

Ingeniería de datos

Big Data



Introducción

La ingeniería de datos construye la base para la ciencia de datos y la analítica de datos en producción.



Introducción


Antes del Big Data ya había ingeniería de datos, entendida como las operaciones necesarias para permitir el acceso a los flujos de información, mediante procesos ETL, pero con el auge del Big Data y la ingente cantidad de herramientas disponibles su importancia se ha multiplicado.



Definición


La **ingeniería de datos** trata del movimiento, manipulación y gestión de los datos.

Se centra en el desarrollo, implementación y mantenimiento de los sistemas y procesos que recuperan los datos en crudo y produce información consistente y de alta calidad que da soporte a los diferentes casos de uso, como pueden ser la analítica de datos o el ML.



Definición

El encargado de gestionar su ciclo de vida es el **Ingeniero de Datos o Data Engineer**, recuperando los datos desde los sistemas origen y sirviendo los datos a los futuros consumidores, ya sean:

- científicos de datos
 - herramientas de visualización
 - modelos de IA
- 

Científico de datos o Data Scientist

La relación existente entre un **Data Engineer** y un **Data Scientist**, es que el primero le deja los datos preparados al segundo, obteniendo los datos y dándoles valor para que el científico realice la analítica y la ciencia de datos.



Roles


Un **Data Engineer** trabaja con plataformas de Big Data como Spark/Databricks o Snowflake (antiguamente todo se basaba en el ecosistema Hadoop), bases de datos relacionales y NoSQL.

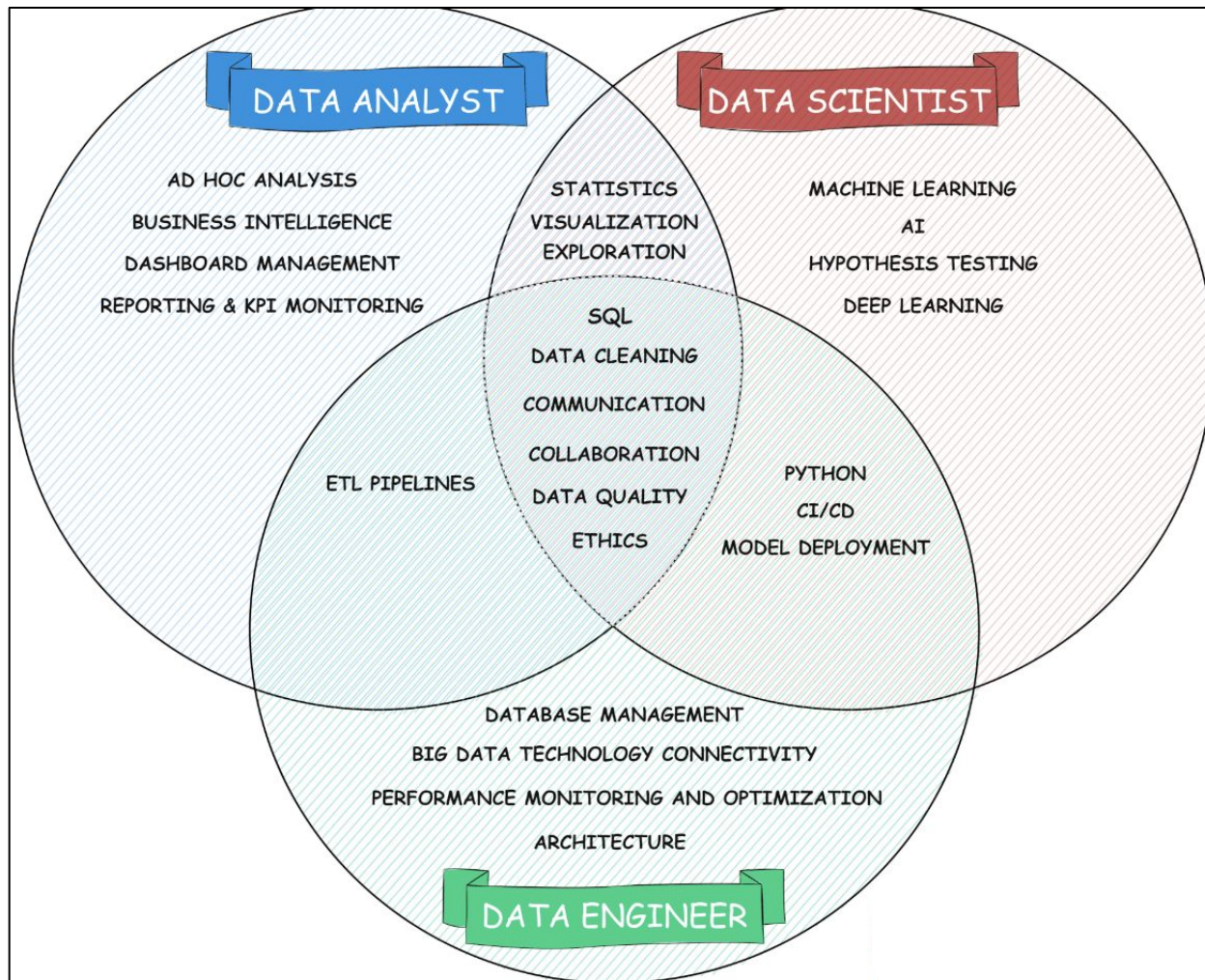
Además, implica tener destrezas en modelado de datos, integración de datos, transformación de datos, calidad y gobernanza del dato.



Roles

Además del **Data Engineer**, existen otros roles específicos:

- **Product Data Engineer:** encargado de instalar, configurar y mantener los productos del equipo de ingeniería de datos, como pueden ser Airflow, Kafka o Spark
 - **Pipeline Data Engineer:** encargado de trabajar con el flujo de datos, con conocimiento de Python, SQL, Spark así como trabajar con Data Lakes.
 - **BI Data Engineer:** SQL y herramientas de visualización como Power BI o Tableau, para mostrar analíticas.
- 



DATA ANALYST



Collect and organize data.



Clean, transform and process data.



Use statistics to identify patterns.



Create reports/visuals to communicate insights.

DATA SCIENTIST



Build predictive models/algorithms.



Use machine learning to predict trends.



Solve complex problems with models.



Find new data sources/opportunities for insights.

Machine Learning

- Classification
- Regression
- Reinforcement Learning
- Deep Learning
- Clustering
- Dimensionally reduction

Programming Language

- Python
- R
- Java

Data Visualization

- Tableau
- Power BI
- Matplotlib
- GG Plot
- Seaborn

Data Analysis

- Feature Engineering
- Data Wrangling
- EDA

Data Science



IDE

- Pycharm
- Jupyter
- Colaboratory
- Spyder
- R-Studio

Math

- Statistics
- Linear Algebra
- Differential Calculas

Deploy

- AWS
- AZURE

Web Scraping

- Beautiful Soup
- Scrapy
- URLLIB