

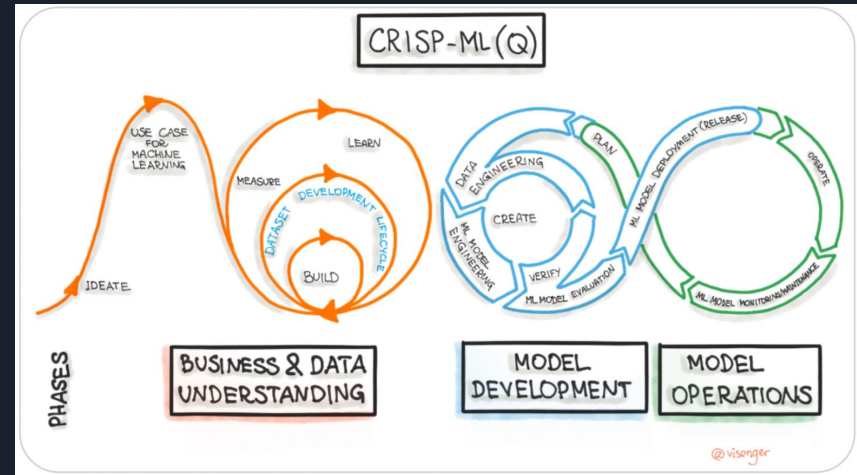


MLOps Платформа

Задача ML

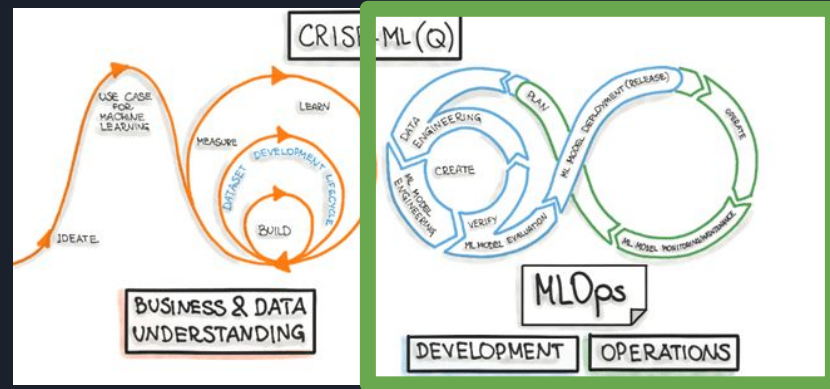
Задача ML в соответствии с методологией CRISP-ML (Cross Industry Standard Process for Machine Learning).

- Понимание бизнеса (Business Understanding) — определение целей и требований проекта с точки зрения бизнеса.
- Понимание данных (Data Understanding) — сбор и анализ данных для понимания их структуры и качества.
- Подготовка данных (Data Preparation) — очистка, трансформация и отбор данных для последующей работы.
- Моделирование (Modeling) — выбор и обучение моделей машинного обучения.
- Оценка (Evaluation) — оценка качества модели и проверка, отвечает ли она требованиям.
- Внедрение (Deployment) — развертывание модели в производственную среду.



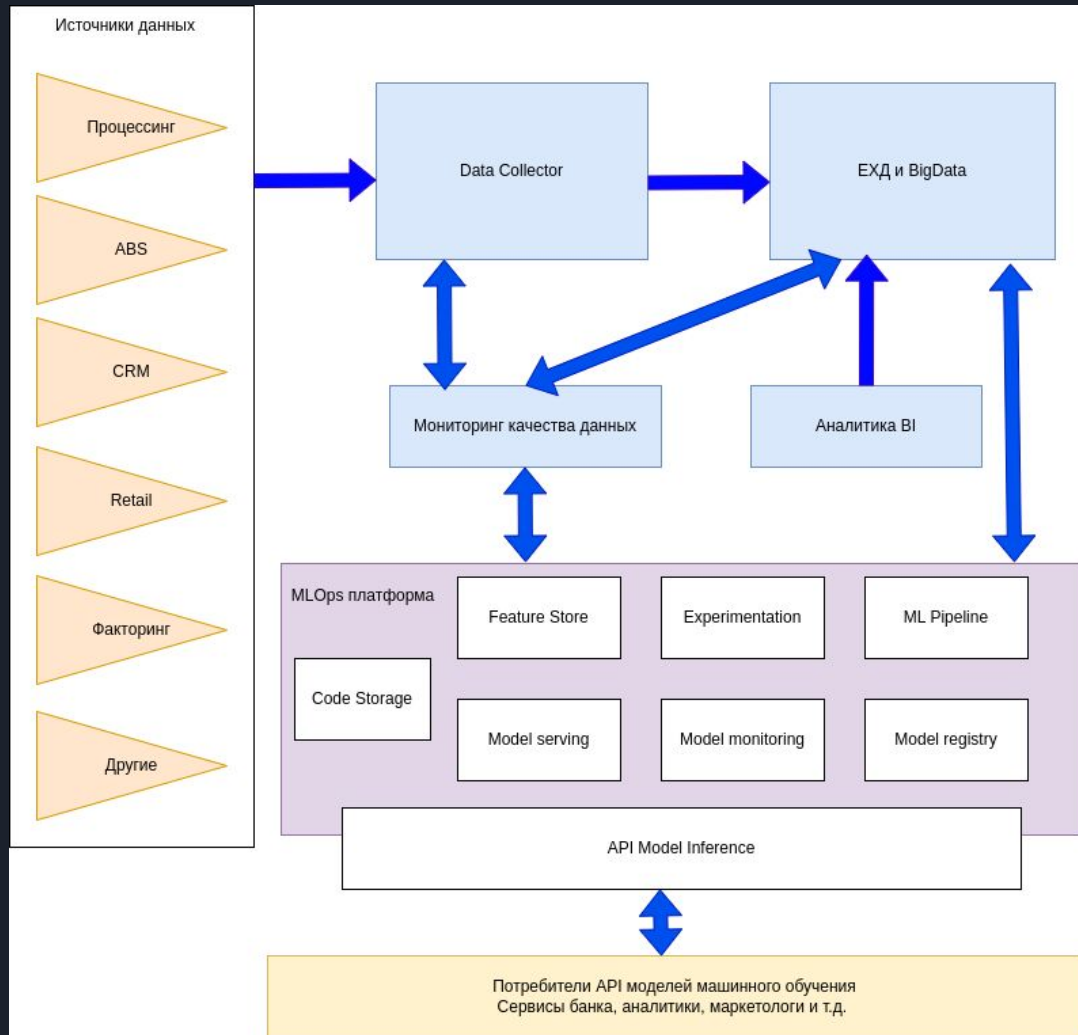
Задачи, которые решает MLOps

- Автоматизация разработки и внедрения моделей
- Управление вычислительными ресурсами
- Аутентификация и авторизация
- Мониторинг и логирование
- Управление данными и признаками
- Развертывание моделей
- Обеспечение надежности и безопасности
- Управление версиями моделей
- Масштабирование
- Документирование процессов
- Оценка и улучшение моделей
- Обратная связь и оптимизация моделей



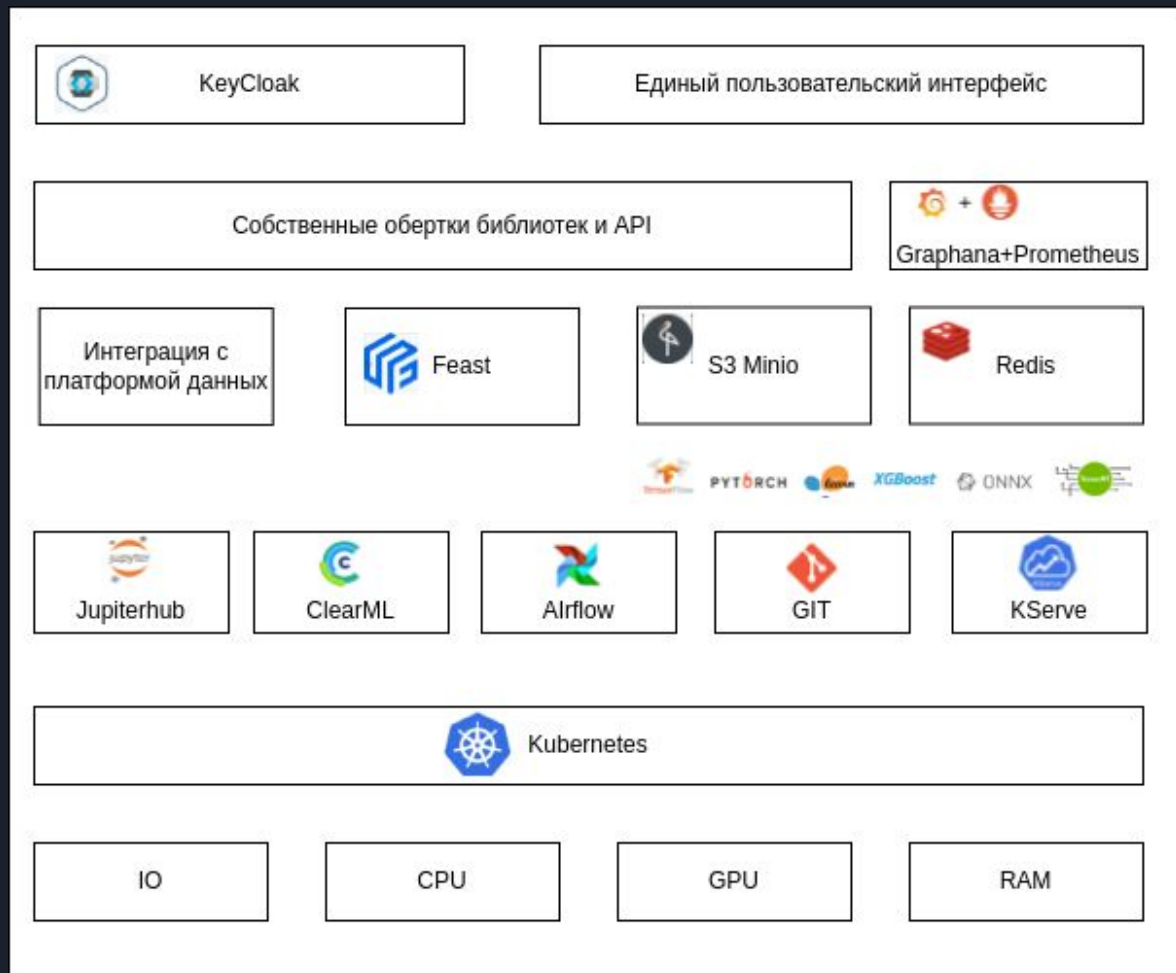
MLOps платформа в системе банка

MLOps платформа в системе банка автоматизирует развертывание, мониторинг и управление моделями машинного обучения, обеспечивая их стабильную и безопасную работу. Она интегрируется с хранилищем данных (DWH) и предоставляет потребителям API моделей машинного обучения.



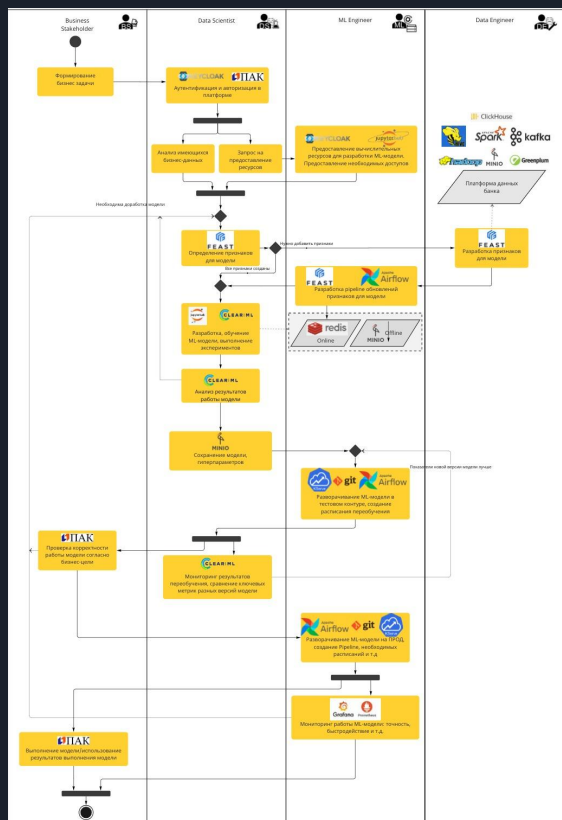
Архитектура MLOps платформы

MLOps платформа, интегрируя лучшие open-source решения, такие как JupyterHub, ClearML и KServe, дополнена собственными кастомными обертками для библиотек и API, предоставляет единый пользовательский интерфейс, обеспечивает гибкое разделение прав доступа и эффективное управление вычислительными ресурсами (CPU, GPU, RAM, IO) в облачной среде Kubernetes. Платформа выделяет ресурсы для разработки в разрезе каждого пользователя и поддерживает четко очерченные контуры для разработки, тестирования и эксплуатации.



Ожидаемые результаты создания MLOps платформы

- **Внедрение современных методов машинного обучения:**
 - Использование передовых алгоритмов и технологий для повышения точности и эффективности моделей.
- **Создание условий для быстрого развёртывания моделей (модель как сервис):**
 - Сокращение времени на интеграцию и масштабирование новых моделей.
- **Внедрение CI (Continuous Integration) для автоматизации процессов разработки и развёртывания:**
 - Обеспечение стабильной и беспроблемной интеграции кода для быстрого выпуска новых версий моделей.
- **Использование современных инструментов и технологий (Kubernetes для развёртывания моделей на Python):**
 - Облегчение управления развёртыванием и масштабированием, обеспечение надёжной инфраструктуры.
- **Создание Feature Store для эффективного управления признаками для моделей:**
 - Стандартизация и повторное использование признаков, ускорение подготовки данных и повышение качества моделей.
- **Улучшение качества данных и моделей:**
 - Тщательное управление признаками и автоматизация проверок данных повысят качество входных данных и, соответственно, точность моделей.
 - Создание среды разработки для анализа результатов экспериментов и выбора наилучшего решения.





Архитектура MLOPs платформы

JupyterHub – это мощная платформа для многопользовательского доступа к Jupyter Notebooks. В контексте MLOps, она позволяет управлять и предоставлять совместные среды для разработки и анализа данных. JupyterHub интегрируется с различными аутентификационными системами и масштабируется на кластеры, что делает его идеальным для использования в корпоративных сетях и образовательных учреждениях.

ClearML – это платформа для управления экспериментами, оркестрации задач и мониторинга в реальном времени. Она помогает отслеживать все этапы жизненного цикла машинного обучения, от подготовки данных до развертывания моделей. ClearML облегчает реплицируемость и анализ экспериментов, интегрируясь с различными инструментами и библиотеками, что делает процесс разработки моделей прозрачным и управляемым.

KServe (бывший KFServing) – это Kubernetes-нативная платформа для развертывания и управления моделями машинного обучения в продакшене. Она поддерживает масштабирование, автообновления и мониторинг моделей, предоставляя единую платформу для работы с различными фреймворками. KServe упрощает процесс развертывания моделей, обеспечивая высокую производительность и надежность.

MinIO – это высокопроизводительное, масштабируемое объектное хранилище, совместимое с Amazon S3 API. В контексте MLOps MinIO используется для хранения больших объемов данных и моделей машинного обучения. Его можно легко интегрировать с различными инструментами и платформами для обеспечения надежного и надежного хранения данных, что критично в процессе разработки, тестирования и развертывания моделей.



Архитектура MLOPs платформы

Apache Airflow – это платформа для оркестрации рабочих процессов, которая позволяет планировать, запускать и мониторить сложные потоки данных и вычислений. В MLOps, Airflow используется для автоматизации разработки, обучения и развертывания моделей, обеспечивая управление зависимостями и очередностью выполнения задач.

Feast (Feature Store) – это платформа для управления признаками (features) машинного обучения. Она позволяет централизованно хранить, обслуживать и использовать признаки как для обучения моделей, так и для их применения на этапах предсказания. Feast интегрируется с различными источниками данных и MLOps платформами, обеспечивая непрерывный поток данных высокого качества для моделей машинного обучения.

Redis – это высокопроизводительная база данных NoSQL в памяти, используемая для кэширования и временного хранения данных. В MLOps Redis можно использовать для хранения промежуточных данных и метаданных, что позволяет ускорять процессы обучения и предсказания моделей. Благодаря низкой задержке и высокой скорости доступа, Redis становится незаменимым инструментом для задач, требующих моментального отклика.

Grafana – это открытая платформа для аналитики и мониторинга, которая визуализирует и исследует метрики в реальном времени. В MLOps Grafana используется для мониторинга производительности и состояния моделей машинного обучения, отображения метрик эксплуатации и анализа потенциальных проблем. Интеграция с такими системами, как Prometheus, позволяет создавать наглядные и информативные дашборды для мониторинга жизненного цикла моделей.



Архитектура MLOPs платформы

Prometheus – это система мониторинга и оповещения с открытым исходным кодом, оптимизированная для сбора и хранения временных рядов данных. В контексте MLOps Prometheus используется для мониторинга системных ресурсов, производительности моделей и других метрик в режиме реального времени. В сочетании с Grafana он позволяет создавать наглядные дашборды и настраивать расширенные механизмы оповещения для оперативного реагирования на сбои и отклонения.

Keycloak – это решение с открытым исходным кодом для управления идентификацией и доступом (IAM), которое позволяет реализовать единый вход (SSO), аутентификацию и авторизацию для различных приложений и сервисов. В контексте MLOps Keycloak используется для обеспечения безопасности и управления доступом к ресурсам и сервисам машинного обучения.

Эти инструменты в комплексе создают надежную инфраструктуру для полного жизненного цикла моделей машинного обучения, от разработки до развертывания и мониторинга, помогая обеспечить эффективность, масштабируемость и управляемость процессов MLOps.