**Resumen prueba 2 Optativo**

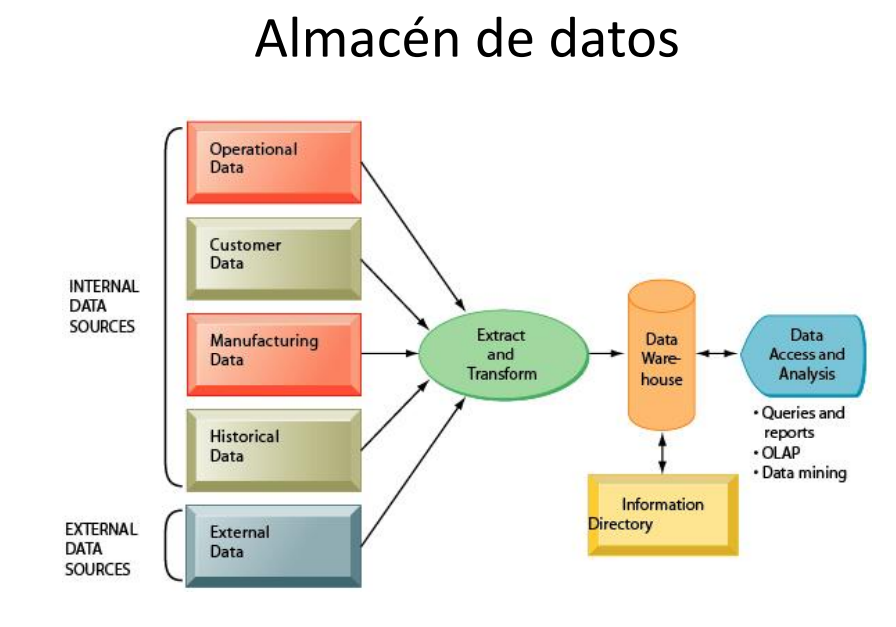
**En el diseño de un data warehouse hay que partir de una serie**

**de características:**

•Administra grandes cantidades de información

•Guarda histórico de datos

•Condensa y agrega información

•Integra y asocia información de varias fuentes

**Este tipo de modelo de datos consta principalmente de dos**

**tipos de elementos:**

**– DIMENSIONES:** Representan factores por lo que se analiza un

determinada área del negocio. Son pequeñas y usualmente están

desnormalizadas.

**– HECHOS:** Son el objeto de los análisis y están relacionados con las

dimensiones. Son tablas muy grandes y suelen estar desnormalizadas.

En un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones)

**Hechos:** colección de ítems de datos y datos de contexto. Cada hecho representa un ítem

de negocio, una transacción o un evento

• Los hechos se registran en las tablas CENTRALES del DW

**Dimensiones**

• Una dimensión es una colección de miembros o unidades o individuos del mismo tipo

• Cada punto de entrada de la tabla de HECHOS está conectado a una DIMENSION

• Determinan el contexto de los HECHOS

• Se utilizan como parámetros para los análisis OLAP

• Dimensiones habituales son: – Tiempo – Geografía – Cliente – Vendedor

**Medida:** es un atributo numérico de un hecho que representa la performance o comportamiento del negocio relativo a la dimensión

• Ejemplos:

– Ventas en $$

– Cantidad de productos

– Total de transacciones, etc.

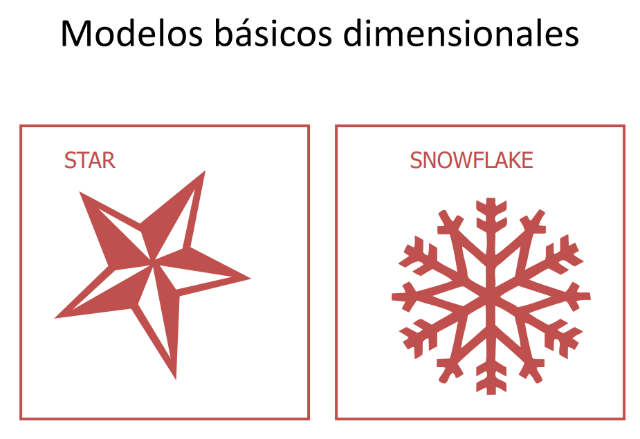
Dependiendo de sus propiedades podemos tener tres medidas:

➢ **Aditivas**, pueden ser combinadas a lo largo de cualquier

dimensión. Por ejemplo, “temperatura”, que puede estar dada por las dimensiones estación, región y fecha.

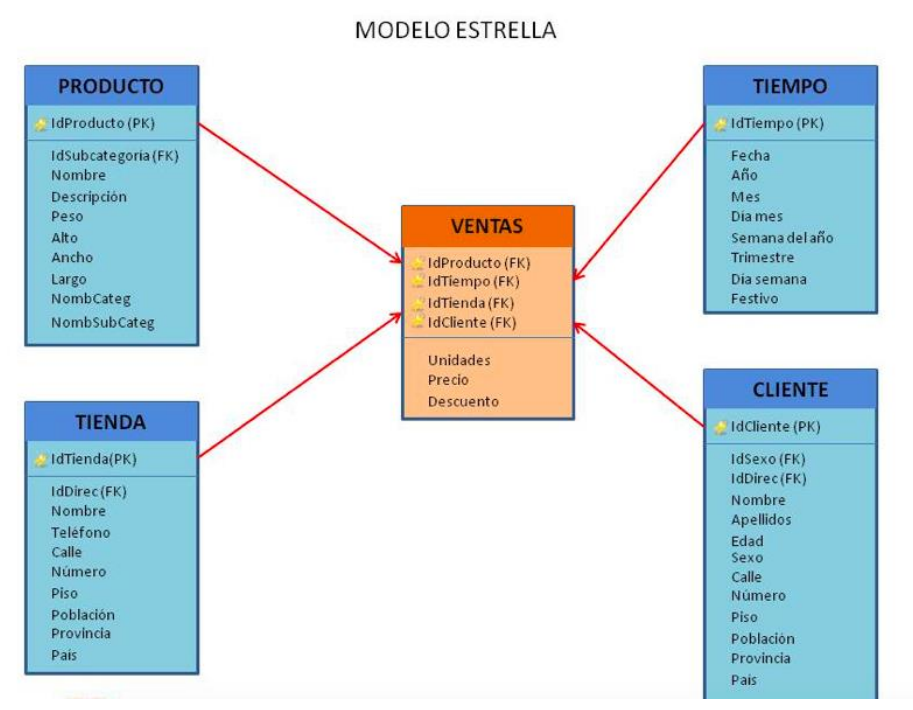
➢ **Semi-aditivas**, pueden no ser combinadas a lo largo de una o más dimensiones. Por ejemplo, “nomina” que puede estar dada por las dimensiones empleados y tiempo, pero no producto.

➢ **No aditivas**, no pueden combinarse a lo largo de ninguna dimensión. Por ejemplo, “cantidad de producto”, que únicamente puede estar dada por la dimensión producto.



**Esquema Estrella**

• También llamado dimensional, contiene una tabla de hechos que es aquella que contiene toda la información y tiene varias tablas de dimensiones que contienen el catálogo de la información.



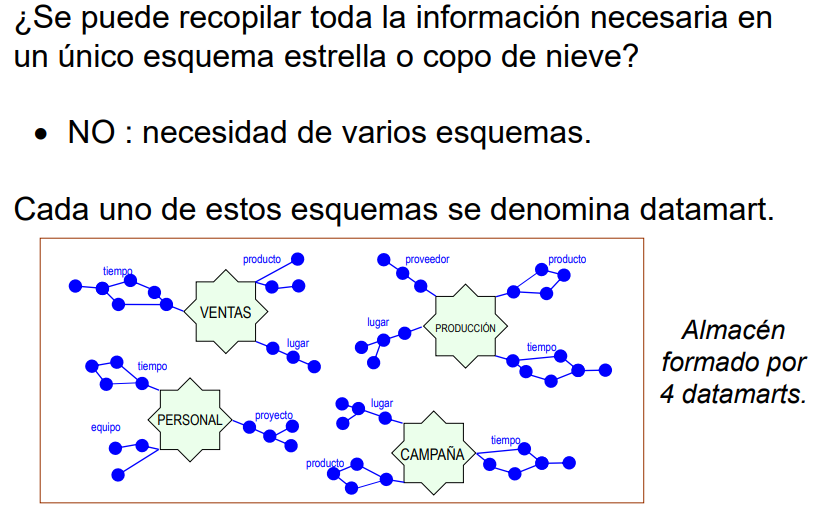
**Esquema Copo de Nieve**

• El diseño es ad hoc y es difícil de cambiar (aplica también en estrella).

• Es una variante del esquema de estrella, el cual normaliza todas las tablas de dimensiones.

• Tiene algunas mejoras de espacio, pero en ocasiones las consultas son más lentas.





**Se pueden plantear dos tipos de datamarts:**

**OLAP**

Se basan en los populares cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

**OLTP**

Es un simple extracto del datawarehouse, no obstante, lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa.

**Ventajas del datamarts**

• Poco volumen de datos

• Mayor rapidez de consulta

• Consultas SQL y/o MDX sencillas

• Validación directa de la información

• Facilidad para la historización de los datos

Modelos de Datamarts

* **Top Down:** tiene como base un sistema de Data warehouse para toda la empresa y a partir de este se desarrollan los Data marts para las divisiones o departamentos.
* **Modelo Bottom up:** se refiere a la arquitectura de un almacén de datos, en el cual se definen previamente los Data Marts y luego se integran en un Data Warehouse centralizado.
* **El Modelo Paralelo:** Consiste en llevar una construcción paralela entre ambas estructuras, sin definir como padre la Data warehouse sobre el Datamart, ni viceversa.
* **Modelo Top down con Retroalimentación:** incluye la posibilidad de agregar datos que no estén presentes en la Data warehouse y que se requieran para cumplir con los requerimientos de un análisis específico.
* **Modelo Bottom up con Retroalimentación:** Permitir que la integración de los datos e información dada en los Data marts se pueda implementar durante la construcción de la Data warehouse.
* **Modelo Paralelo con Retroalimentación:** Data mart, ya que su inicio depende de un período de ajuste entre éste y el modelo de datos de la Data warehouse.

**Datawarehouse se caracteriza por ser:** Integrado, Temático, Histórico, No Volátil

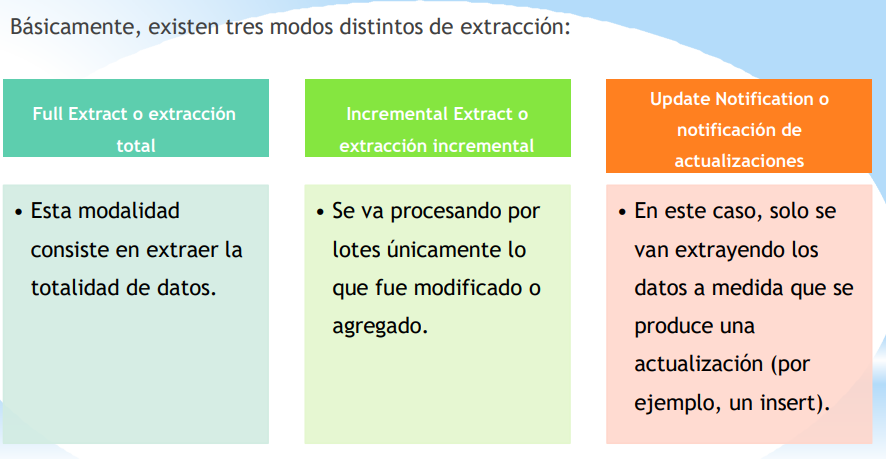
**Característica del datawarehouse**

• Es que contiene metadatos, es decir, datos sobre los datos. Los metadatos permiten saber la procedencia de la información, su periodicidad de refresco, su fiabilidad, forma de cálculo... etc.

FASE DE EXTRACCIÓN

La primera parte del proceso ETL consiste en extraer los datos desde el sistema de origen.

Esta fase consta de tres pasos:

* Identificación de los archivos
* Extracción de Datos
* LIMPIEZA

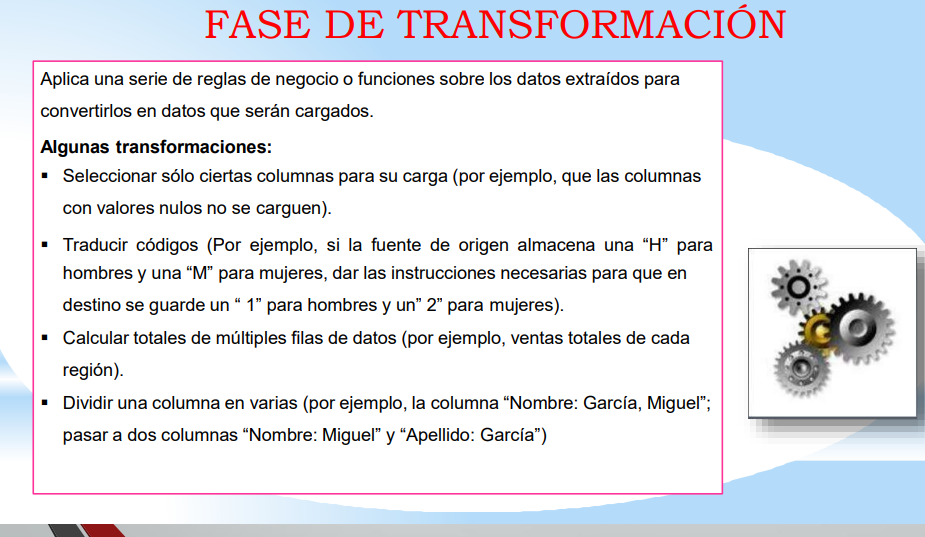
La limpieza de datos se divide en distintas etapas, que vamos a describir a continuación:

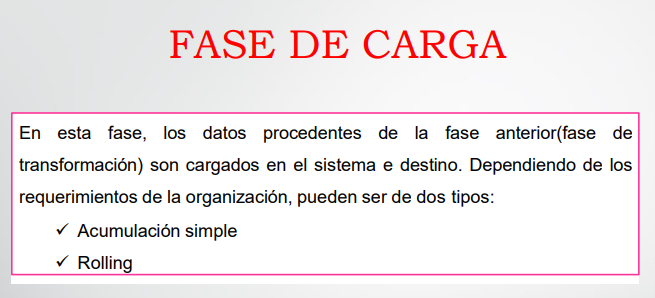
Corregir: Por ejemplo: comprueba una dirección y el código postal correspondiente.

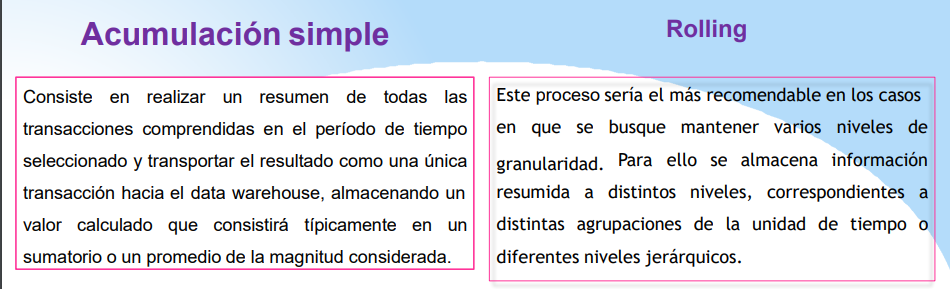
Estandarizar: Formatos definidos. Por ejemplo: trato de Sr., Sra.

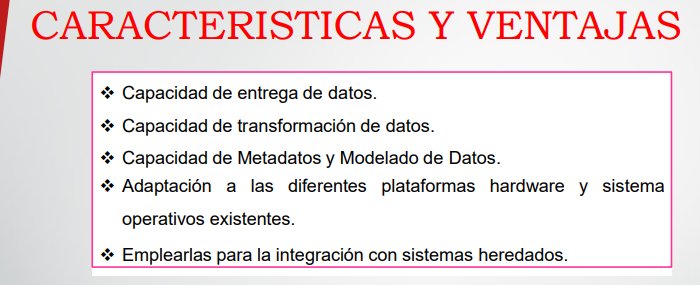
Relacionar: Eliminar duplicados. Por ejemplo: identificando nombres y direcciones iguales.

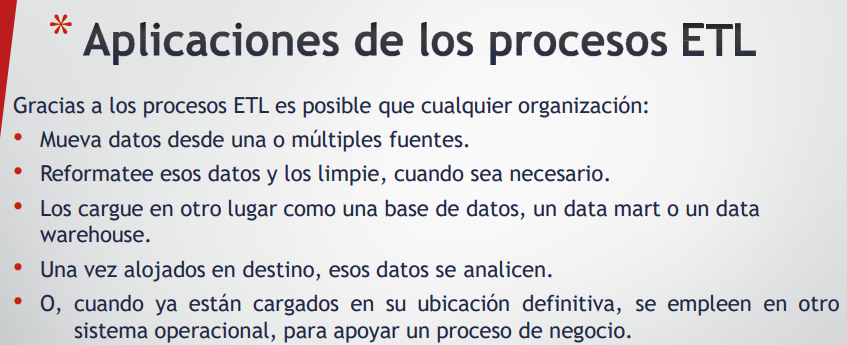
Consolidar: Analiza e identifica sola representación.

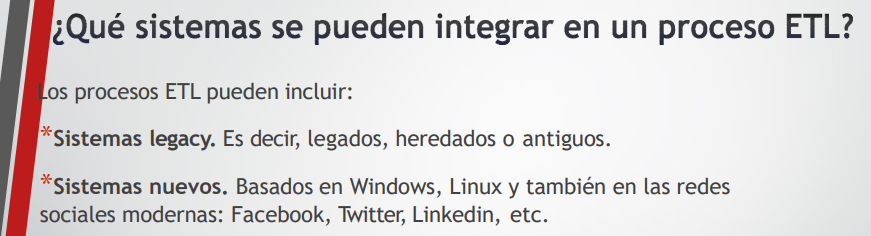


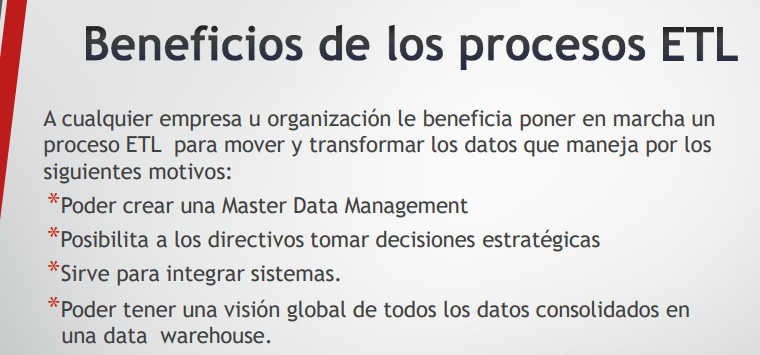












**Wrapper:** programa especializado que identifica los datos de interés sobre la base de varias reglas gramaticales, que luego de extraerlos los transforma a una estructura de datos apropiada para su posterior manipulación

**Granularidad:** especificidad a la que se define un nivel de detalle en una tabla, es decir, si hablamos de una jerarquía la granularidad empieza por la parte más alta de la jerarquía, siendo la granularidad mínima, el nivel más bajo. En Data Warehouse, no solo existe granularidad para las tablas de hechos, también existe granularidad en las dimensiones.

**Definiciones y Conceptos OLAP**

OLAP (procesamiento analítico en línea / OnLine Analytical Processing) es una solución utilizada en la Inteligencia de negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos para extraer algún tipo de información útil.

**Tipos de Sistemas OLAP**

1. **ROLAP** (Procesamiento Analítico Relacional en Línea): Son sistemas en los cuales los datos se encuentran almacenados en una base de datos relacional.

2. **MOLAP** (procesamiento Analítico multidimensional en Línea): En estos sistemas los datos se encuentran almacenados en una estructura de datos multidimensional.

3. **HOLAP** (Procesamiento Analítico Híbrido en Línea): En estos sistemas mantienen los registros detallados en la base de datos relacional, mientras que los datos resumidos o agregados se almacenan en una base de datos multidimensional separada.

**Utilidades de OLAP**

• Tienen acceso a grandes cantidades de datos.

• Analizan las relaciones entre muchos tipos de elementos empresariales.

• Involucran datos agregados.

• Comparan datos agregados a través de periodos jerárquicos.

• Presentan datos en diferentes perspectivas.

• Involucran cálculos complejos entre elementos de datos.

• Pueden responder con rapidez a consultas de usuarios.

**Requerimientos Funcionales de los Sistemas OLAP**

•Construir un Datawarehouse y/o Datamart.

• Uso de herramientas de Almacenamiento (bases de datos multidimensionales), Herramientas de Extracción y Colección, Herramientas para Reportes de Usuario Final y Herramientas para Análisis Inteligentes.

• Un servidor que sea altamente escalable.

• Un segundo servidor para las herramientas de consulta de datos.

**Lo realmente interesante de las herramientas OLAP son sus operadores de refinamiento o**

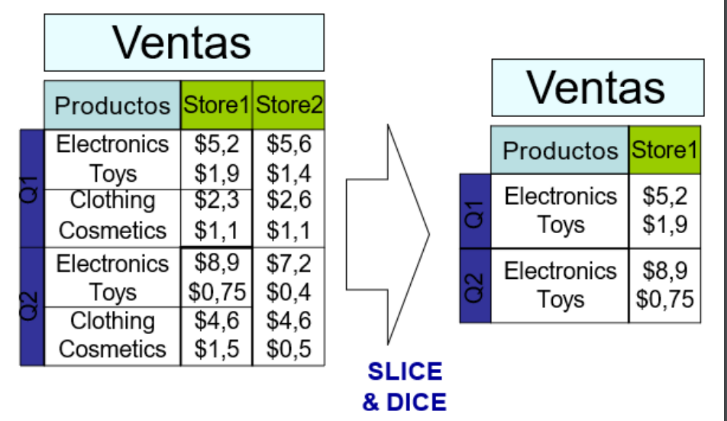
**manipulación de consultas.**

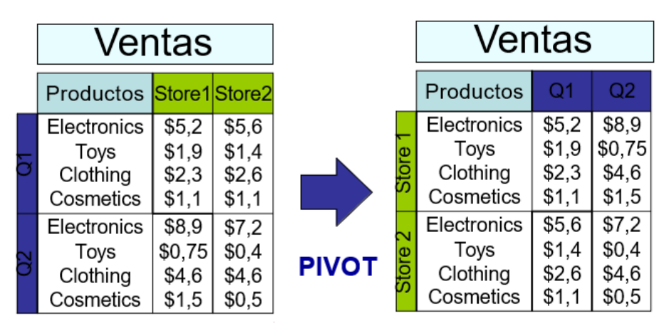
• DRILL ROLL • SLICE & DICE PIVOT

**El carácter agregado de las consultas en el Análisis de Datos aconseja la definición de nuevos operadores que faciliten la agregación (consolidación) y la disgregación (división) de los datos:**

•agregación (roll): permite eliminar un criterio de agrupación en el análisis, agregando los grupos actuales.

• disgregación (drill): permite introducir un nuevo criterio de agrupación en el análisis, disgregando los grupos actuales.

• SLICE & DICE: seleccionar y proyectar datos en el informe.

• PIVOT: reorientación de las dimensiones en el informe.

Las herramientas de OLAP se caracterizan por:

* Ofrecer una visión multidimensional de los datos (matricial).
* No imponer restricciones sobre el número de dimensiones.
* Ofrecer simetría para las dimensiones.
* Permitir definir de forma flexible (sin limitaciones) sobre las dimensiones: restricciones, agregaciones y jerarquías entre ellas.
* Ofrecer operadores intuitivos de manipulación: drill-down, roll-up, sliceand- dice, pivot.
* Ser transparentes al tipo de tecnología que soporta el almacén de datos (ROLAP o MOLAP).