

ML, КОТОРЫЙ РАБОТАЕТ ПО-ВЗРОСЛОМУ

О КОМПАНИИ

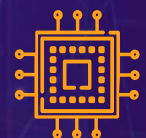
«Неофлекс» создает ИТ-платформы для цифровой трансформации бизнеса, помогая заказчикам получать устойчивые конкурентные преимущества.

Мы фокусируемся на заказной разработке программного обеспечения и внедрении сложных информационных систем, используя передовые технологии и подходы.

Наш отраслевой опыт и технологическая экспертиза, усиленная собственными акселераторами разработки, позволяют решать бизнес-задачи любого уровня сложности.



ОСНОВАНА
В 2005



4 ЦЕНТРА РАЗРАБОТКИ



500+
СОТРУДНИКОВ

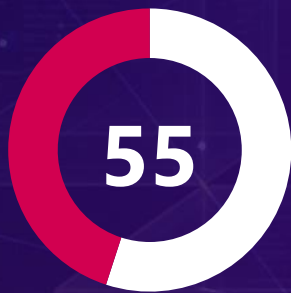
НАШИ ЗАКАЗЧИКИ

9 из 10

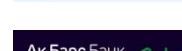
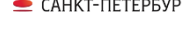
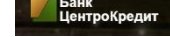
Заказчиков продолжают сотрудничество с «Неофлекс» после первого проекта



Зарубежных банков
из TOP-100



Российских банков
из TOP-100



ПАРТНЁРЫ





Wine Quality Model

Модель классификации

Каждому вину присваивается оценка «качества» от 0 до 10:

- «хорошего качества» (оценка 7 или выше)
- «плохого качества» (оценка ниже 7)

Качество вина будем определять по 11 переменным:

- Фиксированная кислотность
- Летучая кислотность
- Лимонная кислота
- Остаточный сахар
- Хлориды
- Свободный диоксид серы
- Общий диоксид серы
- Плотность
- pH
- Сульфаты
- Алкоголь





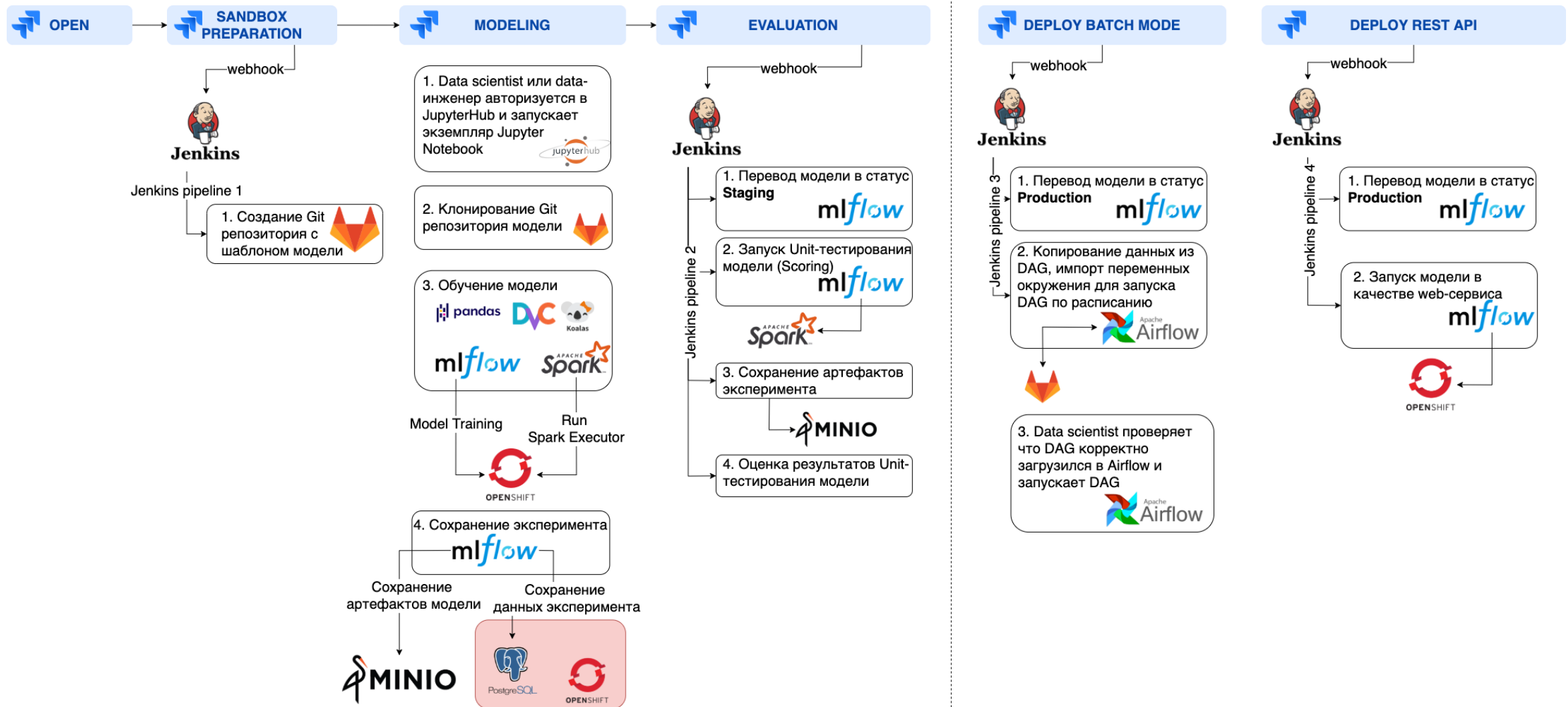
MLOps

Инструменты и процессы для непрерывной разработки
и внедрения моделей машинного обучения

АЛЕКСЕЙ ЖАДАН

Ведущий специалист DevOps

Процесс MLOps



Многопользовательский Jupyter Notebook сервер с открытым исходным кодом

- Работает как в облаке, так и на ноутбуке специалиста по обработке данных или на виртуальной машине
- Подходит для обслуживания различных пользовательских интерфейсов (Jupyter, Rstudio, nteract и др.)
- Доступ пользователей по протоколам аутентификации (Oauth, GitHub и т.д.)
- Может быть развернут с помощью контейнерных технологий, работает на Kubernetes



Платформа с открытым исходным кодом для управления жизненным циклом ML-моделей.

- Позволяет специалистам data science работать с любой ML-библиотекой (H2O, keras, mlearn, PyTorch, SKLearn, TensorFlow)
- Встроенный реестр для хранения и версионирования моделей
- Возможность создавать переиспользуемые и воспроизводимые модели
- Запуск, отслеживание и визуализация экспериментов, использующих Python, REST, R или JAVA APIs



Инструменты MLOps

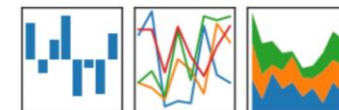
Python библиотека для обработки и анализа данных.

- Быстрый и эффективный DataFrame объект для манипуляции с данными, встроено индексирование
- Инструменты для чтения и записи данных между in-memory структурами данных и различных форматов: CSV, тестовые файлы, Microsoft Excel, БД SQL, HDF5
- Встроенная обработка недостающих данных, выравнивание данных на основе меток и вычислений
- Возможность объединять, совмещать и группировать данные
- Работает с временными рядами
- Оптимизирован для высокой производительности



pandas

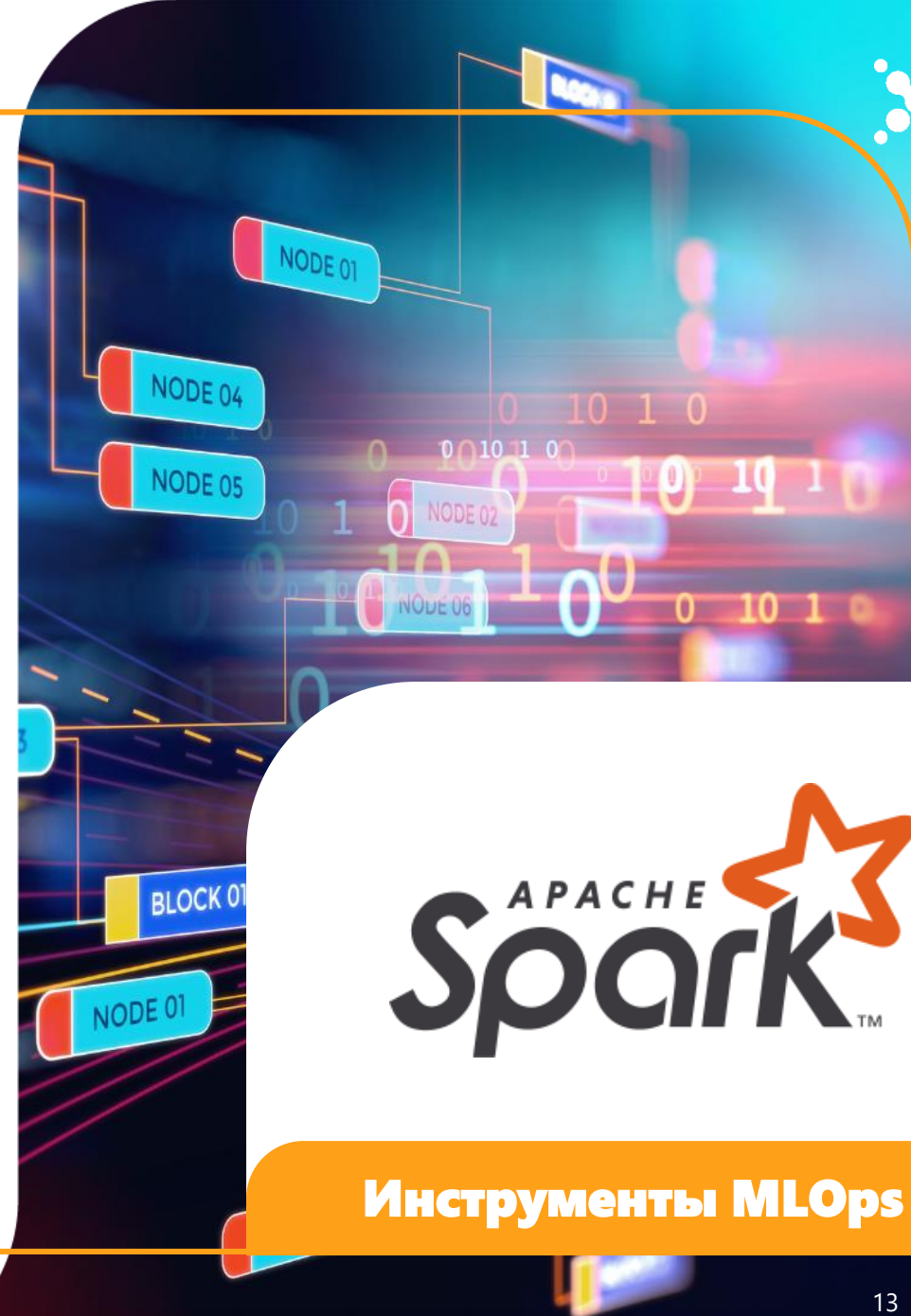
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Apache Spark

Инструмент потоковой обработки неструктурированных и слабо-структурированных данных в режиме реального времени

- Работает в парадигме in-memory вычислений (быстрая обработка данных в оперативной памяти)
- Дополняет возможности Hadoop, может выступать его заменой (способен выполнять пакетную обработку)
- Возможность многократного доступа к загруженным в память пользовательским данным
- Предоставляет API для Java, Scala, R и Python (через инструмент PySpark)
- Открытый исходный код



Инструменты MLOps

Apache PySpark

- Позволяет работать с RDD (устойчивый распределенный набор данных) на языке Python
- Предоставляет PySpark Shell для связи Python API с ядром Spark, инициализирует контекст Spark



Инструменты MLOps

Инструмент интеграции pandas DataFrame API поверх Apache Spark

- Единая кодовая база (поддержка распределенных наборов данных Spark и тестов, datasets pandas)
- Работает с pandas и PySpark



Koalas

Инструменты MLOps

Apache Airflow

Платформа для создания, планирования и мониторинга рабочих процессов

- Создает рабочие процессы (Airflow pipelines) в виде направленных ациклических графов
- Динамическая генерация Airflow pipelines в виде кода на Python (можно написать алгоритм, который будет динамически создавать экземпляры пайплайнов)
- Веб-интерфейс для визуализации пайплайнов, отслеживания прогресса и устранения неполадок
- Параметризация сценариев пайплайна встроена в ядро Airflow с помощью движка шаблонов Jinja
- Масштабируемость (модульная архитектура, использует очередь сообщений для оркестрации рабочими узлами)



Инструменты MLOps

Data Lake (озеро данных) – хранилище данных, в котором разные данные хранятся в неупорядоченном, неструктурированном и необработанном виде.

- Огромное хранилище
- Принимает любые данные (картинки, видео, звук. Файлы. Документы) всех форматов из любого источника (CRM, ERP, продуктовые каталоги, банковские программы и т.д.)
- Информация структурируется на выходе из «озера»
- Анализ данных не влияет на сами данные в «озере»
- Гибкость (можно записывать любые данные, не перестраивая «озеро»)
- Подходят для хранения неочищенной информации, которая может пригодиться с будущем

S3 совместимое программно-определяемое хранилище данных для проектов машинного обучения.

- Защита данных с помощью встроенного стирающего кодирования (позволяет восстанавливать целые пакеты данных в случае их потери)
- Защита от деградации/распада данных (Bitrot защита)
Поврежденные данные не считываются, а захватываются и восстанавливаются на лету с помощью алгоритма HighwayHash
- Поддержка различных схем шифрования (AES-256-GCM, ChaCha20-Poly1305, AES-CBC.), совместим с хранилищами ключей (например, HashiCorp Vault)
- Управление идентичностью
- Непрерывная репликация и другие функции

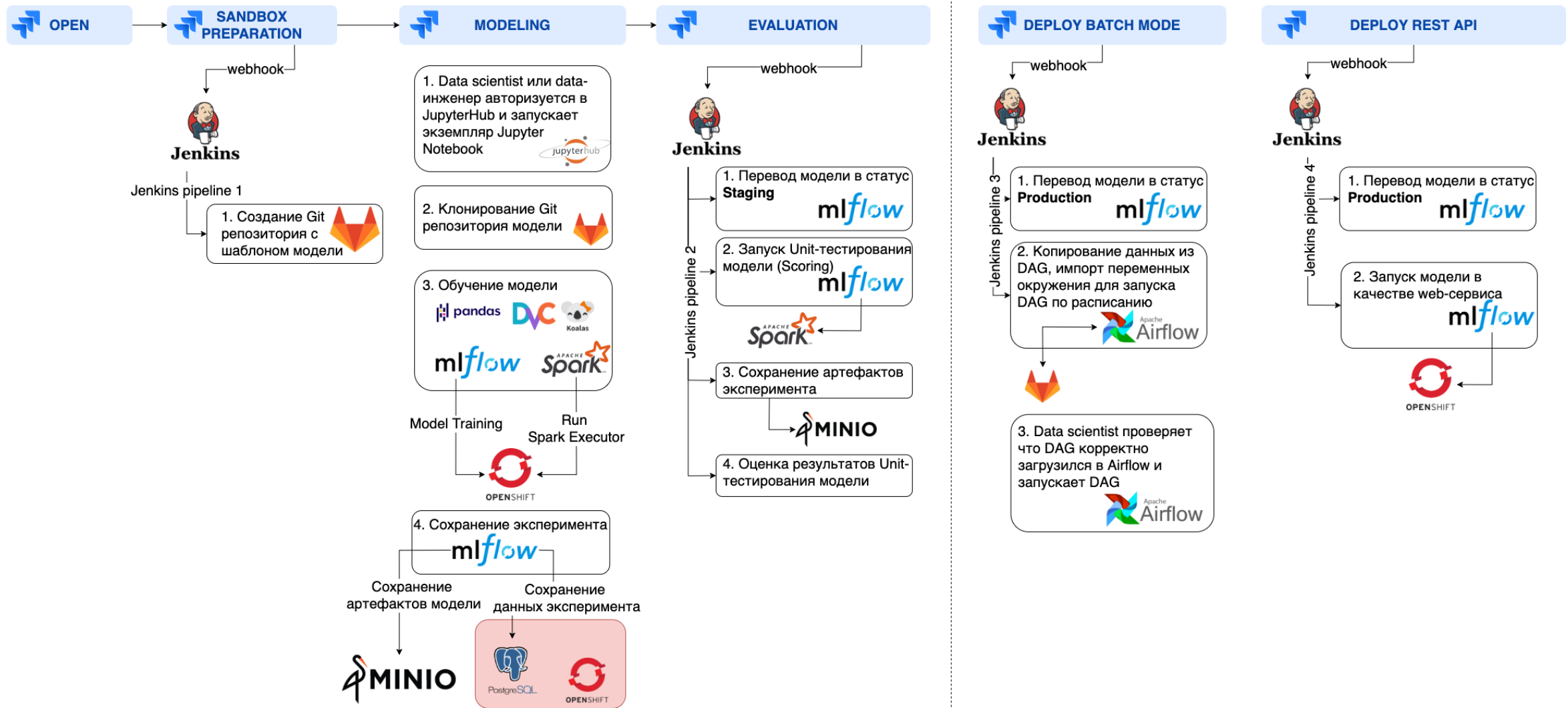


Open-source система контроля версий для проектов машинного обучения. Возможности:

- Git-совместимая система (GitHub, Gitlab и др.)
- Можно использовать любой тип хранилища (Amazon S3, Microsoft Azure Blob Storage, Google Drive, Google Cloud Storage, Aliyun OSS, SSH/SFTP, HDFS, HTTP б дисковое или сетевое)
- Гарантирует воспроизводимость экспериментов
- Поддерживает low friction ветвление (Git ветвление даже с большими файлами, данные не дублируются)
- Отслеживает метрики и неполадки
- Поддержка Hive, Spark jobs
- Подходит для работы с разными языками программирования и файлами (Python, R, Julia, Scala Spark, custom binary, Notebooks, flatfiles/TensorFlow, PyTorch и др.)



Процесс MLOps





ПОЧЕМУ «НЕОФЛЕКС»?

- Быстрый старт и сжатые сроки разработки
- Снижение затрат на разработку и обслуживание
- Использование всех преимуществ передовых IT-технологий

ПОЧЕМУ МЫ ДЕЛАЕМ БЫСТРО И КАЧЕСТВЕННО?

- Понимаем бизнес-процессы и обладаем глубокой технологической экспертизой
- Применяем самые современные технологии (микросервисы, Big Data, AI, machine learning) и процессы (DevOps, Agile)
- Используем готовые программные компоненты – акселераторы разработки



ГЕОГРАФИЯ



19

СТРАН

80+

ЗАКАЗЧИКОВ