



Módulo Profesional: Big Data Aplicado

DBT
Data Build Tool

ÍNDICE

1. Introducción

- La importancia de las herramientas de Transformación de flujo de datos
- ¿Qué es una herramienta de Transformación de flujo de datos?

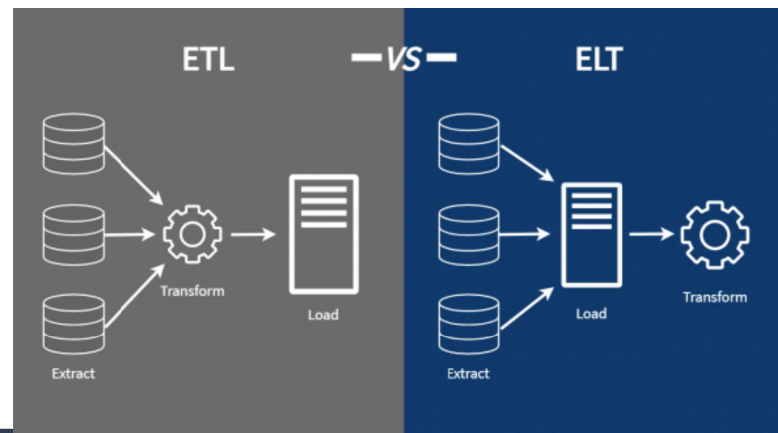
2. DBT

- ¿Qué es DBT?
- Ventajas de utilizar DBT
- DBT Products
- DBT Core

3. Ejercicios

La importancia de las herramientas de Transformación

- La transformación de datos es el paso en el que los datos pasan de ser brutos y no estructurados a ser refinados, organizados y procesables.
- El proceso de transformación es complejo y a menudo requiere un importante esfuerzo de ingeniería, scripts personalizados y supervisión manual.
- La solución: Herramientas de Transformación de flujo de datos.
- Las herramientas de Transformación flujo de datos se utilizan normalmente en las cadenas de datos de extracción, transformación y carga (ETL).



¿Qué es una herramienta de Transformación?

- Una herramienta de Transformación está diseñada para automatizar y gestionar el proceso de transformación de datos brutos en información valiosa y lista para la empresa.
- Algunas ventajas de utilizar Herramientas de Transformación:

**Transformación
automática**

**Garantizar la
coherencia de
los datos**

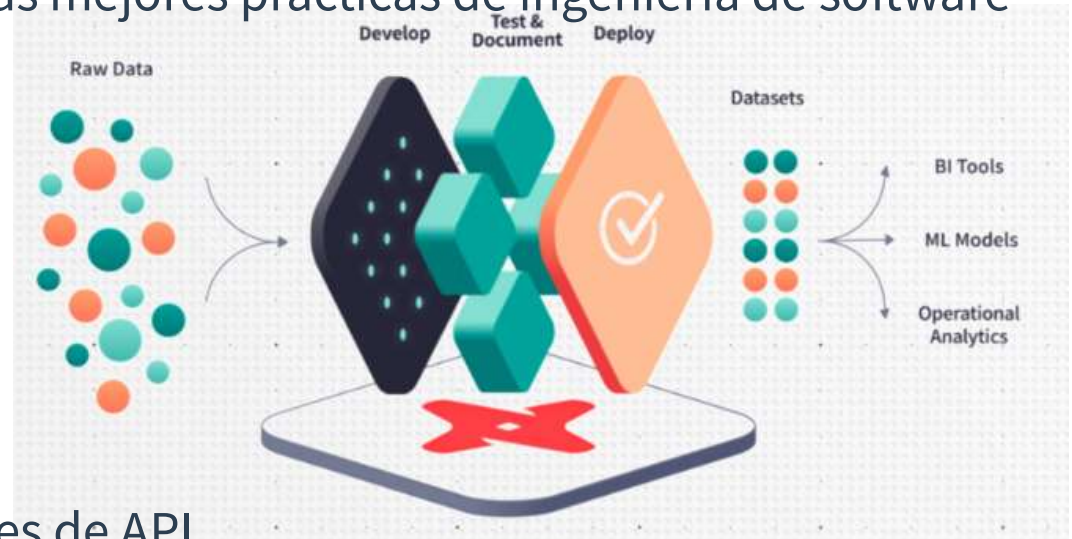
**Control de
versiones y
colaboración**

**Mejorar
Escalabilidad**

**Supervisar y
testear datos**

¿QUÉ ES DBT?

- Transformación de datos “T” ETL
- Despliegue de código analítico siguiendo las mejores prácticas de ingeniería de software
 - Control de versiones
 - Modularidad
 - Portabilidad
 - CI/CD
 - Documentación
- Algunos casos de uso
 - Creación de vistas para BI
 - Preparación de datos para puntos finales de API
 - Preparación de datos para Machine Learning



Ventajas de utilizar DBT

Escribe lógica de negocio que devuelva el conjunto de datos que necesitas y dbt se encarga de la materialización

Construir modelos de datos fiables y escalables

Evitar la consulta de datos brutos

Reducir los tiempos de ejecución de las consultas

Colaborar en modelos de datos, versionarlos y probar y documentar sus consultas

Control y visibilidad

Conexión con plataformas de datos compatibles mediante adaptadores dbt

DBT Products

dbt Cloud

- La forma más rápida, fiable y escalable de implantar dbt.
- Servicio único y totalmente gestionado a través de una interfaz de usuario (UI) basada en web.

dbt Core

- Una herramienta de línea de comandos de código abierto
- Adecuada para usuarios que prefieren configurar dbt manualmente y mantenerlo localmente.

DBT Cloud

The screenshot displays the DBT Cloud interface, which is divided into several sections:

- Jobs:** A table listing various jobs and their execution status.
- Production:** A section for managing production environments.
- File Explorer:** A sidebar showing the project's file structure.
- SQL Editor:** A central area for writing and editing SQL queries.
- Lineage Graph:** A visual representation of the data flow between models.

Job	Last Run	Next Run
Daily Job	Today At 6:01 PM UTC Success	Today At 6:00 PM UTC in 3 hour
CI Testing Job	Today At 12:22 PM UTC Error	Not Scheduled
Incremental Jobs	05/11/2022 UTC Success	Not Scheduled
Snapshot Example	Today At 12:00 AM UTC Success	Today At 11:00 PM UTC in 8 hours
Disabled New Job	08/04/2022 UTC Success	Not Scheduled
Daily Job (dbt build)	Today At 4:02 PM UTC Success	Today At 8:00 PM UTC in 3 hour
Run fct_orders with dependencies	01/30/2022 UTC Success	Not Scheduled
Source Freshness	04/13/2022 UTC Success	Not Scheduled
Update Customer Table Daily	Today At 12:03 PM UTC Success	Tomorrow At 12:00 PM UTC in 17 hours
Source Freshness Job	07/10/2022 UTC Error	Not Scheduled

```
1 {{
2   config(
3     materialized = 'view',
4   )
5 }}
6 WITH src_hosts AS (
7   SELECT
8     *
9   FROM
10    {{ ref('src_hosts') }}
11 )
12 SELECT
13   host_id,
14   NVL(
15     host_name,
16     'Anonymous'
17   ) AS host_name,
18   is_superuser,
19   created_at,
20   updated_at
21 FROM
22   src_hosts
23
24 Limit 100
25
26
27 -- dim_hosts_clean
28 -- dim_listings_clean
29
```

The lineage graph shows the following flow: `airbnb.hosts` → `src_hosts` → `dim_hosts_clean` → `dim_listings_with_hosts`.

DBT Core

- Herramienta de línea de comandos de código abierto
- Está diseñada para ayudar a analistas e ingenieros a transformar y probar sus datos como código
- El código puede ser controlado por versiones, revisado por pares y probado automáticamente.

Conceptos básicos:

Modelos	Tests	Documentación	Control de versiones
<ul style="list-style-type: none">○ SQL files que definen las transformaciones.○ Pueden ser simples sentencias select o consultas analíticas complejas.	<ul style="list-style-type: none">○ Garantizar la calidad del dato.○ Test pueden validar la unicidad, restricciones no nulas y la integridad referencial	<ul style="list-style-type: none">○ Genera un sitio web documentación.○ Proporciona un contexto y comprensión de las transformaciones datos.	<ul style="list-style-type: none">○ dbt integra sistemas de control como git.○ Desarrollo colaborativo y seguimiento de cambios

Ejercicio práctico: dbt Core Instalación y Project Set Up

CLI Instalación

- Requisitos: Python
- Pasos:
 - Instalar dbt Core
 - Instalar Database adapter
 - Crear y set up Project

Ejercicio práctico: Jaffle Shop

Jaffle Shop es una tienda de ventas de sandwich.

- **Crear vistas y tablas para una tienda con datos sobre:**
 - **Clientes (Customers)**
 - **Orden (Orders)**
 - **Pagos (Payments)**
- **Elementos esenciales de dbt:**
 - **Modelos**
 - **Test**
 - **Documentación**



