コスプレ

Pipeline de Datos

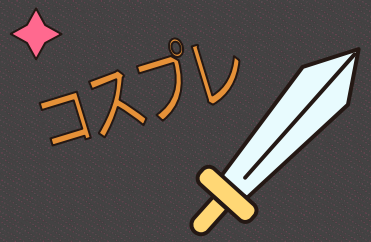
Abstracción del Flujo de Datos

Pipeline de Datos

Idea General

- Generación y recolección de los datos
- Entrada al Data Lake
- Pre-procesamiento de datos para su consumo interno
- Entrada a la Base de Datos
- Consumo de la Base de Datos
- Pre-procesamiento para cualquier proceso (Pipeline de Ciencia de Datos)
- Ejecución del proceso
- Generación de resultados y entrada a Base de Datos
- Consumo final (o repetir)

Conceptos Generales para Científiques de Datos



Datalake	Es un repositorio centralizado y escalable donde se almacenan grandes volúmenes de datos en bruto, sin procesar y en su forma original, permitiendo un análisis flexible y ágil de los datos de una organización.
Base de Datos	Es un sistema estructurado que organiza y almacena datos de manera eficiente, aplicando un esquema predefinido y una estructura de tablas relacionales.
Data Warehouse	Es una base de datos especializada que se utiliza para almacenar y organizar datos empresariales estructurados y procesados.
Pre-procesamiento de Datos	Toda transformación de datos que no afecte las decisiones o supuestos estadísticos
Procesamiento de Datos	Toda transformación de datos que afecte o tenga supuestos estadísticos
Algoritmo de Datos	Proceso abstracto que aplicaremos a los Datos (lo que vamos a entrenar)
Modelo de Datos	Proceso abstracto ya entrenado con los Datos (un algoritmo ya entrenado)

Pipeline de Datos

01

ETL

Extract Transform Load (Extraer Transformar y Carga)

02

EDA

Exploratory Data Analysis (Analisis Exploratorio)

03

Entrenamiento

Entrenamiento del modelo (utilización de los datos)

04

Produccion

Automatizar todo el proceso para su consumo

Atencion!

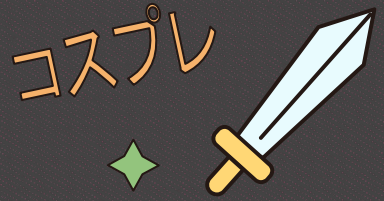
Todos los conceptos que veamos son abstracciones que nos sirven para darnos ideas, ninguna definición es perfecta o inmutable!

Nos estaremos enfocando en procesos de Ciencia de Datos, pero esto aplica para otras cosas.
No es un proceso lineal, pueden estar saltando entre ciclos con las mismas etapas.

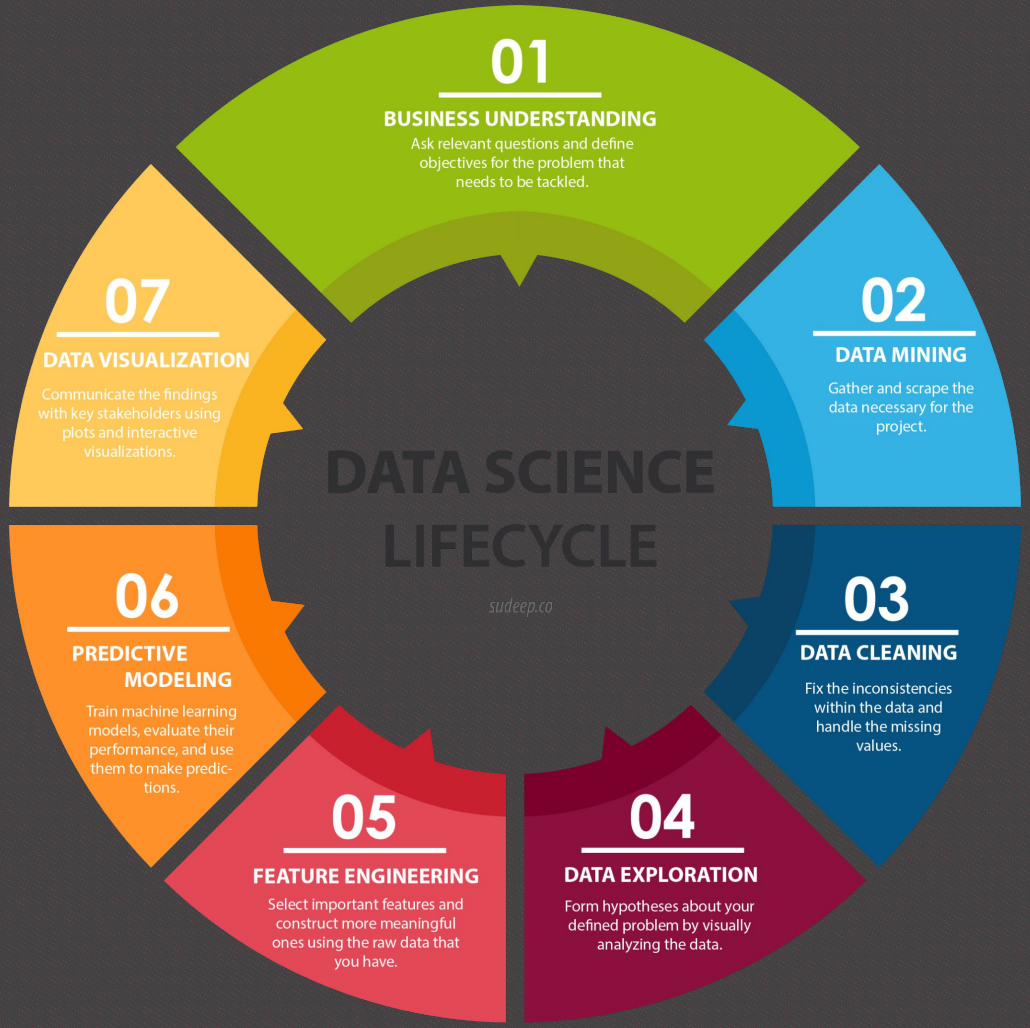
CrowdFlower, provider of a “data enrichment” platform for data scientists, conducted a survey of about 80 data scientists and found that data scientists spend –

- 60% of the time in organizing and cleaning data.
- 19% of the time is spent in collecting datasets.
- 9% of the time is spent in mining the data to draw patterns.
- 3% of the time is spent in training the datasets.
- 4% of the time is spent in refining the algorithms.
- 5% of the time is spent in other tasks.

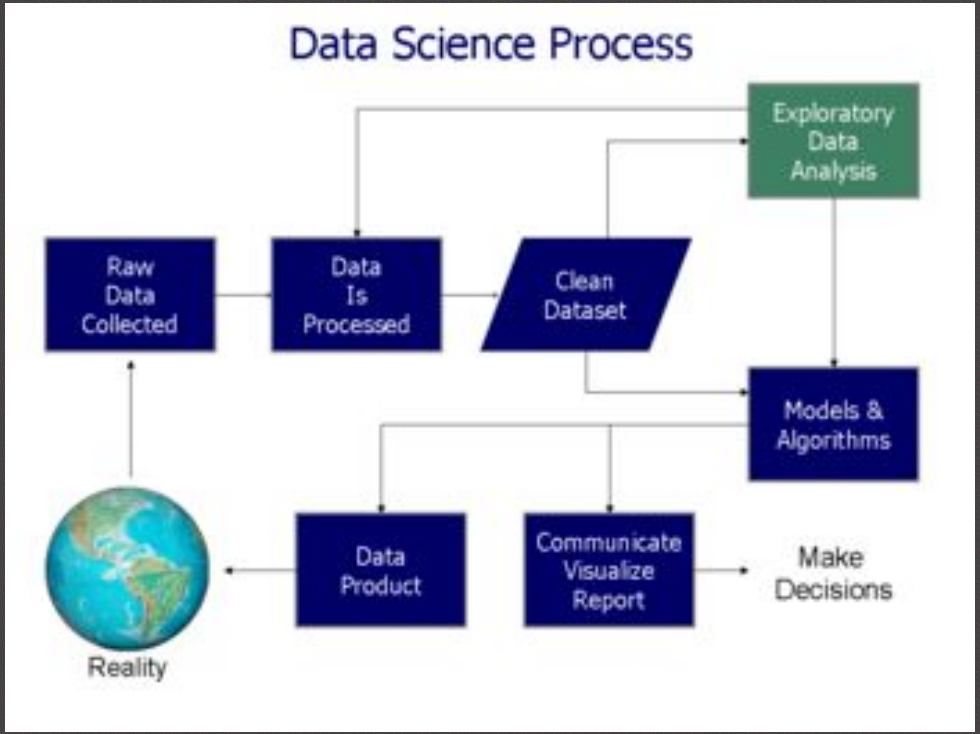
La mayoría del tiempo lo ocuparemos limpiando, organizando y conociendo nuestros datos.



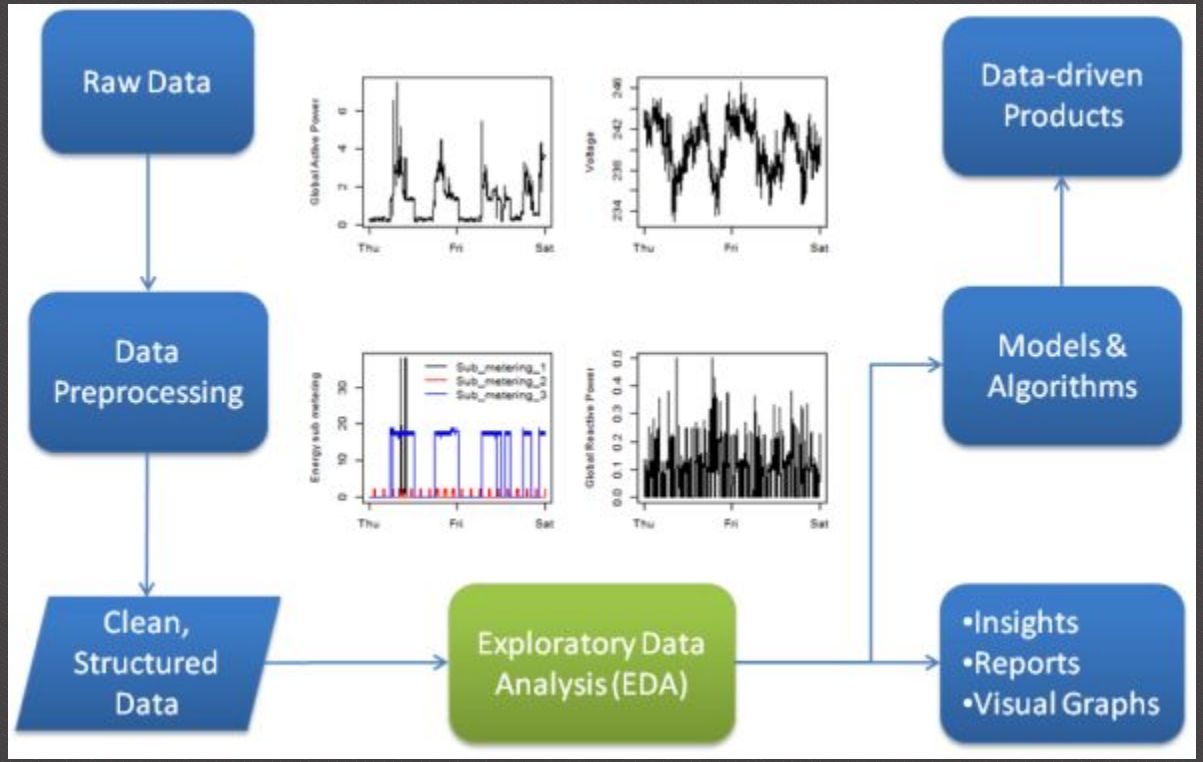
Fuente:
https://www.projectpro.io/article/why-data-preparation-is-an-important-part-of-data-science/242


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Caricatura Empresarial



Caricatura Empresarial



“Es un error capital teorizar antes de contar con los datos. Imperceptiblemente, uno comienza a torcer los hechos para que se ajusten a las teorías, en lugar de ajustar las teorías a los hechos.”

–Arthur Conan Doyle, Sherlock Holmes

01

ETL

Extraccion Transformacion y Carga

Que es?

Es un proceso fundamental en la gestión y análisis de datos que permite integrar y preparar datos de diversas fuentes para su posterior análisis.

Es utilizado para integrar datos desde múltiples fuentes en un sistema de almacenamiento o base de datos o en algún lugar que lo prepare para su consumo.

Extract (Extraccion)

Que es?

Consiste en obtener los datos de diferentes fuentes, como bases de datos, archivos, APIs, servicios web, entre otros.

Fuentes de Datos

Pueden variar ampliamente, incluyendo bases de datos relacionales, archivos CSV, JSON, XML, sistemas en tiempo real, registros de eventos, redes sociales, entre otros.

Conectividad

Se establecen conexiones para acceder y guardar los datos. Esto implica configurar credenciales, protocolos de comunicación y mecanismos de autenticación.

Preprocesamiento

Como filtrado, muestreo o eliminación de datos no deseados. Esto ayuda a reducir el volumen de datos y mejorar su calidad antes de pasar a la etapa de transformación.

Validación y verificación

Garantizar la integridad de los datos extraídos, cómo comprobar la coherencia de los valores, la integridad referencial y la detección de errores.

Metadatos

Registro de metadatos que describen la fuente de datos, la fecha de extracción, el método utilizado y cualquier otro detalle relevante.

Transformación

Que es?

implica aplicar reglas y manipulaciones a los datos extraídos para limpiarlos, combinarlos, filtrarlos y cambiar su estructura.

Que se busca?

La estandarización de formatos para asegurar que los datos estén coherentes, consistentes y preparados para su análisis posterior.

Procesamiento

Recuerden especificar qué decisiones estadísticas tomaron. Si el proceso está diseñado para guardar datos brutos, eviten decisiones estadísticas

Validación y verificación

Esta etapa también puede incluir la validación de datos, la detección y manejo de valores nulos o faltantes, así como la eliminación de duplicados o valores atípicos.

Metadatos

Registro de metadatos que describen las transformaciones aplicadas.

Datos Tabulares y Tidy Data

Datos Tabulares

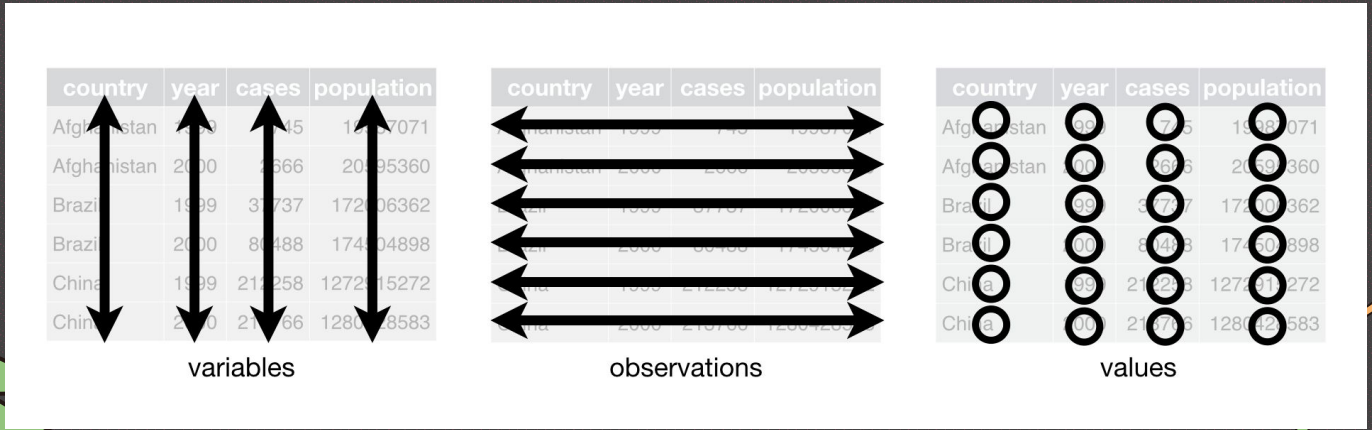
Los datos tabulares son una forma **estructurada** de almacenar información en filas y columnas, similar a una tabla.

Tidy Data

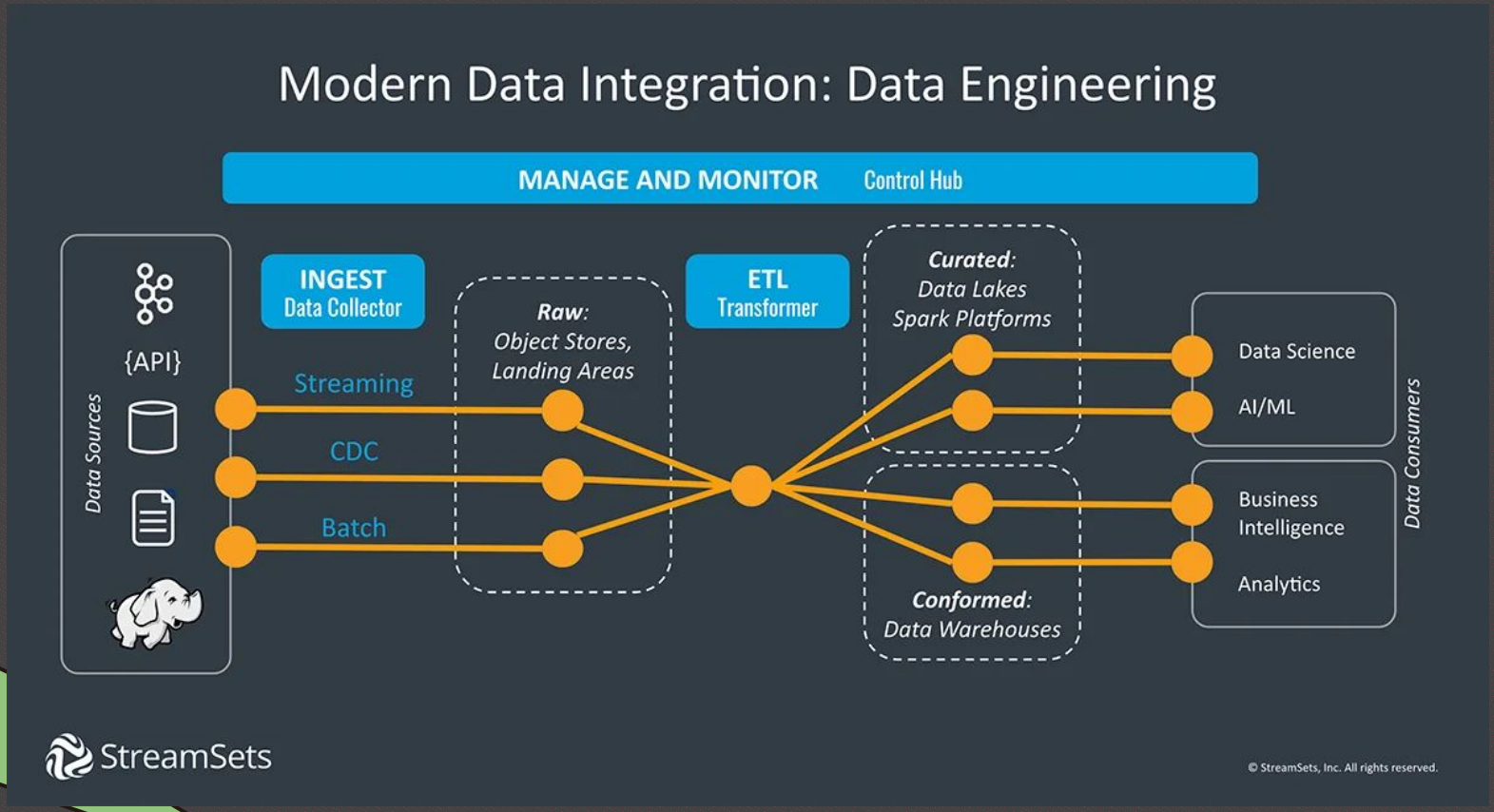
Cada fila representa una instancia o registro, mientras que **cada columna corresponde a un atributo o variable específica.**

Uso

Bases de Datos SQL (Postgres MySQL), Excel u hojas de cálculo, Pandas.



Caricatura Empresarial



EDA

Exploratory Data Analysis

Que es el EDA?

Consiste en examinar y comprender los datos en bruto (y no en bruto) de una manera sistemática y visual, utilizando técnicas estadísticas y herramientas gráficas.

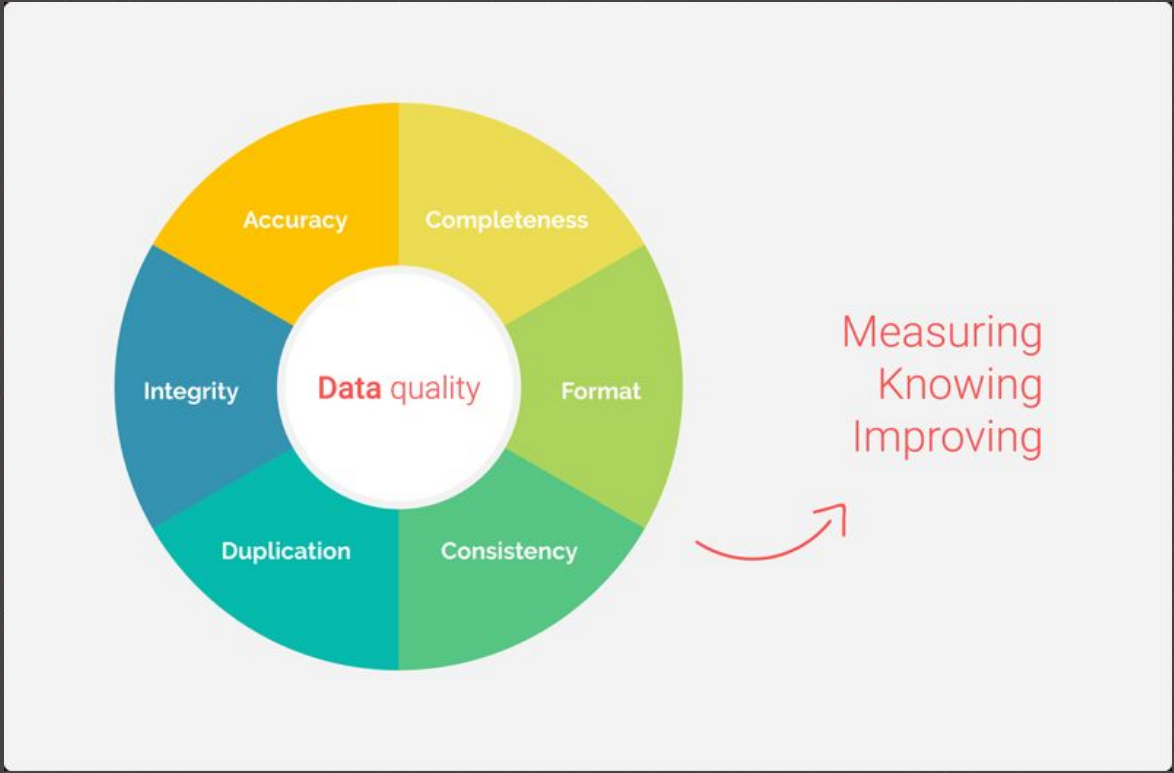
El EDA permite descubrir patrones, identificar valores atípicos, detectar relaciones y obtener conocimientos iniciales sobre los datos antes de realizar un análisis más profundo.

Que es el EDA?

Se realizan tareas como la visualización de datos mediante gráficos, tablas o diagramas, el cálculo de medidas descriptivas, la identificación de valores faltantes o inconsistentes, la exploración de correlaciones entre variables y la realización de pruebas estadísticas preliminares.

Proporciona una comprensión profunda de los datos, permitiendo a los analistas obtener ideas, generar nuevas hipótesis y validar suposiciones antes de realizar un análisis más profundo.

EDA: Calidad de los Datos



EDA

Viabilidad

En esta etapa se busca entender si el proyecto es viable o no. La única forma de saberlo es conociendo los datos.

Conocimiento

En esta etapa se busca obtener conocimiento del negocio (o proyecto) y entender qué está pasando.

Visualizacion

Se busca generar visualizaciones que nos ayuden a entender cómo son los datos y lo que podemos lograr.

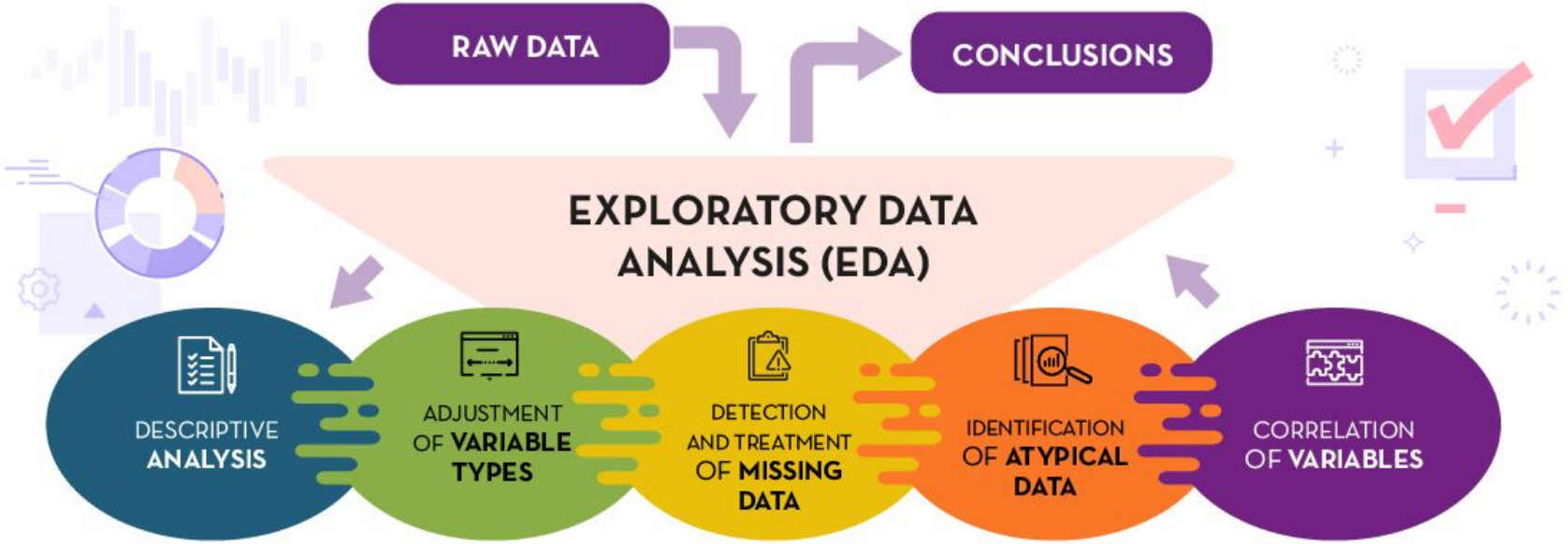
Herramientas

Pandas, graficas (Tableau, ggplot, seaborn, plotly). Teoría estadística y estadísticos.

Al finalizar

También se busca generar preguntas que podamos hacer a expertes.

Caricatura Empresarial



Posiciones

コスプレ

Posiciones

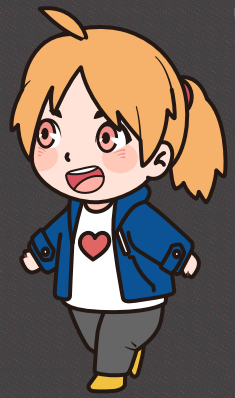
- **Ingeniere de datos**: Especialista en el diseño, implementación y mantenimiento de arquitecturas y pipelines de datos para almacenar, procesar y gestionar grandes volúmenes de datos de manera eficiente y escalable.

- **Científique de datos**: Experto en aplicar técnicas estadísticas, algoritmos de aprendizaje automático y análisis de datos para descubrir patrones, construir modelos predictivos y extraer conocimientos significativos de los datos.

Posiciones

- **ML Engineer (Ingeniera de aprendizaje automático):** Especialista en desarrollar, implementar y optimizar modelos de aprendizaje automático y sistemas de aprendizaje automático en producción, garantizando su rendimiento, escalabilidad y eficiencia.

- **Analista de datos:** Profesional encargade de examinar y analizar datos existentes para identificar patrones, tendencias y obtener información valiosa que pueda respaldar la toma de decisiones empresariales.

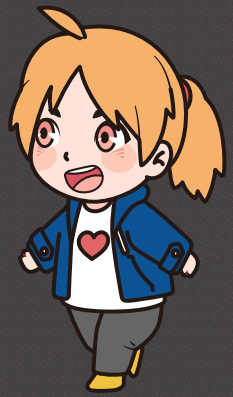


Posiciones

- **MLOps (Operaciones de Aprendizaje Automático):** Es la combinación de prácticas y tecnologías utilizadas para gestionar, implementar y operar modelos de aprendizaje automático en producción de manera eficiente y escalable. Los profesionales de MLOps colaboran estrechamente con científicos de datos y desarrolladores para garantizar la reproducibilidad, el monitoreo y la gobernanza de los modelos de ML.
- **DevOps (Desarrollo y Operaciones):** Es una metodología que combina las prácticas de desarrollo de software y operaciones de TI para agilizar y automatizar el ciclo de vida de desarrollo de aplicaciones. Los profesionales de DevOps se enfocan en la colaboración, la integración continua, la entrega continua y la automatización de pruebas y despliegues, mejorando la eficiencia y la calidad del software.

Posiciones

- Developer (Desarrolladrx)**:** Profesional de desarrollo de software que crea aplicaciones y herramientas para recopilar, procesar, visualizar y analizar datos, utilizando lenguajes de programación y tecnologías adecuadas. Normalmente se divide en Frontend y Backend



(Posiciones) Backend & Frontend

- **Backend Developer** (Desarrolladorx de backend): Se especializa en la implementación y gestión de la **lógica y funcionalidad de la parte del servidor** de una aplicación o sistema, que se encarga del procesamiento y almacenamiento de datos.
-
- **Frontend Developer** (Desarrolladorx de frontend): Responsable de crear la **interfaz de usuario y la experiencia de usuario** de una aplicación o sistema, que se encarga de la presentación y visualización de los datos. Tambien hay UX/UI (User Experience & User Interface)

Cómo lo ve el desarrollador Back-end...

Cómo lo ve el desarrollador Front-end...

Posiciones

- **Project Manager (Gerente de Proyecto):** Profesional encargado de supervisar el desarrollo y ejecución exitosa de proyectos de ciencia de datos. Gestiona los recursos, el cronograma y los riesgos, garantizando la entrega del proyecto dentro de los plazos y presupuestos establecidos. Generalmente es el encargado de utilizar las técnicas Agiles/Agile **También es la persona encargada de presionar a los demás, así que deberán aprender a no ceder ante esta presión y establecer límites.**

Posiciones

- **Product Owner (Dueño de Producto):** Responsable de definir y priorizar los requisitos y funcionalidades del producto o servicio basado en ciencia de datos. Colabora con el equipo para asegurar que las necesidades del negocio se satisfagan, y se encarga de la visión, la estrategia y el éxito del producto, trabajando de cerca con los stakeholders y el equipo de desarrollo.

Tarea siguiente clase 22 Enero

- Repasar lo visto en clase
- Ver los siguientes videos sobre NLP & LLMs
- Crear cuenta de ChatGPT y de Google Bard
- Crear cuenta de Google Colab (videos)

コスプレ

Material Adicional