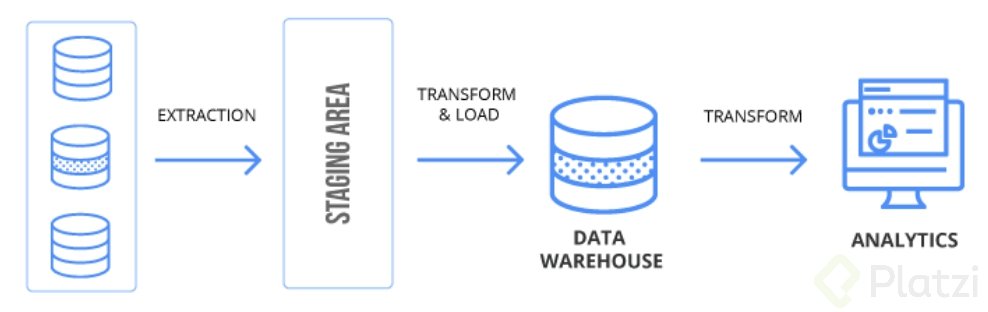
**TAREA 31: DICCIONARIO**

* **ETL (EXTRACT TRANFORM LOAD):**



Normalmente los datos obtenidos para las analíticas no están preparados para ser analizados, por que están en diferentes formatos, tienen valores erróneos, caracteres inválidos etc. Para solucionarlo, se abre el proceso ETL, que por sus siglas en ingles significa extraer los datos crudos desde su origen (source), transformarlos según nuestras necesidades de la analítica o la estructura que deseamos y cargarlos a una base de datos orientada a procesos analíticos (Target).

* EXTRACT ( Proceso de Extracción):

Nuestra fuente (source) puede contener múltiples fuentes de datos: Bases de datos, archivos CSV…Hay que tener bien claro el origen de nuestros datos para buscar la mejor manera de extraerlos.

Si tenemos que extraer millones de datos de una única base de datos, si lo hacemos con un “select” estresaremos el motor de datos, por lo que mejor ejecutar un proceso por lotes, para interferir lo mínimo posible. El éxito de una buena extracción es que cause el menor impacto en el origen de los datos. Hay dos formas de extracción:

1. Total: Cada ejecución extrae en una única llamada la totalidad de los datos.
2. Incremental: Se extraen los datos en pequeños lotes múltiples veces. Por ejemplo, una ETL que solo consulta los datos del día anterior a su ejecución para trabajar con ellos.

* TRANSFORM (TRANSFORMAR): Se aplican las reglas para realizar un buen proceso de analítica que se necesitan en cada momento. Estas incluyen procesos como:

1. Filtrar filas por ciertas características
2. Eliminar duplicados
3. Transformar datos (p.ej. cambiar nombres países por números)
4. Calcular datos nuevos (edad con la fecha de nacimiento)
5. Unir o combinar datos de distintas fuentes
6. Pivotar las tablas
7. Dividir columnas (Nombre completo en nombre y apellido)

Estas transformaciones no se realizan ni en source ni en target, se realizan en el área de “Staging”, un repositorio temporal para procesar los datos, funcionar por medio de tablas o archivos planos dependiendo de la ETL. Una vez los datos pasan al destino, este repositorio temporal es eliminado.

* LOAD (CARGA): Nuestros datos ya están transformados y listos en el área staging, así que ahora debemos proceder a cargarlos a nuestra base de datos analítica. Comúnmente es un DATAWAREHOUSE en donde conviven distintos repositorios de datos de una manera no normalizada (como en una base de datos relacional) y con una estructura lista para realizar el análisis de datos.

En este punto también debemos tener en cuenta que motor corre nuestra base de datos de destino, pues esto definirá la mejor manera de cargar los datos. Por ejemplo, una BBDD en Redshift es mucho más eficiente haciendo un “Copy” basados en un archivo plano y no realizando tareas de “insert” registro a registro.

Como los datos son cargados en destino, hay que intentar resumir los datos obtenidos. Hay dos maneras:

1. ACUMULACION SIMPLE: Realizar un resumen de transacciones como sumatorio o promedio, o utilizar una periodicidad como compras/mes.
2. ROLLING: Interesante cuando queremos incluir diferentes niveles de granularidad, como totales diarios, semanales… o queremos almacenar diferentes jerarquías entre las dimensiones de nuestra BBDD, como ventas por periodo, ventas por producto…

* **BASES DE DATOS:**

Conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. Se pueden clasificar según diferentes propiedades:

1. SEGÚN SU FLEXIBILIDAD: Bases de datos dinámicas, que pueden actualizarse sus datos y estáticas.
2. SEGÚN SU FORMA DE ORGANIZACIÓN: Bases de datos jerárquicas, organizadas en forma de árbol al revés, bases de datos de red, como las jerárquicas, pero un hijo puede tener varios padres, bases de datos relacionales donde los datos se organizan en tablas que pueden asociarse entre ellas, bases de datos deductivas o lógicas, que se utilizan generalmente en buscadores y multidimensionales, conceptualmente la idea de un cubo donde las informaciones se almacenan en la intersección de tres o mas atributos.

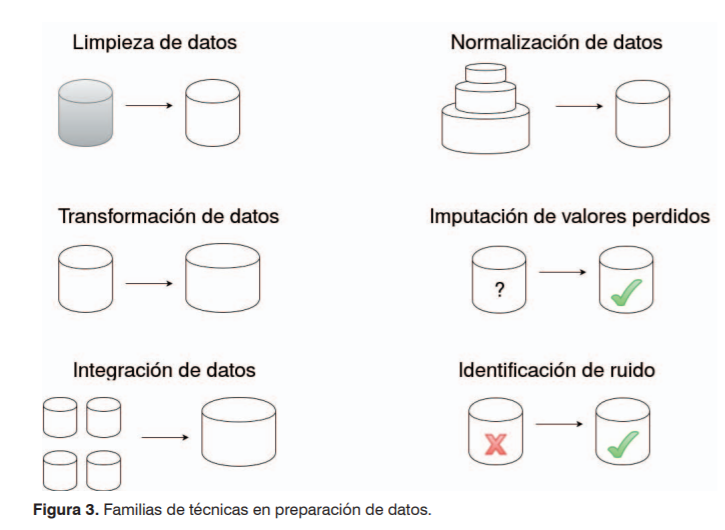
* **DATA PREPROCESSING:**

El procesamiento s de datos es una etapa esencial del proceso de descubrimiento de información o KDD (Knowledge Discovery in Databases). Esta etapa se encarga de la limpieza de datos, su integración, transformación y reducción para la siguiente fase de minería de datos.

Debido a que normalmente el uso de datos de baja calidad implica un proceso de minería de datos con pobres resultados, se hace necesaria la aplicación de técnicas de preprocesamiento.

Después de la aplicación de la fase de preprocesamiento, el conjunto resultante puede ser visto como una fuente consistente y adecuada de datos de calidad para la aplicación de algoritmos de minería de datos. El preprocesamiento incluye un rango amplio de técnicas que podemos agrupar en dos áreas: preparación de datos y reducción de datos.

La preparación de datos esta formada por una serie de técnicas que tienen el objetivo de inicializar correctamente los datos que servirán de entrada para los algoritmos de minería de datos. Este tipo de técnicas pueden clasificarse como de uso obligado, ya que sin ellas los algoritmos de extracción de conocimiento no podrían ejecutarse u ofrecerían resultados erróneos. En esta área se incluye la transformación de datos y normalización, integración, limpieza de ruido e imputación de valores perdidos.



Las técnicas de reducción de datos se orientan a obtener una representación reducida de los datos originales, manteniendo en la mayor medida posible la integridad y la información existente en los datos.

Es por esta razón que la aplicación de técnicas de reducción no se considera estrictamente obligatoria. Sin embargo, cuando el tiempo de ejecución de un algoritmo o el tamaño de los datos son prohibitivos para los algoritmos de extracción, las técnicas deben ser aplicadas para obtener conjuntos de datos mas pequeños y de calidad.

En esta área las técnicas de reducción mas relevantes son: la selección de atributos (Feature selection, FS), la selección de instancias (Instance selection, IS) o la discretización.

