Fundamentos Ing. De datos

lunes, 13 de febrero de 2023 22:37

Ciclo de vida Ciencia de datos:

- Hacer una pregunta interesante (plantearse hipótesis)
- Obtener los datos
- Explorar los datos
- Modelar los datos
- Comunicar y visualizar los resultados

Ingeniero de datos: Apoya haciendo que los datos estén en un ambiente productivo y útil. Asegura la calidad del código desarrollado y hace que el mismo pueda ser código escalable y fiable.

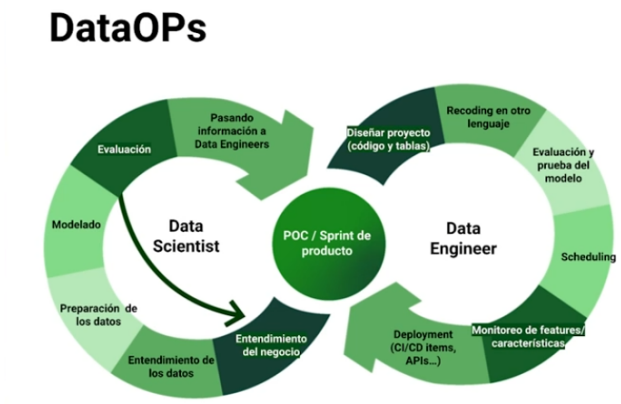
Toma los datos crudos de diferentes fuentes de datos para almacenarlos en bases de datos y que los mismos estén disponible para el software en producción. Para ello crean pipelines ETL.

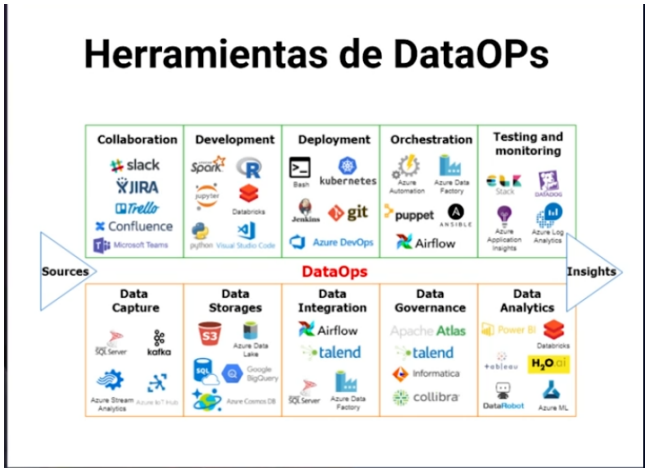
DataOps: Entregar a la organización los datos adecuados de manera rápida.

Debe de ser Ágil, de desarrollo y entrega continua (CI/CD).

Secuencia de proceso y filosofía. Se compone de Agil, DevOps y Lean.

- Ágil: Entrega de valor continua.
- DevOps: Todas las actividades que acompañan al desarrollo para hacerlo disponible.
- Lean: Foco en los procesos de valor, mapeo de los mismos. Llegar a un proceso lo más ligero posible. Reducir a lo máximo los residuos generados por los proyectos.





Tareas:

- Escalar el modelo ML de pruebas a un modelo que pueda ser viable en producción.
- Evaluar, observar y monitorear el funcionamiento del modelo en producción, que el mismo corra correctamente.
- Automatización de tareas de re entrenamiento del modelo.
- Disponibilidad los datos.

Planeación: Planear por medio de Agile (Sprints)

Regla de Beetlejuice:
Esperar a que la misma tarea surja 3 veces antes de proceder a automatizarla.

Bases de datos SQL aseguran
A: Atomicidad
C: Consistencia
I: Integridad
D: Durabilidad

Bases de datos NO SQL:
- Útiles para guardar objetos flexibles.
- Información más fácil de consultar desde un lenguaje de programación.

APIS: Consumen información de otras plataformas y sirven para comunicar diferentes plataformas.

Estructuración de la información por medio de **modelos de datos** (relacional, documental, etc).

Un PipeLine de datos corresponde a un proceso de ETL
Airflow (automatización de Pipeline) {
¿Qué código se necesita?
¿En qué secuencia? ¿Qué código va a correr después de que?
¿Cada cuánto?
}

Airflow hace uso de un DAG, un grafo el cual determina la secuencialidad de las tareas a ejecutar.

Spark {
Util para: Procesamiento en parelelo
-> Procesamiento en Batch (por bloques)
-> Procesamiento en Stream (en vivo)
}

Cron: Lenguaje con el cual se definen tareas que tienen que correr con determinada periodicidad.

Docker: Hace que las aplicaciones corran en un ambiente aislado y controlado (contendor).
Orquestación: Solución al problema de gestión de contenedores y saber que contenedores corren y como corren.

Dockerizar los procesos de datos.

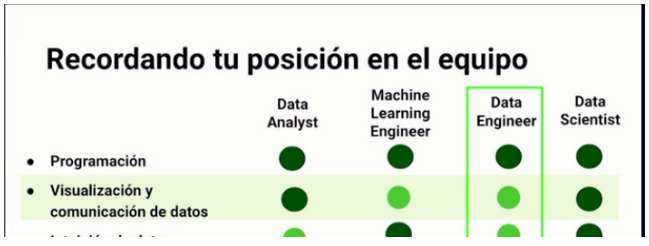
(build): Contribución la cual lleva un orden de etapas (branches), da una forma de aplicar una validación adecuada al código.

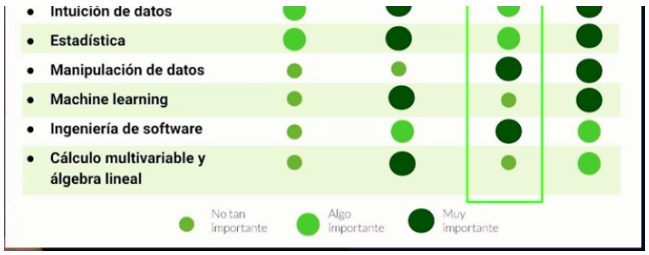
Testing en data: Hacer una función o serie de funciones que verifiquen cada uno de los procesos que hay.
- Testeo de APIS
- Testeo de flujos de datos (pipelines): Se testea que todos los procesos se cumplan.
- Testeo de calidad de datos: https://servian.dev/data-quality-and-testing-frameworks-316c09436ab2

Nube: Lugar para montar pipelines de datos para que den valor.

ML Engineer: Hacen las cosas especificas del modelo
Data Engineer. Trabajar en produccion y con bases de datos, ademas del modelo.

Visibilidad:
- A nivel de dashboard
- A nivel de alertas y notificaciones.




Intuición de datos
Estadística
Manipulación de datos
Machine learning
Ingeniería de software
Cálculo multivariable y álgebra lineal
No tan importante
Algo importante
Muy importante