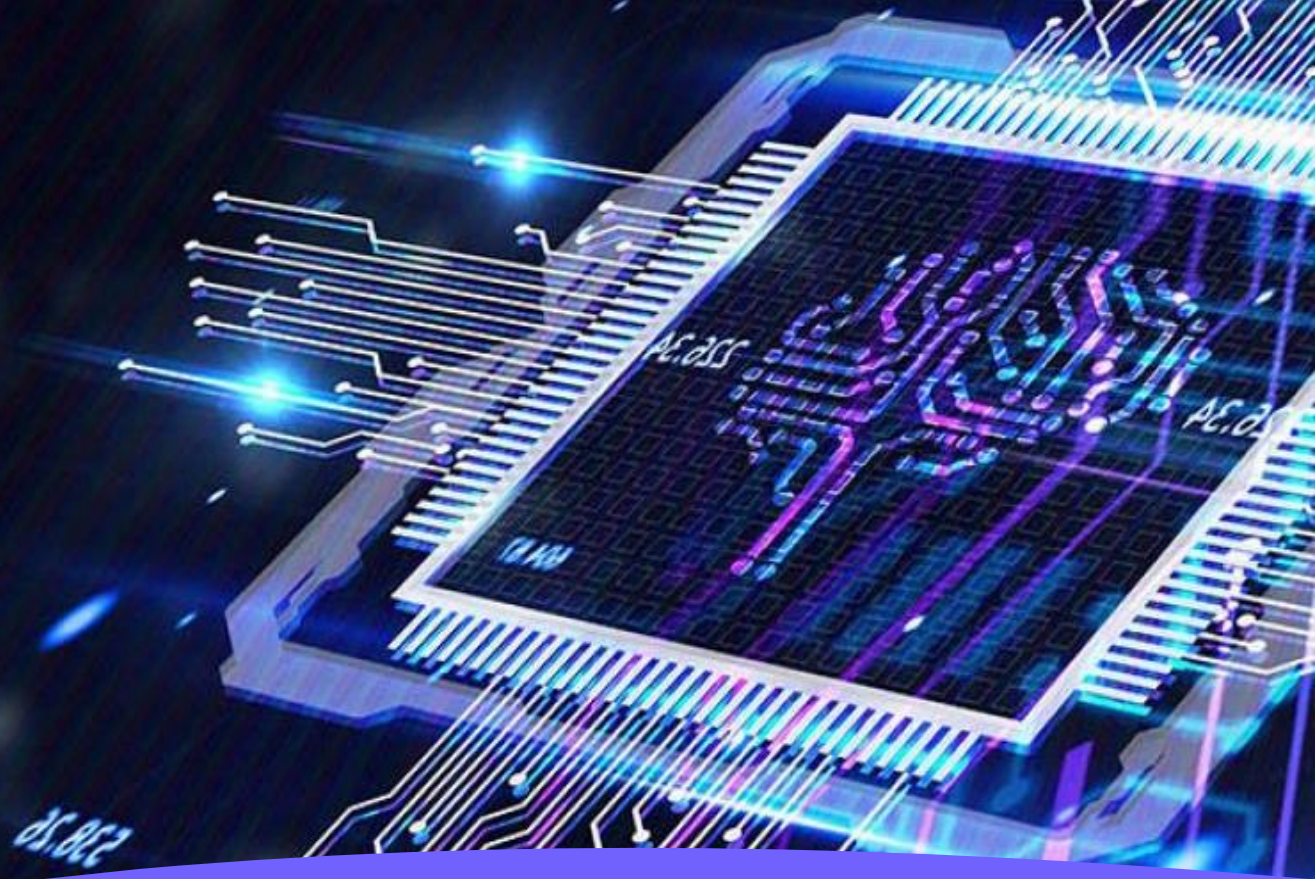




**FACULTAD
DE INGENIERIA**
Universidad de Buenos Aires



Aprendizaje de máquina II

Carrera de
Especialización en
Inteligencia Artificial

Agenda



- Roles dentro de la industria de datos
- Buenas prácticas de programación
- Código para producción
- Presentación TP integrador

Roles dentro de la industria de datos



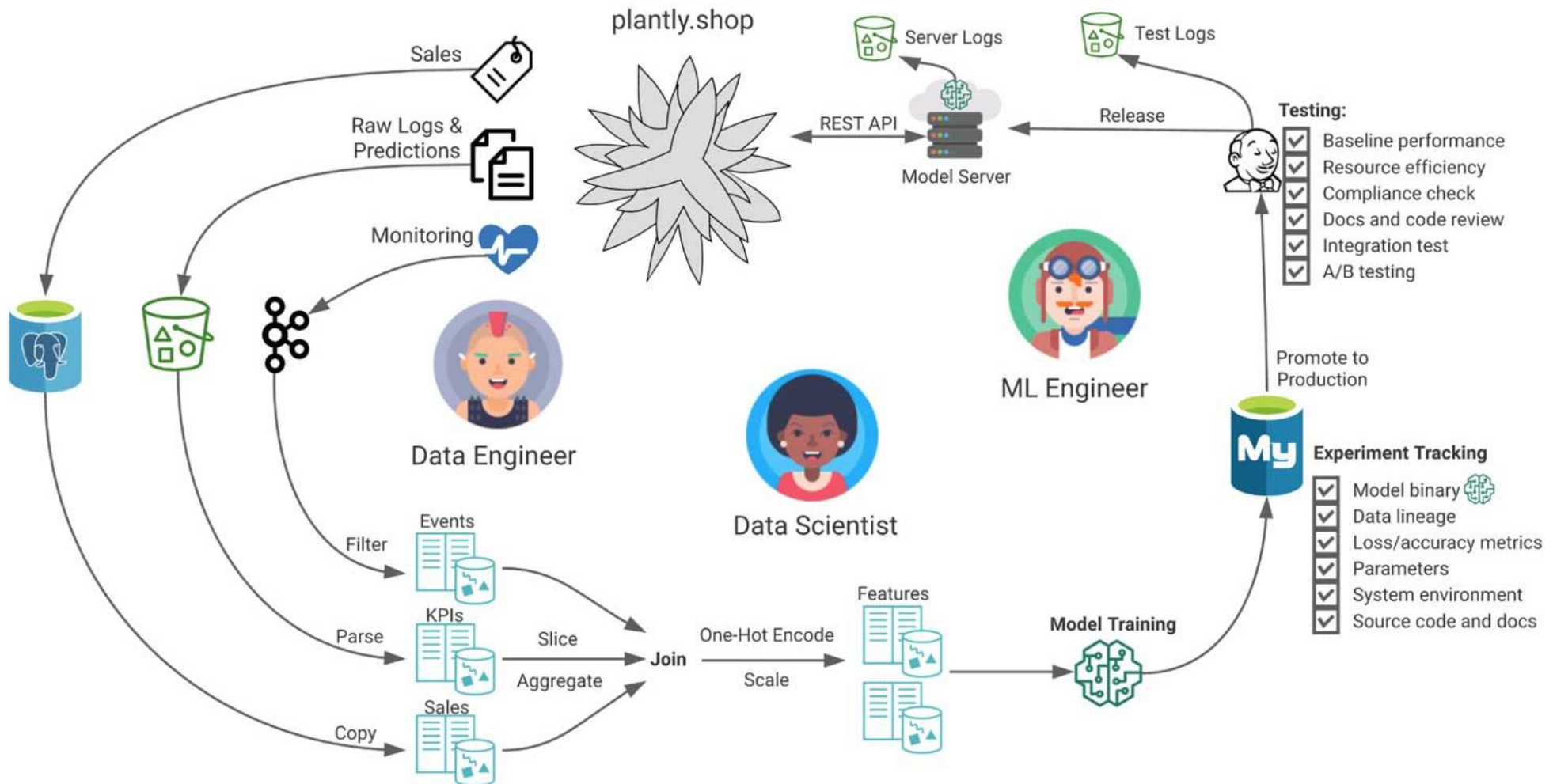
Ciclo de vida de un proyecto de ML

A lo largo del ciclo de vida de un proyecto de ML deben intervenir varios participantes para que el desarrollo se lleve a cabo de la mejor manera posible. La distribución de tareas en los distintos roles pueden variar según cada una de las organizaciones, pero de manera general podemos definir las siguientes:

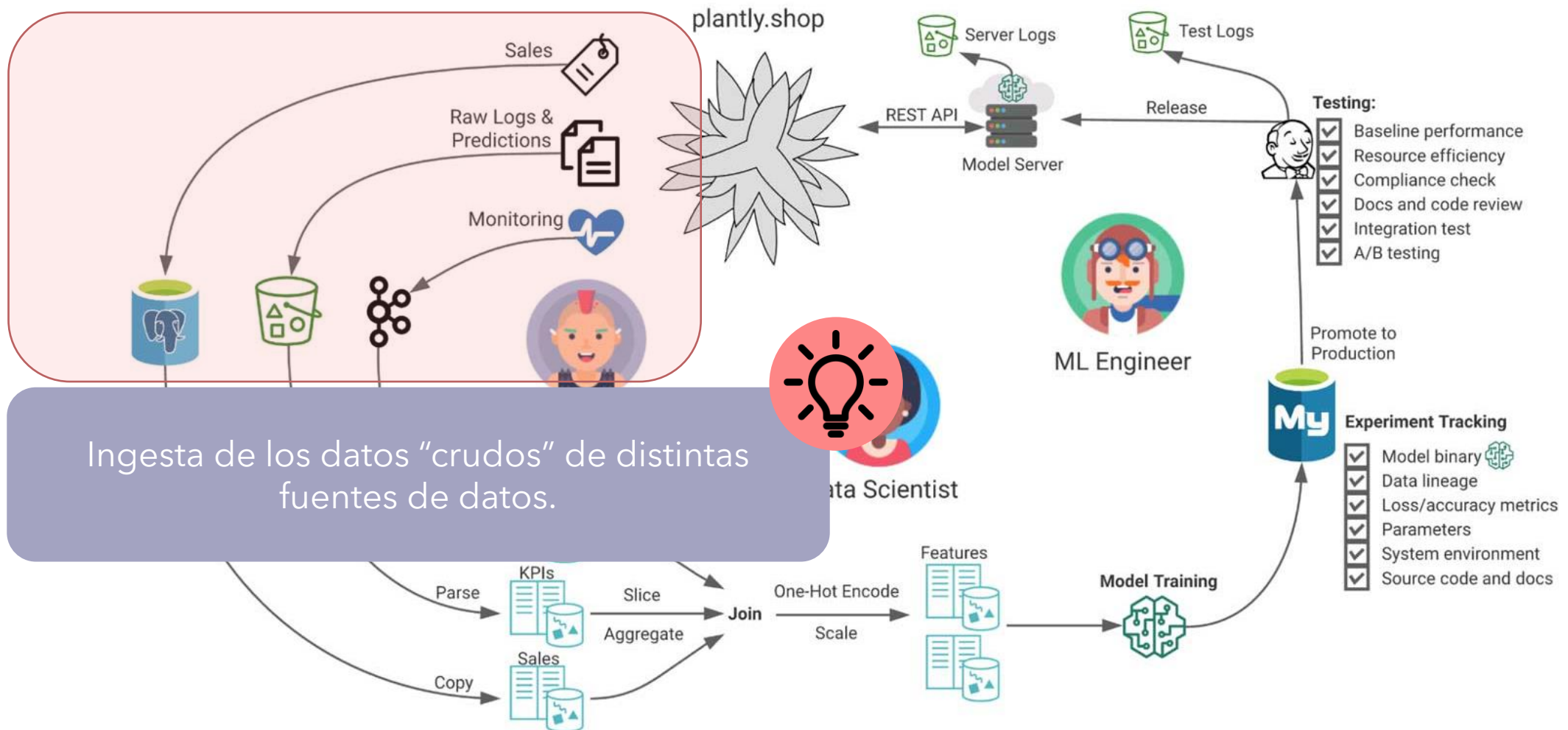
- Data engineer
- Data scientist
- Machine learning engineer



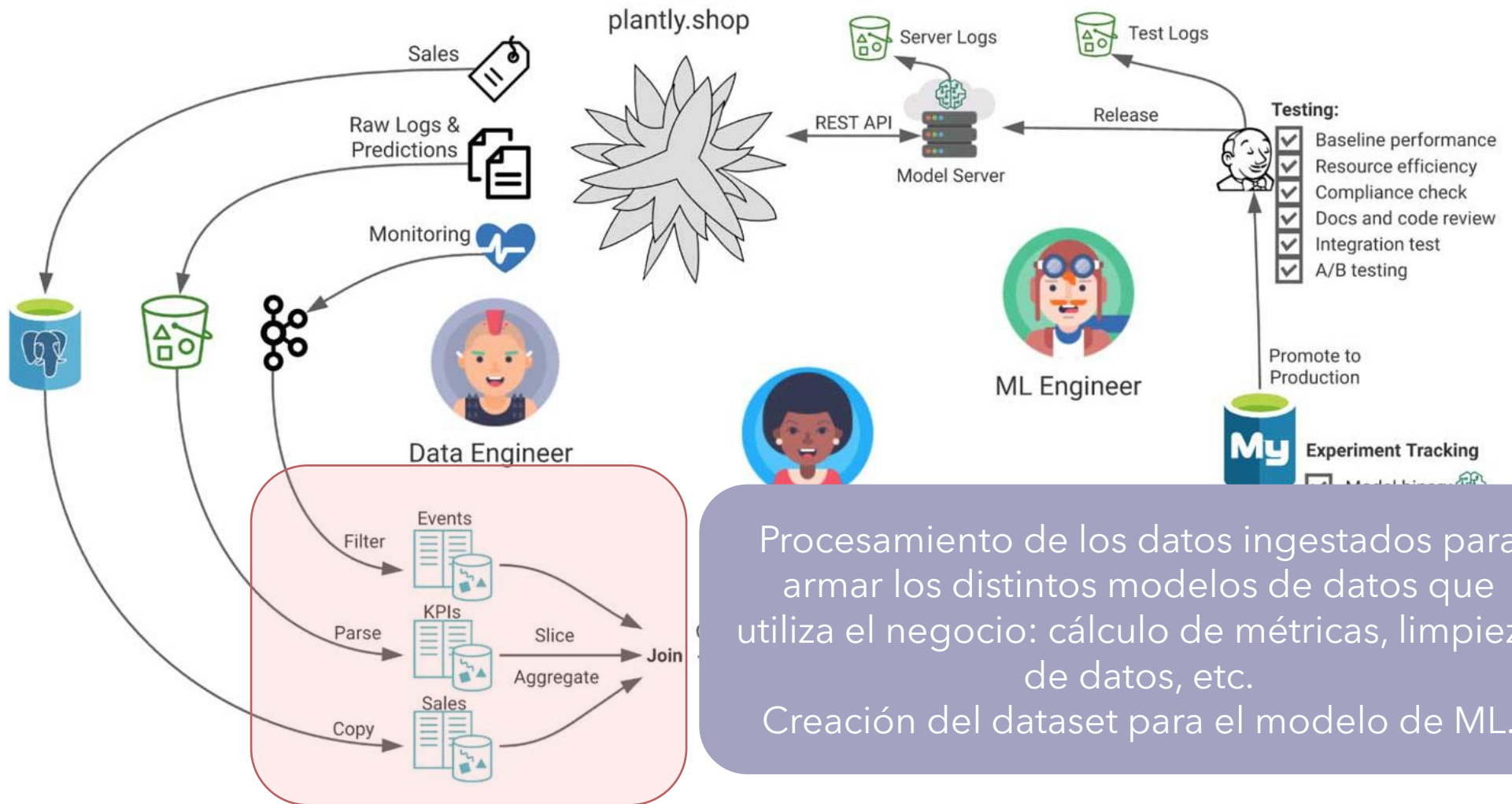
Relación entre los distintos roles de trabajo



Flujo de trabajo típico en un equipo de ciencia de datos



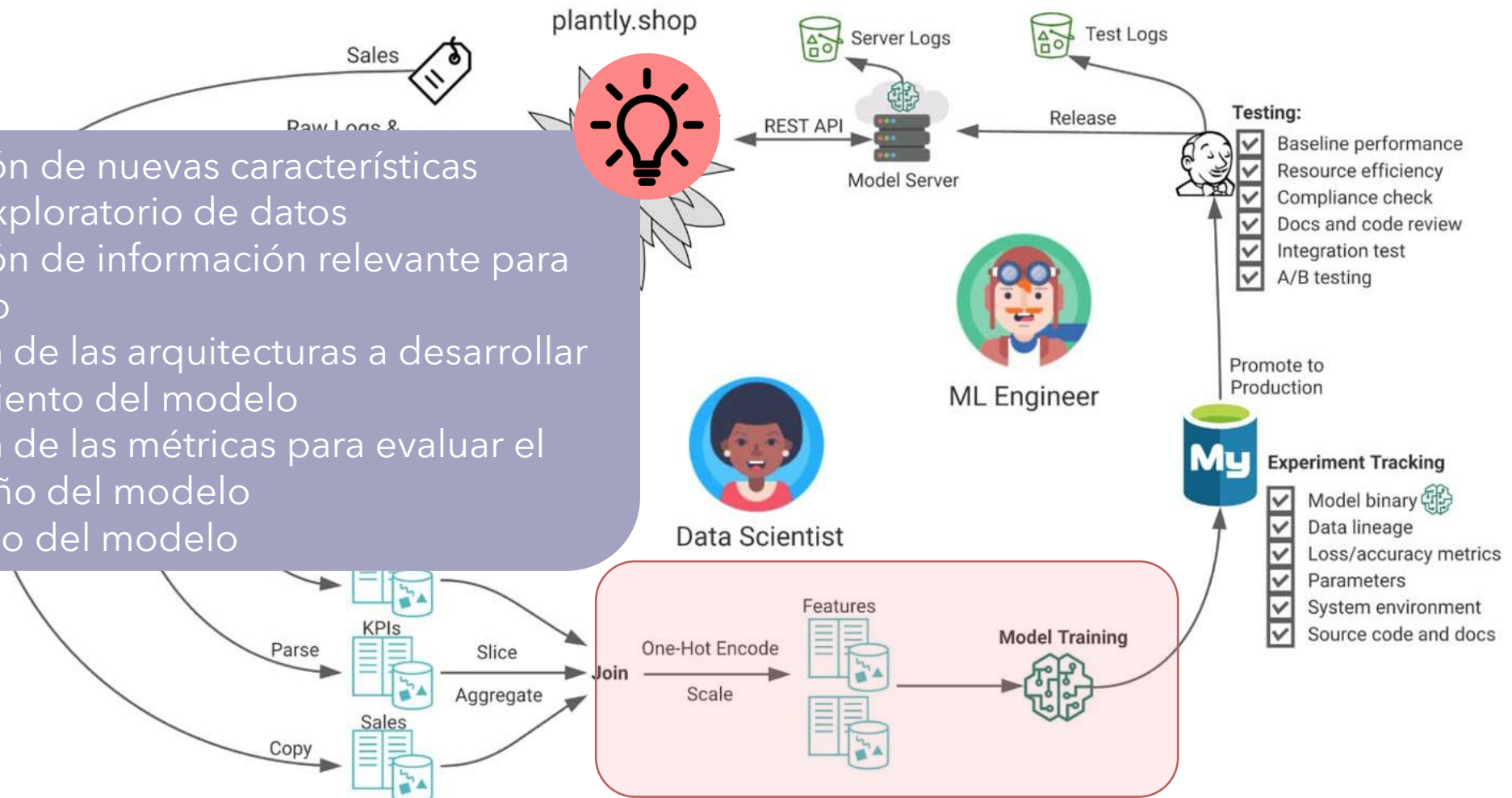
Flujo de trabajo típico en un equipo de ciencia de datos



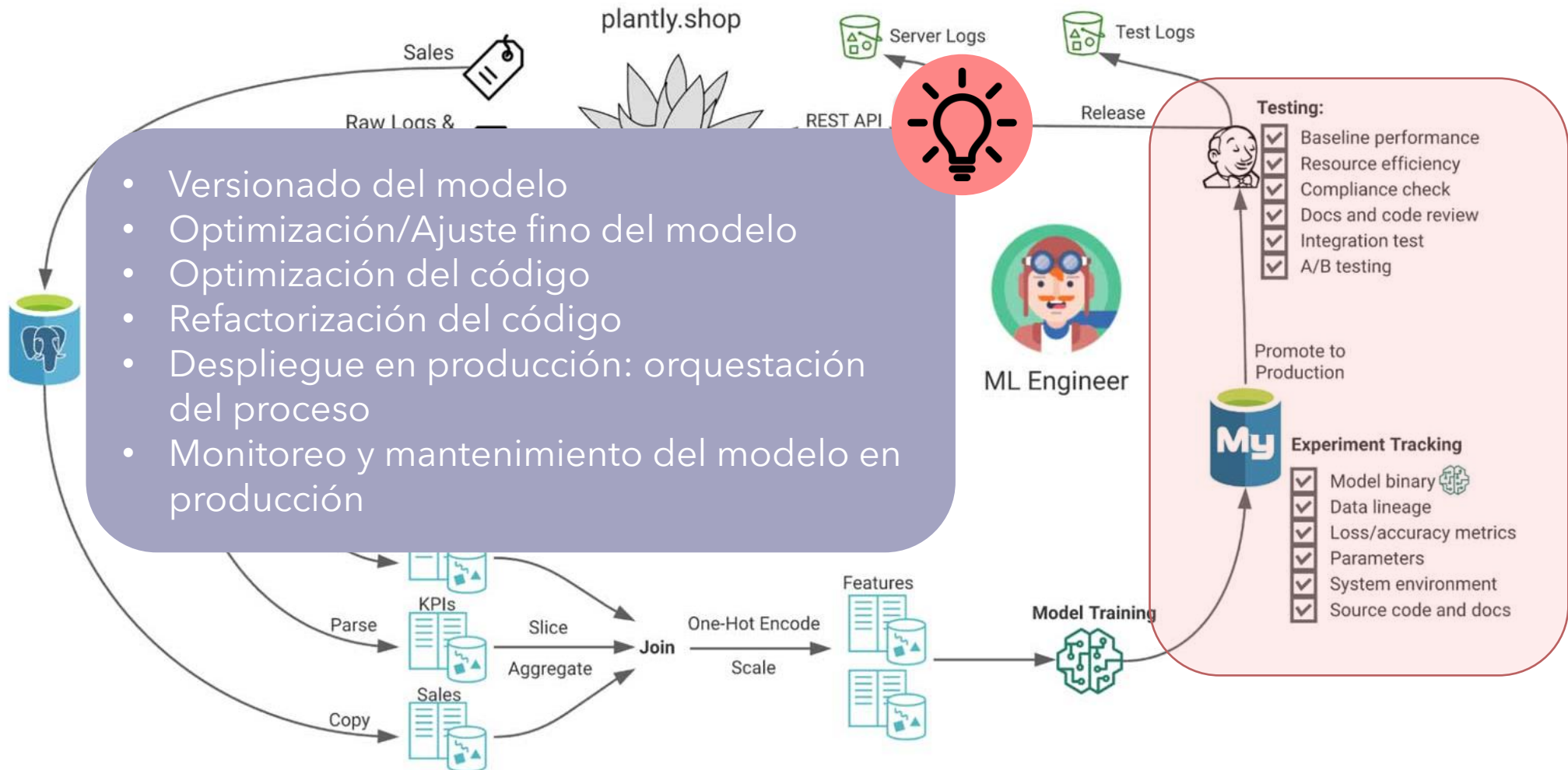
Procesamiento de los datos ingestados para armar los distintos modelos de datos que utiliza el negocio: cálculo de métricas, limpieza de datos, etc.
Creación del dataset para el modelo de ML.

Flujo de trabajo típico en un equipo de ciencia de datos

- Generación de nuevas características
- Análisis exploratorio de datos
- Generación de información relevante para el negocio
- Definición de las arquitecturas a desarrollar
- Entrenamiento del modelo
- Definición de las métricas para evaluar el desempeño del modelo
- Versionado del modelo



Flujo de trabajo típico en un equipo de ciencia de datos



DATA ENGINEER

El data engineer es responsable de la **preparación y limpieza de los datos**, la creación de **pipelines de datos** y la integración de diferentes fuentes de datos. Su trabajo también incluye la selección de las herramientas y tecnologías adecuadas para la gestión de datos y la implementación de soluciones de **almacenamiento y procesamiento de datos escalables**.

DATA SCIENTIST

El data scientist se encarga de **definir y crear modelos de machine learning** que permitan hacer predicciones a partir de los datos. Su trabajo implica seleccionar los algoritmos adecuados, entrenar los modelos y optimizar su rendimiento. Los data scientists también pueden participar en la identificación de variables relevantes y en la **exploración de los datos para encontrar patrones y tendencias**.

DATA ANALYST

El data analyst trabaja con datos para descubrir patrones y tendencias que puedan ser útiles para la **toma de decisiones empresariales**. Su trabajo implica realizar análisis estadísticos y visualizaciones de datos para **entenderlos mejor** y hacer recomendaciones sobre cómo pueden utilizarse **para mejorar el negocio**.

MACHINE LEARNING ENGINEER

El machine learning engineer es responsable de llevar los modelos de machine learning a **producción** y asegurarse de que estén funcionando correctamente. Su trabajo implica seleccionar la infraestructura adecuada para el **despliegue de los modelos**, integrar los modelos con otras aplicaciones y sistemas, y **supervisar el rendimiento de los modelos**.

Roles y su alcance

DATA ENGINEER

El data engineer es responsable de la **preparación y limpieza de los datos**, la creación de **pipelines de datos** y la integración de diferentes fuentes de datos. Su trabajo también incluye la selección de las herramientas y tecnologías adecuadas para la gestión de datos y la implementación de soluciones de **almacenamiento y procesamiento de datos escalables**.

DATA SCIENTIST

El data scientist se encarga de **definir y crear modelos de machine learning** que permitan hacer predicciones a partir de los datos. Su trabajo implica seleccionar los algoritmos adecuados, entrenar los modelos y optimizar su rendimiento. Los data scientists también pueden participar en la identificación de variables relevantes y en la **exploración y análisis de datos**.

DATA ANALYST

El data analyst trabaja con datos para descubrir patrones y tendencias que puedan ser útiles para la **toma de decisiones empresariales**. Su trabajo implica realizar análisis estadísticos y visualizaciones de datos para **entender mejor los datos** y hacer recomendaciones sobre cómo pueden utilizarse **para mejorar el negocio**.

MACHINE LEARNING ENGINEER

El machine learning engineer es responsable de llevar los modelos de machine learning a **producción** y asegurarse de que estén funcionando correctamente. Su trabajo implica seleccionar la infraestructura adecuada para el **despliegue de los modelos**, integrar los modelos con otras aplicaciones y sistemas, y **supervisar el rendimiento de los modelos**.



Mientras que el Data Analyst trabaja con datos estructurados para identificar patrones y tendencias en los **datos existentes**, el Data Scientist se enfoca en construir **modelos predictivos** utilizando técnicas de aprendizaje automático y estadística para crear soluciones a problemas complejos.

Roles y su alcance

DATA ENGINEER

El data engineer es responsable de la **preparación y limpieza de los datos**, la creación de **pipelines de datos** y la integración de diferentes fuentes de datos. Su trabajo también incluye la selección de las herramientas y tecnologías adecuadas para la gestión de datos y la implementación de soluciones de **almacenamiento y procesamiento de datos escalables**.

DATA SCIENTIST

El data scientist se encarga de **definir y crear modelos de machine learning** que permitan hacer predicciones a partir de los datos. Su trabajo implica seleccionar los algoritmos adecuados, entrenar los modelos y optimizar su rendimiento. Los data scientists también pueden participar en la identificación de variables relevantes y en la **exploración de los datos para encontrar patrones y tendencias**.

DATA ANALYST

El data analyst trabaja con datos para descubrir patrones y tendencias que puedan ser útiles para la **toma de decisiones empresariales**. Su trabajo implica realizar análisis estadísticos y visualizaciones de datos para **entender mejor los datos** y hacer recomendaciones sobre cómo pueden utilizarse **para mejorar el negocio**.

MACHINE LEARNING ENGINEER

El machine learning engineer es responsable de llevar los modelos de machine learning a **producción** y asegurarse de que estén funcionando correctamente. Su trabajo implica seleccionar la infraestructura adecuada para el **despliegue de los modelos**, integrar los modelos con otras aplicaciones y sistemas, y **supervisar el rendimiento de los modelos**.

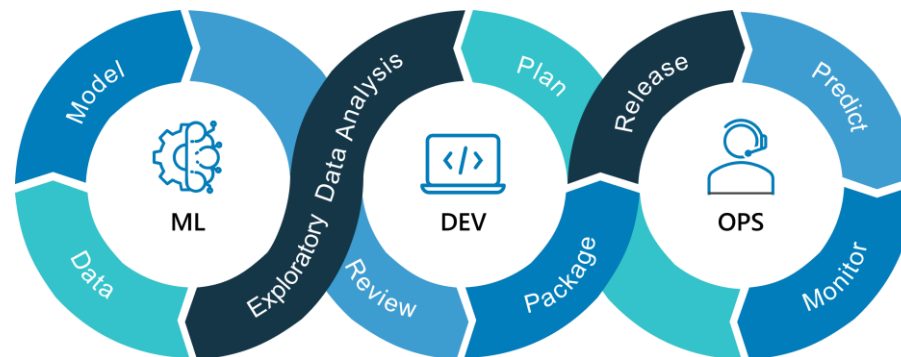
Roles y su alcance

¿Qué es MLOps?

MLOps

MLOps, o Machine Learning Operations, es un término que se refiere a las prácticas y herramientas utilizadas para gestionar y desplegar modelos de machine learning a gran escala en producción de manera efectiva y eficiente.

MLOps es una disciplina emergente que se enfoca en la gestión de los modelos de machine learning en producción y busca establecer procesos y herramientas para garantizar que los modelos de machine learning sean precisos, escalables y adaptables a diferentes situaciones.



¿Qué es producción?

ENTORNO DE DESARROLLO

Entorno donde comienzan a gestarse los proyectos, se realizan los primeros análisis exploratorios de datos y POCs.

Es un entorno donde podemos hacer **pruebas** sin miedo a que si nos equivocamos, afectemos un proceso crítico.

Debe ser lo más **parecido** posible **al entorno productivo**.

ENTORNO PRODUCTIVO

Entorno donde se ejecutan los procesos que ya fueron **validados por el negocio**.

Hay más tareas que se ejecutan de manera **automática**, por ejemplo: predicciones, tests unitarios sobre funciones, etc.

Es un entorno más **estable** que el de desarrollo.

Entornos de desarrollo y de producción

Desarrollo vs. producción

PROPÓSITO

Un entorno de desarrollo se utiliza para desarrollar y probar nuevas aplicaciones y funcionalidades. En cambio, un entorno de producción se utiliza para alojar aplicaciones y servicios que están siendo utilizados por los usuarios finales.

ESCALA

Un entorno de desarrollo se suele ejecutar en una sola máquina o en un pequeño grupo de máquinas, mientras que un entorno de producción suele tener múltiples máquinas y una mayor capacidad para manejar grandes volúmenes de datos.

CONFIGURACIÓN

En un entorno de desarrollo, la configuración es más flexible y menos rigurosa, y los desarrolladores pueden hacer cambios y ajustes según sea necesario. En un entorno de producción, la configuración es más rígida y estándar para garantizar la estabilidad y seguridad de los sistemas.

Desarrollo vs. producción

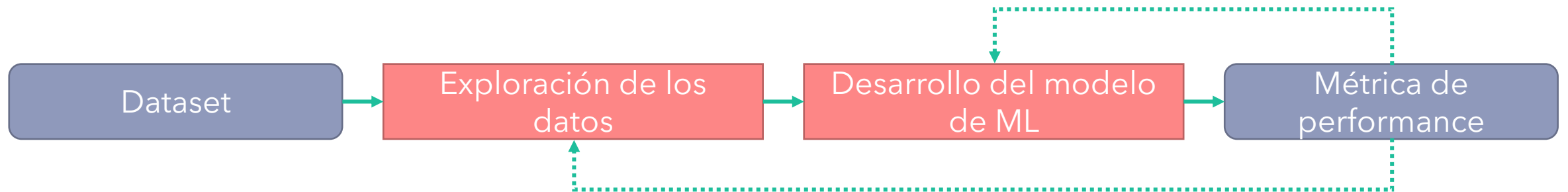
ACCESO

En un entorno de desarrollo, los desarrolladores tienen un acceso completo y libre para modificar y probar el sistema. En cambio, en un entorno de producción, el acceso se limita solo a aquellos usuarios que necesitan interactuar con el sistema para cumplir con sus roles y responsabilidades.

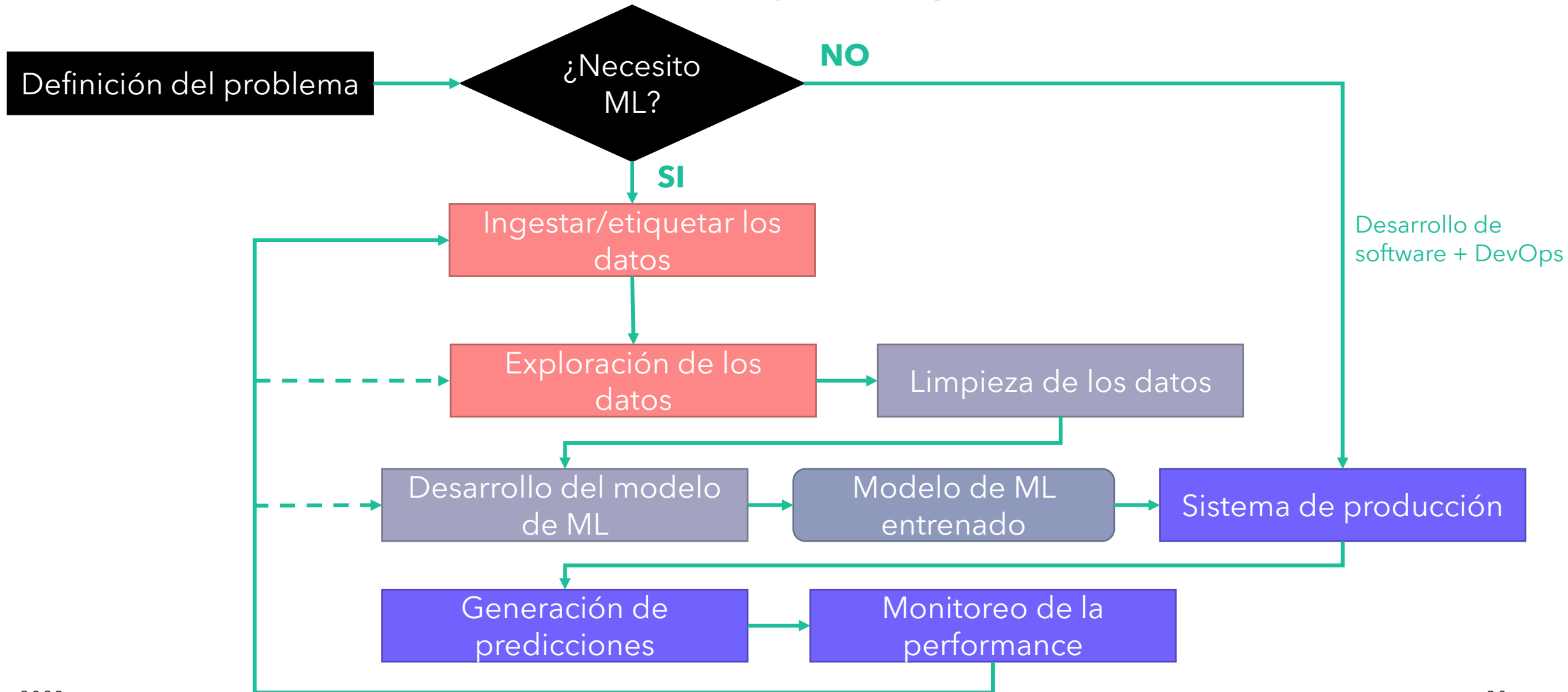
MANTENIMIENTO

En un entorno de desarrollo, los desarrolladores son responsables de mantener el sistema y corregir los errores que se encuentran durante el proceso de desarrollo y pruebas. En cambio, en un entorno de producción, el equipo de operaciones y soporte son responsables de mantener el sistema y corregir los errores en un ambiente de producción en vivo.

Enfoque académico/Competición de Kaggle



Machine learning en producción



Consideraciones para aplicaciones en la industria



Producción

Para que el modelo pueda entregar valor al negocio debe estar productivo.



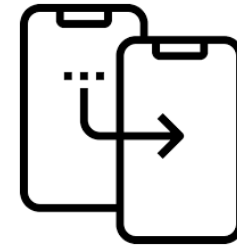
Usabilidad

Un modelo con 70% de accuracy en producción produce mucho más valor que uno con 100% de accuracy que no se puede usar.



Dependencia

Los modelos en producción requieren mantenimiento para prevenir el drift en los datos o en el target.



Escalabilidad

El proceso debe ser implementado para que otras personas del equipo lo entiendan, debe ser transparente y replicable.

Pipelines/flujos de trabajo reproducibles dentro de ML

¿Qué es un pipeline?

Los pipelines son una manera de organizar nuestro trabajo. Para ello dividimos el desarrollo general en secciones o módulos que se ejecutan de manera secuencial.

Esto nos permite **encapsular el código**, hacerlo más **legible**, más **ordenado**, **estandarizar** y **automatizar** los procesos, entre otros beneficios.

→ *Es una de las claves para lograr un desarrollo **escalable, eficiente y reproducible**.*

Los pipelines de machine learning permiten a los equipos de datos iterar rápidamente sobre diferentes modelos y ajustes y mejorar continuamente el rendimiento del modelo.

Los pipelines están conformados por **componentes** y por **artefactos** (artifacts).

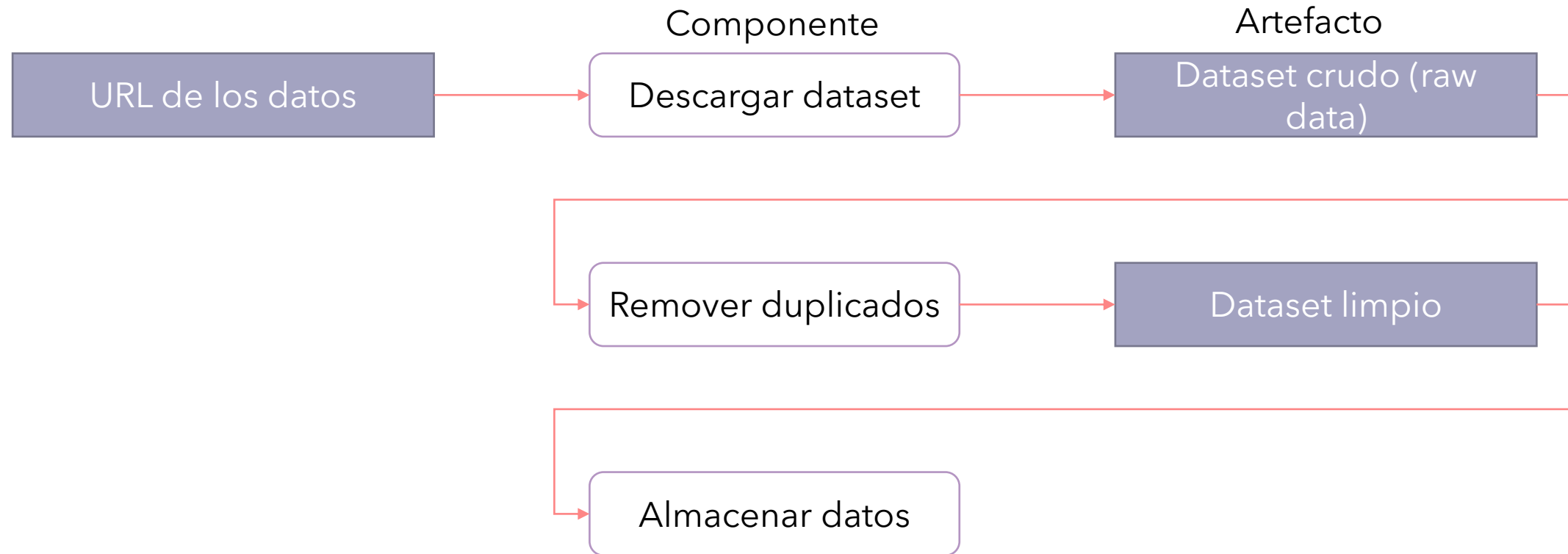
Componentes

Los componentes o pasos de un pipeline son piezas de código reutilizables y modulares que reciben una o varias entradas y producen una o varias salidas. Pueden ser scripts, notebooks u otro ejecutable.

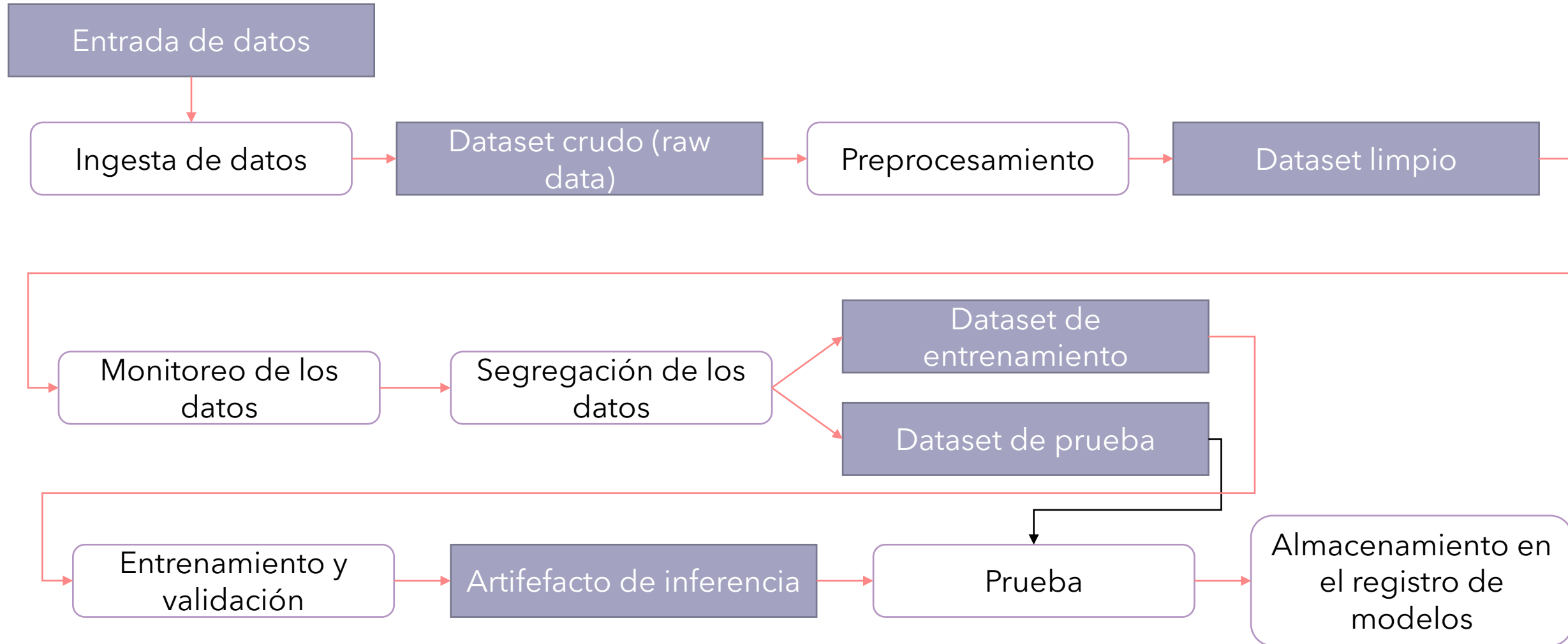
Artifact

Los artifacts son el resultado de los componentes, su salida. Estos pueden convertirse en la entrada de uno o más componentes para unir los distintos pasos de un pipeline. Los artifacts deben ser trackeados y versionados.

Ejemplo de pipeline de ETL



Ejemplo de un pipeline de entrenamiento



Niveles de MLOps

Los 3 niveles de MLOps

Frecuentemente dentro de la industria se pueden encontrar diferenciados tres niveles de MLOps. Estos niveles se diferencian en cuanto a la cantidad de herramientas/prácticas de MLOps que incluyen dentro de su funcionamiento.

- Nivel 0
- Nivel 1
- Nivel 2

Nivel 0 de MLOps

En este nivel no hay prácticas de MLOps en el proceso. Es adecuado para proyectos personales, cuando se está probando algún concepto/arquitectura nueva, para MVPs/POCs, etc.

En estos casos las ventajas de MLOps se dejan de lado debido a los tiempos de entrega o presupuestos destinados para esas etapas de desarrollo.

Características de esta etapa:

- **El código es monolítico:** se compone de uno o pocos scripts/notebooks que tienen una reusabilidad muy limitada.
- **El objetivo del desarrollo es el modelo y sus métricas,** no un pipeline de ML.
- **El foco del equipo no es la puesta en producción** del modelo, si se decide llevarlo a producción tal vez sea tarea de otro equipo de trabajo.
- **No hay conocimiento de la necesidad de monitoreo y reentrenamiento** del modelo.

Nivel 1 de MLOps

Este nivel de MLOps es importante cuando ya pasamos por la etapa de POCs y el equipo empieza a pensar en pasaje a producción, por lo que se deben considerar procesos más maduros para un desarrollo de software.

Características de esta etapa:

- **El objetivo de esta etapa es un pipeline de ML** que sea reproducible y que, por ejemplo, facilite el re-entrenamiento sobre nuevos datos.
- El pipeline es desarrollado con **componentes reutilizables**.
- El código, los artefactos y experimentos se comienzan a seguir (**tracking**) para generar **reproducibilidad** y **transparencia**.
- **La salida del pipeline de ML es un artefacto de inferencia** que contiene los pasos de preprocesamiento.
- Se incorpora el **seguimiento/monitoreo** del modelo en producción.

Nivel 1 de MLOps

Ventajas de implementar el Nivel 1

Con respecto a la implementación manual del nivel 0, al implementar el nivel 1 de MLOps podemos obtener las siguientes ventajas:

- Estandarización de los procesos
- Desarrollo más rápido de prototipos: reutilización de código
- Rapidez en llevar al mercado nuevos productos de datos
- Evitar model drift



Nivel 2 de MLOps

Este nivel de MLOps está pensado para compañías o proyectos de ML de gran escala, largo alcance y con un nivel de madurez muy avanzado. En esta etapa se cambia el foco del trabajo de desarrollar el pipeline de ML a mejorar sus componentes.

El nivel 2 de MLOps asume que ya se cuenta con múltiples pipelines de ML productivos y continúa aumentando el nivel de automatización aún más.

Características de esta etapa:

- **Integración continua (CI):** cada vez que un componente es modificado se ejecutan pruebas de integración para verificar que el componente funciona de la forma esperada.
- **Despliegue continuo (CD):** cada componente que pasa satisfactoriamente las pruebas es desplegado de manera automática y comienza a ejecutarse en producción como parte de los pipelines de ML.
- **Entrenamiento continuo:** cuando un componente cambia o cuando ingresan datos con nuevas distribuciones, se disparan las ejecuciones de los pipelines de entrenamiento y el proceso de CI/CD es ejecutado nuevamente.

Nivel 2 de MLOps

Ventajas de implementar el Nivel 2

Con respecto a la implementación del nivel 1, al implementar el nivel 2 de MLOps podemos obtener las siguientes ventajas:

- Iteración más rápida para llevar pipelines a producción
- Es más sencillo implementar A/B testing sobre los cambios
- El trabajo cooperativo entre grandes equipos de personas se facilita
- Se comienza a trabajar en la mejora continua del proceso productivo.



Comparación entre los distintos niveles de MLOps

	Objetivo	Re-entrenamiento	Equipo de trabajo	Aplicación	Producción	Reutilización	Infraestructura	Dificultad
Nivel 0	Model	Difícil, manual	1-5	POCs	NO	NO	Poca	Fácil
Nivel 1	Pipeline	Fácil, manual o mediante un disparador	1-20	Pequeña/mediana escala	SI	SI	Intermedia	Media
Nivel 2	Pipeline	Fácil, automático	+10	Mediana/gran escala	SI	SI	Mucha	Difícil

Para comenzar a pasar a producción uno de nuestros modelos de aprendizaje automático, el código debe cumplir con ciertos estándares de buenas prácticas de programación.

Buenas prácticas de programación



Buenas prácticas de programación


¿Qué vamos a ver?

- Cómo escribir código limpio y de manera modular
- Refactorización de código
- Optimización
- Cómo documentar el código
- Estándar PEP8 y Linting

Buenas prácticas de programación

Código limpio

Cuando nuestro código va a ser potencialmente usado en producción, debe cumplir con ser **legible**, **simple** y **conciso**.



"Uno puede observar que existe gran presencia de nubes que tapan el cielo azul y como consecuencia de ello hay una alta probabilidad de que esta tarde hayan precipitaciones en forma de agua."



"El cielo está nublado y es probable que a la tarde llueva."