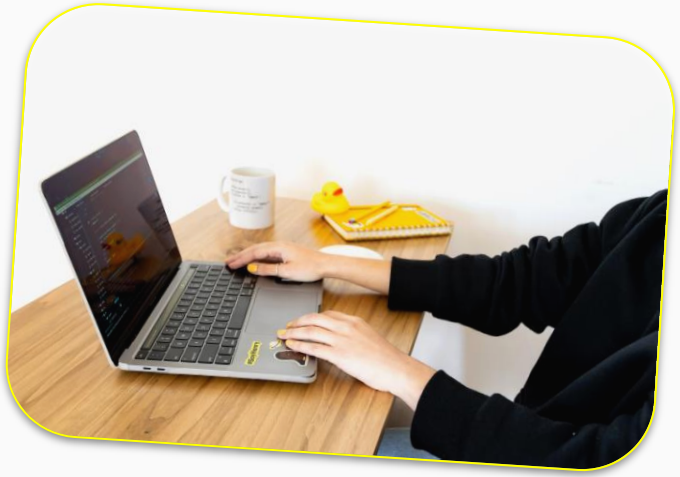
HENRY

Calidad del dato

Data Science

Agenda



- → **Calidad** de los datos
- → **Mala calidad** en los datos
- → **Criterios** para la calidad de los datos

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

- → **Comprender** el concepto de la **Calidad del Dato** y sus implicaciones.
- → **Conocer** el concepto del **Modelado de Datos**.

Calidad de los datos

¿Para qué?

La **calidad de los datos** es fundamental para asegurar la confiabilidad de los análisis.

El uso de **datos no confiables** puede llevar a conclusiones erróneas y decisiones incorrectas.

HENRY

- Fuentes críticas de datos
- Métodos para garantizar la calidad
- Recuperación de información perdida o incompleta
- Resolución de conflictos en los datos

Causas de la mala calidad de los datos

HENRY

- Carga de datos en forma manual o Data Entry.
- Carga de datos externos sin los recaudos correctos para su adecuación.
- Problemas de carga originados en los sistemas transaccionales utilizados como fuente de datos.
- Implementación de nuevas aplicaciones en la organización, implica nuevos orígenes de datos, que necesitan ser congruentes con los datos ya existentes.
- Cambios en las aplicaciones existentes o migraciones de sus bases de datos.

Criterios para garantizar la calidad del dato

LOS CRITERIOS SON

- Actualización
- Fiabilidad
- Consistencia
- Completitud
- Accesibilidad

Criterios de actualización

- Los datos deben estar **actualizados**. Debe existir referencias de la fecha de confección o de la fecha de última actualización.
- **Por ejemplo:** Información de deuda sin una referencia en cuanto a la fecha de actualización.

Criterios de completitud

- Los datos deben estar **completos**. Puede parecer obvio pero es una de las situaciones más habituales.
- **Por ejemplo:** tablas con datos filiatorios y de contacto con campos vacíos aleatoriamente.

Criterios de fiabilidad

- La **procedencia** y la **trazabilidad** del dato son características que hacen a la fiabilidad.

Criterios de accesibilidad

- Los datos deben ser accesibles con **bajo nivel de esfuerzo.**
- Deben estar en **lugares previsibles** y ser fácilmente ubicables y elegibles.
- **Ejemplo**: Una tabla con nombres de campos numerados: Campo1, Campo2, etc...
- **Ejemplo 2:** Un reporte que se aloja en una ubicación poco habitual.

Experiencia y conocimiento

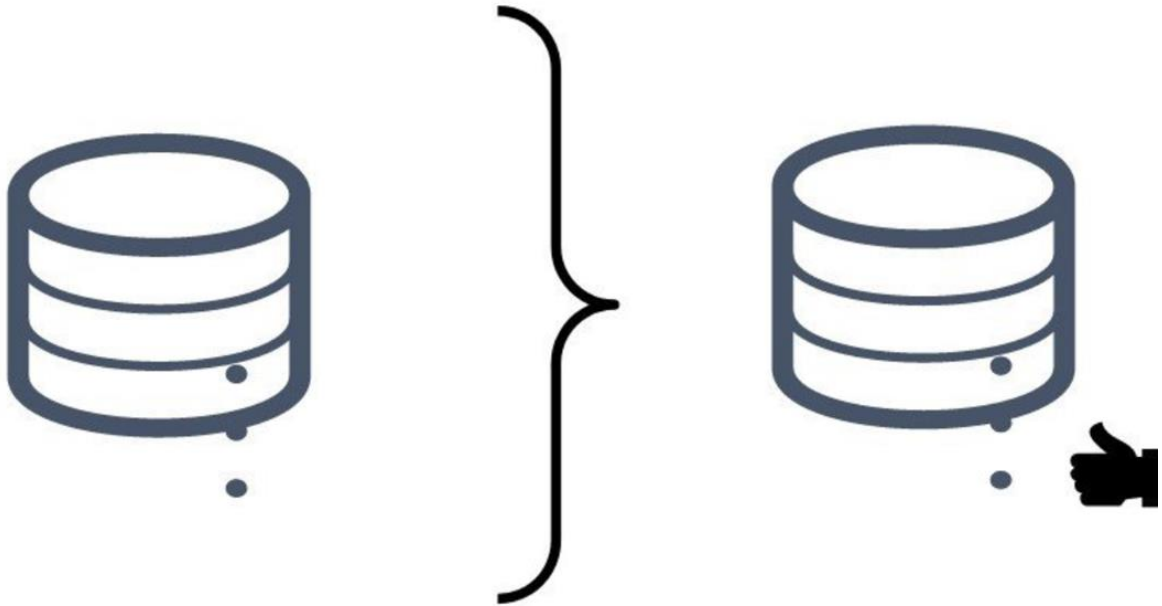
- **Interna**: Calidad de caracteres y de lo que se guarda en los campos.
- **Externa**: Calidad de interdependencia y racionalidad de los campos.
- **Ejemplo**: Las claves primarias y las claves foráneas deben ser consistentes y permitir la relación entre tablas.

Preparación de los datos

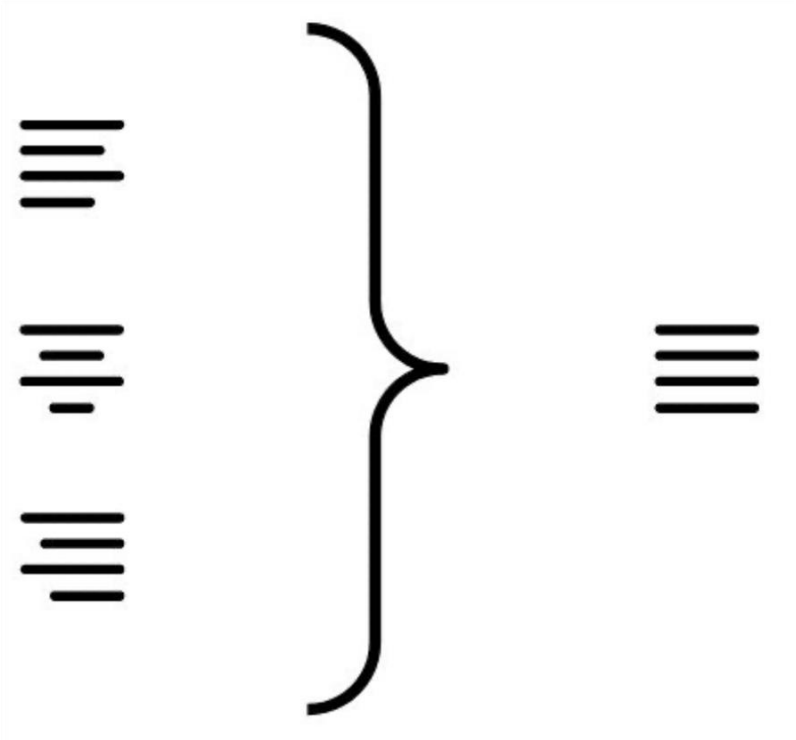
- Integración
- Limpieza del ruido
- Transformación
- Imputación de valores faltantes

Integración

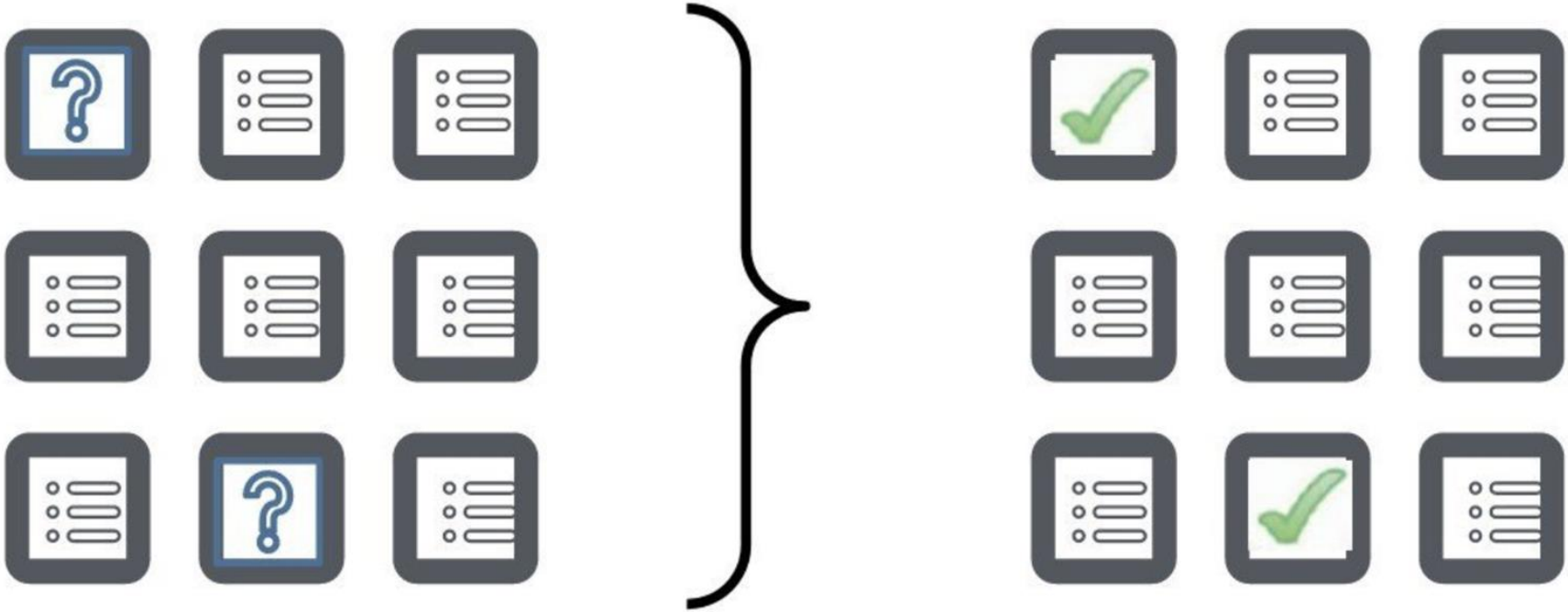
Limpieza



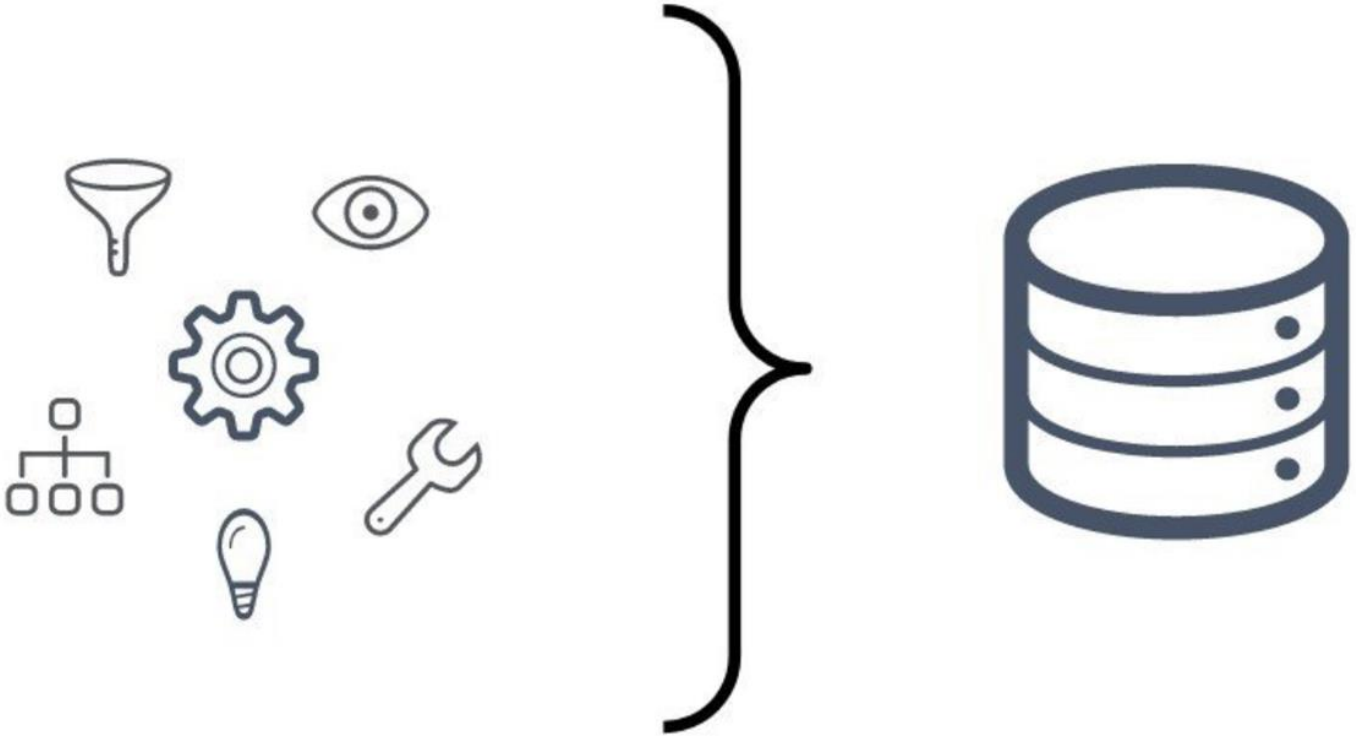
Normalización



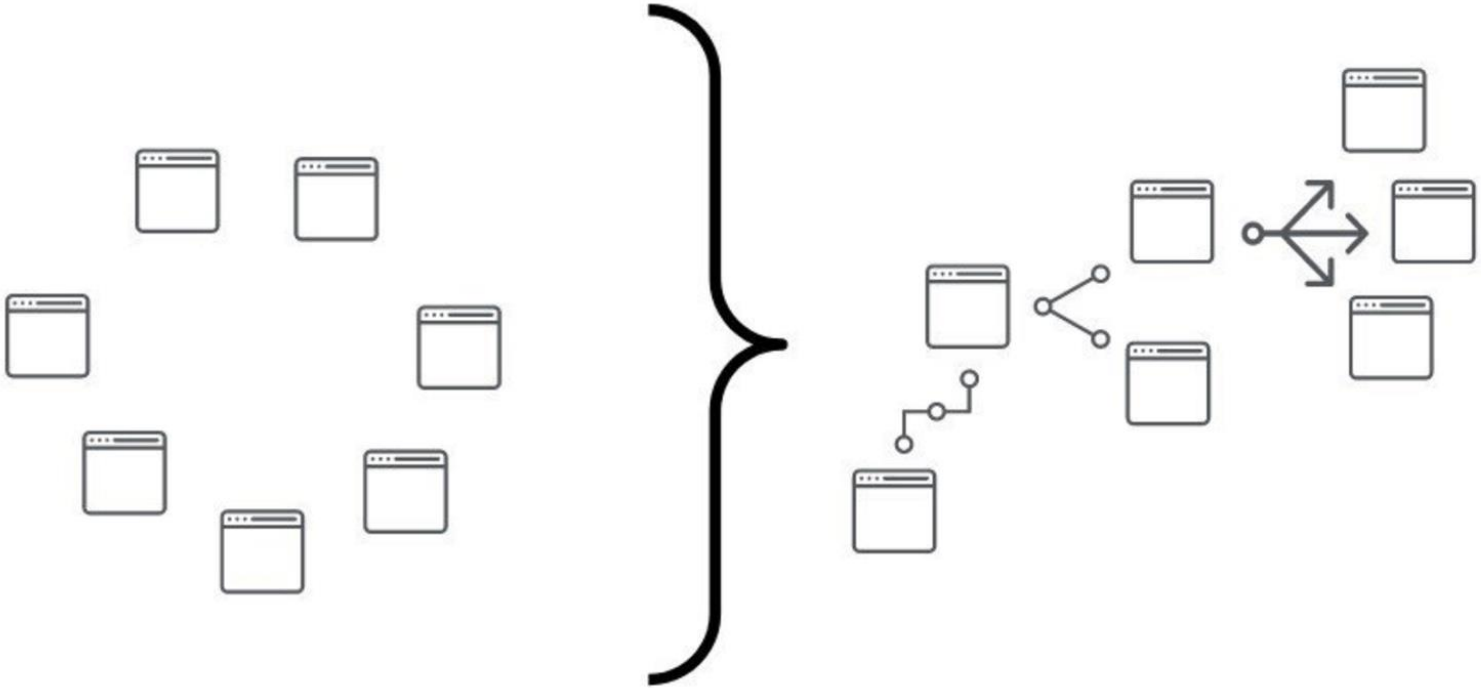
Imputación de Valores Faltantes



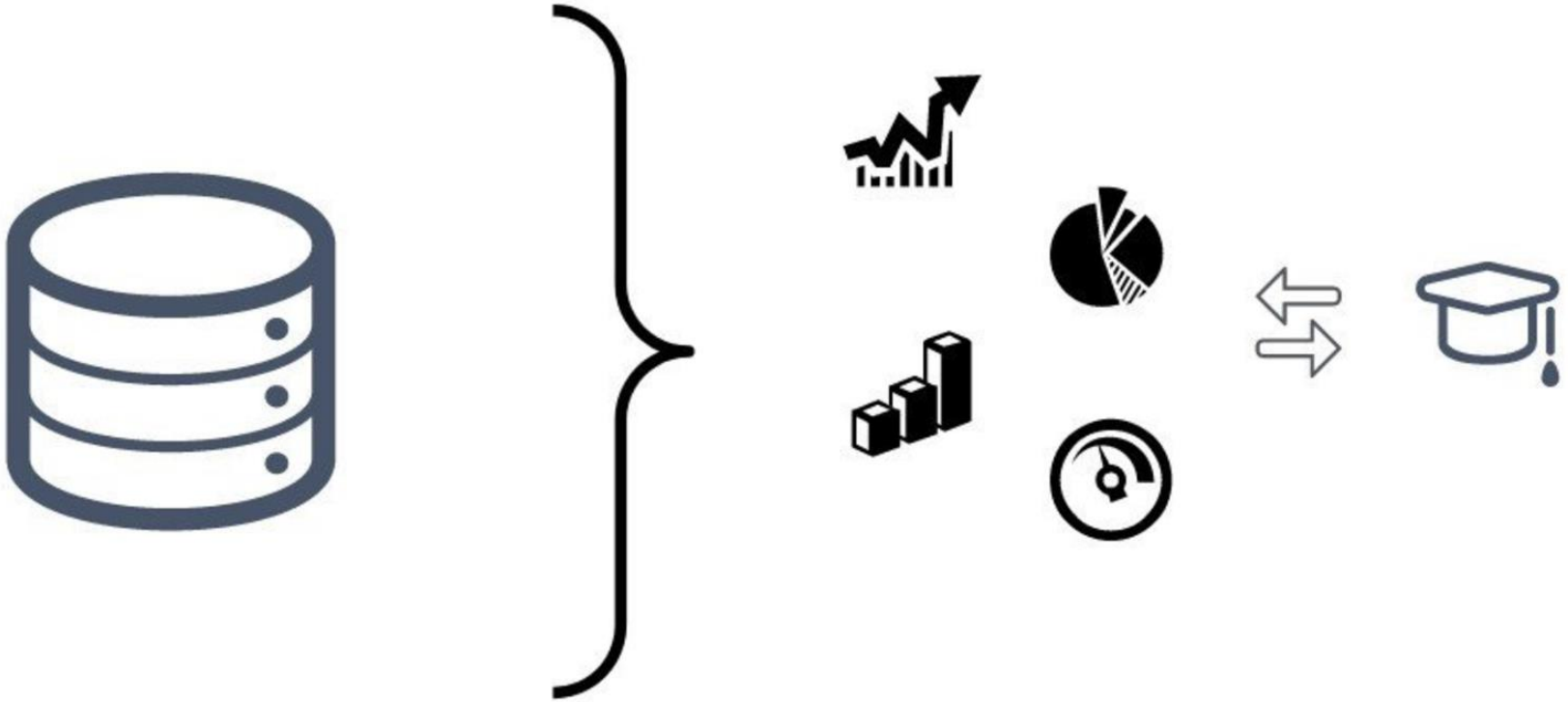
Transformación de datos



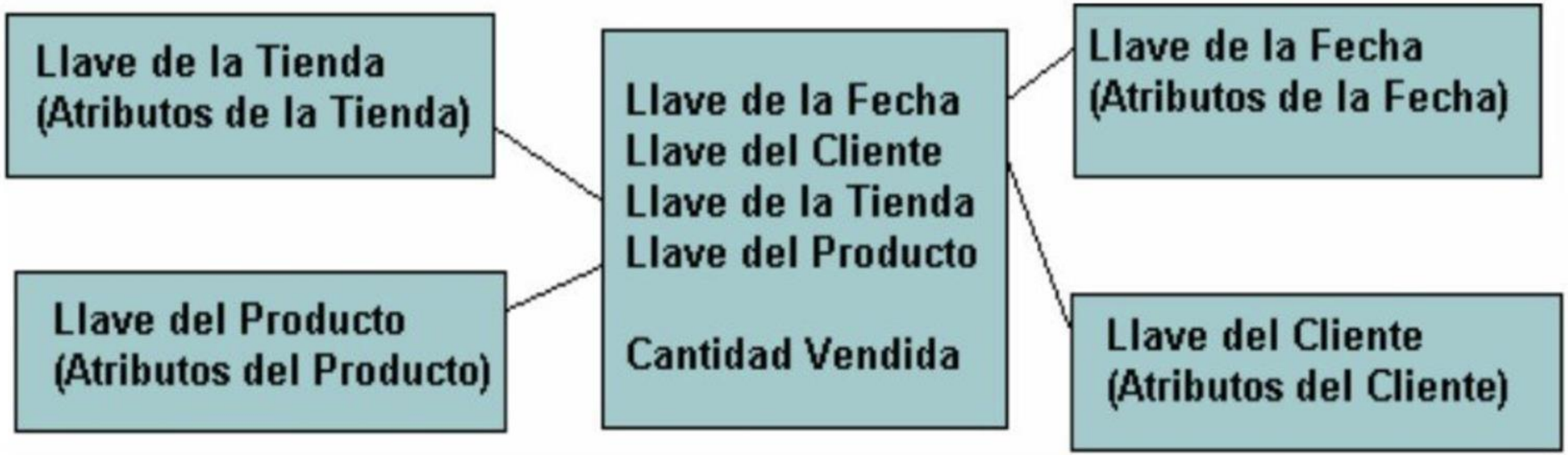
Modelado de Datos



Reportes y Visualización



Hechos/Dimensiones



Claves Subrogadas

Una **clave subrogada** es un identificador único que se asigna a cada registro de una tabla. Puede obtenerse a partir de la conjunción de columnas ya preexistentes.

¿PREGUNTAS?

Resumen

¿Alguien dijo Homework?

HENRY

Próxima lecture

ETL


HENRY