



Владимир Новоженев,
руководитель продуктового
направления Data Platform, VK Tech

VK Data Lakehouse

Единый контур данных для быстрой
и дешевой промышленной
AI/ML-разработки



 **Cnews**
FORUM КЕЙСЫ

Эффективное внедрение ML Ops

Для эффективного построения ML Ops циклов требуется понимание бизнес-целей



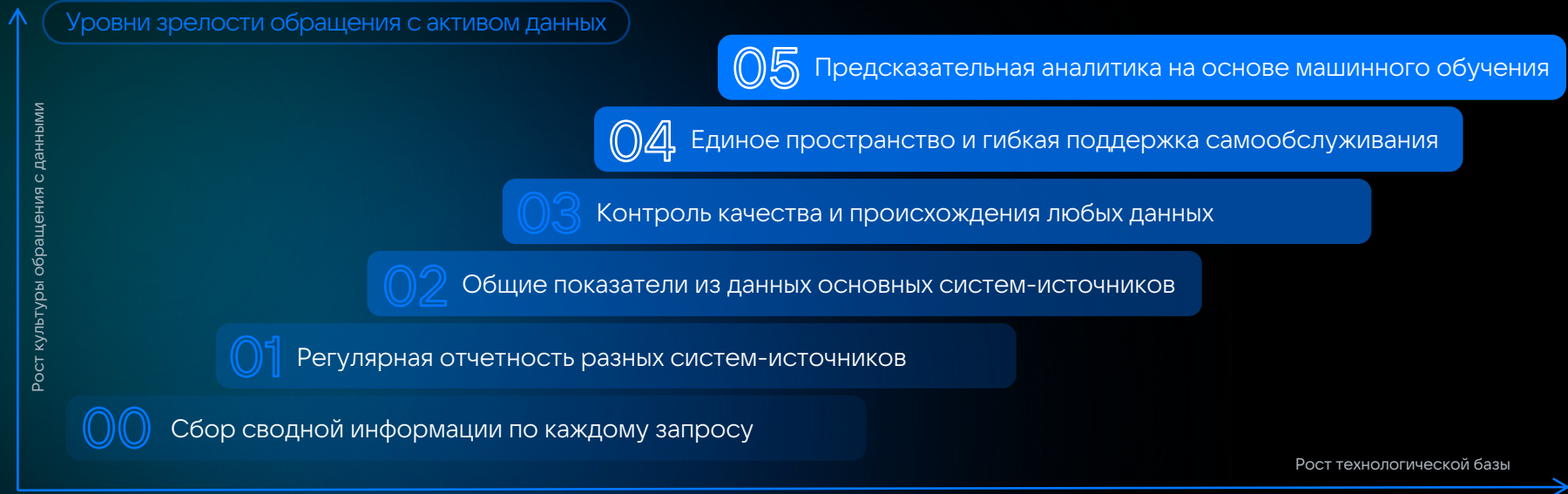
Необходимо определить ключевые бизнес-задачи, которые нужно решать с помощью ML

И оценить финансовый эффект



Необходимо убедиться, что все участники процесса (Data Scientists, инженеры, DevOps, бизнес-команда) понимают цели и согласованы между собой

Принципы обращения с данными для построения ML Ops



доступность
для бизнеса



обеспечение
достоверности



своевременность



повторяемость
использования



скорость
разработки

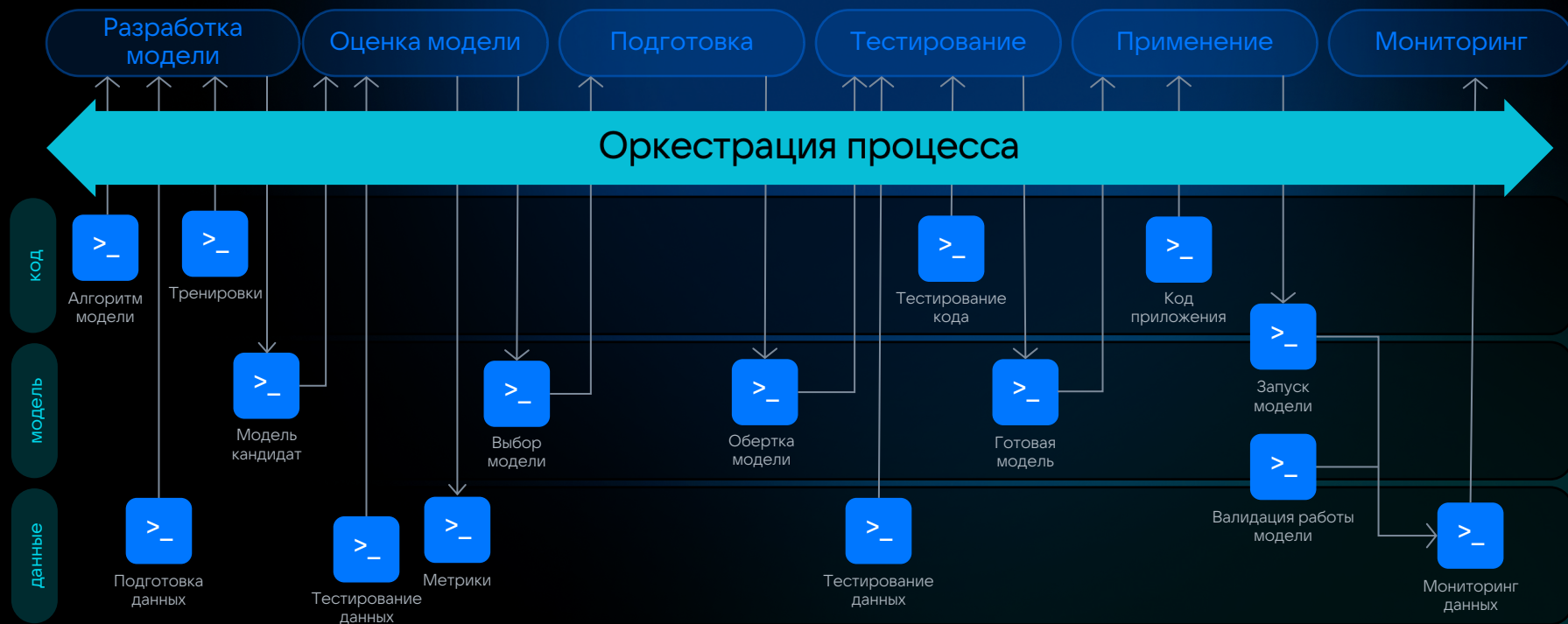


оптимизация
хранения и обработки



надежность
и безопасность

ML Ops — Карта процесса



Что тормозит развитие AI и ML?



Разрозненные источники данных (data silos)



Дорогие DWH-решения



Медленные и сложные ETL-процессы



Зависимость от централизованных ИТ-ресурсов



Деградация lake в data swamp:
нет схем, нет контроля качества,
нет доверия



Отсутствие компетенций

Результат



Медленный time-to-market



Высокая стоимость владения



Низкая масштабируемость
и качество моделей

Как решать эти проблемы?



Навести порядок в бизнесе

Это не к нам

Найти компетенции

Это уже ближе



Навести порядок
в инфраструктуре

А вот это по-нашему

Единая платформа – основа продуктов данных

Проблематика

Разнородность платформ
и хранилищ у направлений
и сложность получения
доступа к данным

01

Ручной ввод
и трансформация данных,
множество данных в excel

02

Отсутствие
прослеживаемости
происхождения данных
и истории изменений

03

Пропуски, выбросы,
противоречивость
данных
в источниках

04

Сложность
в идентификации владельцев
данных
и разные методики

05

Отсутствие
технологической базы
для реализации новых
расчетных задач
и машинного обучения

06

Цифровая Платформа - общий технологический фундамент для эффективной работы с данными и выстраивания сквозной аналитики в предприятии

1

Создание общего достоверного источника данных, унификация используемых технологических компонент сбора, хранения и обработки данных

2

Единые инструменты
и подходы к управлению
данными и их качеством

Доступная инфраструктура
и инструменты разработки
информационно-расчетных
задач и машинного обучения

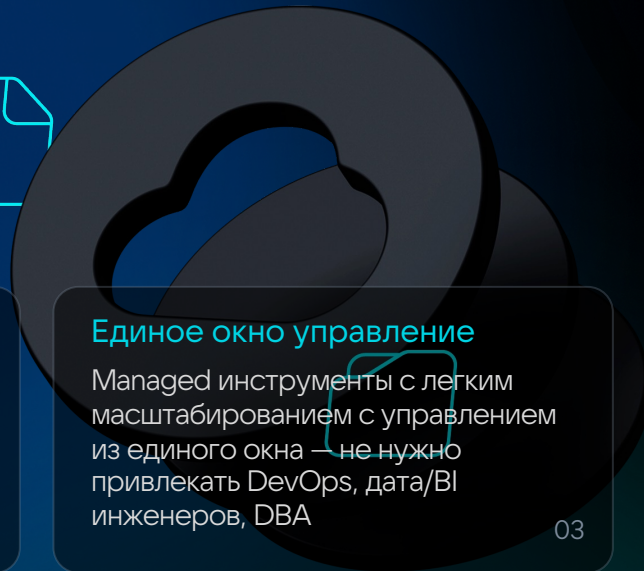
3





VK Data Lakehouse

для порядка в инфраструктуре →



Хранение в S3

Дешевле хранение на порядок, в отличие от DWH — ML инициативы будут лучше окупаться

01

ACID для S3-хранилища

Iceberg для версионности, ACID и каталогизации — озеро не превратится в болото

02

Единое окно управления

Managed инструменты с легким масштабированием с управлением из единого окна — не нужно привлекать DevOps, дата/BI инженеров, DBA

03

Работа напрямую с любыми источниками данных

Trino напрямую обращается к различным источникам без сложных ETL трансформаций — быстрее запускаем модели

04

Разделение storage и compute

Меньше расходов за счет отдельной оплаты. Меньше конфликтов с командой DWH за ресурсы и стабильность хранилища.

05



ML-сценарии использования VK Data Lakehouse

Финансовый сектор и банки

Скоринг и антифрод

- Объединение транзакционных логов, клиентских профилей и внешних источников
- Обучение и валидация моделей скоринга и детекции мошенничества на «живых» данных
- Использование time-travel в Iceberg для воспроизводимости моделей

Результат

Сокращение времени на подготовку данных с недель до часов, повышение точности моделей за счет доступа к полным историческим данным



Ритейл и Ecom

Персонализированные рекомендации и динамическое ценообразование

- Интеграция данных о покупках, поведении на сайте, логистике и отзывах
- Быстрое формирование фич с помощью Spark/Trino
- Реализация full-cycle ML: от фичей до inference через MLflow

Результат

Рост конверсии и среднего чека, автоматизация принятия решений в real-time



Промышленность и логистика

Предиктивное обслуживание оборудования

- Сбор телеметрии с датчиков, событий SAP/MES и техобслуживания
- Построение моделей прогнозирования отказов и оптимального времени ТО
- Визуализация и реакция через BI-инструменты на Trino

Результат

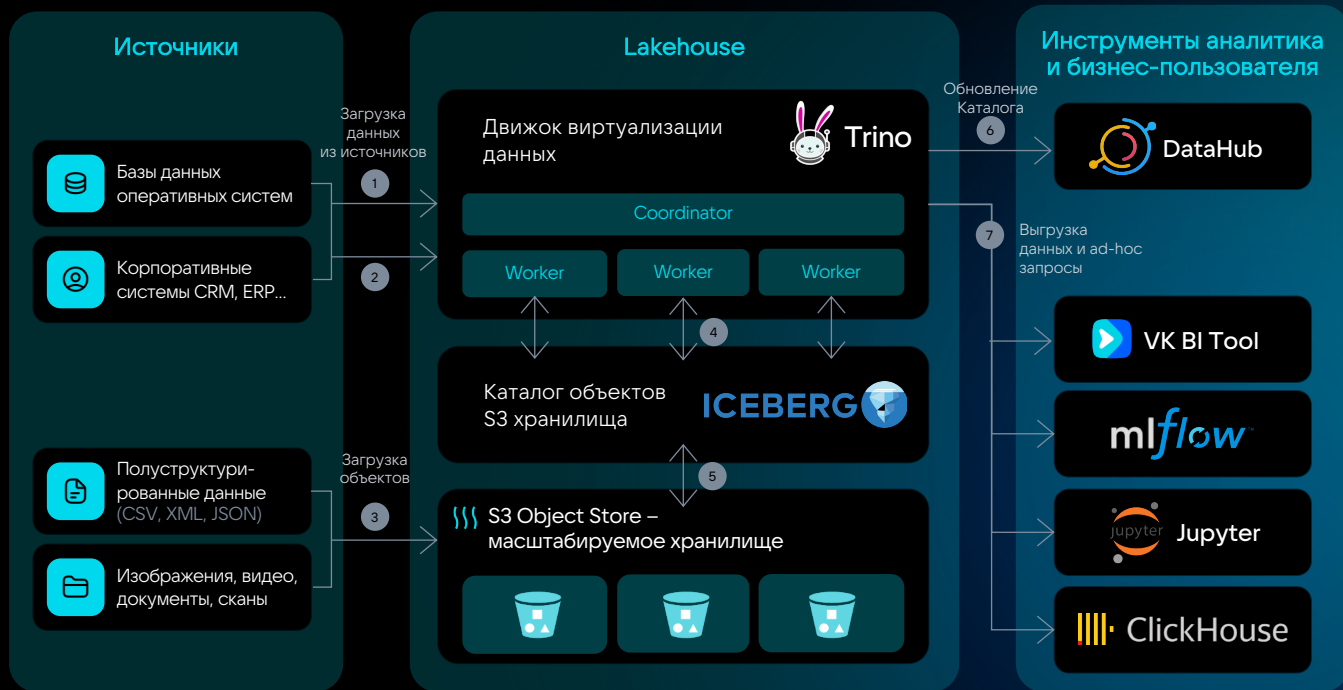
Снижение простоев, увеличение ресурса оборудования, экономия на ремонт



Data LakeHouse как основа Хранилища Данных



Реализация Lakehouse в VK Data Platform



- 1 Структурированные данные поступают из источников в Trino
- 2 Источники могут запрашиваться напрямую через Trino
- 3 Неструктурированные данные поступают потоком или блоками в хранилище
- 4 Trino через Iceberg читает и сохраняет данные в объектном хранилище
- 5
- 6 Информация об источниках и витрине публикуется в каталоге данных
- 7 Trino и Iceberg предоставляют данные для отчетов, аналитики BI, машинного обучения напрямую или через промежуточные хранилища

Trino

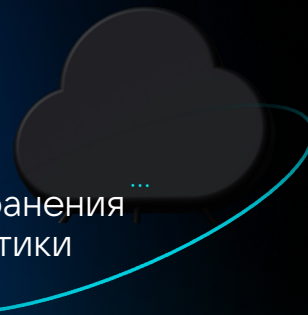
движок виртуализации данных:
• параллельного исполнения SQL запросов
ко множеству источников и хранилищ

• перемещения данных ETL/ELT для
формирования Lakehouse в объектном S3
хранилище

VK Data Lakehouse тестируют и внедряют

в 10 раз

Уменьшение стоимости хранения
данных и ускорение аналитики



Крупный ритейлер

Public



VK Data Lakehouse

Сценарий

Импортозамещение
Oracle Exadata.
Инсталляция 10 ТБ

Тестирование

Крупная IT-компания

Public



VK Data Lakehouse

Сценарий

Переезд с Hadoop.
Инсталляция 300 ТБ

Внедрение

Нефтедобывающая
компания

On-Premise



VK Data Lakehouse

Сценарий

Импортозамещение SAP.
Инсталляция 100 ТБ

Внедрение

VK Data Lakehouse доступен в любом формате поставки

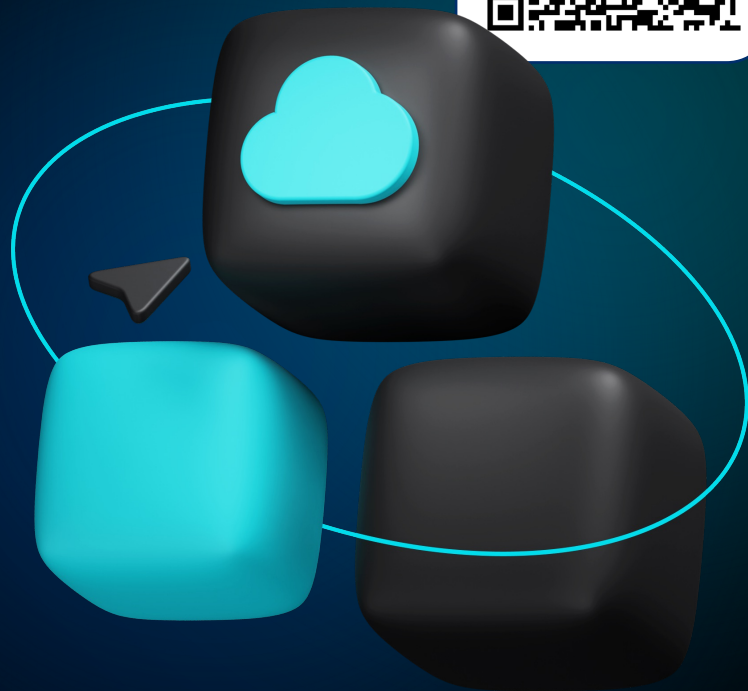




VK Data Lakehouse

Приходите тестировать в облаке

150 000 бонусных рублей на тестирование
VK Data Lakehouse в облаке с поддержкой
архитекторов



tech