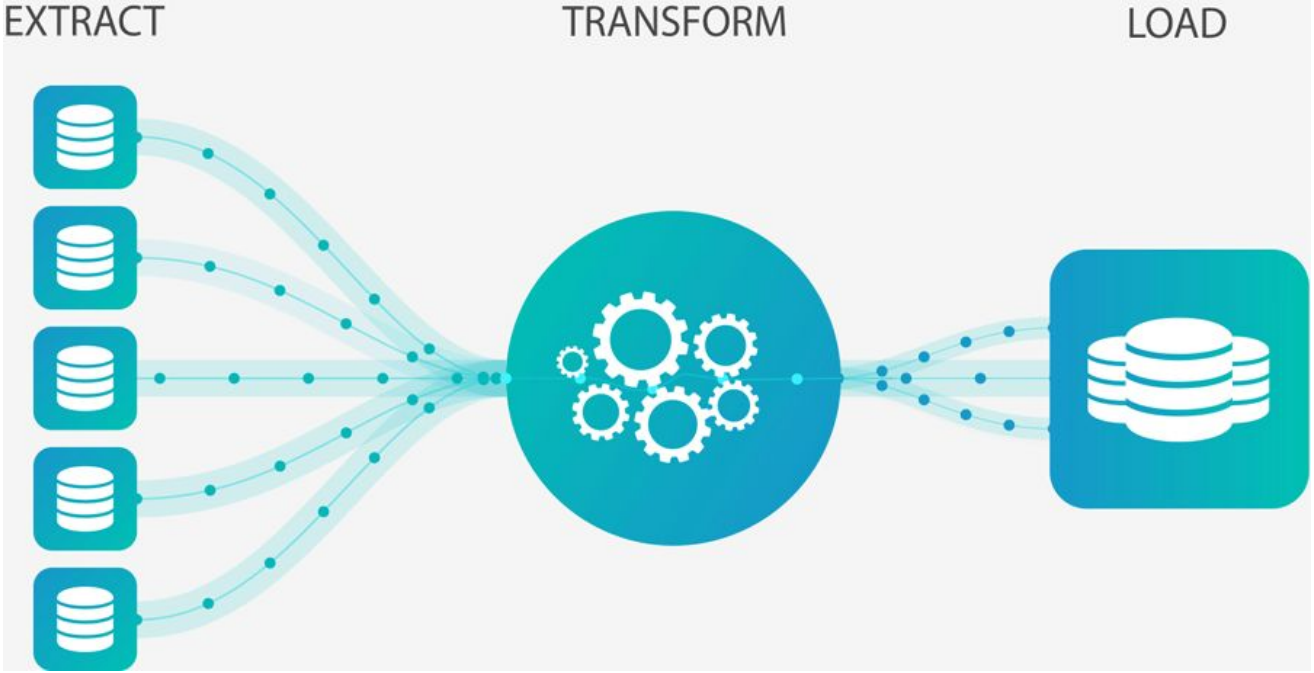
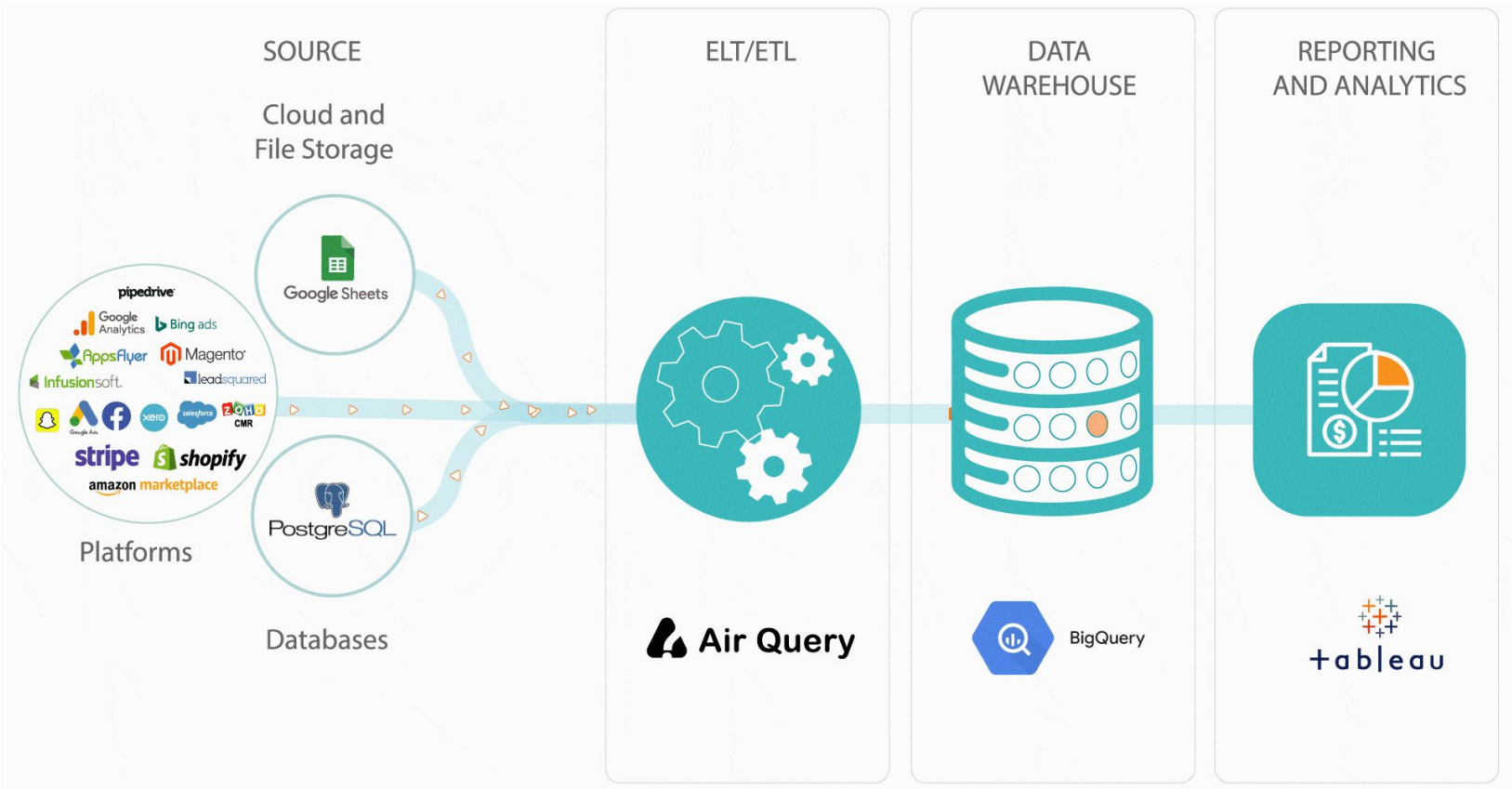
Big Data Aplicado

Curso de especialización en Inteligencia Artificial y Big Data.

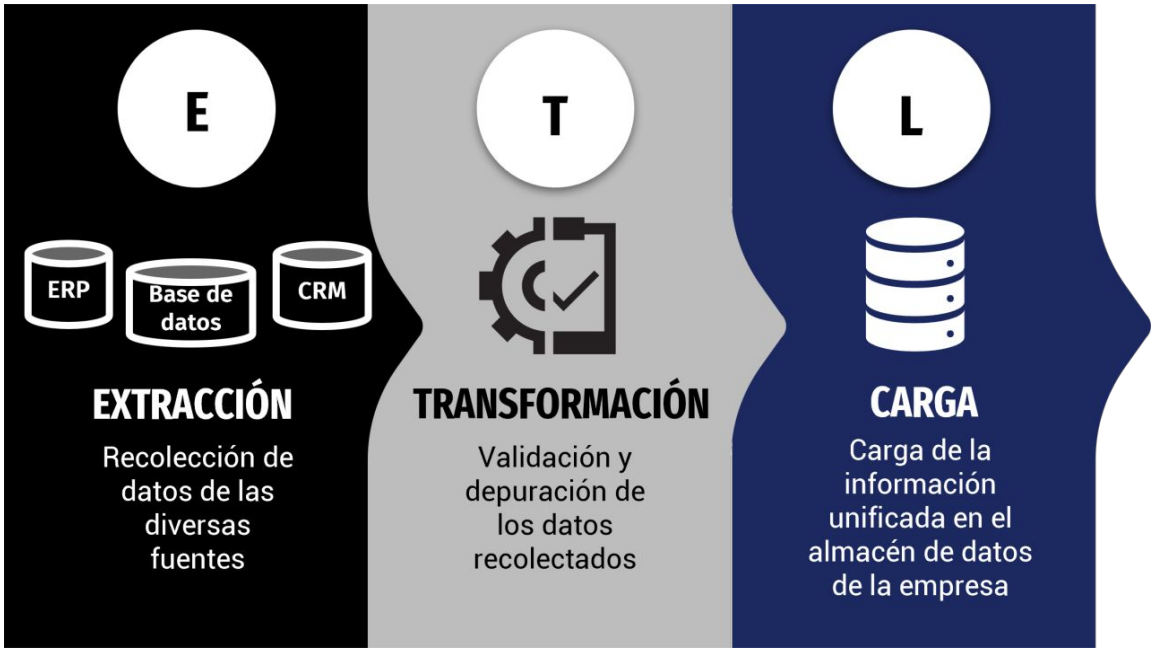
El proceso ETL. Introducción



El proceso ETL. Introducción



El proceso ETL. Introducción



Almacenes de datos. Diseño e implementación. El proceso de Extracción, Transformación y Carga (ETL)

El proceso ETL es el encargado de extraer, limpiar e integrar los datos provenientes de las fuentes de datos para alimentar el almacén de datos. Este proceso también es el encargado de alimentar la capa de datos reconciliados en la arquitectura de tres capas. El proceso ETL tiene lugar cuando se prueba el almacén de datos y se lleva a cabo cada vez que el almacén de datos se actualiza. A continuación, se describen detalladamente cada una de las fases de las que consta este proceso.

El proceso de Extracción, Transformación y Carga (ETL)

E**xtracción:** Etapa que consiste en la lectura de los datos de las distintas fuentes de las que provienen.

1. Cuando un almacén de datos se rellena por primera vez, se suele utilizar la técnica de extracción estática, la cual consiste en extraer una instantánea de los datos operacionales.
2. A partir de entonces, se utiliza la extracción incremental para actualizar periódicamente los datos del almacén de datos, recogiendo los cambios aplicados desde la última extracción. Para ello, **se utiliza el registro mantenido por el SGBD** que, por ejemplo, asocia una marca de tiempo (timestamp) a los datos operacionales para registrar cuando fueron modificados y agilizar el proceso de extracción.

En la actualidad, **existe una gran cantidad de conjuntos de datos o data sets públicos, conocidos bajo el nombre de Open Data**, que abarcan una gran cantidad de dominios y con los que es posible trabajar para construir soluciones big data.

El proceso de Extracción, Transformación y Carga (ETL)

Open Data: Se trata de datos que han sido generados por una fuente en particular, que abarcan un dominio temático o disciplinar y tienen atributos, dentro de los cuales está la frecuencia de actualización. Además, cuentan con una licencia específica que indica las condiciones de reutilización de los mismos.

El proceso de Extracción, Transformación y Carga (ETL)

La **fuente de los datos** es en muchos de los casos datos **del estado nacional, provincial, municipal u organizaciones comerciales**. En otras ocasiones, l**a fuente de los datos es fruto del estudio o medición por parte de particulares. Los atributos de los conjuntos de datos deben especificar cómo fueron obtenidos, incluyendo fechas de obtención, actualización y validez, así como el público involucrado, la metodología de recogida o muestreo, etc.**

Algunas de las fuentes más utilizadas en la actualidad para la obtención de datos abiertos provienen de los centros nacionales e internacionales de estadísticas, como son:

- el Instituto Nacional de Estadística de España (INE),
- eurostat,
- la oficina europea de estadísticas,
- la Organización Mundial de la Salud (OMS)...

El proceso de Extracción, Transformación y Carga (ETL)

Por otra parte, existen repositorios públicos de datos abiertos y a disposición de los usuarios como UCI o Kaggle, que es una comunidad de científicos de datos donde empresas y organizaciones suben sus datos y plantean problemas que son resueltos por los miembrc de la comunidad.

- https://www.ine.es
- https://ec.europa.eu/eurostat
- https://www.who.int/es/data/gho/publications/world-health-statistics
- https://archive.ics.uci.edu/ml/index.php
- https://www.kaggle.com

Figura 2.4: Fuentes para la obtención de datos abiertos.

El proceso de Extracción, Transformación y Carga (ETL)

Transformación:

La etapa de transformación es la fase clave para **transformar los datos operativos en datos con un formato específico para alimentar un almacén de datos.** En esta etapa, **los datos se limpian y se transforman, añadiéndoles contexto y significado.** En caso de implementar un almacén de datos siguiendo una arquitectura de tres capas, el proceso de transformación es el encargado de obtener la capa de datos reconciliados.

> La etapa de **transformación** engloba todos los procesos de limpieza y manipulación de los datos, con el objetivo de transformar los datos operativos propios de sistemas relacionales (OLTP) en datos preparados para ser incluidos dentro del almacén de datos (OLAP).

El proceso de Extracción, Transformación y Carga (ETL)

Transformación:

La limpieza de los datos o data cleaning engloba todos aquellos procedimientos necesarios para **detectar y resolver situaciones problemáticas con los datos de partida que pudieran suponer problemas potenciales a la hora de analizarlos.** Así pues, **los datos de partida pueden ser incompletos, es decir, pueden contener atributos sin valor o valores agregados, incorrectos que incluyan errores o valores sin ningún significado,** lo cual es común cuando los datos se introducen manualmente en el sistema, o inconsistentes cuando los cambios no son propagados a todos los módulos del sistema, los rangos de un determinado atributo son cambiantes, existen datos duplicados...

El proceso de Extracción, Transformación y Carga (ETL)

Transformación:

En general, se distinguen dos tipos de situaciones cuando existen valores perdidos:

- datos perdidos completamente aleatorios y
- datos perdidos de forma no completamente aleatoria.

En el segundo caso, puede ser interesante **intentar analizar la razón de la pérdida de los datos,** la cual puede ser indicativa. En muchas ocasiones, **los valores perdidos tienen relación con un subconjunto de variables predictoras y no se encuentran aleatoriamente distribuidos por todas ellas.**

Por todo ello, las aproximaciones más comunes a la hora de gestionar datos perdidos son: **Eliminación de instancias que contengan valores perdidos y Asignación de valores fijos, Asignación de valores de referencia, Imputación de valores perdidos**

El proceso de Extracción, Transformación y Carga (ETL)

Transformación:

Eliminación de instancias que contengan valores perdidos: la eliminación de instancias que contienen valores perdidos implica establecer un umbral de porcentaje, y si una instancia tiene un número de valores perdidos que supera este umbral, se elimina. Esta técnica es útil para conjuntos de datos grandes con pocos valores perdidos, pero debe usarse con precaución, ya que puede resultar en una pérdida significativa de información en conjuntos de datos más pequeños, con muchas instancias con valores perdidos o con un umbral muy bajo.

Asignación de valores fijos: consiste en asignar un valor fijo a todos los valores perdidos de todas las variables. Este valor puede ser el número 0 o incluso un valor desconocido Unknown o no numérico en función del lenguaje de programación utilizado.

El proceso de Extracción, Transformación y Carga (ETL)

Transformación:

Asignación de valores de referencia: **asigna un valor de referencia a los valores perdidos para cada variable.** Estos valores de referencia suelen ser medidas de centralización como la media o la mediana de los valores de cada variable.

Imputación de valores perdidos: consiste en la aplicación de técnicas más sofisticadas, como pueden ser técnicas estadísticas o de aprendizaje automático para predecir o averiguar los valores que se han perdido.

El proceso de Extracción, Transformación y Carga (ETL)

Finalmente, **la etapa de limpieza de datos** también se encarga de la **detección de valores anómalos o outliers**. Se trata de valores que se han introducido de forma errónea o bien a una deformación en la distribución de valores.

El proceso de detección de anomalías consiste, fundamentalmente, en dos etapas:

- **en primer lugar, asumir que existe un modelo generador de datos, como podría ser una distribución de probabilidad.**
- **En segundo lugar, considerar que las anomalías representan un modelo generador distinto, que no coincide con el original.** Existen multitud de técnicas para detectar y descartar o imputar valores anómalos, como lo son técnicas estadísticas basadas en la desviación y el rango intercuartílico o técnicas de aprendizaje automático.

El proceso de Extracción, Transformación y Carga (ETL)

Procesos de transformación más comunes:

- Estandarización de códigos y formatos de representación.
- Conversiones y combinaciones de campos.
- Correcciones.
- Integración de varias fuentes.
- Eliminación de datos y/o registros duplicados.
- Escalado y centrado.
- Discretización.
- Selección de características.

Documento → Ebook Procesos ETL - La base de la Inteligencia de Negocio

El proceso de Extracción, Transformación y Carga (ETL)

Carga:

Se trata de la **última fase de cara a incluir datos en el almacén de datos**. La carga inicial de los datos puede requerir bastante tiempo al cargar de forma progresiva todos los datos históricos, por lo que es normal realizarla en horas de baja carga de los sistemas. Una vez que el almacén de datos ha sido inicialmente cargado, las sucesivas cargas de datos se pueden realizar de dos formas:

Refresco: El almacén de datos se reescribe completamente, de forma que se reemplazan los datos antiguos. El refresco es utilizado habitualmente junto con la extracción estática y suele ser una estrategia muy utilizada para la carga inicial del almacén de datos, aunque puede también realizarse a posteriori.

Actualización: Se añaden al almacén de datos solamente aquellos datos nuevos que se pretenden incluir, sin eliminar ni modificar los datos ya existentes. Esta técnica es utilizada frecuentemente junto con la extracción incremental para actualizar regularmente el almacén de datos.

El proceso de Extracción, Transformación y Carga (ETL)

Frameworks ETL

Bubbles (http://bubbles.databrewery.org): Se trata de un framework ETL escrito en Python, aunque es posible utilizarlo desde otros lenguajes.

Apache Camel (https://camel.apache.org): Este framework escrito en lenguaje Java y de acceso abierto se enfoca en facilitar la integración entre distintas fuentes de datos, haciendo el proceso más accesible a los desarrolladores, ofreciendo distintas herramientas para dar soporte al proceso ETL.

Keetle (https://community.hitachivantara.com/s/article/data-integration-kettle): Es el framework de Pentaho para dar soporte al proceso ETL. Keetle es de acceso abierto.