



**Diseño de Base de Datos II**

**Unidad II - Tema I**

# **BASES DE DATOS CORPORATIVAS**

Ing. Carina Cozzolino



# Agenda

## **BASES DE DATOS CORPORATIVAS**

Conceptos

Arquitectura - Componentes

Fases de Implementación

Proceso de Modelado

Dimensiones

Tabla de Hechos

Práctica

# Bases de Datos Corporativas

## Data Warehouse

- Repositorio centralizado
- Almacena toda la información proveniente de uno o más sistemas
- Cuyos datos atraviesan un proceso de limpiado y transformación
- Para realizar un análisis sobre ellos





# DATA WAREHOUSE

## PROCESO

No es un producto.  
Es un conjunto de  
herramientas.

## TÉCNICA

Para consolidar y  
administrar datos  
provenientes de varias  
fuentes.

## TOMA DE DECISIONES

Responde preguntas de  
negocios, brindando un  
panorama generalizado.  
Identifica patrones y  
tendencias.



Un **DATA WAREHOUSE** permite almacenar, gestionar y acceder a grandes cantidades de datos empresariales de manera eficiente.

OBJETIVO INICIAL: satisfacer la demanda de los gestores para obtener una Visión Integrada de la Empresa y su Entorno.



# ARQUITECTURA

- **Fuentes de Datos:**

- Fuentes *Internas*: sistemas y aplicaciones internas de la empresa (sistemas de ventas, CRM, ERP, etc.)
- Fuentes *Externas*: datos de proveedores, de redes sociales, datos de mercado, etc.

- **ETL** (Extracción, Transformación y Carga):

- *Extracción*: etapa de recopilación de datos de diversas fuentes. Se extraen para su posterior procesamiento.
- *Transformación*: los datos extraídos se limpian, se transforman y se enriquecen para garantizar su calidad y coherencia.
- *Carga*: los datos procesados se almacenan en el data warehouse, de acuerdo a un esquema estrella, diseñado para este fin.

- **Motor de Consulta:** para ejecutar consultas SQL y operaciones de análisis sobre los datos almacenados.

- **Capa de Presentación:** reportes y herramientas de visualización, para crear informes, paneles de control y visualizaciones para el análisis de los datos.

- **Seguridad y Control de Acceso:** se implementan medidas de seguridad para proteger los datos confidenciales, y garantizar el acceso a personas autorizadas a cada conjunto de datos.

- **Programación y Automatización:** se configuran tareas para la actualización automática y ejecución de procesos ETL.



## DATA WAREHOUSE

# DATA WAREHOUSE

## Técnicas Analíticas



### OLAP

#### On Line Analytical Processing

Es una estructura multidimensional para realizar análisis avanzados, y explorar los datos desde distintas dimensiones.

Permiten explorar los datos de manera interactiva mediante consultas y operaciones, aplicando diferentes filtros.

**Dimensión:** características o atributos que describen los datos.



### DSS - EIS

#### Decision Support System - Executive Information System

Son herramientas de ayuda para la toma de decisiones.

**DSS:** proporcionan herramientas para la consulta, modelado de datos, simulación y generación de informes. Permite evaluar diferentes escenarios antes de tomar una decisión.

**EIS:** sistema de información diseñado para brindar información de alto nivel a los ejecutivos de una organización. Proporcionan resúmenes y visualizaciones claves de datos relevantes para la alta dirección. Monitorean el rendimiento de la organización. Permite tomar decisiones estratégicas basadas en información actualizada y comprensible.



### DATA MINING

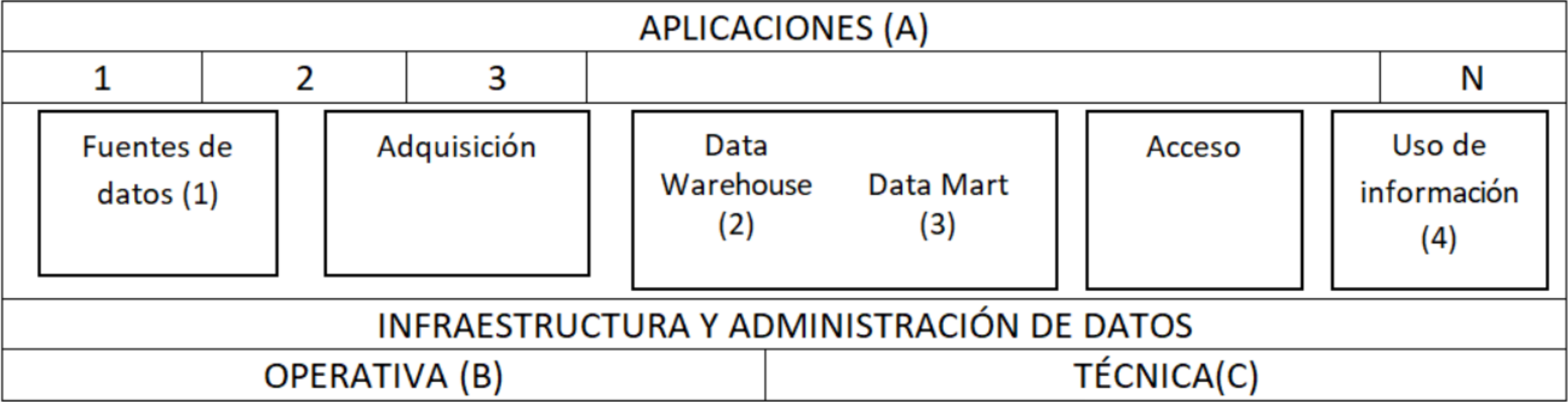
Proceso de descubrir patrones, tendencias, relaciones o información significativa a partir de grandes conjuntos de datos.

Este proceso implica el uso de técnicas y algoritmos estadísticos, matemáticos y de inteligencia artificial para analizar los datos y extraer conocimientos útiles que pueden ayudar en la toma de decisiones.



# Data Warehouse

## ARQUITECTURA



Un Data Warehouse está formado por cuatro bloques:

- Fuente de Datos
- Data Warehouse
- Data Mart
- Uso de la Información

El objetivo fundamental de la construcción de un Data Warehouse es **transformar Datos en Conocimiento**.

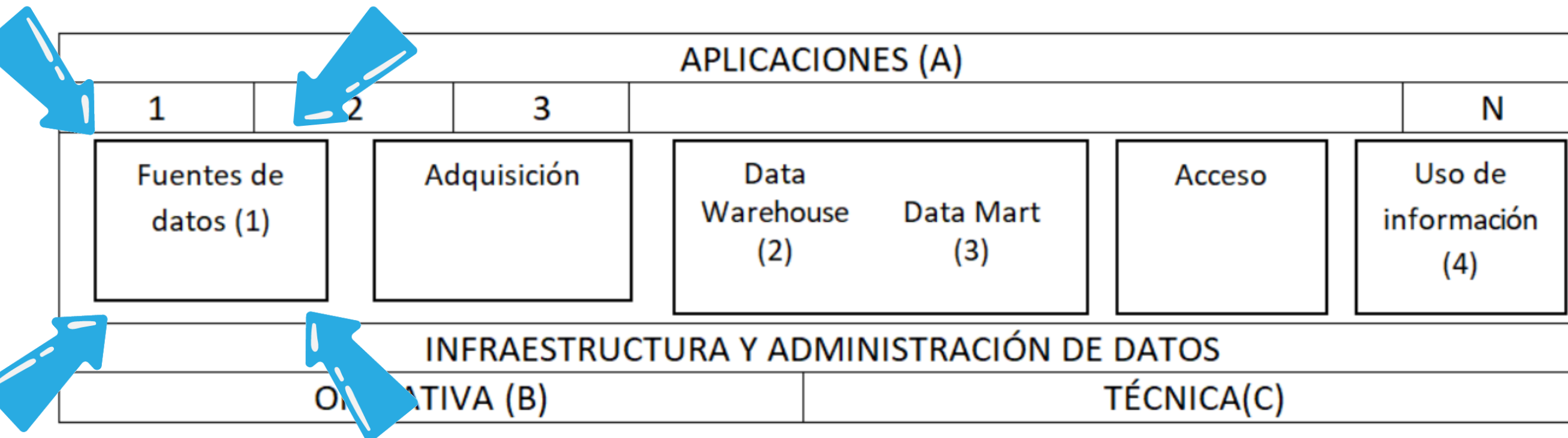


Para eso es necesario ensamblar varias **Fuentes (1)** en un depósito que es el **Data Warehouse (2)** o el **Data Mart (3)**, para que los usuarios puedan acceder a su contenido y obtener **Conocimiento (4)**



# Data Warehouse

## ARQUITECTURA



### Fuentes de Datos (bloque 1)

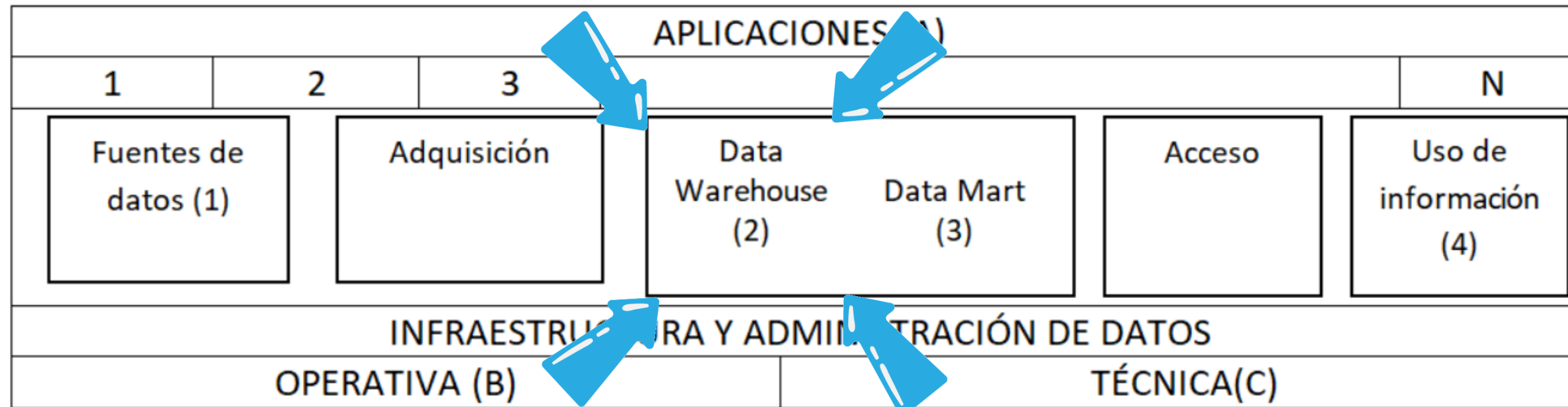
Comprende los siguientes elementos:

- Bases de Datos Operacionales: contienen datos provenientes de aplicaciones transaccionales.
- Datos de Herencia: son datos No Operacionales, pero importantes por su valor histórico
- Fuentes Externas: datos relacionados al Data Warehousing, pero que surgen de organismos especializados en temas financieros, bursátiles, etc. y deben ser adquiridos.



# Data Warehouse

## ARQUITECTURA



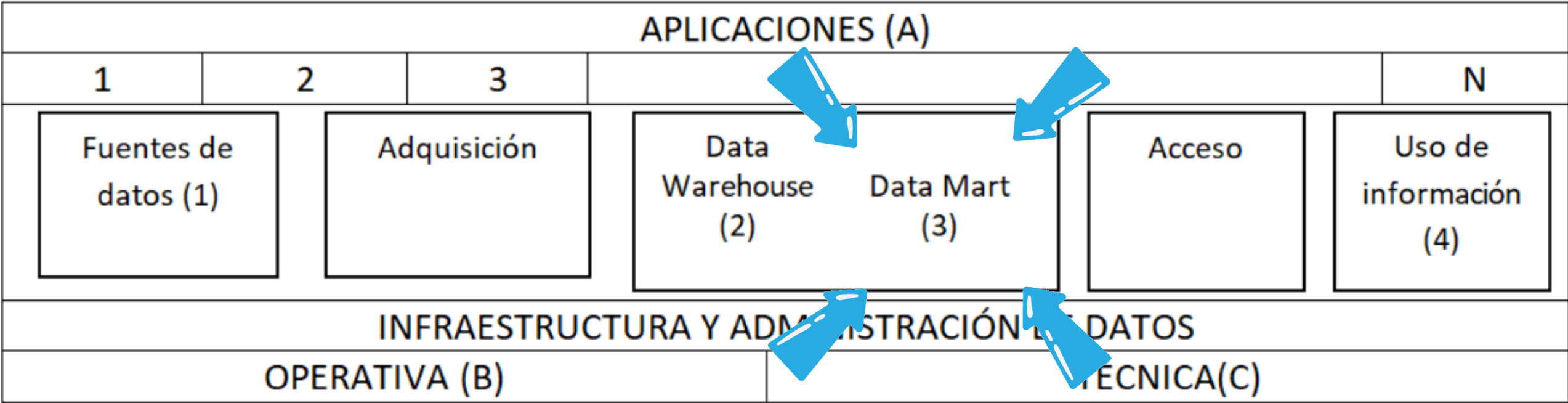
### Data Warehouse (bloque 2)

Se compone de los siguientes elementos:

- Componentes de Refinamiento: su función es estandarizar los datos, filtrar y pulir, registrar la fecha de de la fuente de datos y verificar la calidad de los datos.
- Componentes de Reingeniería: su función se orienta a exponer lo datos de tal forma que puedan servir para un análisis adecuado del usuario final.
  - ☐ Integra y segmenta datos
  - ☐ Resume información según reglas definidas previamente
  - ☐ Incorpora campos adicionales resultantes de cálculos
- Componentes de Generación de Data Warehouse:
  - ☐ Modelado de información y condensación de información voluminosa en magnitudes manejables.

# Data Warehouse

## ARQUITECTURA



### Data Mart (bloque 3)

Representa una implementación de Data Warehouse, pero referida a un *ámbito de datos y funciones más limitado*. Puede tener como usuario a un único Departamento dentro de la Corporación.

La organizaciones empresariales pueden disponer de un Data Warehouse y, opcionalmente, de uno o varios Data Mart.

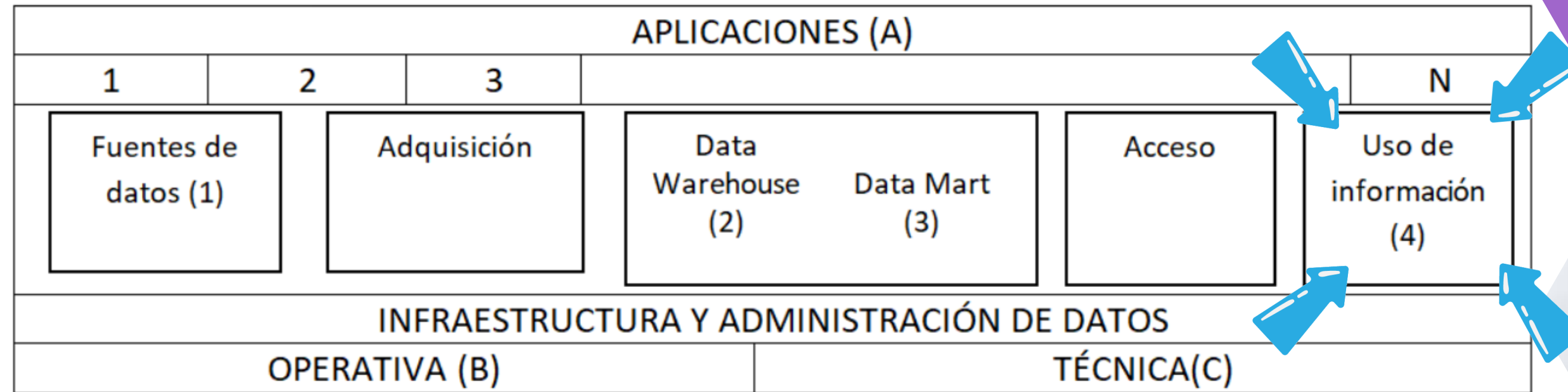
También puede ocurrir que las corporaciones implementen varios Data Mart sin estructurar un Data Warehouse. En las organizaciones donde conviven el Data Warehouse y el Data Mart, éste último opera a partir del contenido del primero.

Los componentes para construir un Data Mart son similares a los del Data Warehouse. El refinamiento y reingeniería se aplican al filtrado y ajuste de esa información, generación de nuevos resúmenes y asignación de fechas de los nuevos datos generados.



# Data Warehouse

## ARQUITECTURA



## Uso de Información (bloque 4)

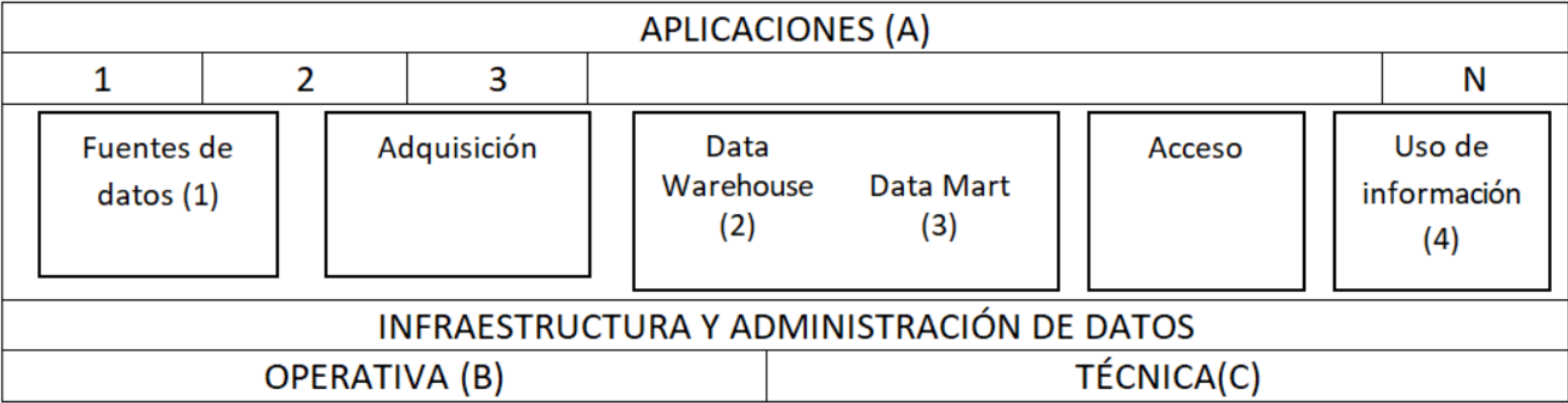
Se compone de:

- Acceso y Recuperación: Se pueden transformar los datos extraídos en “Vistas Multidimensionales”, o almacenarlos en bases de datos multidimensionales para un análisis posterior.
  - Dimensión: es un eje de análisis que corresponde a los temas de interés del Data Warehouse (tiempo, cliente, producto, etc.). Cada dimensión tienen jerarquía (por ej. país, región, provincia, ciudad).
- Análisis y Reporte: son herramientas para informes y soporte de decisiones.



# Data Warehouse

## ARQUITECTURA



### Capas Horizontales

(A) Aplicaciones: Son las distintas partes (iniciativas o aplicaciones) en que se descompone cada tema tratado en el Data Warehouse.

- ❑ *Iniciativa*: proyecto de decisión que se incorpora de forma iterativa en la construcción de un Data Warehouse.
- ❑ *Aplicaciones*: definidas en cuanto a objetivos, frecuencia y periodicidad de análisis.

**Infraestructura y Administración de Datos**: soporta los bloques descriptos previamente. Sobre esta capa se apoyan las aplicaciones de decisión.

La infraestructura se presenta en dos niveles:

(B) Operativo: procedimientos para administrar los datos

(C) Técnico: comprendelos productos (programas) que instalan las tecnologías que se aplican.

# COMPONENTES

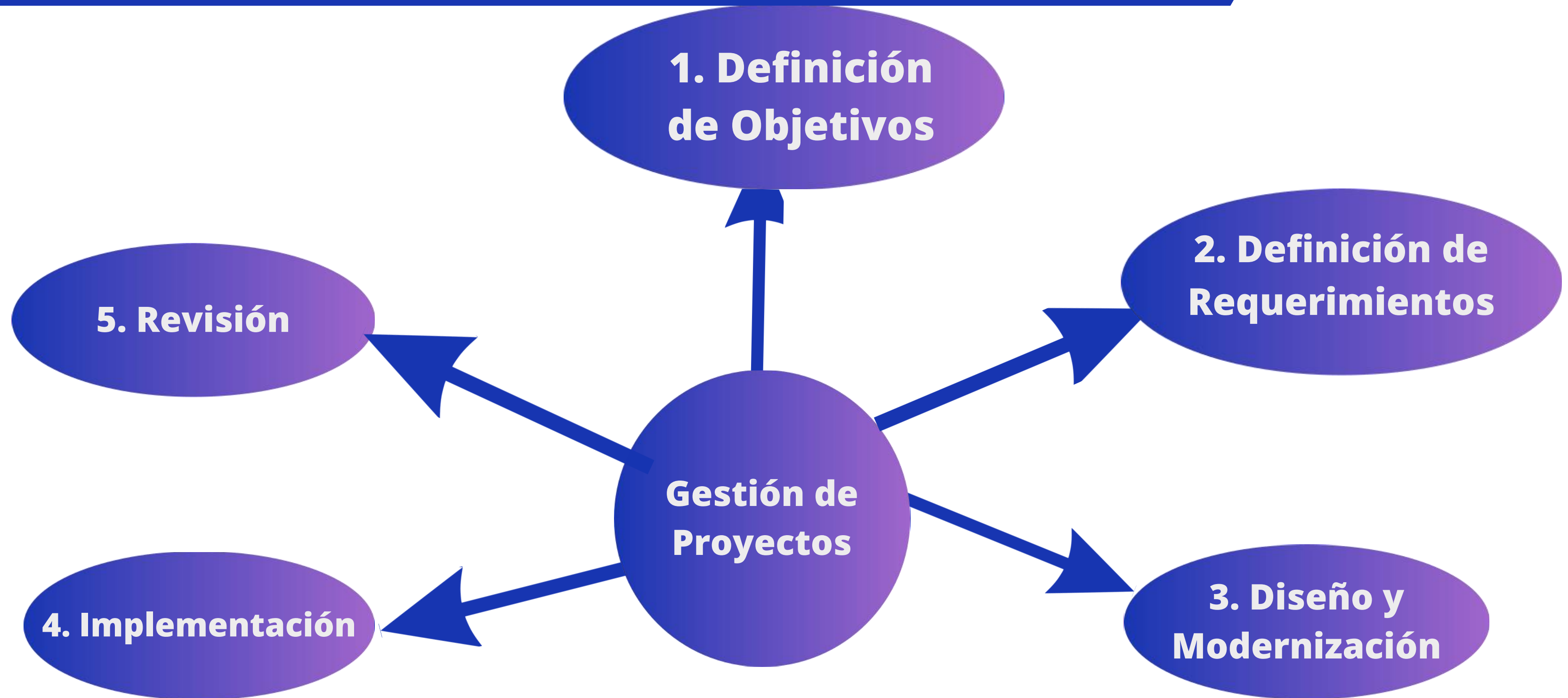
- Hardware:
  - ☐ Pocos usuarios con grandes necesidades
  - ☐ Consultas complejas e imprevistas
  - ☐ Gran cantidad de información
  - ☐ Escalabilidad
- Software de Almacenamiento (SGBD):
  - ☐ Atender consultas multidimensionales
  - ☐ Los datos se ven como cubos de información
  - ☐ Procesan a gran velocidad mucha información
- Software de Extracción y Manipulación de Datos:
  - ☐ Herramientas para controlar y actualizar el Data Warehouse
  - ☐ Gestión Integrada de Data Warehouse y Data Marts existentes
  - ☐ Acceso a gran variedad de fuentes de datos
  - ☐ Manejo de excepciones
  - ☐ Planificación, logs, interfaces a terceros
  - ☐ Soporte de explotación del Data Warehouse
- Herramientas de Middleware: son herramientas que se ubican entre el Sistema O y las aplicaciones de usuario.
  - ☐ Herramientas que proveen conectividad entre entornos diferentes
  - ☐ Analizadores y aceleradores de consultas.

# Data Warehouse



# Data Warehouse

## Fases de Implementación







## 1. Definición de Objetivos



Se deben definir todos los objetivos que se quieren lograr con el Data Warehouse, qué es lo que se quiere y hasta donde se puede llegar.

## 2. Definición de Requerimientos de Información



Analizar las necesidades.

El usuario juega un papel destacado en esta fase del proyecto.

## 3. Diseño de Modelización



Identificar las fuentes de datos y realizar las transformaciones necesarias para obtener el modelo lógico de datos, que está formado por entidades relacionadas que permiten resolver las necesidades del negocio.

## 4. Implementación



La implementación lleva los siguientes pasos:

Extracción de datos y transformación de los mismos.

Carga de los datos validados, planificada con periodicidad.

Explotación del Data Warehouse mediante diversas técnicas (OIS, DSS, Visualización, Data Mining).

## 5. Revisión



Revisar lo implementado, planteando preguntas que luego de puesto en marcha, pueda definir los aspectos a mejorar.

Un Data Warehouse es una tarea iterativa.





### 6. Diseño de Estructura de Cursos de Formación

Deben diseñarse cursos a medida, cuyo objetivo es proporcionar la formación estadística necesaria para el mejor aprovechamiento de la funcionalidad incluida en la aplicación.

Se deben realizar prácticas, las cuales permitirán fijar los conceptos adquiridos.



# Proceso de Modelado de Data Warehouse

## 1. Proceso de Negocio

*UN PROCESO A LA VEZ*

Definir el proceso de negocio que se va a analizar y los datos que se pueden obtener de los sistemas fuentes.

## 2. El Grano

*EL ANÁLISIS MÁS PEQUEÑO AL QUE SE DEBE LLEGAR*

El grano establece el mínimo nivel en que los datos son capturados y almacenados en la tabla de Hechos.

La elección del grano depende de los objetivos analíticos.

## 3. Dimensiones

*FILTRADO DE DATOS*

Las dimensiones añaden contexto a las métricas obtenidas de los eventos del proceso de negocio. Las tablas de dimensiones contienen atributos que permiten a las métricas ser filtradas y agrupadas.

## 4. Hechos

*EL DATO QUE SE BUSCA ANALIZAR*

*Los hechos son medidas resultantes de los eventos del proceso de negocio.*

*En su mayoría son numéricos.*

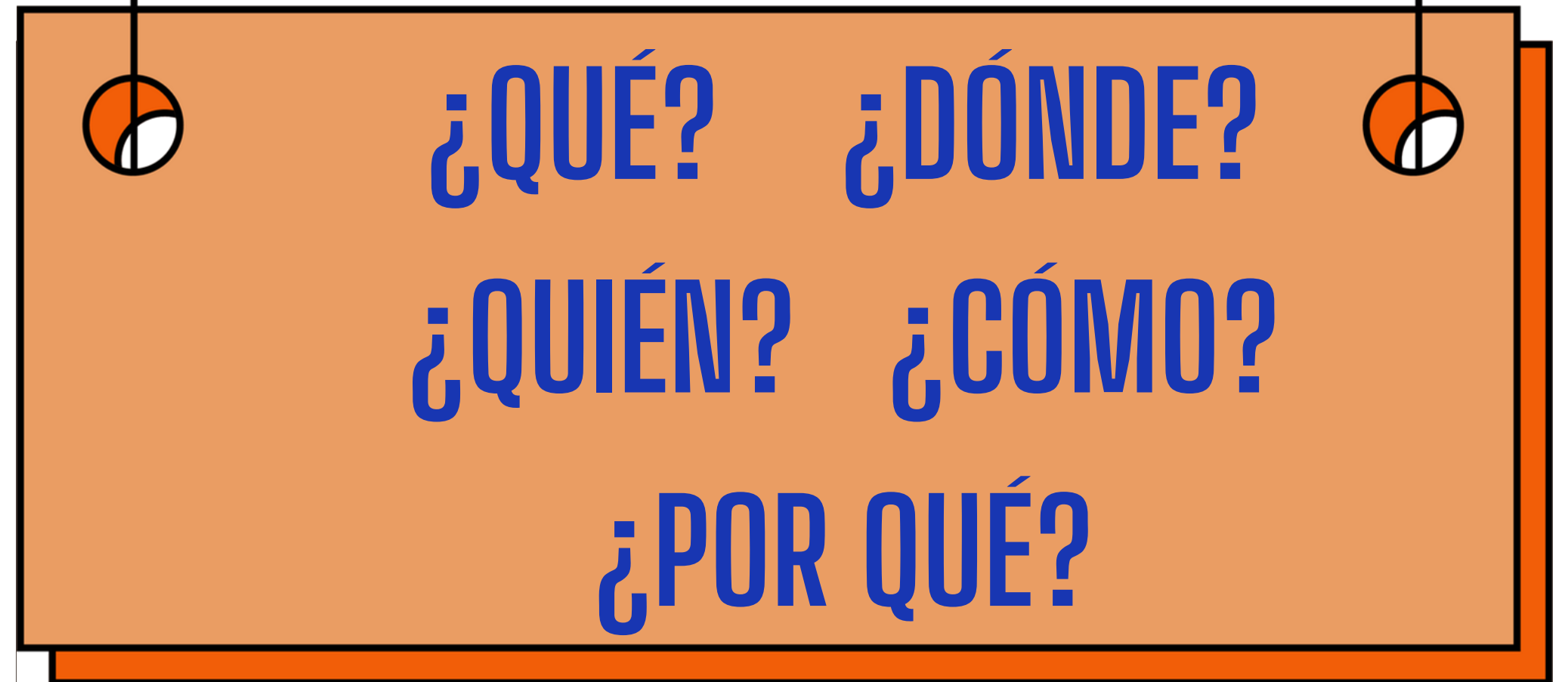
*Cada fila solo permite medidas basadas en el grano*



# DATA WAREHOUSE

## Dimensiones

Proveen de contexto a los eventos que responden preguntas.



Cada pregunta representa una dimensión diferente.

Pueden existir combinación de preguntas.

# Dimensiones

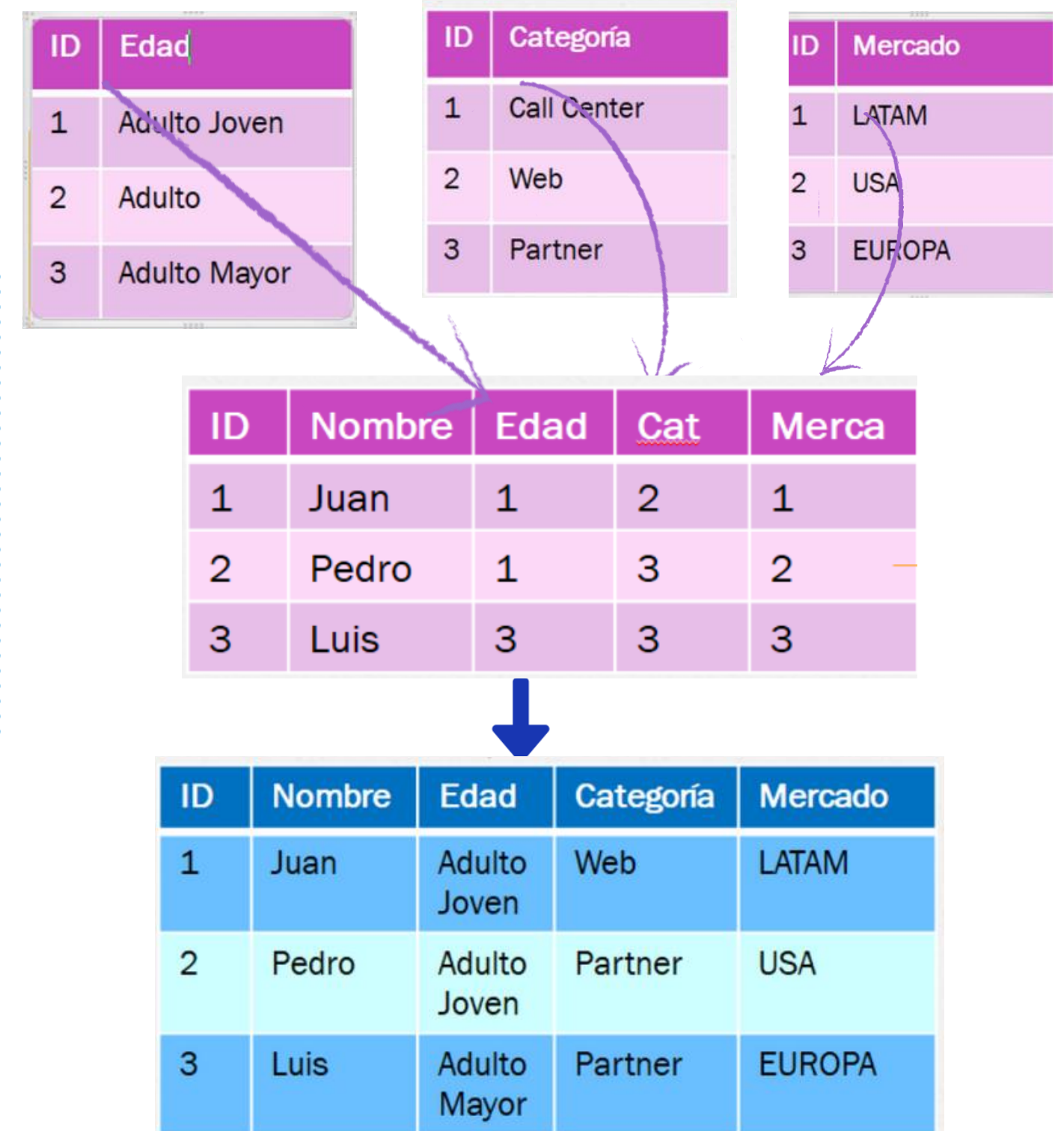
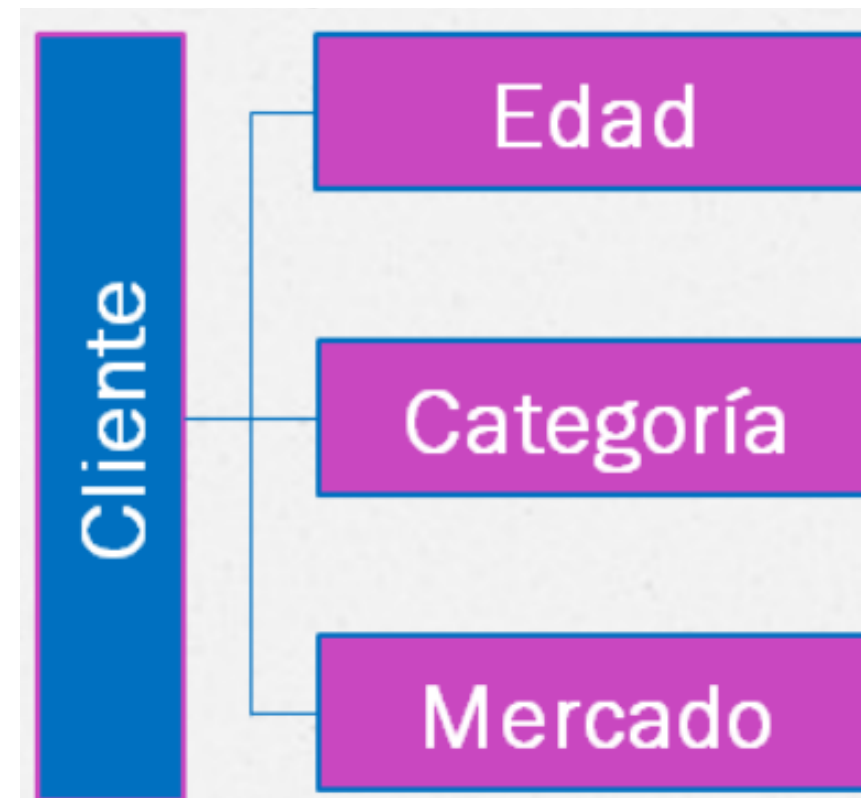
## Consideraciones

*Se debe definir como se guardarán los datos obtenidos del sistema origen:*

- *Desnormalización*
- *Llaves Primarias*
- *Datos Nulos*
- *Cambios en el Origen*

## Desnormalización

En el sistema relacional una tabla relacionada (por claves foráneas) con otras que aportan atributos, en el Data Warehouse pasan a ser una sola dimensión, que contiene los datos de ese conjunto de tablas en una.



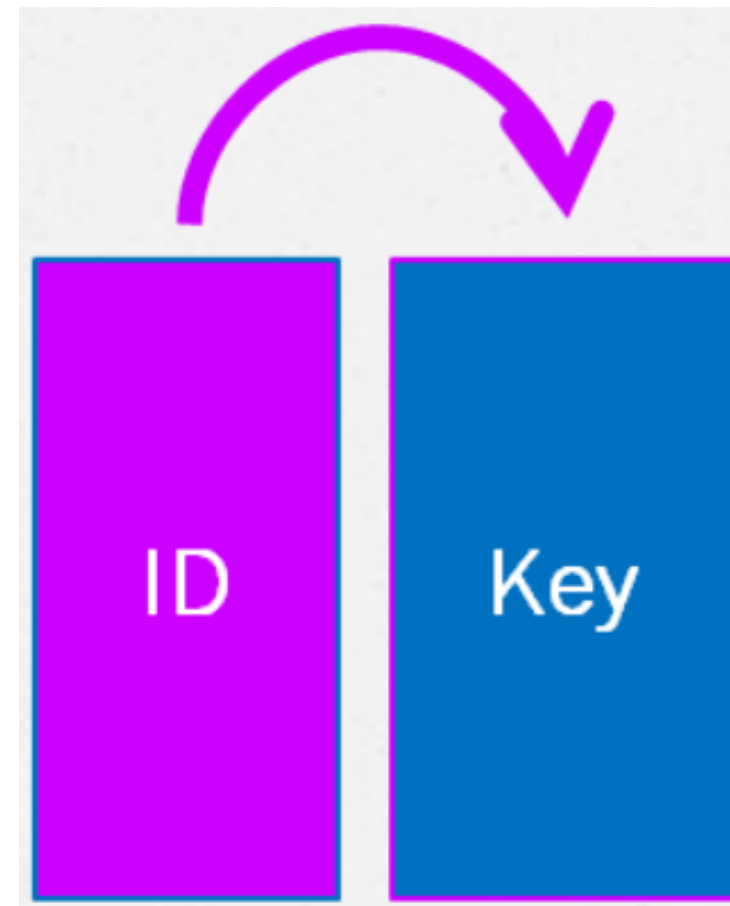
# Dimensiones

## Llaves Primarias

Cada dimensión debe tener una columna que sirva como llave primaria.

**Subrogate Key**: es el identificador que crea el Data Warehouse (uno para cada fila dentro de cada dimensión).

Estas llaves deben ser números enteros consecutivos.



ID	Nombre	Edad	Agen	Merca
1	Juan	1	2	1
2	Pedro	1	3	2
3	Luis	3	3	3



ID	Key	Nombre	Edad	Agencia	Merca
1	1	Juan	Adulto Joven	Web	LATAM
2	2	Pedro	Adulto Joven	Partner	USA
3	3	Luis	Adulto Mayor	Partner	EUROPA
4	1	Juan	Adulto	Web	LATAM
5	1	Juan	Adulto Mayor	Call Center	USA

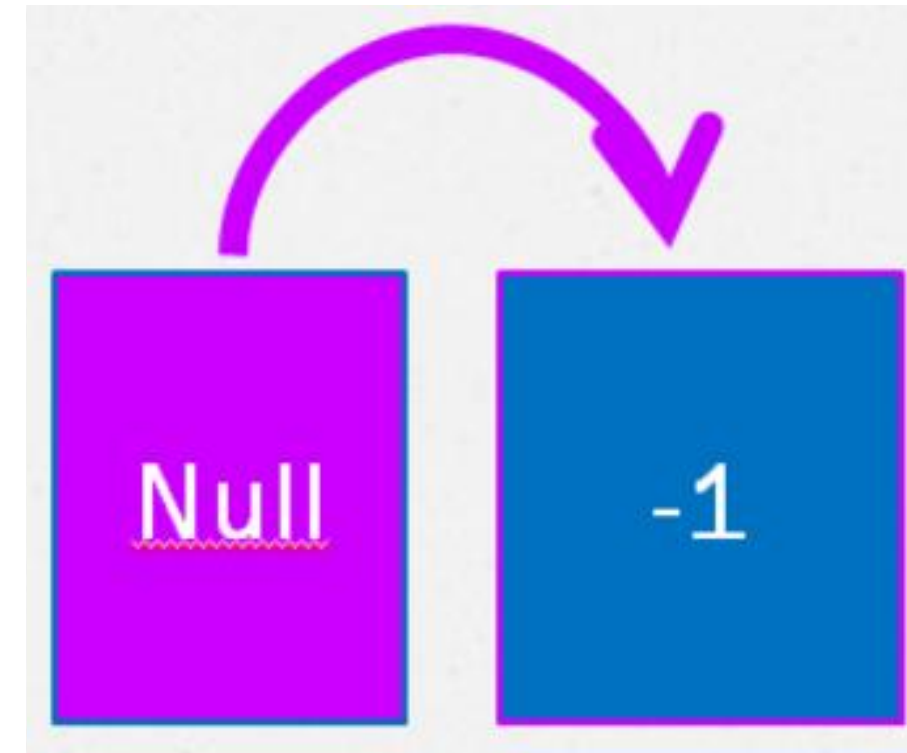


# Dimensiones

## Datos Nulos

- Deben evitarse los atributos nulos o vacíos dentro de una dimensión.
- Eventos registrados en la tabla de hechos pueden no traer el dato de la dimensión en cuestión.

**Recomendación:** crear una fila “default” con los valores como “No Disponible”, “Desconocido” o “No Aplica”.



ID	Key	Nombre	Edad	Categoría	Mercado
-1	-1	No Disponible	No Disponible	No Disponible	No Disponible
1	1	Juan	Adulto Joven	Web	LATAM
2	2	María	No Disponible	Partner	USA
3	3	Luis	Adulto Mayor	No Disponible	UE

# Dimensiones

## Cambios en el Origen Slowly Changing Dimension (SCD)

- Los datos de los sistemas origen cambian constantemente por lo que el Data Warehouse requiere un método para reflejar estos cambios dentro de las dimensiones.
- Existen 4 formas de manejar los cambios.



# Dimensiones

## Cambios en el Origen - Tipo 0 (SCD0)

- Una vez ingresado el valor a la dimensión, no importa cuántos cambios hayan en el origen, estos serán descartados.
- La dimensión siempre tendrá los valores originales.
- Se utiliza cuando hay pocos cambios.

The diagram illustrates the SCD0 (Type 0 Slowly Changing Dimension) process. It shows three tables connected by arrows, demonstrating how changes in the source data are handled in the dimension table.

**Table 1 (Initial State):**

ID	Nombre	Cat	Age	Merca
1	David Smith	1	2	1
2	Melinda Gates	1	3	2
3	Jhon Doe	3	3	3

**Table 2 (After Changes):**

ID	Nombre	Cat	Age	Merca
1	David Smith	2	3	2
2	Melinda Gates	1	3	2
3	Jhon Doe	3	2	3

**Table 3 (Dimension Table):**

ID	Key	Nombre	Cat	Age	Merca
1	1	David Smith	Adulto Joven	Web	LATAM
2	2	Melinda Gates	Adulto Joven	Partner	USA
3	3	Jhon Doe	Adulto Mayor	Partner	EU

**Table 4 (Final State):**

ID	Key	Nombre	Cat	Age	Merca
1	1	David Smith	Adulto Joven	Web	LATAM
2	2	Melinda Gates	Adulto Joven	Partner	USA
3	3	Jhon Doe	Adulto Mayor	Partner	EU



# Dimensiones

## Cambios en el Origen - Tipo 1 (SCD1)

- Cuando se registra un cambio en el sistema origen, la dimensión **sobreescribe** el atributo viejo por el nuevo valor.
- No queda ningún registro del valor anterior y todos los datos siempre muestran los valores más actualizados.

ID	Nombre	Cat	Age	Merca
1	David Smith	1	2	1
2	Melinda Gates	1	3	2
3	Jhon Doe	3	3	3



ID	Nombre	Cat	Age	Merca
1	David Smith	2	3	1
2	Melinda Gates	1	3	2
3	Jhon Doe	3	2	3

ID	Key	Nombre	Cat	Age	Merca
1	1	David Smith	Adulto Joven	Web	LATAM
2	2	Melinda Gates	Adulto Joven	Partner	USA
3	3	Jhon Doe	Adulto Mayor	Partner	EU



ID	Key	Nombre	Cat	Age	Merca
1	1	David Smith	Adulto	Partner	USA
2	2	Melinda Gates	Adulto Joven	Partner	USA
3	3	Jhon Doe	Adulto Mayor	Call center	EU

# Dimensiones

## Cambios en el Origen - Tipo 2 (SCD2)

- Cada cambio en el sistema origen genera una nueva fila en la dimensión.
- Cada nuevo registro genera su propia llave.

ID	Key	Nombre	Cat	Age	Merca	StartDate	EndDate	Flag
1	1	David Smith	Adulto Joven	Web	LATAM	01/01/1999	31/12/2999	True
2	2	Melinda Gates	Adulto Joven	Partner	USA	01/01/1999	31/12/2999	True
3	3	Jhon Doe	Adulto Mayor	Partner	EU	01/01/1999	31/12/2999	True




ID	Key	Nombre	Cat	Age	Merca	StartDate	EndDate	Flag
1	1	David Smith	Adulto Joven	Web	LATAM	01/01/1999	18/06/2005	False
2	2	Melinda Gates	Adulto Joven	Partner	USA	01/01/1999	31/12/2999	True
3	3	Jhon Doe	Adulto Mayor	Partner	EU	01/01/1999	25/03/2017	False
4	1	David Smith	Adulto	Partner	USA	18/06/2005	31/12/2999	True
5	3	Jhon Doe	Adulto Mayor	Call Center	EU	25/03/2017	31/12/2999	True

# Dimensiones

## Cambios en el Origen - Tipo 3 (SCD3)

- Los cambios en el sistema origen son preservados en las dimensiones como un nuevo atributo.
- En la misma fila está el valor actual y el histórico.

I D	Key	Nombre	Cat	Cat_Hist	Age	Age_Hist	Merca	Merca_ Hist
1	1	David Smith	Adulto Joven	Adulto Joven	Web	Web	LATAM	LATAM
2	2	Melinda Gates	Adulto Joven	Adulto Joven	Partner	Partner	USA	USA
3	3	Jhon Doe	Adulto Mayor	Adulto Mayor	Partner	Partner	EU	EU



I D	Key	Nombre	Cat	Cat_Hist	Age	Age_Hist	Merca	Merca_ Hist
1	1	David Smith	Adulto	Adulto Joven	Partner	Web	USA	LATAM
2	2	Melinda Gates	Adulto Joven	Adulto Joven	Partner	Partner	USA	USA
3	3	Jhon Doe	Adulto Mayor	Adulto Mayor	Call Center	Partner	EU	EU



# DATA WAREHOUSE

## Fact o Hechos

 **EVENTOS**

Cada tabla de Hechos sólo debe hacer referencia a un tipo de evento.

**NUMÉRICAS**

Contiene medidas obtenidas dentro de un evento operativo.

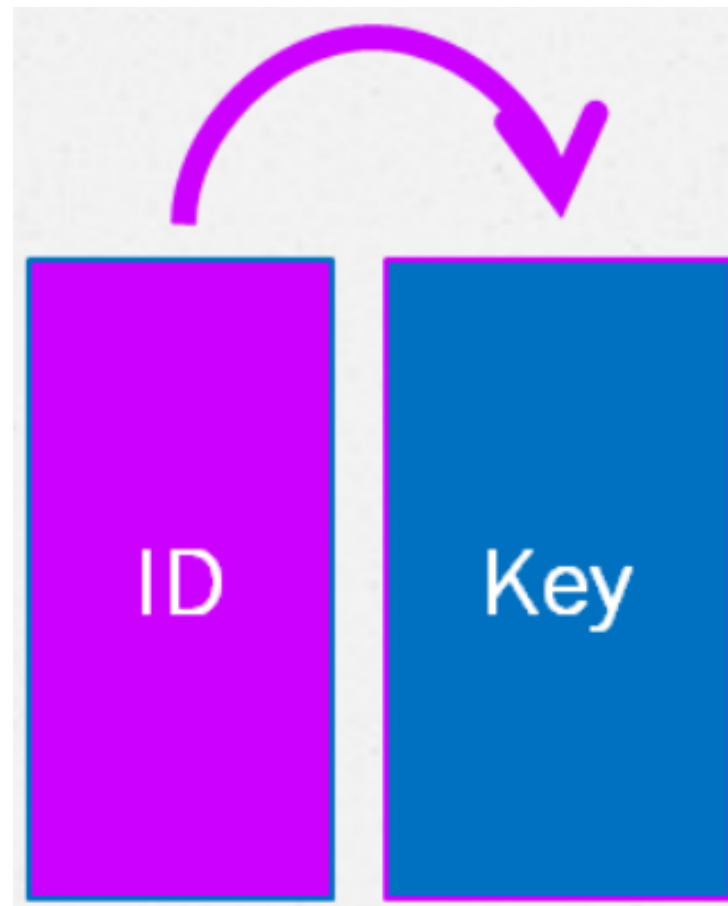
**MEDIDAS**

Las medidas SIEMPRE deben ser NUMÉRICAS (sumarizaciones y agregaciones)

# Tabla de Hechos

## Llaves Primarias

- Es el identificador de la tabla y como se relaciona con las dimensiones.
- Se recomienda que el Data Warehouse se encargue de crear las llaves durante el proceso **ETL**.

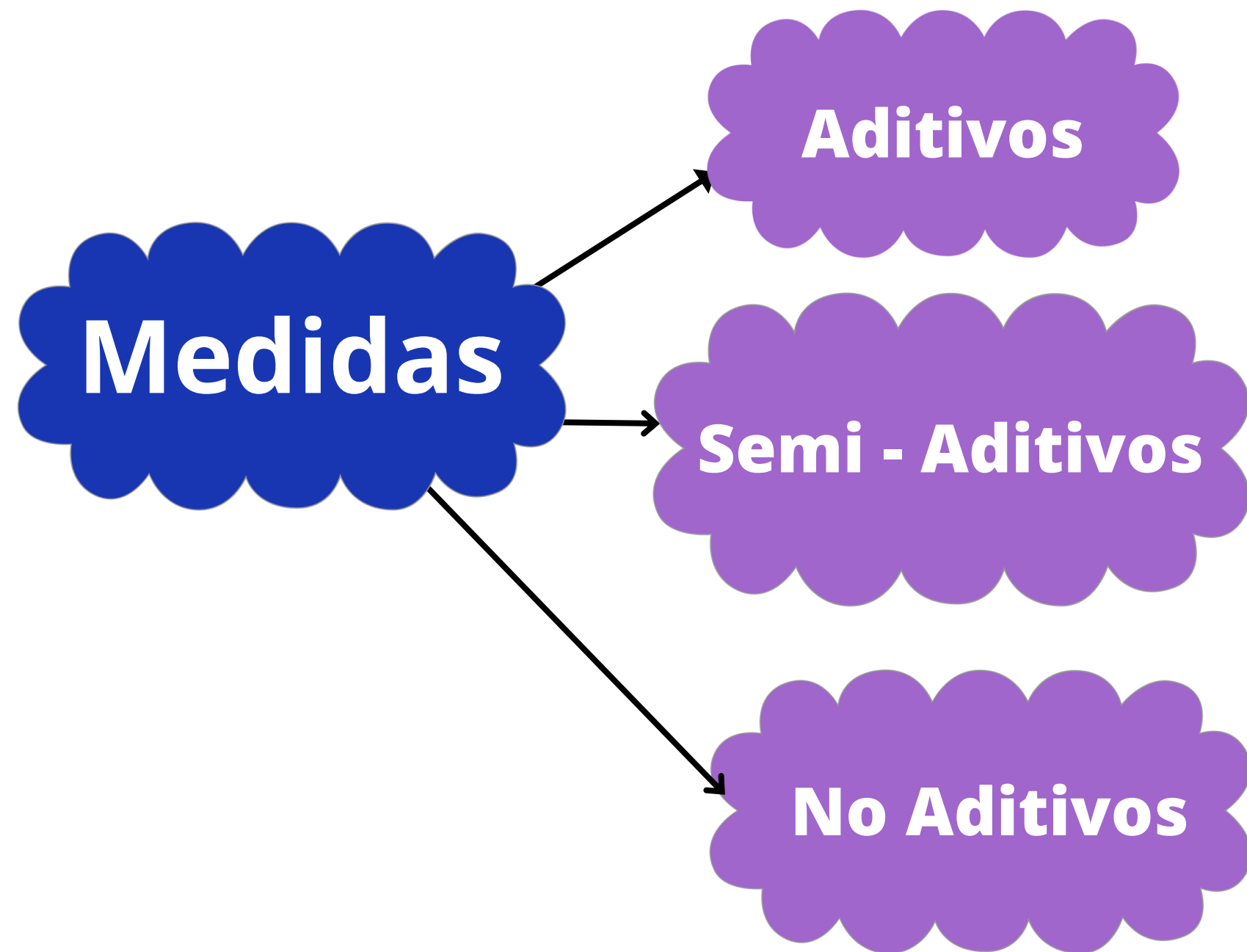


ID	Order	Date	Sale	Customer
1	8237834	01/03/2005	263	1
2	8384589	16/05/2007	373	2
3	3494503	31/12/2015	395	3



ID	Key	Order	Date	Sale	IDCustomer
18	1	8237834	01032005	263	5
19	2	8384589	16052007	373	7
20	3	3494503	31122015	395	10

# Tabla de Hechos



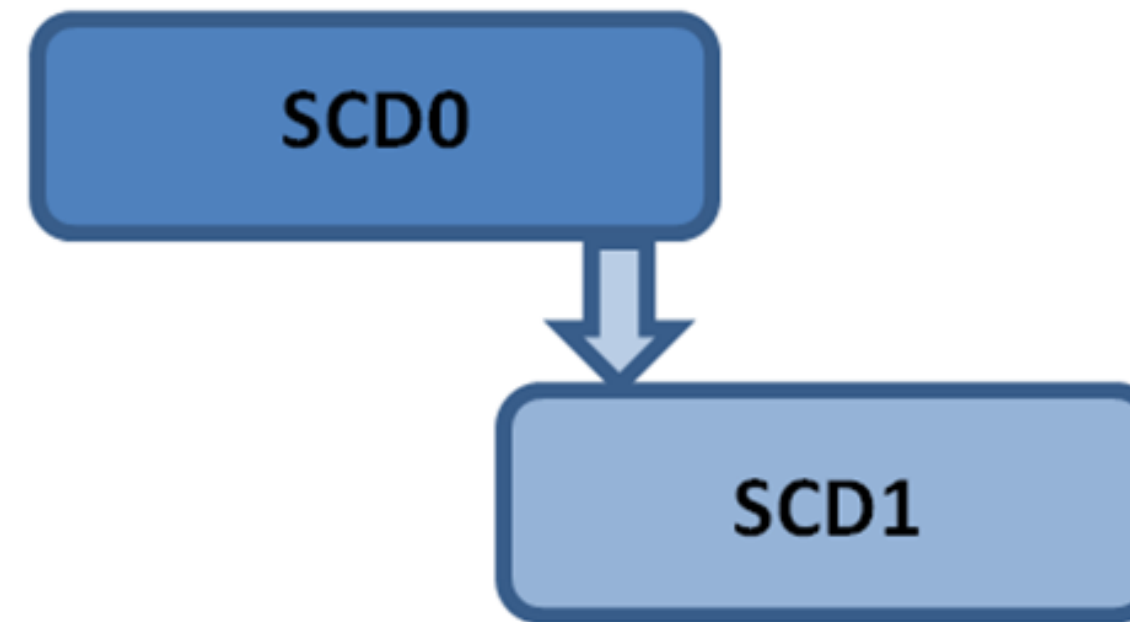
- Las medidas se sumarizan en cualquier dimensión. Por ejemplo: Totales de una venta
- Las medidas se sumarizan en algunas dimensiones. Por ejemplo: Balances.
- Las medidas no pueden ser sumarizadas en ninguna dimensión. Por ejemplo. Márgenes de utilidad.



# Dimensiones

## Cambios en el Origen

- Los eventos están propensos a cambiar en el tiempo.
- La forma en que se manejan los cambios dependen del negocio.
- Suelen almacenarse el evento tal como sucedió SCD0, o sobre escribirse los datos antiguos para dejar los más actuales.



ID	Order	Date	Customer	Sale
1	837634	01/12/2002	2	782
2	763476	18/03/2005	3	231
3	763489	23/09/2016	3	934



ID	Order	Date	Customer	Sale
1	837634	25/12/2002	3	782
2	763476	18/03/2005	3	231
3	763489	23/09/2016	3	570

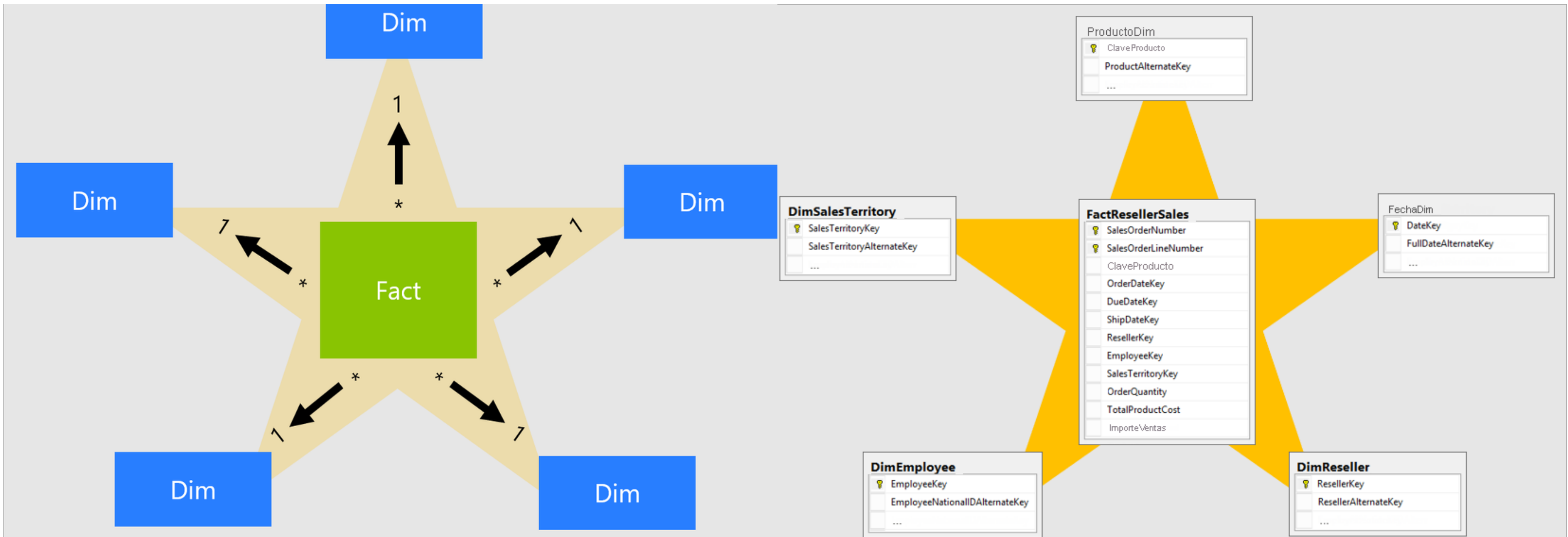
# Práctica

Pasar de un Diagrama de Entidad-Relación a un Diagrama Estrella.

*Grupos de hasta 4. Un correo por grupo. Fecha límite: Miércoles 27/9,*

*Mail: carina.cozzolino@um.edu.ar*

Ejemplo:





**Data Warehouse**

# ¡Gracias!

**Ing. Carina Cozzolino**