DATAOPS PLATFORM

КТО СЕГОДНЯ ВЕЩАЕТ

Наджим Мохаммад

CPO Платформы
(как бы бизнес)

Макс Бартенев

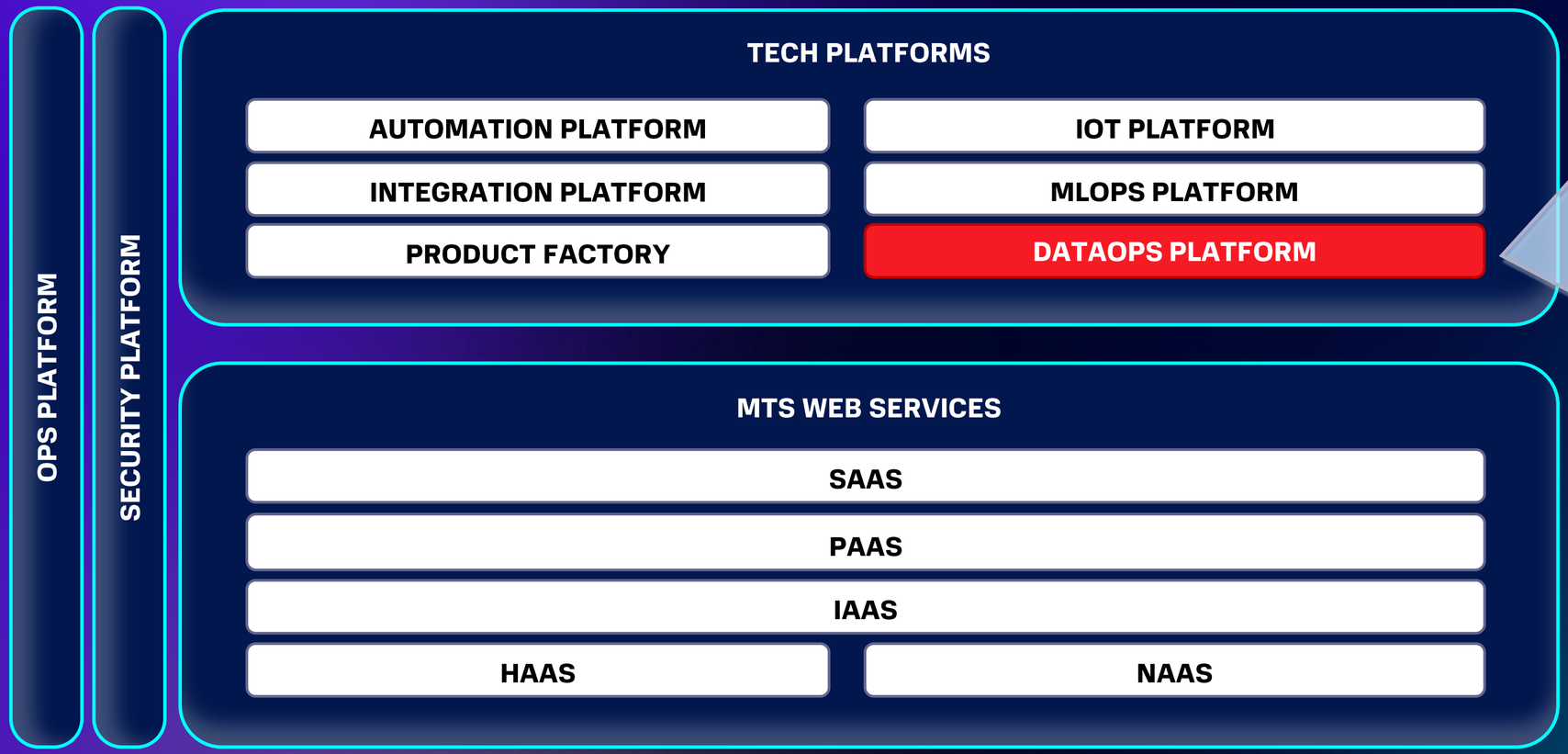
CTO Платформы
(типа технарь)

Дима Бодин

Лид Customer Happiness
(про потребителей)

DATAOPS PLATFORM – одна из ключевых платформ, согласно технологической стратегии МТС

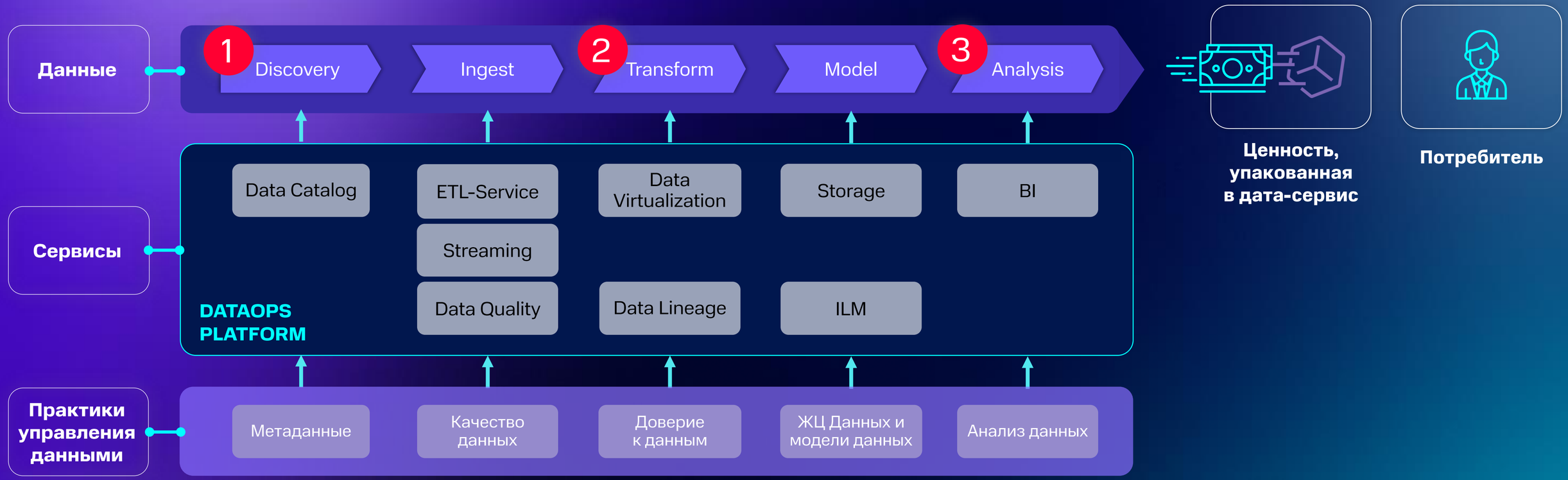
DATAOPS PLATFORM обеспечивает продукты экосистемы современными инструментами для потребления, хранения, преобразования, визуализации данных в соответствии с практиками DataOps



DATAOPS PLATFORM

- Сокращение T2M
- Управление качеством и доступностью данных
- Снижение совокупной стоимости владения инструментов управления данными
- Предоставление стандартизованного инструментария
- Снижение требований к квалификации инженеров
- Встроенная интеграция с платформенными сервисами

ЧТО ТАКОЕ DATAOPS PLATFORM?



	1	2	3
AS IS	В среднем на поиск и получение доступов к данным, уходит **более 2 недель**. Основной драйвер – поиск нужных данных	Загрузка данных, извлечение данных и преобразование становятся «узким горлышком» любой продуктовой команды, так как нужны специфичные знания **SAS** и т.д.	**80%** времени работы с данными уходит на их поиск, загрузку, очистку и т.д. и только **20%** на анализ
TO BE	Каталог данных на базе описанной метамодели, позволяет найти нужный набор данных **1-3 минут**	Команды могут самостоятельно реализовать загрузки, извлечения и преобразования, минимальные знания: **SQL**	Сокращение времени на поиск, загрузку, очистку и т.д. до **30%**.

КРАСИВЫЕ ЦИФРЫ

20+

ПБ данных

100+

Продуктов

2000+

Исследователей и инженеров данных

ТЕХНОЛОГИЧЕСКИЙ ЛАНДШАФТ

ГДЕ ХРАНИМ ДАННЫЕ

STORAGE

HADOOP - Сервис на базе собственного дистрибутива Apache Hadoop, предназначенного для хранения и обработки многотерабайтных массивов слабоструктурированных или неструктурированных данных

GREENPLUM – реляционная СУБД, имеющая массово-параллельную (massive parallel processing) архитектуру без разделения ресурсов (Shared Nothing). Или проще говоря, это система управления данными из мира big data. Она нужна тем, кто анализирует и обрабатывает десятки терабайтов информации и кому тесно и некомфортно работать с обычными СУБД.

CLICKHOUSE — аналитическая база данных, способная за секунду обработать огромное количество запросов: от сотен миллионов до более миллиарда строк и десятки гигабайт данных на один сервер в зависимости от конфигурации инфраструктуры. Команда предоставляет доступ к базе данных на ресурсах заказчика или внутреннем клауде МТС (Ocean)

ЦЕЛИ

→ Дать возможность пользователям **самостоятельно получить готовый кластер** Hadoop**,** GreenPlum, ClickHouse или других сервисов

ЧТО УМЕЕТ ПРОДУКТ:

Мы умеем **разворачивать и обеспечивать высокую доступность** высоконагруженным кластерам Hadoop, ClickHouse и Greenplum на BareMetal и виртуальной инфраструктуре

Мы оказываем услугу DBAaaS, это значит, что мы полностью администрируем развернутые нашей командой хранилища данных на ресурсах заказчика

Сервис	Краткое описание
On-premise инсталляция	Установка и настройка дистрибутива Hadoop, GreenPlum, Clickhouse (сборка МТС) на серверах клиентов
Обеспечение высокой доступности и надежности	Организация отказоустойчевого решения (в том числе с георезервированием), резервного копирования и оперативное восстановление данных в случае сбоев
Поддержка и обслуживание	Мониторинг, расширение, решение проблем пользователей, включая устранение неполадок и сбоев
Оптимизация производительности	Настройка параметров конфигурации и управление ресурсами
Безопасность данных	Управление доступом, сбор логов, аудит

ТЕХНОЛОГИИ hadoop GREENPLUM DATABASE ClickHouse

КАК ЗАГРУЖАЕМ ДАННЫЕ

BATCH

Сервис для загрузки и преобразования данных в DataOps Platform.

Наш продукт – это инструменты и сервисы для реализации процессов выгрузки, загрузки и трансформации данных (ETL\ELT).

ТЕХНОЛОГИИ

ЦЕЛЬ

→ Пользователи могут самостоятельно разрабатывать ETL-процессы, что значительно ускорит процесс аналитики данных, а также сократит T2M

ИНСТРУМЕНТЫ, КОТОРЫЕ МЫ ПРЕДОСТАВЛЯЕМ:

OnETL

- универсальная Python библиотека для любых задач Extract/Load
- на базе Apache Spark и других open-source инструментов/библиотек

JupyterHub

Единый интерфейс запуска Jupyter Notebooks и доступа к ним на edge-нодах кластеров BigData

Replick

Сервис репликации Hive таблиц между Hadoop кластерами

EverProject

Шаблон проекта и CI, который предлагается к использованию в DS/ETL репозиториях, так и всего набора инструментов и подходов для этой задачи

MTSpark

Python-библиотека для упрощения работы со Spark в окружении MTS Big Data

STREAMING

Инструмент самообслуживания для потоковой обработки данных Kafka2Kafka, позволяющий даже начинающему разработчику в сжатые сроки реализовать streaming-процесс

Основная задача:

создать набор инструментов для дата-инженеров, которые позволят им заниматься реализацией и тестированием бизнес-логики без погружения в изучение фреймворков, CI/CD, выделения и заказа ресурсов

ЦЕЛИ

→ Дать возможность пользователям самостоятельно реализовывать Streaming-процессы

→ Сокращение T2M
Проект предназначен для быстрой разработки и вывода в эксплуатацию потоковой обработки данных

ЧТО УМЕЕТ ПРОДУКТ:

Возможность использования ресурсов MTS Ocean

Возможность локальной отладки

Возможен стриминг больших объемов данных

Дата инженер/разработчик реализовывает только бизнес-логику
без погружения в изучение фреймворков, CI/CD и т.д.

От разработчика не требуется большая экспертиза в потоковой обработке данных

ТЕХНОЛОГИИ

kafka

Spark Streaming

Flink

КАК УПРАВЛЯЕМ ДАННЫМИ

DATA QUALITY

Self-service инструмент для контроля качества данных. Под качеством данных принято понимать степень их пригодности для решения конкретной задачи. Степень пригодности характеризуется рядом метрик (доступность, полнота, точность т.д.).

SDQ позволяет эти метрики:

- создавать и запускать их
- анализировать результаты
- отслеживать динамику изменений
- получать алерты о неуспешных проверках

ЦЕЛИ

→ Предоставить единый инструмент и подходы к проверке качества данных

→ Централизованное хранение информации о качестве данных во всех ИС ПАО МТС

ЧТО УМЕЕТ ПРОДУКТ:

Централизованное хранение информации о качестве данных во всех ИС компании

Подключение к различным реляционным и NoSQL источникам данных. Полный список - 03 - Data Quality (DQ Neo)

Стандартный набор метрик с возможностью создания кастомных

Продвинутый программный API с **возможностью оптимизации проверок** и code-first настройки

Web интерфейс для создания и запуска проверок

Визуализация результатов метрик и проверок на дашбордах Grafana

Оповещения об инцидентах:

- через почту
- интеграция с JIRA
- интеграция с Remedy

ТЕХНОЛОГИИ

kafka, Spark, JDBC, Grafana

Сервис	Краткое описание
DQ Neo	Доступность сервиса
Подключение новых источников данных	Сроки подключения новых источников данных по запросу продуктовых команд

DATA CATALOG

Сервис для поддержки процессов data governance:

- каталогизация
- описание
- администрирование данных

Система, которая является точкой входа для вопросов по данным,
и содержит информацию
о том, какие данные есть в МТС как экосистеме, каковы их характеристики и откуда эти данные можно получить

В качестве Каталога данных, как компонента DataOps платформы, в МТС выбрано opensource решение, которое называется DataHub

ЦЕЛИ

- → Ведение описания данных и их поиск
- → Обеспечение процессов data governance
- → Избавиться от вендорской зависимости: Alation

ЧТО УМЕЕТ ПРОДУКТ:

Извлечение метаданных и **поиск** источников данных
ряционные/нереляционные БД, API, Kafka

Поиск и заведение бизнес сущностей
Термины глоссария

Описание данных
физических и бизнес

Нахождение источников, физически хранящих данные о бизнес сущностях
Терминах

Содержит информацию о владельцах данных, теги, описание данных, ссылки на источники данных

ТЕХНОЛОГИИ

DataHub

Сервис	Краткое описание
Каталогизация данных	Прием мета данных источников из инджесторов. Пользовательский интерфейс для работы с элементами ФМД, Терминами глоссария, связями

КАК АНАЛИЗИРУЕМ ДАННЫЕ

BI TOOLS

Предоставление **стандартов визуализации** и нескольких BI инструментов

В основе DataOps.BI лежит Apache Superset - это открытое программное обеспечение **для исследования и визуализации данных**, ориентированное на большие данные

DATAOPS.BI (SUPERSET):

- Гибкий инструмент для работы с большим объемом данных
- Большой набор визуализаций из коробки и возможность подключить собственные
- Расширенная модель безопасности
  роли, RLS
- Удобный импорт/экспорт объектов инфопанели | наборы данных | визуальные элементы
- Импорт данных из CSV файлов
- Экспорт данных в CSV и JSON
- Возможность получить встраиваемые дашборды с использованием RLS
- Можно работать с SQL запросами прямо в интерфейсе приложения, создавать из результатов запроса отчеты или извлекать из визуализации SQL выборку
- Инструмент постоянно развивается, добавляются новые возможности и расширяются имеющиеся
- Доработки ведутся как opensource-сообществом, так и нашей командой

ТЕХНОЛОГИИ

Superset

ЦЕЛИ

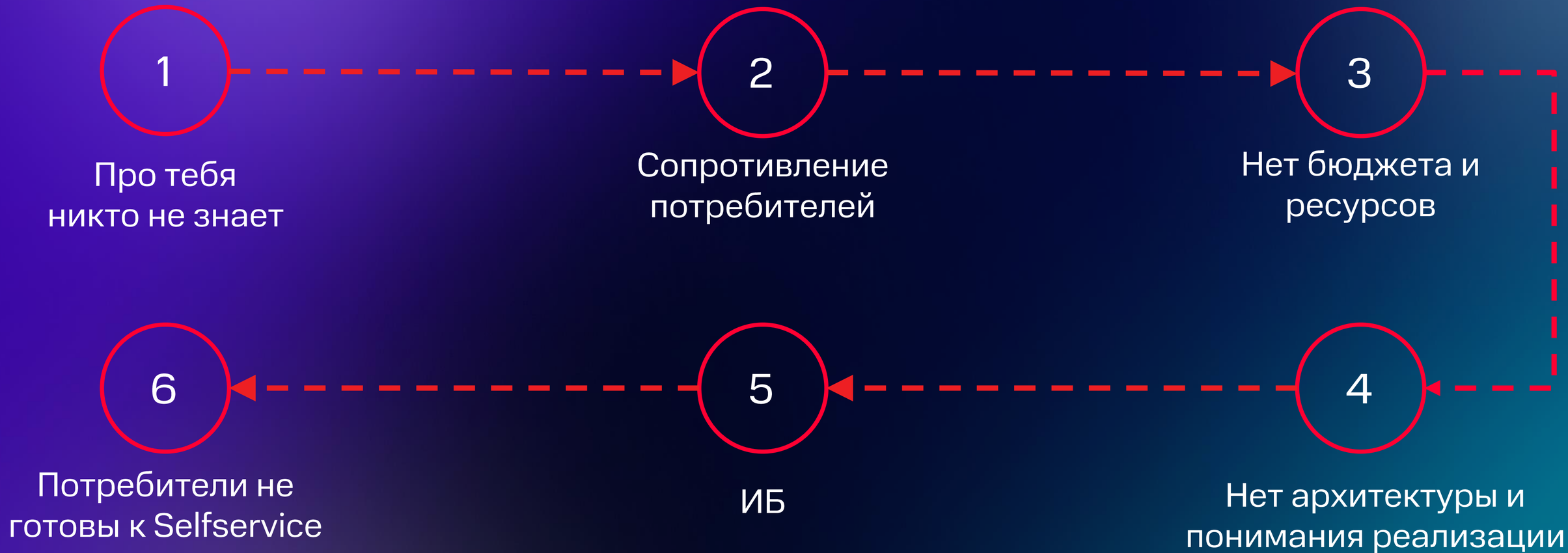
→ Предоставление пользователям инструмента с возможностью самостоятельной работы с BI

→ Избавиться от зависимости текущего вендорского стека: Tableau, PowerBI

Сервис	Краткое описание
DataOps **BI Superset**	Компонент платформы DataOps разрабатываемый на основе открытых исходников Apache Superset для исследования и визуализации данных

ЧТО В ИТОГЕ ИМЕЕМ

С ЧЕМ СТОЛКНУЛИСЬ?

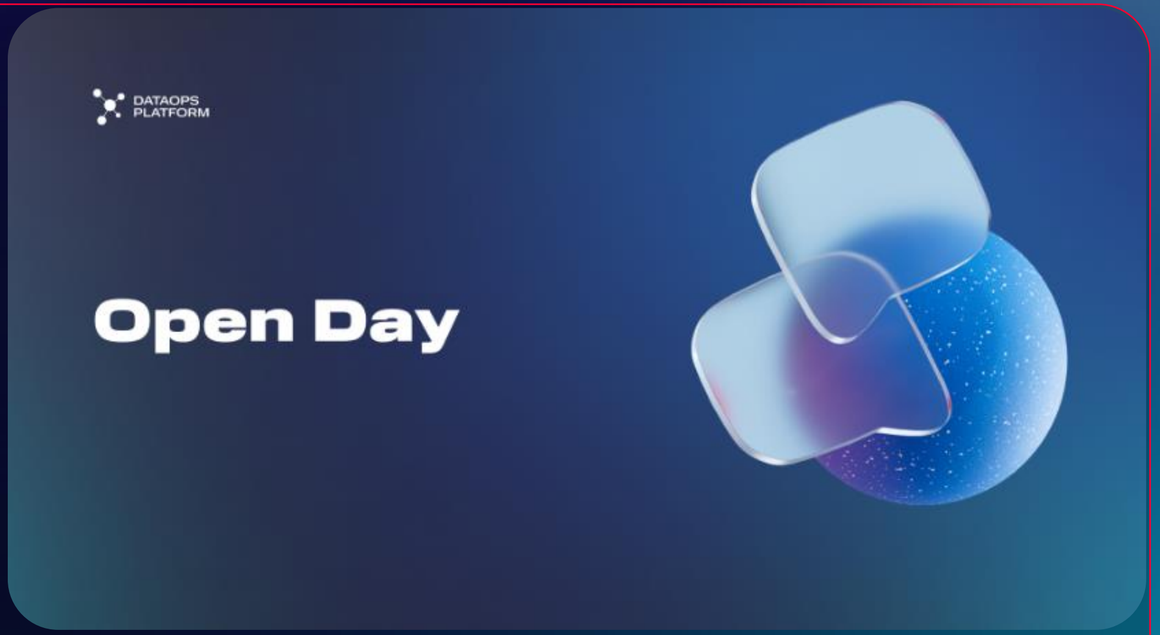


С ЧЕМ СТОЛКНУЛИСЬ?
- ПРО НАС НИКТО НЕ ЗНАЕТ

- Публикации в каналах ТГ, на внутреннем портале, в почтовых рассылках
- Митапы продуктов
- OpenDay на всю компанию

С ЧЕМ СТОЛКНУЛИСЬ?
- ПРО НАС НИКТО НЕ ЗНАЕТ

- Сделали рассылку с приглашением на всю компанию
- На OpenDay DataOps platform провели демо по каждому инструменту платформы
- Собрали обратную связь, провели сессию вопросов - ответов



В результате получили +200 новых потребителей за 2 дня

С ЧЕМ СТОЛКНУЛИСЬ?
- СОПРОТИВЛЕНИЕ ПОТРЕБИТЕЛЕЙ

Customer Happiness – внутренняя команда консалтинга

Сопровождаем до успешного подключения к платформе

Единый Service Desk всех компонентов платформы

С ЧЕМ СТОЛКНУЛИСЬ?
- СОПРОТИВЛЕНИЕ ПОТРЕБИТЕЛЕЙ

Step 1 По всем вопросам идите в Customer Happiness!

Step 2

1. Определяем цели и ограничения проекта
2. Проводим архитектурную сессию с потребителем
3. Согласовываем HLD целевой архитектуры с учетом требований ИБ
4. Разрабатываем ресурсную модель и план реализации проекта
5. Сопровождаем реализацию проекта совместно с SME потребителей
6. Осуществляем поддержку решений на протяжении всего времени эксплуатации

С ЧЕМ СТОЛКНУЛИСЬ?
- НЕТ БЮДЖЕТА И РЕСУРСОВ

Выделяем ресурсы во внутреннем Cloud

Защита бюджета по внутренним процедурам компании

Наращивание компетенций внутри команды потребителя

С ЧЕМ СТОЛКНУЛИСЬ?
- НЕТ БЮДЖЕТА И РЕСУРСОВ

Для инструментов DataOps platform еще и сервера нужны?

Мы не закладывали деньги на сервера

А сколько нужно Гигабайт, чтобы сохранить 1 млн строк?

Нам нужна виртуалка или Bare metal?

Хотим сразу в ПРОД, чтобы 99,9999

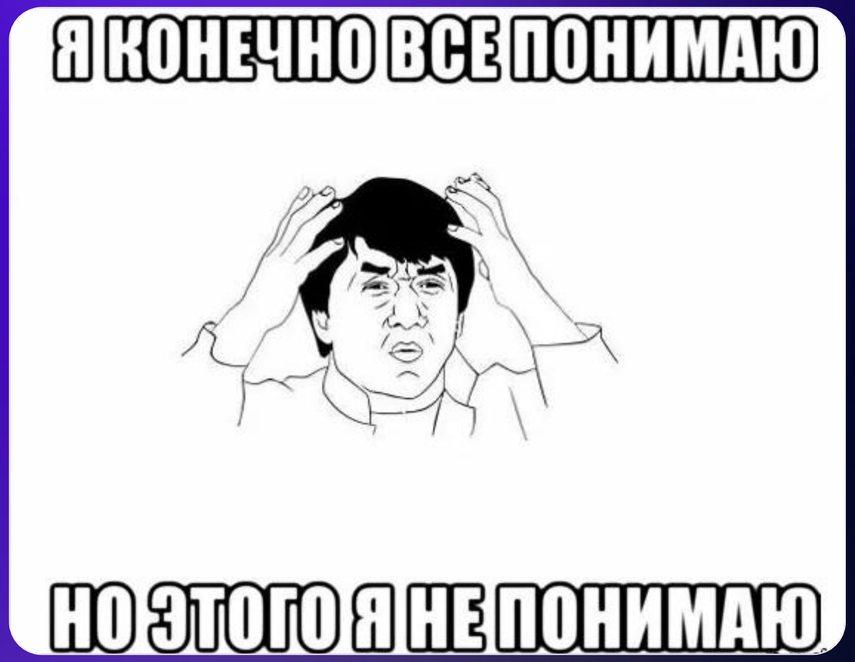
Как защитить бюджет на сервера?

CAPEX? OPEX? Что это?

Защитить бюджет не так просто, а можно только посмотреть? (MVP)

С ЧЕМ СТОЛКНУЛИСЬ?

НЕТ АРХИТЕКТУРЫ И ПОНИМАНИЯ РЕАЛИЗАЦИИ

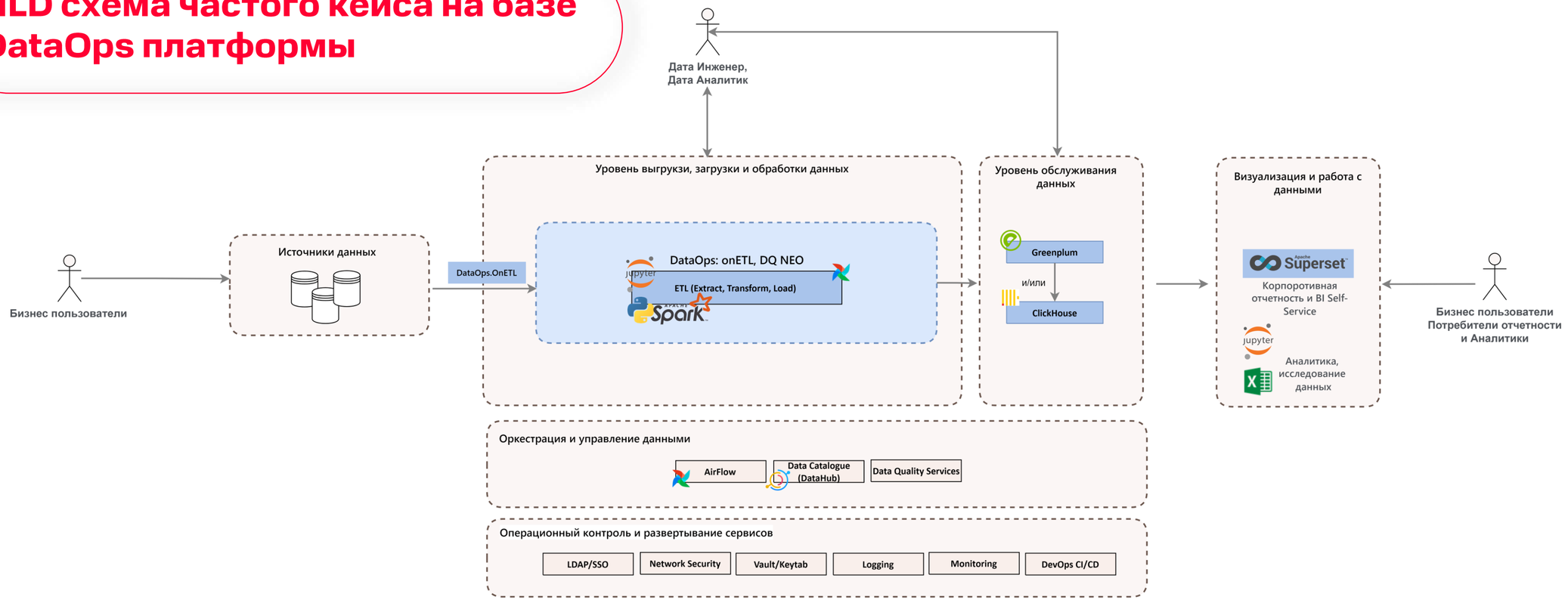


- Подключаем Solution Architect на архитектурной сессии
- Предлагаем оптимальное решение
- Готовим HLD и согласовываем с ИБ

С ЧЕМ СТОЛКНУЛИСЬ?

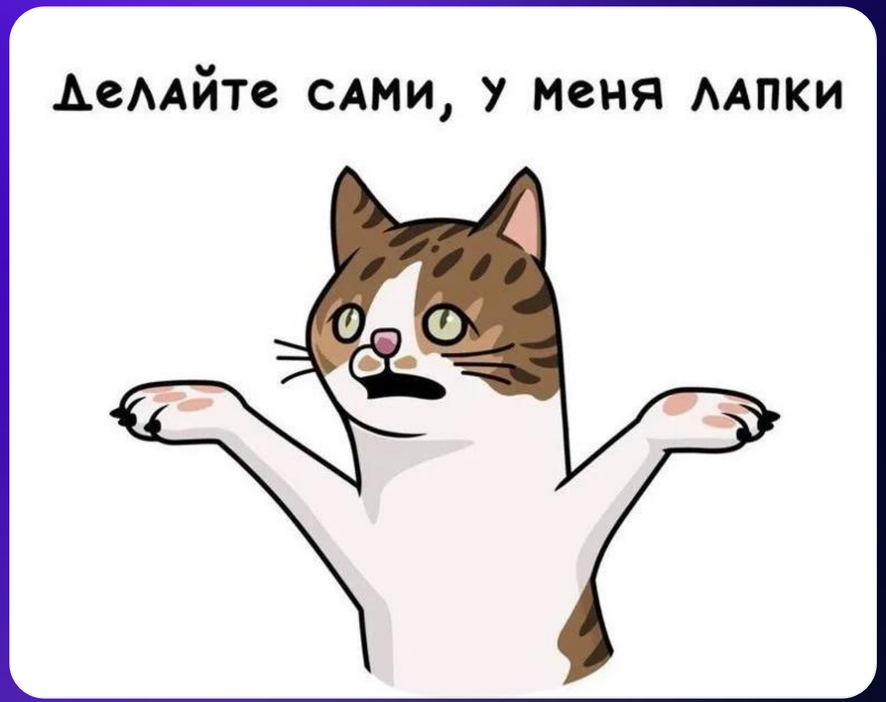
НЕТ АРХИТЕКТУРЫ И ПОНИМАНИЯ РЕАЛИЗАЦИИ

HLD схема частого кейса на базе DataOps платформы



С ЧЕМ СТОЛКНУЛИСЬ?

ПОТРЕБИТЕЛИ НЕ ГОТОВЫ К СЕЛФСЕРВИСУ



- Фреймворк обучения компонентам платформы
- Каждый проект ведет Delivery Manager
- Подключаем DevOps из команды Customer Happiness
- Консультируем в соответствии с Best Practice

С ЧЕМ СТОЛКНУЛИСЬ?
ПОТРЕБИТЕЛИ НЕ ГОТОВЫ К СЕЛФСЕРВИСУ

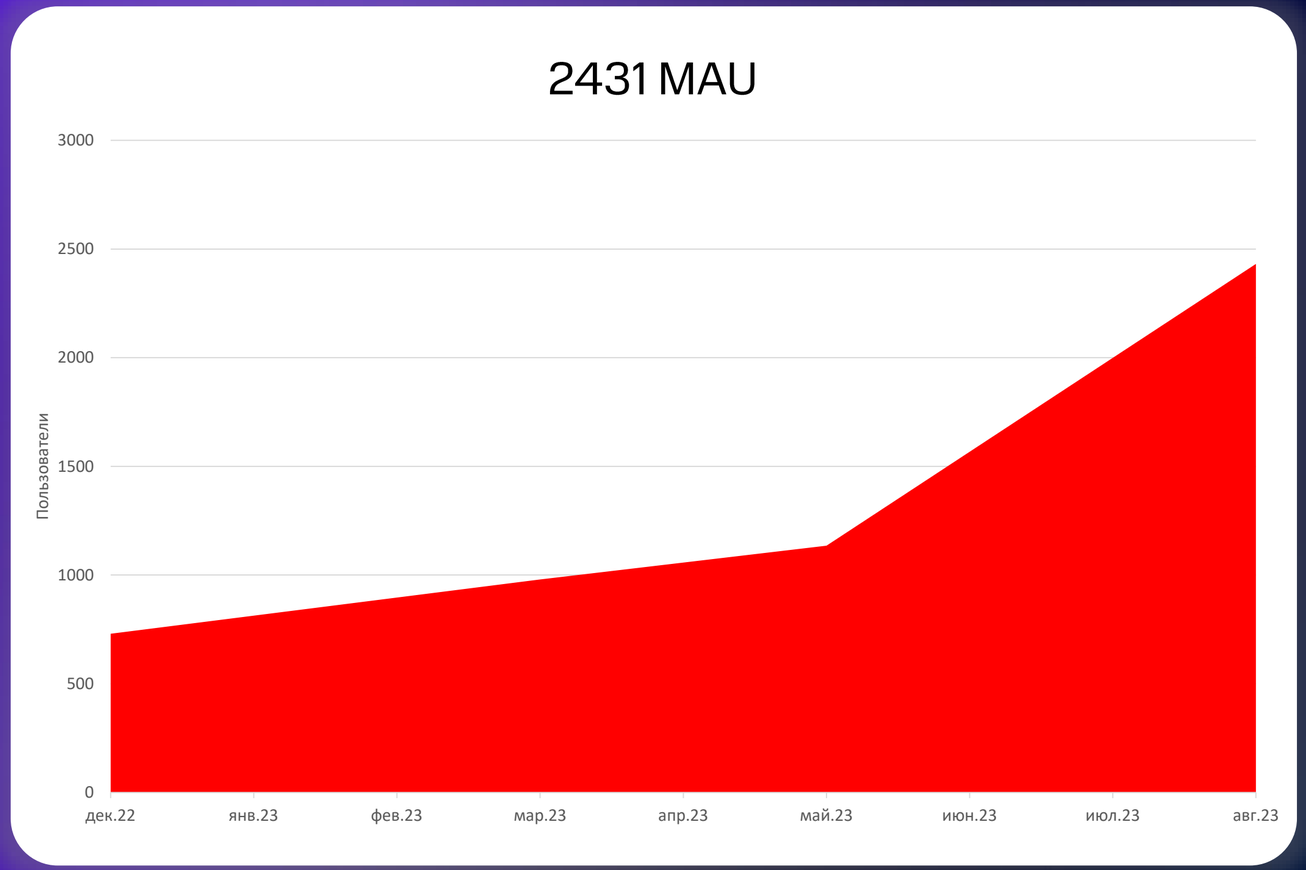
1.	Первичные вводные от потребителя	DOP-3947	ОТКРЫТО	=
2.	Заведение проекта в CRM и Jira	DOP-3949	ОТКРЫТО	=
3.	Проведение установочной встречи с потребителем	DOP-3948	ОТКРЫТО	=
4.	Регистрация продукта в PPinfo и SOL	DOP-4212	ОТКРЫТО	=
5.	Описание источников данных + подготовка сэмплов данных	DOP-3952	ОТКРЫТО	=
6.	Проектирование сетевой схемы	DOP-3958	ОТКРЫТО	=
7.	Проведение встречи СН + Потребитель + ИБ +ОЗКИ	DOP-3953	ОТКРЫТО	=
8.	Составление план-графика + DoD	DOP-3956	ОТКРЫТО	=
9.	Заполнение опросного листа	DOP-3957	ОТКРЫТО	=
10.	Проектирование схемы ИП	DOP-3959	ОТКРЫТО	=
11.	Описание ИП	DOP-3960	ОТКРЫТО	=
12.	Проектирование архитектуры системы защиты при создании ИС.	DOP-3961	ОТКРЫТО	=
13.	Организация новой ИС	DOP-3962	ОТКРЫТО	=

Всегда актуальный и прозрачный статус в план-графике:
сроки, ответственный, какой командой выполняется задача, к кому можно обратиться с вопросами и т.д.

Ответственный за реализацию проекта интеграции — менеджер проекта в СН

Недостаток компетенции внутри команды потребителя – привлекаем специалистов и роли из других подразделений

МЫ НАУЧИЛИСЬ РАБОТАТЬ С ПОТРЕБИТЕЛЯМИ



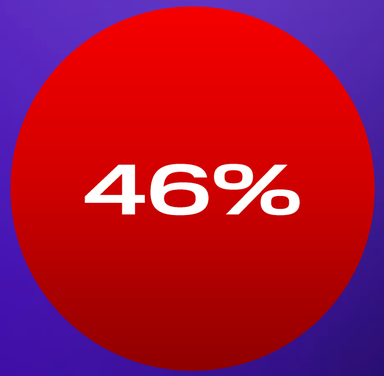
86

Проектов интеграции

80

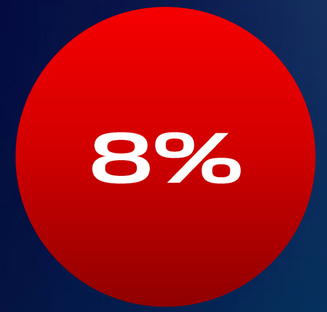
Продуктов потребителей

РАЗБИВКА ПРОДУКТОВ DATAOPS PLATFORM ПО ПОПУЛЯРНОСТИ, Q3 2023



BI Tools

Storage



DQ

ETL

Data Catalog

Data Virtualization

КОНТАКТЫ

Макс

tg: @max_bartenev

Наджим

tg: @nadzhimeski

Дима

tg: @DmitryB1

Спасибо за внимание!