

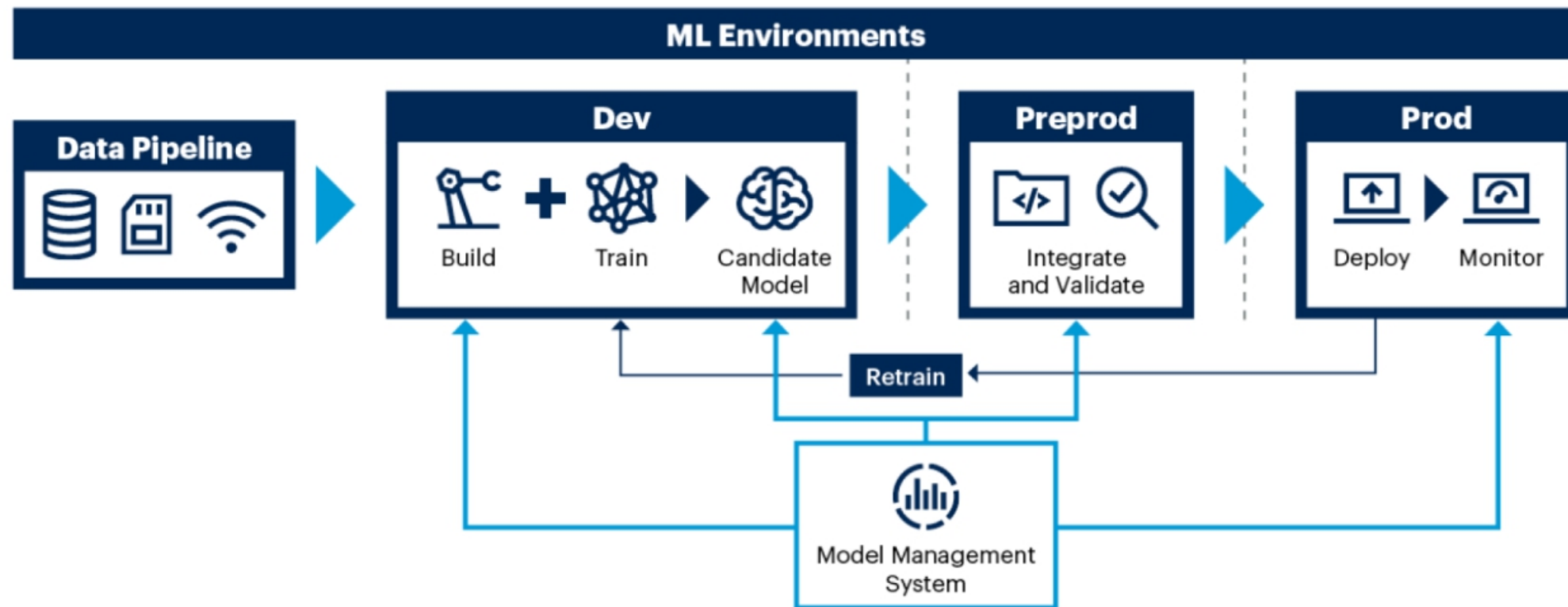
Полный цикл проекта по машинному обучению

Елена Кантонистова

ВШЭ, 2023

Схема проекта по машинному обучению

Typical ML Pipeline



Source: Gartner

718951_C

Схема проекта по машинному обучению



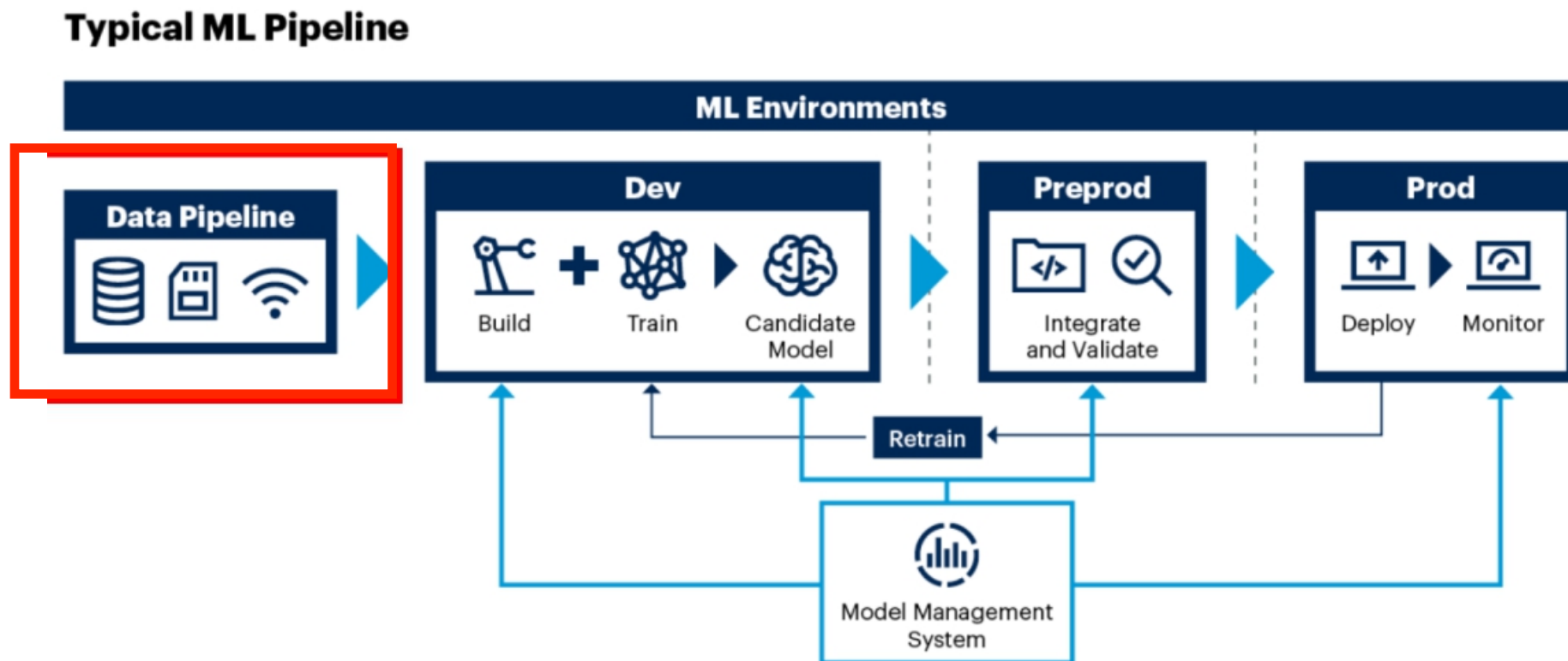
1. Постановка задачи
2. Работа с данными
3. Обучение и валидация модели
4. Тестирование модели на новых пользователях
5. Внедрение модели и мониторинг
6. Оркестрация процессов

1. Постановка задачи



- Что нужно сделать ?
- Какие есть данные ?
- Где хранятся данные ?
- Какие метрики качества решения ?
- Когда и в каком виде предоставить решение ?
- Какие технологии необходимо использовать в проекте ?

2. Этап работы с данными



Source: Gartner

718951_C

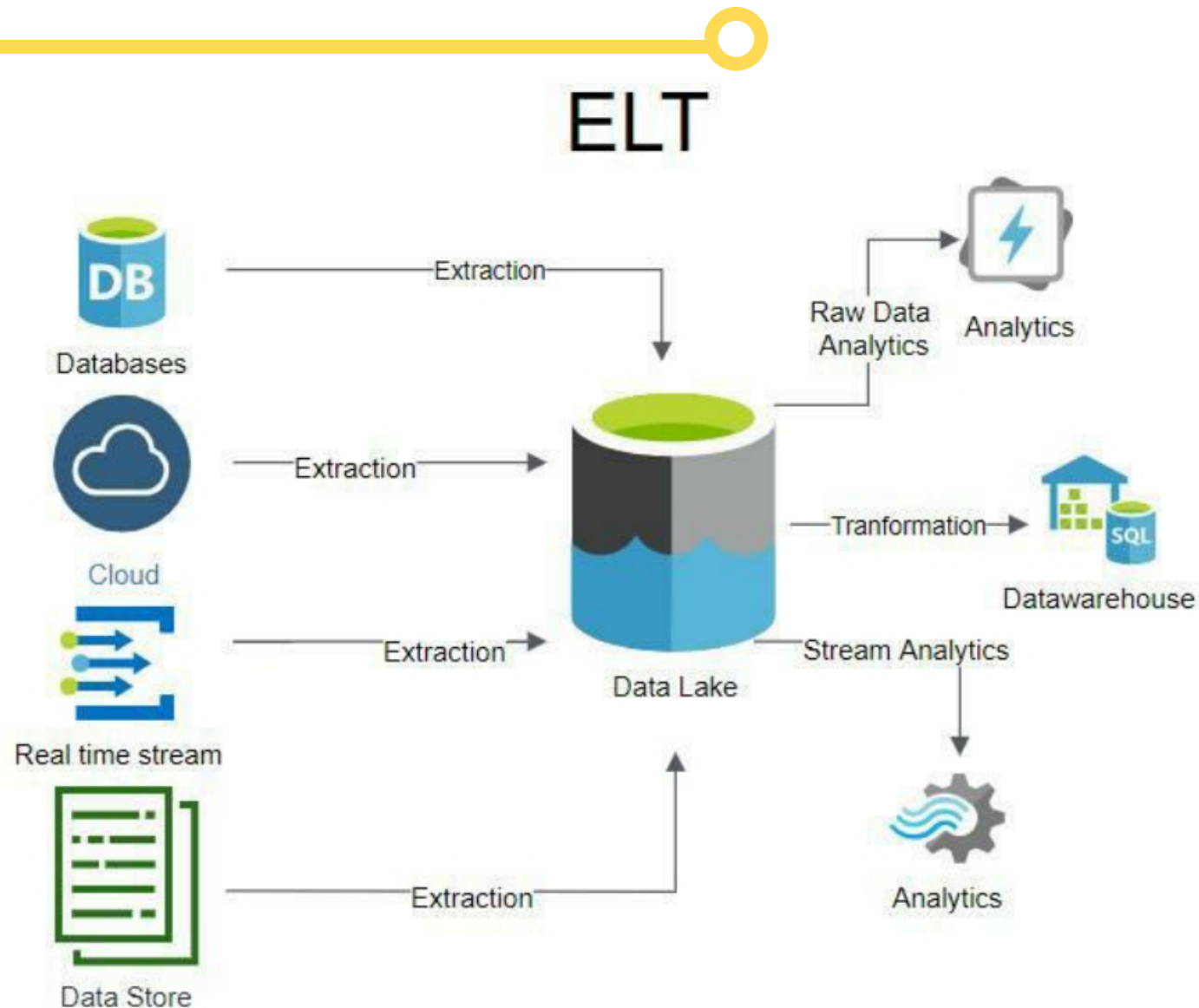
2. Этап работы с данными

1. *Сбор данных* : в каких источниках хранятся данные ? Есть ли к ним доступы ?
2. *Обработка данных* :
 - Проверка качества данных
 - Очистка данных
 - Feature engineering
 - Агрегация данных
3. *Загрузка данных в хранилище*

2. Этап работы с данными

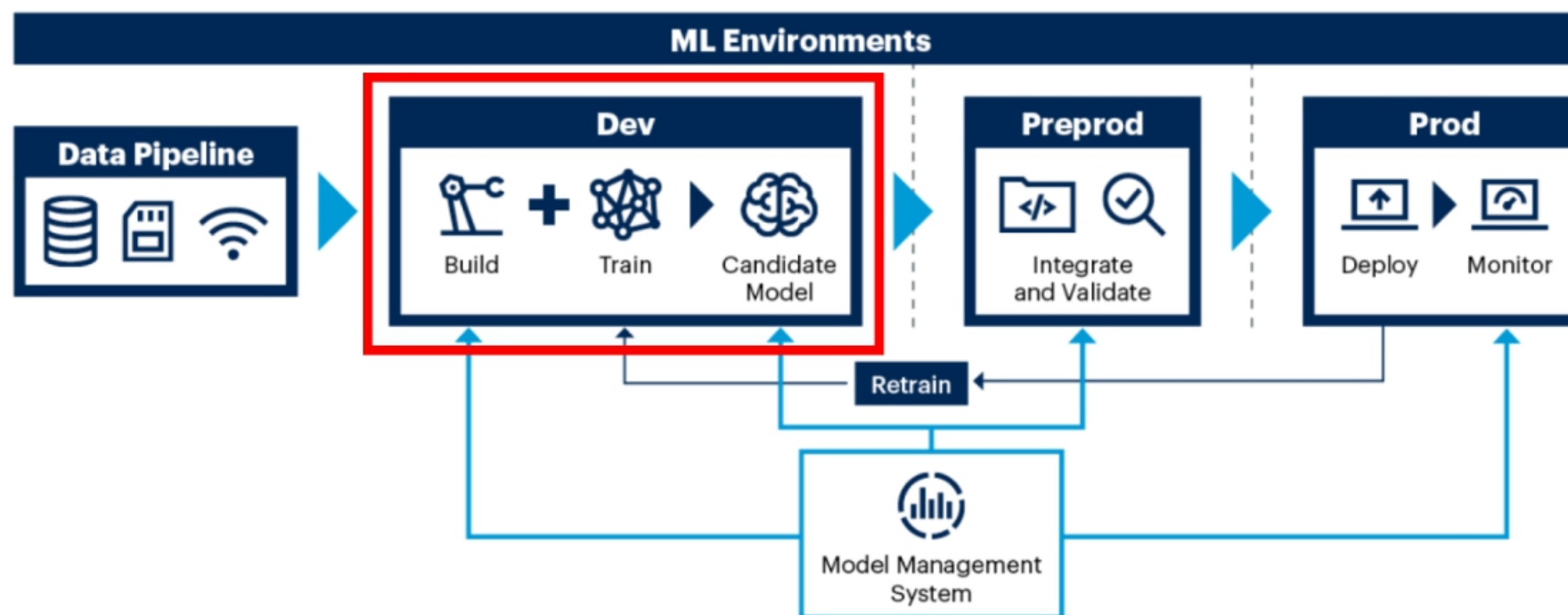


2. Этап работы с данными



3. Обучение и валидация модели

Typical ML Pipeline



Source: Gartner

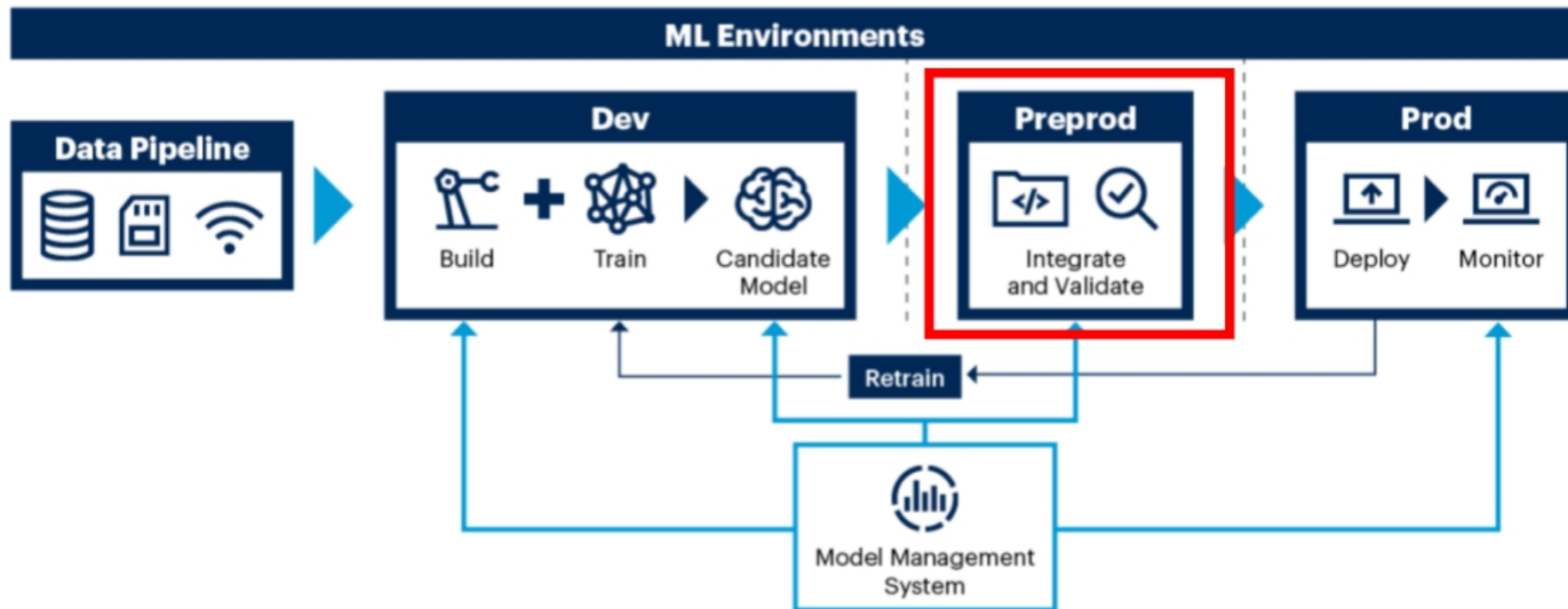
718951_C

3. Обучение и валидация модели

1. *Выбор модели* (линейные модели, деревья, бустинги , нейронные сети)
2. *Обучение модели*
3. *Валидация модели* (оценка качества модели на тестовых данных)
4. *Подбор гиперпараметров модели*
5. *Выбор наилучшей модели*

4. Тестирование модели

Typical ML Pipeline

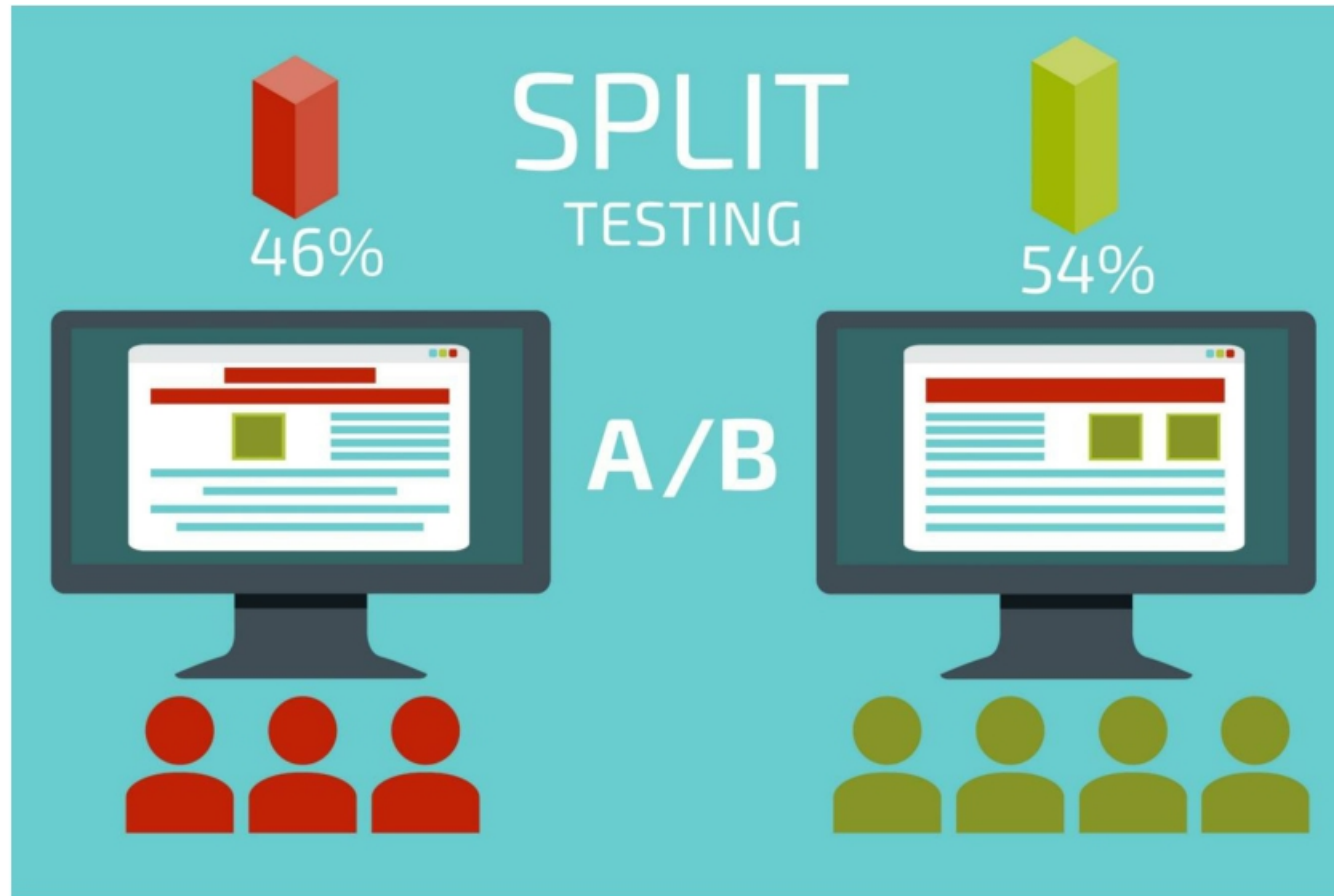


Source: Gartner

718951_C

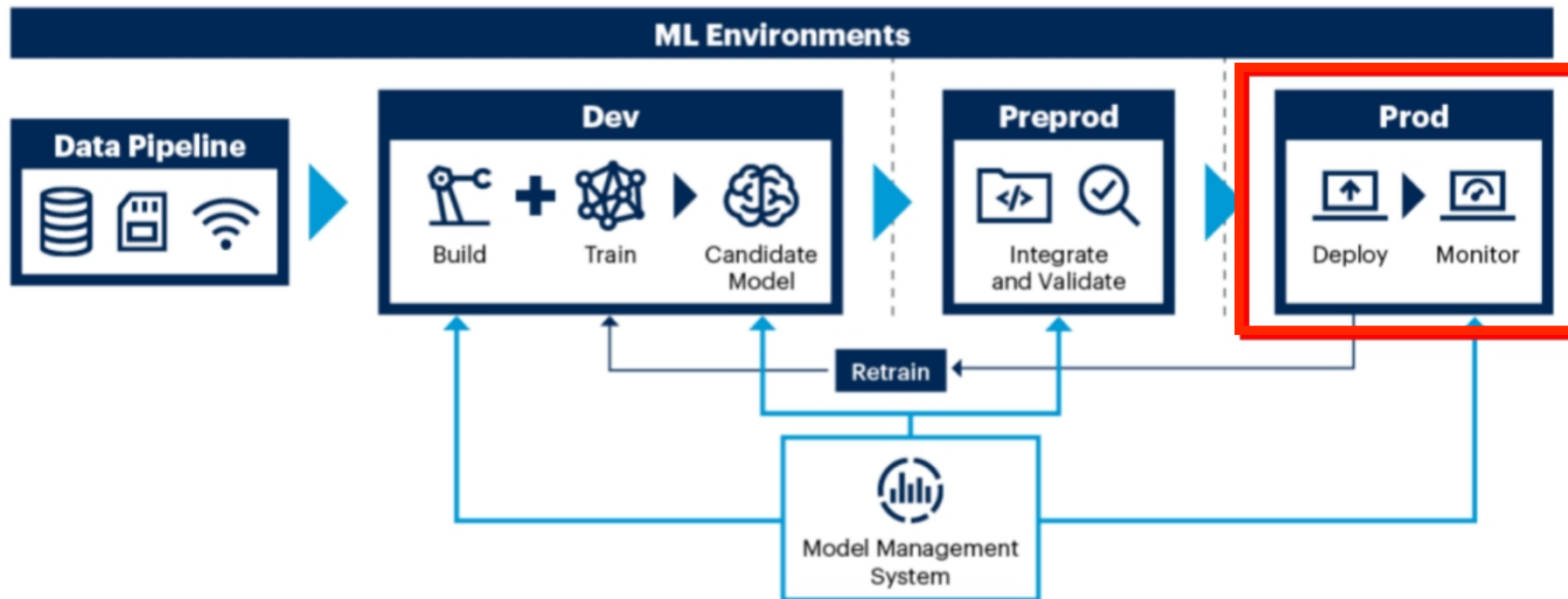
4. Тестирование модели

A/B- тестирование модели на НОВЫХ ПОЛЬЗОВАТЕЛЯХ



5. Внедрение модели и мониторинг

Typical ML Pipeline



Source: Gartner

718951_C

5. Внедрение модели и мониторинг

Внедрение модели :

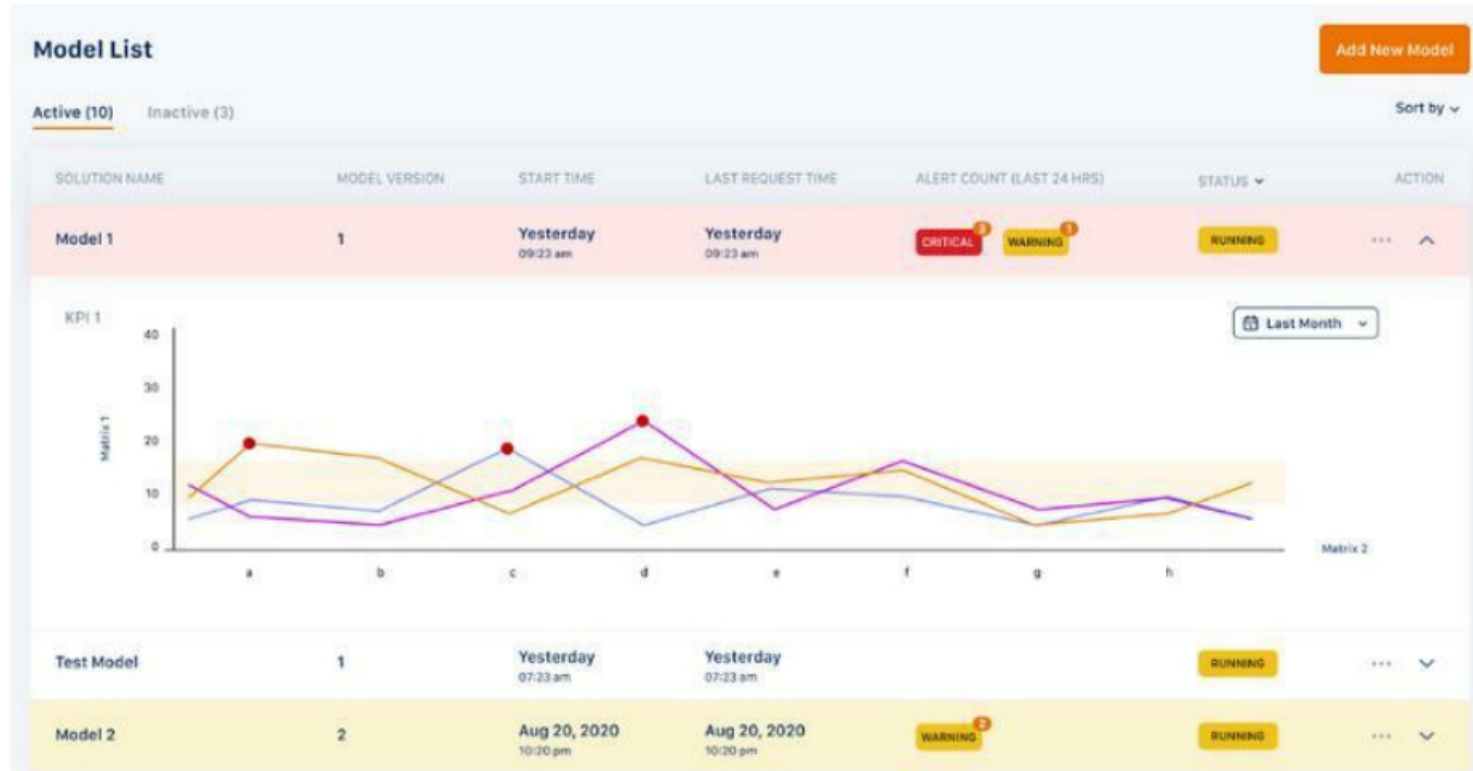
в зависимости от бизнес-целей это может быть

- сервис, применяющий модель ([пример](#)) - например FastApi
- телеграм-бот с моделью
- использование специальных serving-инструментов (например, [Seldon](#))

5. Внедрение модели и мониторинг

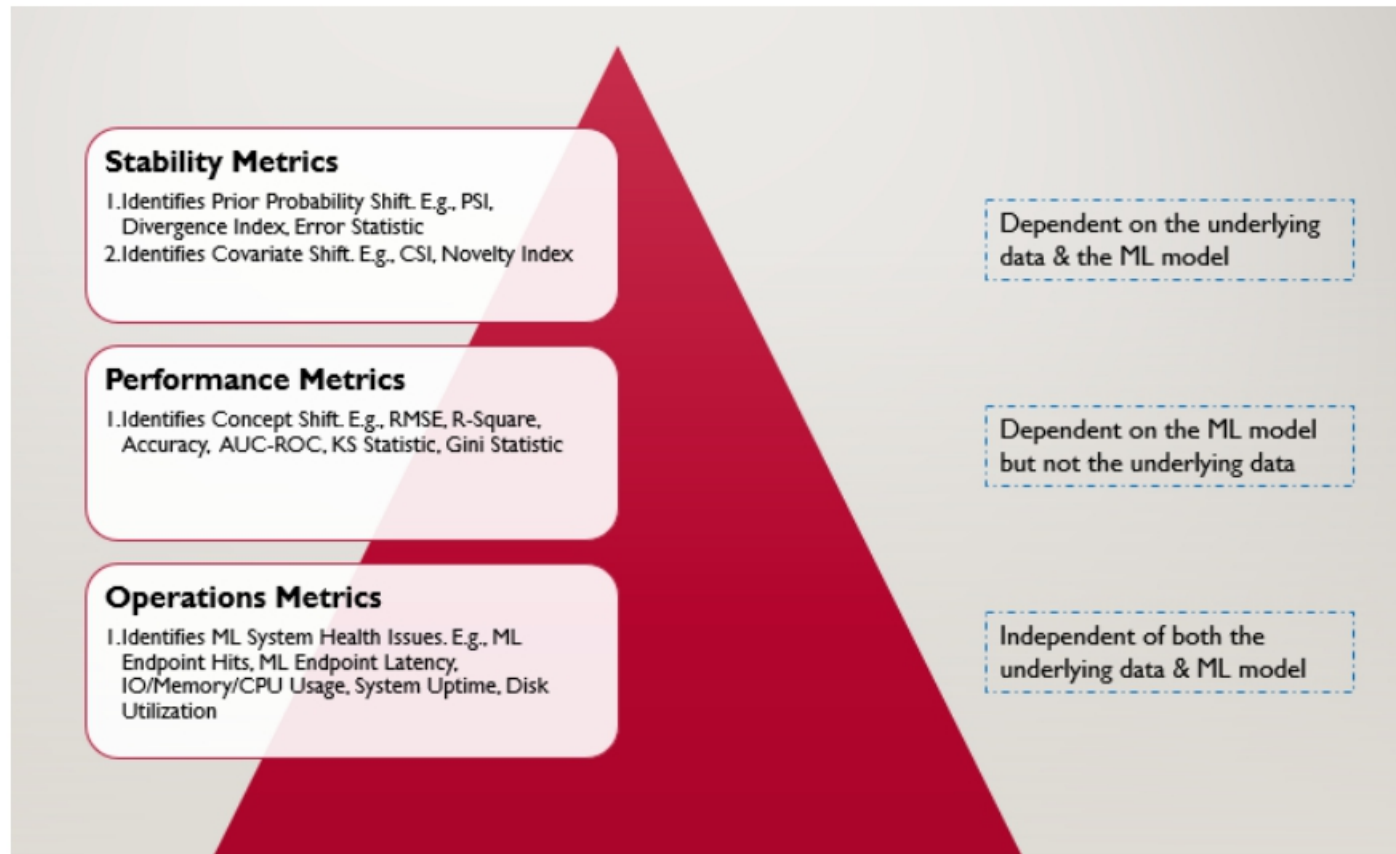
Мониторинг модели

- Цель состоит в том, чтобы отслеживать модели по различным метрикам

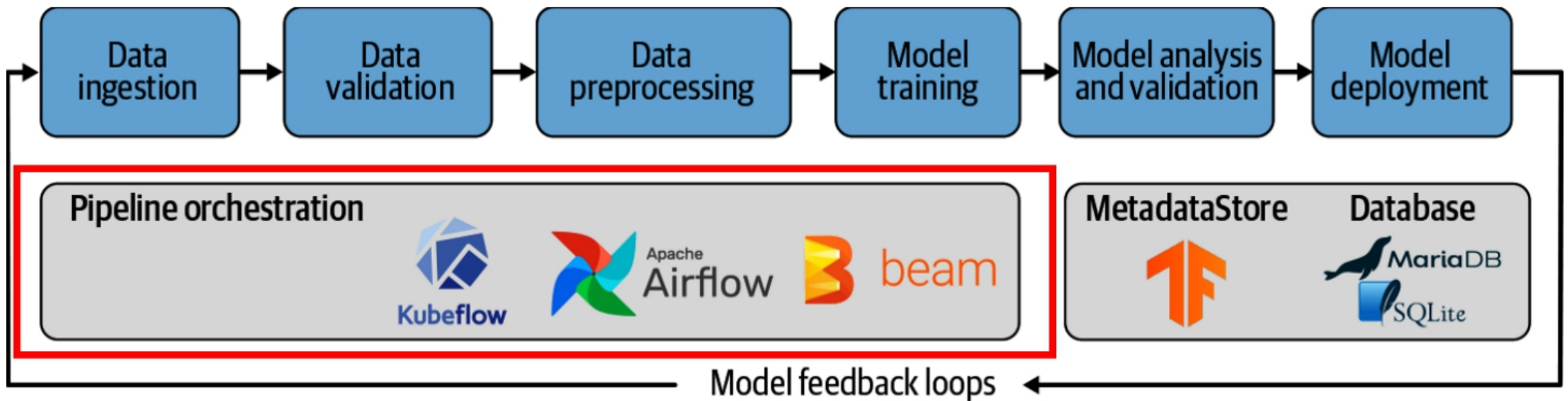


5. Внедрение модели и мониторинг

Мониторинг модели — какие показатели измеряем :



6. Оркестрация процессов



Оркестрация пайплайна



Оркестрация

Будем использовать инструмент Apache Airflow.

С его помощью можно:

- Запланировать регулярные запуски пайплайна
- Оценивать успешность выполнения шагов пайплайна и их время



Планирование времени запуска

- Регулярный запуск пайплайна осуществляется при помощи Cron.
- Формат времени в Cron:



[перевод времени онлайн в Cron](#)