Curso de

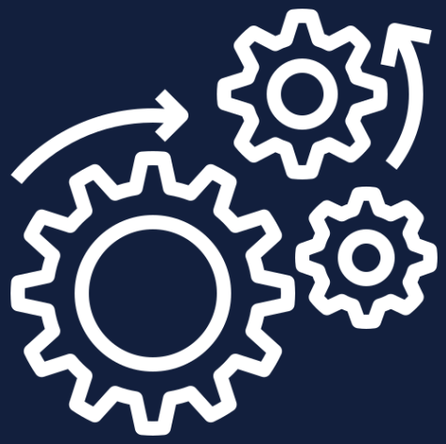
Fundamentos de ETL con Python y Pentaho

Carlos Alarcón

¿Qué es un ETL en ingeniería de datos?

ETL

Extraer
(Extract)



Transformar
(Transform)

Cargar
(Load)

Extraer datos de diferentes fuentes, transformarlos para que cumplan con los requisitos de calidad y formato esperados y, finalmente, cargarlos en un sistema de almacenamiento centralizado.

¿Para qué sirve?

⚙️ ETL

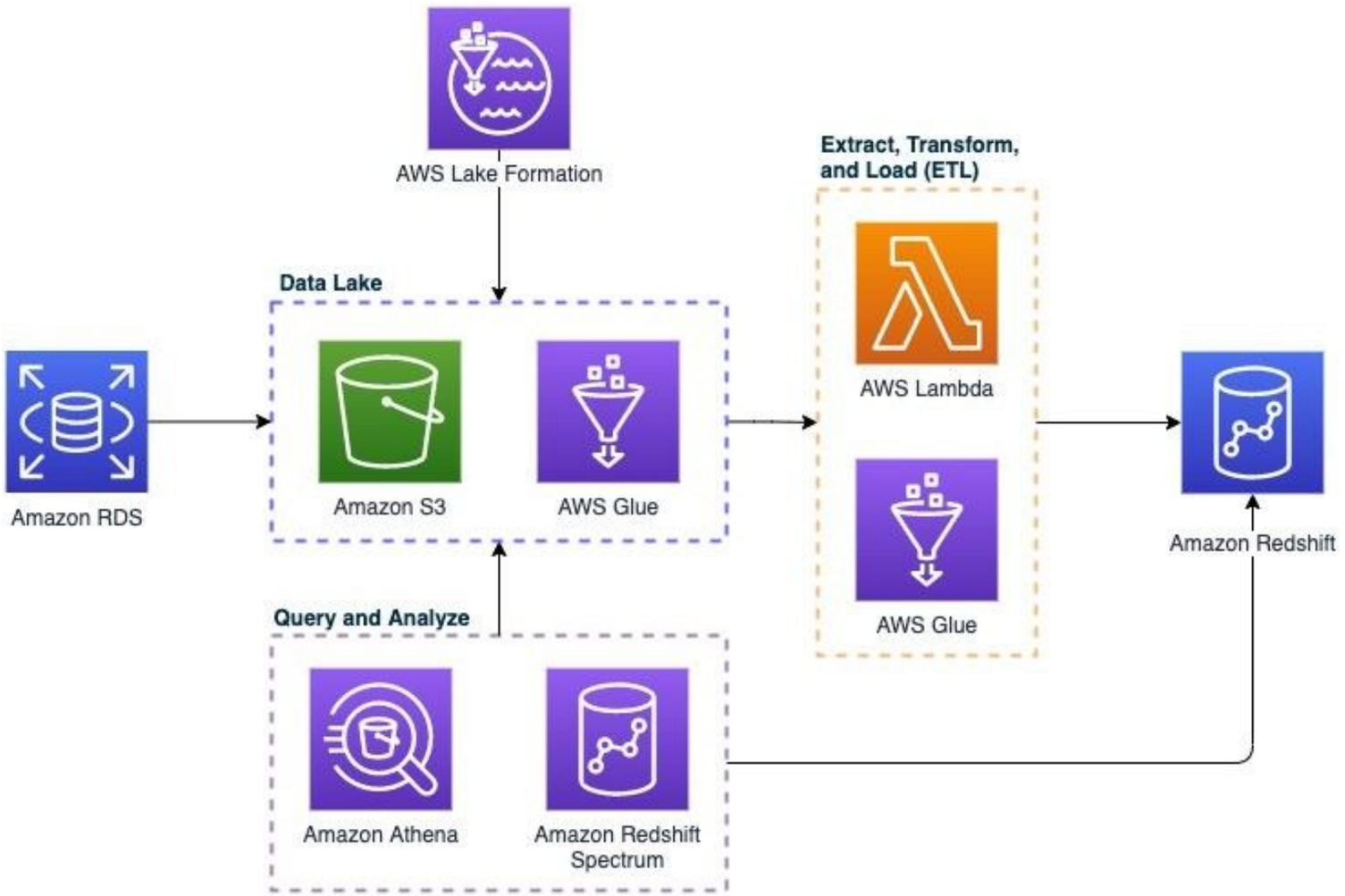
- Vista **unificada** y **coherente** de los datos.
- Tomar decisiones más **precisas**.
- Asegura la **calidad de los datos** al limpiar y normalizar.

⚙️ ETL

- Detección de **anomalías**.
- **Fuentes confiables** para machine learning y data science.

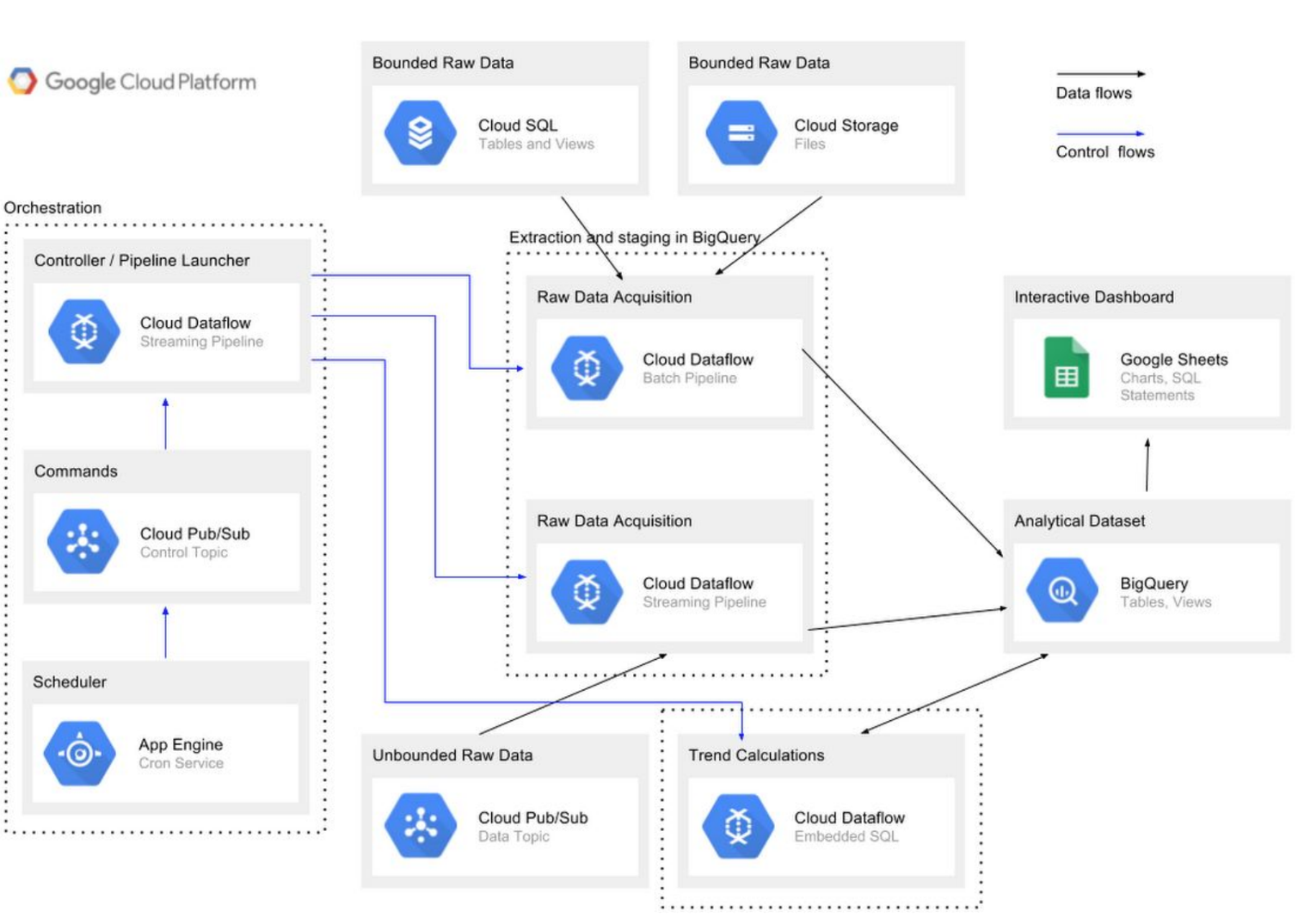
¿Cómo se usa actualmente?

ETL - AWS



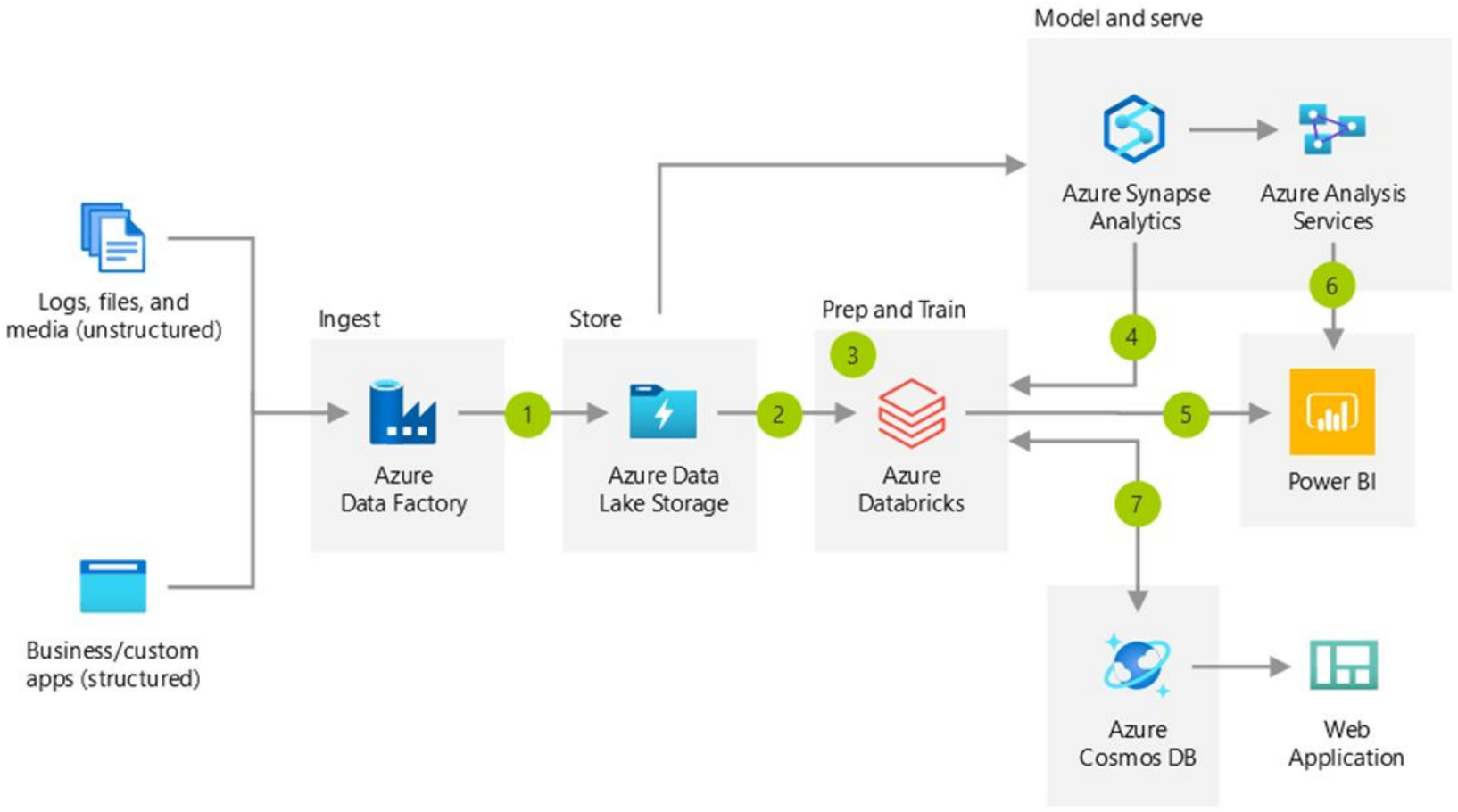
Fuente: Amazon Web Services

ETL - GCP



Fuente: Google Cloud Platorm

ETL - Azure



Fuente: Microsoft Azure

ETL - Python

1. Recolección de datos

RIAA Google Spotify kaggle

2. Validación de datos

Diseñar las tablas para guardar datos en CSVs, importar CSVs en una base de datos de SQL. Detectar conflictos en la base de datos.

python™ pandas

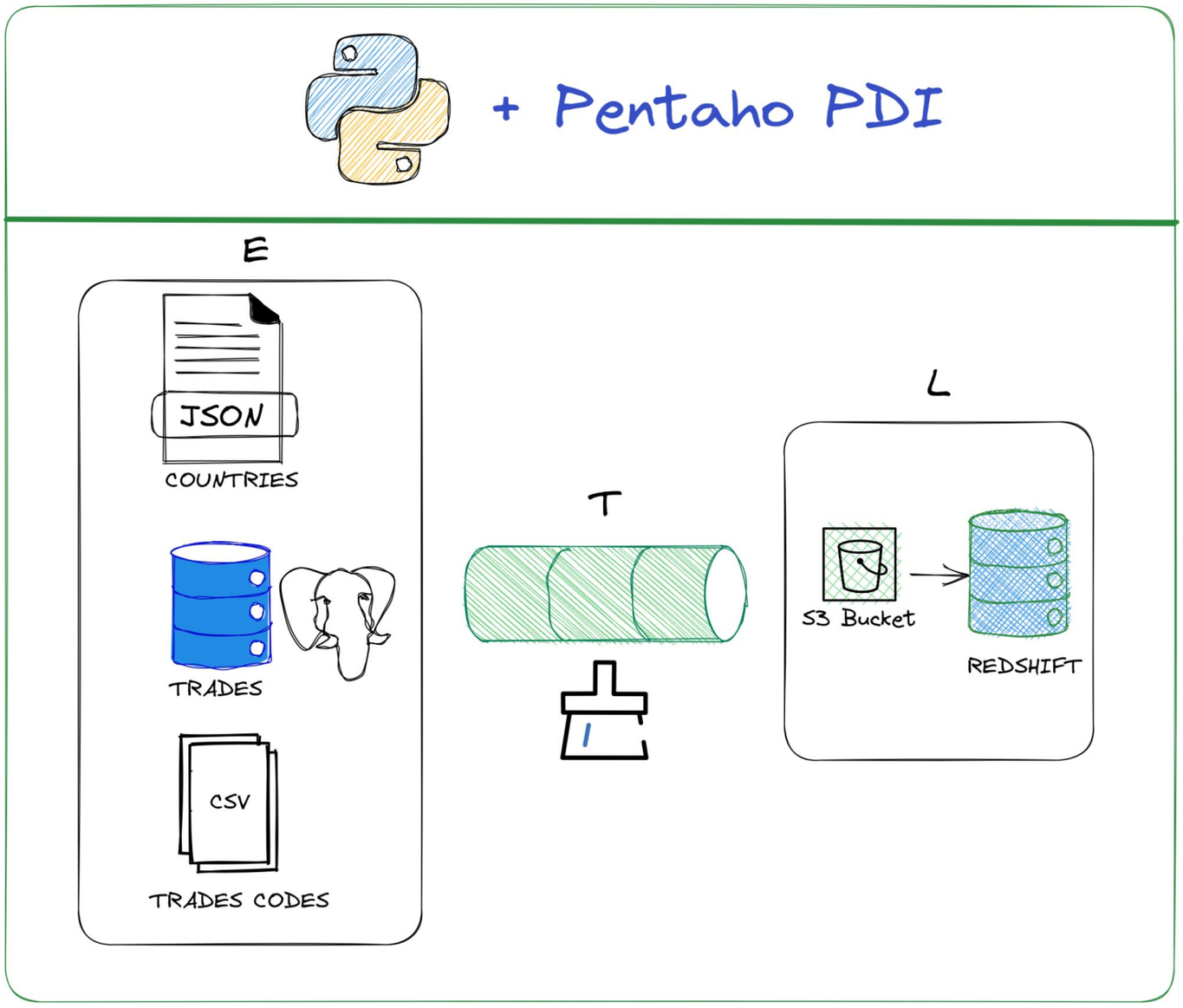
3. Creación de base de datos

Dar un formato común a los datos (estandarización / consistencia). Limpiar duplicados. Llenar datos perdidos vs. borrar datos incompletos.

PostgreSQL

Proyecto

Proyecto


+ Pentaho PDI
E
JSON
COUNTRIES
TRADES
CSV
TRADES CODES
T
L
S3 Bucket
REDSHIFT

Conceptos base de ETL

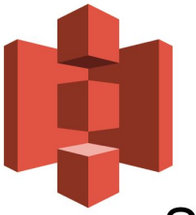
Source

Target

Cloud Datastore

PostgreSQL

amazon REDSHIFT



Amazon S3

Google BigQuery

++

Staging


SQL
RDBMS
NoSQL
DBMS
SaaS
Platforms
Archivos
XML
Transformación
Extracción
Carga
Análisis
Área de
staging
Data
Warehouse

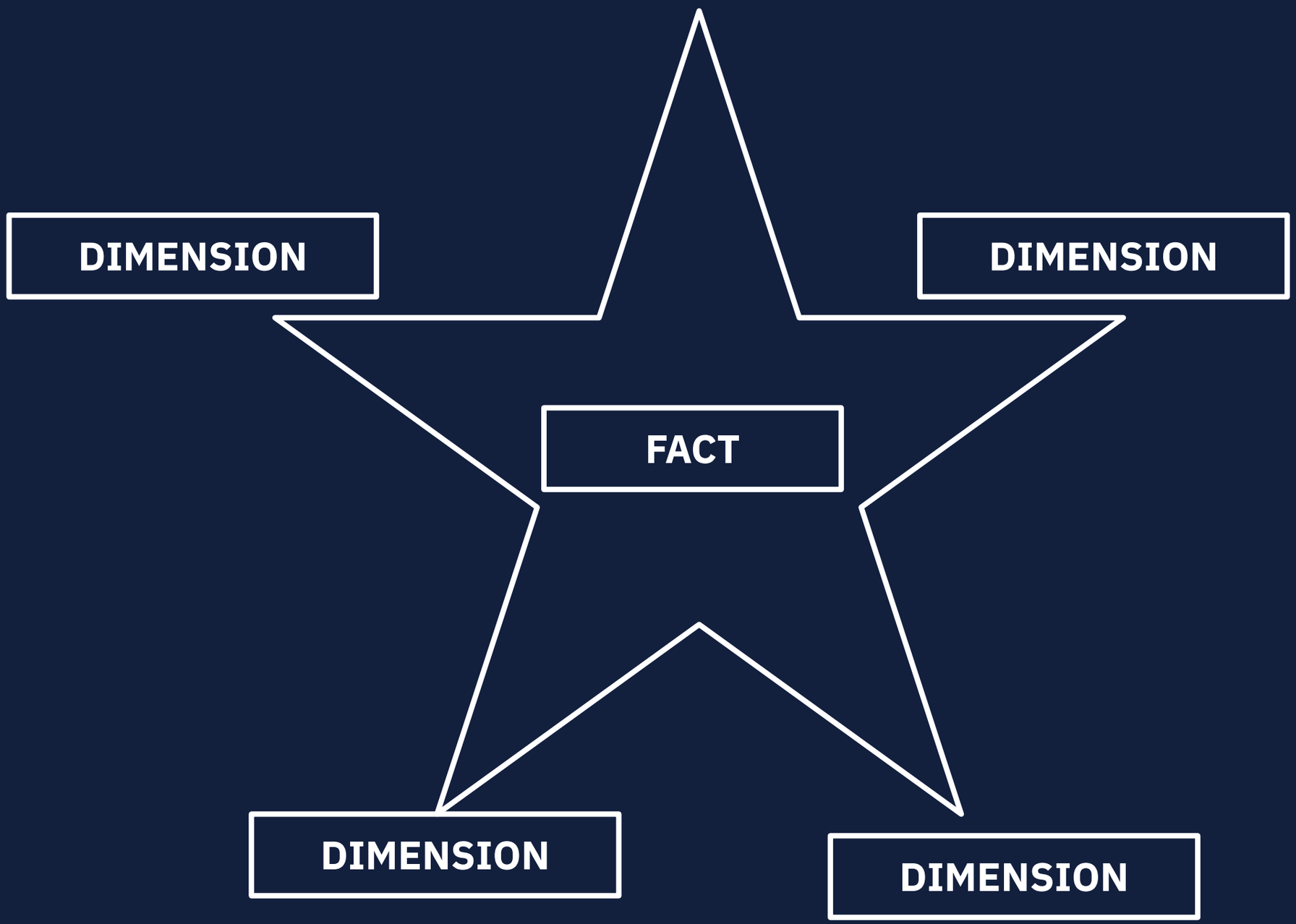
Data Warehouse

Google Big Query

snowflake

amazon REDSHIFT

Star Model



Data Lake


DATA LAKE
0101010001110010011101011011010111000
0111001101110101011000110110101101111001100
0101010001110010011101011011010111000
0111001101110101011000110110101101111001100

Data Lakehouse

Data Warehouse

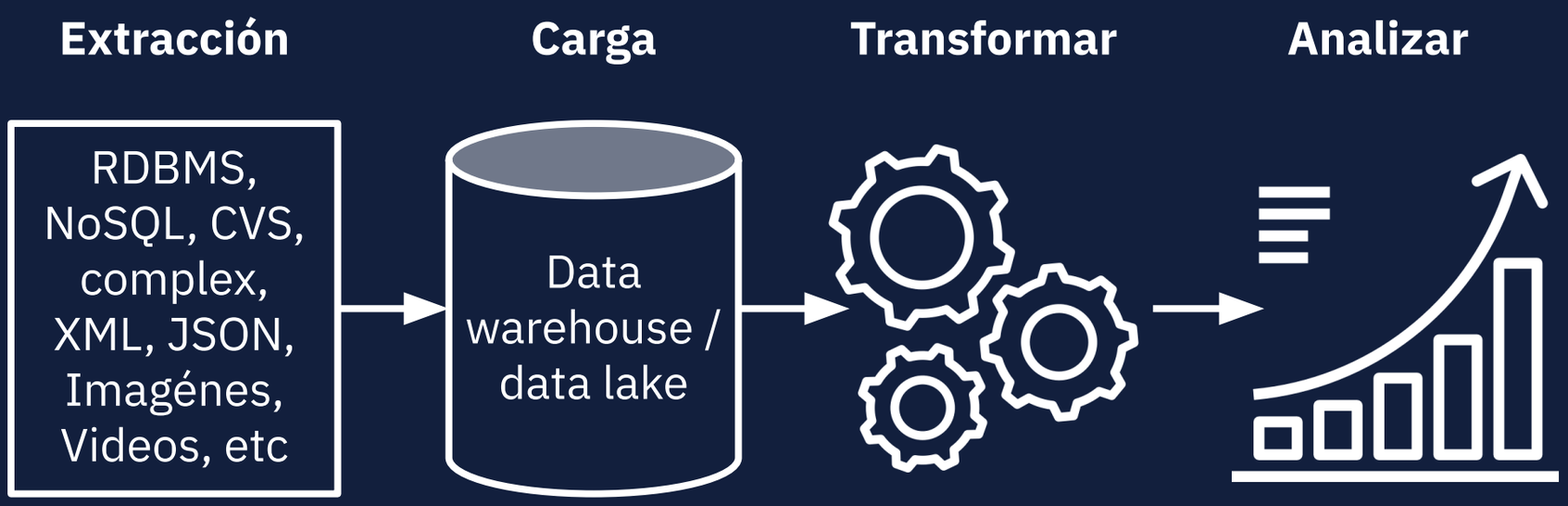
+

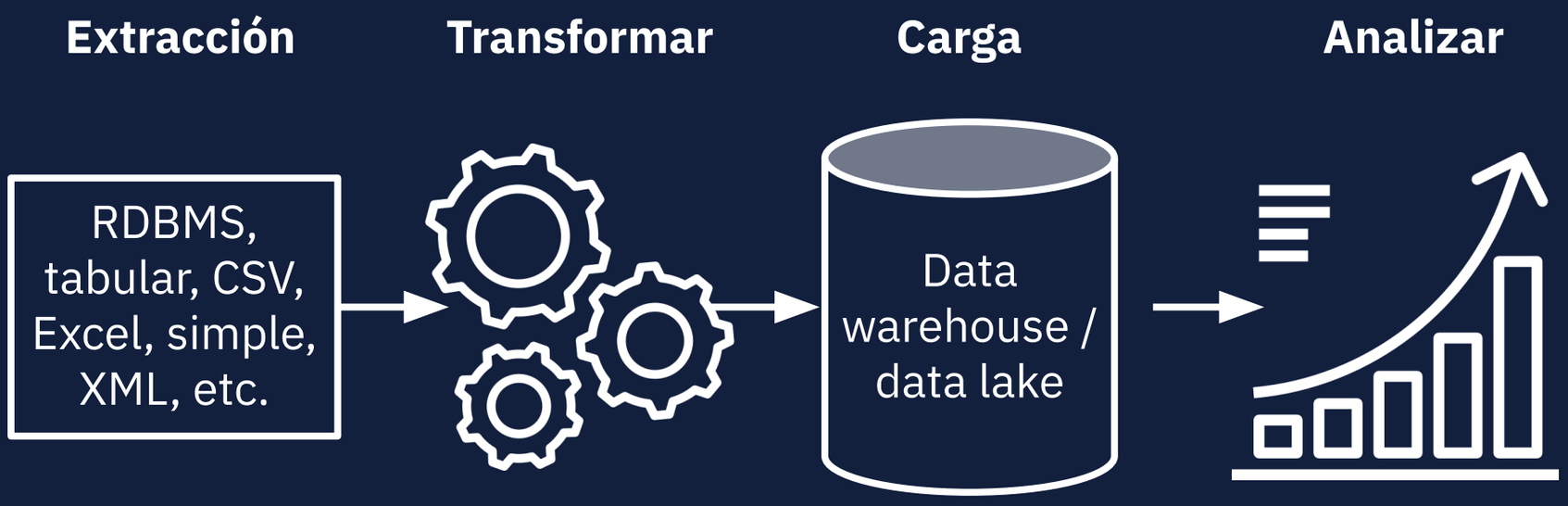
Data Lake

=

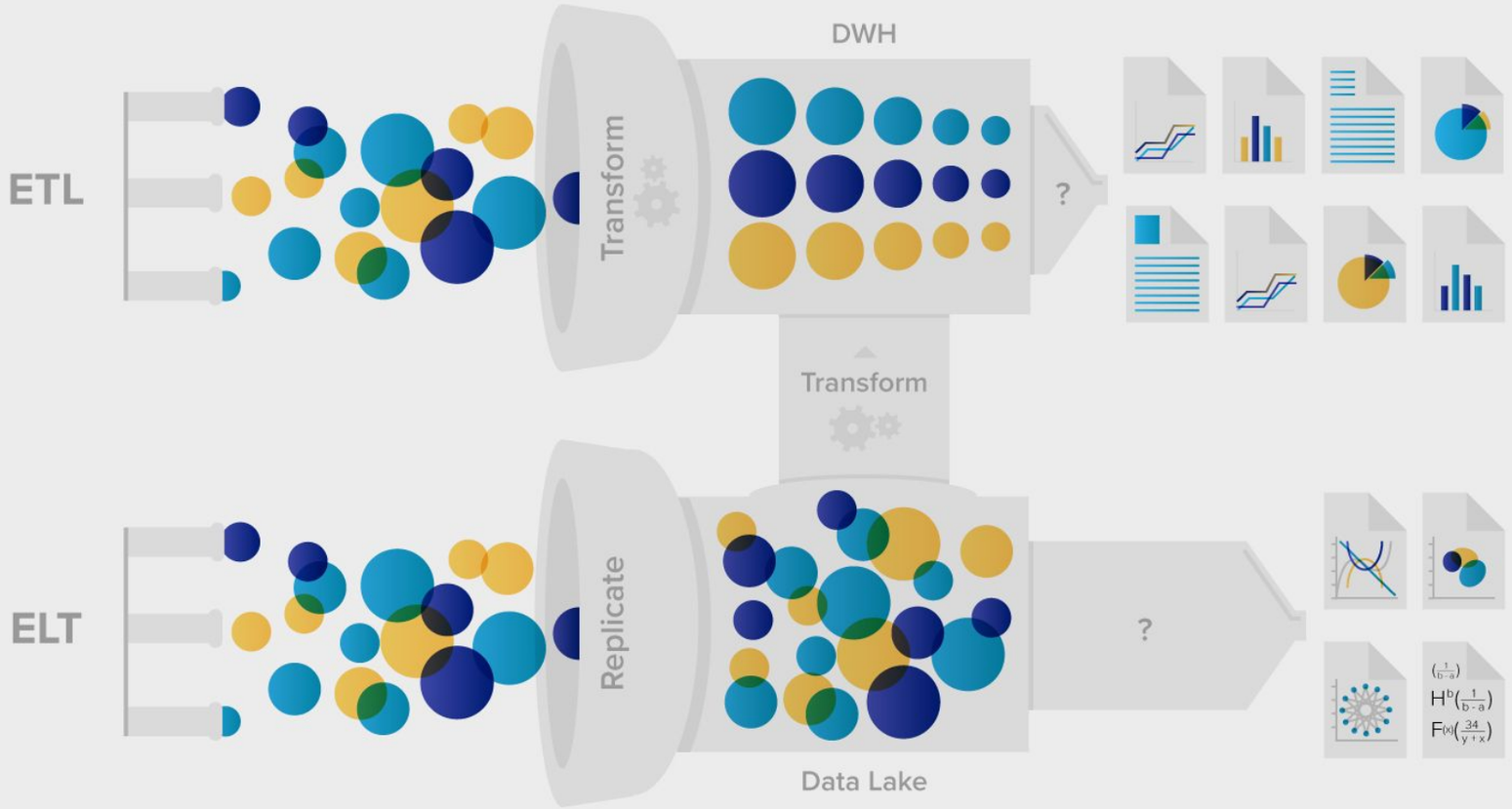
Data Lakehouse

ELT


Extracción
Carga
Transformar
Analizar
RDBMS, NoSQL, CVS, complex, XML, JSON, Imagénes, Videos, etc
Data warehouse / data lake


Extracción
Transformar
Carga
Analizar
RDBMS, tabular, CSV, Excel, simple, XML, etc.
Data warehouse / data lake

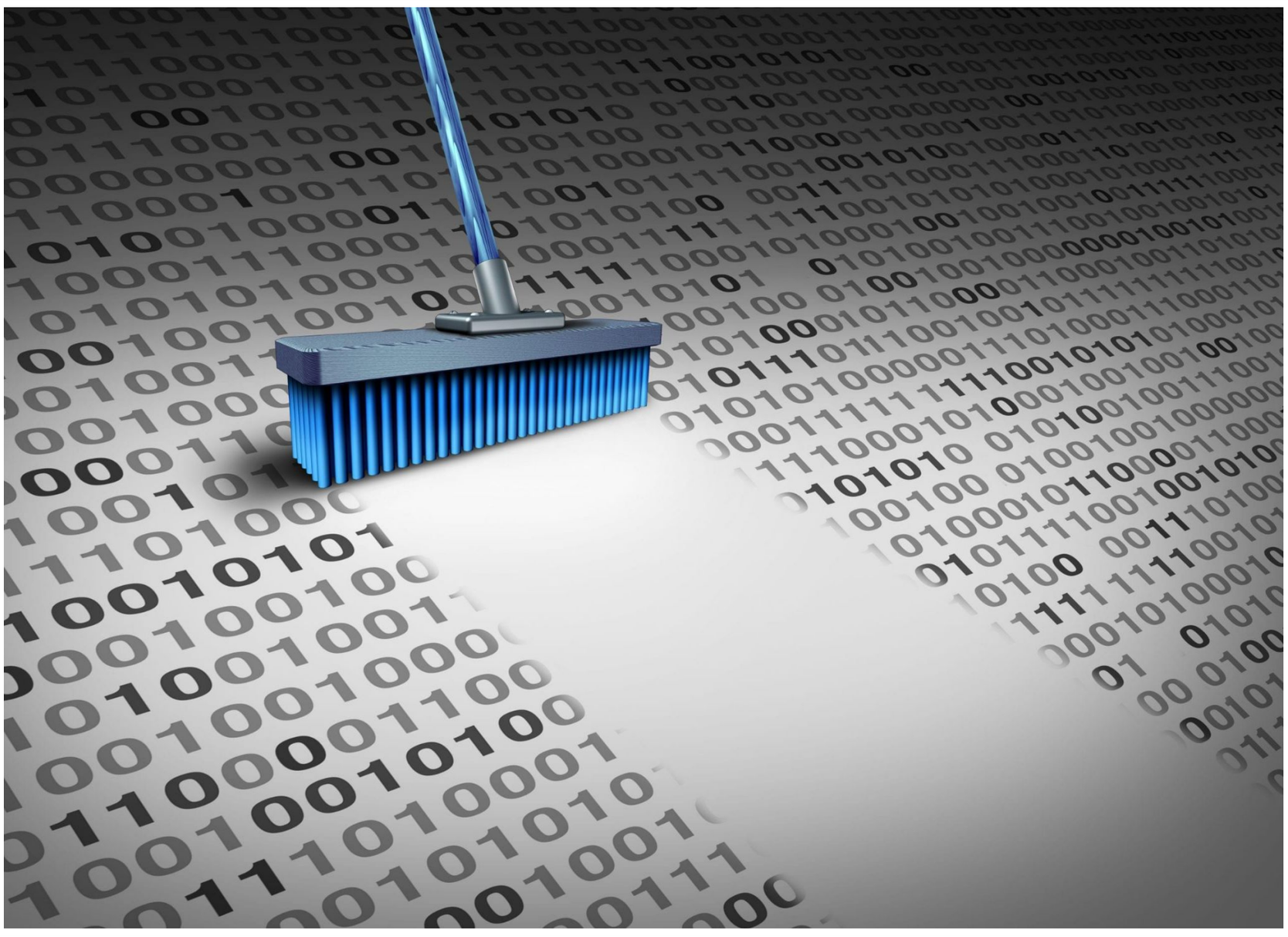
ELT



https://cdn.buttercms.com/qZu81OW0TwGL9mPvDHYs

Consideraciones de ETL

Calidad de los datos



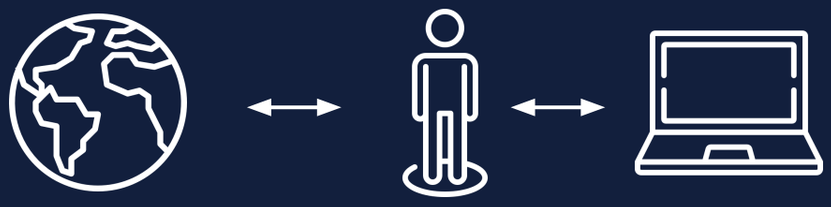
Sources & Target

Fuentes

Objetivos

Batch o Streaming

Procesamiento en lote - 20 min



Información Empleado Sistema

Procesamiento Menos de 1 seg

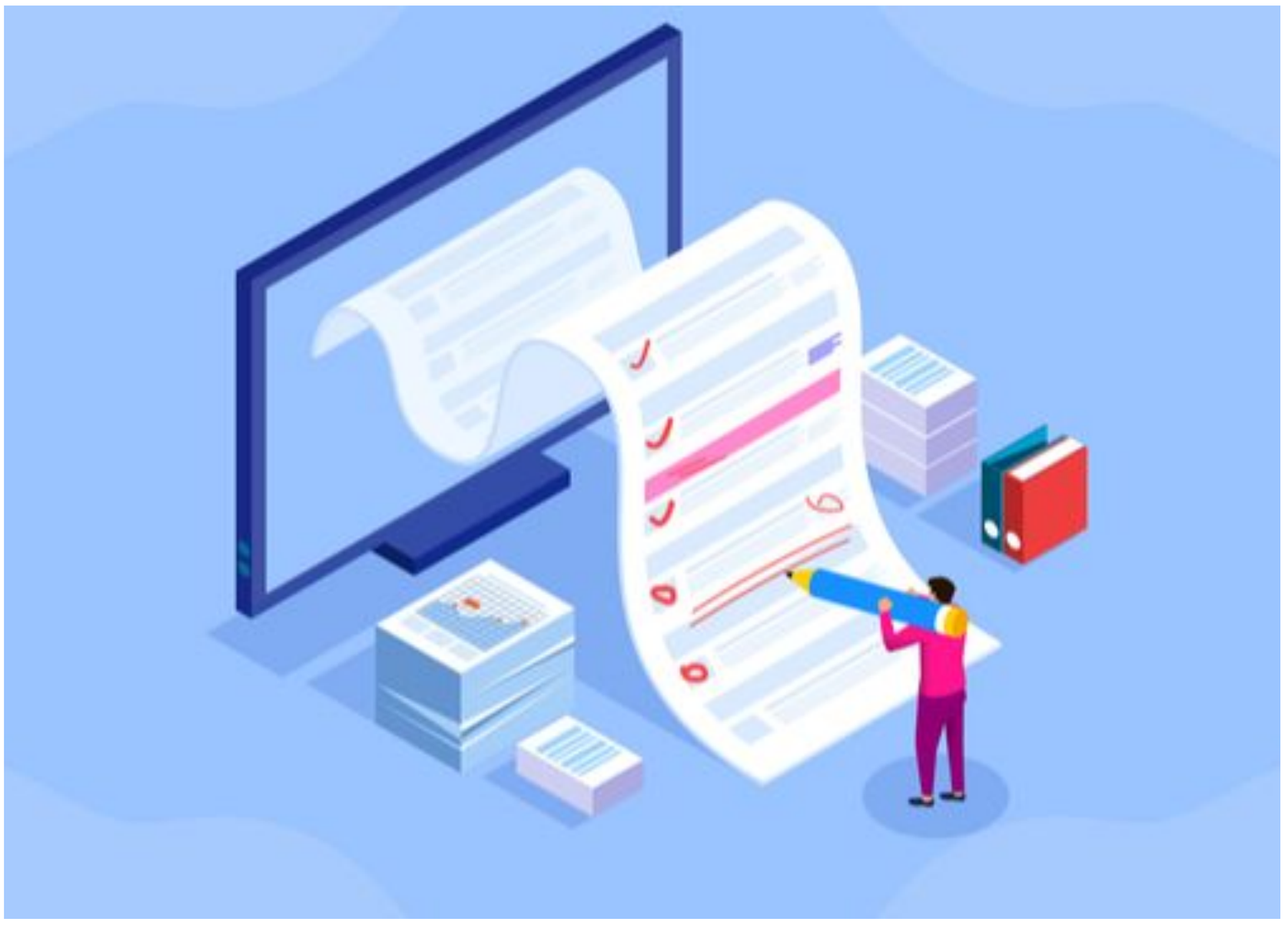
Información Sistema

Incremental o Total

Paso 1: **Carga total**

Paso 2: **Carga incremental**

Documentar



++

Servicios y herramientas

Enterprise

IBM DataStage®

ORACLE DATA INTEGRATOR

Informatica™

sas Data Management

++

Cloud

integrate.io

Google Dataflow

++

Open source

++

Custom

Sources

 Formato

Es importante asegurarse de que los datos estén en un formato compatible con la herramienta de ETL que se está utilizando.

✅ Calidad de los datos

Es necesario verificar la integridad y la precisión de los datos antes de cargarlos.

🔄 Frecuencia de actualización

Debes determinar la frecuencia con la que los datos deben ser extraídos y actualizados.

➡️ Accesibilidad

Debes tener acceso a las fuentes de datos para poder extraerlos y cargarlos en el sistema.

Seguridad

Debes asegurarte de que los datos estén protegidos y de que solo las personas autorizadas tengan acceso a ellos.

⚡ Eficiencia

Debes buscar la manera más eficiente de extraer y cargar los datos, para evitar retrasos y errores.

Escalabilidad

Debes tener en cuenta si la solución de ETL es escalable y si es posible manejar una cantidad creciente de datos en el futuro.

Extracción de datos con Python 

Transform

Estructura final

¿Con qué estructura requiero los datos compatibles con el target?

♾️ Relaciones

¿Cómo debo relacionar los datos de distintas fuentes?

Normalización

Normalizar los datos para lograr consistencia a lo largo del flujo de datos.

👥 Duplicados

¿Con qué estructura requiero los datos compatibles con el target?

🔎 Datos faltantes

¿Qué ocurre con los datos faltantes en mi data?

🔎 Agregaciones

Debo agrupar los datos por alguna característica y buscar agregaciones como suma, promedio, máximo y demás.

Transformación de datos con Python 

Load

 Formatos de datos aceptables

Garantizar que solo se reciban datos relevantes y coherentes en la estructura necesaria por el warehouse o target.

✅ Permisos

Se deben tener todos los **permisos necesarios** para escribir sobre el destino y modificar archivos o datos de ser necesario.

🔎 Auditar

Comparar los datos recibidos con los datos de referencia permite **detectar errores, problemas de calidad y duplicados** o demás errores en el proceso.

⚡ Eficiencia

Debes buscar la **manera más eficiente** de extraer y cargar los datos para evitar retrasos y errores.

⚠️ Control de errores

Es importante establecer un plan de acción en caso de presentarse un error: **¿revertir todo el proceso o solo corregir los fallos y continuar con el proceso?**

Carga de datos con Python

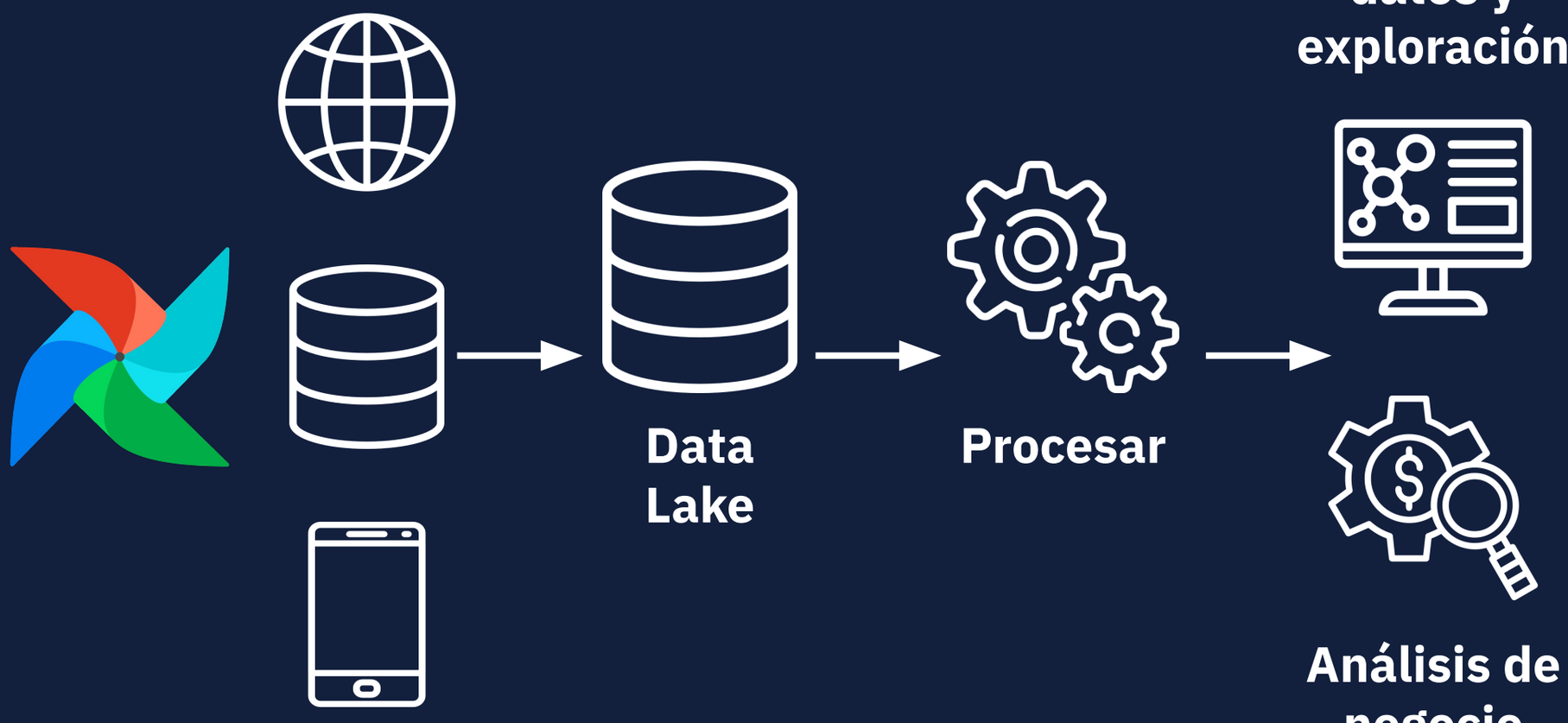
Extracción de datos con Pentaho

Transformación de datos con Pentaho

Carga de datos con Pentaho

Siguientes pasos

Orquestar


Ciencia de datos y exploración
Data Lake
Procesar
Análisis de negocio (Business intelligence)