

Maestría en Ciencia de Datos

Facultad de Ingeniería, Diseño y Ciencias

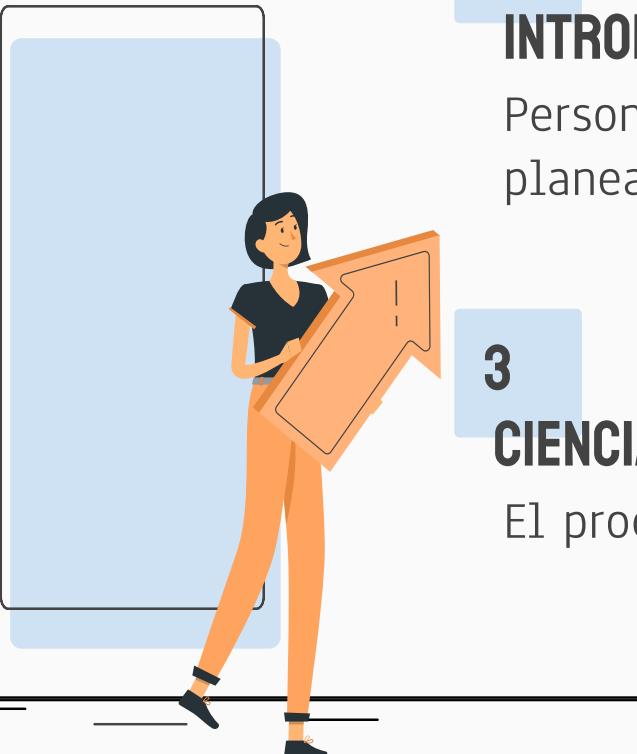


INFRAESTRUCTURA Y ARQUITECTURA DE

TI

Ángela Villota Gómez
apvillota@icesi.edu.co





I **INTRODUCCIÓN**

Personas, programa,
planeación, herramientas

3 **CIENCIAS DE DATOS**

El proceso

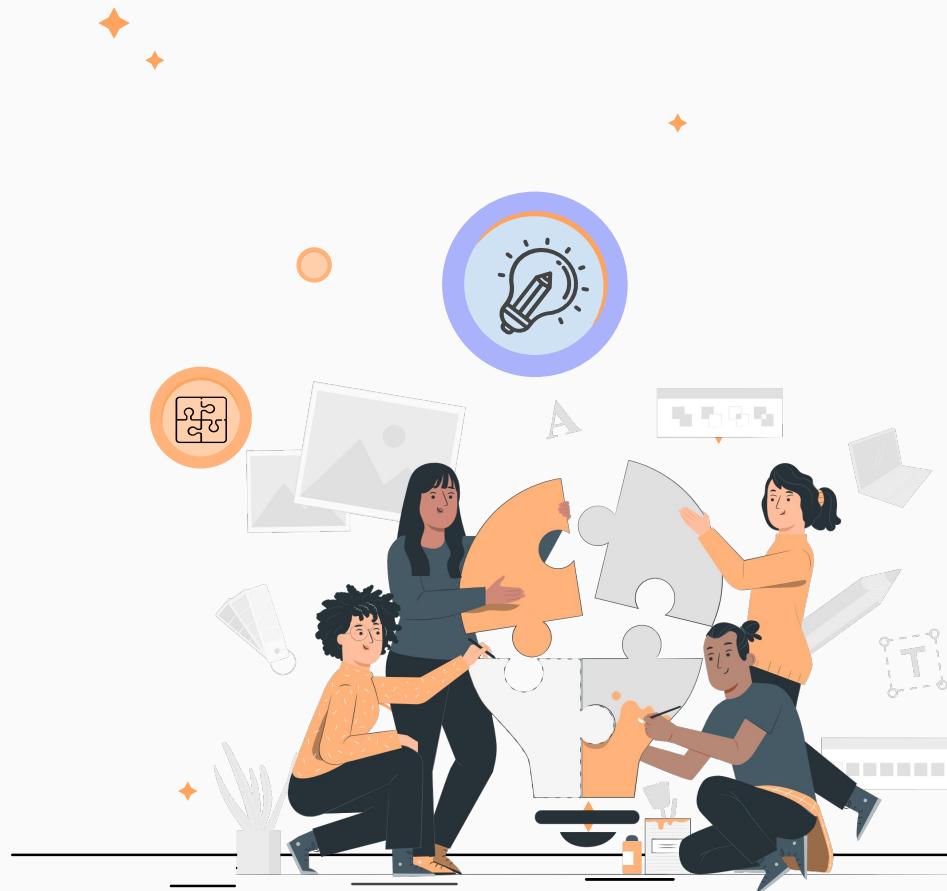
2 **CONTEXTO DEL CURSO**

En qué parte del proceso
de ML estamos?

4 **REVISIÓN DE CONCEPTOS**

Actividad

INTRODUCCIÓN



Ingeniera de Sistemas de la Universidad del Valle, PhD en Ciencias de la computación (2022).

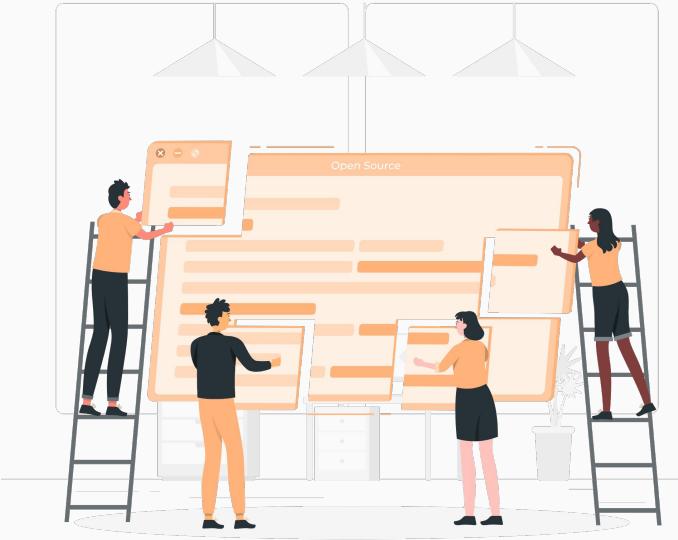
Soy una entusiasta de las competencias de programación, de la ingeniería de lenguajes de programación, la aplicación de métodos formales y las técnicas de solución de problemas por medio de algoritmos.

En lo personal, me caracterizo por ser una persona de buen humor y empática. En mi tiempo libre leo y practico Karate (Shotokan)

Angela Villota, apvillota@icesi.edu.co
Perfil de [LinkedIn](#)



PRESENTACIÓN DEL CURSO PROGRAMA



TIC - 60155-Infraestructura y Arquitectura de TI.

Créditos: 3

Intensidad semanal : 4 horas

Profesoras: Mónica Rojas, Ángela Villota

Capacitar al estudiante en los conceptos, técnicas y algunas herramientas para la gestión de los datos en un proyecto de ciencias de datos.

OBJETIVO GENERAL

OBJETIVOS TERMINALES



OT1

Distinguir los diferentes patrones de arquitecturas disponibles para implementar un proyecto de analítica de datos.



OT2

Diferenciar las características y tecnologías asociadas a los modelos de datos OLTP y OLAP con el fin de procesar los datos como insumo para el modelado y análisis.



OT3

Realizar procesos de ETL-ELT para procesar datos de diversas fuentes y dejarlos disponibles como insumo para las tareas de analítica de datos.



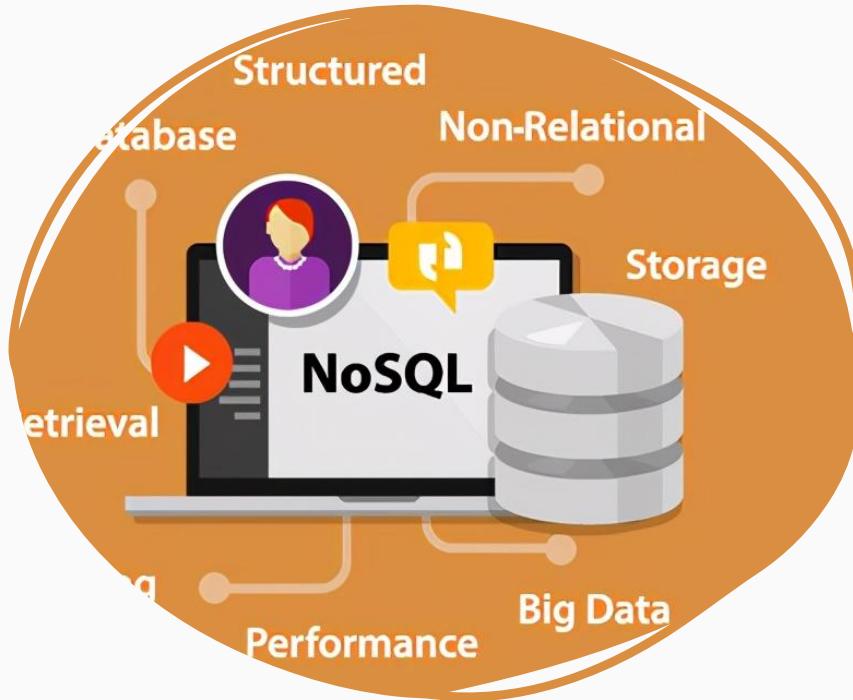
UNIDAD I – PATRONES DE ARQUITECTURA PARA LA GESTIÓN DE DATOS

-
- Explicar el ciclo de vida de los datos en un proyecto de Ciencia de Datos
 - Explicar los diferentes patrones de arquitecturas e infraestructuras disponibles para implementar un proyecto de analítica de datos.

UNIDAD 2 – ALMACENAMIENTO Y PROCESAMIENTO DE DATOS

- Explicar los diferentes tipos de procesamiento sobre los repositorios de datos (OLTP y OLAP) incluyendo las implicaciones tecnológicas y sus aplicaciones.
- Diferenciar las características de los modelos SQL y NoSQL y utilizar los lenguajes de consulta sobre dichos modelos.
- Extraer e interpretar la información de meta-modelado subyacente a un repositorio de datos.



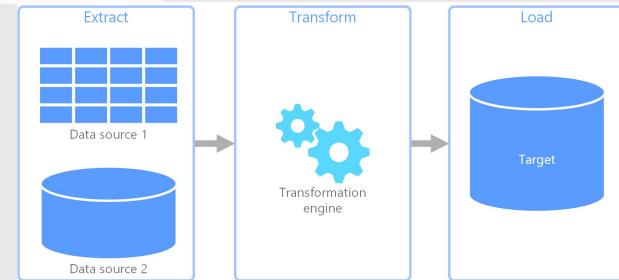


UNIDAD 3 – MODELOS NOSQL

- Explicar de forma general las características de los modelos NoSQL tales como orientados a documentos, a columnas, y a grafos entre otros.
- Escribir consultas que permitan hacer tareas de ETL sobre los modelos NoSQL.

UNIDAD 4 – PROCESOS ETL

- Aplicar las funciones y comandos del lenguaje SQL para construir consultas que permitan recuperar datos en una base de datos relacional.
- Construir procesos ETL utilizando datos de bases de datos relacionales y noSQL como fuente de procesos de analítica de datos.



Enlace a la planeación detallada del curso

PLANEACIÓN

Actividad	Porcentaje
Quices	10%
Evaluación - unidad 2	30%
Evaluación - unidad 3	30%
Evaluación - unidad 4	30%
Total	100%

EVALUACIÓN

Medios oficiales

- Intu
- correos
- miro
- carpeta del material enlazada en Intu

No-oficial

- Grupo wa??
- Discord

MATERIAL Y COMUNICACIÓN

SOFTWARE

Intu: material, el seguimiento y entrega de trabajos

MongoDB

Oracle

Pentaho Data Integration

Postgresql

LOS ESTUDIANTES

- MIRÓ

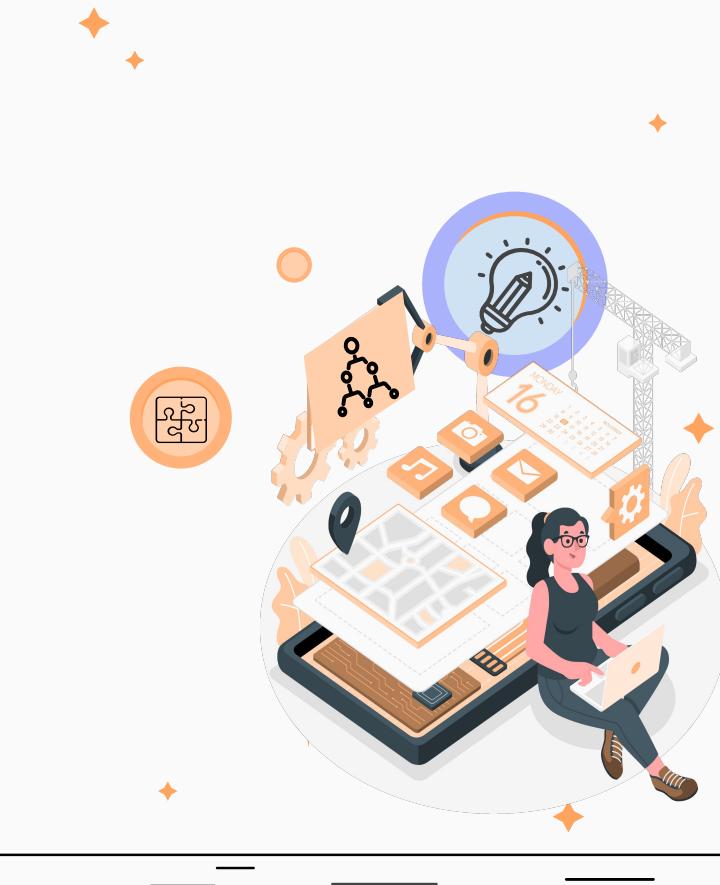


BIBLIOGRAFÍA

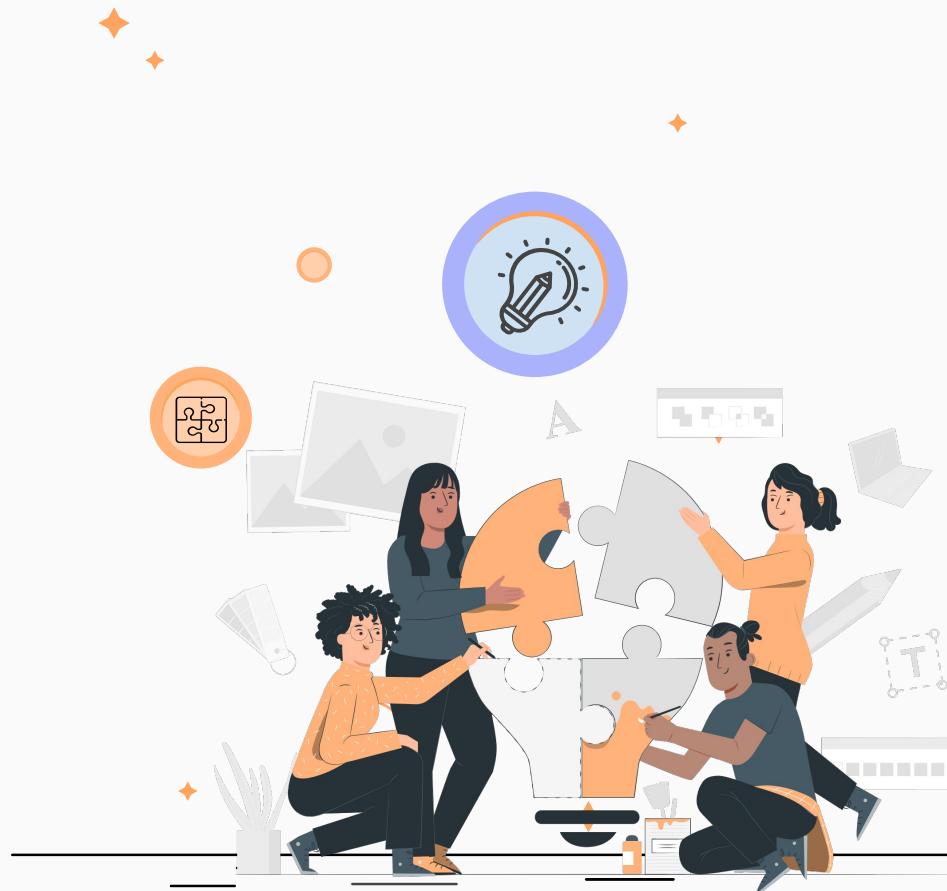
Algunos libros, enlaces y herramientas relevantes:

- Harrison, G. (2015). Next Generation Databases: NoSQL and Big Data. Apress.
- Carpenter, J., & Hewitt, E. (2020). Cassandra: the definitive guide: distributed data at web scale. O'Reilly Media.
- Sadalage, P. J., & Fowler, M. (2013). NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Pearson Education.
- V. Kale, Parallel Computing Architectures and APIs IoT Big Data Stream Processing, Taylor & Francis Group, 2019, p. 380.
- T. Soyata, GPU Parallel Program Development Using CUDA, Taylor & Francis Group, 2018, p. 440.
- Thomas Connolly, Carolyn Begg. Database Systems a practical approach to design, implementation and management. Addison Wesley, 2009
- Toby Teorey. Database modeling and design: logical design (5th ed). Morgan Kaufmann, 2011.
- Cielen, D., Meysman, A., & Ali, M. (2016). Introducing Data Science: Big Data. Machine Learning and More, Using Python Tools. Manning, Shelter Island, US, 322

PROCESO DE CIENCIA DE DATOS



INTRODUCCIÓN



OBJETIVOS TERMINALES



OT1

Distinguir los diferentes patrones de arquitecturas disponibles para implementar un proyecto de analítica de datos.



OT2

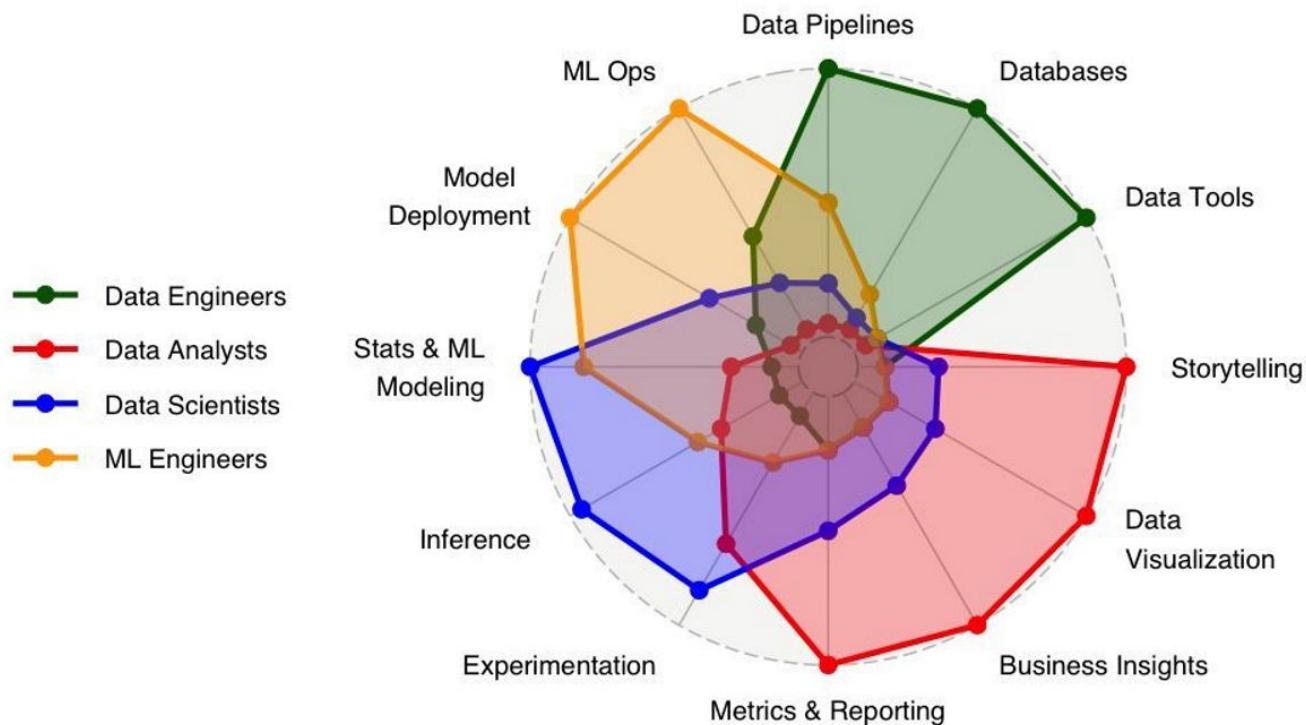
Diferenciar las características y tecnologías asociadas a los modelos de datos OLTP y OLAP con el fin de procesar los datos como insumo para el modelado y análisis.



OT3

Realizar procesos de ETL-ELT para procesar datos de diversas fuentes y dejarlos disponibles como insumo para las tareas de analítica de datos.

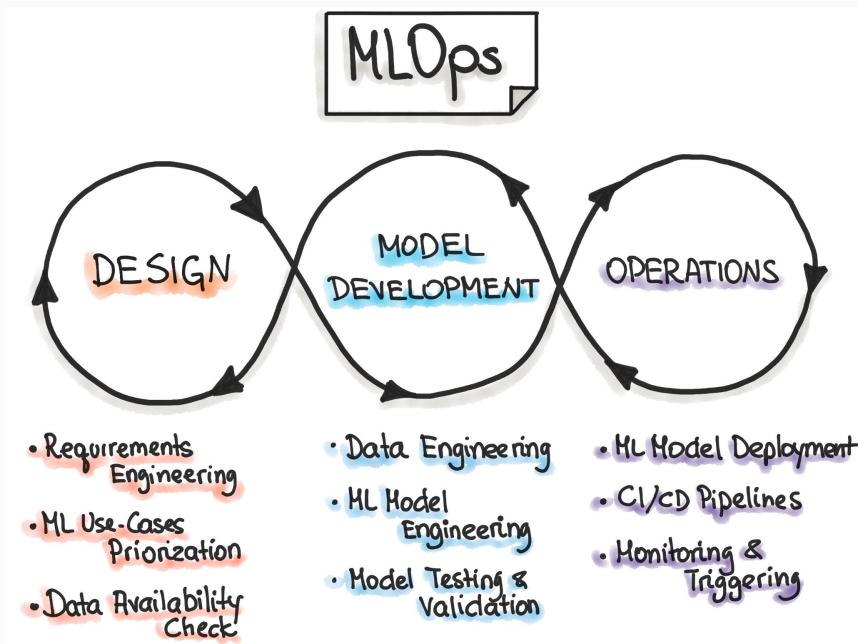
THE CONTEXT OF DATA SCIENTISTS



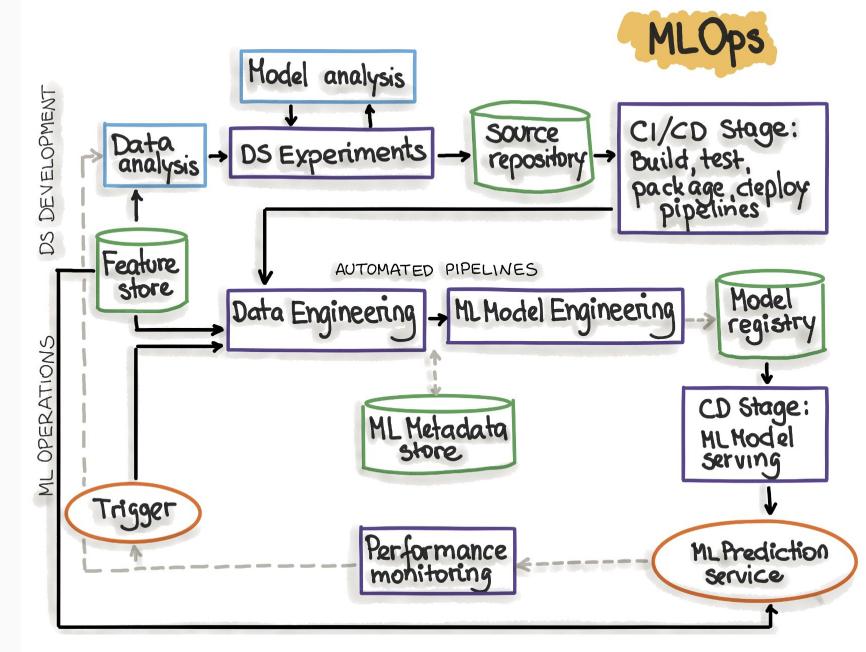
Source:

- <https://newsletter.theaiedge.io/>

MODELOS DE PROCESO



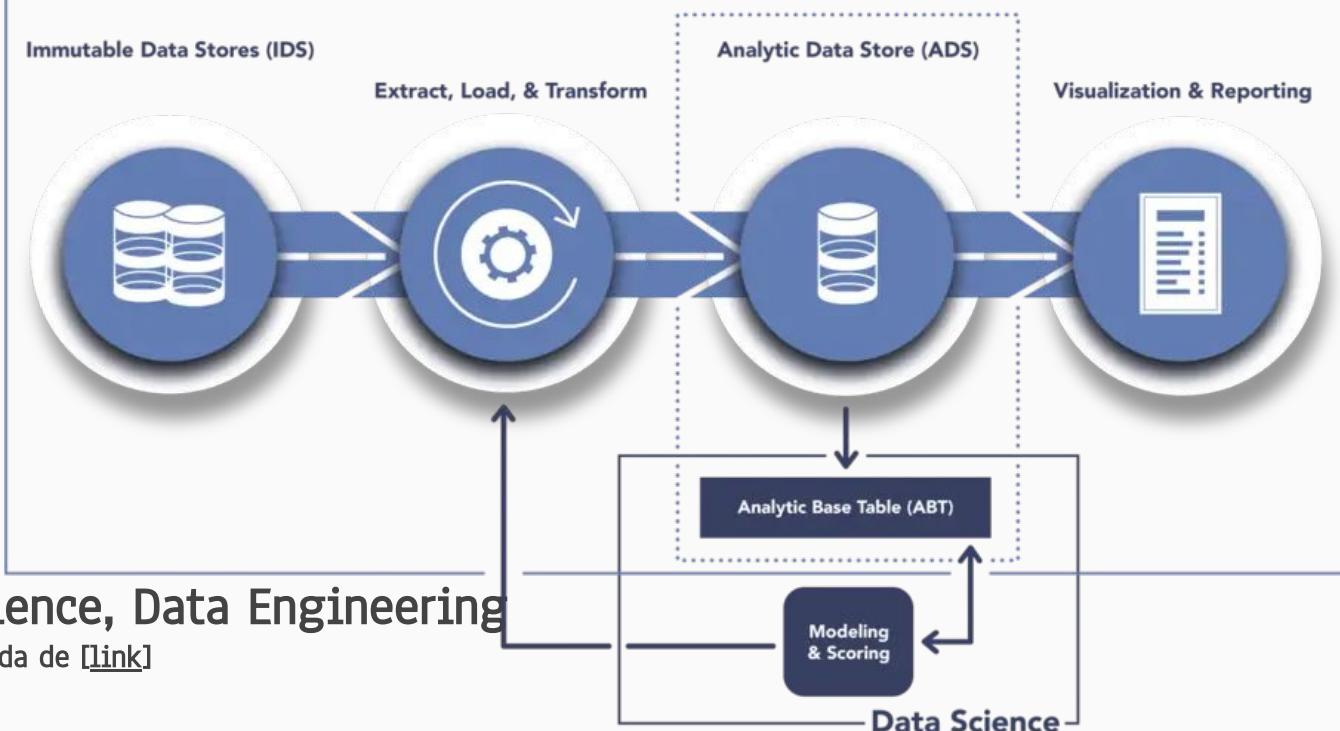
Machine Learning Model Operationalization Management (MLOps)
Imagen tomada de [\[link\]](#)



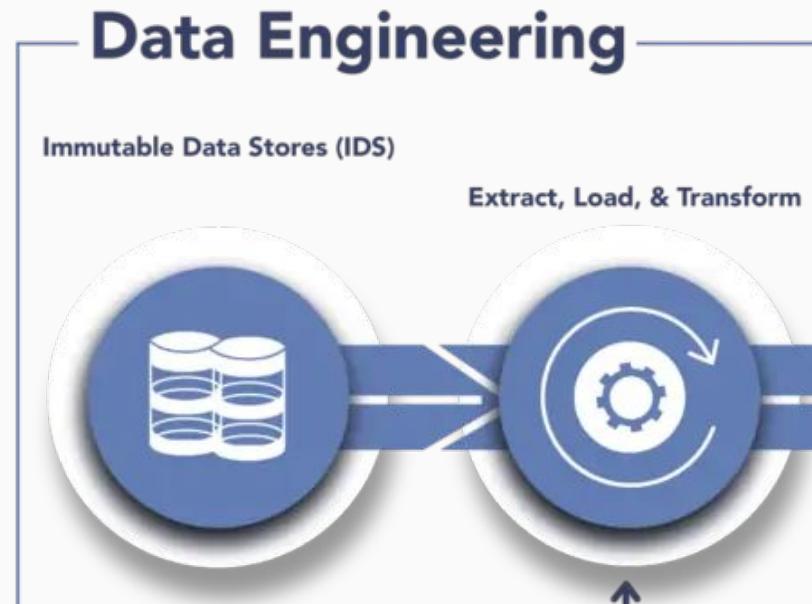
MLOps - automated ML pipeline
Imagen tomada de [\[link\]](#)

CONTEXTO

Data Engineering

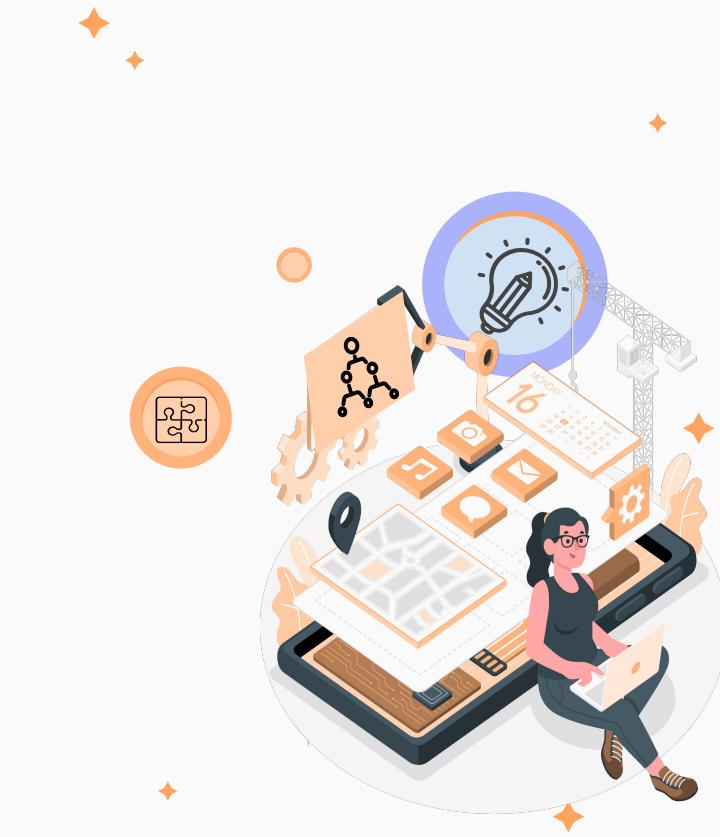


CONTEXTO



Almacenamiento de datos, carga y transformación.

PROCESO DE CIENCIA DE DATOS



TIPOS DE DATOS

- Estructurado
- No estructurado
- Lenguaje natural
- Generado por máquina
- Basado en grafos o redes
- Audio, video e imágenes
- Streaming

[1]



DATOS ESTRUCTURADOS

Son dependientes del modelo de datos y manejan campos (atributos) fijos.

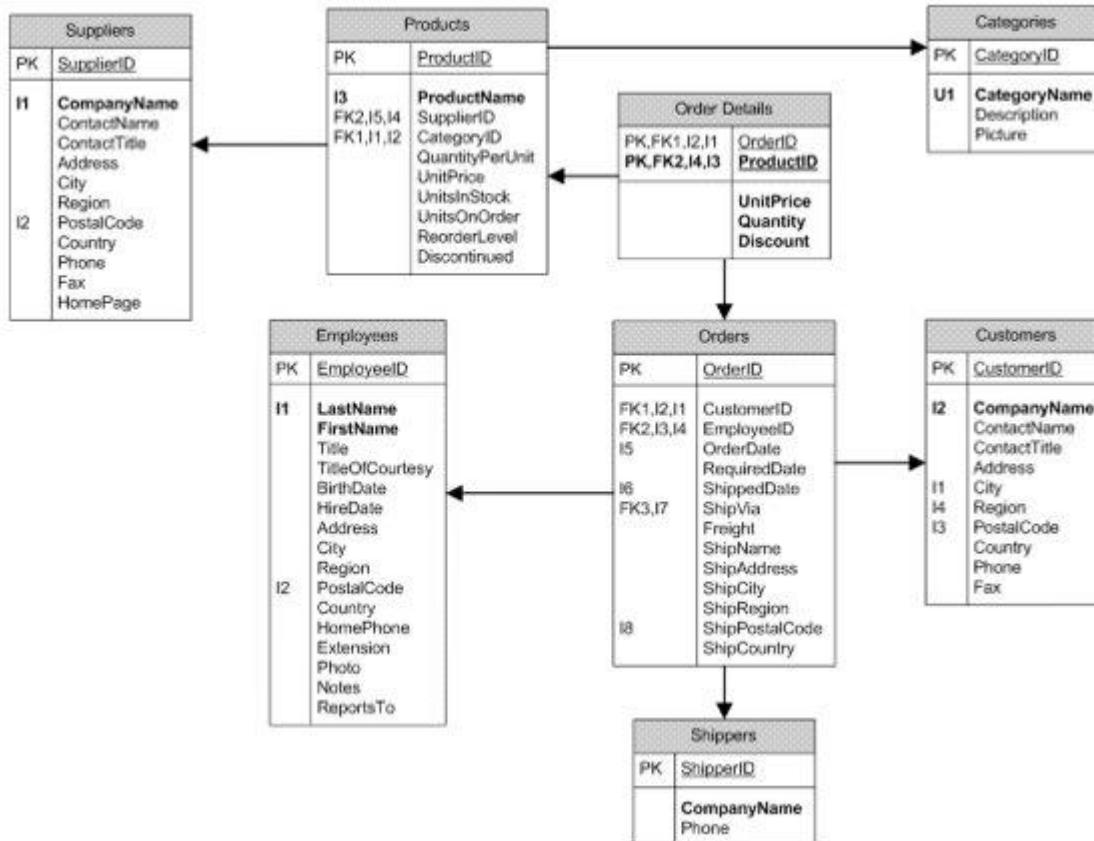
Ejemplos:

BD Relacionales

Archivos

Hojas de cálculo

Bases de datos tradicionales provenientes de CRM, ERP, etc.



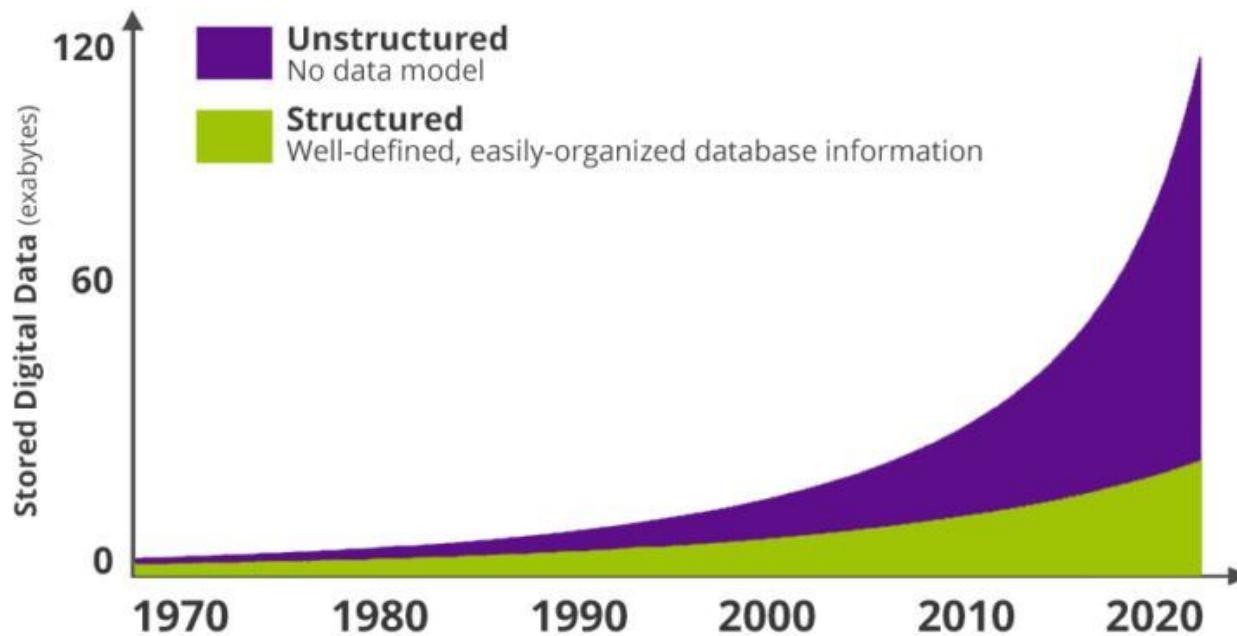
PRODUCTS

ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitPrice
1	Chai	1	1	10 boxes x 20 bags	18
2	Chang	1	1	124 - 12 oz bottles	19
3	Aniseed Syrup	1	2	12 - 550 ml bottles	10
4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22
5	Chef Anton's Gumbo Mix	2	2	36 boxes	21.35
6	Grandma's Boysenberry Spread	3	2	12 - 8 oz jars	25
7	Uncle Bob's Organic Dried Pears	3	7	12 - 1 lb pkgs.	30
8	Northwoods Cranberry Sauce	3	2	12 - 12 oz jars	40
9	Mishi Kobe Niku	4	6	18 - 500 g pkgs.	97
10	Ikura	4	8	12 - 200 ml jars	31
8	Northwoods Cranberry Sauce	3	2	12 - 12 oz jars	40
9	Mishi Kobe Niku	4	6	18 - 500 g pkgs.	97
10	Ikura	4	8	12 - 200 ml jars	31

CATEGORIES

CategoryID	CategoryName	Description	Picture
1	Beverages	Soft drinks, coffees, tea...	43 b... beverages.gif
2	Condiments	Sweet and savory sauces, ...	58 b... condiments.gif
3	Confections	Desserts, candies, and sw...	35 b... confecti...nctions.gif
4	Dairy Products	Cheeses	7 b... diary.gif
5	Grains/Cereals	Breads, crackers, pasta, ...	35 b... cereals.gif
6	Meat/Poultry	Prepared meats	14 b... meat.gif
7	Produce	Dried fruit and bean curd	25 b... produce.gif
8	Seafood	Seaweed and fish	16 b... seafood.gif
(NULL)	(NULL)	(NULL)	0 Kb...

DATOS NO ESTRUCTURADOS



Tomado de: https://www.komprise.com/glossary_terms/unstructured-data/

DATOS NO ESTRUCTURADOS

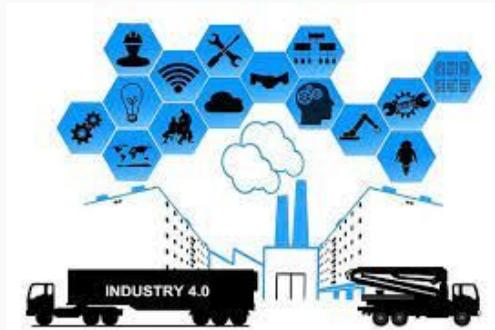
Unstructured data types

			
Text files and documents	Server, website and application logs	Sensor data	Images
			
Video files	Audio files	Emails	Social media data

Structured vs. Unstructured Data: What You Need To Know [\[link\]](#)

GENERADOS POR MÁQUINAS

Datos que son creados por un computador, un proceso, aplicación u otra máquina.



EJEMPLOS

Internet de las cosas
Sensores
Lecturas RFID
Variables meteorológicas



LENGUAJE NATURAL

Datos que son generados por personas.

EJEMPLOS

Correos electrónicos
Búsquedas
Traducciones
Análisis de textos
Notas de texto, audio
Registros médicos



DATOS - EN FORMA DE GRAFOS

Datos en los que dada su naturaleza, se muestran como relaciones entre pares de objetos.

EJEMPLOS

Redes sociales
Intereses



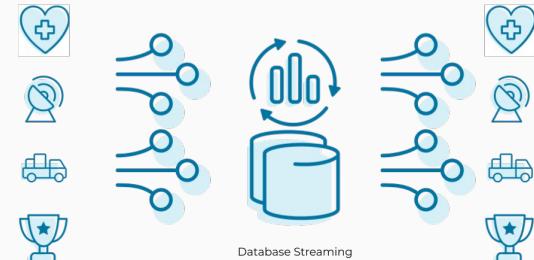
STREAMING

Datos que se generan continuamente por miles de orígenes de datos, que normalmente envían los registros de datos simultáneamente, y en tamaños pequeños.

Fluyen hacia el sistema cuando ocurre un evento.

EJEMPLOS

- Lecturas de sensores
- Log de eventos
- Actividades en un juego
- Clicks



Event-driven use cases

Database Streaming

Real-time applications

THE DATA SCIENCE PROCESS

The six steps of the
data science process

-  Setting the research Goal
-  Retrieving Data
-  Data preparation
-  Data exploration
-  Data modeling
-  Presentation and automation

THE DATA SCIENCE PROCESS

The six steps of the
data science process

 Setting the research Goal

 Retrieving Data

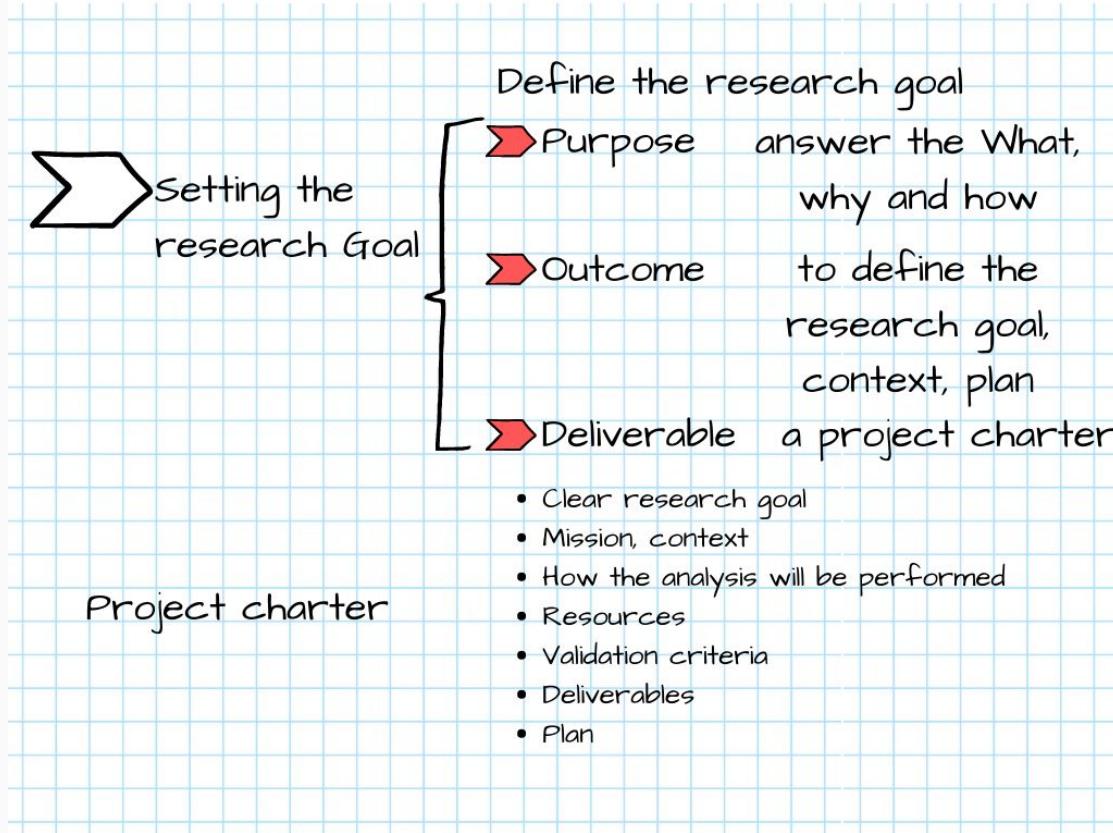
 Data preparation

 Data exploration

 Data modeling

 Presentation and automation

THE DATA SCIENCE PROCESS, STEP 1



Other sources

- The ML Canvas web
[[link](#)]
- The ML Canvas book

THE DATA SCIENCE PROCESS, STEP 2

➤ Retrieving Data

Internal Data, data stored in the company
➤ Databases
Data marts
Data warehouses
Data lakes → Raw format

External Data
➤ Companies collecting valuable info / serv
Provide data / return
Public data

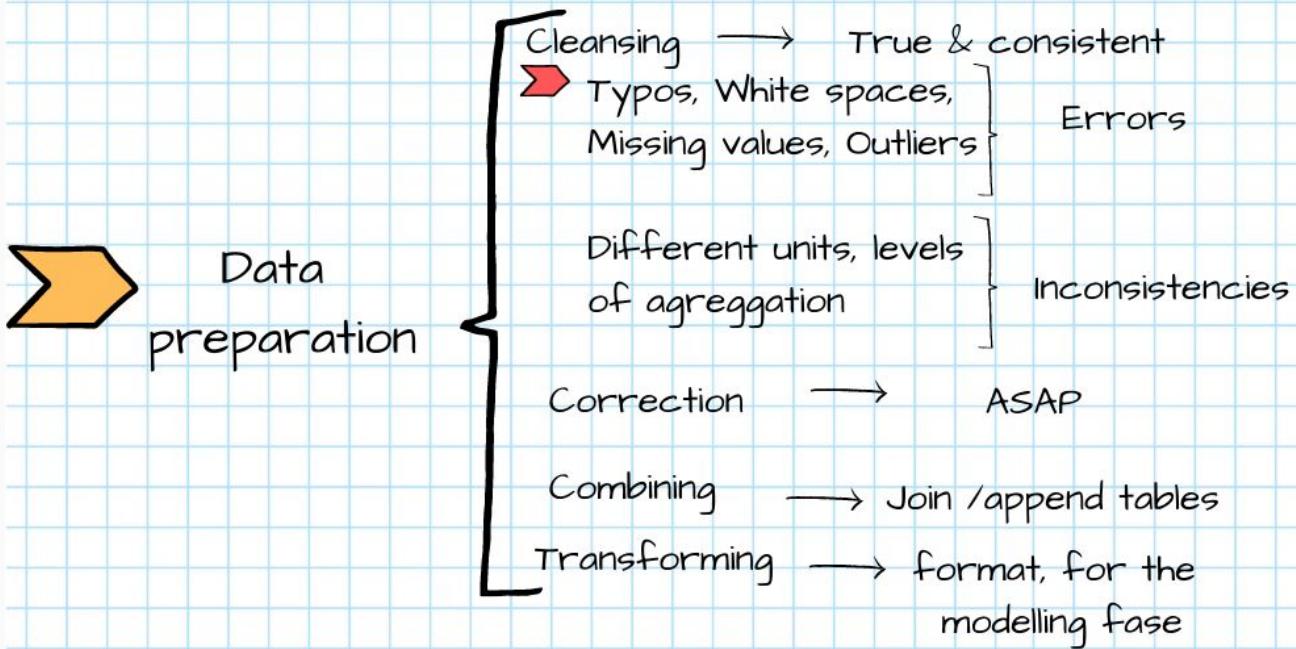
➤ Quality check

Data import
Data preparation → Content of variables
Exploratory analysis → Statistical properties

import correctly
check datatypes

THE DATA SCIENCE PROCESS, STEP 3

Models need the data in a specific format, so data transformation will always come into play



EJERCICIO

En la carpeta compartida del curso encontrará la carpeta datos que contiene 3 archivos: data.txt, data1.txt y data3.txt

1. Construya un solo archivo consolidando los datos que provienen de las 3 fuentes

Tenga en cuenta:

→ Los datos pueden tener alguno de los siguientes errores:

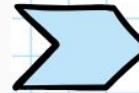
- Errores de digitación
- Espacios en blanco redundantes
- Valores imposibles
- Outliers
- Valores faltantes
- Diferentes unidades de medida
- Diferentes niveles de agregación

→ Es necesario combinar los datos usando un joining y appending

2. Proponga dos o más columnas dummies a partir de los datos obtenidos en el punto anterior.
3. Si tuviera que ejecutar las tareas anteriores con la ayuda de una herramienta, ¿qué herramienta usaría y por qué?

THE DATA SCIENCE PROCESS, STEP 4

A picture is worth a thousand words



Data
Exploration

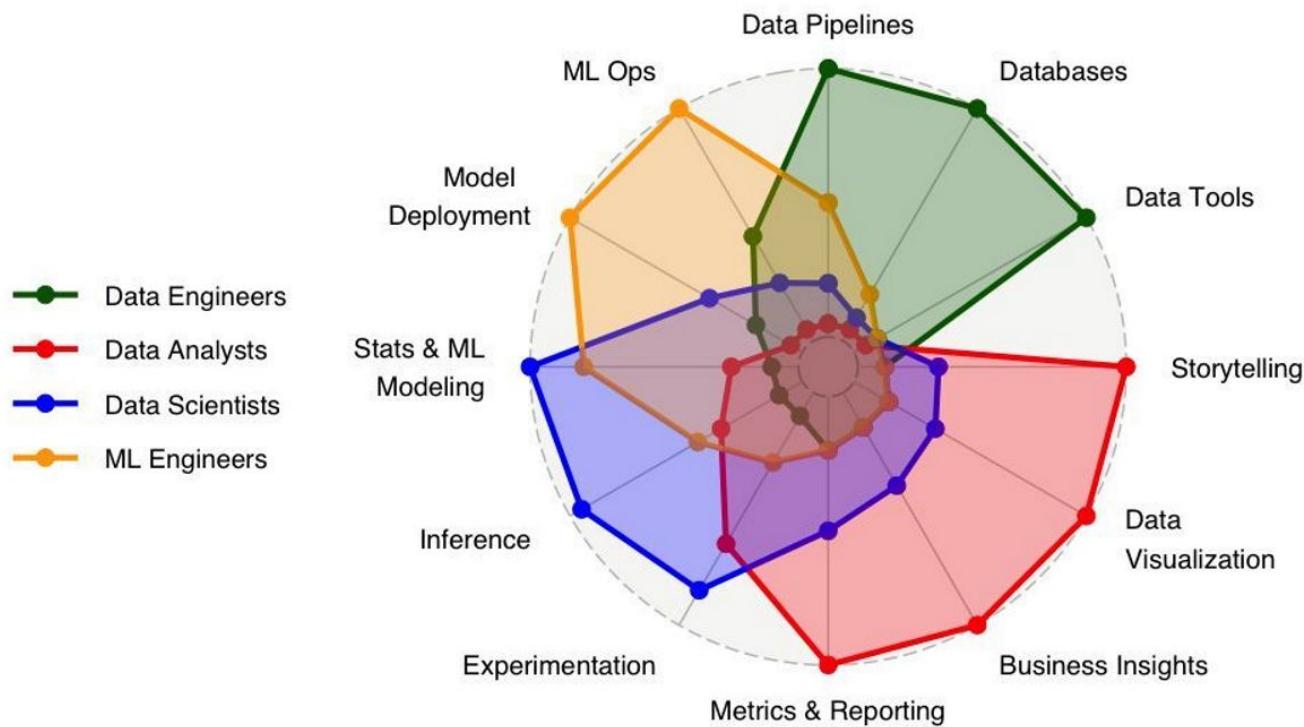
Goal → use graphical techniques to gain an understanding of your data and the interactions between variables

► Techniques

- Simple graphs
- Combined graphs
- etc..

In practice techniques such as tabulation, clustering, and other modeling techniques can also be a part of exploratory analysis

THE CONTEXT OF DATA SCIENTISTS



Source:

- <https://newsletter.theaiedge.io/>

THE DATA SCIENCE PROCESS

The six steps of the
data science process

-  Setting the research Goal
-  Retrieving Data
-  Data preparation
-  Data exploration
-  Data modeling
-  Presentation and automation

THE DATA SCIENCE PROCESS

The six steps of the
data science process

 Setting the research Goal

 Retrieving Data

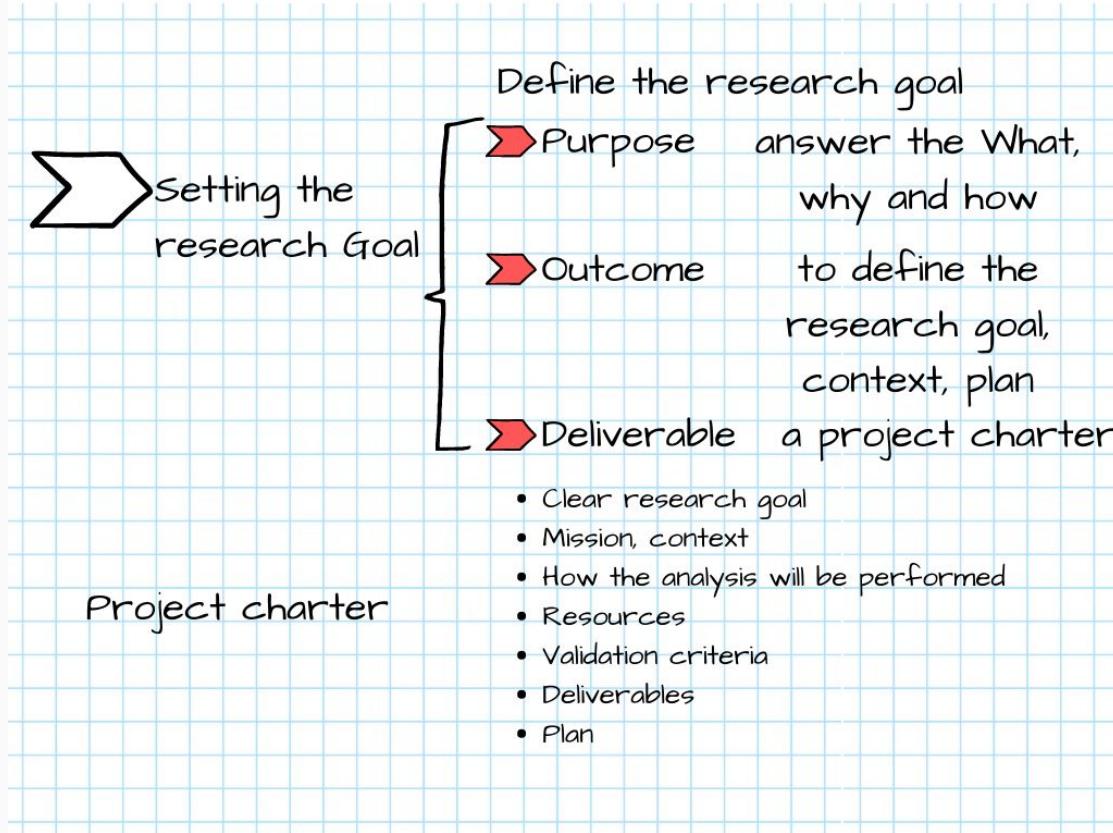
 Data preparation

 Data exploration

 Data modeling

 Presentation and automation

THE DATA SCIENCE PROCESS, STEP 1



Other sources

- The ML Canvas web [[link](#)]
- The ML Canvas book

THE DATA SCIENCE PROCESS, STEP 2

Retrieving Data

Internal Data, data stored in the company
➡ Databases
Data marts
Data warehouses
Data lakes → Raw format

External Data
➡ Companies collecting valuable info / serv
Provide data / return
Public data

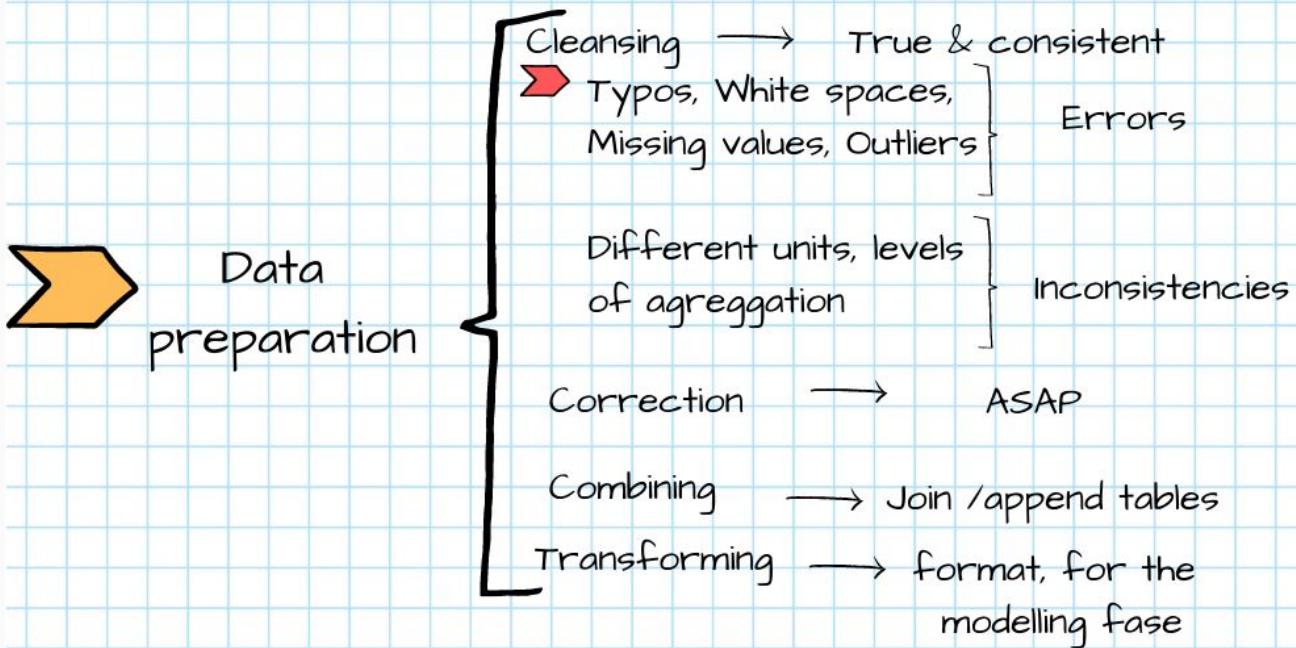
Quality check

Data import
Data preparation → Content of variables
Exploratory analysis → Statistical properties

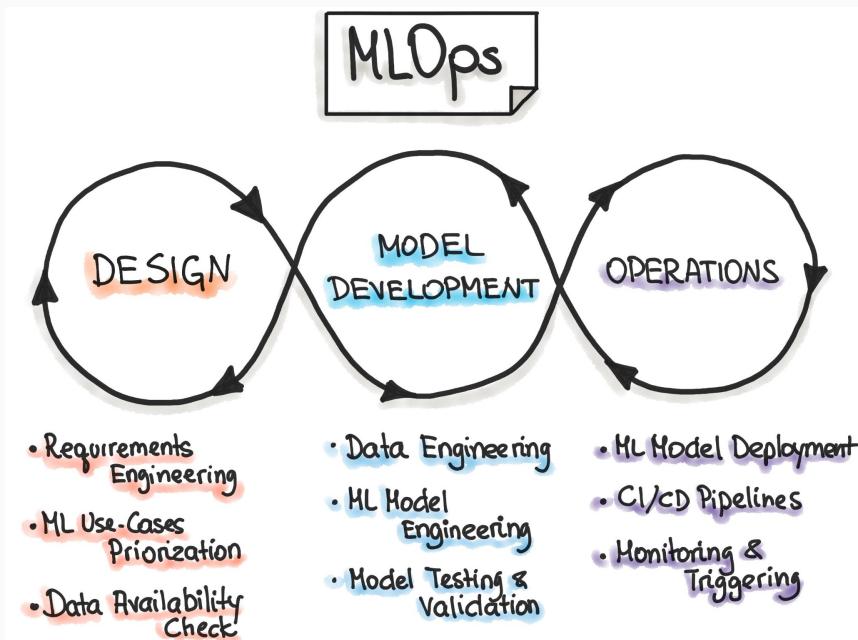
import correctly
check datatypes

THE DATA SCIENCE PROCESS, STEP 3

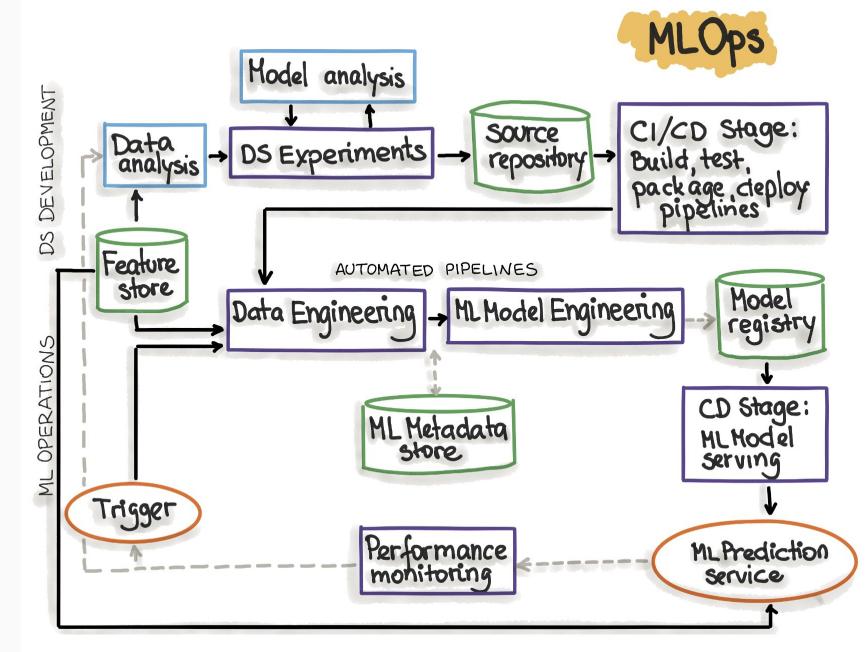
Models need the data in a specific format, so data transformation will always come into play



MODELOS DE PROCESO



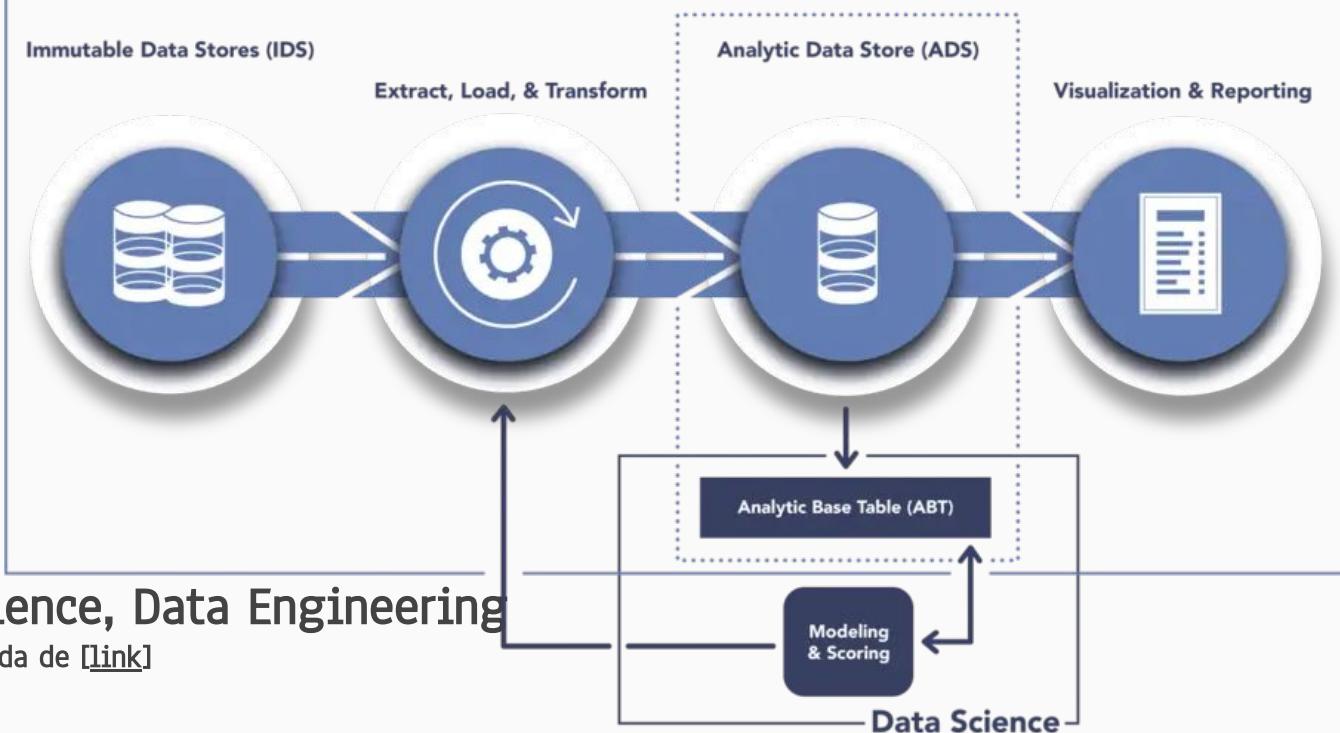
Machine Learning Model Operationalization Management (MLOps)
Imagen tomada de [\[link\]](#)



MLOps - automated ML pipeline
Imagen tomada de [\[link\]](#)

CONTEXTO

Data Engineering

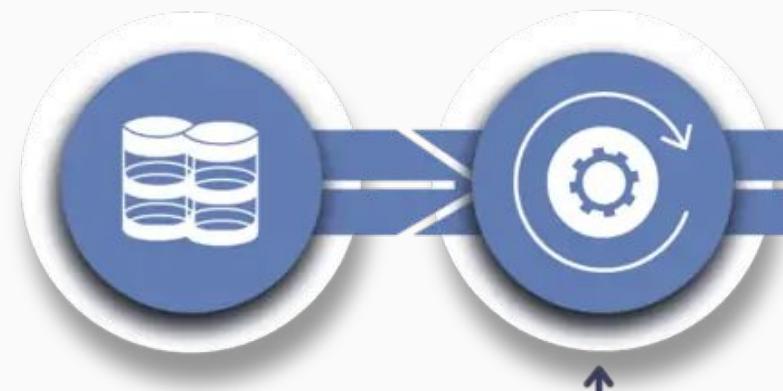


CONTEXTO

Data Engineering

Immutable Data Stores (IDS)

Extract, Load, & Transform



Almacenamiento de datos, carga y transformación.



CIERRE DE LA CLASE



LA PRÓXIMA SEMANA

Fecha: Sábado, 20 de abril de 2024

Tema: Infraestructura para el almacenamiento y procesamiento de datos tradicional y en la nube.

Asignación: Lectura [[link](#)] Tablero [[Miró](#)]

REFERENCES

- [1] Cielen, D., & Meysman, A. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Simon and Schuster. Chapter 1 and 2 [[libro online](#)]
- [2] Crockett, E. (2023, 9 febrero). Structured vs Unstructured Data: Key Differences Explained. Datamation.
<https://www.datamation.com/big-data/structured-vs-unstructured-data/>

THANKS

apvillota@icesi.edu.co

mmrojas@icesi.edu.co

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), and infographics & images by [Freepik](#) and illustrations by [Storyset](#)

Does anyone have any
questions?

