Arenadata Hadoop

Экосистема для хранения, обработки и анализа неограниченного объёма данных любого типа

ARENADATA

План презентации

1. О компании Arenadata
2. Обзор Arenadata Hadoop
3. Экосистема Arenadata Hadoop
4. Сравнение ADH vs Open Source
5. Платформа сбора и хранения данных Arenadata Enterprise Data Platform

О компании Arenadata

ARENADATA

Arenadata

лидер в области платформ больших данных для цифровой трансформации

Greenplum · hadoop · ClickHouse · PostgreSQL · TARANTOOL · kafka · nifi · Open Metadata

СТАБИЛЬНОСТЬ

АДС Холдинг:

- **6 лет** на рынке
- **430** человек
- **4,0 млрд**. **₽** выручка ’23

80+ КЛИЕНТОВ

- ГПН, Росатом, ММК, Норникель, Сибур;
- X5, Магнит, Ашан;
- Тинькофф, ВТБ, ПСБ, ГПБ.
- ФНС, ДИТ МСК, Минздрав;

NO VENDOR LOCK

Мы создаем **тиражируемый продукт, совместимый с upstream** версиями opensource проектов.

ONPREM & CLOUD

- Onpremise и ПАКи;
- CLoud: VK Cloud; Cloud.ru, MTS Cloud, Tl Cloud, Croc Cloud, Beeline Cloud;
- Гетерогенные ландшафты;

ФОКУС БИЗНЕСА

100% бизнеса Arenadata – создание СУБД на базе Open Source, поддержка 1, 2 и 3 линии. Архитектурный надзор от вендора.

ЭКСПЕРТЫ

- **60+** технологических партнеров и интеграторов;
- **2.500+** курсов проведено для специалистов и экспертов.

COMMUNITY

- **#1** в мире в сообществе контрибьютеров в Greenplum;
- **#1** в РФ в сообществе контрибьютеров в Clickhouse;
- Профессиональная документация.

РЕГУЛЯТОРЫ

- Отечественные ОС;
- Реестр;
- ФСТЭК;
- ГОСТЕХ, ГЕОП.

Наш успех – доверие клиентов

80+ клиентов размещают свои хранилища, витрины, озера данных на продуктах Arenadata

Retail & FMCG

X5Group, Магнит, Ашан, РИВ ГОШ, Комус, Danone, Hoff, Unilever, Burger King

Банки, страховые и телеком

ВТБ, Газпромбанк, Росбанк, ПСБ, ДОМ РФ, Банк Санкт-Петербург, Ренессанс Кредит, Ак Барс Банк, ТТК, МТС, Тинькофф, ОТП Банк, РСХБ, Синара банк (новое имя СКБ-банка), Kapital Bank, PAŞA Həyat, Halyk Bank, АЭБ, ГТЛК, Ингосстрах

Госсектор

ФНС, ДИТ Москвы, Гостех, НСУД (Национальная система управления данными), Социальный фонд, Россельхознадзор, Счетная палата, Минтруд России, Минздрав, Минцифры, Казначейство, Росреестр, Роструд

Пром, Энерго, Добыча, Транспорт

Газпром нефть, КазМунайГаз, Россети, Почта России, Норникель, Магнитогорский металлургический комбинат, Евраз, Зарубежнефть, Сибур, Уралхим, ЦППК, Мосэнергосбыт, Мосэнерго, Алроса, Полюс, УГМК-Холдинг, Росатом

Arenadata – платформа данных для цифровой трансформации

Предиктивные ремонты

Next Best Action

Кредитный калькулятор

Быстрые витрины

AI ML

Расчет мотивации

Ядро цифровой трансформации – платформа данных

Быстрые витрины
ADQM on Clickhouse

Аналитическое хранилище
ADB on Greenplum

Озеро данных для ML\AI
ADH
on Hadoop & Impala

Интеграция
ADSK on Kafka

Интеграция
ADSN on NiFi

Безопасность
ADPS: Arenadata Platform Security

In-Memory Data Grid
ADG on Picodata

OLTP СУБД
ADPG on PostgreSQL

Каталог данных, Бизнес глоссарий, Качество данных, Дата линедж
ADC on Open Metadata

ИТ-Оркестрация, Мониторинг, Эксплуатация
ADCM: Arenadata Cluster Manager

Activity based costing

Промышленный IoT

Customer journey

Снижение оттока

Прогноз спроса

Обзор Arenadata Hadoop

ARENADATA

Корпоративный дистрибутив для распределённой масштабируемой обработки данных

Arenadata Hadoop (ADH) — корпоративный дистрибутив на базе Apache Hadoop, предназначенный для хранения и обработки слабоструктурированных и неструктурированных данных.

Наиболее эффективен для следующих сценариев:

- обработка больших наборов данных в средах, где размер данных превышает доступную память
- пакетная обработка с задачами, использующими операции чтения и записи на диск
- создание инфраструктуры для анализа больших данных
- анализ исторических и архивных данных
- непрерывный сбор различных метрик и журналов
- распределённое хранилище оперативных данных

Развитие Arenadata Hadoop

Решаемые задачи

Хранение данных любого типа, включая неструктурированные

- Распределённая обработка данных
- Системы управления документами и контентом
- Хранение и регистрация событий
- Данные датчиков, каталоги товаров
- Резервное копирование других СУБД

Построение озёр и фабрик данных

- Единый центр всех данных компании
- Быстрое развёртывание «песочниц» для пилотных проектов и проверки статистических гипотез
- Работа со всеми аналитическими инструментами в единой среде

Машинное обучение и искусственный интеллект

- Обучение моделей на больших данных
- Распределённое машинное обучение на базе Spark
- Эффективная эксплуатация моделей в SQL-среде с помощью встроенных функций Madlib

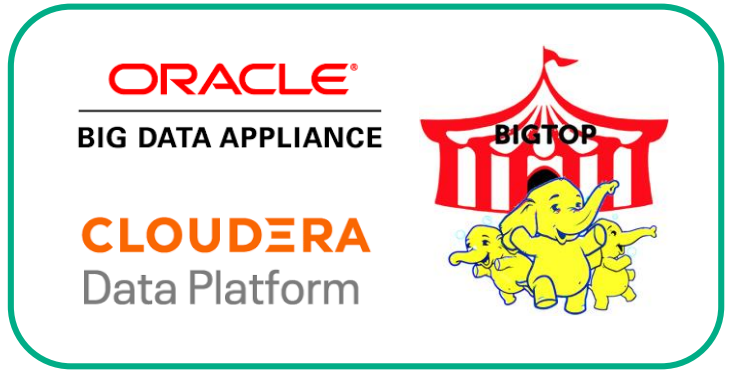
Импортозамещение и разгрузка иностранных систем

- Миграция с иностранных систем (Oracle BDA, Cloudera)
- Прозрачная методика перехода, минимум рисков и сохранение всех преимуществ

Минимизируйте санкционные и валютные риски

Arenadata Hadoop высоко зарекомендовал себя на российском рынке. Продукт эффективно замещает такие иностранные системы как Cloudera CDP, Oracle BDA, а также «ванильные» сборки Hadoop и др.

Наши программные решения стали частью проектов миграции с зарубежного ПО в десятках крупнейших организаций, включая ФНС РФ, Счётную палату РФ, Банк ВТБ и ПАО «Газпром нефть».



Преимущества Arenadata Hadoop

1

Корпоративная платформа хранения и обработки данных

Широкая экосистема data-сервисов Arenadata позволит заказчикам использовать не только Hadoop, но и полнофункциональные решения для обработки структурированных и потоковых данных.

2

Сборка на базе открытых проектов Apache

ADH является полностью open-source дистрибутивом Hadoop, поэтому нашим клиентам никогда не придётся столкнуться с такой проблемой, как vendor lock-in.

3

Пакет утилит для полной офлайн-установки

Arenadata Hadoop обеспечивает полный набор возможностей и инструментов для автоматических установки и настройки компонентов как на «голом железе», так и на виртуальных машинах (в облаке). Средства мониторинга и управления конфигурацией кластера позволяют оптимизировать производительность для всех компонентов системы.

4

Интеграция с FreeIPA

Поддержка Kerberos аутентификации и Ranger авторизации на базе FreeIPA (централизованной системы по управлению идентификацией пользователей), как альтернатива интеграции с Active Directory.

Преимущества Arenadata Hadoop

5

Собственная система управления

Мы предоставляем открытую систему автоматического развёртывания и управления Arenadata Cluster Manager. Она является Multi-cloud системой, и может быть развёрнута на любой имеющейся инфраструктуре, включая публичные облака.

7

Набор типовых пакетных сервисов по планированию, установке и аудиту системы

Наши специалисты настроят Arenadata Hadoop (удалённо или on-site), а впоследствии проведут аудит системы и помогут разработать шаги для решения поставленных задач.

6

Возможность влиять на развитие системы

Наши клиенты могут влиять на планы по развитию системы и экосистемы в целом, а мы можем взять на поддержку полноценный компонент платформы.

8

Поддержка российских ОС

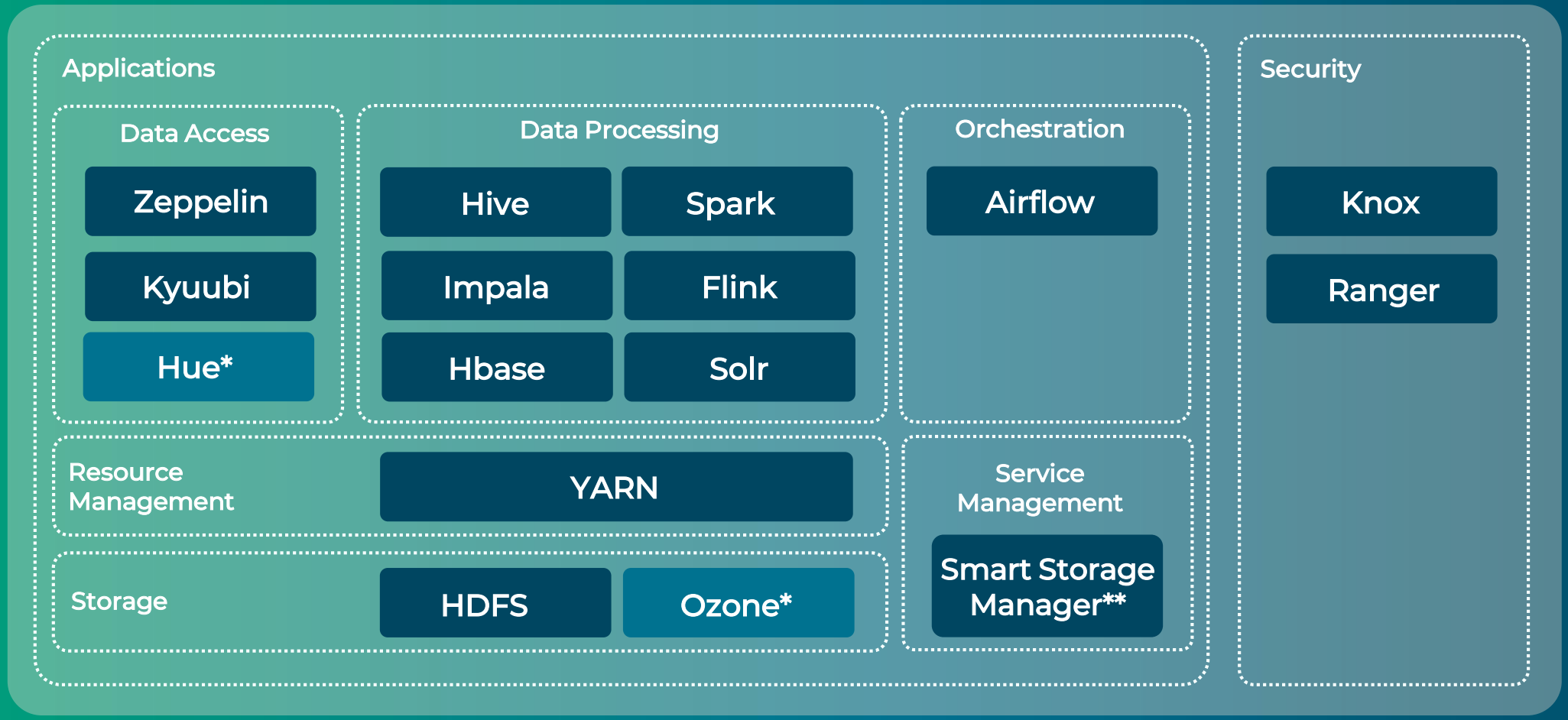
Поддержка российских сертифицированных ОС AltLinux 8 SP (Альт 8 СП), Astra Linux 1.7 SE «Орёл».

Экосистема Arenadata Hadoop

Компоненты и сервисы

ARENADATA

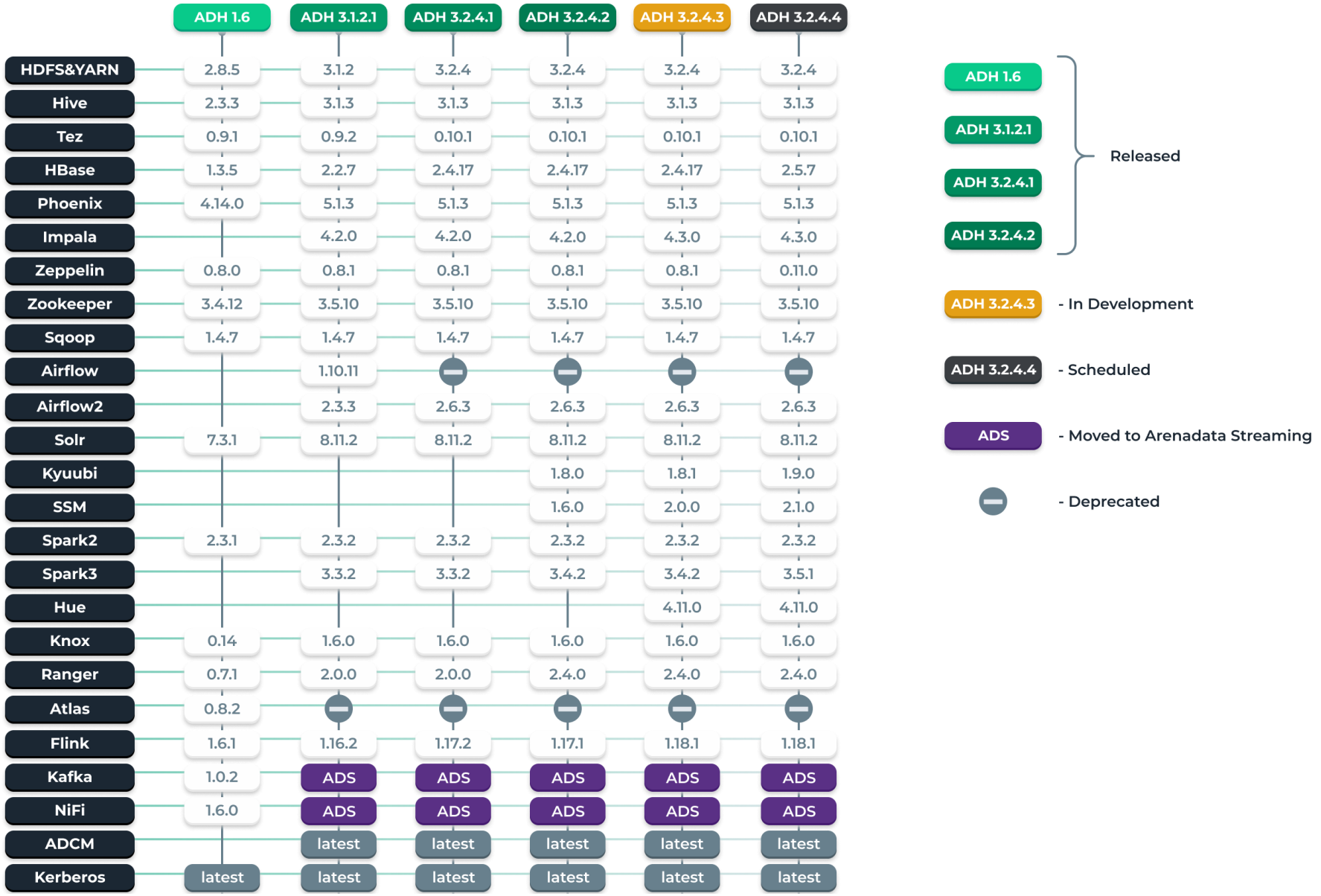
Экосистема Arenadata Hadoop



* В разработке
** Technology Preview

ARENADATA HADOOP

Состав компонентов текущей и будущих версий



Система управления групповыми политиками безопасности

Arenadata Platform Security (ADPS) – система централизованного управления политиками безопасности кластера ADH. Включает:

- аутентификацию с использованием Kerberos, интеграцию с LDAP/Active Directory
- интеграцию с Apache Ranger и Apache Knox для обеспечения безопасного доступа к кластерам Hadoop
- журналы аудита и отчеты

Поставляется как бесплатное дополнение к Enterprise редакции ADH версии 2.1+.

Комплексный подход к безопасности, включающий защиту периметра, управление доступом на основе политик, авторизацию и безопасный доступ к платформе и ее сервисам.

Защита конфиденциальных данных и соответствие нормативным требованиям.

Функции безопасности Arenadata Platform Security

Безопасность периметра

- Apache Knox
- Gateway

Аутентификация

- Kerberos
- LDAP/AD
- FreeIPA
- MIT KDC

Защита данных

- SSL
- Шифрование RPC
- Шифрование at Rest

Аудит и мониторинг

- Запросы доступа
- Операции обработки данных
- Изменение данных

Авторизация

- Контроль доступа HDFS, YARN, Hbase
- Контроль доступа на уровне баз данных, таблиц, столбцов для наборов данных Hive
- Контроль доступа Knox
- Контроль доступа к коллекциям Solr

Impala: MPP-СУБД в экосистеме Hadoop

Impala — распределённая система исполнения SQL-запросов в экосистеме Hadoop. Сервис предназначен для интерактивной обработки данных на сверхбольших объёмах данных.

Основные преимущества

Высокая скорость обработки запросов в озере данных

Низкая задержка и высокий уровень параллелизма в экосистеме Hadoop, что позволяет эффективнее решать задачи self-service аналитики и ad-hoc запросов

Простое внедрение в имеющуюся инфраструктуру

Impala использует те же метаданные, форматы файлов и драйверы подключения, что и Hive

Снижение стоимости обработки данных

Достигается за счёт оптимального использования аппаратного обеспечения

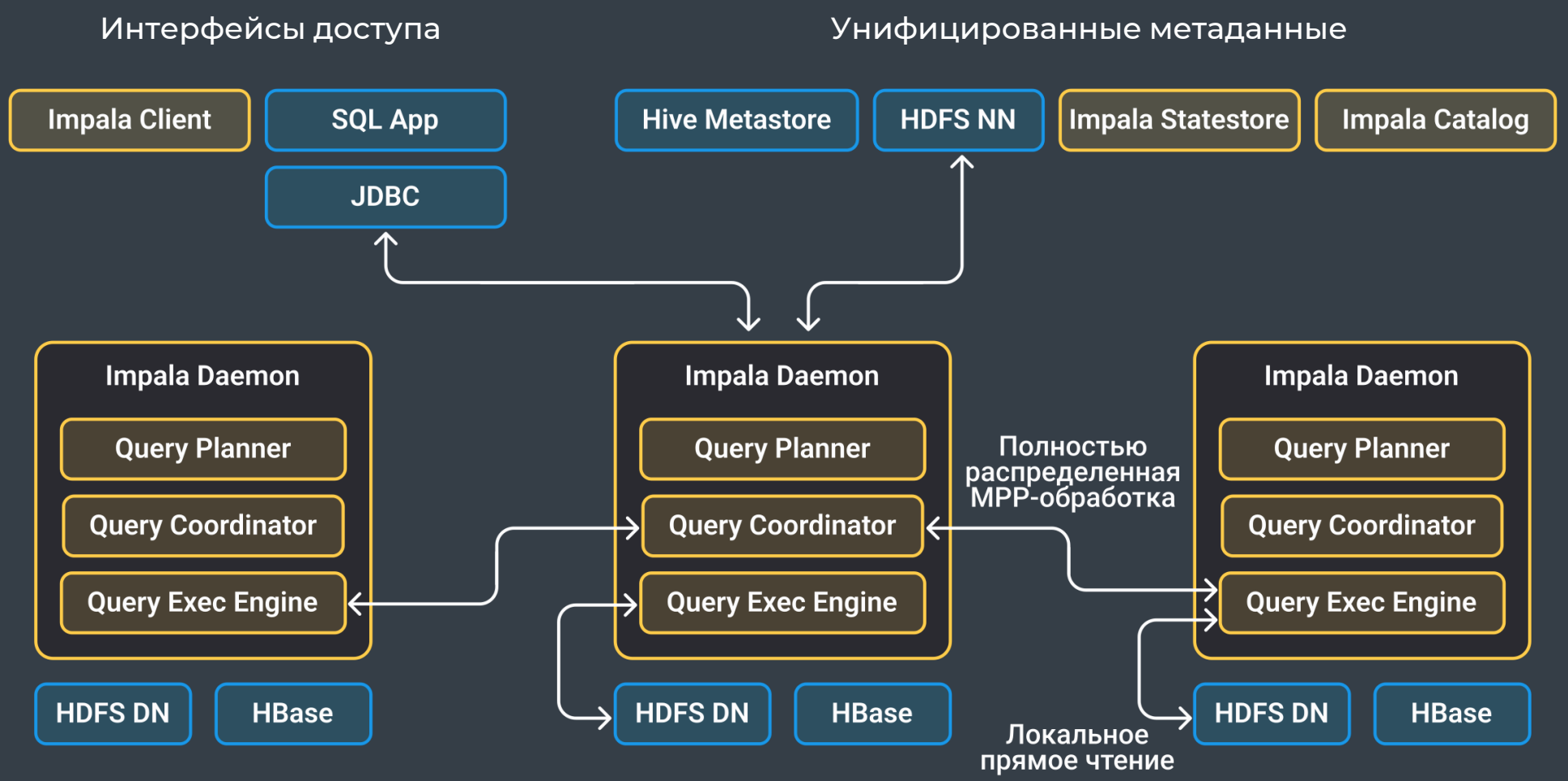
Масштабирование аналитической нагрузки

Развёртывание Impala вне основного кластера позволяет исключить конкуренцию за аппаратные ресурсы

Оптимизация ландшафта

Благодаря локальной реализации отдельных сценариев ad-hoc и self-service аналитики

Архитектура Impala



Почему Impala? Технические аргументы

Быстрый SQL – ELT на больших данных

(TPC-DS/H SF>5000)

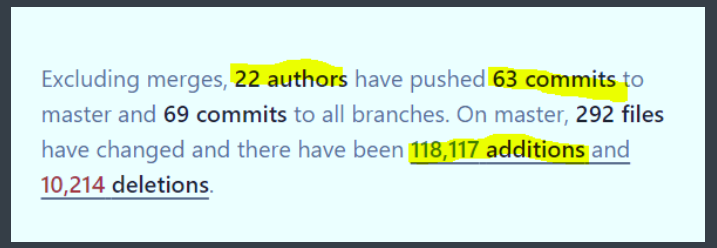
- MPP-архитектура на Hadoop
- Механизм Short-Circuit (только C++ библиотеки)
- Кэширование данных
- Кэширование метаданных
- Вычисления в памяти

Конкурентный доступ – Ad Hoc-analysis

- Минимальная утилизация ОП
- Механизм кэширования данных ОС

Перспективный opensource проект

- Активно развивается
- Hadoop Native (HDFS, Ozone)
- Проверенный production ready сервис



Варианты развёртывания Impala

1. Совместно с HDFS на одном узле данных
2. Отдельный аналитический кластер Impala

1

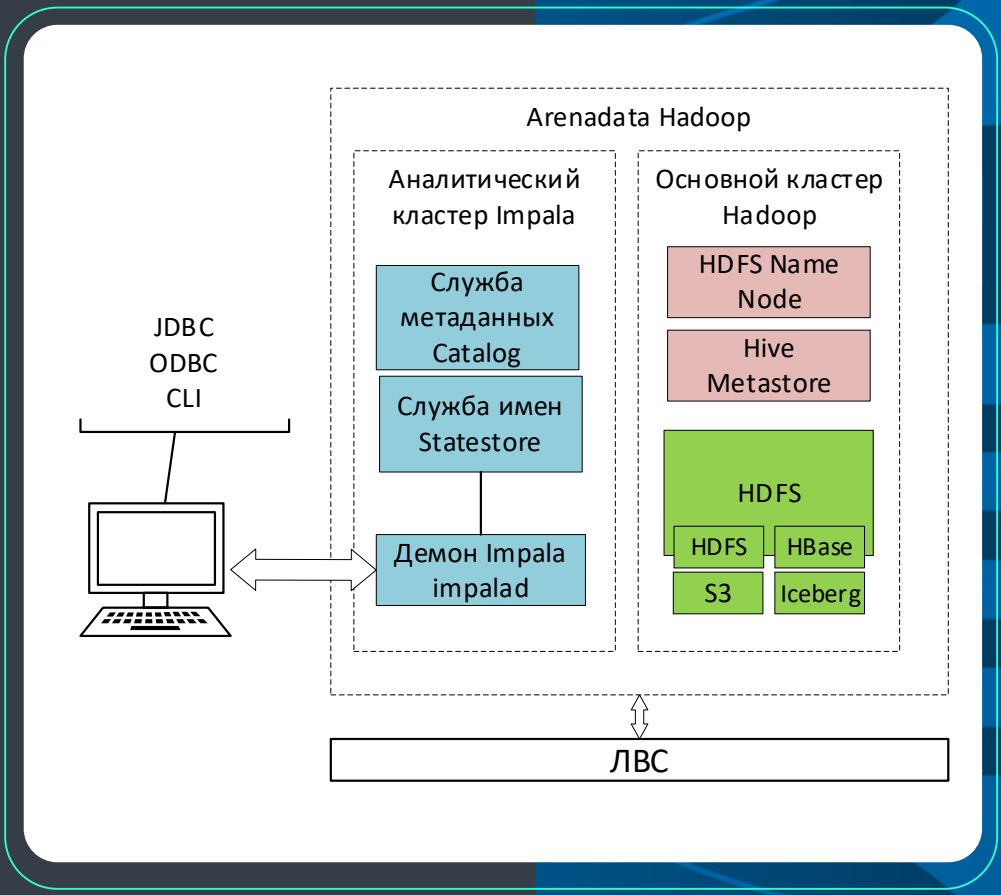
Совместное с HDFS развертывание это:

- максимально производительный MPP
- использование преимуществ механизма прямого доступа к блокам данных (Short-Circuit) и HDFS cache

2

Развертывание отдельного аналитического кластера Impala позволит:

- использовать существующую инфраструктуру хранения и обработки
- дополнить производственный кластер быстрой Ad Hoc-аналитикой
- ускорить критические ELT-процессы



Smart Storage Manager: интеллектуальное управление данными в экосистеме Hadoop

Температурное хранение данных

Оптимизация управления данными в зависимости от их востребованности:

- перемещение наиболее горячих данных в кеш
- перемещение горячих данных в SSD
- архивация холодных данных

Асинхронная репликация

Репликация данных между Hadoop-кластерами или между Hadoop-кластером и облачным хранилищем:

- отслеживание изменений данных
- синхронизация в реальном времени
- реализация сценариев аварийного восстановления

Настройка политик Erasure Coding и сжатия данных

- гибкая настройка включения Erasure Coding и управление файлами с разными политиками EC с помощью правил
- сжатие данных в HDFS без ограничения доступа к ним для внешних приложений

Оптимизация работы с небольшими файлами

- сжатие небольших файлов в файл-контейнер, хранящийся в HDFS
- данные доступны для приложений верхнего уровня

80%

вычислительных нагрузок приходится на обработку

20%

данных

Smart Storage Manager: преимущества сервиса

- Снижение стоимости хранения холодных данных
- Повышение производительности чтения горячих данных
- Простая настройка и управление функцией асинхронной репликации
- Снижение накладных расходов и повышение производительности записи и чтения небольших файлов HDFS
- Экономия места в хранилище

Оптимизация затрат

Повышение производительности

Надёжность

Kyuubi: SQL gateway в экосистеме Hadoop

Kyuubi — распределённый многопользовательский шлюз для предоставления SQL для DWH и DataLake.

Kyuubi создает распределённые механизмы SQL запросов поверх различных вычислительных платформ, например, Apache Spark, Flink, Hive, Impala и др., чтобы получать и обрабатывать большие наборы данных из разнородных источников.

Основные возможности

Многопользовательский доступ

Сквозная поддержка доступа нескольких пользователей к данным через единую систему аутентификации и авторизации

Высокая доступность (HA)

Балансировка нагрузки через Zookeeper обеспечивает высокую доступность enterprise-уровня и неограниченно высокий уровень параллелизма клиентов

Несколько рабочих нагрузок

Поддержка разнородных рабочих нагрузок в рамках одной платформы, одной копии данных и одного интерфейса SQL

Сферы применения Kyuubi

Интерактивная аналитика

- Быстрая аналитика для интерактивного анализа больших данных
- Распределённые механизмы SQL-запросов поверх различных вычислительных платформ (Spark, Flink, Impala и др.)
- Доступ через JDBC/ODBC
- Возможность генерации запросов через SQL или инструменты BI
- Совместное использование ресурсов и быстрый отклик за счёт распараллеливания запросов

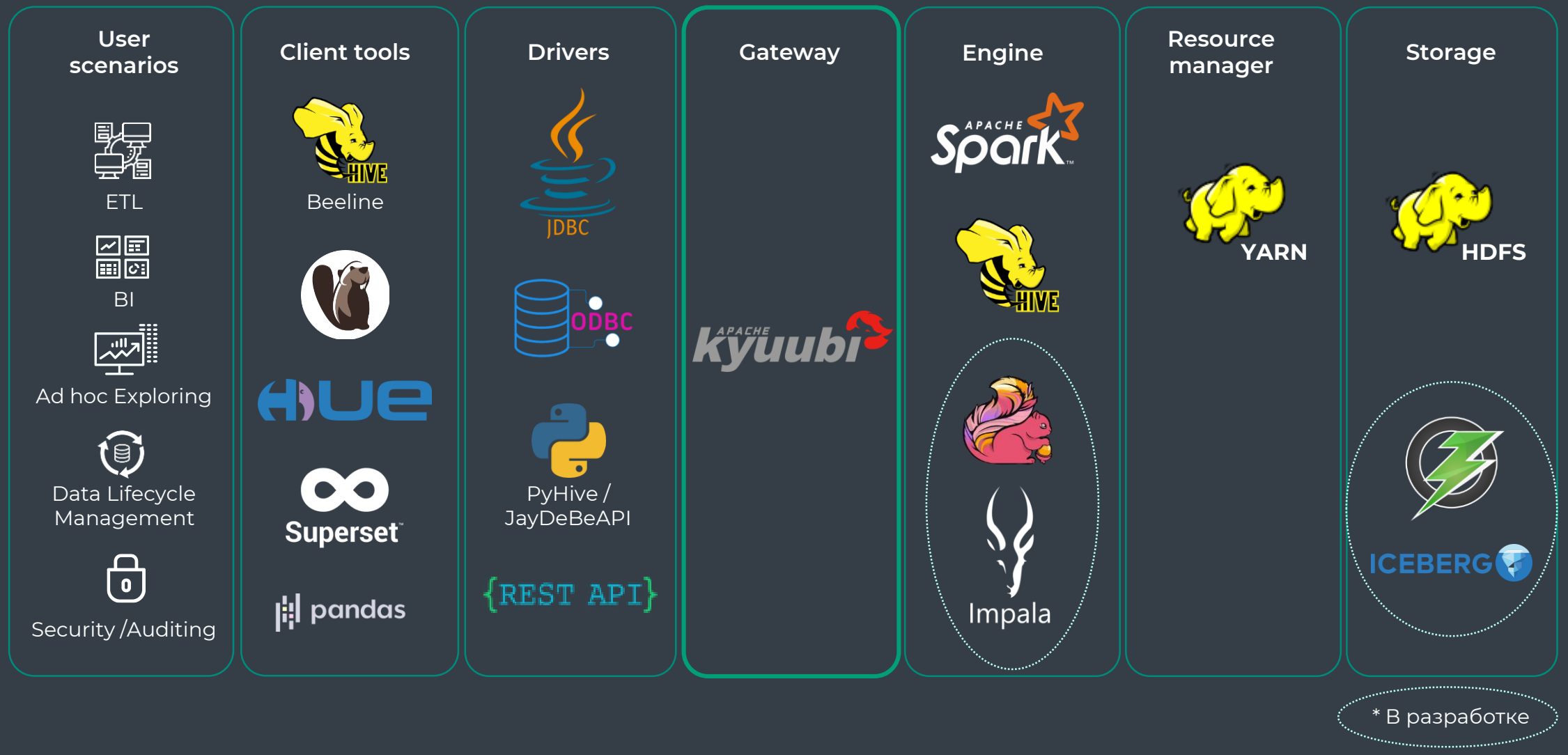
Пакетная обработка

- Kyuubi предоставляет интерфейс SQL, удобный для пакетной обработки ETL
- Kyuubi и его движки работают с многочисленными источниками данных независимо от хранилища
- Изоляция вычислительных ресурсов

Data Lake & Lakehouse

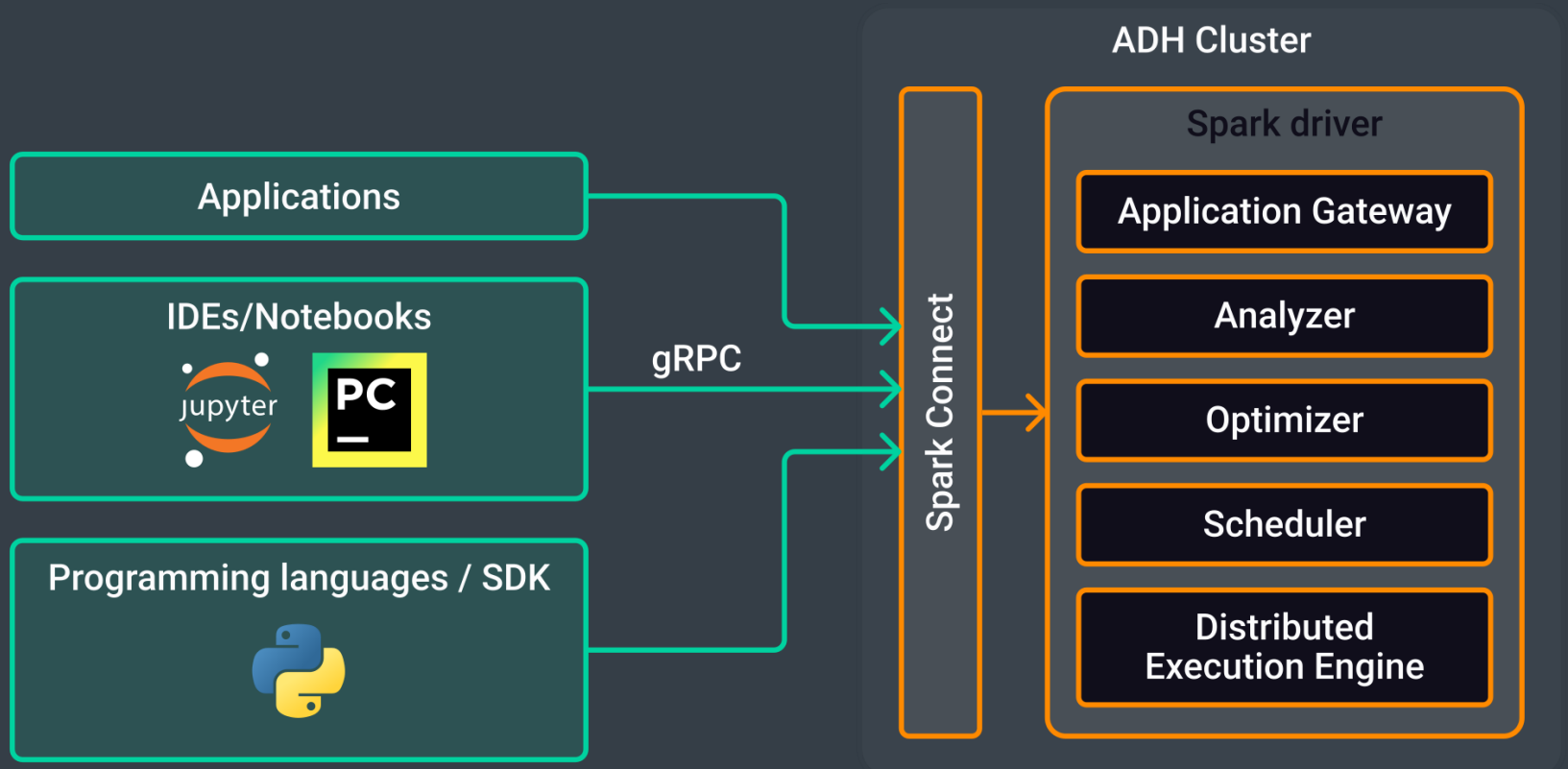
- Возможность выполнения запросов к традиционным хранилищам (Hive/HFDS) и современным озёрам данных.
- Централизованная картина данных
- Возможность запрашивать разнородные источники данных
- Аутентификация и авторизация (поддержка Kerberos)

Kyuubi: место в платформе данных



Spark Connect: удалённое управление кластером Spark

Spark Connect — компонент сервиса Spark3, который, выполняя функции тонкого клиента, обеспечивает удалённое подключение к кластерам Spark. С помощью Spark Connect можно удалённо управлять кластером Spark, например, используя привычную IDE на обычном пользовательском ноутбуке.



Верхнеуровневая архитектура Spark Connect

Arenadata Hadoop vs “ваниль”

Преимущества Arenadata Hadoop в сравнении
с Open Source Hadoop

ARENADATA

Преимущества Arenadata Hadoop в сравнении с Open Source

Качественная сборка совместимых компонентов

Hadoop — это множество сервисов, призванных взаимодействовать между собой. Самостоятельная сборка из исходников требует существенных вложений в RnD, либо будет выполнена без оглядки на совместимость, что скажется на стоимости эксплуатации и повлечет за собой простои.

Техническая поддержка от вендора с SLA, подтвержденным контрактом

Гарантии на поддержку платформы от вендора, включая штрафы, указанные в договоре, в сравнении с мотивацией команды собственных экспертов.

Универсальный оркестратор гибридного ландшафта

Предоставляем Arenadata Cluster Manager, объединяющий Hadoop и комплиментарные сервисы. ADCM можно развернуть на любой инфраструктуре, включая публичное облако.

Дополнительные возможности Enterprise версии

Высокопроизводительные коннекторы позволяют интегрировать Hadoop, ClickHouse, Greenplum и Tarantool между собой, а также с внешними системами.

Безопасность

Единая, интегрированная во все компоненты платформы, система безопасности Arenadata Platform Security на базе Kerberos, Ranger и Knox — в сравнении с частным решением, которое нужно постоянно дорабатывать и обновлять.

Преимущества Arenadata Hadoop в сравнении с Open Source

Набор типовых пакетных сервисов по планированию, установке и аудиту системы

Вам не придётся самостоятельно проводить оценку оборудования, которое потребуется для решения поставленной задачи. Наши специалисты настроят Arenadata Hadoop (удалённо или on-site), проведут аудит системы и помогут определить дальнейшие шаги.

Реестр отечественного ПО

Наличие регистрации продукта в реестре отечественного ПО не только отражает возможность снижения внешних рисков, но и напрямую влияет на стоимость для клиента за счёт экономии на НДС.

Пакет утилит для полной офлайн-установки

Arenadata Hadoop включает набор инструментов для автоматической установки и настройки компонентов, как на «чистом железе», так и в облаке. Средства мониторинга и управления конфигурацией кластера позволяют оптимизировать производительность для всех компонентов системы.

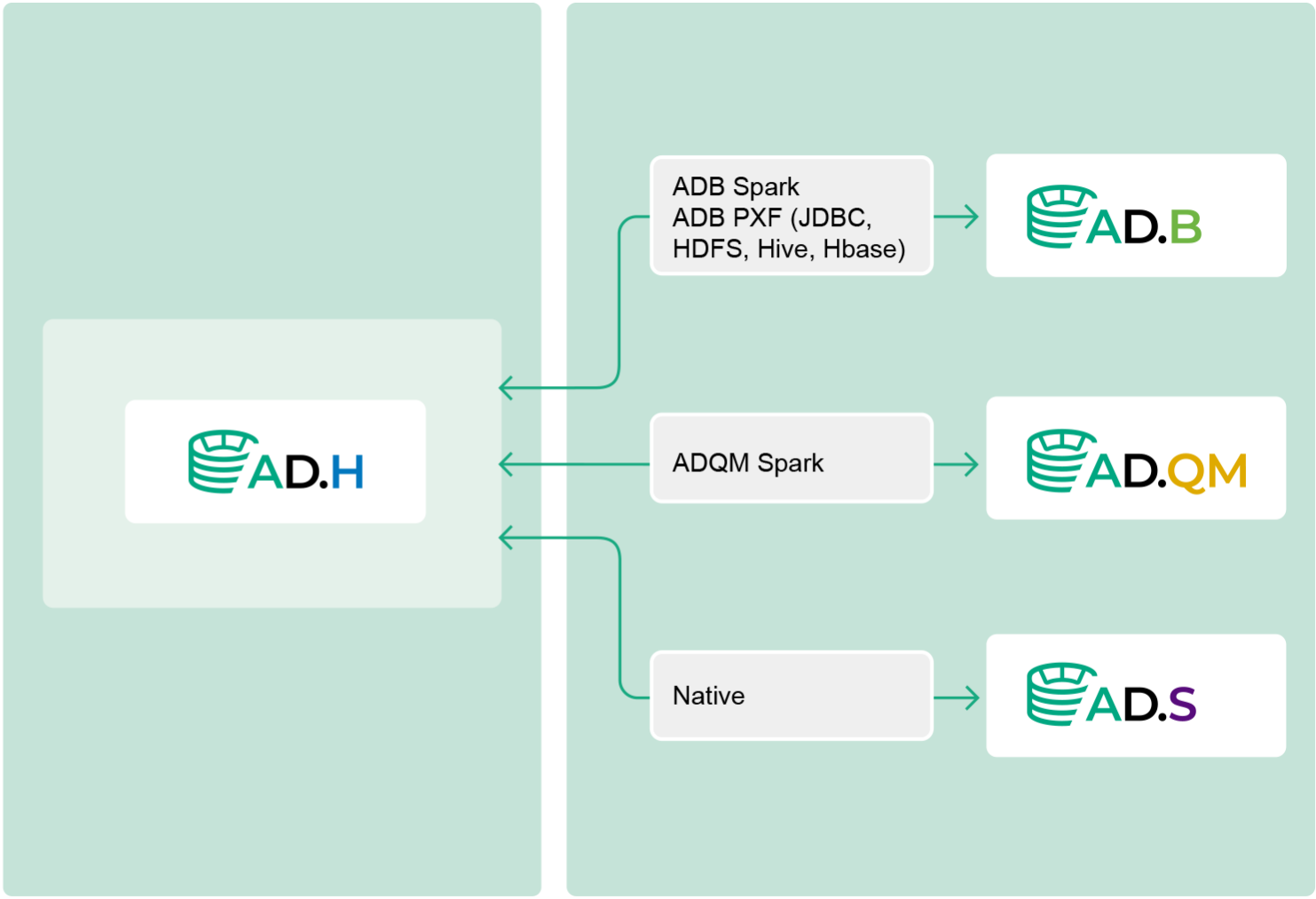
Документация

Оригинальная документация на русском и английском языках поможет облегчить процесс планирования, установки и настройки кластера Hadoop.

Коннекторы в платформе

Продукт Arenadata Hadoop обеспечен всеми необходимыми коннекторами для работы с другими компонентами платформы Arenadata EDP:

- **Arenadata DB**
- **Arenadata Streaming**
- **Arenadata QuickMarts**



* ADB PXF Hive и HBase коннектор работают только на чтение

Коннекторы в платформе

ADB Spark3 Connector

Обмен данными между Apache Spark и Arenadata DB

Интеграционное решение обеспечивает высокоскоростной параллельный обмен данными между Spark 3 и DWH на базе Arenadata DB (ADB).

ADQM Spark3 Connector

Обмен данными между Apache Spark и Arenadata QuickMarts

Многофункциональный коннектор с поддержкой параллельных операций чтения/записи между Spark 3 и Arenadata QuickMarts (ADQM).

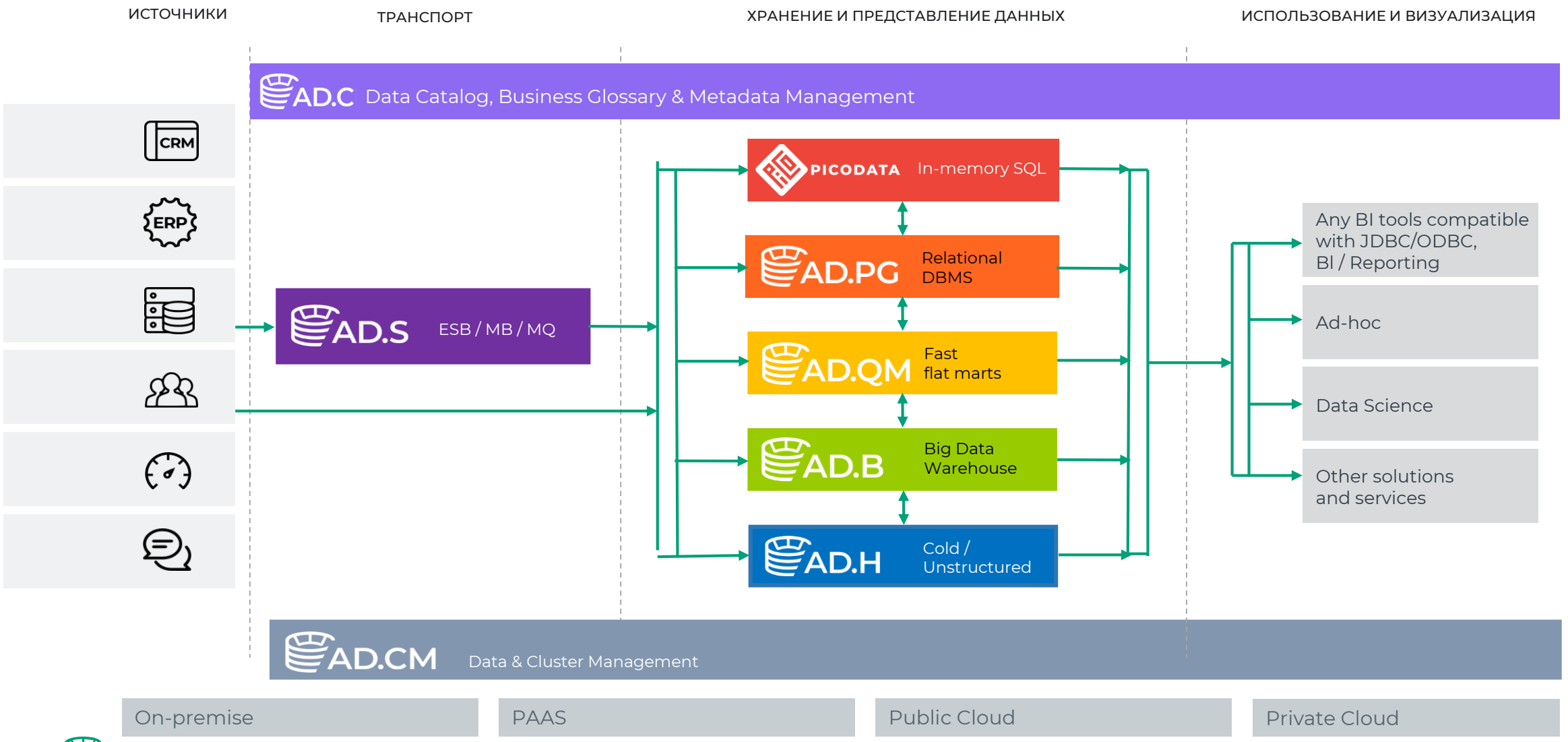
Возможности:

- высокая скорость передачи данных
- автоматическое формирование схемы данных
- гибкое партиционирование
- поддержка push-down операторов
- поддержка batch-операций

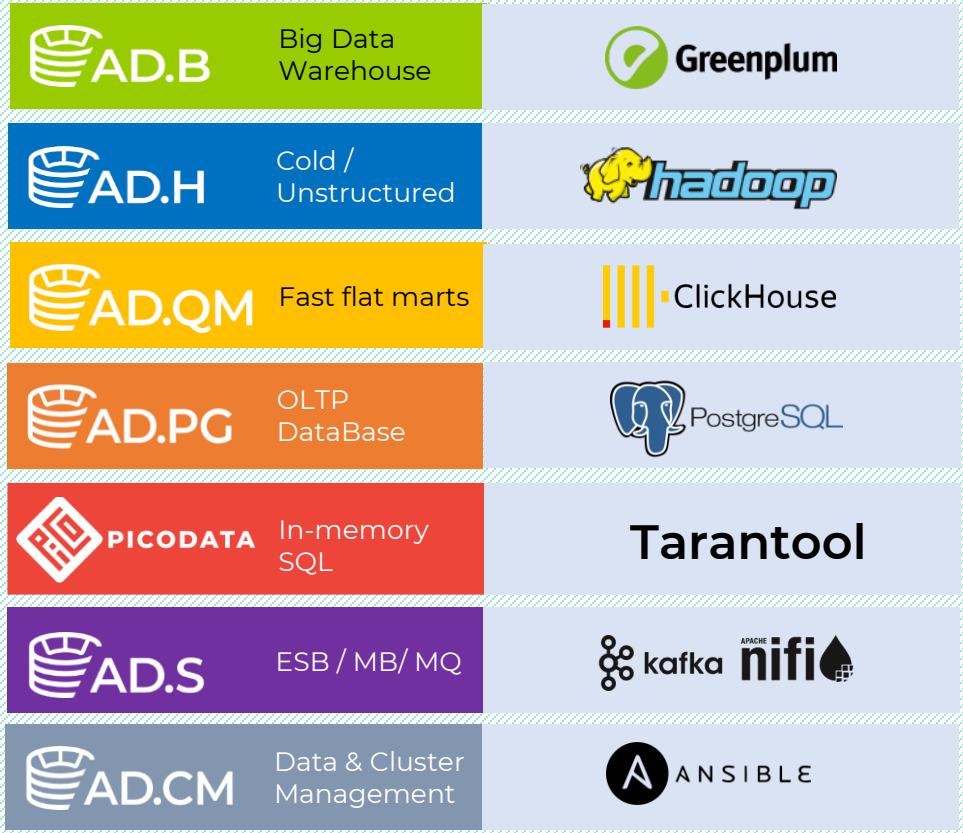
Arenadata EDP

ARENADATA

Arenadata Enterprise Data Platform



Что внутри?

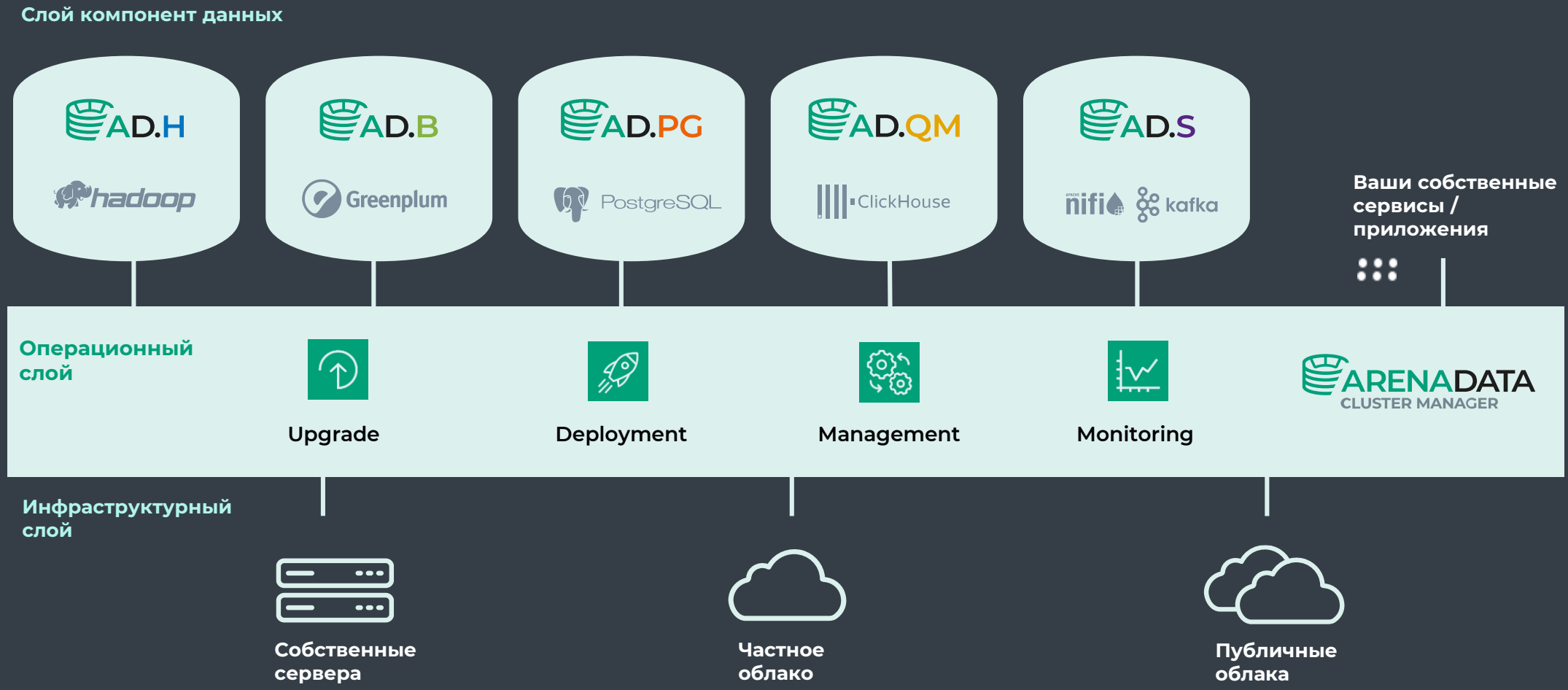


Каждый добавляемый компонент платформы доработан, чтобы стать её частью

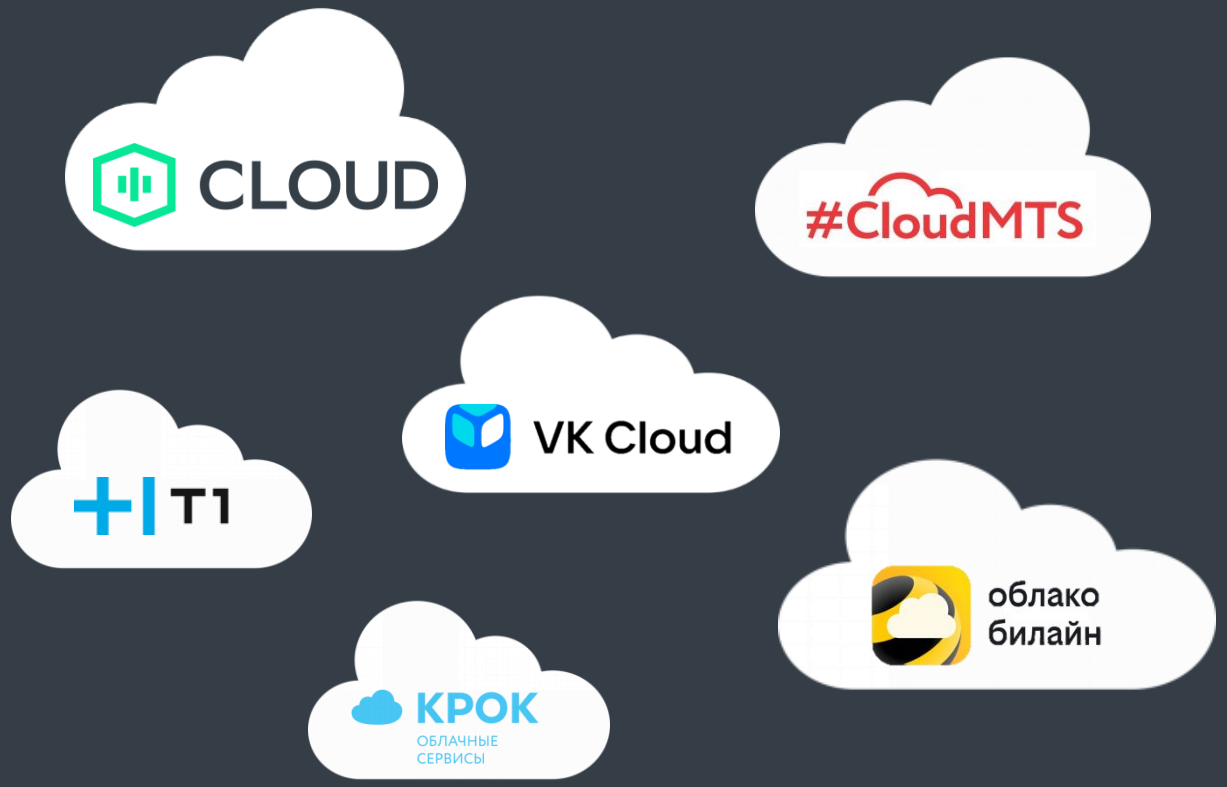
Производится:

- Анализ кода
- Интеграция с модулем мониторинга
- Проверка работоспособности заявленного функционала
- Подбор компетентного персонала поддержки и погружение в продукт архитекторов Professional Services
- Разработка коннекторов для взаимодействия с другими компонентами платформы
- Формирование дорожной карты доработок продукта

Гибридный корпоративный ландшафт



Enterprise Multi-cloud платформа данных



- Вы можете развернуть любой компонент платформы Arendata EDP как на bare-metal оборудовании, так и в облаке, или же воспользоваться востребованным сегодня сценарием — гибридной ИТ-инфраструктурой или Multi-clouds.

- Установите компоненты платформы распределённо на разные инфраструктуры, обеспечив между ними сетевую доступность.

Оставайтесь на связи

Наши новости
в телеграм-канале

ARENADATA