

## **ETL - PARTE 1**

**Eric Gustavo Coronel Castillo**

[youtube.com/c/DesarrollaSoftware](https://youtube.com/c/DesarrollaSoftware)

[www.desarrollasoftware.com](http://www.desarrollasoftware.com)

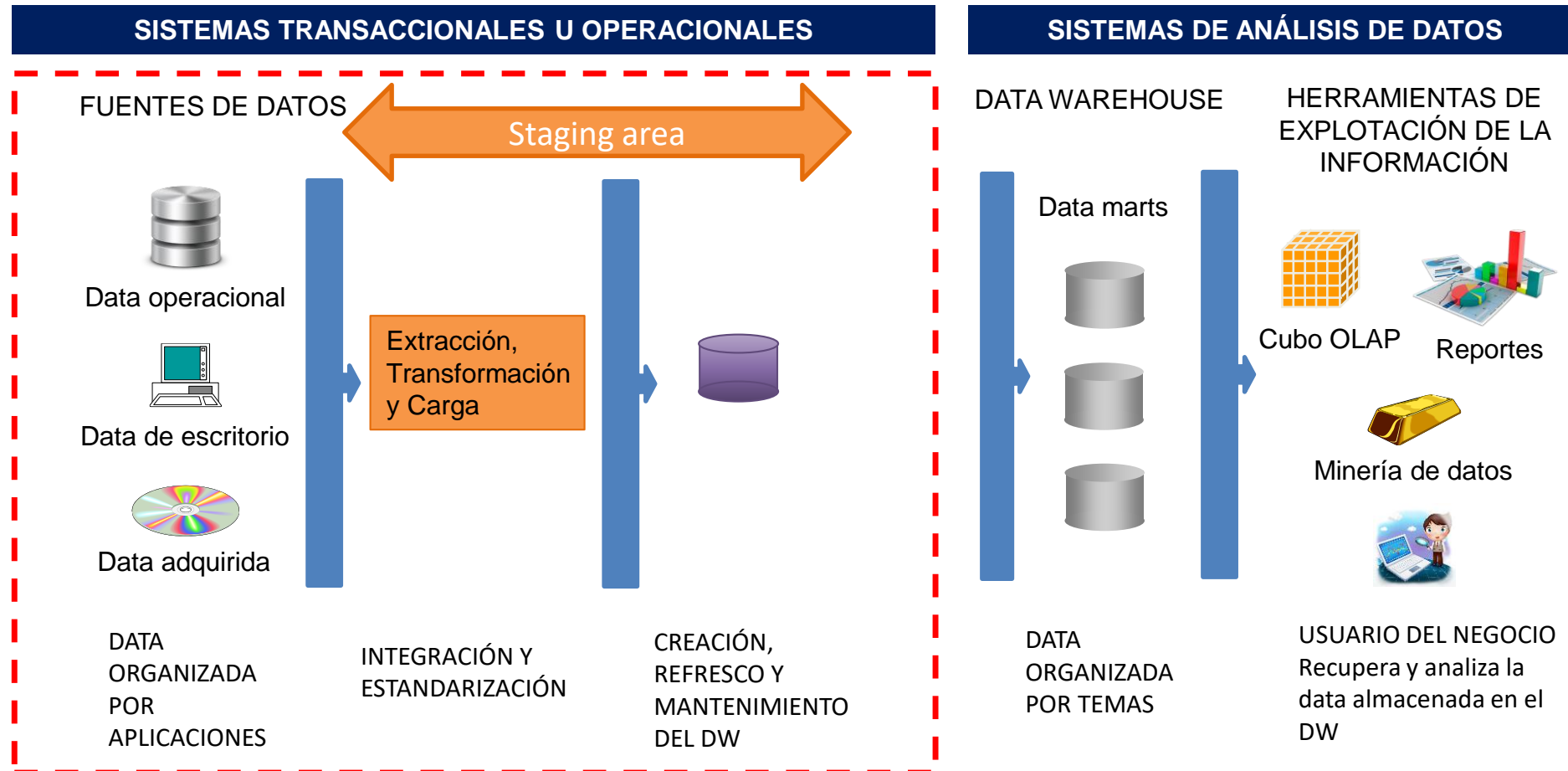
[gcoronel@uni.edu.com](mailto:gcoronel@uni.edu.com)

# Logro Esperado

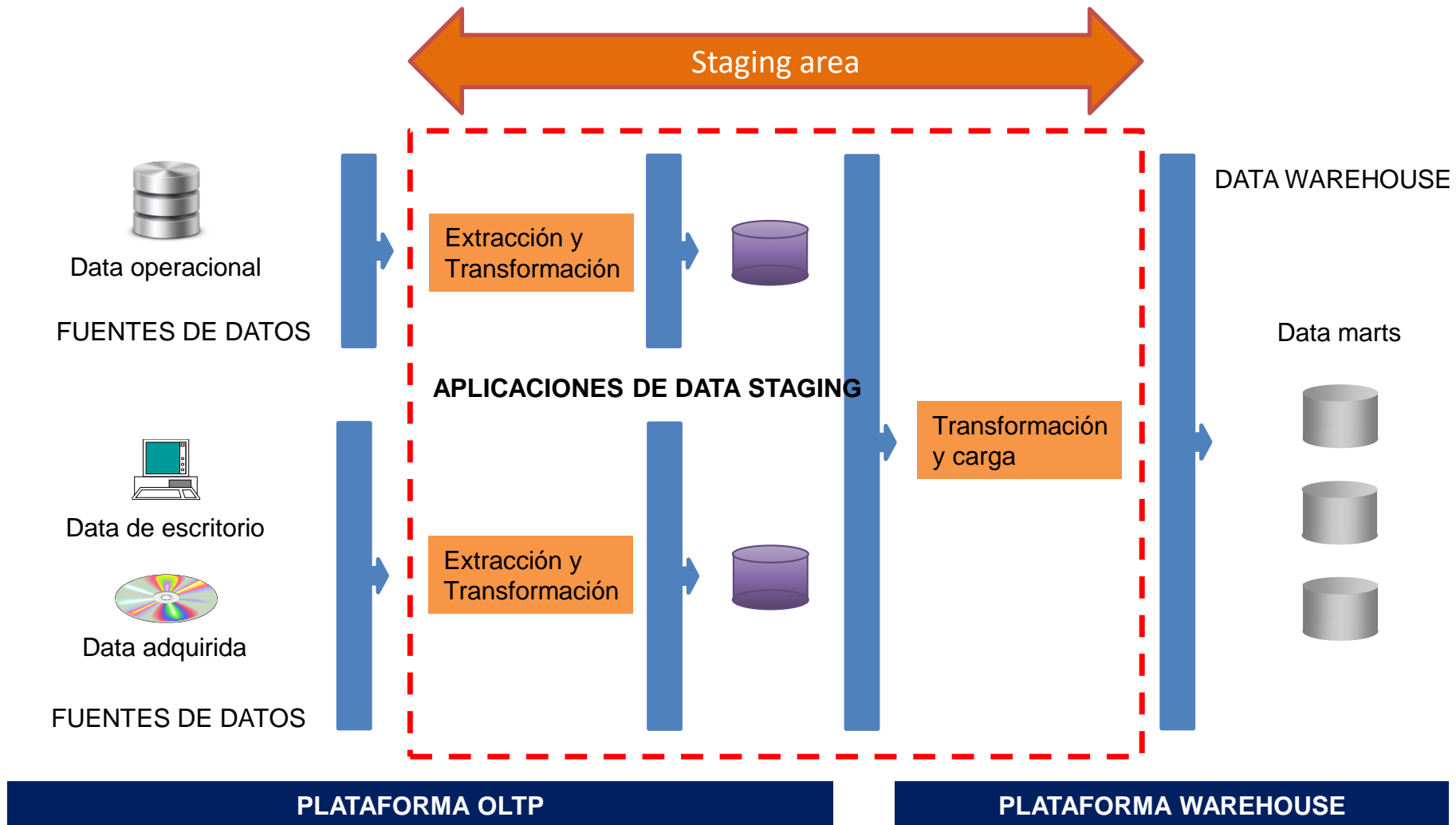
Al finalizar esta presentación, se espera que el participante aplique SSIS para crear procesos ETL Básicos.



# ESCENARIO DE OPERACIONES ETL



# STAGING AREA

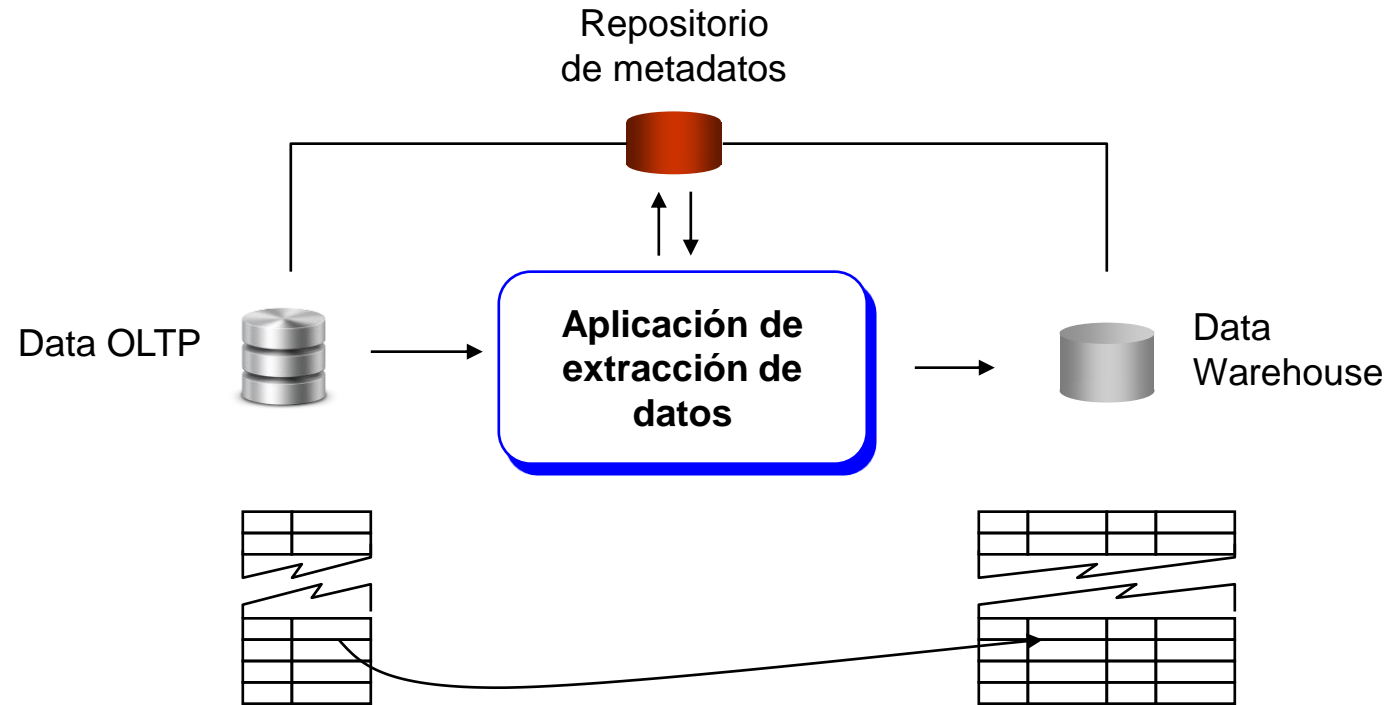


# STAGING AREA

---

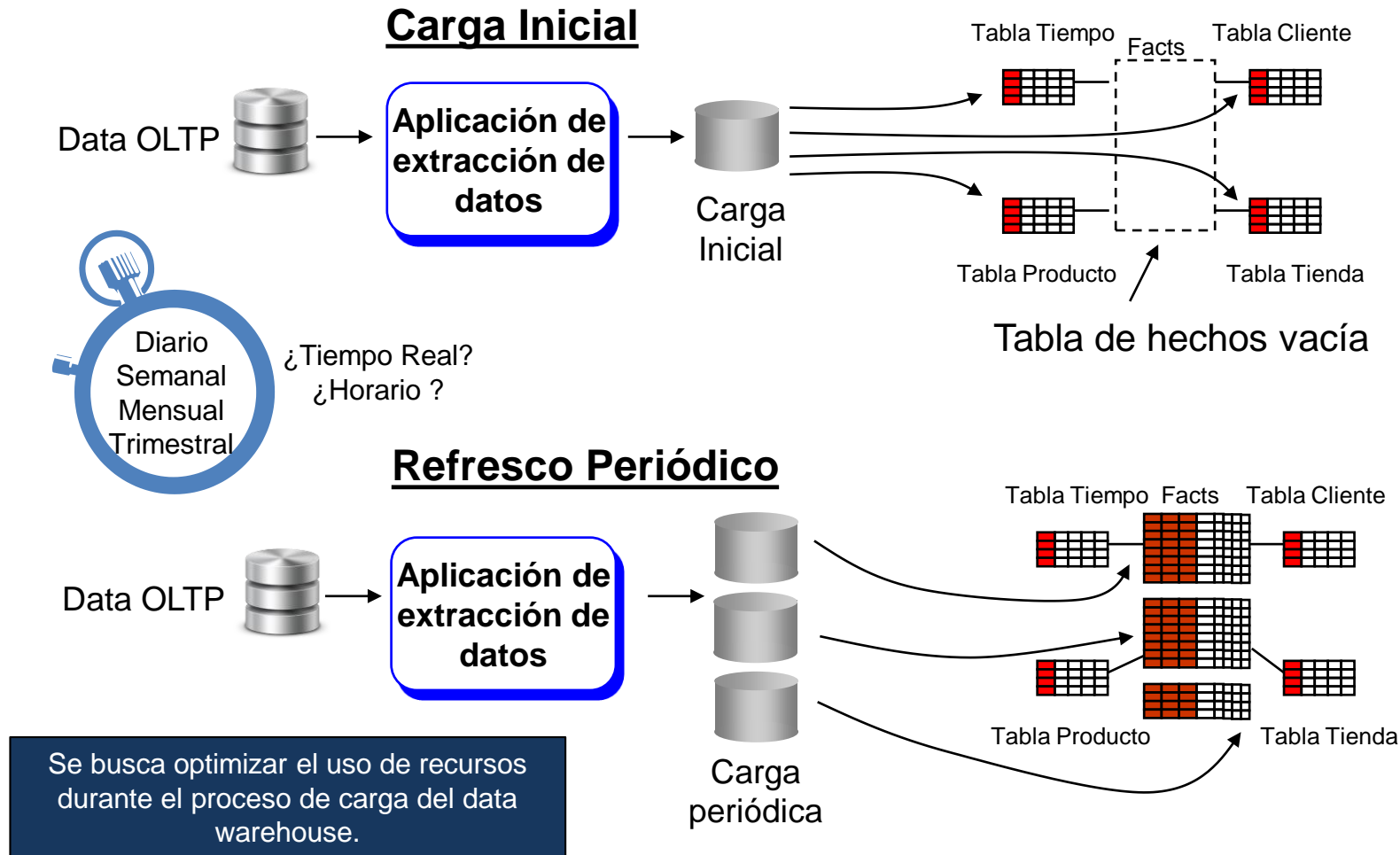
- Área de almacenamiento intermedio utilizada para el procesamiento de calidad de datos durante las operaciones de extracción, transformación y carga.
- Contiene tablas normalizadas y archivos usados para
  - Staging e integridad de datos del Sistema OLTP.
  - Validación.
  - Referencias cruzadas y búsquedas.
- La data OLTP fluye a través del staging area y es eventualmente almacenada en un esquema estrella (dimensional).
- El acceso a la data por parte de los usuarios de negocio está siempre direccionado a la base de datos dimensional.

# EXTRACCIÓN DE DATOS: Mapeo de datos



Verificar en las fuentes de datos, qué atributo o atributos se deben leer para poblar cada atributo del data warehouse.

# EXTRACCIÓN DE DATOS: El proceso de refresco



# ESTANDARIZACIÓN DE DATOS: Inconsistencia en estándares

Valores flag

*	
X, x	
1	0
Y, y	N, n
T, t	F, f
	U

Se utiliza distintas denominaciones y formatos para el mismo dato.

Valores de código

10
FIN
Fin
Finanzas

Valores desconocidos

NULL
UN
?

Valores numéricos

-8	Signo negativo en posición fija
8-	Signo negativo en posición fija
(8)	Valor negativo entre paréntesis
800	Separador decimal implícito
-8,00	Separador decimal explícito
8E2	Notación científica

Valores de fecha

MMDDYY	
DDMMYY	
YYMMDD	
CCYYMMDD	Estándar ISO
	Fecha desconocida
0	Fecha desconocida
00 00 00	Fecha desconocida



# ESTANDARIZACIÓN DE DATOS: Semántica (significado) de los datos



Mi nombre es Tomy



Mi nombre es Tomás

Diferente nombre, mismo significado



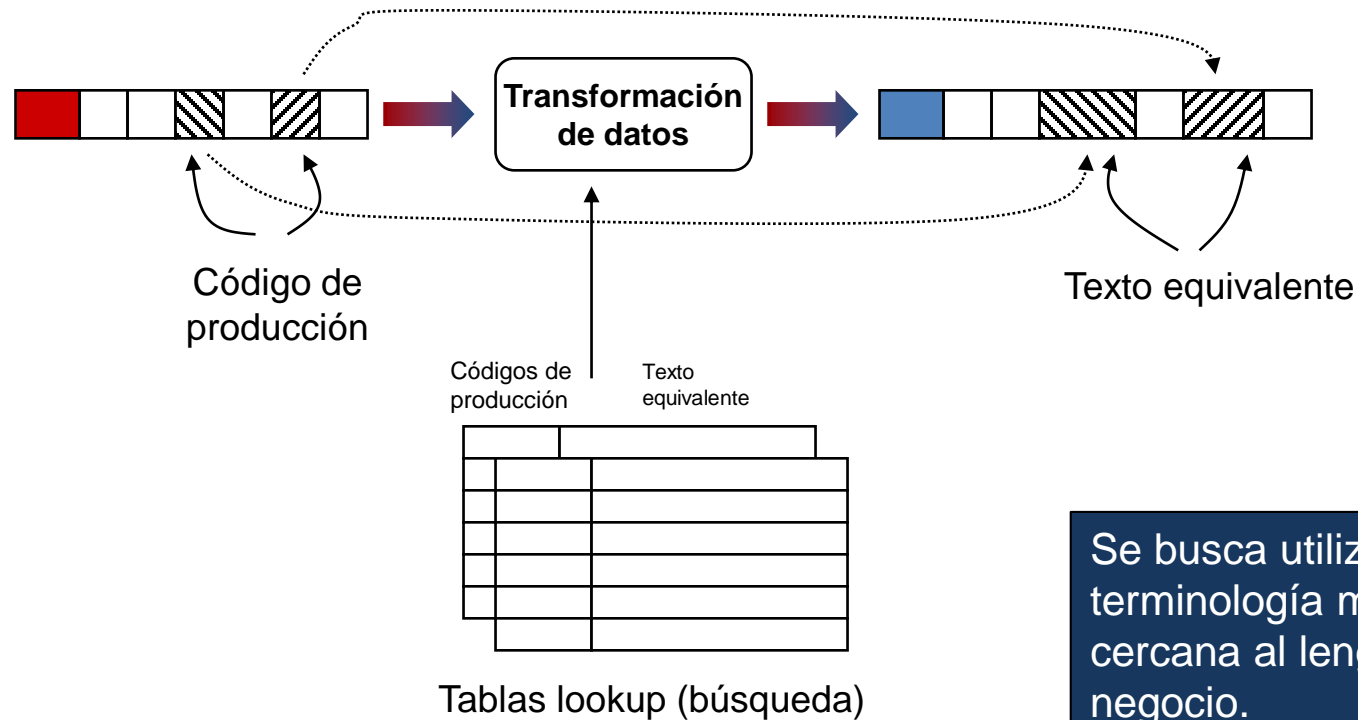
Mi nombre es Brenda



Mi nombre es Brenda

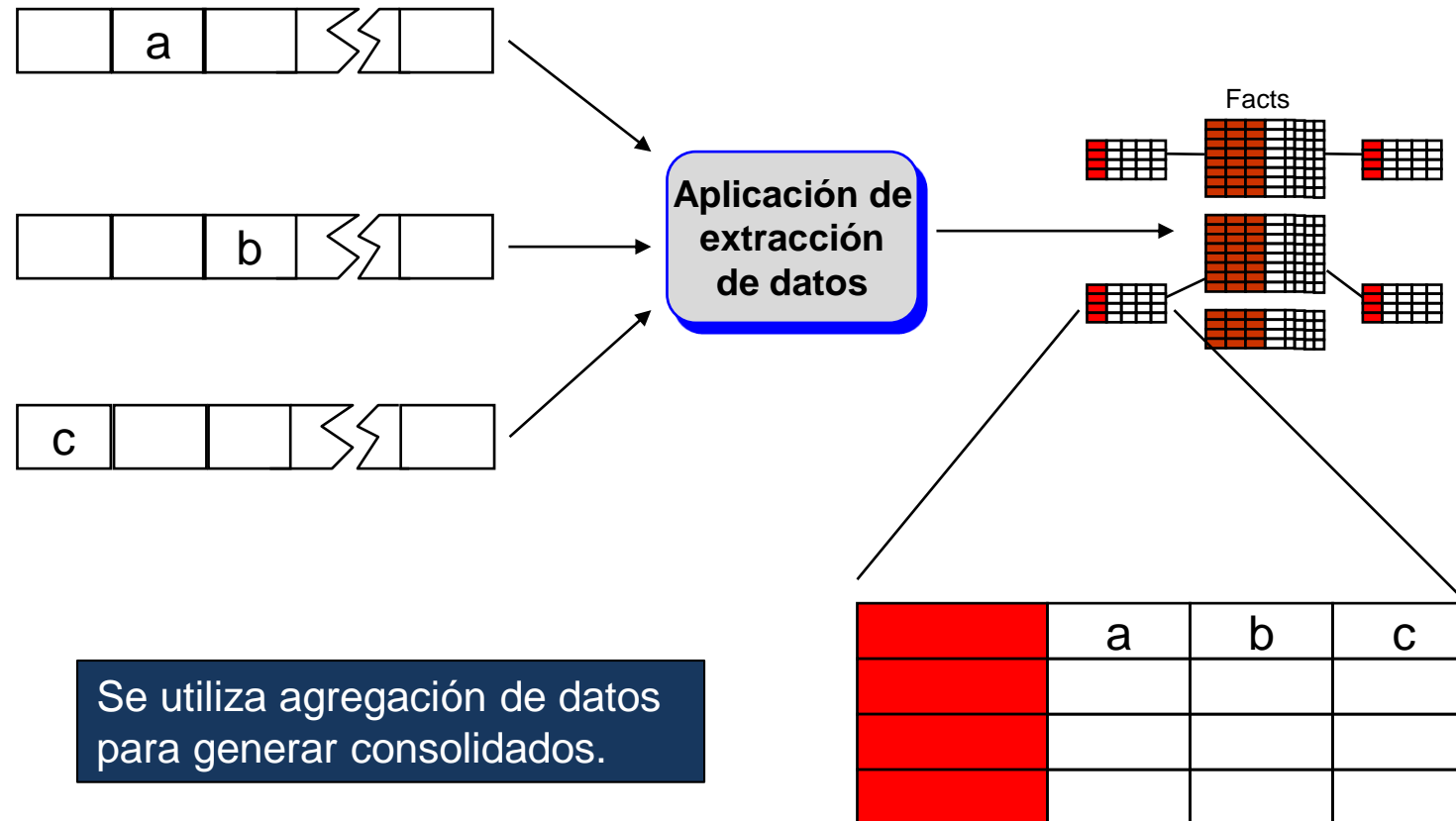
Mismo nombre, diferente significado

# TRANSFORMACIÓN DE DATOS: Códigos de producción

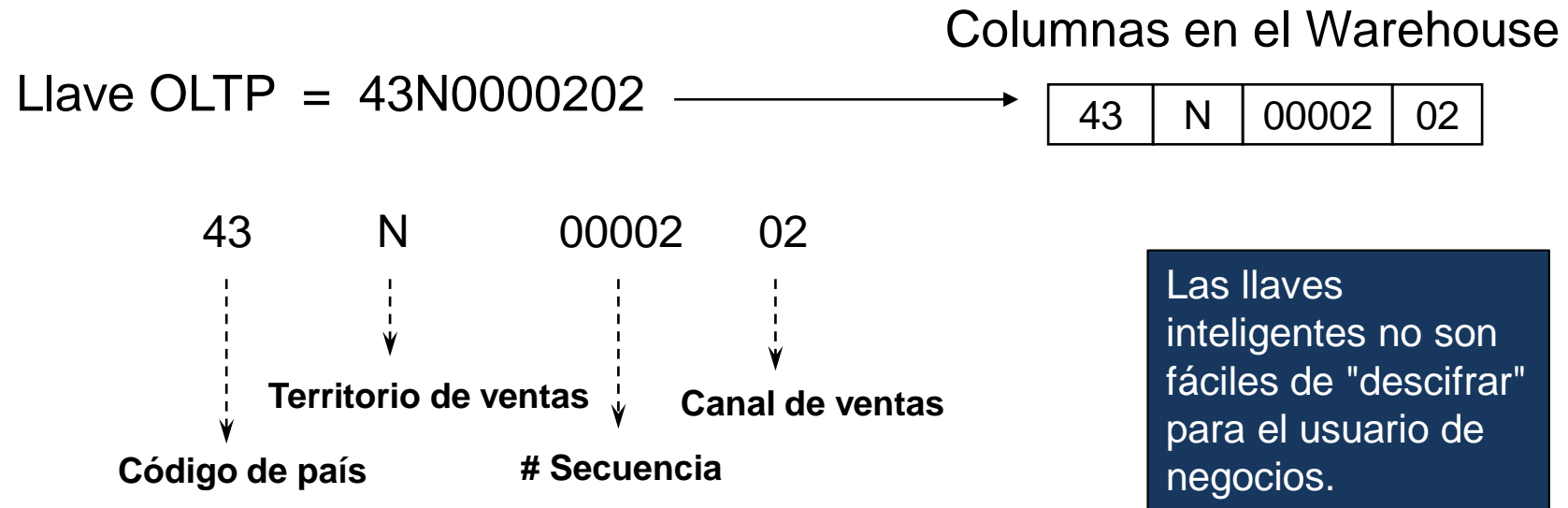


Se busca utilizar terminología más cercana al lenguaje del negocio.

# TRANSFORMACIÓN DE DATOS: Consolidación de datos



# TRANSFORMACIÓN DE DATOS: Llaves inteligentes

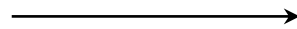


**Llaves con inteligencia embebida**

# TRANSFORMACIÓN DE DATOS: Separación de datos

Campo de Entrada

XYZ
-----



Campo de Salida

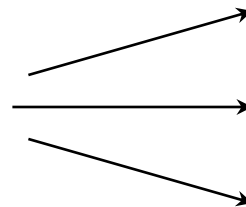
A	B	C
---	---	---



Un campo de entrada dividido en varios campos de salida.

Registro de Entrada

Llave	X	Y	Z
-------	---	---	---



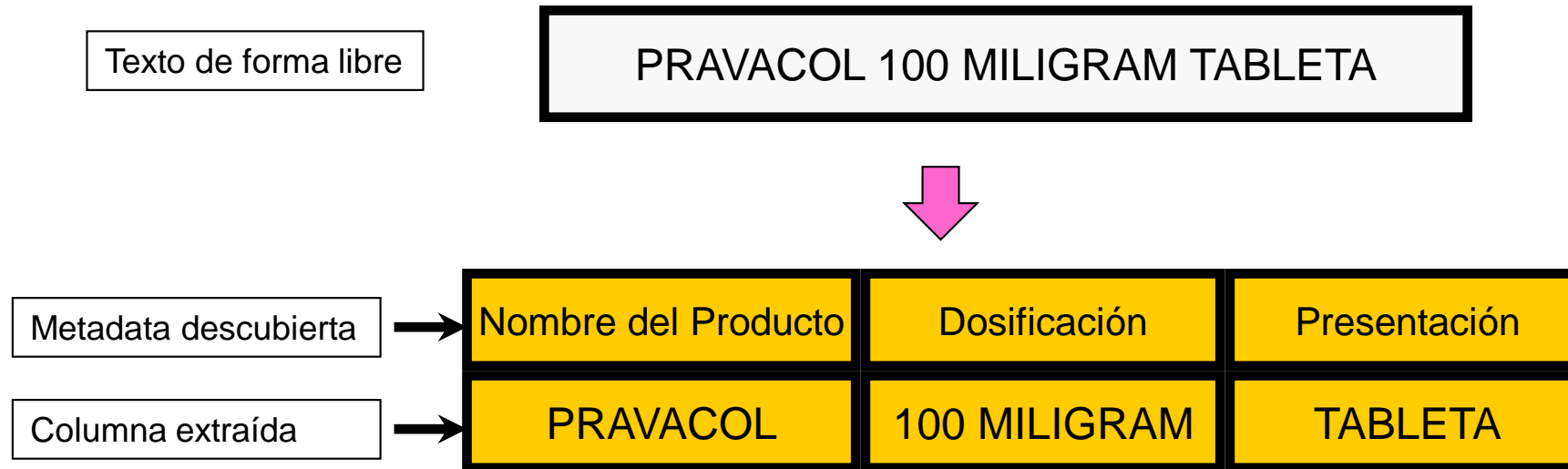
Registro de Salida

Llave1		X	
Llave2	Y		
Llave3		Z	



Un registro de entrada dividido en varios registros de salida.

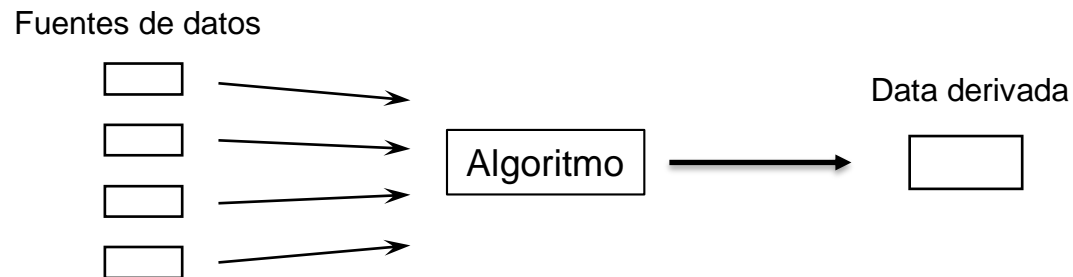
# TRANSFORMACIÓN DE DATOS: Separación de datos – Campo de texto libre



Se busca darle estructura a la data en el campo de texto libre.

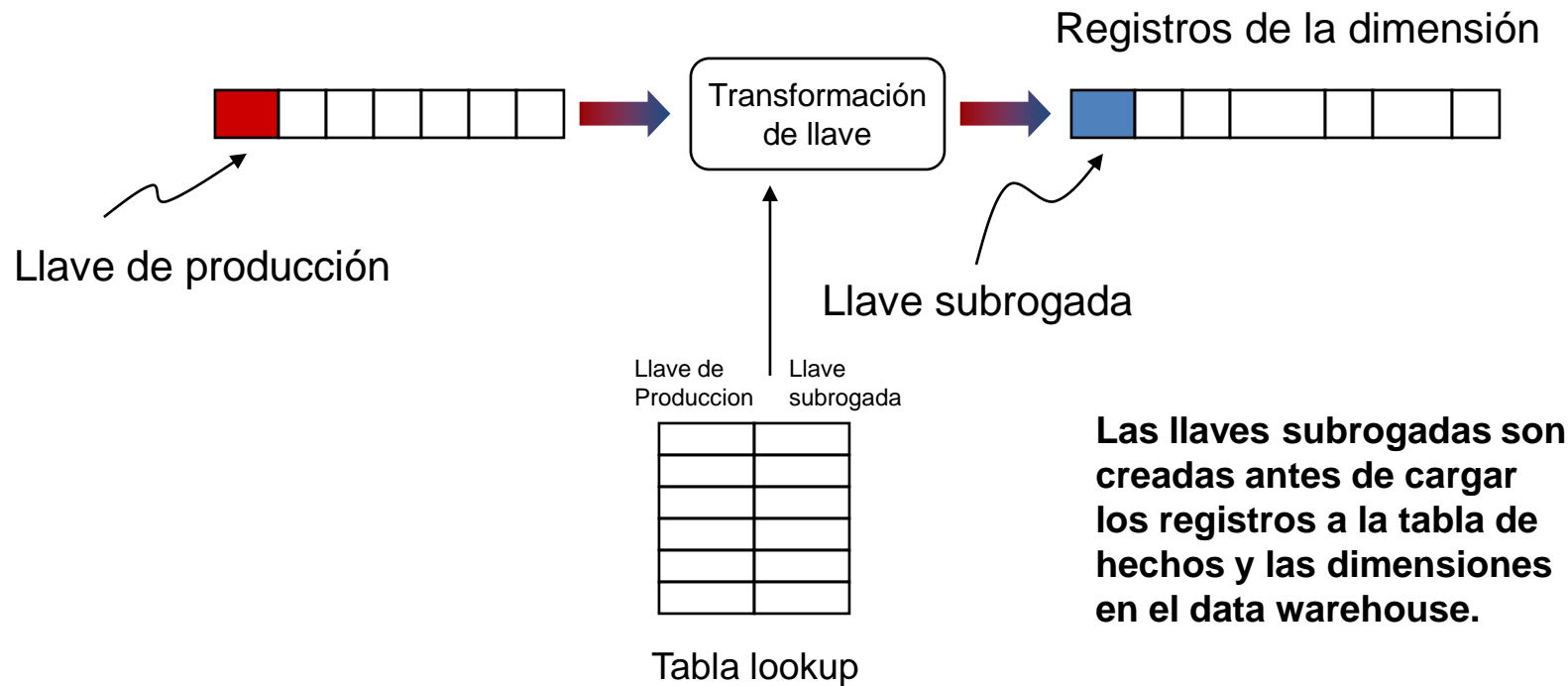
# TRANSFORMACIÓN DE DATOS: Data derivada

- Ranking de clientes
- Cálculos de cierre de año
  - Total sueldos, Total de impuestos, Total beneficios sociales, etc.
- Calculada una vez – siempre disponible



Obtención de un nuevo dato a partir de datos existentes mediante la aplicación de una fórmula o algoritmo.

# TRANSFORMACIÓN DE DATOS: Asignación de llaves



Clave subrogada: identificador único en una tabla, y que no tiene ningún significado para el negocio. Su objetivo es diferenciar a un registro de los otros.



## CONCLUSIONES

---

- Los datos de una organización no siempre tienen la calidad requerida para realizar **análisis de información**, por lo que es necesario someterlos a un proceso de **calidad de datos** para estandarizarlos e integrarlos.
- Durante el proceso de extracción de datos se debe identificar **qué problemas presentan los datos**, para luego aplicarles las operaciones de transformación que permiten que los datos sean utilizables para análisis de información.