

Curso de  
**Fundamentos  
de Ingeniería  
de Datos**

Ricardo Alanís  
Director of Data Science  
en Nowports



# ¿Quién soy yo?



Ing. Químico, M.Sc. Energía.



Enfocado en el uso de datos.



En compañías tech desde 2014.



Director of Data Science en Nowports.



Me encanta la inteligencia artificial.

# ¿Qué es ingeniería de datos?

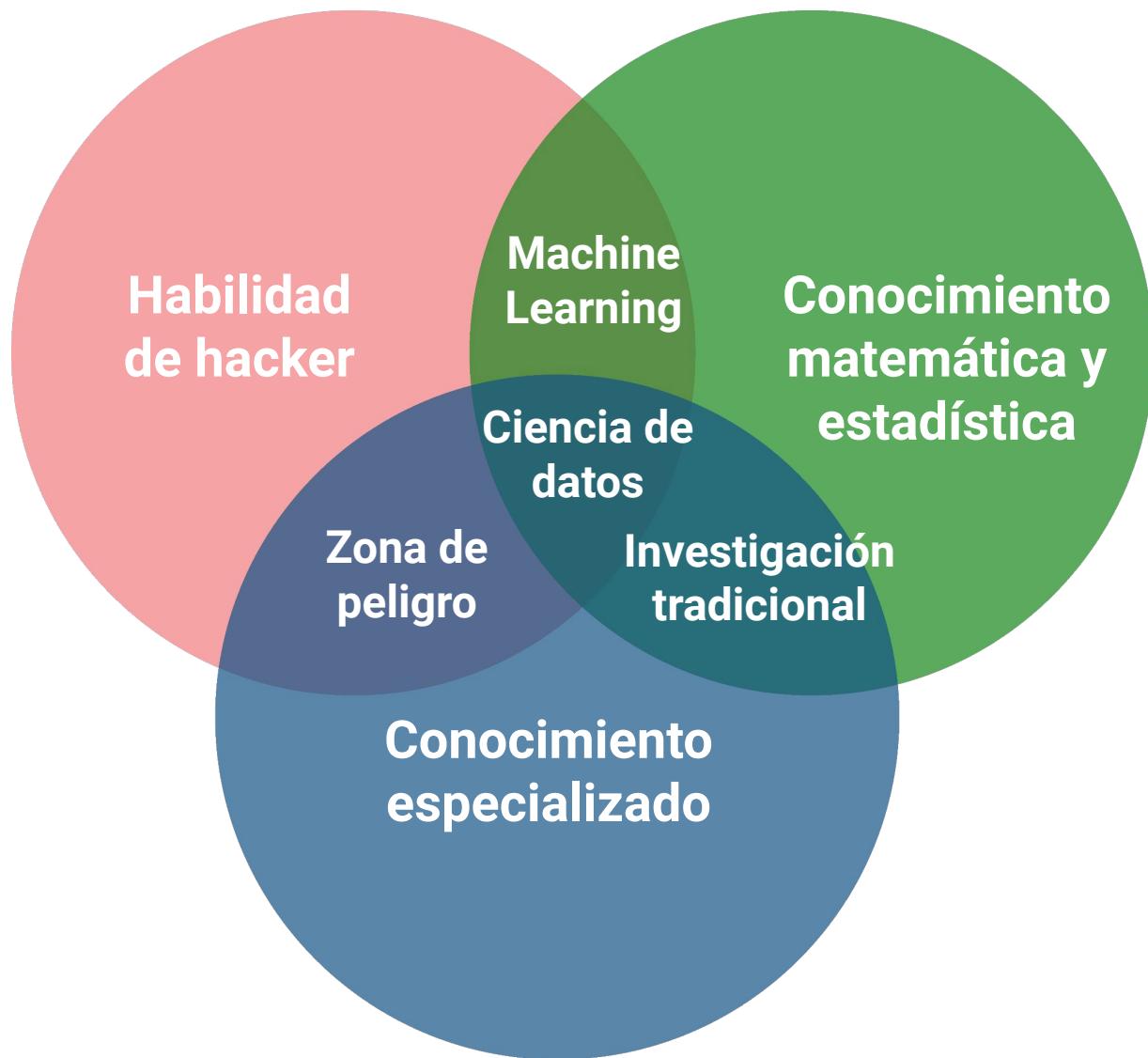
# **Érase una vez**

## Un mundo de datos

# Un unicornio



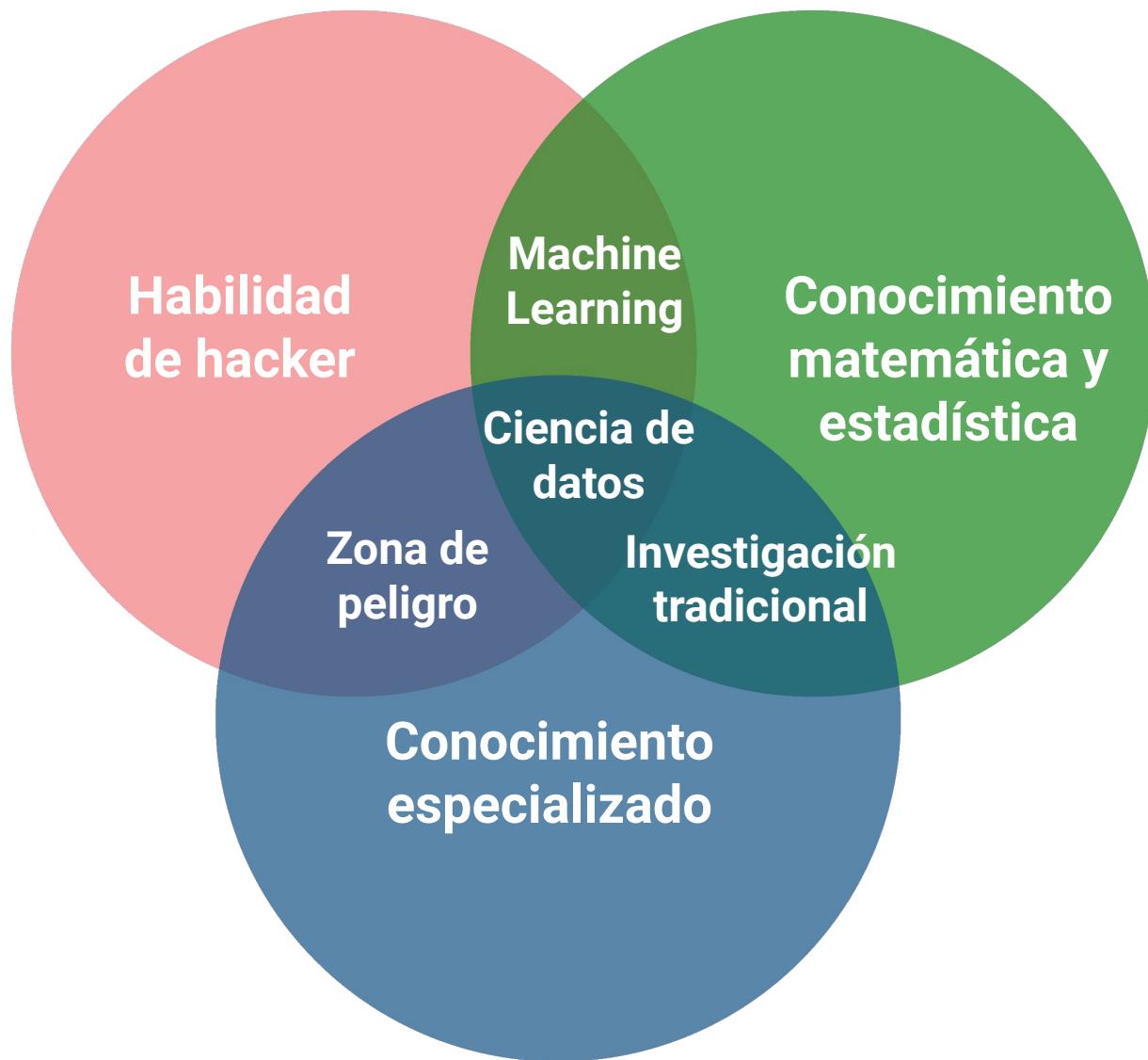
# Un unicornio



ERROR

404

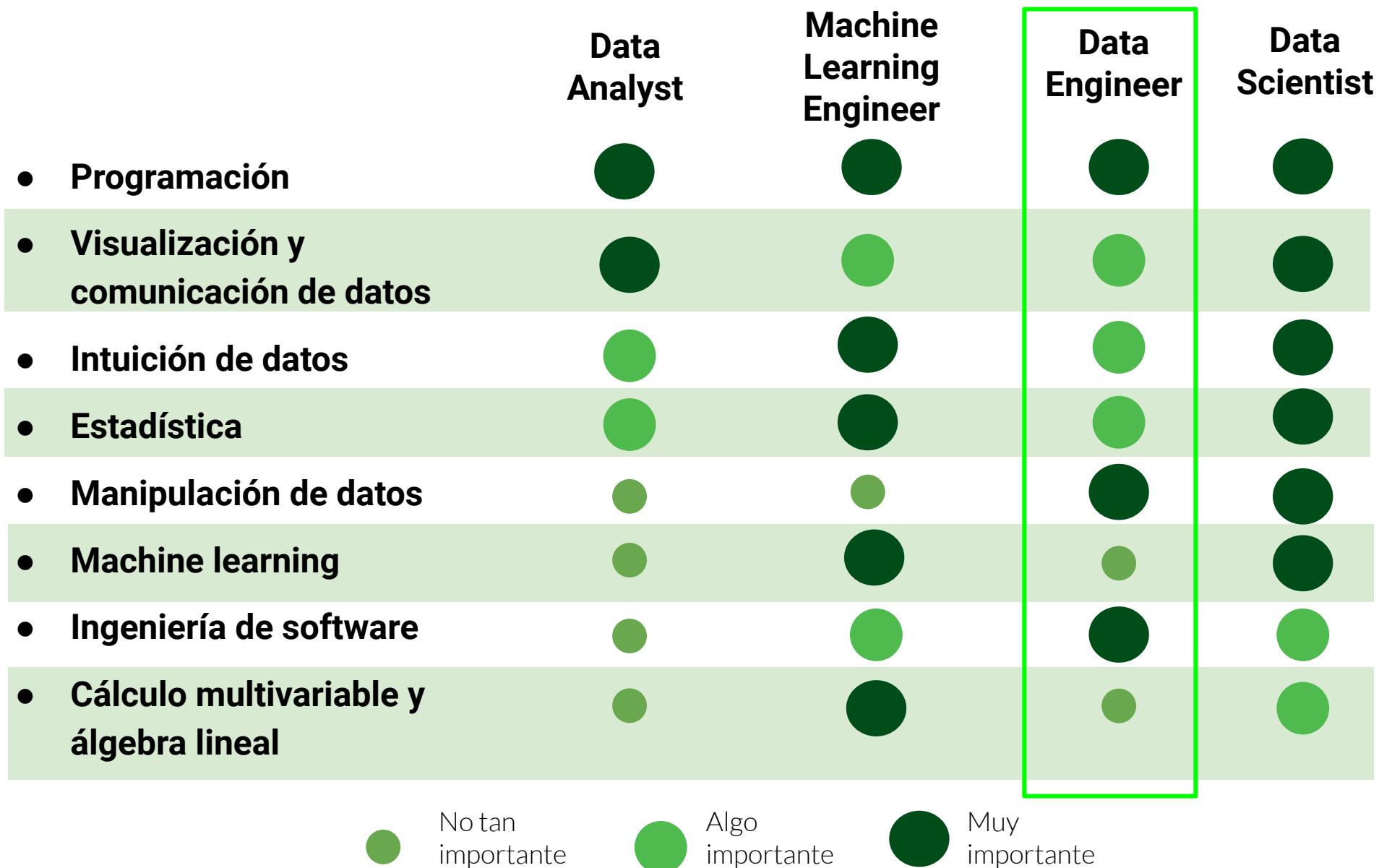
# Ciencia de datos - Habilidades



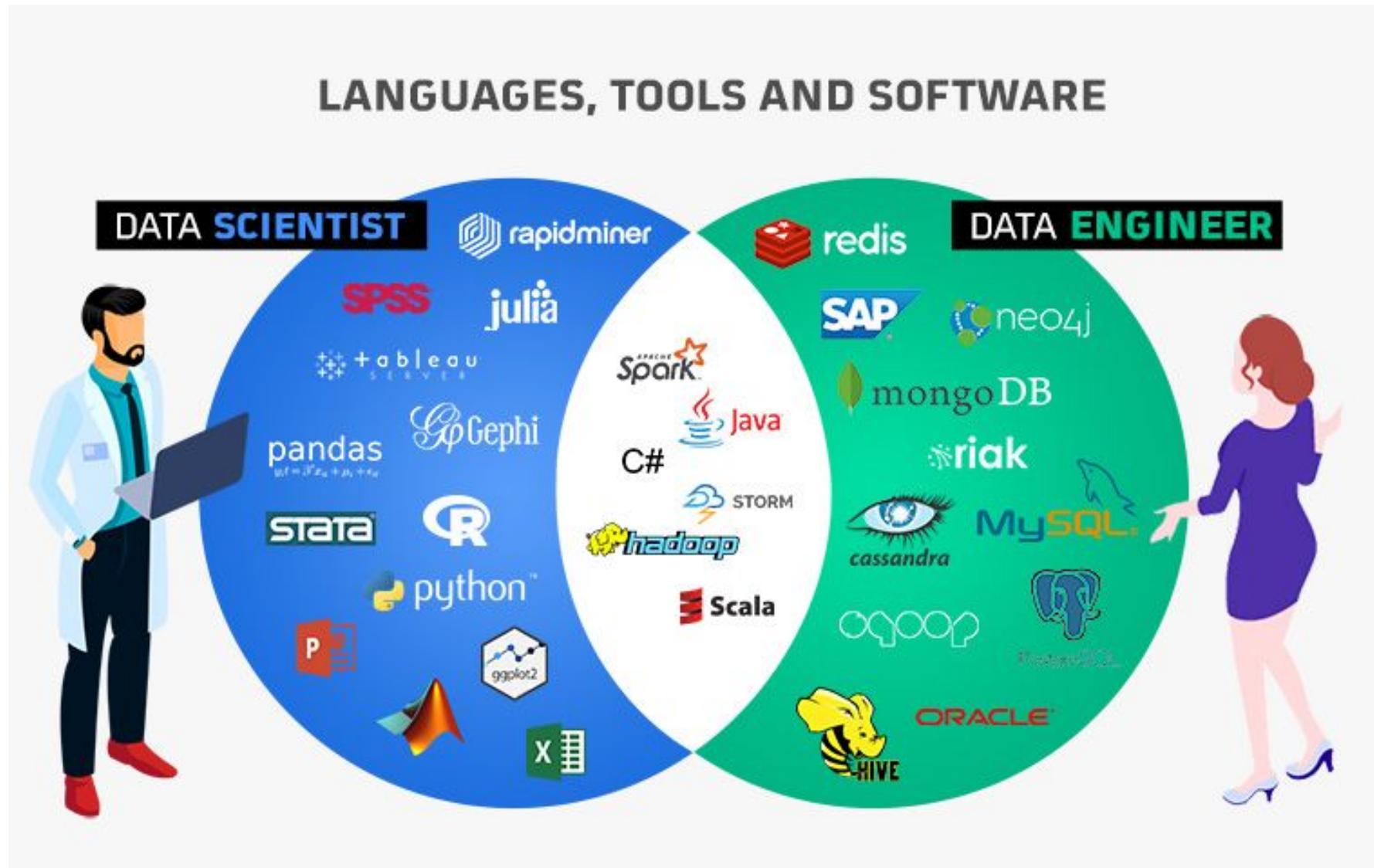
# Proceso de ciencia de datos



# Ciencia de datos - Equipo

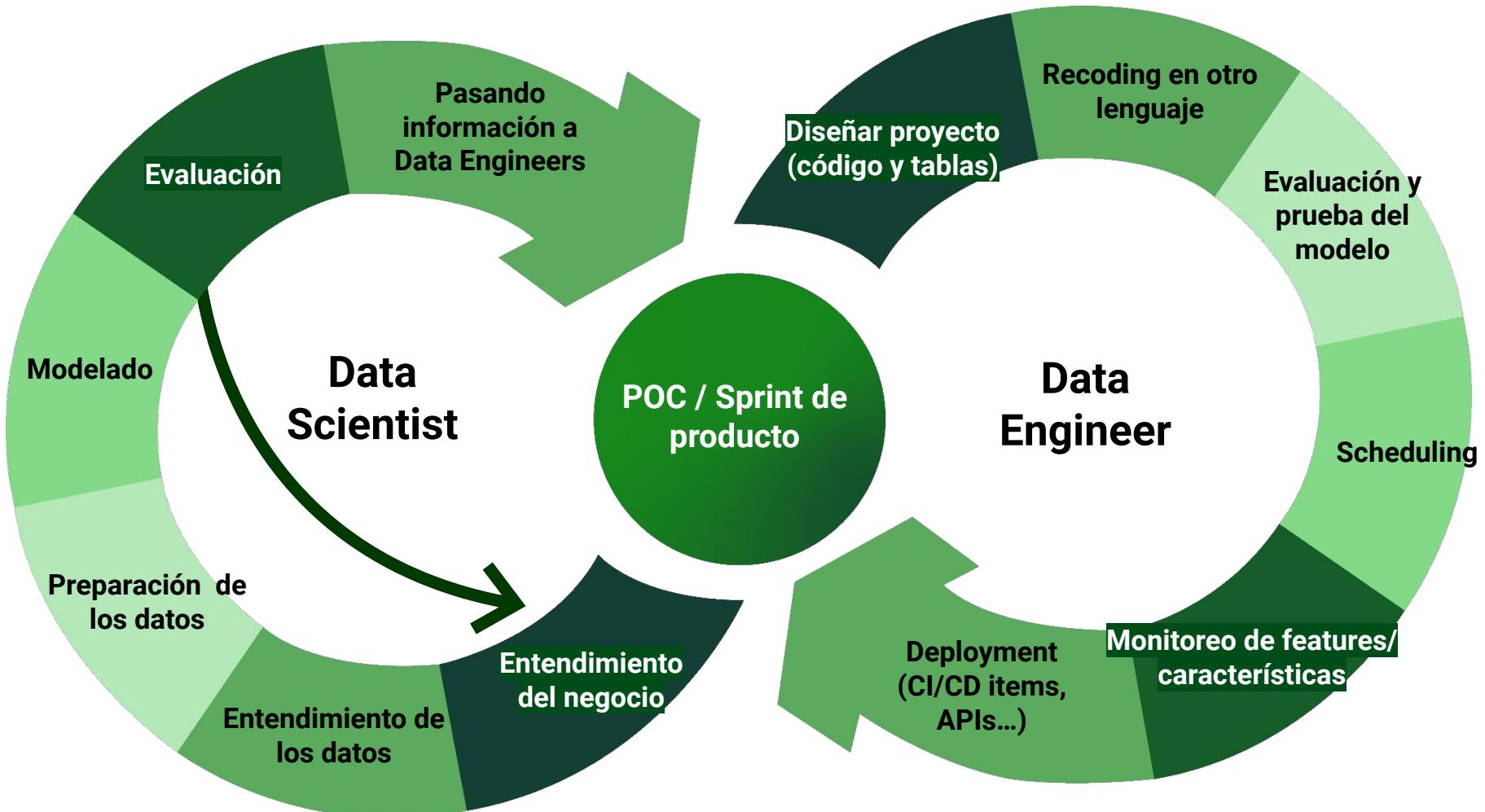


# Ingeniería de datos



# Ingeniería de datos

## Proceso



**¿Qué te parece? ¿Listo?**



# ¿Cómo convertirte en data engineer?

Introducción a la Ingeniería de Datos

# Los frentes de la ingeniería de datos

# Conocimiento de data science



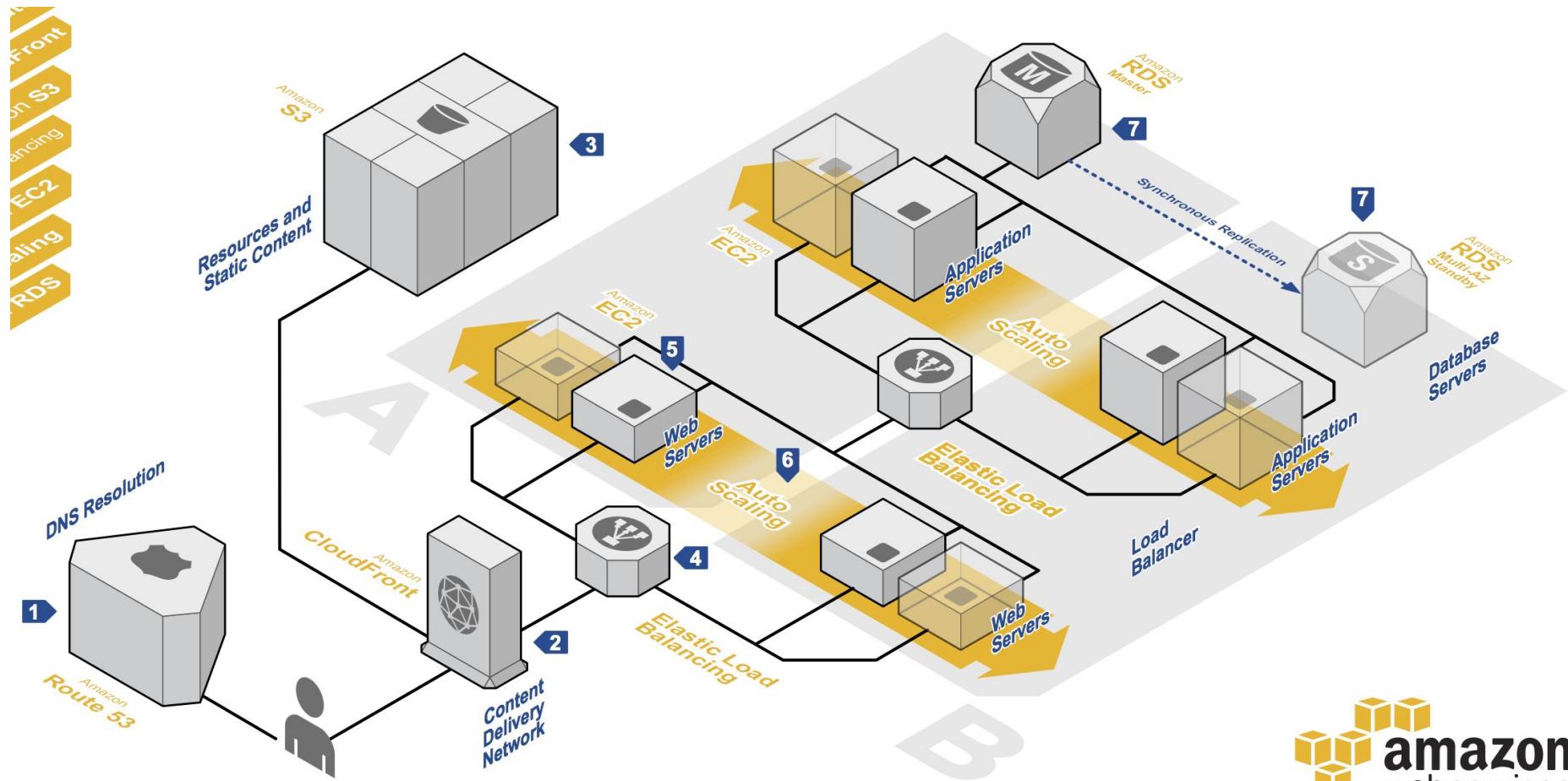
# Ingeniería de software

```
# -*- coding: utf-8 -*-
from odoo import http
from odoo.exceptions import ValidationError
from datetime import datetime
import calendar
import math
import pytz
import io, base64

class AdmissionExtensionOnlineController(http.Controller):
    @http.route('/get/type_wise_program', website=True, auth='none')
    def type_wise_program(self, **kwargs):
        if len(kwargs['types'])<=0:
            return "None"
        types = kwargs['types']
        program_list = []
        domain = []

        if types == 'local_bachelor_program_hsc':
            domain = [('course_id.is_local_bachelor_program_hsc', '=')]
        elif types == 'local_bachelor_program_a_level':
            domain = [('course_id.is_local_bachelor_program_a_level', '=')]
        elif types == 'local_bachelor_program_diploma':
            domain = [('course_id.is_local_bachelor_program_diploma', '=')]
        elif types == 'local_masters_program_bachelor':
            domain = [('course_id.is_local_masters_program_bachelor', '=')]
```

# Arquitectura e infraestructura



# **¡Reto!**

Conociendo las vacantes

**Instrucciones:**

Busca en LinkedIn oportunidades  
para data engineers.

# Dónde ejercer como data engineer

Introducción a la Ingeniería de Datos

**El campo de juego es vasto**

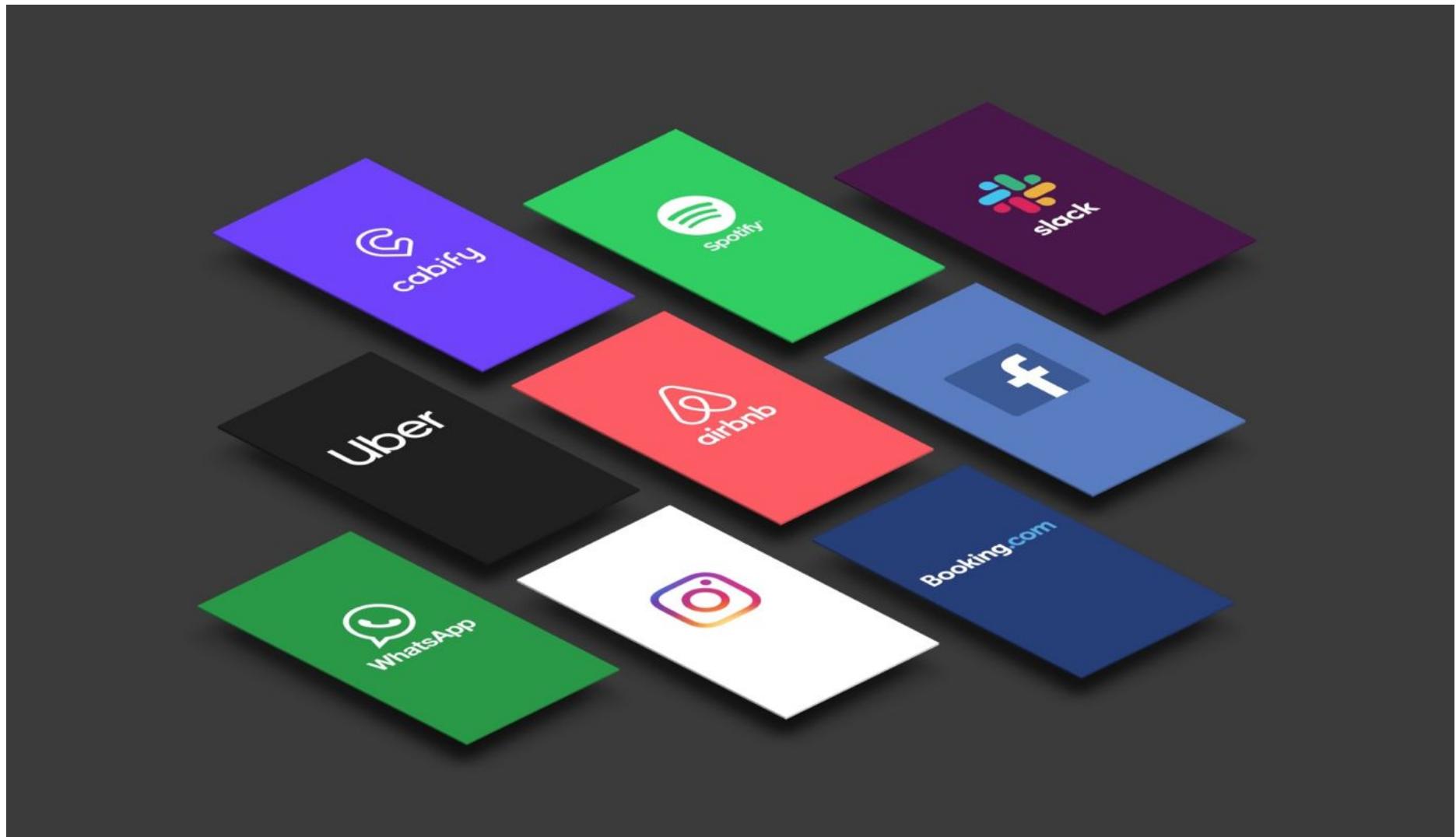
# Startups de producto



Deepnote



# Empresas de software



# Corporaciones y empresas



# ¡Reto!

Conociendo la empresa  
de tus sueños

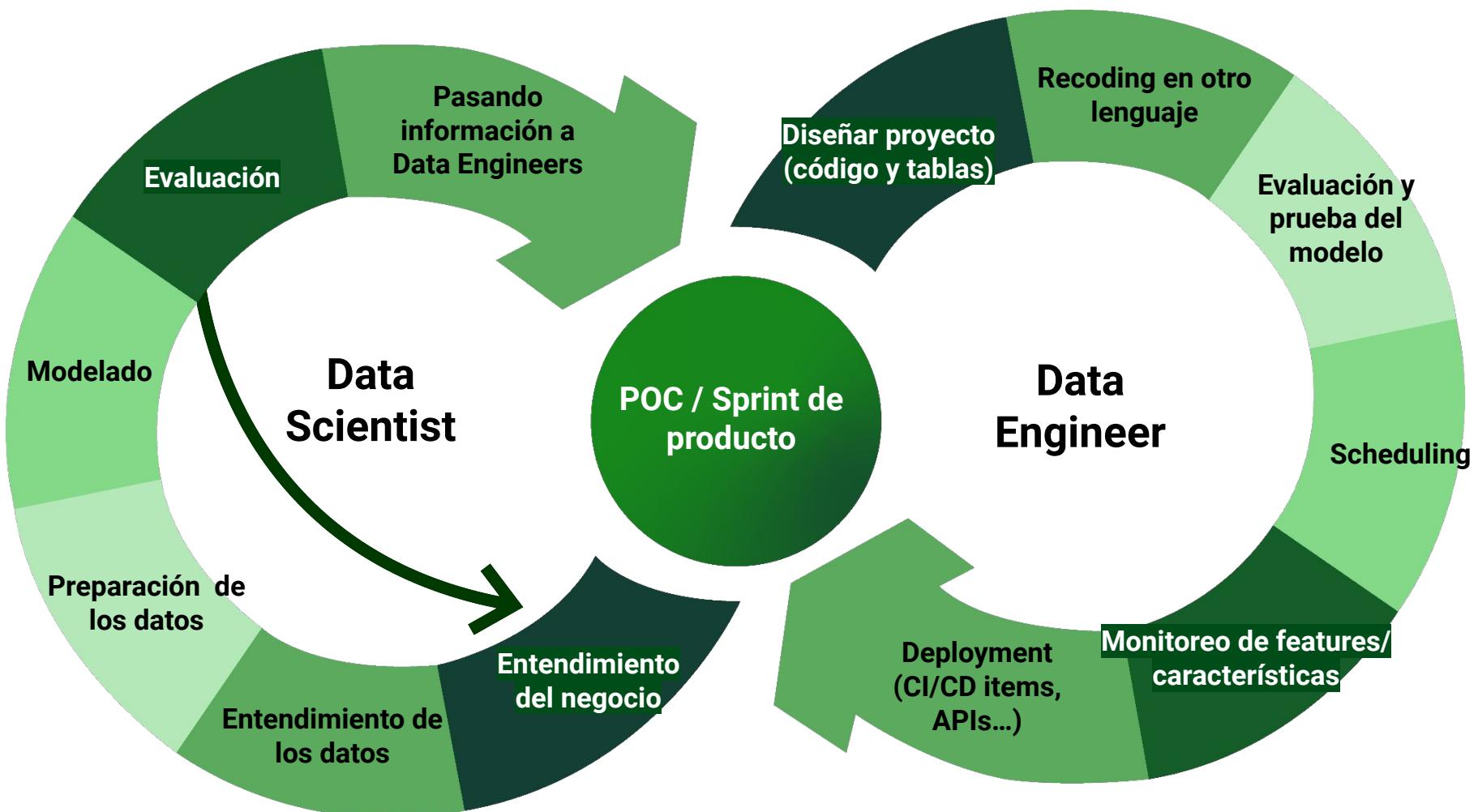
## Instrucciones:

Lista 3 empresas donde te gustaría trabajar  
como Data Engineer. Comenta por qué.

# Tareas de data engineer:

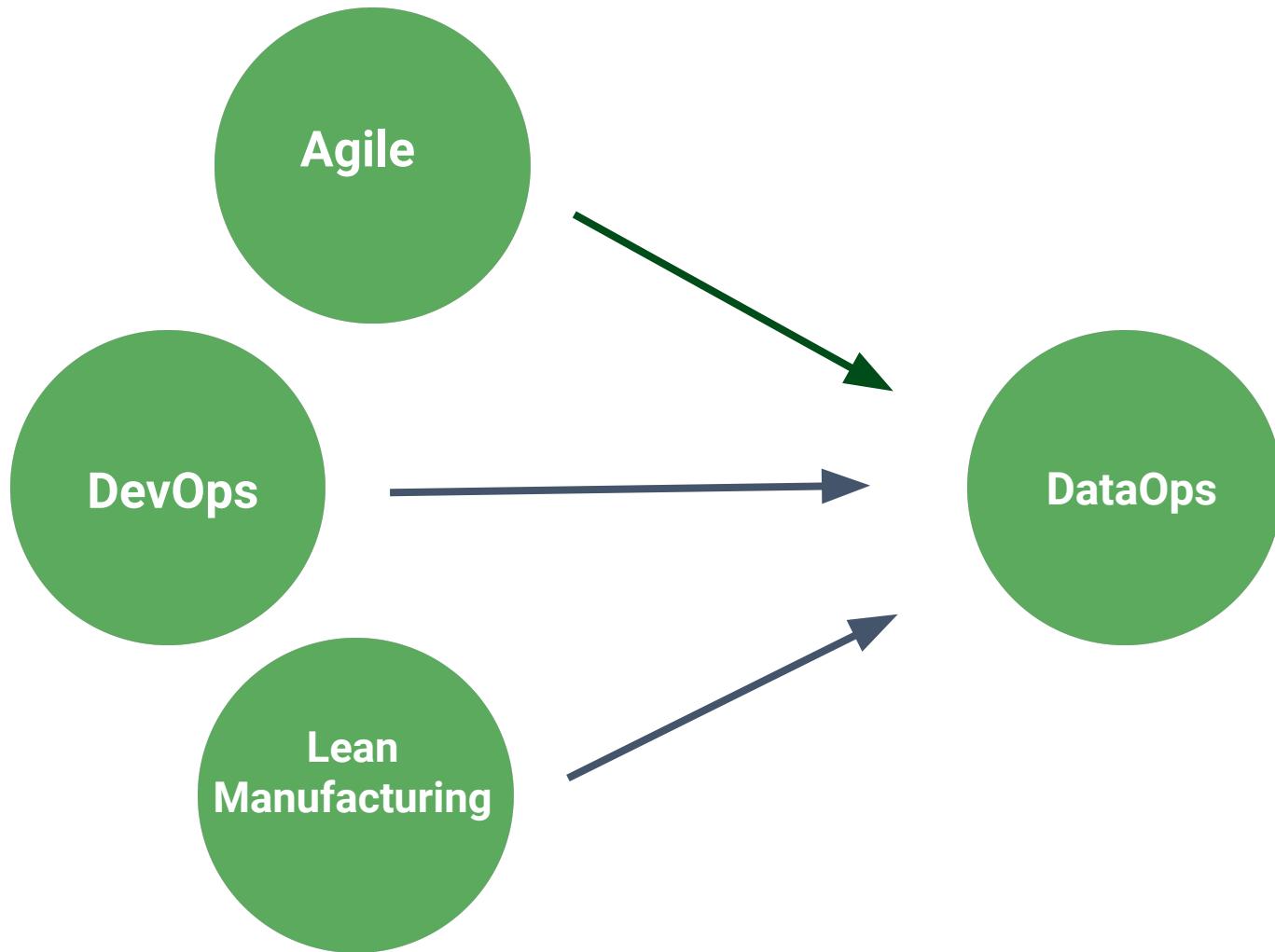
Data OPs

# DataOPs

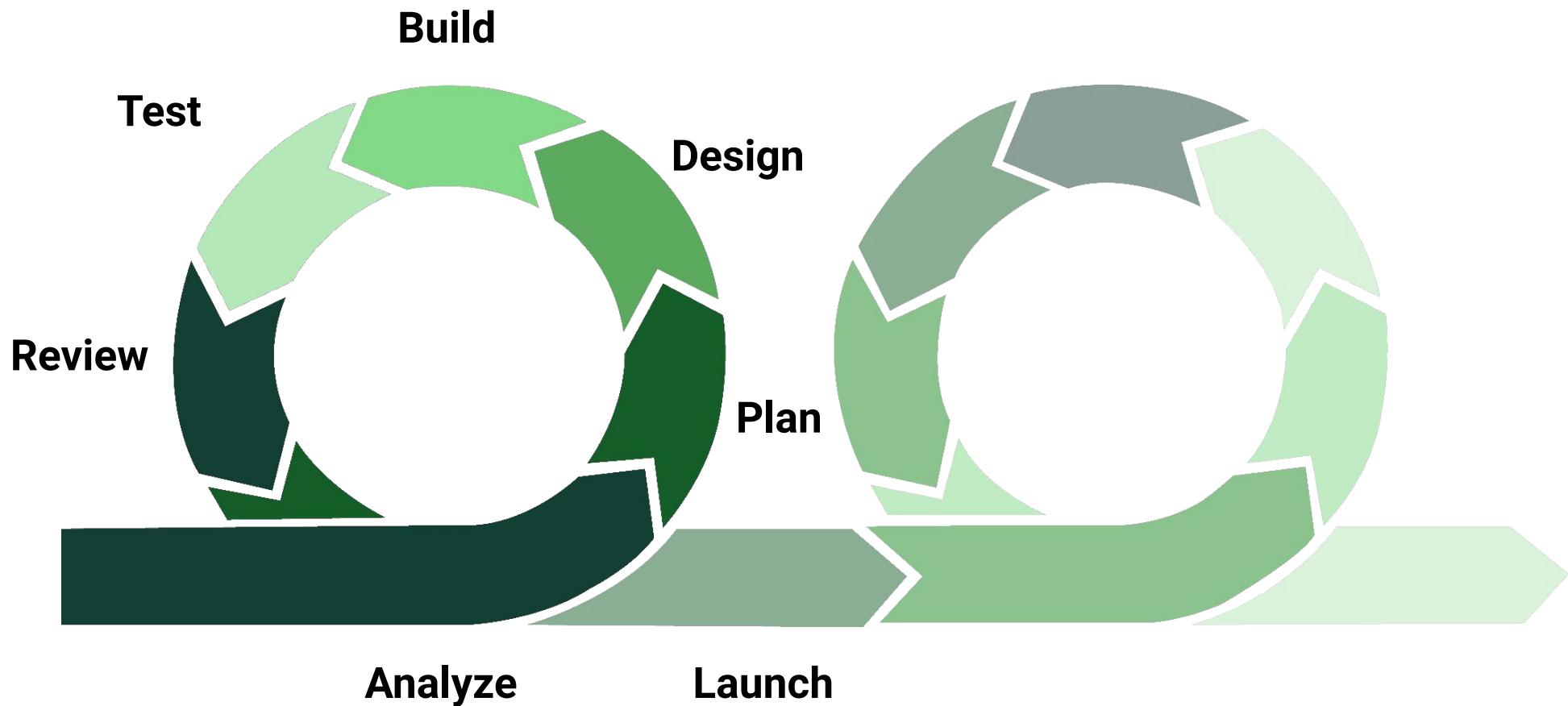


# Filosofía de DataOPs

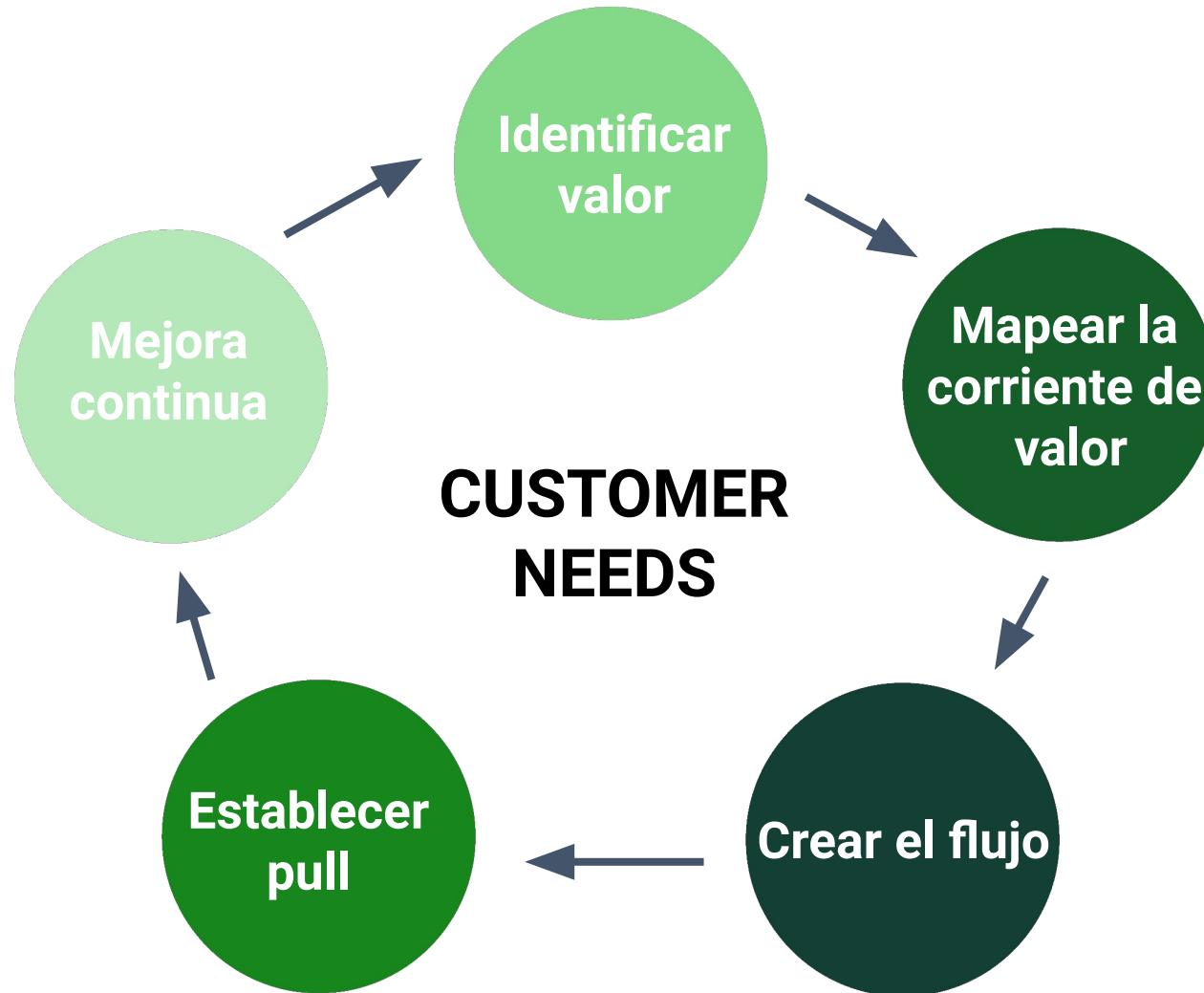
# Los componentes ideológicos



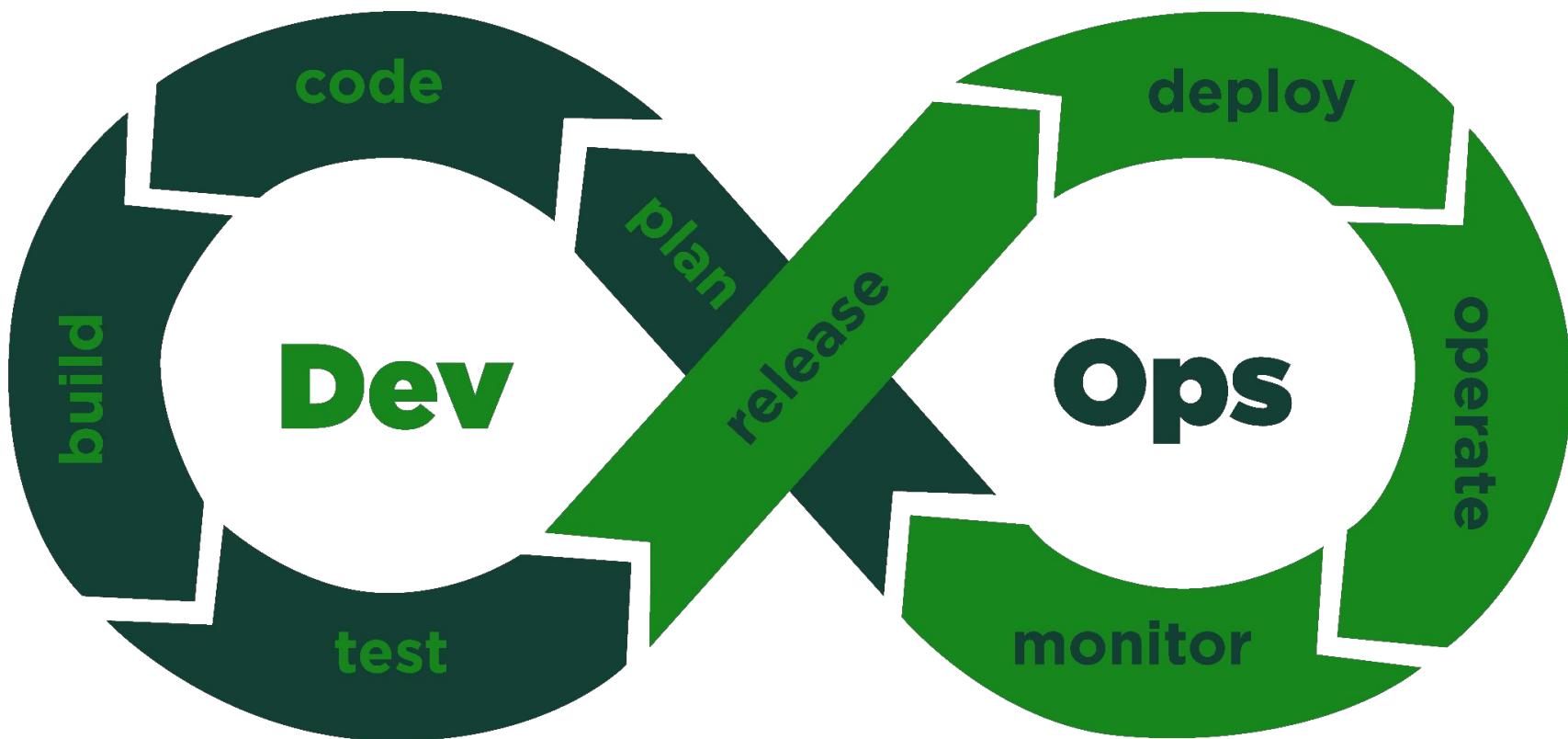
# Agile



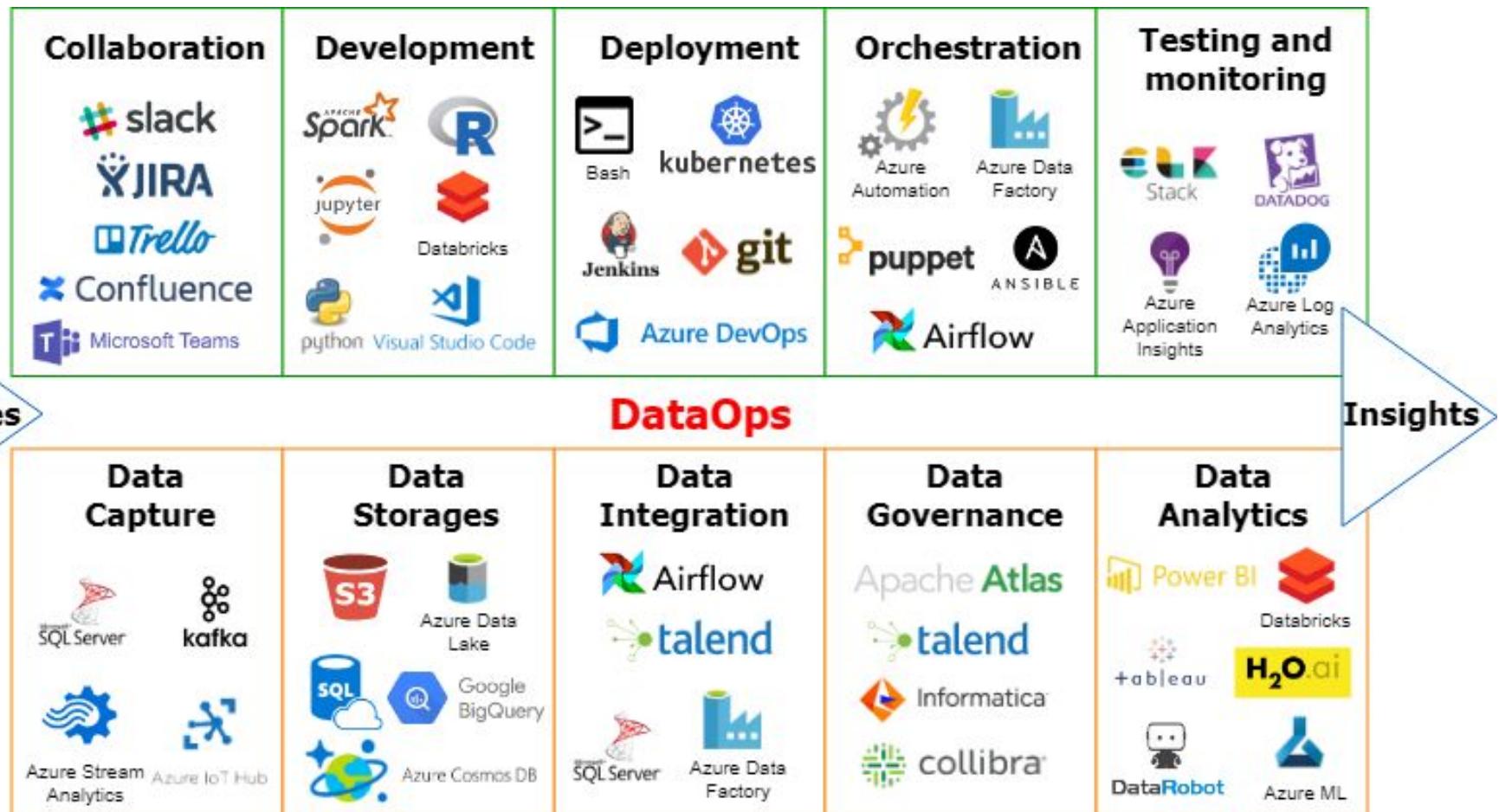
# Lean



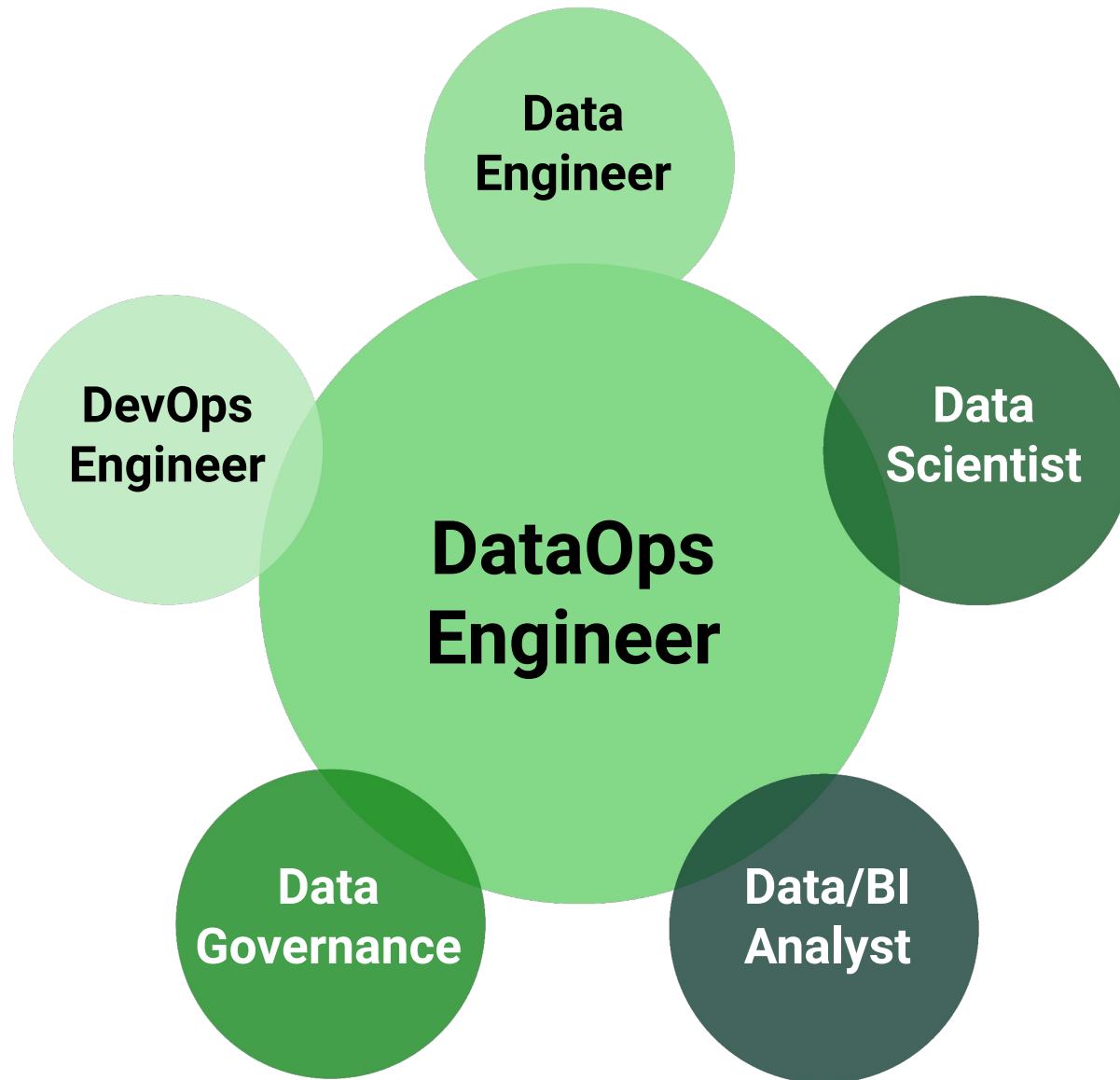
# DevOps



# Herramientas de DataOPs



# Equipos de DataOPs



# ¡Reto!

Reconoce tu camino

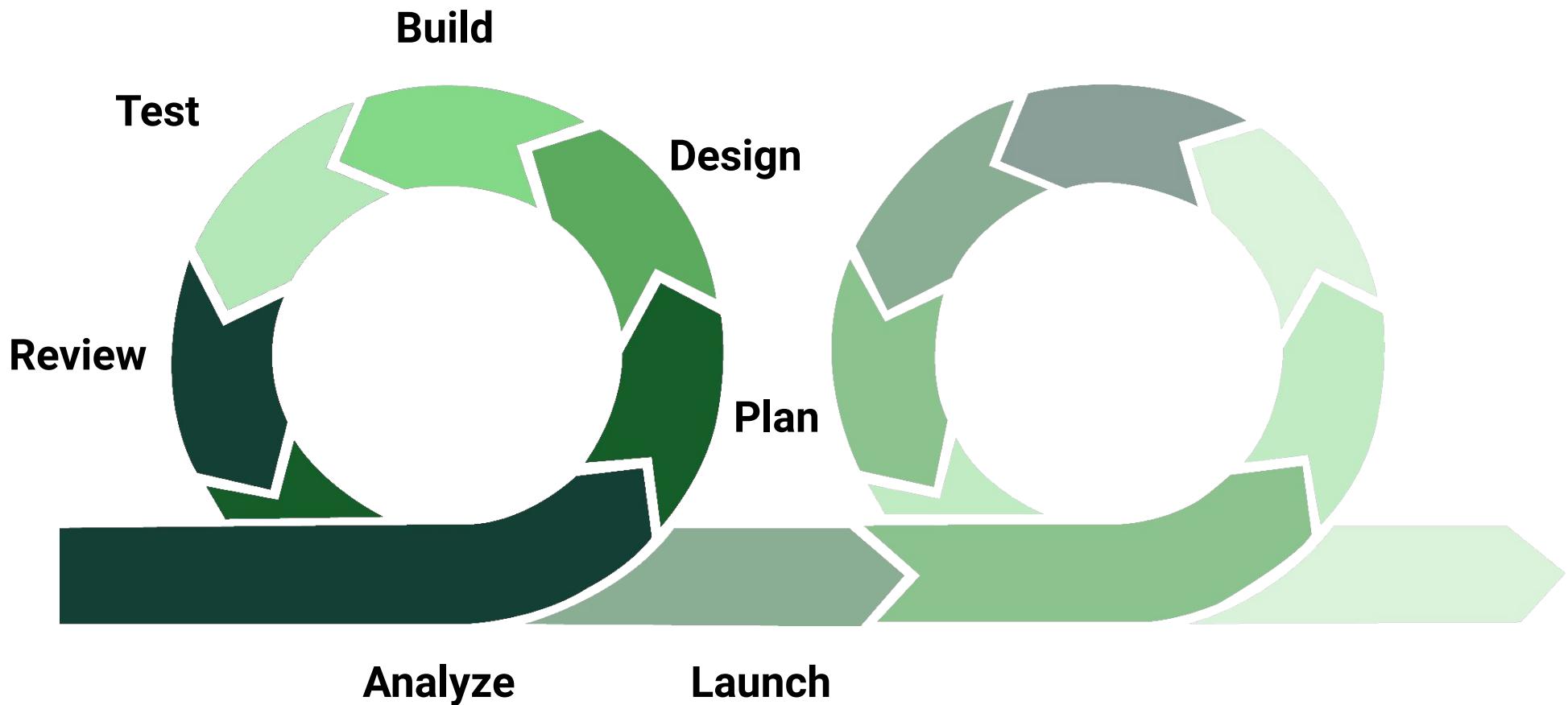
## Instrucciones:

Recapitula tus habilidades.  
¿Qué herramientas has utilizado  
previamente?

# Agile en Ingeniería de Datos

Planea

# Retomando Agile



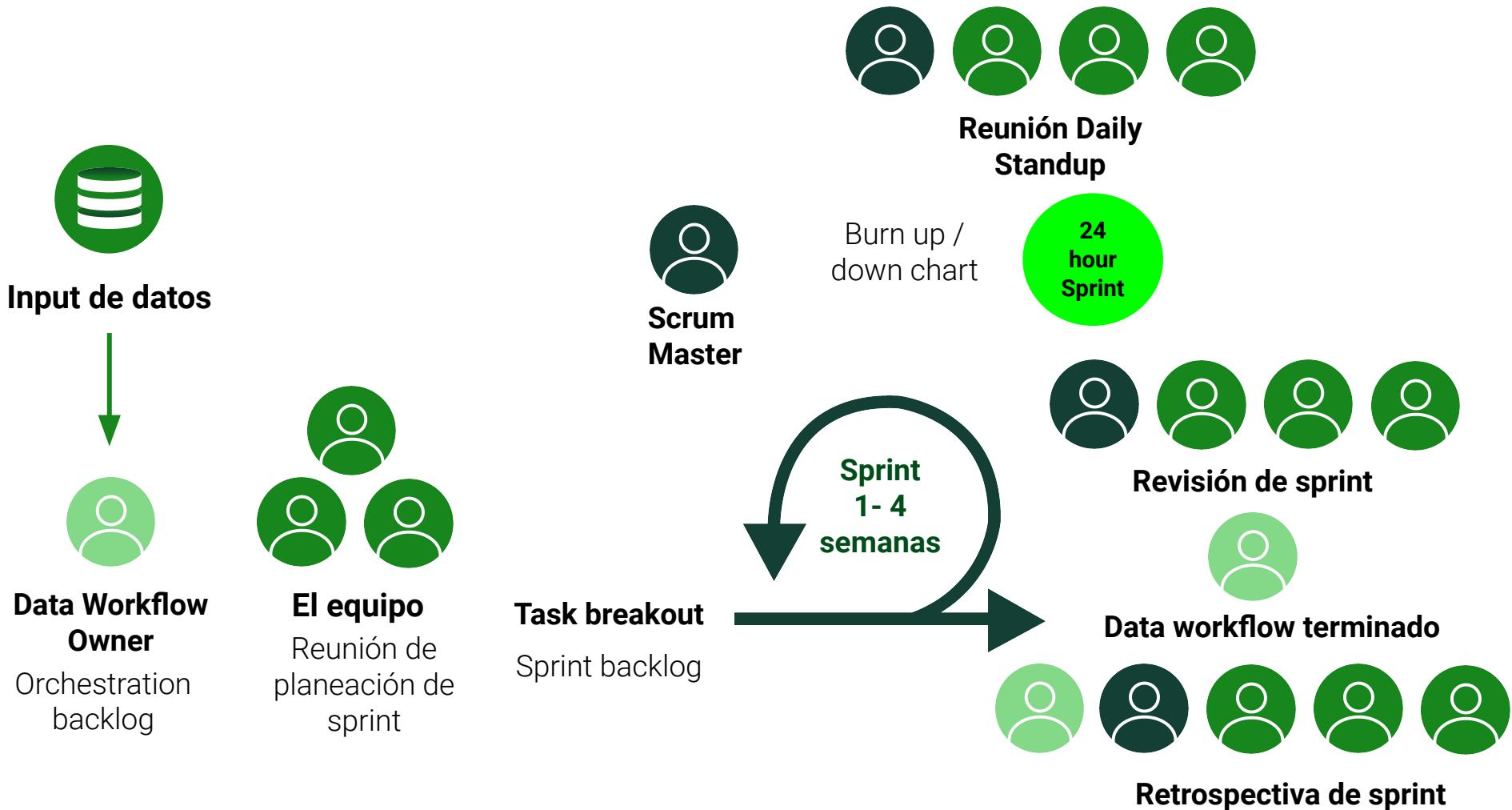


# Una advertencia

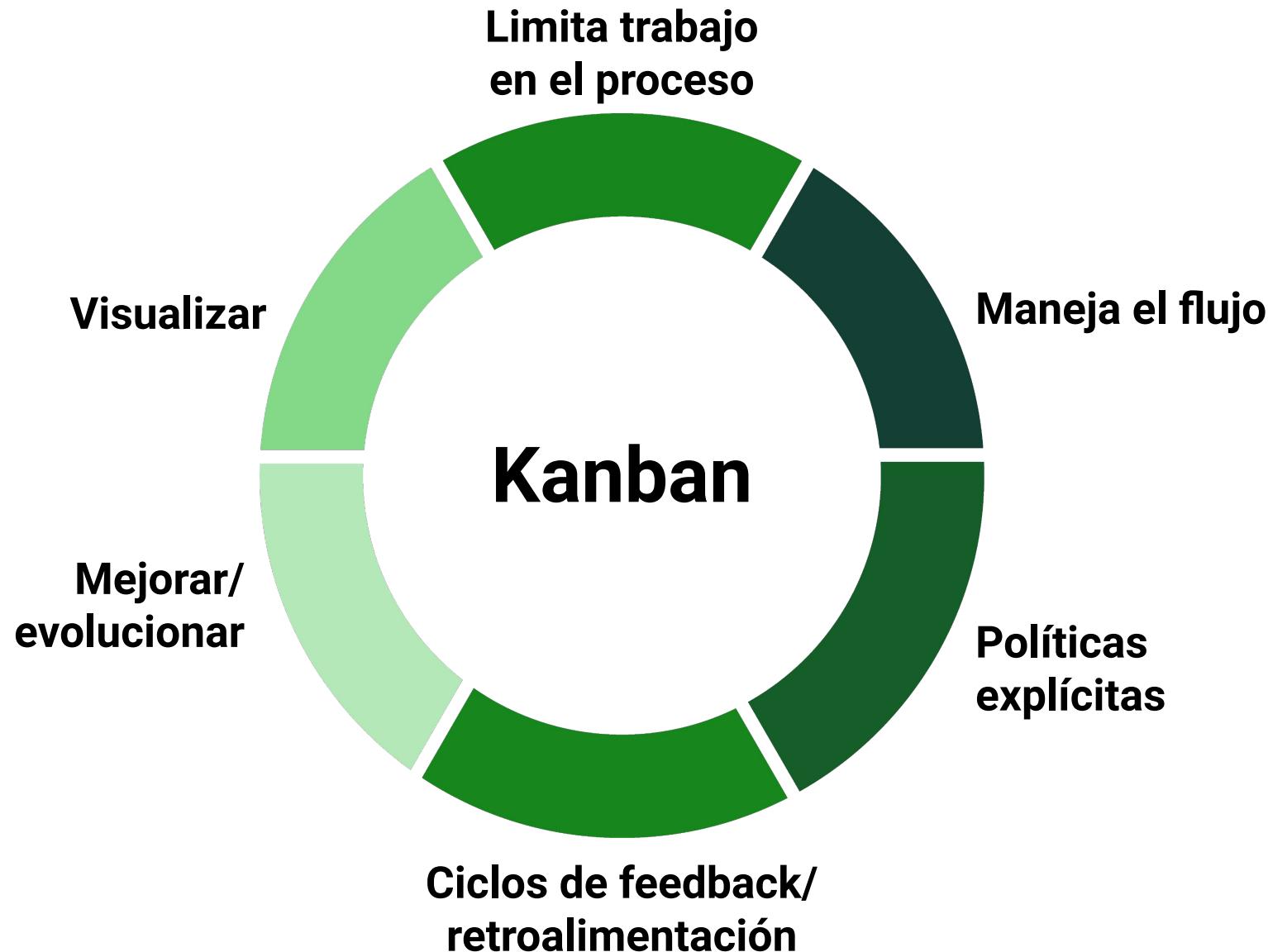
Muchos temas de Agile tienen **muchísima opinión**. Y en algunas ocasiones, todas tienen la “razón” con **posturas distintas**.

**Usa tu criterio, usa lo que te sirve.** 

# Scrum



# Kanban



# ¡Reto!

Kanban vs. Scrum

## Instrucciones:

¿Qué diferencias encuentras entre ellos?  
¿Qué ventajas y desventajas observas en  
cada uno?

# Lenguajes de Programación en Ingeniería de Software

Develop

# La importancia de programar

# La programación en producción



**Maquetar código  
vs.  
Código en producción**

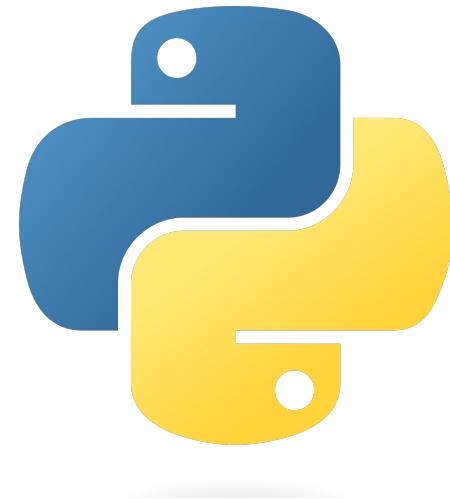
# Lenguajes de programación



# Python

**Nuestro caballo de batalla actual.**

- Librerías de código científico.
- Puede ser más lento que otras opciones.
- Mi favorito y el de muchas personas por su sencillez.



# R

## Elegido por estadistas.

- Herramienta estadística y antecesor de Python.
- Adelantado en modelos.
- Importante para analistas.



# Scala

**Spark se implementa sobre él.**

- Usa Java de base.
- Con la implementación y optimización de PySpark bajó su necesidad.
- Interesante para programación funcional.



# Java

Potente lenguaje multiplataforma.

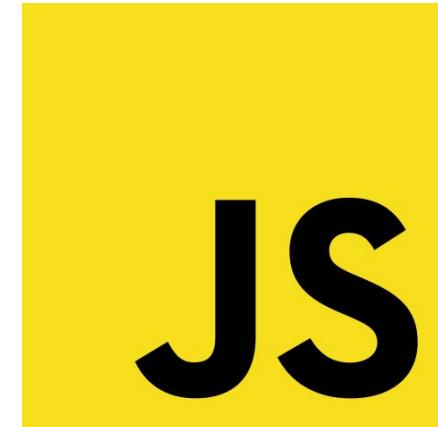
- Scala corre sobre Java.
- Su escalabilidad es envidiable.
- Puede ser un cómodo siguiente paso.



# JavaScript

**La navaja suiza para web developers.**

- Súper flexible y útil para muchos ámbitos.
- ImpONENTE por la cantidad de herramientas que tiene.
- Visualizaciones de datos más bellas posibles.



# C++ y derivados

**Columna vertebral de muchos proyectos.**

- Curva de aprendizaje potente.
- Muchas herramientas usan C en el fondo.
- Implementaciones modernas ayudan a que no sea tan difícil de implementar.





# **Lo importante de los lenguajes**

**No es que los colecciones, sino que te sientas  
con comodidad implementando en diversos  
paradigmas de programación.**

**Lo que vale es llevar tus ideas hasta generar  
valor.** 

# ¡Reto!

## Python para Data Engineer

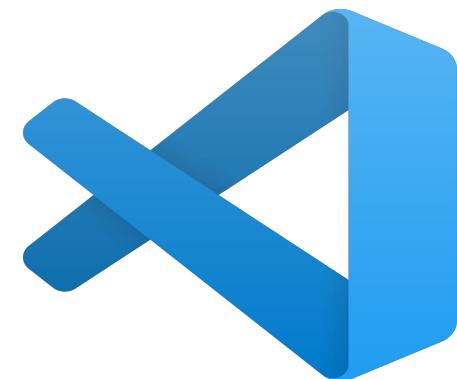
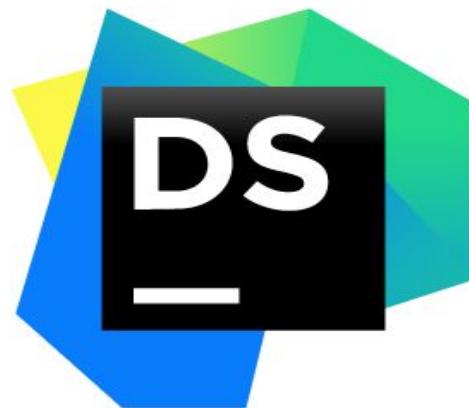
### **Instrucciones:**

Busca herramientas/librerías en Google que se usan en ingeniería de datos

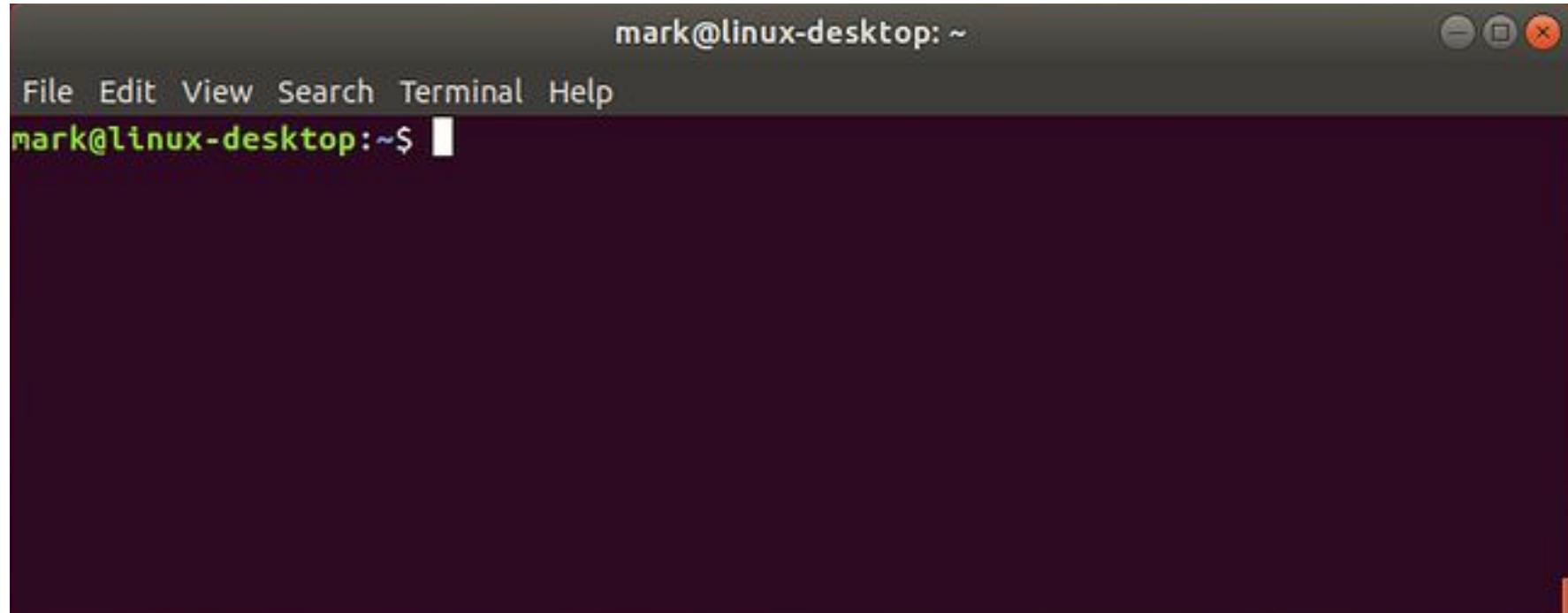
# ¿Dónde y cómo escribir tu código?

Develop

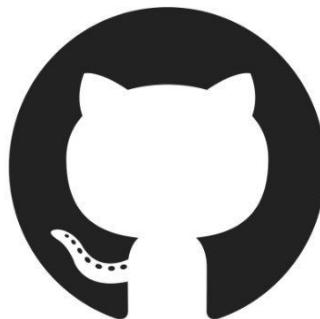
# Jupyter Notebooks vs. IDE vs. Editor de código



# La terminal es tu amiga



# Git también te quiere



git



# ¡Reto!

Jupyter Notebooks vs. IDE vs.  
Editor de código

## Instrucciones:

Busca ventajas y desventajas  
de estas tres herramientas.

# Automatización y scripting

Develop

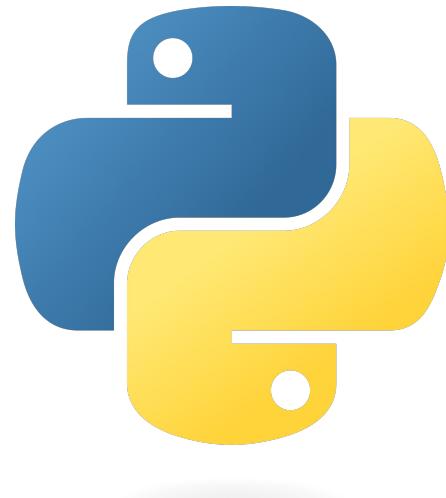


# Automatizar las tareas

- Repetir una tarea
- Trabajar de manera inteligente
- Optimizar el proceso
- Utilizar recursos externos

# ¿Por qué Python?

- Lenguaje sencillo
- Librerías variadas
- Comunidad activa
- Multiplataforma



# Ejemplos de automatización



Wholesale factory x3 pro smartphones 5600mah big...

**ARS 4,271.38 -**

**ARS 5,155.11**

1 piece (MOQ)



Factory direct inventory new 55-inch TCL smart TV

**ARS 17,674.66 -**

**ARS 19,147.55**

1 piece (MOQ)



Hot AMAZON 2022 action sports camera go pro Full...

**ARS 412.41 - ARS 883.74**

1 piece (MOQ)



Bestitalian IPTV lista Iptv Italia senza Italian Reseller Panel...

**ARS 4,418.67 -**

**ARS 5,420.23**

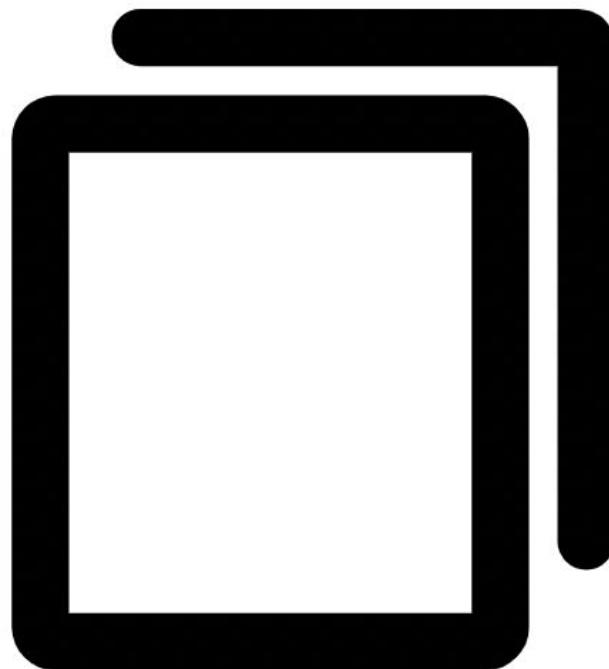
1 piece (MOQ)



3-pro-smartphones-5600mah\_1600357464142.html

# Generar y copiar archivos

Descargar una fuente de datos externa.



# Modificaciones masivas

Editar y limpiar conjuntos de datos.

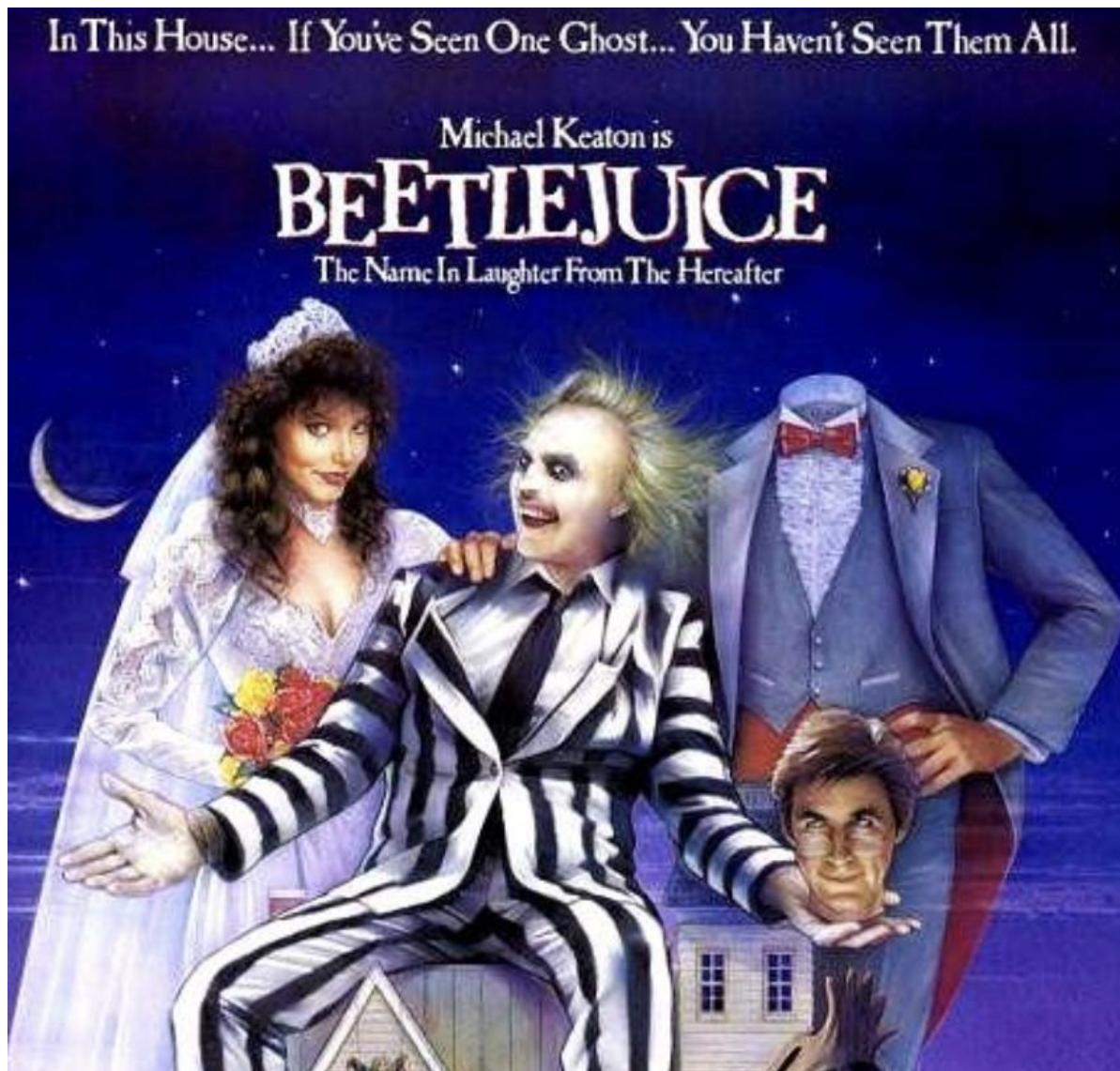
	A1	:	X	✓	f <sub>x</sub>	Date,Open,High,Low,Close,Volu
	A	B	C	D	E	
1	Date,Open,High,Low,Close,Volume					
2	8-Dec-16,61.30,61.58,60.84,61.01,21043447					
3	7-Dec-16,60.01,61.38,59.80,61.37,30808969					
4	6-Dec-16,60.43,60.46,59.80,59.95,19907035					
5	5-Dec-16,59.70,60.58,59.56,60.22,23552658					
6	2-Dec-16,59.08,59.47,58.80,59.25,25515665					
7	1-Dec-16,60.11,60.15,58.94,59.20,34542121					
8	30-Nov-16,60.86,61.18,60.22,60.26,34655435					
9	29-Nov-16,60.65,61.41,60.52,61.09,22366721					
10	28-Nov-16,60.34,61.02,60.21,60.61,20732619					
11	25-Nov-16,60.30,60.53,60.13,60.53,8409616					
12	23-Nov-16,61.01,61.10,60.25,60.40,21848913					
13	22-Nov-16,60.98,61.26,60.80,61.12,23206700					
14	21-Nov-16,60.50,60.97,60.42,60.86,19652595					

# ¡Cuidado!

- Automatizar cosas que no sean repetitivas.
- Automatizar antes de tiempo.



# Regla Beetlejuice



# ¡Reto!

Automatización de tareas

## Instrucciones:

Lista las tareas que haces repetitivamente que podrías automatizar.

# Fuentes de datos: SQL, NoSQL, APIs y web scraping

Develop

# **La centralidad de las bases de datos**

**Data Retrieval**

**Data Redundancy**

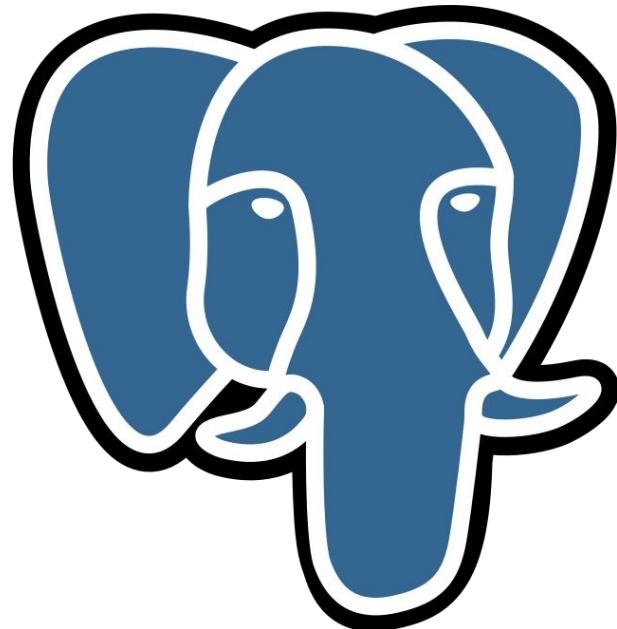
**Data Integrity**

**Data Security**

**Data Indexing**

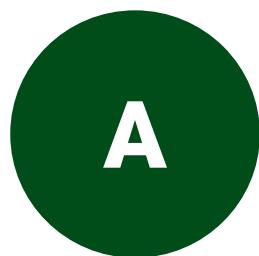
# Utilidad de SQL

Lenguaje de consulta, pero también el nombre con el que se identifican las bases de datos.

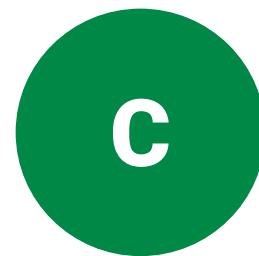


# Utilidad de SQL

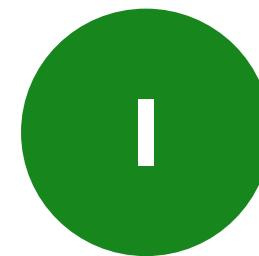
Excelentes para transacciones por los principios ACID.



Atomicity



Consistency



Isolation



Durability

# Utilidad de NoSQL

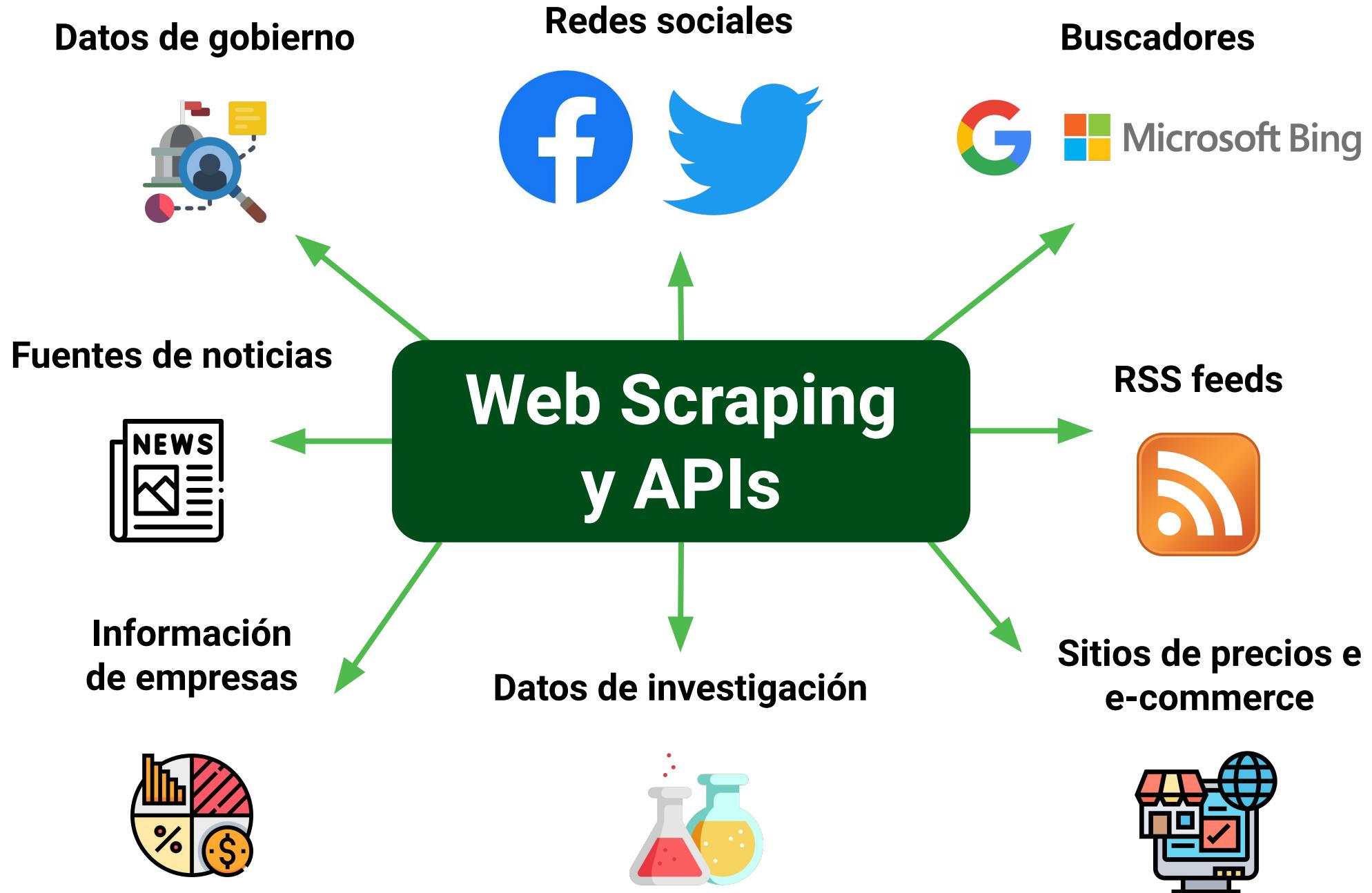
Por su cercanía a los lenguajes de programación, son útiles para guardar objetos flexibles.

{;}

# Utilidad de NoSQL

- La más famosa es MongoDB.
- Otras importantes son Redis, ElasticSearch y HBase.





# API

## Application Programming Interface

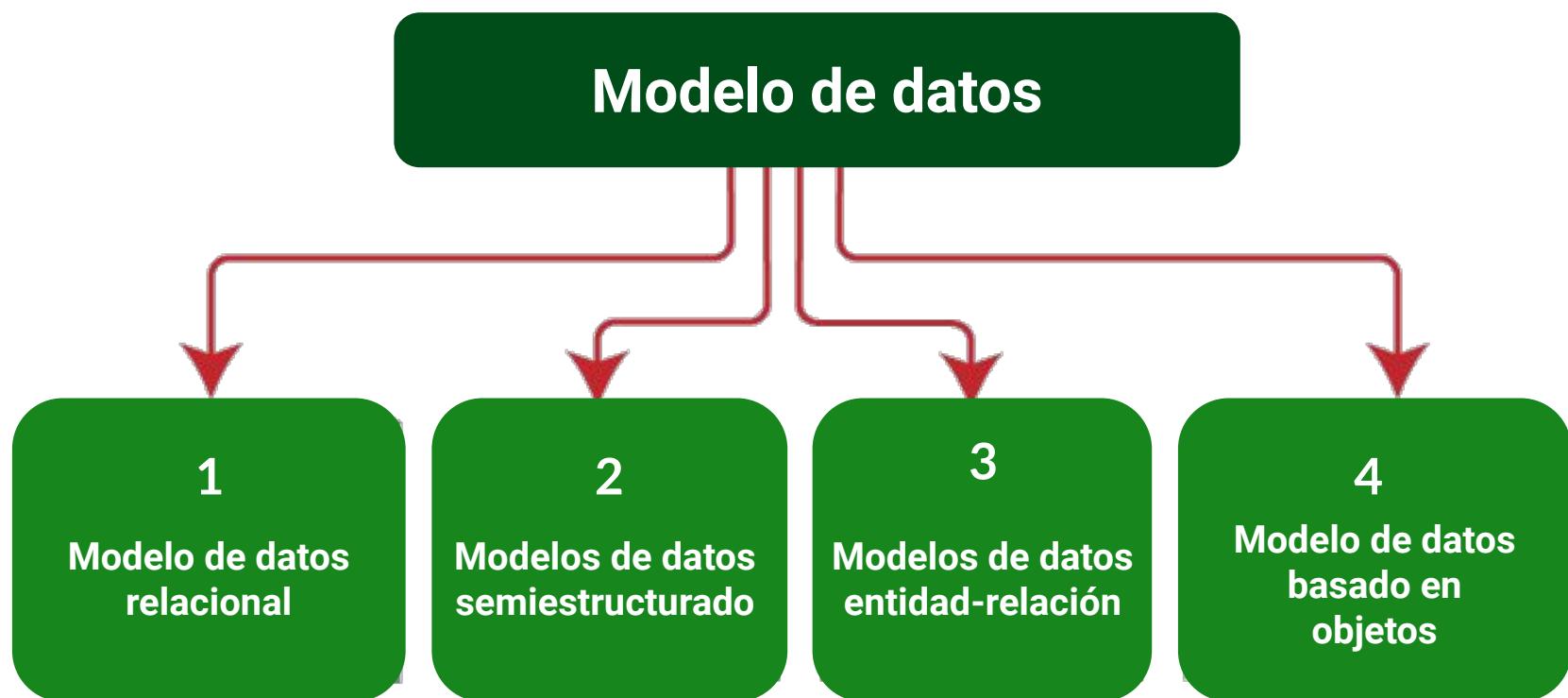
- Consume información de otras plataformas.
- Permite utilizar capacidades mandando un input, recibiendo un output.
- Pueden ser creadas por uno, externas y de paga.

# Web scraping

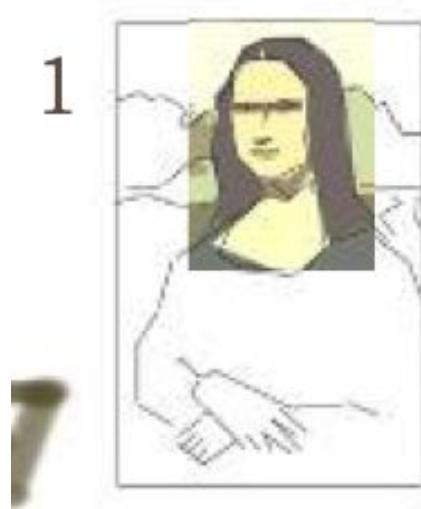


Scrapy

# Diseño basado en modelos



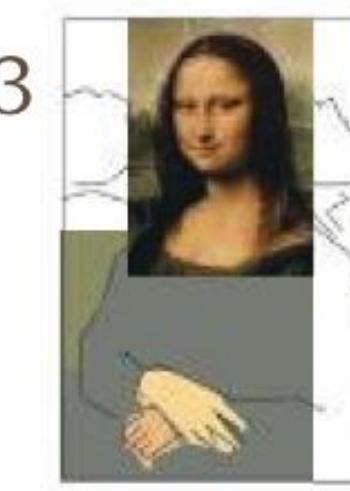
# Generando modelos sostenibles



1



2



3



4



5



6

# ¡Reto!

Bases de datos  
en la cotidianidad

**Instrucciones:**  
Investiga qué bases de datos  
usan tus apps favoritas.

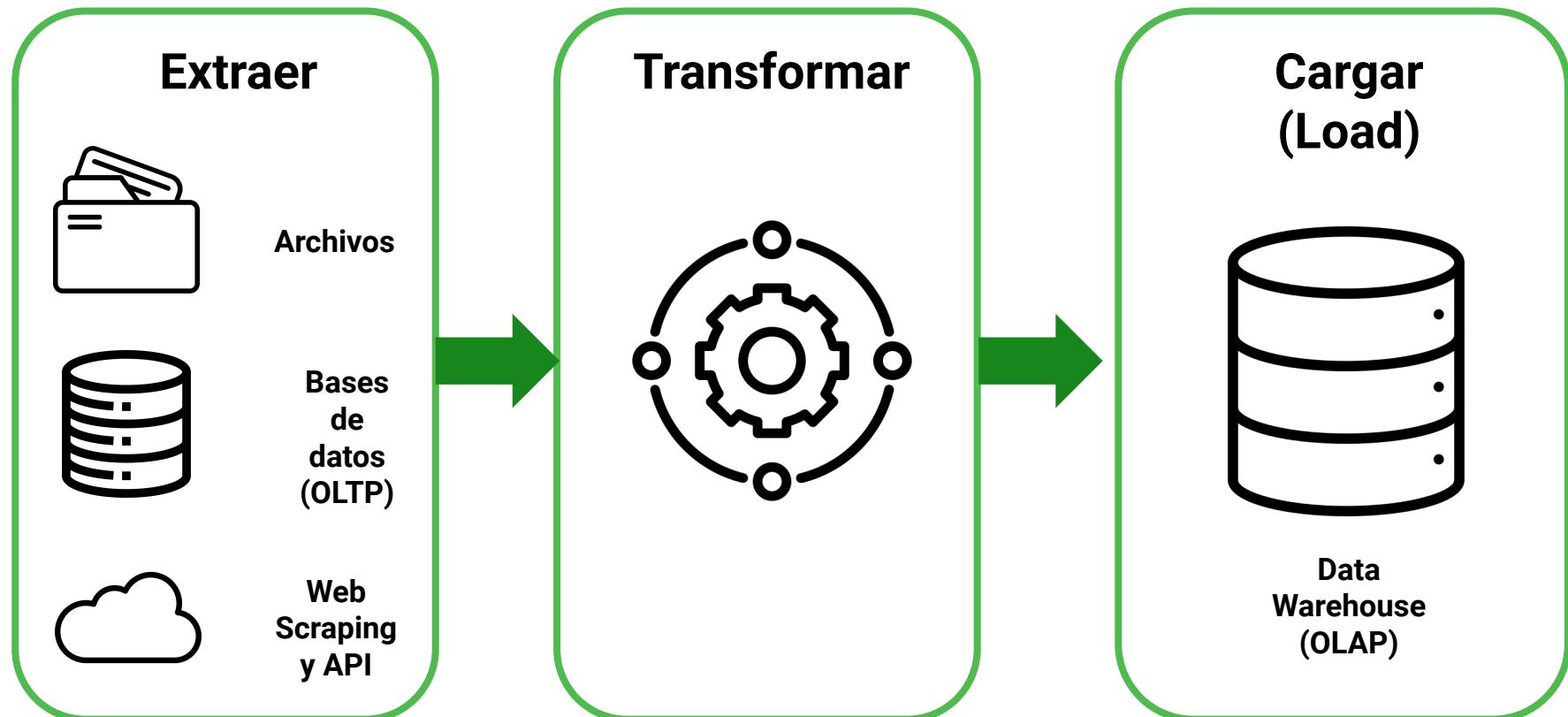
# Procesamiento de datos: pipelines, Spark y cómputo paralelo

Develop

# El escenario de datos

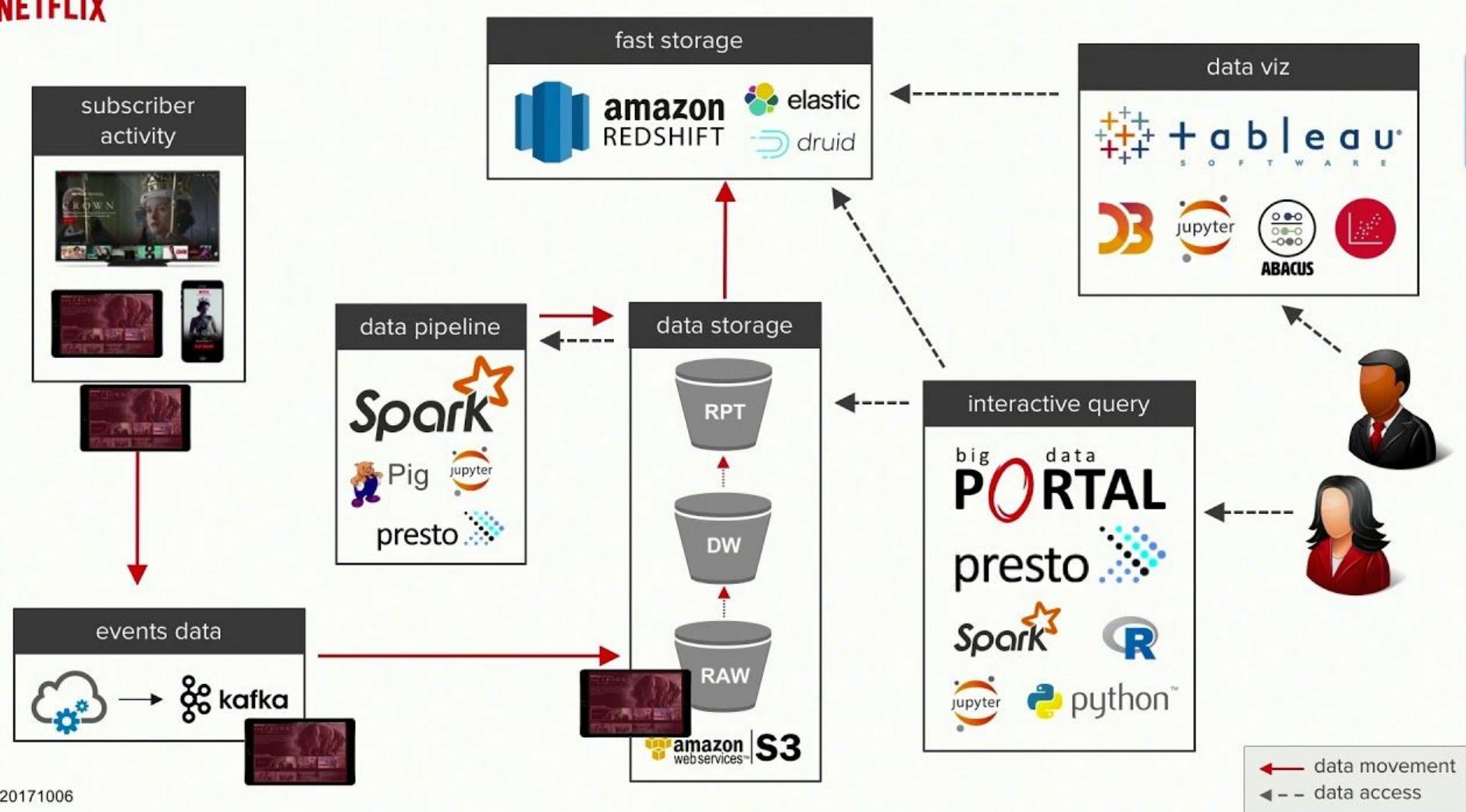


# ¿Qué es un pipeline de datos?



# Ejemplo de pipeline

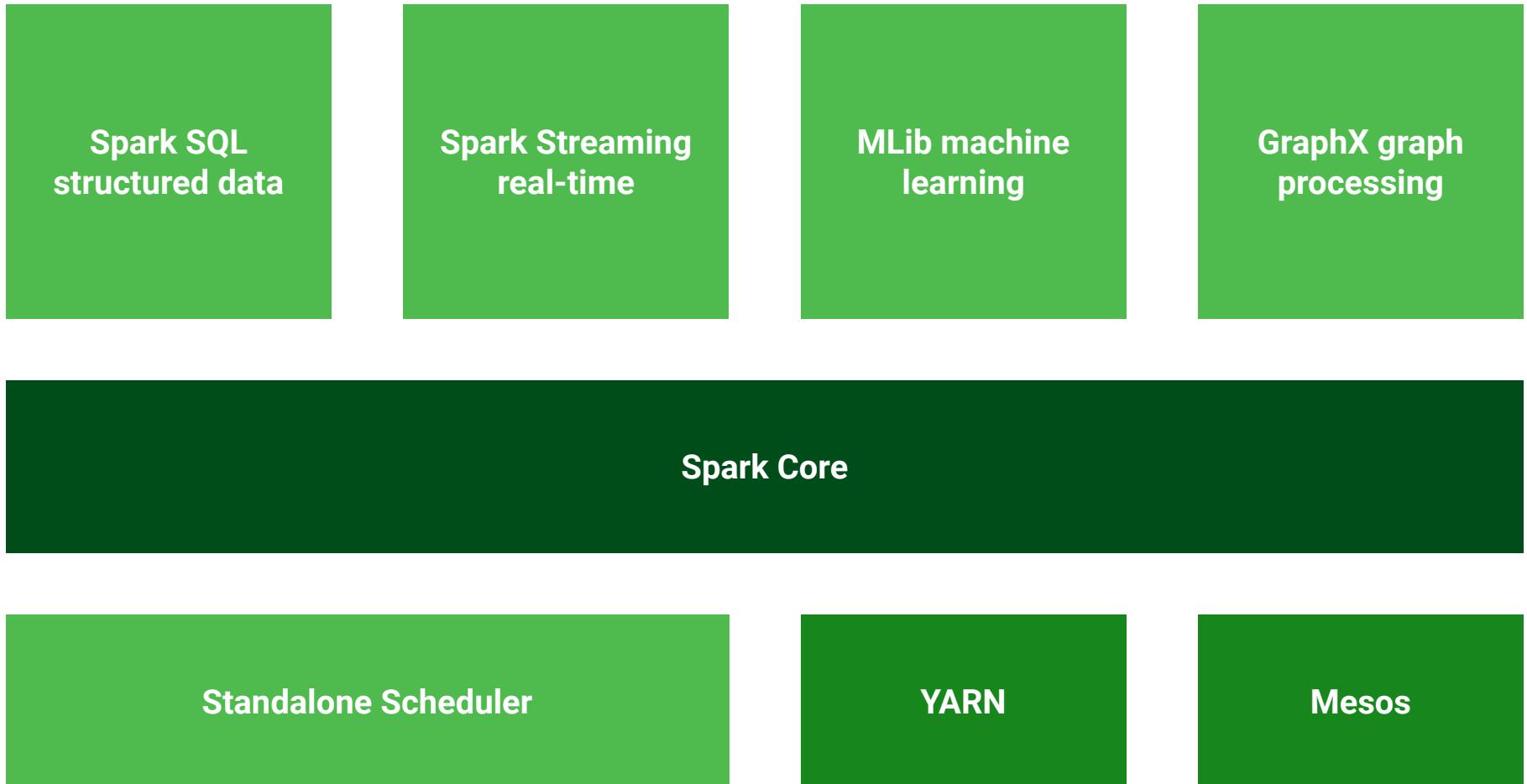
NETFLIX



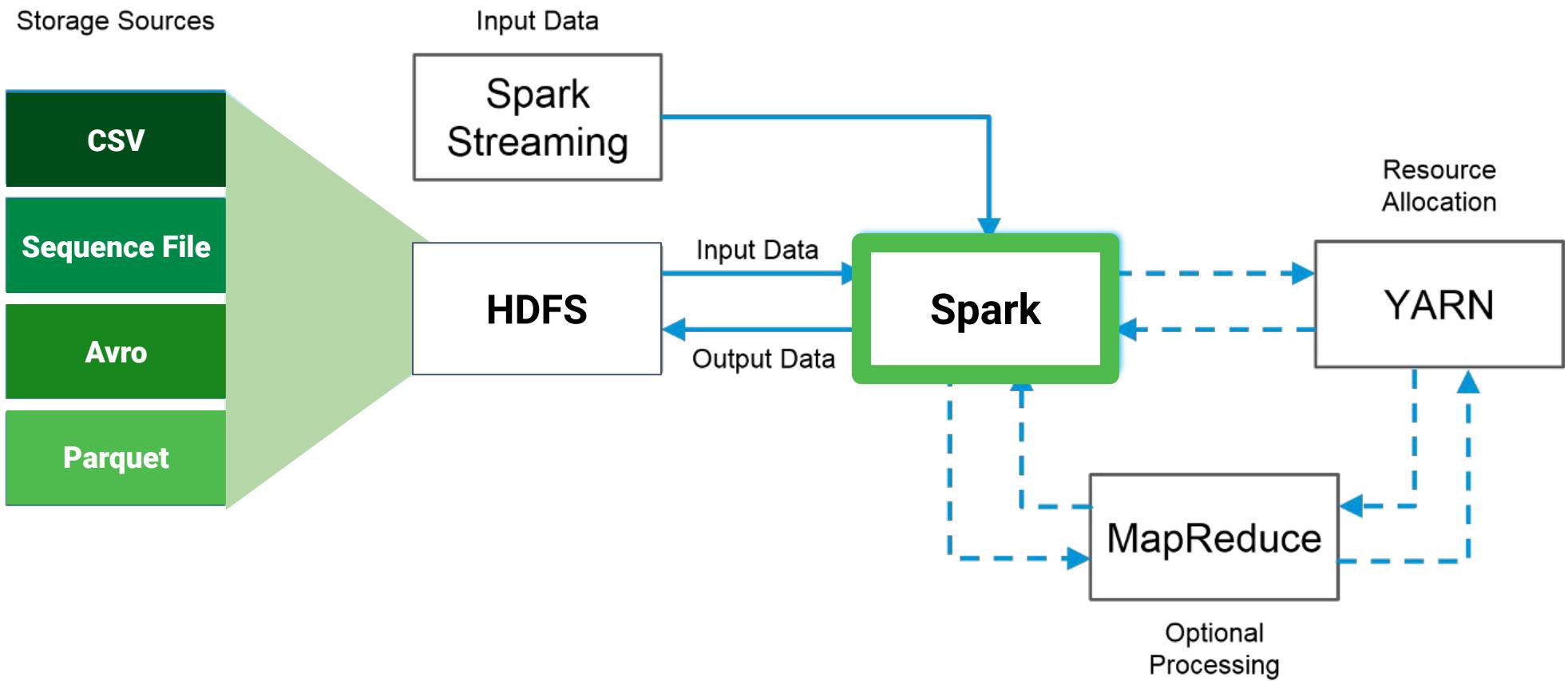
# ¿Qué busca Apache Spark?



# Stack de Apache Spark



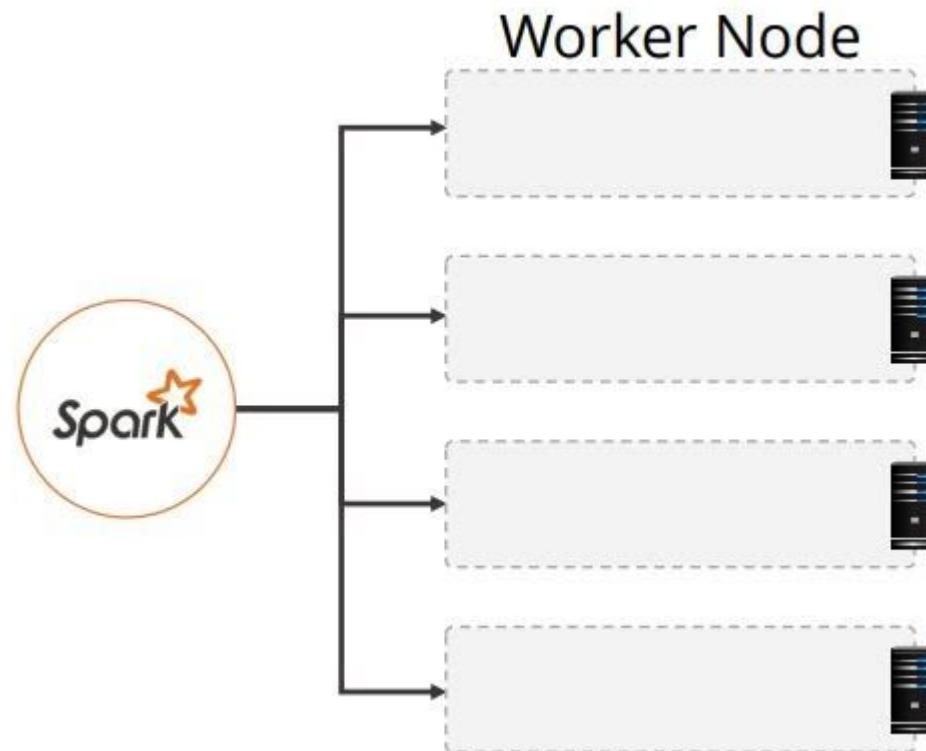
# Procesamiento en Batch



# Procesamiento en Stream



# Apoyándose del procesamiento paralelo



# ¡Reto!

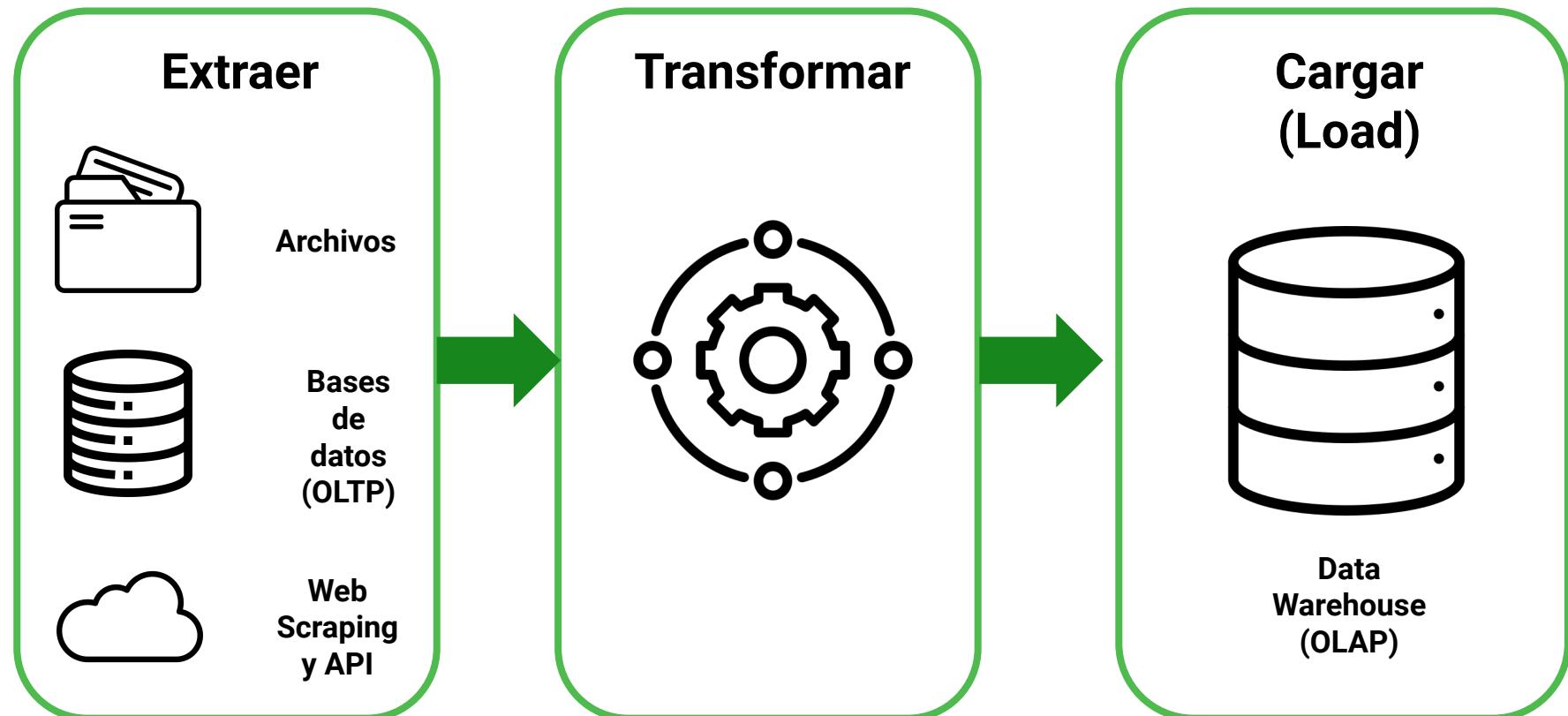
Spark  
en la cotidianidad

**Instrucciones:**  
Investiga qué empresas/apps  
usan Spark

# Automatizar los pipelines: Airflow

## Develop

# Retomemos los pipelines

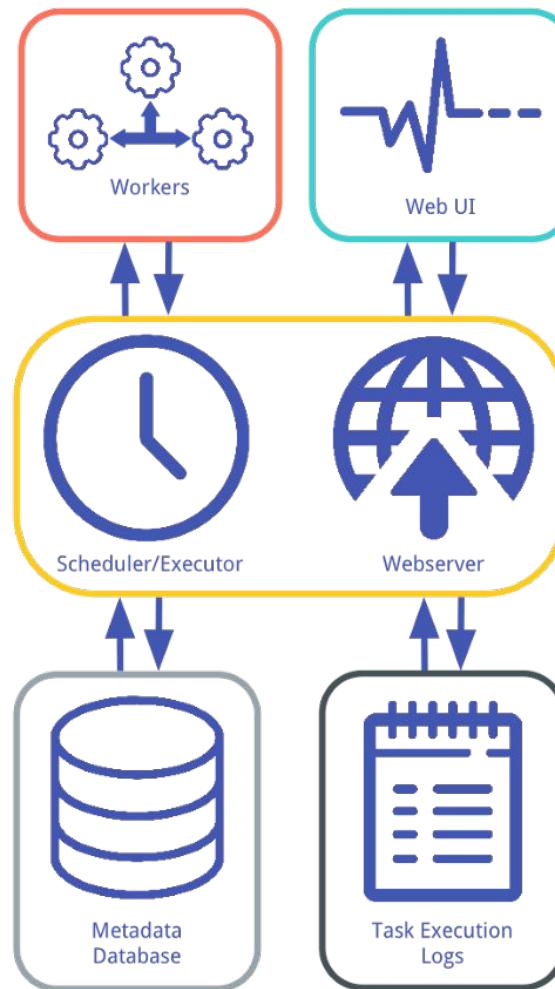


# Tres preguntas importantes

- ¿Qué va a correr?
- ¿En qué secuencia?
- ¿Cuándo y cada cuánto va a correr?

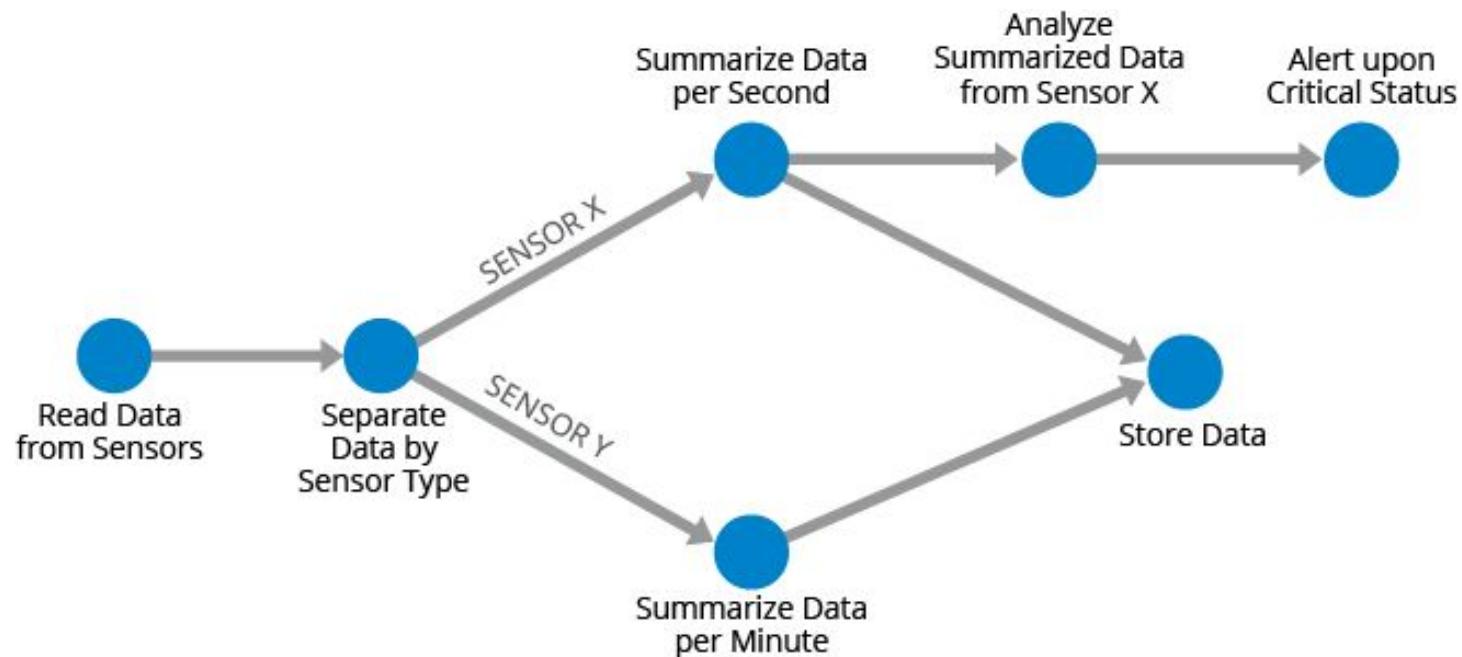
# Airflow

Airflow's General Architecture



# DAGs

***Directed Acyclic Graphs - Grafos Acíclicos Direcionados***



# Acompáñame a conocer un DAG



Apache  
**Airflow**

# ¡Reto!

Conociendo un  
repositorio de Airflow

**Instrucciones:**  
Explora el repositorio del  
proyecto de la clase.

# Containers y empaquetamiento: Docker y Kubernetes

Build

# El problema de los ambientes

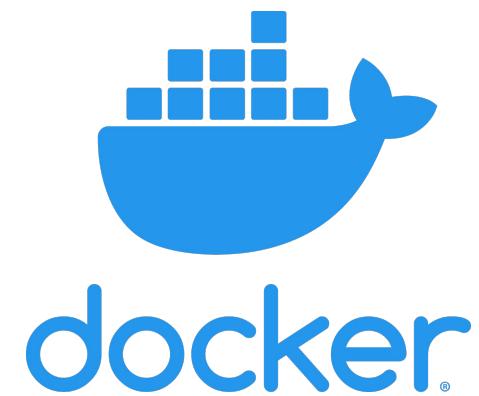
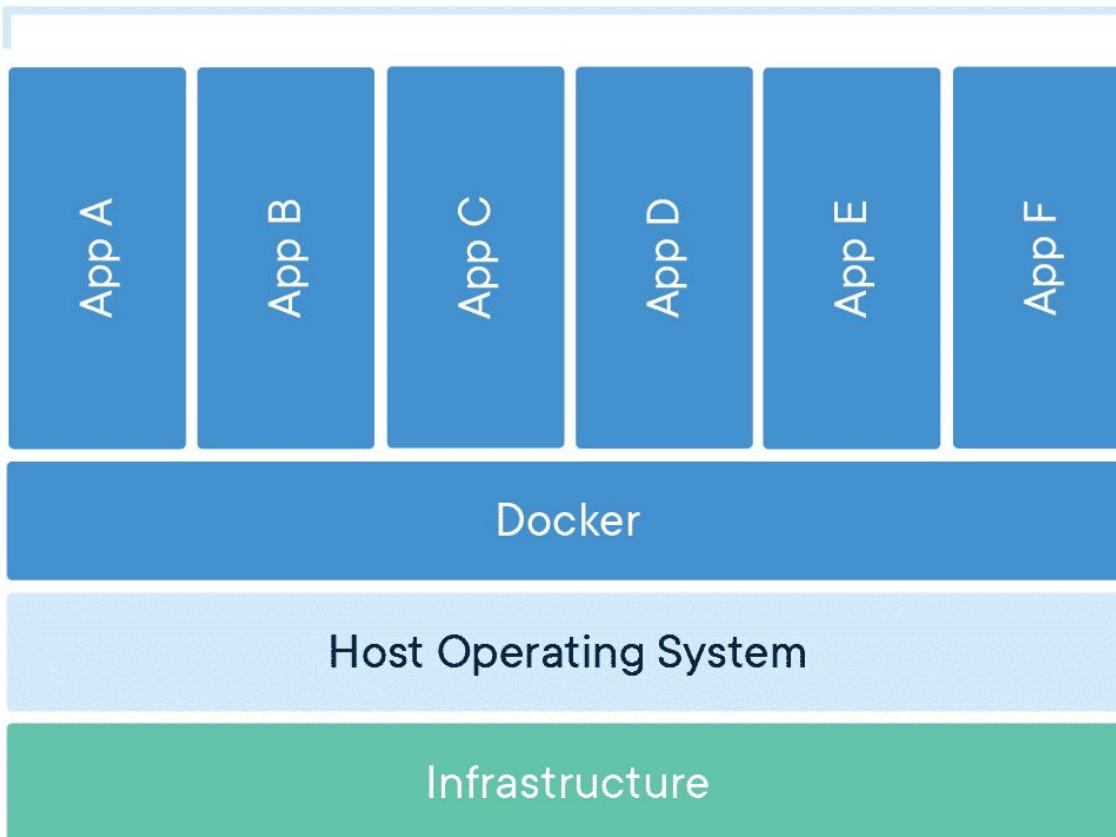
**EN LOCAL**



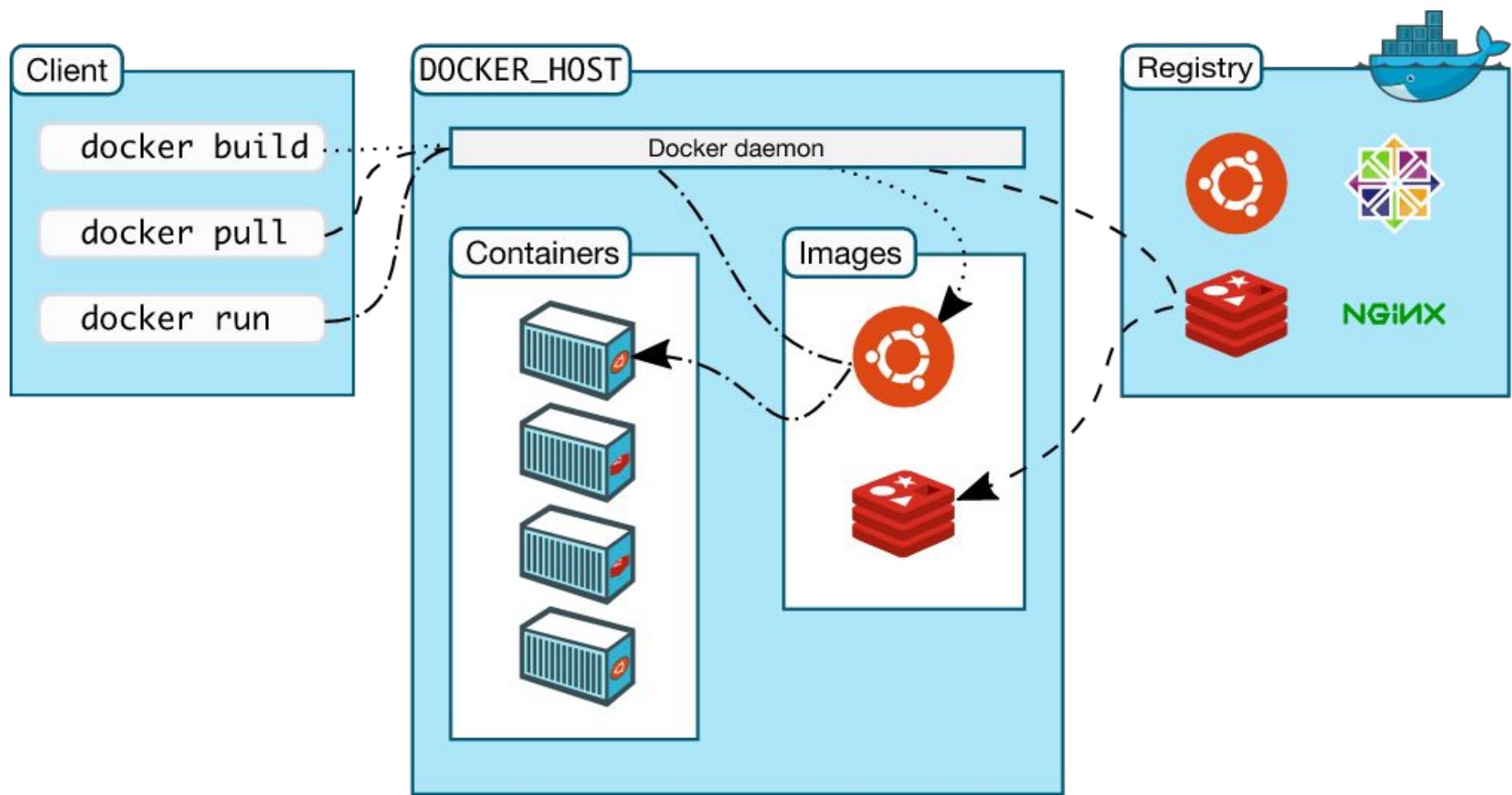
**FUNCIONA**

# Docker

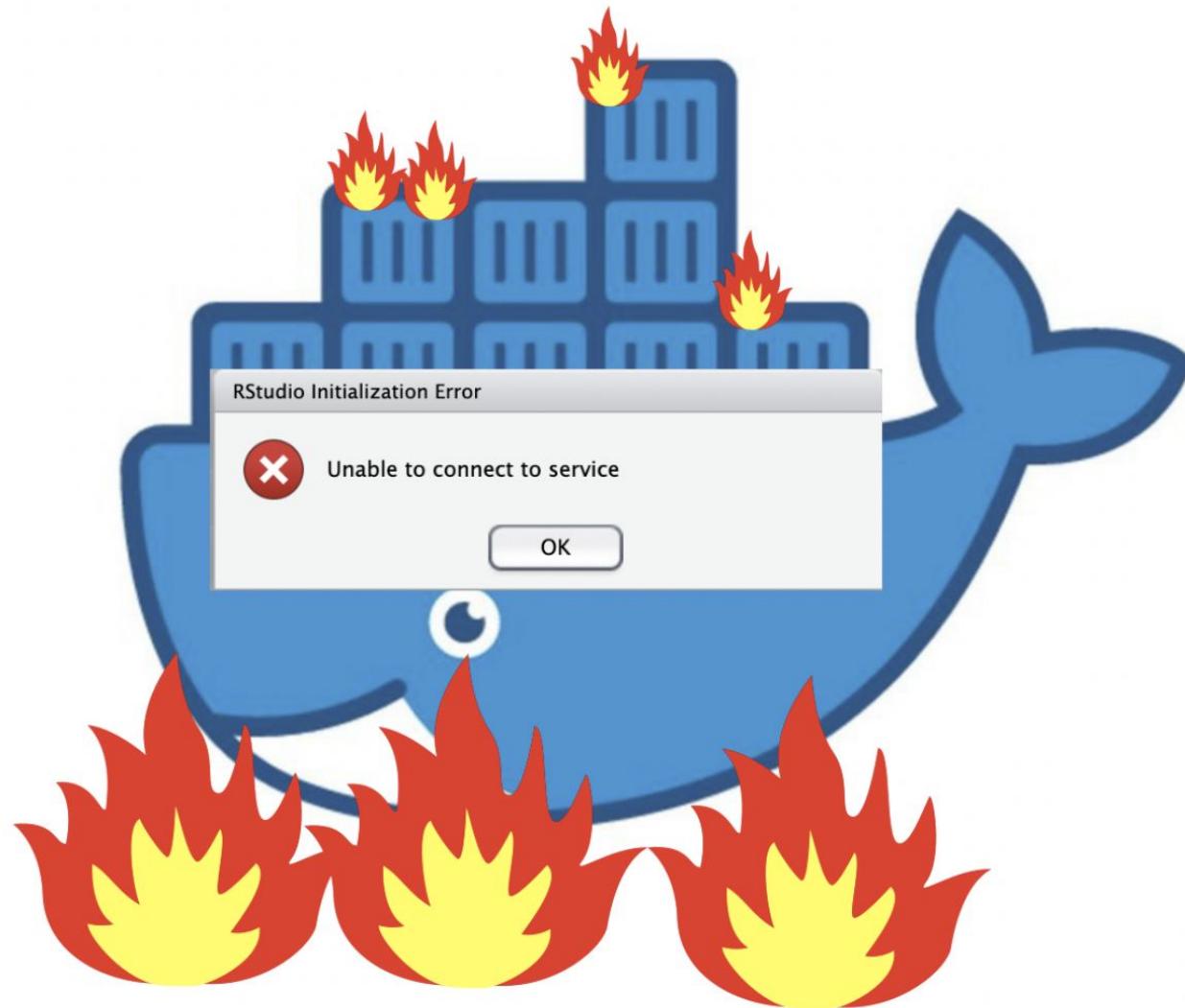
Containerized Applications



# Imágenes y contenedores



# El tema con la virtualización



# Orquestación

Container Orchestration Software  
(Docker, Openshift & Kubernetes)



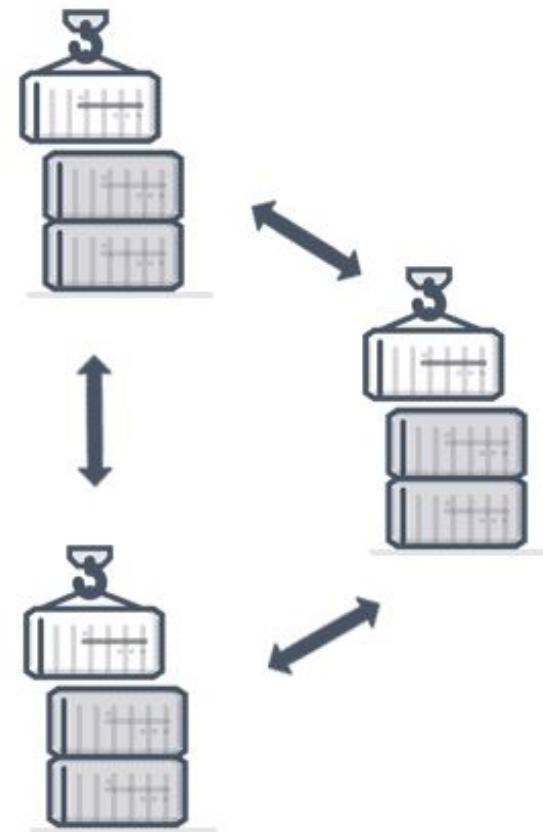
OPENSIFT



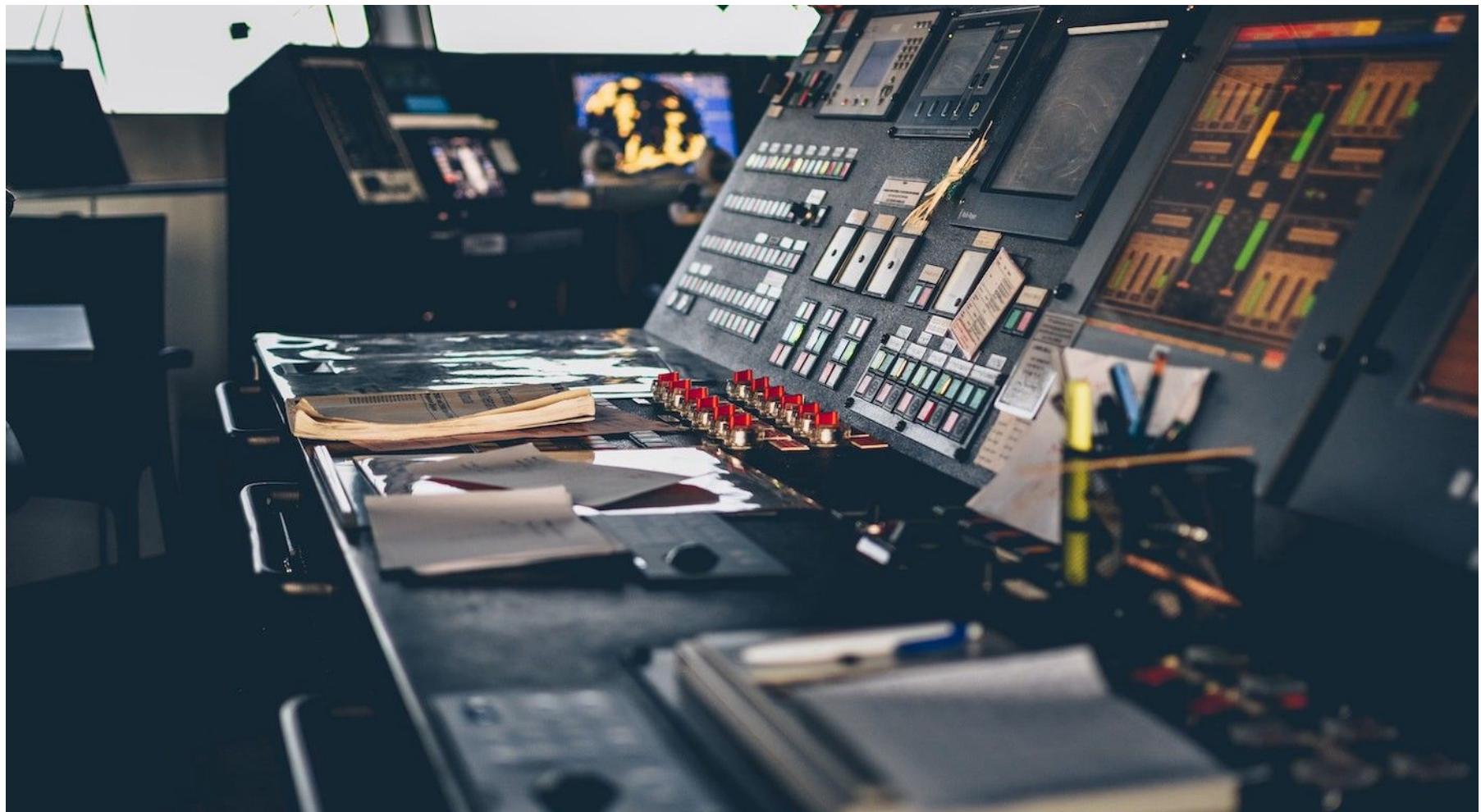
- Automate:**
- Configuration
  - Provisioning
  - Availability
  - Scaling
  - Security
  - Resource allocation
  - Load balancing
  - Health monitoring



Application Environment  
w/ Multiple Containers



# Relación con ingeniería de datos



# ¡Reto!

Imágenes de Docker

**Instrucciones:**  
Explora las imágenes públicas  
en Docker Hub

# Manejo de ambientes para datos

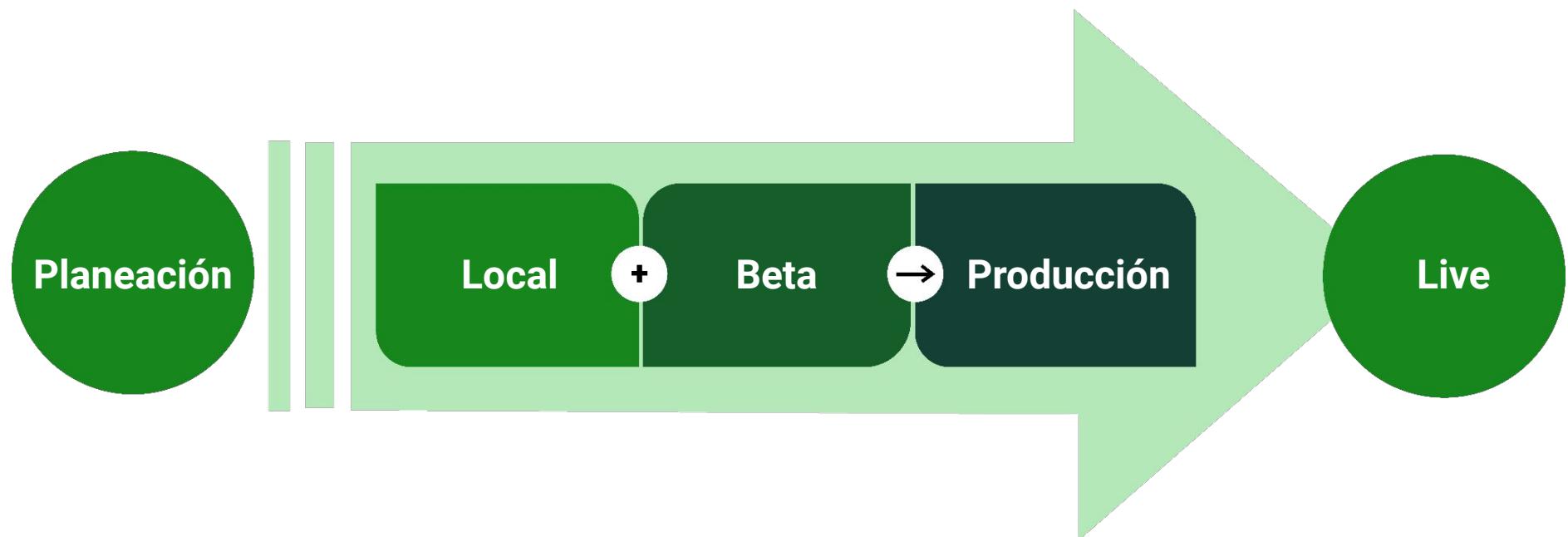
Build

# La maldición de producción



# **Contribución ordenada**

# Ambientes existentes



**Esto no está  
escrito en piedra**

# **¡Reto!**

Ventajas de ambientes

**Instrucciones:**  
Investiga ventajas y desventajas  
de utilizar ambientes.

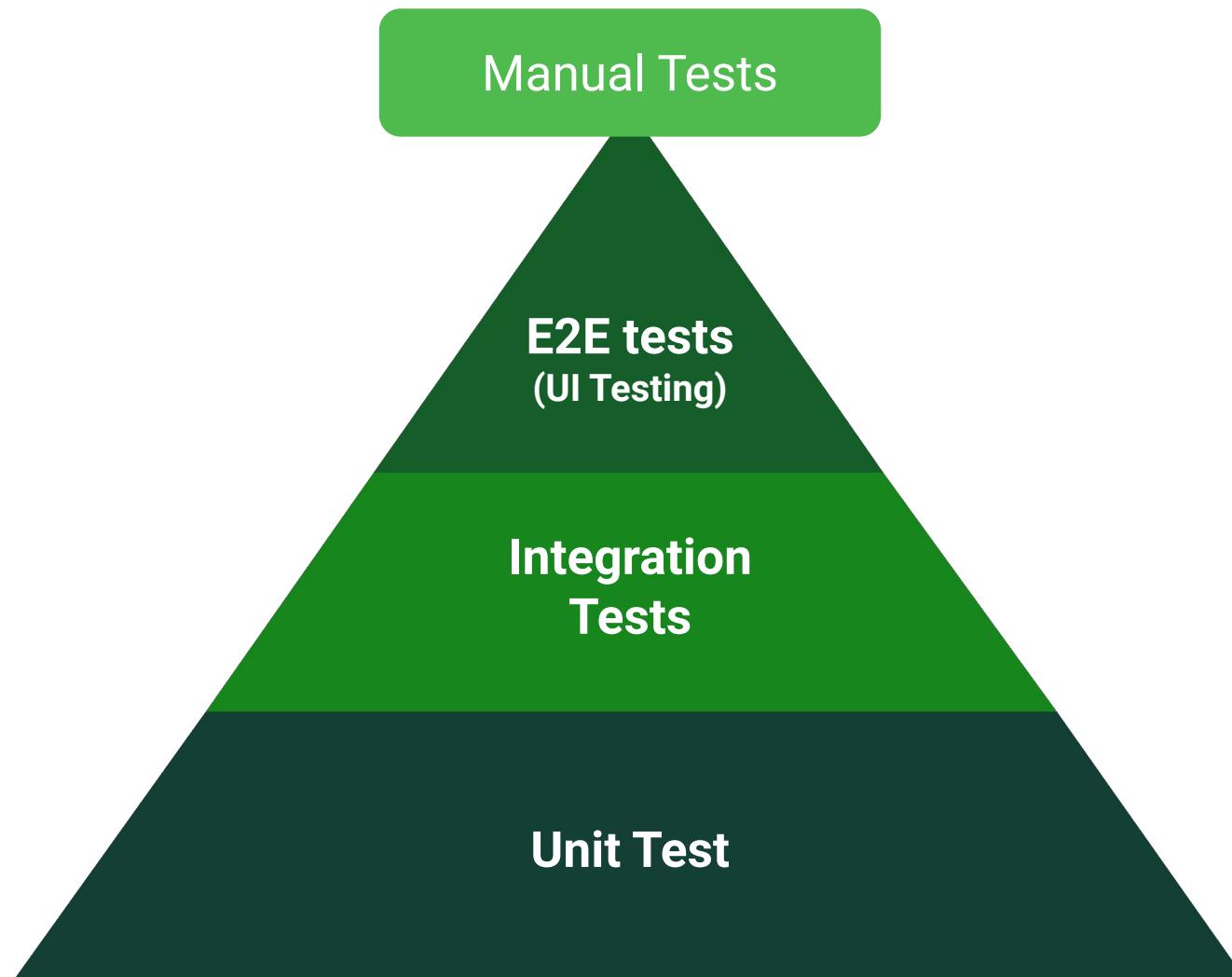
# Testing de software y datos

Build

# **El caso de uso de testing**

# Testing en ingeniería vs. data

# Testing de software



# **¡Reto!**

Librerías de Python para testing

**Instrucciones:**

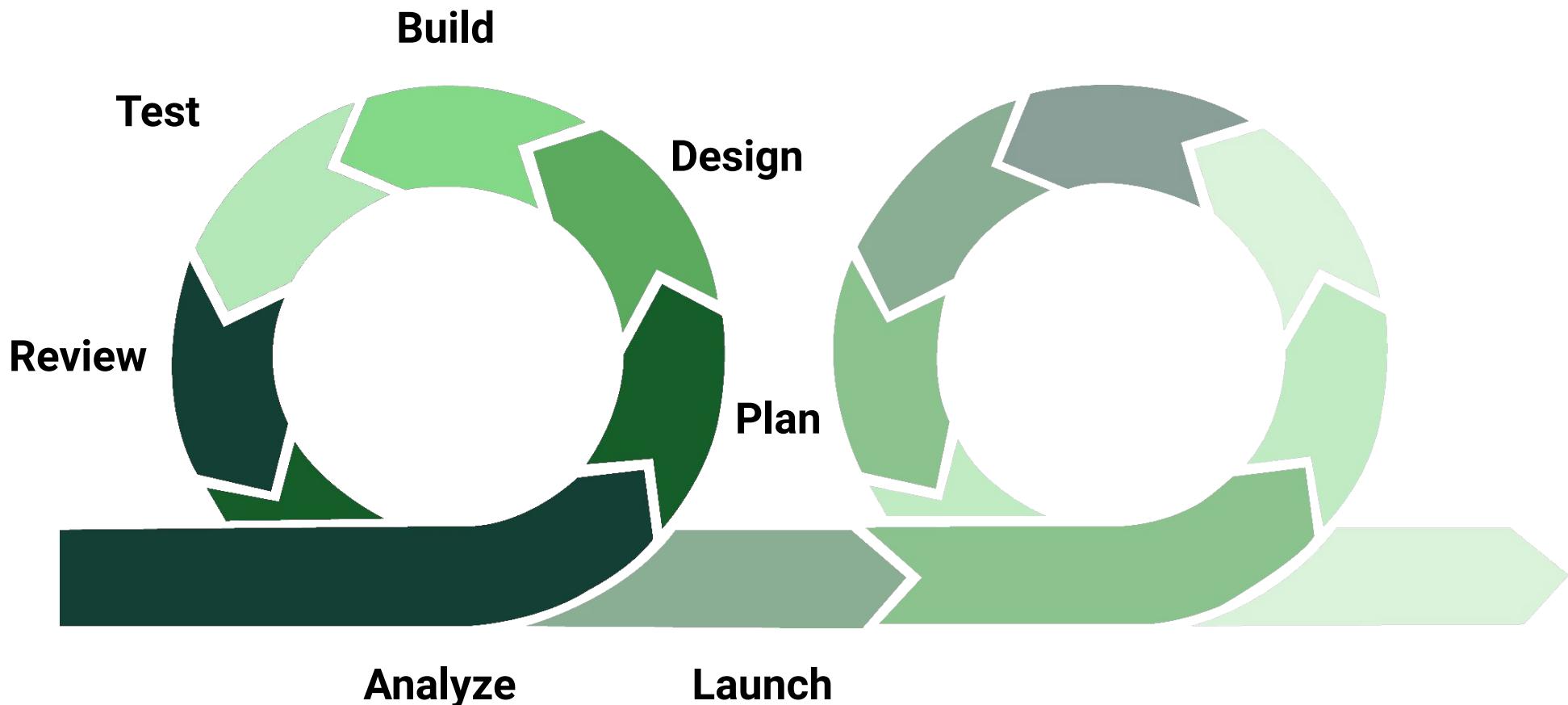
Busca en internet qué librerías existen para testing.

# CI/CD básico

## Release

# **Disclaimer: DevOPs**

# La necesidad de tener código actualizado



# Herramientas en el mercado



# **Flujos de datos con el manejo de ambientes**

# ¡Reto!

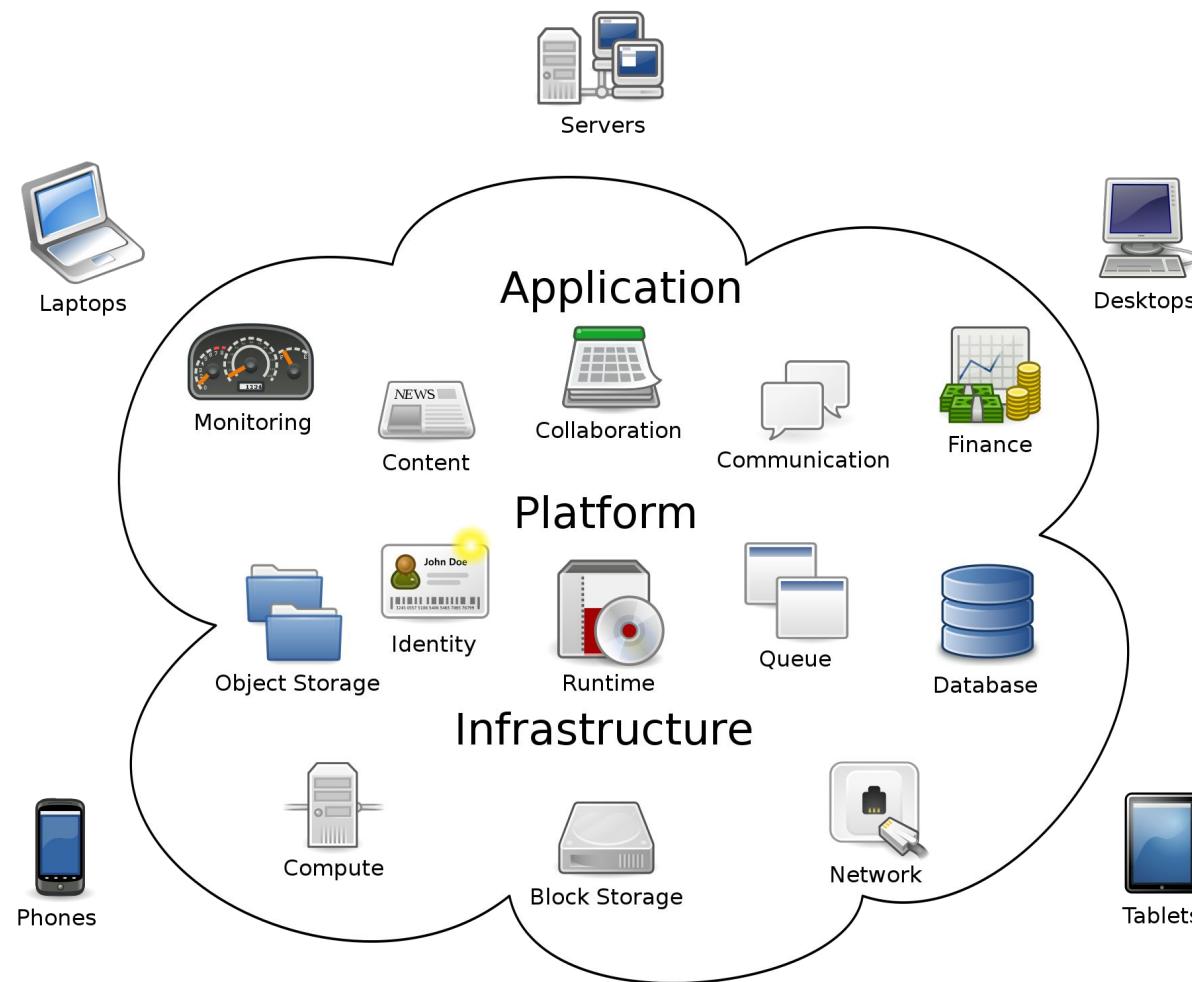
DataOps vs. DevOps

**Instrucciones:**  
Investiga el límite entre ambas.

# Servidores y computación en la nube para data

Release

# El mito de la nube



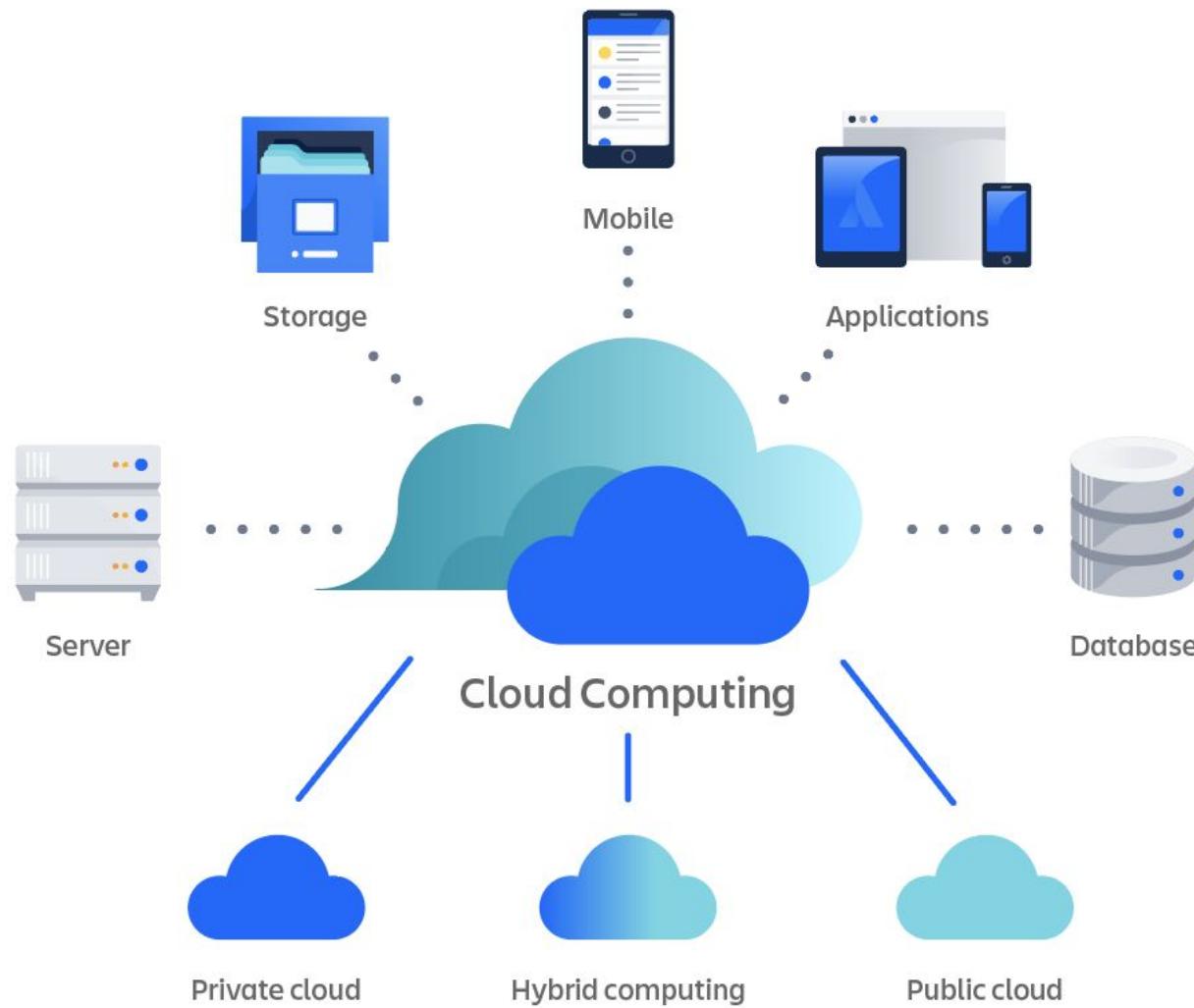
# Alternativas del mercado y sus similitudes



**Nube en casa:  
trabajando en premisas**

# Interacciones importantes

# Manteniendo recursos



# Manteniendo usuarios

# Manteniendo costos

# Distribuyendo carga

# Consideraciones de seguridad

# ¡Reto!

Data en cloud

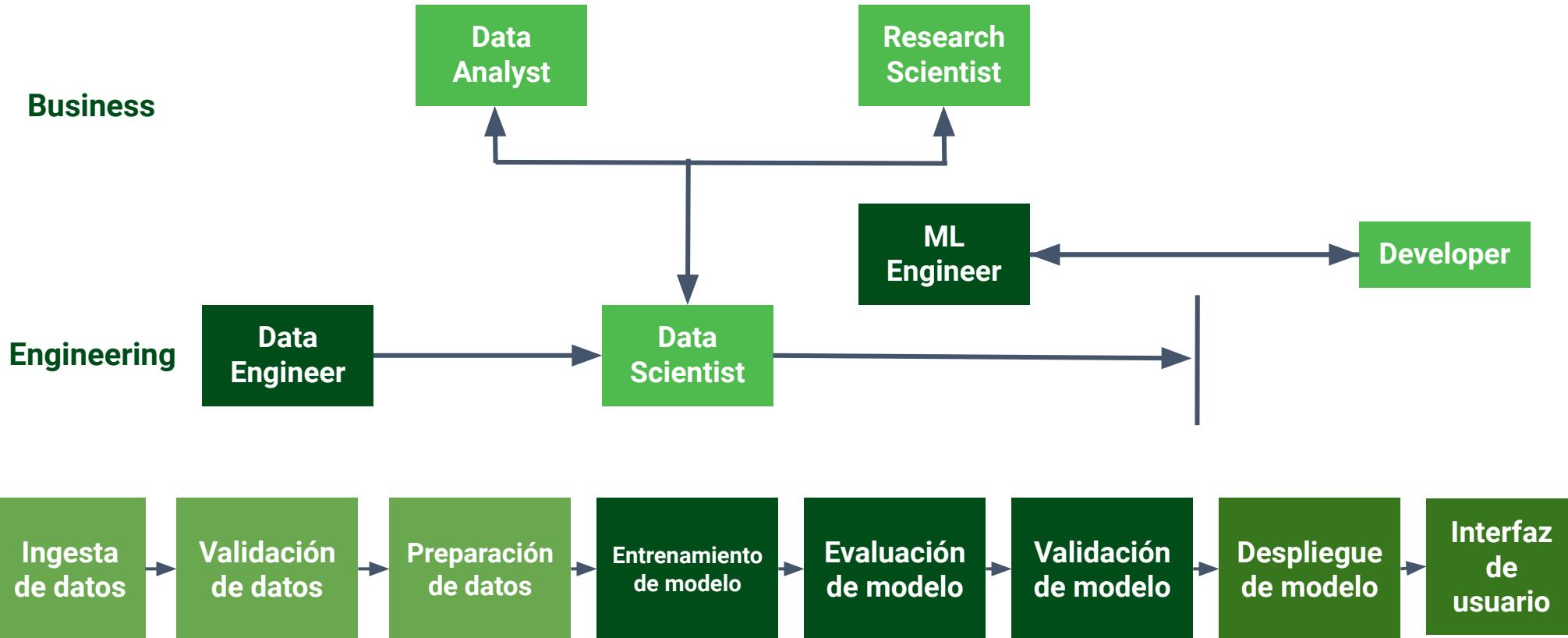
**Instrucciones:**

Investiga qué productos de data ofrece cada proveedor.

# Reentrenamiento y control de salud de servicios

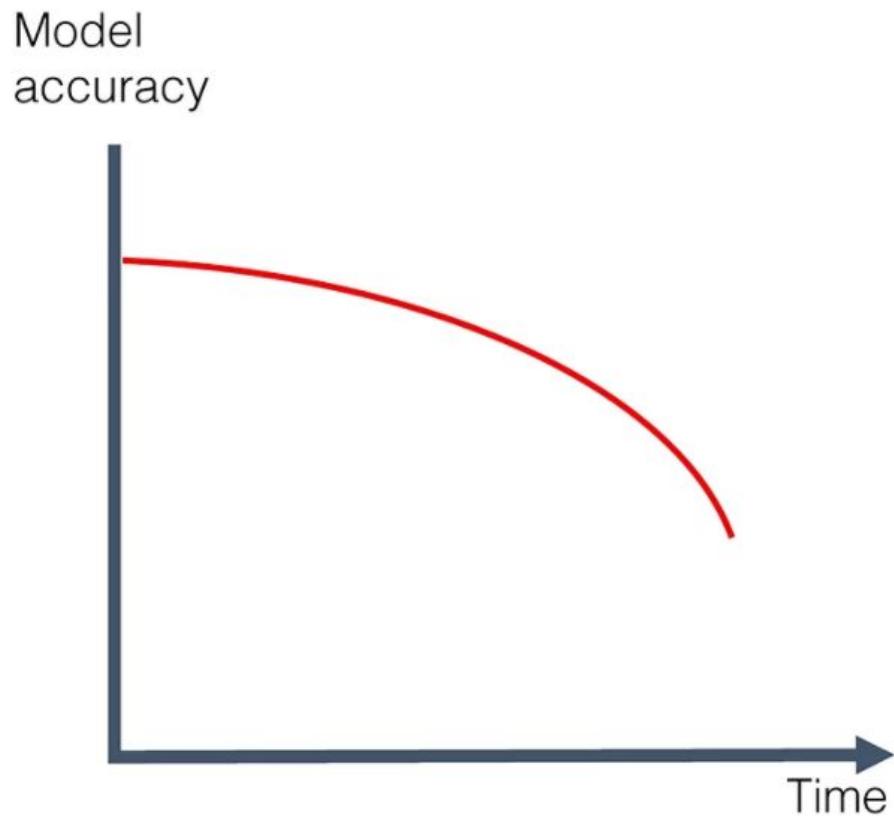
Operate

# Data Engineer vs. ML Engineer

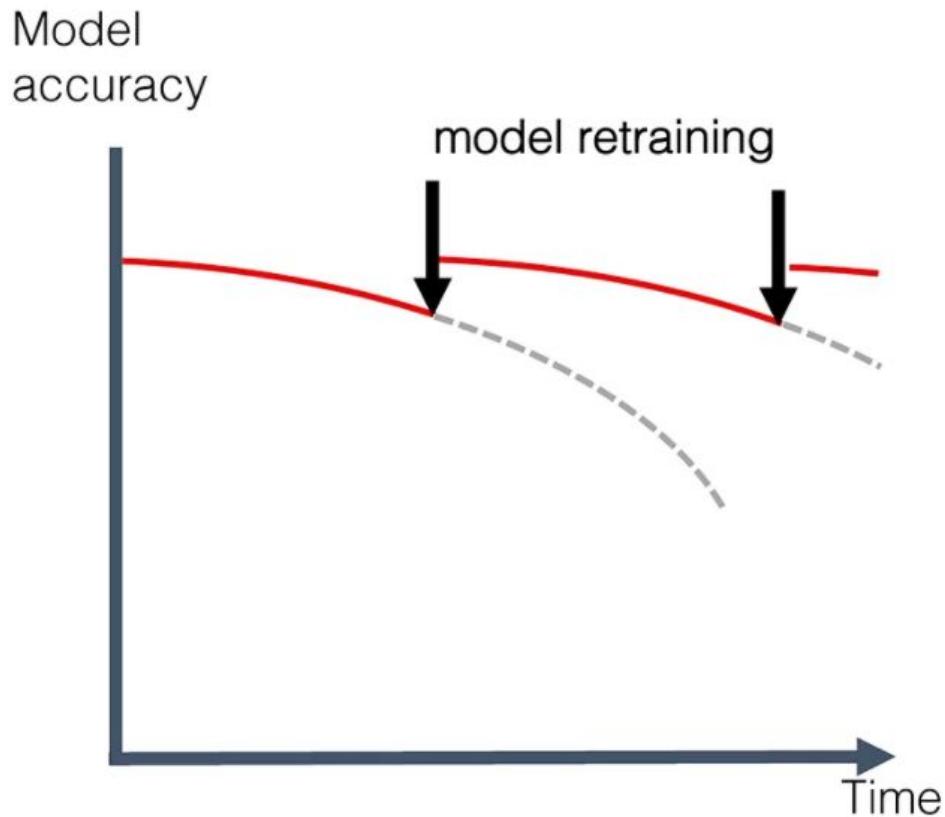


*Pasos y roles en el flujo de trabajo de data science  
(Design Patterns in Machine Learning).*

# Model drifting



Model decay over time



Regularly updated model

# Manteniendo un modelo en el tiempo

# Datos nuevos generados

# Entrega de nuevos modelos

# ¡Reto!

## ML Engineers

**Instrucciones:**

Busca en LinkedIn a una ML Engineer y analiza su trayectoria y habilidades.

# Medición de indicadores y seguimiento a proyectos

Monitor

# **El caso de no medir**

# Puntos de riesgo

# Contexto - API



# Puntos de riesgo

- A la entrada del servicio.
- A la entrada del modelo o punto de valor.
- Durante el proceso.
- Datos de consultas externas.
- A la salida del modelo.
- A la salida del servicio.
- Lo que se le mostró al cliente.

# Visibilidad

**Dashboards**

**Alertas y  
notificaciones**

# ¡Reto!

Prevención con monitoreo

**Instrucciones:**

Investiga desastres que han ocurrido por falta de monitoreo.

**Buscando  
oportunidades como  
data engineer**

**¡Busca y encuentra!**

**Atrévete a buscar  
tu trabajo ideal**

# Prepara tu LinkedIn



## Ricardo Alanis

Director of Data Science, Nowports

Monterrey, Nuevo León, Mexico · [Contact info](#)

500+ connections



Nowports



University of London

# Conecta física y virtualmente

Meetup

## 20 años de conexiones reales— Para **habilidades duraderas**

Tu nueva comunidad te espera. Desde hace 20 años, millones de personas han elegido Meetup para construir conexiones reales en torno a cosas que importan. Crea un grupo hoy.

[Crea un grupo](#)



# Conecta física y virtualmente

The screenshot shows the Platzi website interface. At the top, there's a navigation bar with the Platzi logo, a search bar asking "¿Quéquieres aprender?", a notification bell icon with a red dot, a user profile showing 38.262 pts, and a menu icon.

**LA COMUNIDAD DE PLATZI LLEGA A DISCORD**

**Crea conexiones profesionales con más estudiantes** [SEP]

**ÚNETE A DISCORD**

**Eventos en vivo**

**SEP 28** **12:00 PM** **Mouredev: La senda del freelance y sus secretos** **CLASE VIRTUAL**

**SEP 28** **01:00 PM** **Reto UX/UI mejora la experiencia de usuario en tu...** **CLASE VIRTUAL**

**SEP 28** **04:00 PM** **Utiliza el e-mail marketing para potenciar tus estrategias** **CLASE VIRTUAL**

Below the event cards, there are five small circular dots indicating more content is available.

**Pregunta, ¡no cuesta!**

# Job sites



# Learn English!

# ¡Reto!

## Networking

**Instrucciones:**

- Crea tu LinkedIn
- Participa en un evento

Evolución en el rol:  
ganando seniority como  
data engineer

**Disclaimer: puede variar**

# Primer día: aprende

# Primer mes: contribuye

# Primer año: mentoreo

**No tengas miedo  
de ir por más**

# **¡Reto!**

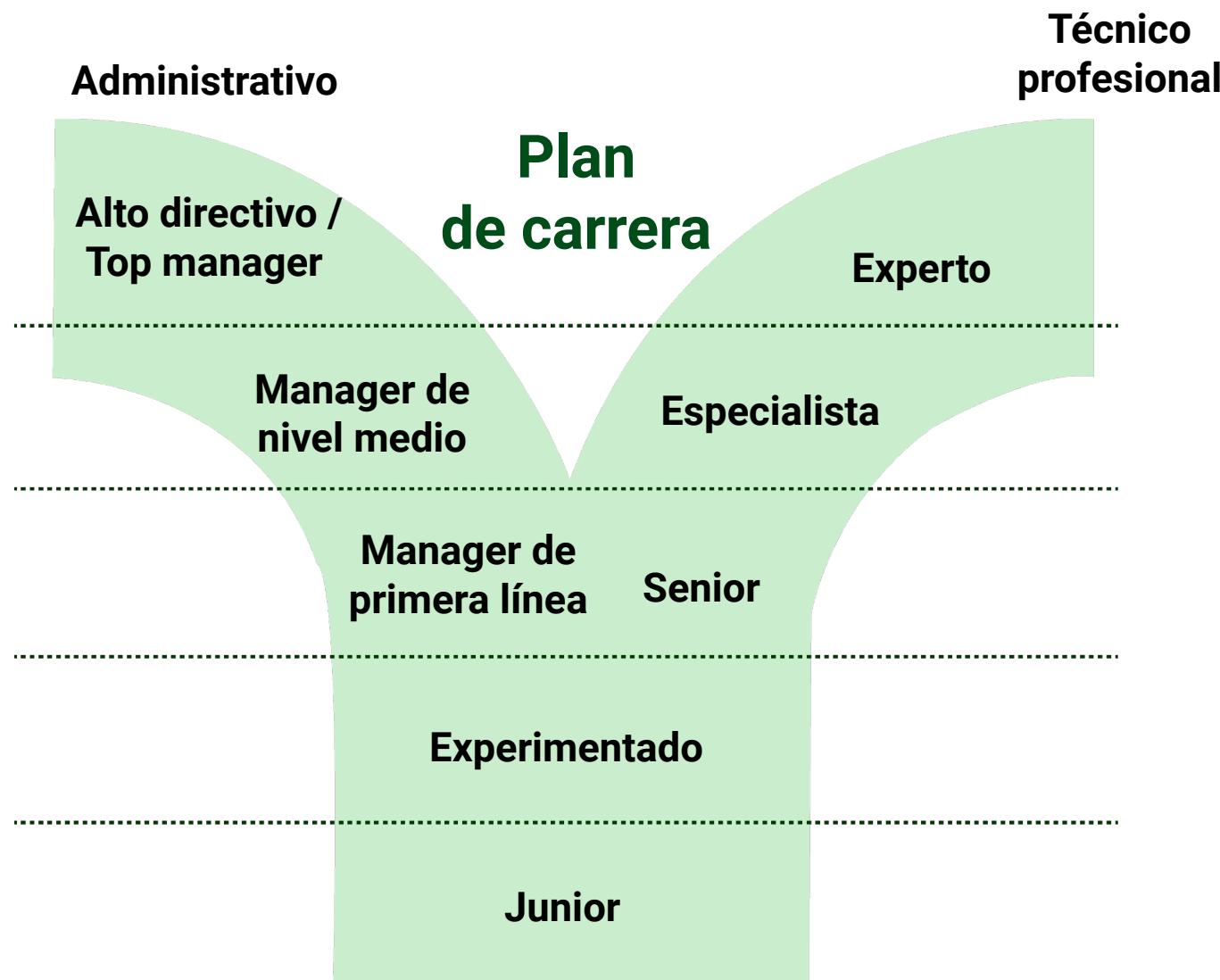
Imagínate ejerciendo

**Instrucciones:**

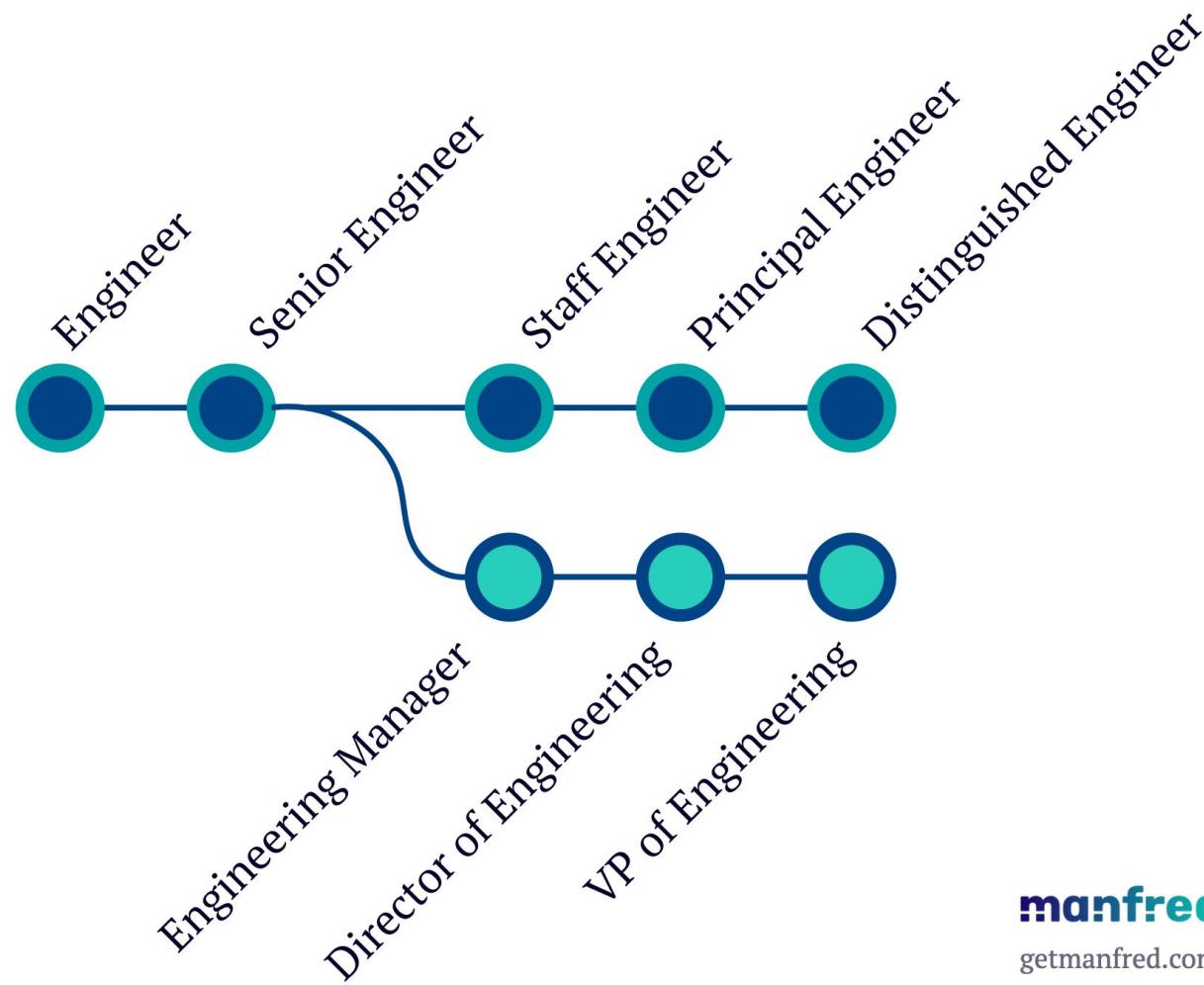
Identifica qué dudas podrías tener en tu primer día al ejercer.

# Evolución en el rol: manager, architect, pivot

# Dos caminos: técnico y administrativo

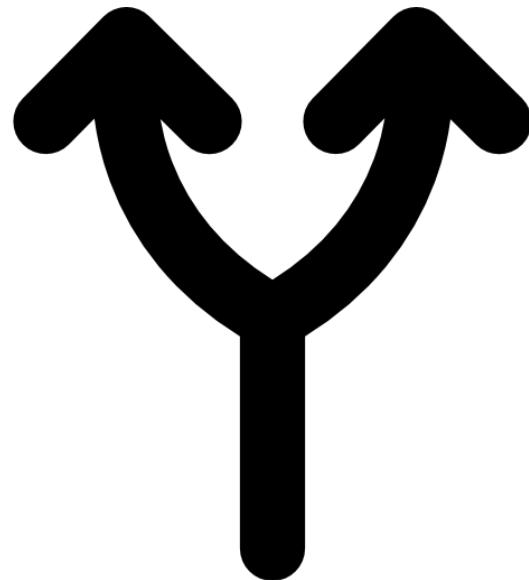


# El camino técnico vs. administrativo



# Tercer camino: cambiar

- Arquitectura
- Ciencia de datos
- DevOPs y cloud



**Cambia y evoluciona  
con confianza**

# ¡Reto!

Los caminos de evolución

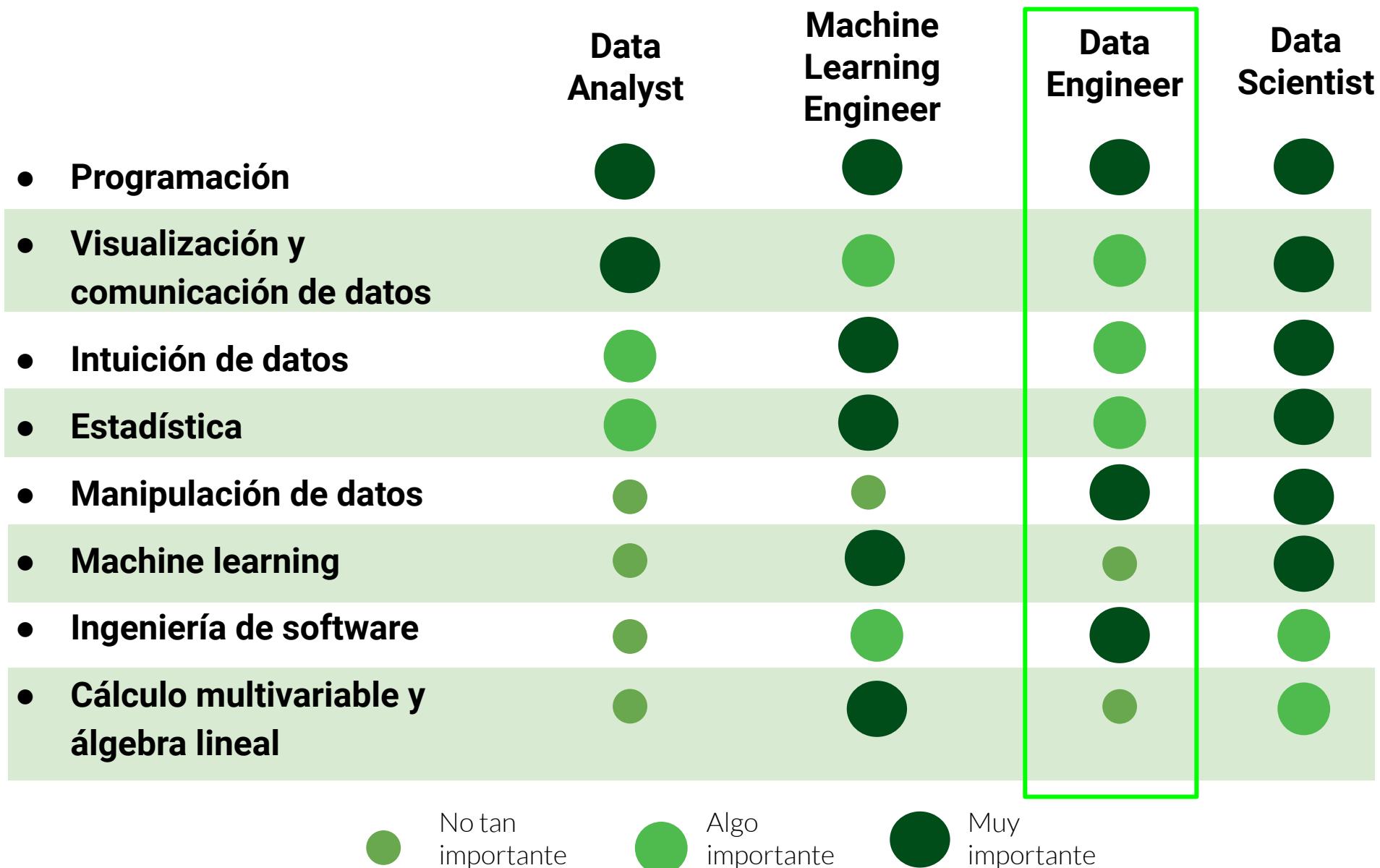
**Instrucciones:**

¿Qué camino te puede  
atraer más?

# Trabajando en equipo

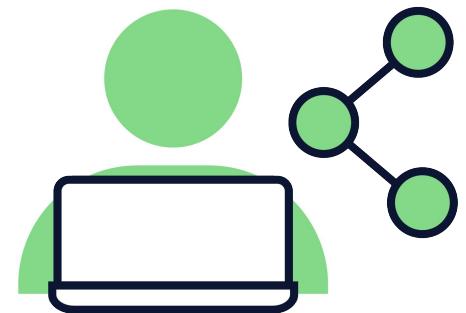
Contribuir como Data Engineer

# Recordando tu posición en el equipo



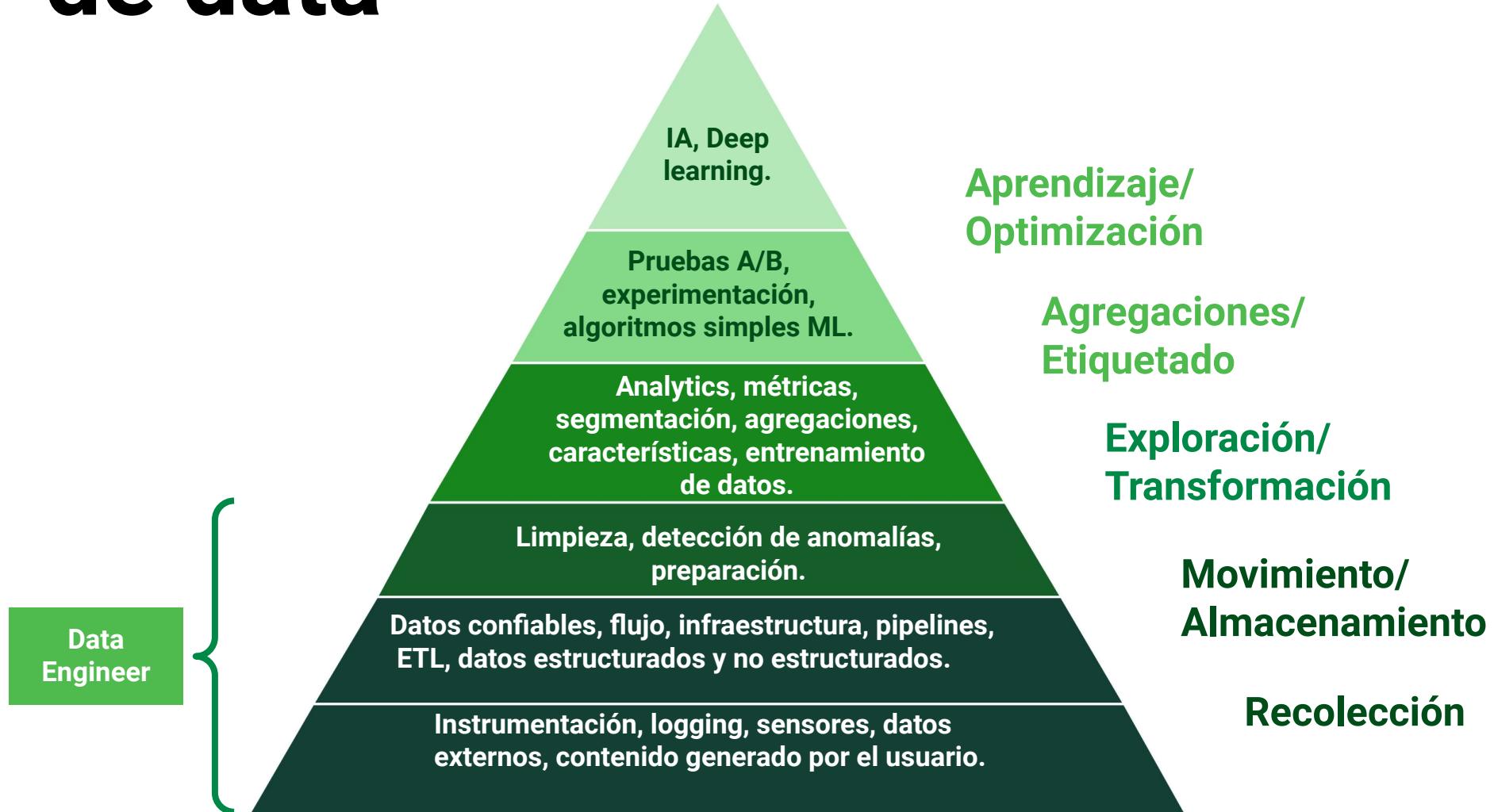
# Colaboración con personas

- Data scientist y analysts.
- Otras data engineers.
- Personas de producto.
- Equipo de desarrollo de aplicaciones.

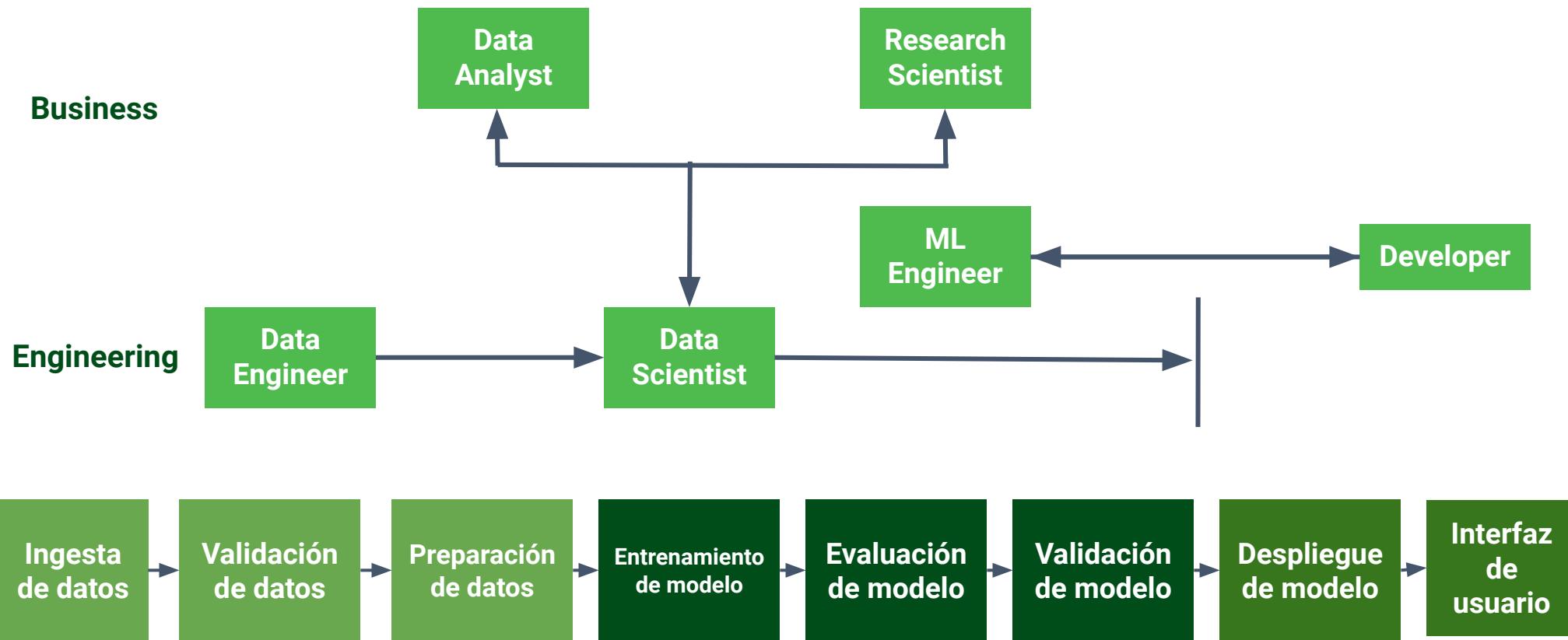


**La ingeniería de datos es muy importante.** 

# Jerarquía de necesidades de data



Referencia: 2. Data Science Hierarchy of needs (Monica Rogati – Hackernoon)



*Pasos y roles en el flujo de trabajo de data science  
(Design Patterns in Machine Learning).*

# ¡Reto!

Discord

**Instrucciones:**

Únete al Discord de Platzi  
y a sus canales de Data

**Compartir con la  
comunidad y seguir  
creciendo**

# Antes de ser data engineer

- Conoce la comunidad.
- Practica y crea proyectos.
- Comparte tus procesos y resultados.

# Ya siendo data engineer

- Ser parte activa de la comunidad.
- Exponer tus resultados.
- Mentorear a otras personas.
- Contribuir a los proyectos de código abierto.

# Conclusiones

# Más allá del mito



# Qué aprendiste

- Sobre el rol de la ingeniería de datos.
- Componentes del proceso de la ingeniería de datos.
- Tu evolución como data engineer.

# ¿Qué empezar a aprender?

- Python para data
- GitHub
- Terminal
- SQL





# Data Engineer

[platzi.com/data-engineer](https://platzi.com/data-engineer)



CERTIFICA A

por haber completado el programa de formación

## DATA ENGINEER

de la Escuela de Data Science e Inteligencia Artificial



A handwritten signature in black ink that reads 'Vander'.

CHRISTIAN VAN DER HENST S  
COO DE PLATZI

CERTIFICACIÓN DE APROBACIÓN ONLINE:

A handwritten signature in black ink that reads 'John Freddy Vega'.

JOHN FREDDY VEGA  
CEO DE PLATZI

# ¡Reto!

Inicia tu carrera como  
Data Engineer

## Instrucciones:

- Sigue la ruta de aprendizaje
- Imagínate qué te gustaría construir

**¡Nunca pares de  
aprender!**