

Arenadata Streaming

Система для потоковой передачи и обработки данных



План презентации

1. О компании Arenadata
2. Arenadata Streaming: общее описание
3. Обзор продукта Arenadata Streaming
4. Типовые архитектурные шаблоны
5. ADS vs Open Source
6. Платформа сбора и хранения данных Arenadata Enterprise Data Platform

О компании Arenadata

Arenadata

лидер в области платформ больших данных для цифровой трансформации

СТАБИЛЬНОСТЬ

ПАО Группа Arenadata:

- 9 лет на рынке
- 600+ сотрудников
- 6 млрд ₽ выручка '24
- 30 млрд ₽ капитализация

150+ КЛИЕНТОВ

- ВТБ, ПСБ, ГПБ, Т-Банк
- ГПН, Росатом, Норникель
- X5, Магнит, Ашан, Hoff
- Мегафон, МТС, НН.ru
- ФНС, ДИТ МСК, Росреестр

ПРОДУКТЫ

- Мы создаем **тиражируемые продукты**, которые внедряют наши партнеры и клиенты
- 170+ релизов в год
- Обеспечиваем совместимость с **upstream open source**

ONPREM & CLOUD

- On-premise и ПАКи
- Cloud: VK Cloud, Cloud.ru, Selectel, MWS, TP Cloud, K2 Cloud, Beeline Cloud
- Гетерогенные ландшафты

#DATA

Дата-платформа в доменах:

- **Аналитика, Транзакции, Интеграция, Data Governance**
- Техническая поддержка 1, 2 и 3 линии и сервисы от вендора

ЭКСПЕРТЫ

- 150+ технологических партнеров и интеграторов
- 4 000+ экспертов подготовлено в центре обучения Arenadata

ВКЛАД В OPENSOURCE

- **Greengage**: преемник Greenplum
- **ClickHouse**: #1 контрибутор в РФ
- **Greengage, Kyuubi, Spark, Kafka, SSM, PXF**: коммитим в ядро
- **ADCM, Picodata**: собственные open source-проекты
- Открытая онлайн [документация](#)

РЕГУЛЯТОРЫ

- Отечественные ОС
- Реестр Минцифры
- ФСТЭК УД4
- ГосТех, ГЕОП



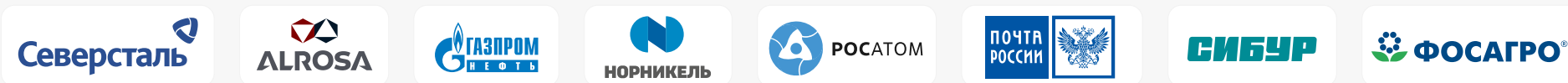
Наш успех – доверие клиентов

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

Банки и страховые: **40+ клиентов**



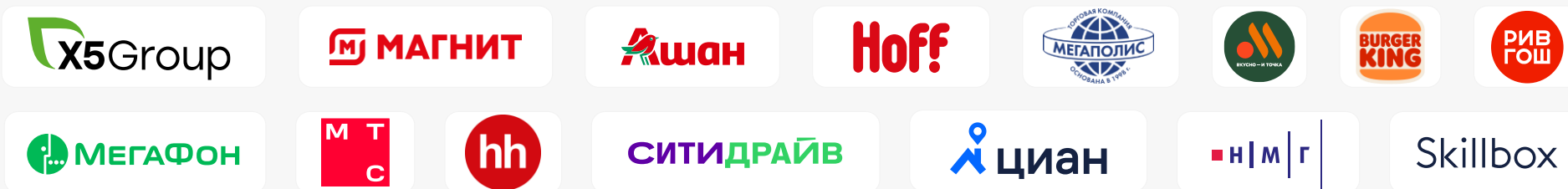
Промышленность, ТЭК, Транспорт: **40+ клиентов**



Госсектор: **30+ клиентов**



Retail, FMCG, Фарма, Telecom, Digital, Media: **30+ клиентов**

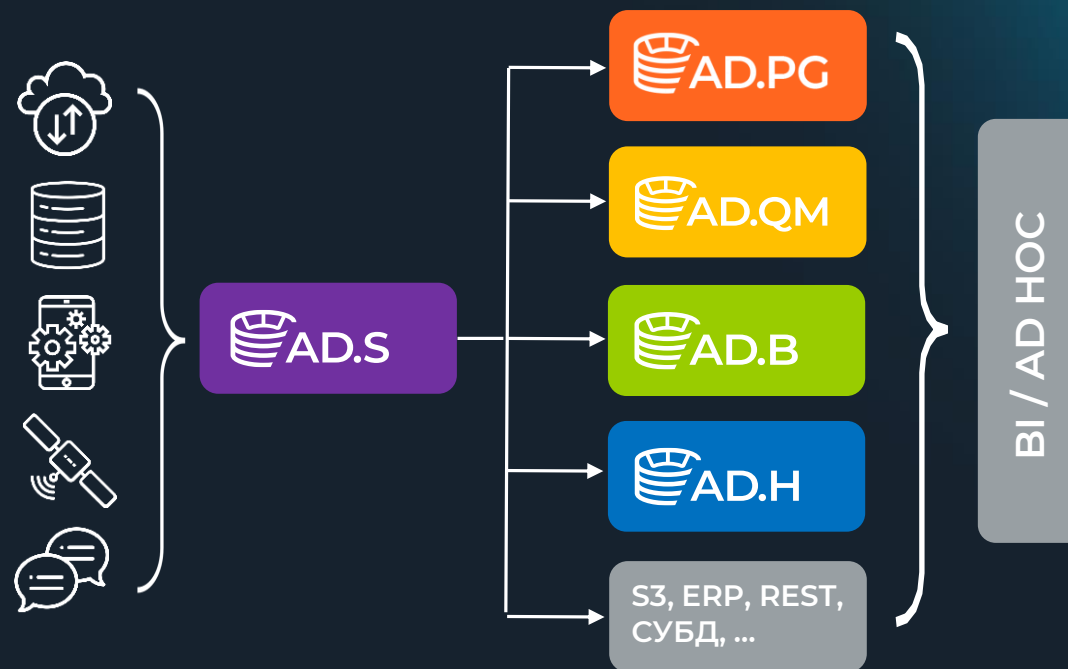


Arenadata Streaming: общее описание

Arenadata Streaming (ADS) — масштабируемая отказоустойчивая система для потоковой обработки данных в режиме реального времени, адаптированная для корпоративного использования и построенная на базе Apache Kafka и Apache NiFi

Сценарии использования:

- Корпоративная шина обмена данными для всех приложений
- Среда консолидации больших потоков данных и управления ими в режиме реального времени
- Инструмент разграничения прав доступа к потокам данных
- Фреймворк для разработки потоковых аналитических приложений в режиме реального времени без единой строки кода



Arenadata Streaming: различные задачи — одно решение



ETL

NiFi
NiFi Registry
MiNiFi

SOA (ESB)

Kafka
Kafka Connect
KSQL

Microservices architecture

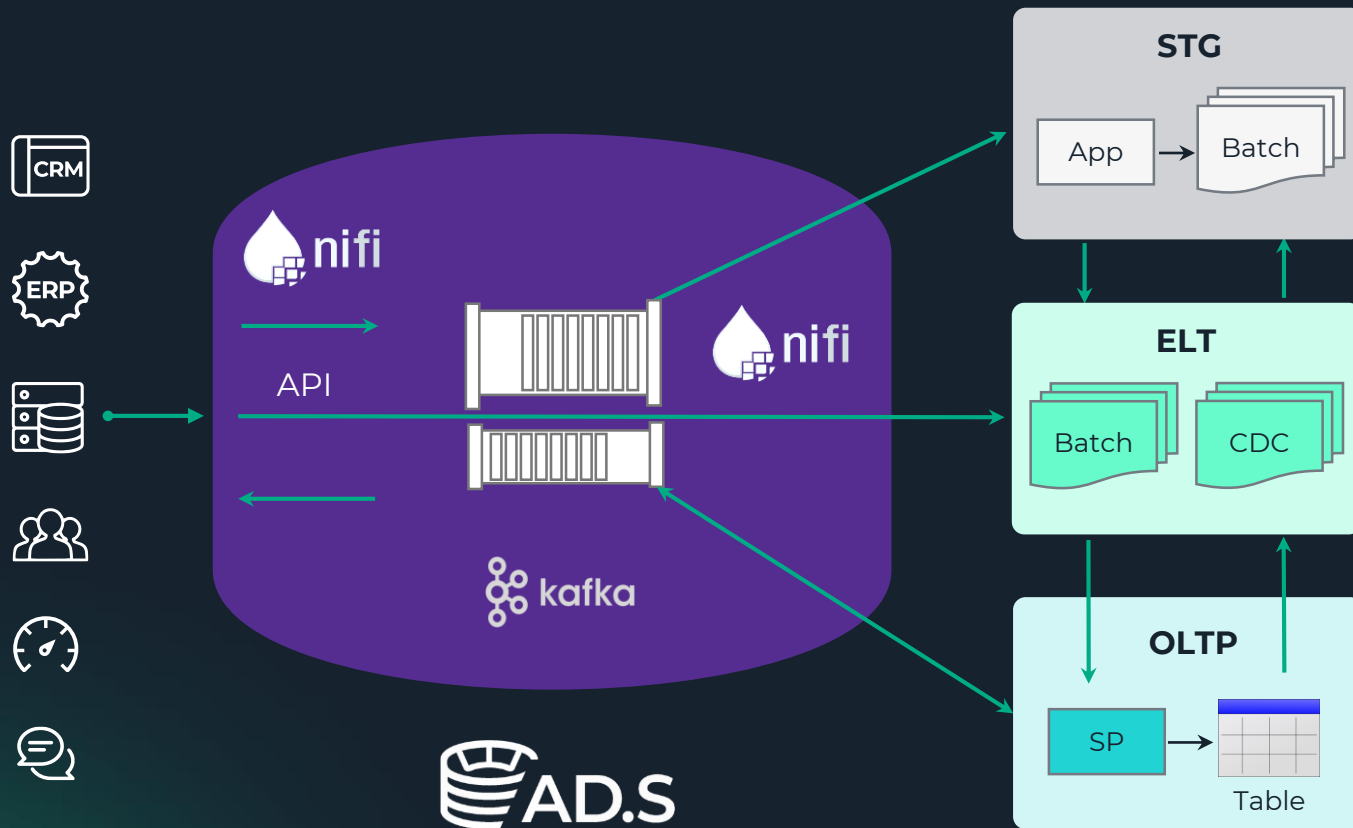
Kafka
Kafka REST Proxy

CDC интеграция

Kafka
Kafka Connect
ADS Control



Arenadata Streaming позволяет



- Строить потоковые конвейеры данных в реальном времени, которые надежно передают данные между системами или приложениями
- Строить потоковые приложения в реальном времени, которые преобразуют или реагируют на потоки данных
- Хранить потоки записей отказоустойчивым долговечным способом

Основные преимущества Arenadata Streaming

Производительность

ADS оперирует сотнями тысяч сообщений в минуту, эффективно потребляя как структурированную, так и неструктурированную информацию.

Надёжность

ADS снижает риски потери данных. Информация в системе доступна в течение любого заданного периода хранения.

Безопасность данных

ADS позволяет разграничивать права доступа к данным для защиты критически важной информации от несанкционированного доступа.

Низкий порог входа

ADS снижает порог входа в технологию, т.к. имеет редактор визуального проектирования и интуитивный GUI для управления и мониторинга и не требует навыков написания программного кода.

Гибкость

В отличие от брокеров сообщений, удаляющих информацию сразу после доставки, в ADS можно задавать любое время хранения данных, а также гибко настраивать другие характеристики, необходимые для корпоративного использования.

Масштабируемость

Система растёт вслед за потребностями бизнеса: можно добавлять новые кластеры или убирать неиспользуемые без приостановки работы.

Отказоустойчивость

Балансировка нагрузки и реализация отказоустойчивости на уровне приложений. Автобалансировка с помощью Kafka Cruise Control.

Экономия ресурсов CPU

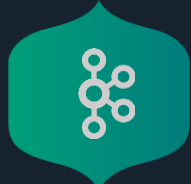
Снижает нагрузку на CPU при выполнении шифрования TLS за счёт поддержки kTLS в Kafka

Универсальность

Доступен широкий набор API для интеграции с внешними системами.

Arenadata Streaming: обзор продукта

Arenadata Streaming 3.9.0.1



Kafka 3.9.0

- Kafka REST Proxy 7.9.0
- KSQL 7.9.0
- Schema Registry 7.9.0
- Kafka Connect 3.9.0
- Kafka Cruise Control

Security

- Kerberos MIT KDC, AD, FreeIPA
- SSL/TLS поддержка
- Авторизация и аудит в Ranger
- Плагин авторизации в Kafka REST Proxy

ADS Control

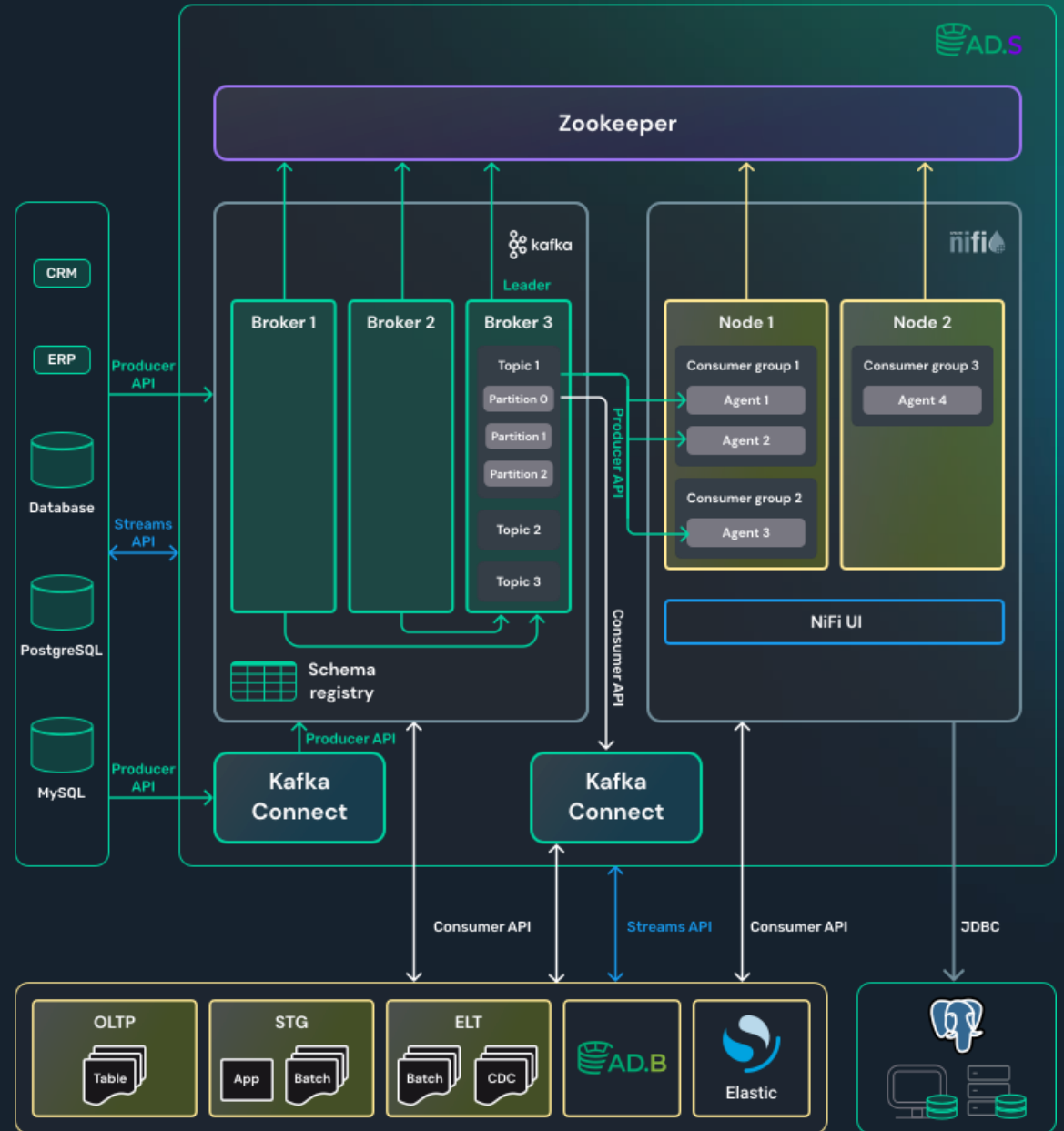
- GUI managing ADS clusters
- MirrorMaker 2
- Iceberg sink connector / Debezium source connectors

NiFi 1.28.0

- NiFi Registry 1.28.0
- MiNiFi 1.28.0
- NiFi ADB Processor

Security

- NiFi LDAP/LDAPS аутентификация
- SSL/TLS поддержка
- Авторизация и аудит в Ranger



Сервисы кластера ADS в интерфейсе ADCM

Cluster Manager

 Clusters

 Hostproviders

 Hosts

 Jobs

 Access manager

 Audit

 Bundles

 admin

 Settings

 Log out

Clusters

Clusters / ADS / Services

ADS 

Overview **Services** Hosts Mapping Configuration Import

Add service

Name	Version	State	Concerns	Actions
Kafka	2.8.1	installed		  
Kafka Connect	2.8.1	installed		  
Kafka REST Proxy	6.2.1	installed		  
Kafka-Manager	3.0.0.5_arenadata5	installed		  
MINIFI	1.20.0_arenadata1	installed		  
NI FI	1.20.0_arenadata1	installed		  
Schema-Registry	6.2.1	installed		  
Zookeeper	3.5.10	installed		  
ksqlDB	6.2.1	installed		  

Show 10 per page >

1 < >

Настройка сервисов в кластере ADS

The screenshot displays the 'Cluster Manager' web interface. The left sidebar contains navigation links: Clusters, Hostproviders, Hosts, Jobs, Access manager, Audit, Bundles, admin, Settings, and Log out. The main content area is titled 'Clusters' and shows the path 'Clusters / ADS / Services / Kafka / Primary Configuration'. The 'Services' tab is active, showing a table with one entry: 'Kafka' (version 2.8.1, 0 / 1 successful components). Below the table, the 'Primary Configuration' section is visible, showing a 'current' configuration group with a timestamp of '08/11/2023 09:32:40' and the state 'init'. A search input field and a 'Show advanced' toggle are present. The configuration tree on the right includes 'Main', 'Log Cleaner', 'kafka-env.sh', 'server.properties', 'Kafka Ranger plugin enabled: false', 'ranger-kafka-audit.xml', 'ranger-kafka-security.xml', and 'ranger-policymgr-ssl.xml'.

Cluster Manager

Clusters

Hostproviders

Hosts

Jobs

Access manager

Audit

Bundles

admin

Settings

Log out

Clusters

ADS

Overview

Services

Hosts

Mapping

Configuration

Import

Kafka

2.8.1

0 / 1 successful components

Actions

Delete

Primary Configuration

Configuration groups

Components

Info

1

current

08/11/2023 09:32:40

init

Search input

Show advanced

Revert

Save

Configuration

Main

Log Cleaner

kafka-env.sh

server.properties

Kafka Ranger plugin enabled: false

ranger-kafka-audit.xml

ranger-kafka-security.xml

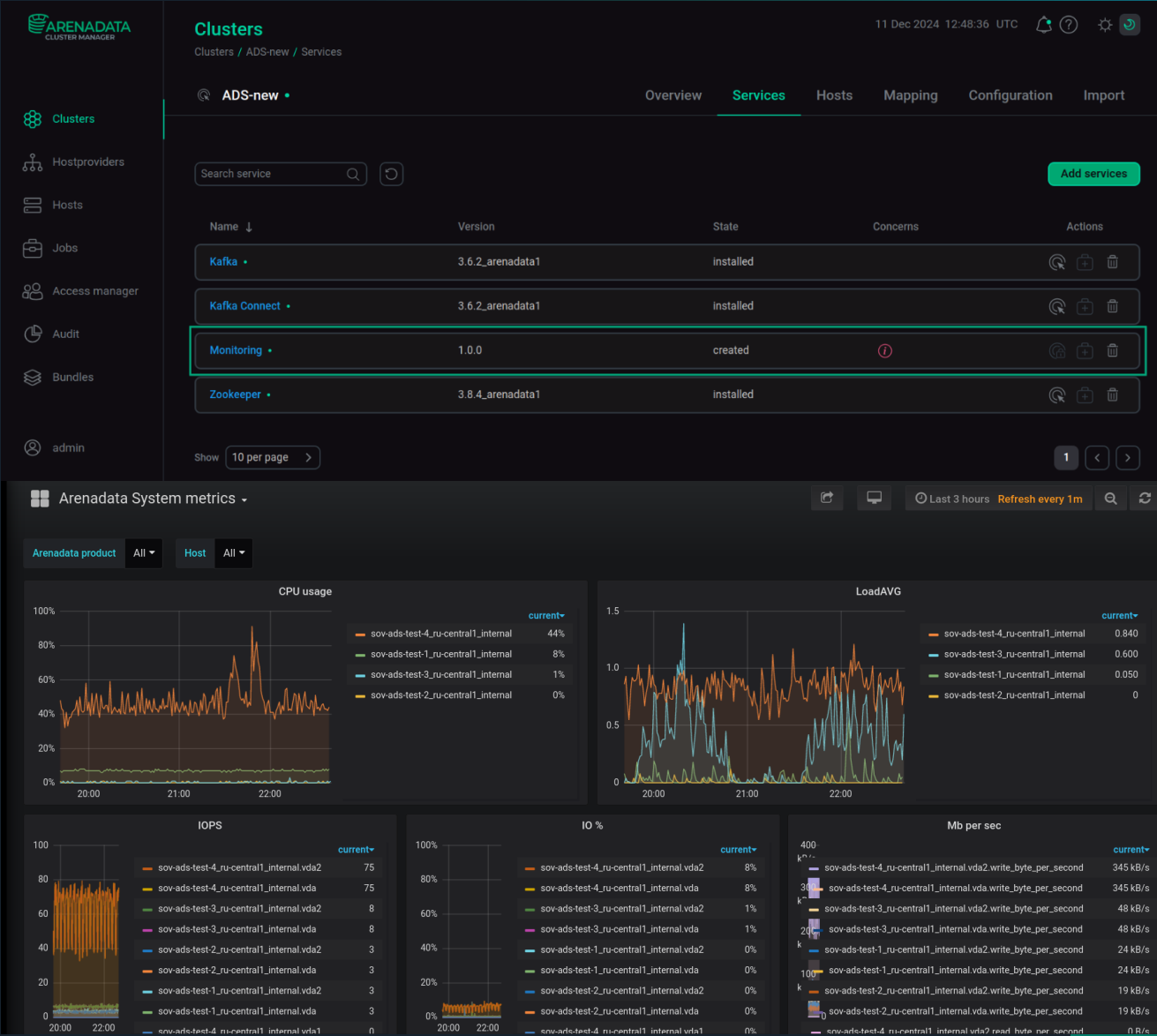
ranger-policymgr-ssl.xml

Мониторинг в кластере ADS

Мониторинг в ADS реализован в два уровня:

- 1. Сбор и хранение системных метрик на уровне хостов.
- 2. Сбор и хранение метрик от Kafka и NiFi и отображение их в разработанных панелях мониторинга.

Также мониторинг может выполняться на уровне приложений NiFi и Kafka.



Настройка сервиса Monitoring

Рекомендуемый способ мониторинга кластера ADS — сервис Monitoring

- Сервис разворачивает свой сервер Prometheus, который собирает и хранит метрики мониторинга кластера ADS
- Есть возможность использования веб-приложения Grafana для визуализации и анализа информации



The screenshot displays the 'ARENADATA CLUSTER MANAGER' interface. The left sidebar contains navigation links: Clusters, Hostproviders, Hosts, Jobs, Access manager, Audit, Bundles, admin, Settings, and Log out. The main panel is titled 'Clusters' and shows the 'ADS-new' cluster selected. The 'Mapping' tab is active, displaying a list of components to be monitored. The components are: Kafka (checked), Kafka Connect (checked), and Monitoring (checked). Under 'Kafka', there is a 'Kafka Broker' section with 1/1 instances, showing 'sov-ads-test-2'. Under 'Kafka Connect', there is a 'Kafka Connect Worker' section with 1/1 instances, also showing 'sov-ads-test-2'. Under 'Monitoring', there are three sections: 'Prometheus Server' (0/1), 'Grafana' (0/1), and 'Node-exporter' (0/0), each with an 'Add hosts' button. The top right corner shows the date and time: 11 Dec 2024 12:51:24 UTC.

Почему Kafka

— Горизонтальная масштабируемость

Возможность добавления новых источников данных и узлов в единый кластер

— Отказоустойчивость

Обеспечивается архитектурой master-slave, механизмом реплицирования и синхронизацией реплик

— Высокая производительность

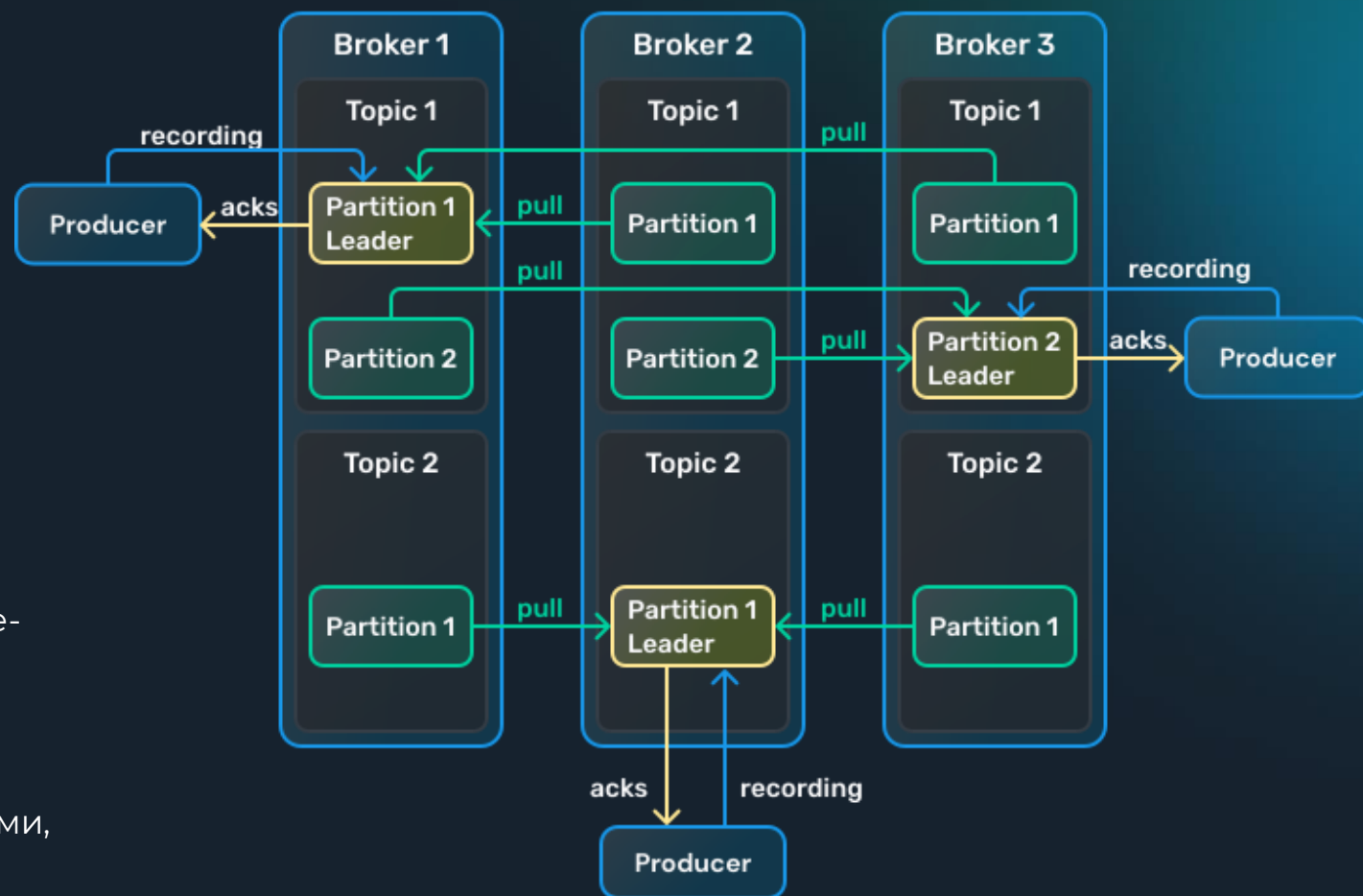
Высокая пропускная способность: чтение и запись более миллиона событий в секунду

— Безопасность

Несколько уровней изоляции для транзакций «чтение-обработка-запись»

— Гибкость

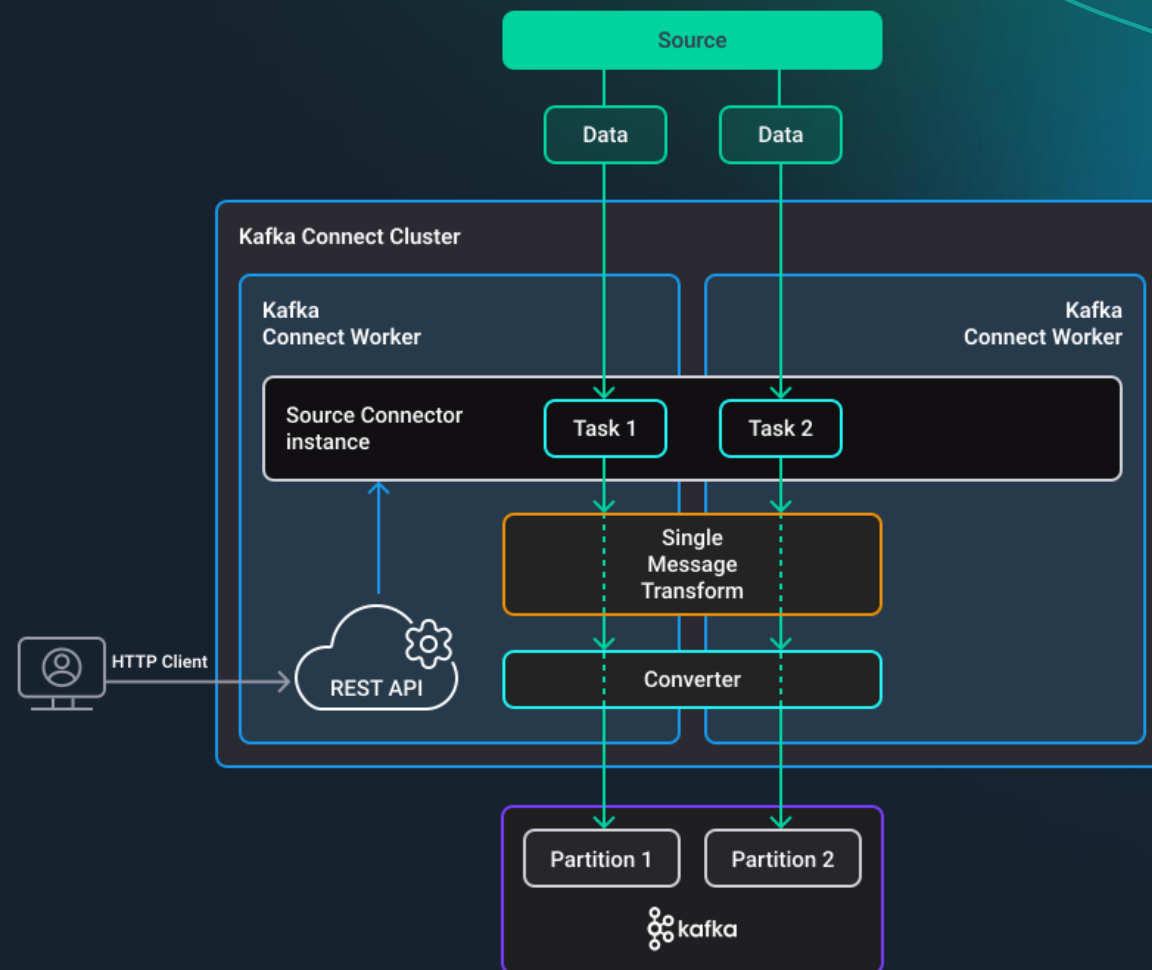
Интегрируемость с различными системами, решениями, структурами и форматами предоставления данных



Экосистема Kafka: Kafka Connect — инструмент для потоковой передачи данных между Kafka и другими системами

Возможности Kafka Connect:

- Копирование данных (создание коннекторов и задач, управление ими) с помощью REST-интерфейса
- Широкие возможности масштабируемости: изменение количества исполнителей, подключение новых коннекторов к имеющимся платформам, управление количеством запущенных задач
- Автоматическое управление смещением данных при работе коннекторов
- Широкий выбор уже существующих плагинов коннекторов, а также возможность создать собственный
- Интеграция потоковой/пакетной обработки данных

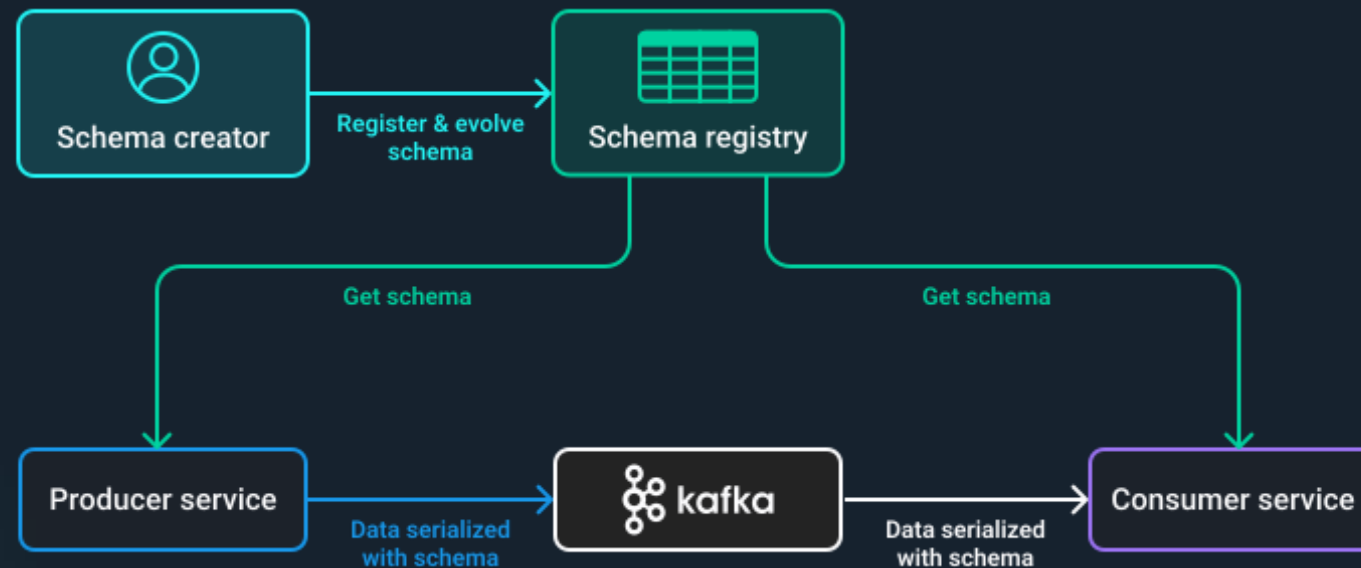


Архитектура Kafka Connect для переноса данных из внешней системы (source) в Kafka

Экосистема Kafka: Schema Registry

Централизованный репозиторий, использующийся для обеспечения согласованности форматов данных между производителями и потребителями сообщений в Kafka

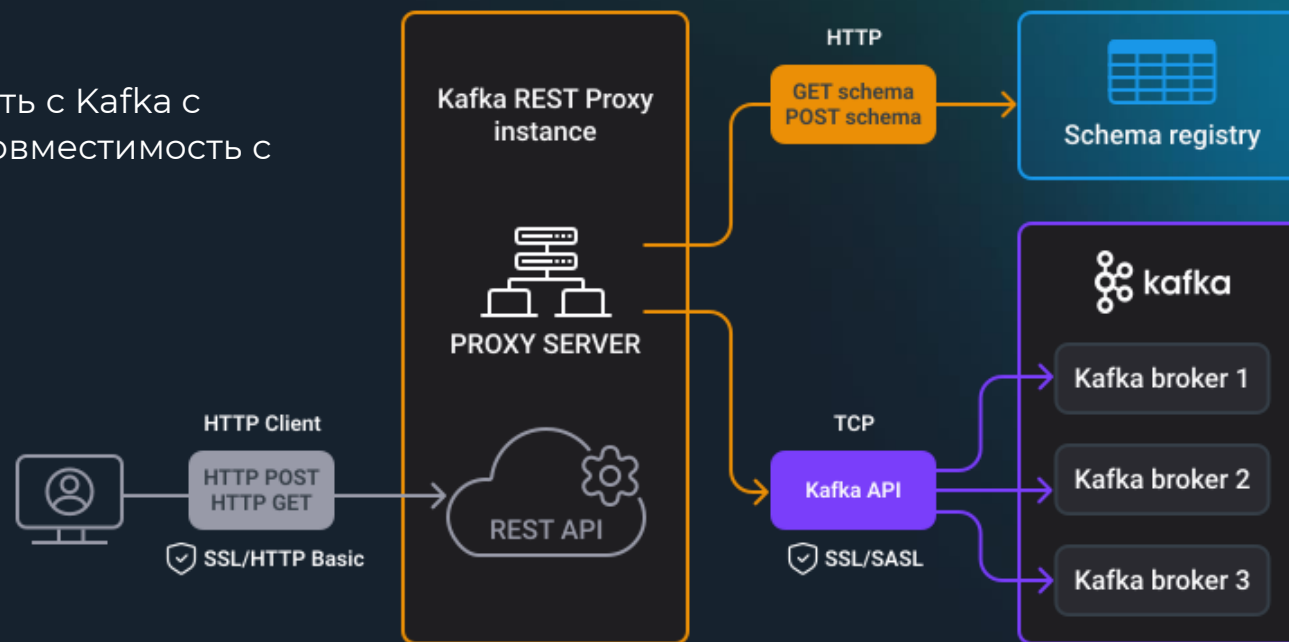
- Обеспечение сериализации и десериализации данных с помощью сериализаторов, предоставляемых клиентам Kafka
- Взаимодействие с клиентами Kafka для записи и чтения сериализованных сообщений
- Регистрация и хранение схем данных для каждого субъекта
- Поддержка эволюции схем с хранением каждой версии с собственным идентификатором и возможностью проверки совместимости версий для каждого субъекта
- Обеспечение доступности для осуществления действий со схемами и субъектами по интерфейсу RESTful с помощью Schema Registry API



Экосистема Kafka: Kafka REST Proxy — прокси-сервер для взаимодействия с кластерами Kafka

Возможности Kafka REST Proxy:

- Интерфейс HTTP RESTful позволяет взаимодействовать с Kafka с помощью стандартных методов HTTP, обеспечивая совместимость с различными языками программирования
- Взаимодействие с Kafka API через прокси-сервер:
 - создание и отправка сообщений в топики Kafka в различных форматах;
 - подписка на топики и чтение сообщений в режиме реального времени;
 - получение данных о кластере и топиках.
- Интеграция со Schema Registry для создания и развития схем данных
- Механизмы аутентификации и авторизации для защиты доступа к кластерам Kafka.



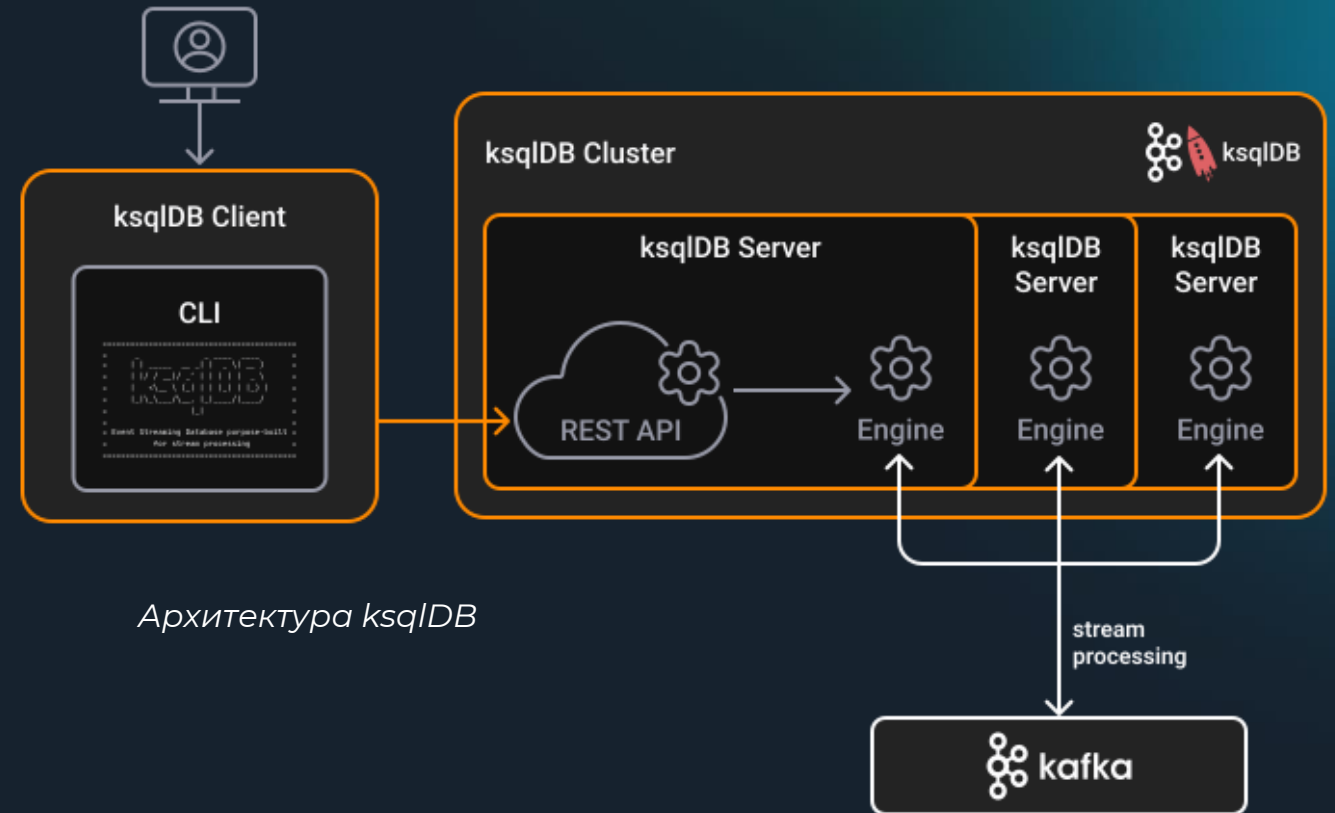
Архитектура Kafka REST Proxy

Экосистема Kafka: ksqlDB — база данных Kafka для потоковой обработки

Обработывает данные, хранящиеся в Kafka, и объединяет Kafka Streams для потоковой обработки и Kafka Connect для сбора и обработки событий из различных источников данных.

Возможности ksqlDB:

- Моделирование данных, хранящихся в Kafka, в виде потоков или таблиц посредством SQL
- Выполнение push-запросов на передачу данных
- Выполнение pull-запросов на получение данных
- Создание материализованных представлений из потоков и таблиц
- Создание коннекторов для интеграции с внешними хранилищами данных

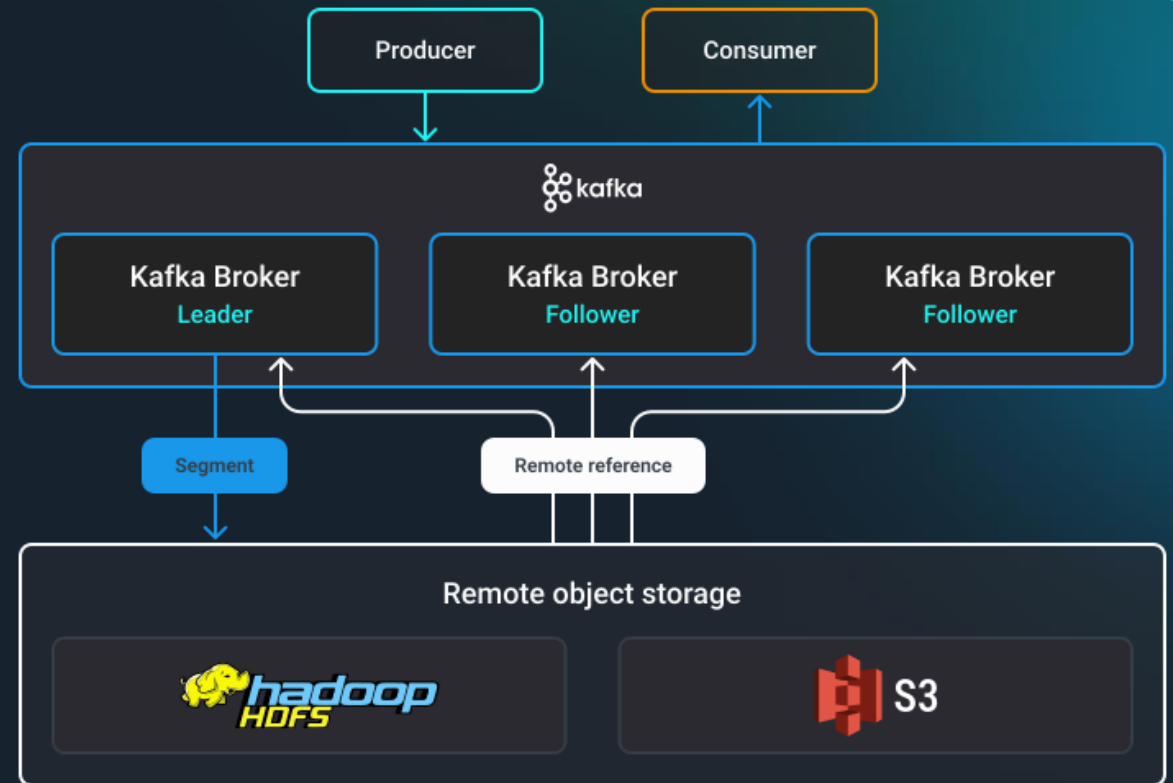


Архитектура ksqlDB

Kafka Tiered Storage — реализация многоуровневого хранения данных в топиках Kafka

Новый слой хранения для исторических данных за пределами файловой системы брокеров:

- Поступающие «горячие» данные располагаются в локальной файловой системе, а далее, согласно retention-политикам, переносятся во внешнюю систему
- В качестве внешней системы, адаптированной для долгосрочного хранения «холодных» данных, поддерживаются HDFS и S3
- Для пользователя интерфейс взаимодействия с данными не меняется, вне зависимости от их физического местонахождения
- Опция может быть включена как глобально для всего кластера, так и для отдельных топиков



Kafka Tiered Storage — верхнеуровневая архитектура

RemoteStorageManager (pluggable)

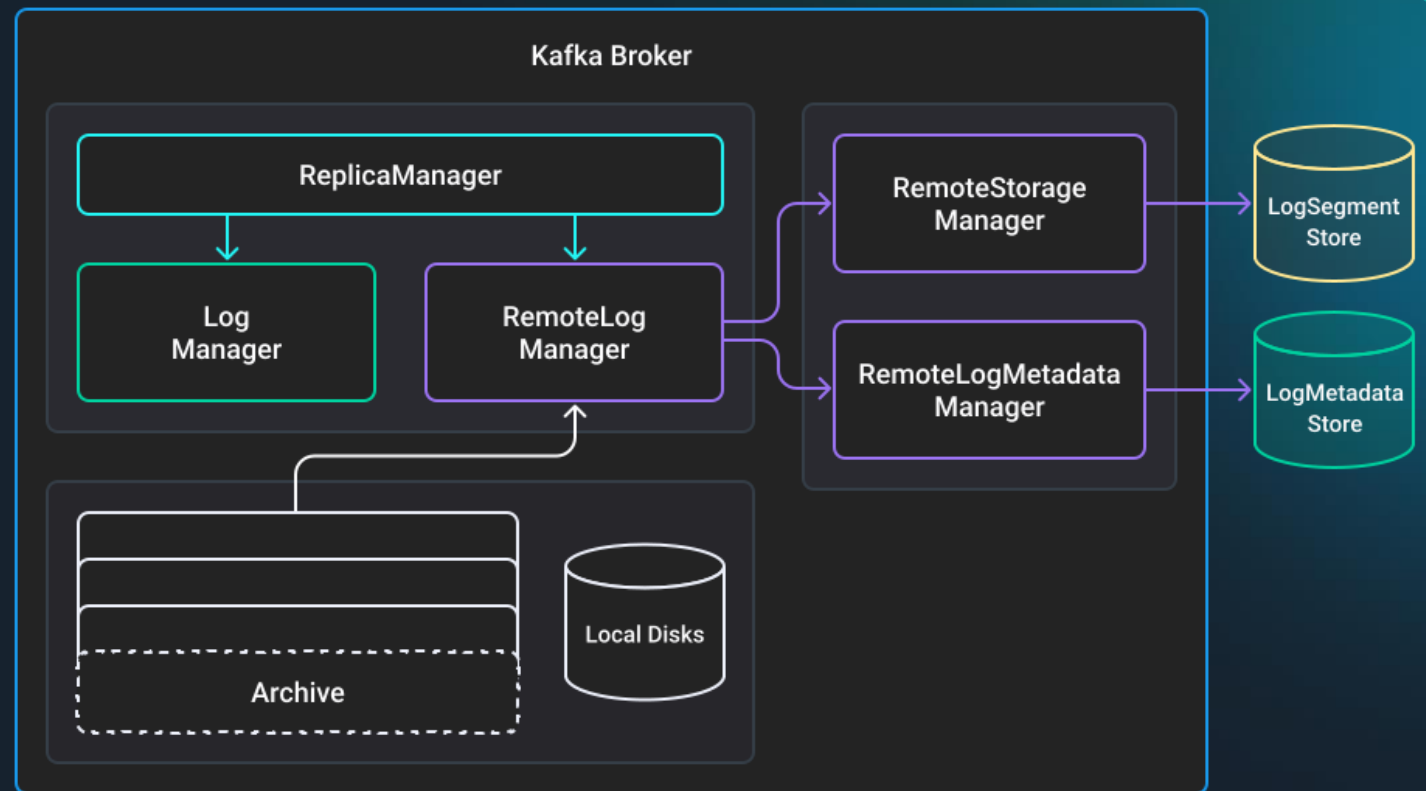
Управление хранением удалённых сегментов и вспомогательных индексов

RemoteLogMetadataManager (pluggable)

Управление хранением метаданных удалённых сегментов

RemoteLogManager (internal)

Точка входа для взаимодействия с удалёнными сегментами и управления — RemoteStorageManager и RemoteLogMetadataManager



Kafka Tiered Storage — мотивация для внедрения

Без Tiered Storage

С Tiered Storage

✓ стоимость хранения	Нет смысла хранить исторические данные на дорогих «быстрых» дисках	Уменьшаем расходы на хранение исторических данных путём их переноса на адаптированные для долгосрочного хранения носители
✓ отсутствие независимого масштабирования	При увеличении хранилища кластера посредством добавления нового узла мы также добавляем ненужные вычислительные мощности	Из-за уменьшенного количества хранимых локально данных, один узел кластера может хранить большее число партиций, тем самым снижая необходимость в масштабировании кластера
✓ операционная сложность на больших кластерах	Время актуализации нового брокера и ребалансировки кластера прямо пропорционально количеству данных	При уменьшении количества локальных данных, снижается время актуализации новых брокеров и ребалансировки кластера

Kafka Tiered Storage: преимущества решения

- ✓ Дешёвое хранение исторических данных
- ✓ Уменьшение стоимости обслуживания кластера
- ✓ Cloud ready
- ✓ Нет негативного влияния на пользователя



Быстрые операции обслуживания

- Процессы восстановления
- Partition reassignment
- Рестарт



Операции изменения топологии кластера

- Быстрое масштабирование в зависимости от вычислительной нагрузки вверх/вниз
- Меньше хостов = дешевле обслуживание и инфраструктура
- Отсутствие (отложенное) необходимости в масштабировании в зависимости от роста данных



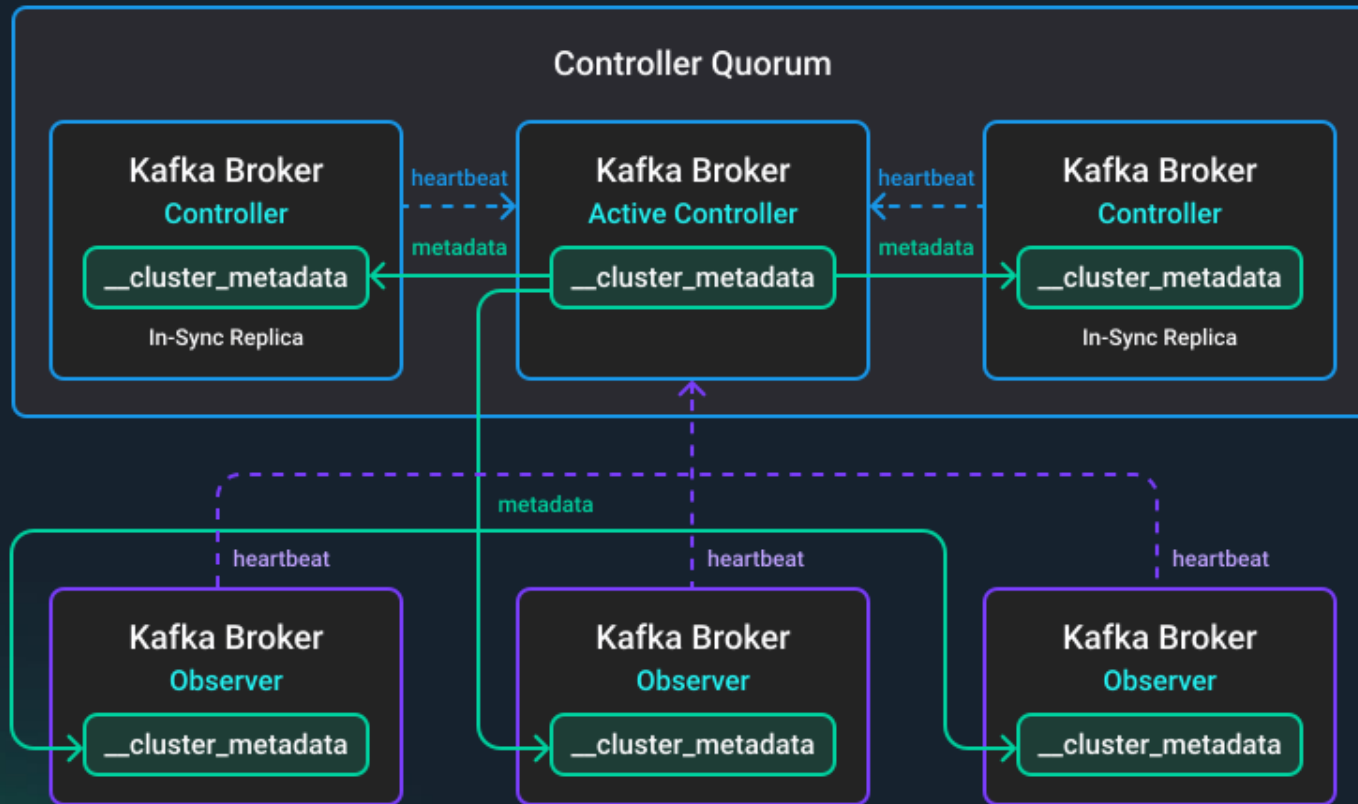
Kafka Raft (KRaft) — протокол, основанный на алгоритме консенсуса Raft, для управления метаданными кластера Kafka

Организует хранение информации о конфигурации кластера, обеспечивает синхронизацию брокеров Kafka при репликации данных без использования механизмов выбора лидера и хранения метаданных, предоставляемых ZooKeeper.

Преимущества KRaft:

- ✓ Быстрое и эффективное хранение и распространение метаданных: обновления отправляются непосредственно в кворум контроллера, устраняя дополнительный узел передачи (ZooKeeper)
- ✓ Брокеры и контроллеры могут хранить метаданные локально в своей кеш-памяти и обращаться к ним, не используя соединение с внешним сервисом
- ✓ Уменьшение задержек при завершении работы брокеров
- ✓ Протокол KRaft позволит значительно увеличить предел количества партиций для одной ноды и всего кластера

Kafka Raft: кворум контроллера в режиме KRaft



- Упрощает администрирование кластера
- Нет зависимостей от внешних сервисов
- Увеличивает масштабирование и снимает ряд ограничений
- Единая модель конфигурации и безопасности для всей системы
- Улучшено время переключения контроллера при отказе

Kafka Cruise Control: оптимизация работы кластеров Kafka

Инструмент для автоматического управления балансировкой рабочих нагрузок и масштабированием кластеров Kafka

Позволяет перераспределять нагрузку и обеспечивать высокую доступность подключений для кластера ADS: Kafka в случае сбоя брокеров. Упрощает управление крупными кластерами, снижая накладные расходы на их поддержку

Возможности Kafka Cruise Control:

Автоматизация

Позволяет автоматически управлять балансировкой нагрузки и перераспределением данных, снижая необходимость ручного вмешательства.

Мониторинг

Отслеживает состояние кластеров, выявляя потенциальные проблемы и обеспечивая их своевременное устранение.

Оптимизация

Помогает повысить производительность и эффективность использования ресурсов за счет интеллектуальной настройки и балансировки.

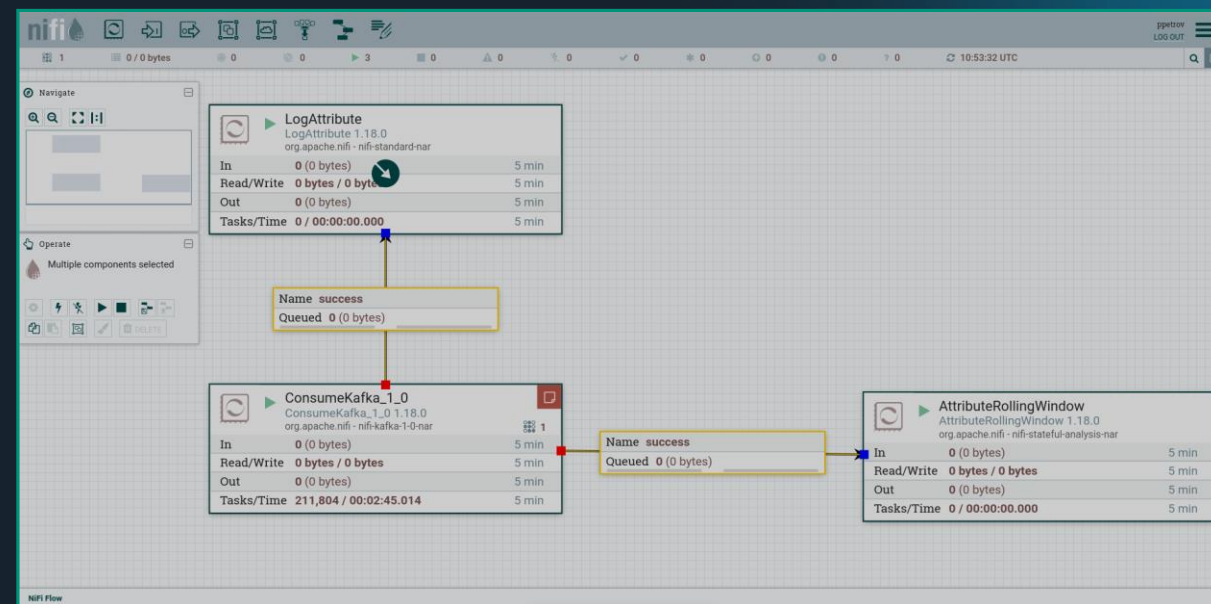
Почему NiFi

- Интуитивно-понятный пользовательский интерфейс
- Обеспечение информационной безопасности
- Мощная и масштабируемая кластерная архитектура
- Простая интеграция с источниками и приёмниками
- Поддержка любых форматов данных



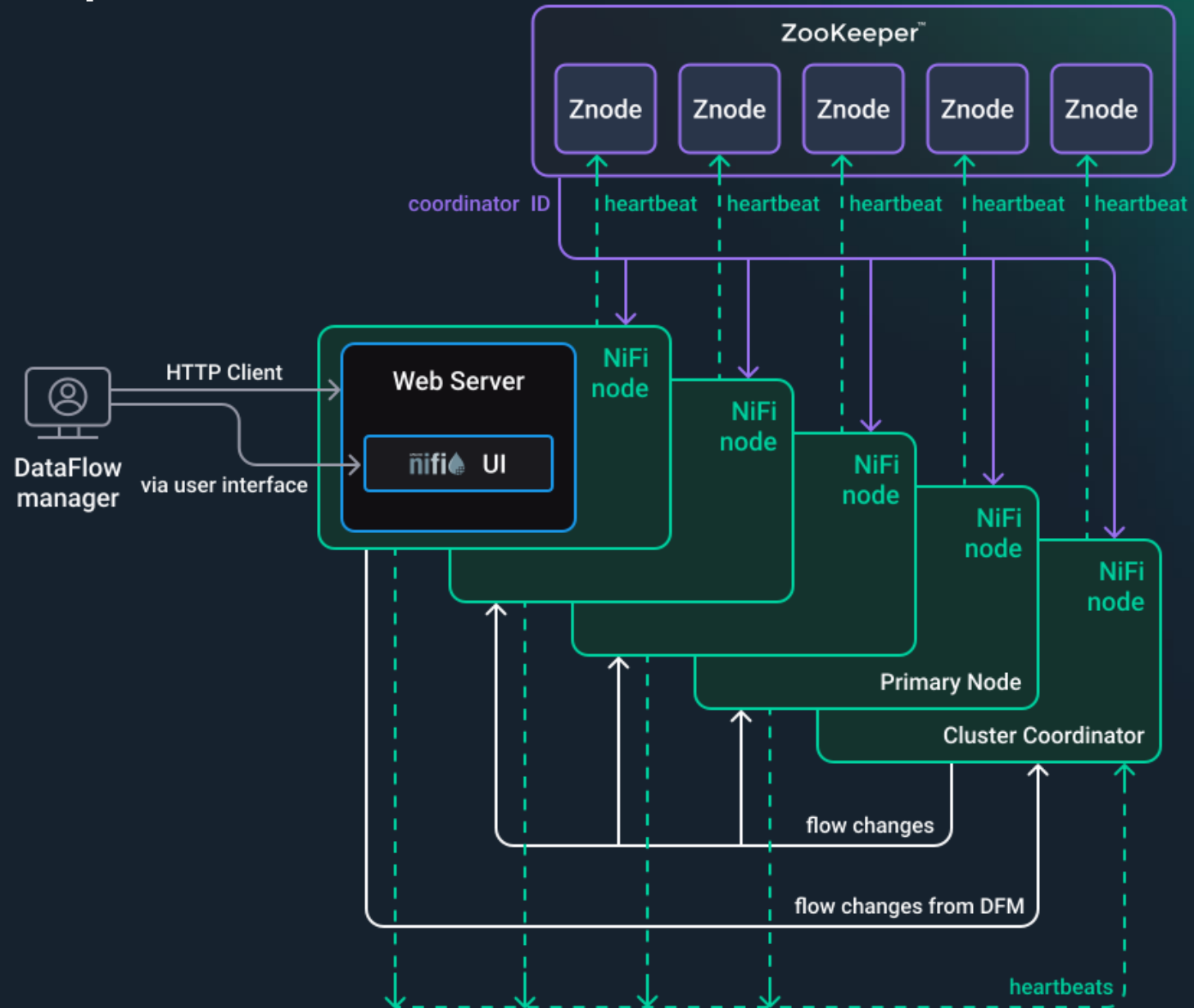
Основные возможности:

- Масштабирование процессов обработки
- Доступ к СУБД через JDBC (драйвер Java)
- Готовые процессоры доступа к данным (ExecuteSQL, PutSQL, PutDatabaseRecord, GenerateTableFetch)
- Собственная разработка: NiFi ADB Processor для высокоскоростной параллельной записи данных в Arenadata DB*
- Работа с большим набором форматов данных (Json, CSV, Avro, Parquet, XML, plain text)
- Широкие возможности трансформаций (XSLT, UpdateRecord, JOLT, JSLT, XSLT, SQL Calcite)



* NiFi ADB Processor доступен в режиме technology preview и не рекомендован для продуктивного использования

Архитектура кластера NiFi

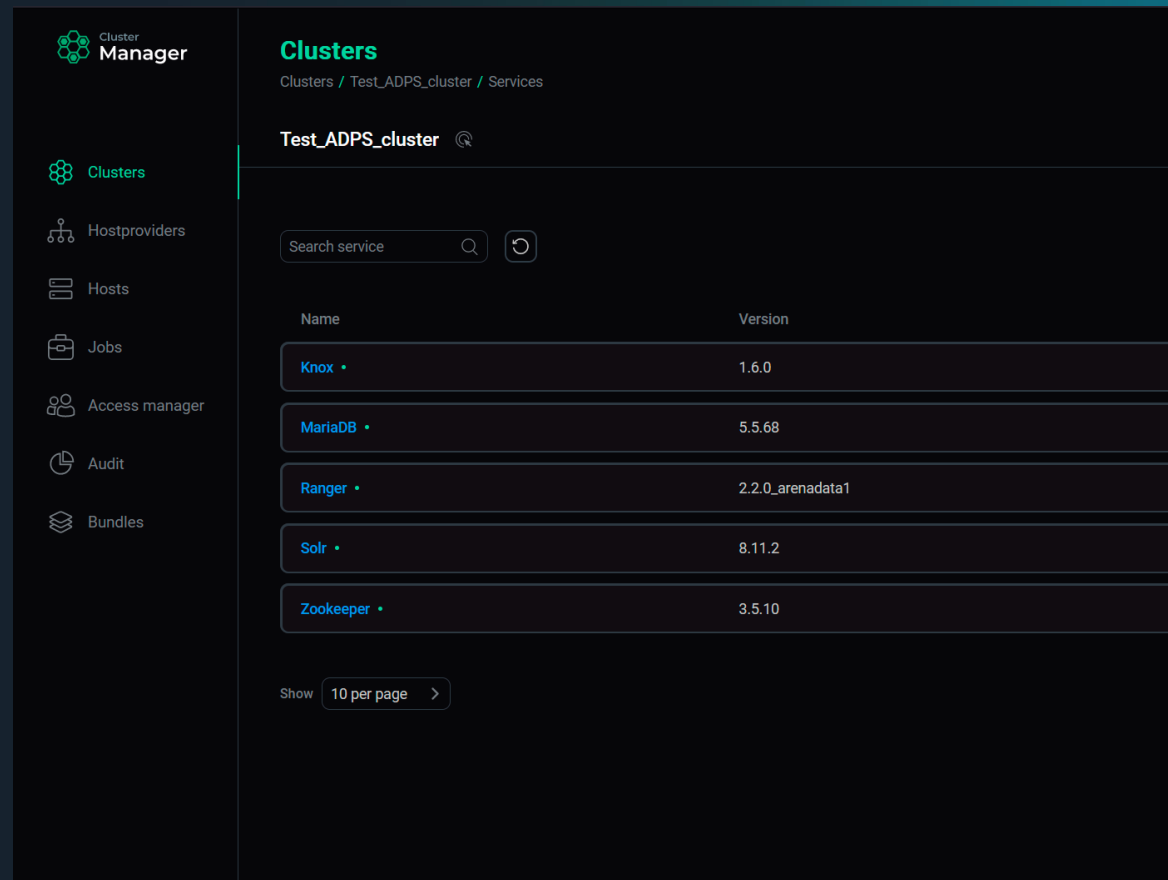


Безопасность в Arenadata Streaming

- Поддержка Apache Ranger для кластеров ADS (Plugins: Kafka, NiFi, NiFi Registry)
- Настройка политик безопасности и ролевого доступа Kafka
- Аудит действий пользователей
- Управление ключами шифрования на базе Ranger KMS
- Поддержка «из коробки» множества кластеров и сред Test/Dev/Prod с использованием Security Zones



Система управления групповыми политиками безопасности



Cluster Manager

Clusters / Test_ADPS_cluster / Services

Test_ADPS_cluster

Search service

Name	Version
Knox	1.6.0
MariaDB	5.5.68
Ranger	2.2.0_arenadata1
Solr	8.11.2
Zookeeper	3.5.10

Show 10 per page

Собственная разработка Arenadata

Функции безопасности Arenadata Platform Security

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.

01

Простая настройка SSL/TLS протокола:

- Заполнить параметры по инструкции в UI ADCM
- Кликнуть Enable SSL

02

Kerberos аутентификация в несколько кликов:

- Заполнить параметры по инструкции в UI ADCM
- Нажать на кнопку Enable Kerberos

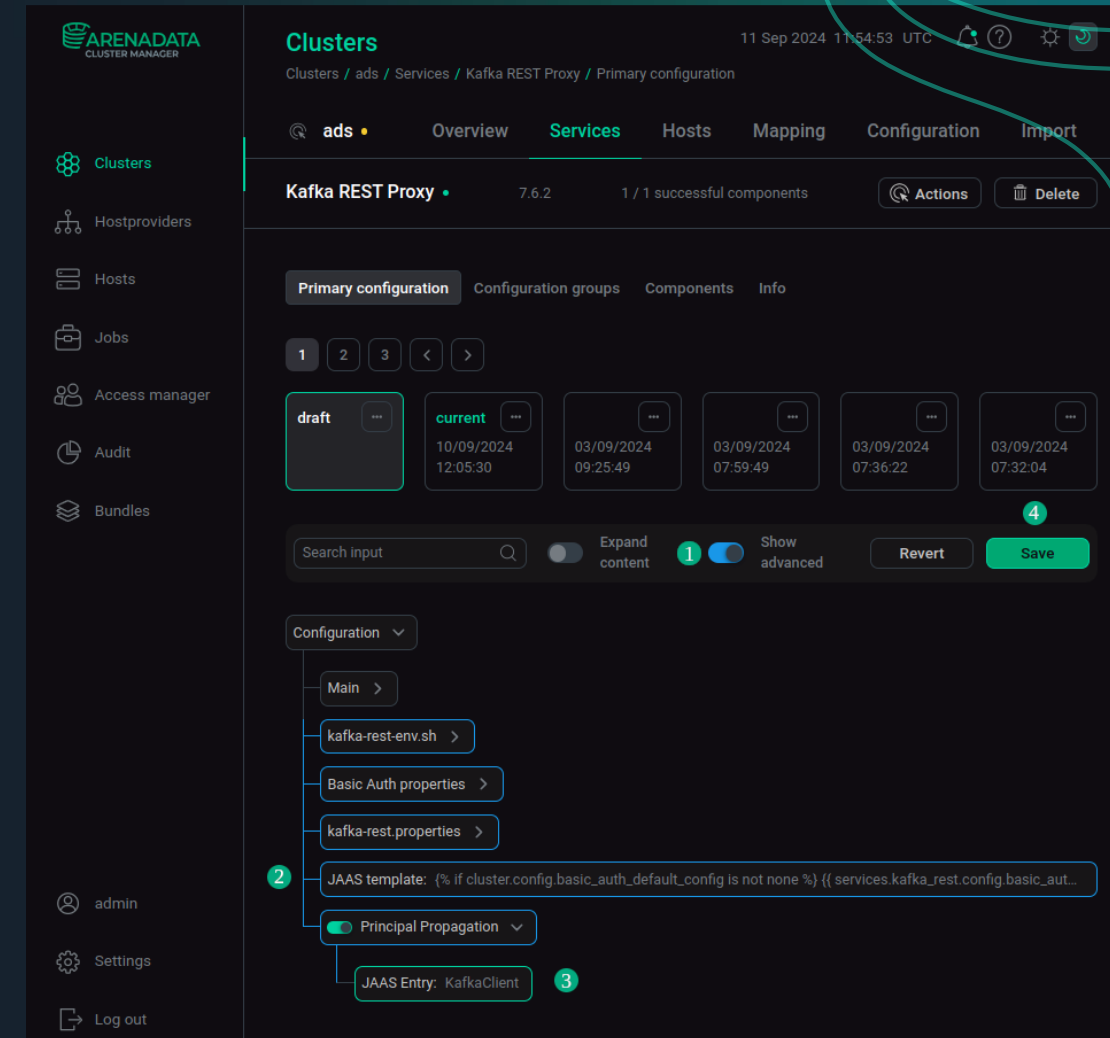
03

Возможность LDAP/LDAPS аутентификации в NiFi

Плагин *kafka-rest-security*

Возможность аутентификации в Kafka с разными принципами при подключении через Kafka REST Proxy:

1. Перехватывает запросы
2. Сопоставляет с соответствующим субъектом, который будет использоваться для подключения к Kafka
3. Извлекает информацию о пользователе
4. Выполняет аутентификацию в Kafka от имени пользователя



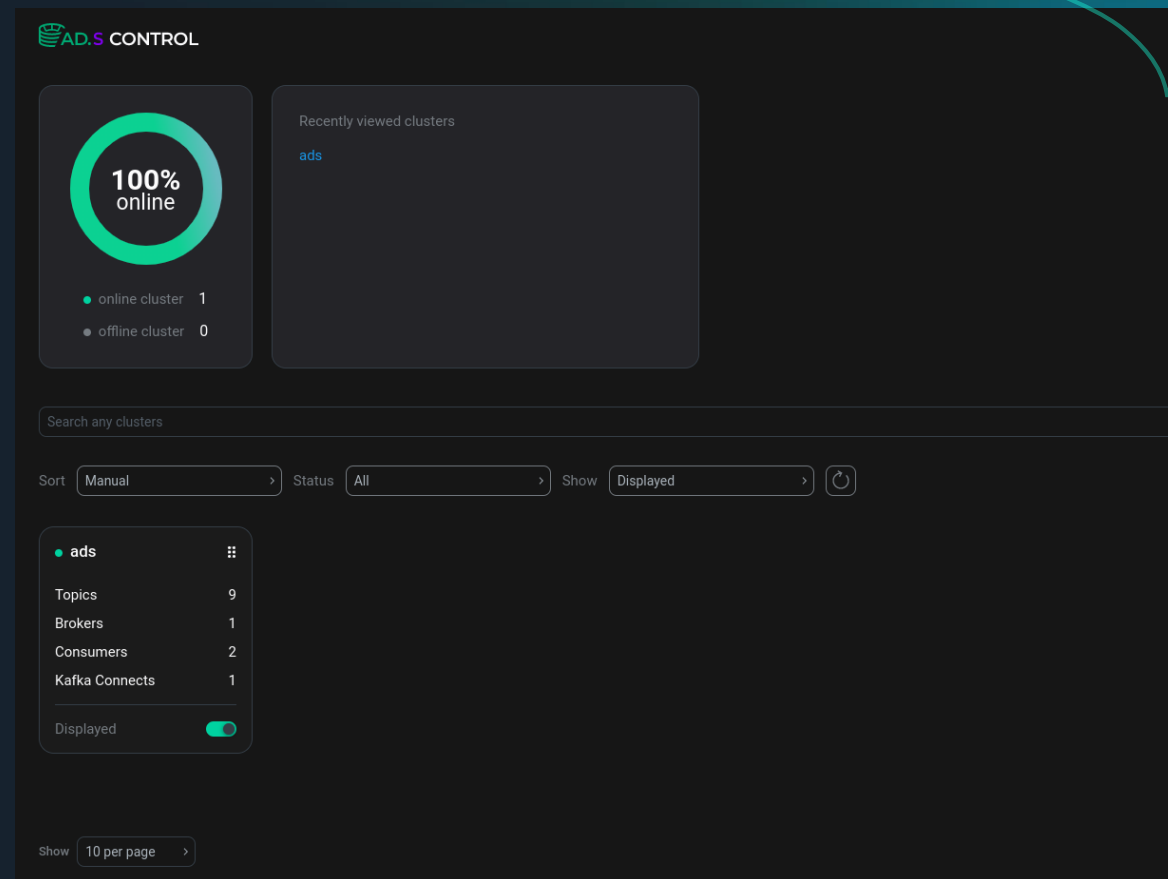
Собственная разработка Arenadata

Arenadata Streaming Control

ADS Control — решение для управления кластерами потоковой передачи

ADS Control предоставляет графический пользовательский интерфейс, позволяющий:

- управлять несколькими кластерами ADS, в которых установлены Kafka и Kafka Connect
- получать быстрый обзор состояния кластера и брокера
- управлять топиками Kafka (просмотр, создание, удаление, редактирование)
- работать с защищёнными SSL/Kerberos кластерами Kafka
- создавать, редактировать и удалять коннекторы Kafka к источникам и приёмникам
- настраивать перенос данных в Kafka и из Kafka
- настраивать коннекторы Kafka Connect
- настраивать Disaster Recovery Kafka



Собственная разработка Arenadata

Кластеры ADS в интерфейсе ADS Control

The screenshot displays the ADS Control dashboard. At the top left is the 'ADS CONTROL' logo. The top right shows the date and time '15/11/2023 09:18:15 UTC' along with settings, refresh, and user icons. A central widget shows a donut chart for '50% online' status, with a legend indicating 1 online cluster and 1 offline cluster. To the right is a 'Recently viewed clusters' section. Below these is a search bar labeled 'Search any clusters'. A filter bar contains 'Sort Manual', 'Status All', 'Show Displayed', a refresh icon, and a 'Manage' button. Two cluster detail cards are shown: 'ads' (0 topics, 0 brokers, 0 consumers, 1 Kafka Connects) and 'ads_test_cluster' (3 topics, 1 broker, 1 consumer, 0 Kafka Connects). Both have 'Displayed' toggle switches. At the bottom, there is a 'Show 10 per page' selector and pagination controls.

ADS CONTROL

15/11/2023 09:18:15 UTC

50% online

- online cluster 1
- offline cluster 1

Recently viewed clusters

Search any clusters

Sort Manual Status All Show Displayed Manage

Cluster	Topics	Brokers	Consumers	Kafka Connects
ads	0	0	0	1
ads_test_cluster	3	1	1	0

Displayed [toggle]

Show 10 per page

Фильтрация кластеров

Поиск кластеров

Обзор кластеров Kafka

Общая информация о кластере в ADS Control

General

GeneralTopics 10Brokers 3Consumers 1Kafka Connects 1

Cluster name

● ads_3.3.2

Status

online

Brokers online

3 of 3

Version

Storage size

499.90 kB

Topics

10

Partitions

99

Consumer groups

1

Services overview



Kafka Connects online

1 of 2

Работа с топиками в ADS Control

Topics

- General
- Topics 11
- Brokers 3
- Consumers 1
- Kafka Connects 1

total topics

11

total partitions

100

Секция с общей информацией о топиках

Search topic



 Show internal topics ☐

Секция общего управления топиками

Create topic

Name ↑	Messages	Internal	Size	Cleanup policy	Retention,ms	Retention,bytes	Partitions	Replicas	Actions
test_topic2	0	false	0 bytes	delete	604800000	-1	3	1	 
test_topic	0	false	0 bytes	delete	604800000	-1	1	1	 
test	0	false	0 bytes	delete	20	5000	2	3	 
mm-connect-status	20	false	7.23 kB	compact	604800000	-1	5	1	 
mm-connect-offsets	0	false	0 bytes	compact	604800000	-1	25	1	 

Работа с топиками в ADS Control

ADS CONTROL

Clusters > ads > Topics > topic-1 > Messages

16/10/2024 09:19:24 UTC

Messages

GeneralTopics 10Brokers 1Consumer groups 1Kafka Connects 1

Name	Messages	Size	Cleanup policy	Retention,ms	Retention,bytes	Partitions	Replicas	Actions
topic-1	2	143.00 bytes	delete	604800000	-1	1	1	 

MessagesPartitionsConfigurationConsumer groups

Search value

PartitionsAll

Start offsetEarliest

1

Max messages read10

Apply



Produce message

Partition	Offset	Timestamp ↑	Key	Value
0	1	10/16/2024, 12:17:16 PM		
0	0	10/15/2024, 1:02:36 PM		

Show10 per page

Вкладка *Messages* предоставляет информацию о сообщениях, содержащихся в топике, отображает (читает) их ключи и значения, а также позволяет записывать сообщения.

Вкладка *Consumer Groups* отображает группы потребителей, которые подключены к топике.

ADS CONTROL

Clusters > ads > Topics > test-topic > Consumer Groups

16/10/2024 12:29:12 UTC

Consumer Groups

GeneralTopics 10Brokers 1Consumer groups 1Kafka Connects 1

Name	Messages	Size	Cleanup policy	Retention,ms	Retention,bytes	Partitions	Replicas	Actions
test-topic	10	711.00 bytes	delete	604800000	-1	5	1	 

MessagesPartitionsConfigurationConsumer groups

Consumer group ID ↓	Reading state	Lag
test-consumer-group1	Stopped	2

Show10 per page

Работа с сообщениями в ADS Control

Messages

General Topics 10 Brokers 3 Consumer groups 0 Kafka connects 1 Schema registry

Name	Messages	Size	Cleanup policy	Retention,ms	Retention,bytes	Partitions	Replicas	Tiered storage	Actions
topic-test	7	490.00 bytes	delete	604800000	-1	1	1	enabled	 

Messages Partitions Configuration Consumer groups



Partitions

All >

Start offset

Earliest >

1

Produce message

Max messages read

10 >

Apply



Partition	Offset	Timestamp	Key	Value
▼ 0	0	2/19/2025, 10:15:48 AM		12
▼ 0	1	2/19/2025, 10:15:52 AM		18
▼ 0	2	2/19/2025, 10:15:57 AM		22
▼ 0	3	2/19/2025, 10:16:01 AM		23

Работа с брокерами в ADS Control

Brokers

General Topics 10 **Brokers 3** Consumers 1 Kafka Connects 1

total brokers
3

online brokers
3

Секция с общей информацией о брокерах

Search broker  Role > Status > 

Идентификатор
брокера

Секция поиска
и фильтрации

Broker ID	Host	Status ↑	Role	Storage size	Rack	Topics	Partitions
● 1001	sa-kafka-3.ru-central1.internal:9092	Online	Leader	42.09 kB		7	42
● 1002	sa-kafka-2.ru-central1.internal:9092	Online	Follower	53.10 kB		7	40
● 1003	sa-kafka-1.ru-central1.internal:9092	Online	Follower	408.90 kB		10	43

Show >

Работа с группами потребителей в ADS Control

ADS CONTROL

Clusters > ● ads > Consumer groups

14/10/2024 11:40:39 UTC

Consumer groups

GeneralTopics 8Brokers 1Consumer groups 2Kafka Connects 1

Total groups2

Stable groups1

Empty groups1

Секция с общей информацией о потребителях

Search consumer group

Имя группы потребителей

StateAll

Секция поиска и фильтрации

Consumer group ID ↑	State	Assigned partitions	Members	Total lag	Actions
● console-consumer-70516	Empty	0	0	0	
● console-consumer-13347	Stable	1	1	7	

Show10 per page

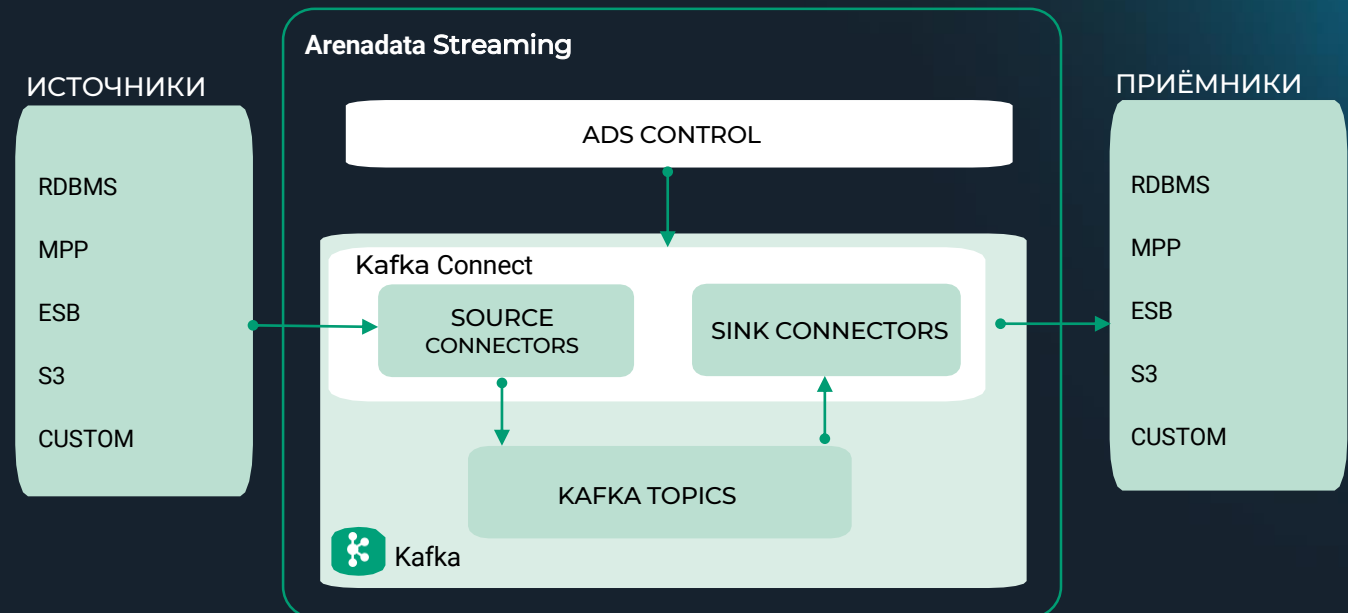
1<>

Управление коннекторами в ADS Control

ADS Control позволяет за считанные минуты настроить передачу данных в реальном времени между источником и приёмником с помощью source и sink коннекторов

Простой доступ к Kafka Connect через UI:

- Создание и управление коннекторами
- Мониторинг состояния
- Работа с множеством кластеров Kafka
- Широкий набор коннекторов
- Добавление собственных коннекторов



Список активных коннекторов в ADS Control

Для конкретного кластера Kafka выводится список настроенных коннекторов



19/07/2023 10:32:33 UTC



● ads > Kafka Connects > ads connector

Topics 9

Brokers 1

Consumers 3

Kafka Connects 1

Connectors

Поиск
коннекторов

Search connectors...



Type



Status



Операции с
коннекторами

Create Connector

Name ↓

Type

Plugin

Topics

Broker

● Tasks

Actions

● 1

sink

org.apache.kafka.connect.file.FileStreamSinkConnector

sov-test-6.ru-central1.internal:8083

1



Show 10 per page



Обзор
коннекторов



Выбор плагина коннектора в ADS Control

● ads > Kafka Connects > ads connector > Kafka connector plugins

Topics 9 Brokers 1 Consumers 2 **Kafka Connects 1**

Kafka Connect Plugins

Search connectors... 🔍 Type > ↻

Выбор типа
коннектора

Подключение
кастомных и OSS
коннекторов

Cancel

Name ↓	Type	Class	Add
FileStreamSinkConnector	sink	org.apache.kafka.connect.file.FileStreamSinkConnector	+
FileStreamSourceConnector	source	org.apache.kafka.connect.file.FileStreamSourceConnector	+
MirrorCheckpointConnector	source	org.apache.kafka.connect.mirror.MirrorCheckpointConnector	+
MirrorHeartbeatConnector	source	org.apache.kafka.connect.mirror.MirrorHeartbeatConnector	+
MirrorSourceConnector	source	org.apache.kafka.connect.mirror.MirrorSourceConnector	+

Настройка параметров коннектора в ADS Control

• ads > Kafka Connects > ads connectors > create

Topics 9

Brokers 1

Consumers 2

Kafka Connects 1

Create Connector FileStreamSinkConnector

enable if you prefer to use JSON format

Save

Cancel

JSON view



Common

Connector name

Globally unique name to use for this connector.

Connector class

org.apache.kafka.connect.file.FileStreamSinkConnector

Tasks max

1

Key converter class

Key converter class

Common

Transforms

Predicates

Error Handling

Other

Advanced

- Группировка параметров
- Режим JSON для глубокой настройки

Доступные коннекторы

SOURCE CONNECTORS

ТИП	название	поставка
Kafka	MirrorMaker 2	В бандле
Debezium	Debezium Source Connectors (PostgreSQL, MS SQL)	В бандле
JDBC	JDBC Source Connector (Oracle, PostgreSQL, MS SQL, MySQL, GreenPlum)	Можно добавить
Elastic	ElasticSearch Source Connector	Можно добавить
S3	Amazon S3 Source Connector	Можно добавить

SINK CONNECTORS

ТИП	название	поставка
Kafka	MirrorMaker 2	В бандле
Iceberg	Iceberg Sink Connector	В бандле
JDBC	JDBC Sink Connector (Oracle, PostgreSQL, MS SQL, MySQL, GreenPlum)	Можно добавить
Elastic	ElasticSearch Service Sink Connector	Можно добавить
S3	Amazon S3 Sink Connector	Можно добавить
MPP	HDFS2 Sink Connector	Можно добавить



- Есть возможность использования доступных в бандле source-коннекторов Debezium и добавления сторонних
- Дополнительные коннекторы устанавливаются путем копирования jar-архива с коннектором на Kafka Connect cluster (путь можно настроить в ADCM)
- Есть возможность подключения собственных разработанных коннекторов, в том числе и платных

Mirror Maker 2

Mirror Maker 2 — основанный на базе Kafka Connect механизм репликации данных из исходного кластера на удалённый. Создаёт специальные коннекторы для включения сложных потоков между кластерами ADS:

- **MirrorSourceConnector** осуществляет репликацию топиков из исходного кластера в целевой кластер
- **MirrorCheckpointConnector** создаёт контрольные точки смещения потребителя и синхронизирует смещение со служебным топиком `__consumer_offsets` исходного кластера
- **MirrorHeartbeatConnector** периодически проверяет подключение между кластерами, создавая сообщения в специальном топике `heartbeats` в исходном кластере через заданный период времени и считывая их в целевом кластере

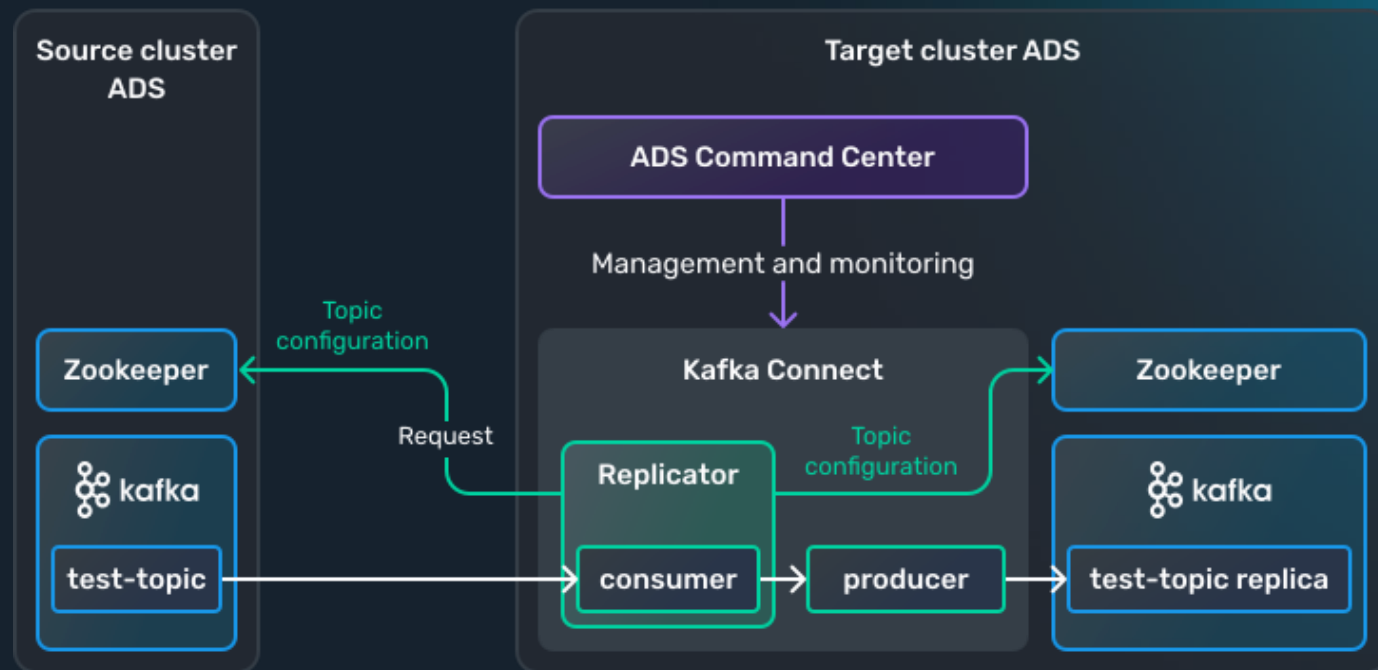


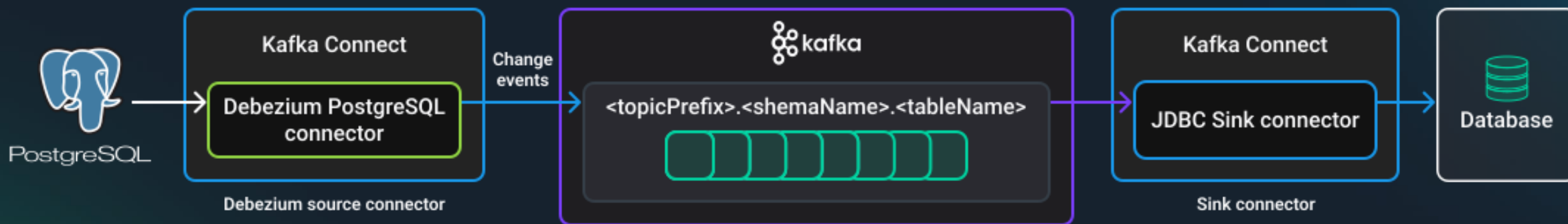
Схема репликации с использованием Mirror Maker 2

Debezium source connectors

Debezium — распределённая платформа, работа которой основана на паттерне захвата изменения данных (change data capture, CDC). Коннектор Debezium собирает изменения в базах данных и передаёт их во внешние приложения.

В составе Arenadata Streaming:

- Debezium source-коннекторы для сервиса Kafka Connect: для PostgreSQL и Microsoft SQL Server*
- возможность настраивать доступные в дистрибутиве коннекторы и добавлять собственные из веб-интерфейса ADS Control



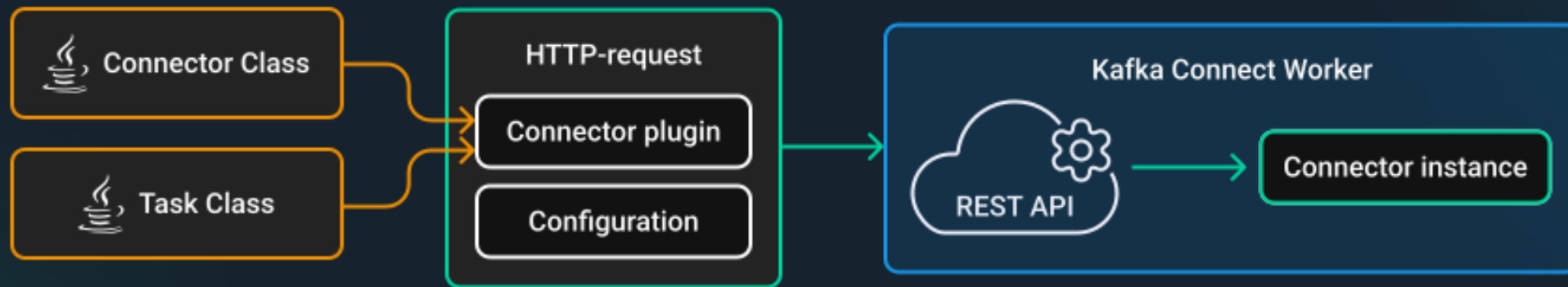
Архитектура сбора изменённых данных на основе Debezium, на примере сервера PostgreSQL

Iceberg sink connector

Коннектор для записи потоковых данных из Kafka в таблицы Apache Iceberg позволяет:

- реализовать полноценный процесс CDC (в связке с Debezium source коннектором)
- обеспечить непрерывную и согласованную интеграцию из оперативных баз данных в табличный формат Iceberg

Подход избавляет от необходимости проектировать и поддерживать сложные ETL-процессы, ускоряя внедрение аналитических решений на основе актуальных данных



Процесс создания экземпляра коннектора в Kafka Connect

1

Обновление версий сервисов

- Kafka 4
- NiFi 2

2

Развитие функционала на базе OSS

- Поддержка io_uring
- Развитие плагинов (Ranger, etc)

3

Разработка нового функционала

- Connector Registry
- Нативная интеграция с ADPS.OpenBao

4

Операционные системы

- Альт СП релиз 10, Astra Linux 1.8 SE, Ubuntu 24.04 LTS

5

Развитие ADS Control

- Управление ACLs
- Интеграция Cruise Control и др.
- Улучшения существующего функционала

6

Развитие мониторинга

- Добавить сбор метрик и дашборды для всех сервисов (Kafka REST, Schema Registry, NiFi, kSQL)
- Улучшить дашборды
- Адаптировать общий бандл мониторинга ADM

Компоненты Arenadata Streaming



	ADS 3.6.2.2	ADS 3.7.2.1	ADS 3.9.0.1	ADS 3.9.1.1	ADS 3.9.1.2
Kafka	3.6.2	3.7.2	3.9.0	3.9.1	3.9.1
Kafka Connect	3.6.2	3.7.2	3.9.0	3.9.1	3.9.1
CMAK	3.1.2	-	-	-	-
Schema Registry	7.6.2	7.7.2	7.9.0	7.9.1	7.9.1
KsqlDB	7.6.2	7.7.2	7.9.0	7.9.1	7.9.1
Kafka REST	7.6.2	7.7.2	7.9.0	7.9.1	7.9.1
ZooKeeper	3.8.4	3.8.4	3.8.4	3.8.4	3.8.4
NiFi	1.26.0	1.26.0	1.28.0	1.28.0	1.28.0
NiFi Registry	1.26.0	1.26.0	1.28.0	1.28.0	1.28.0
MiniFi	1.26.0	1.26.0	1.28.0	1.28.0	1.28.0
C2 Server	1.26.0	1.26.0	1.28.0	1.28.0	1.28.0
NiFi 2	-	-	-	-	2.5.0
Kerberos	latest	latest	latest	latest	latest
ADCM	latest	latest	latest	latest	latest

Предыдущие
версии

ADS 3.6.2.2

ADS 3.7.2.1

Текущая
версия

ADS 3.9.0.1

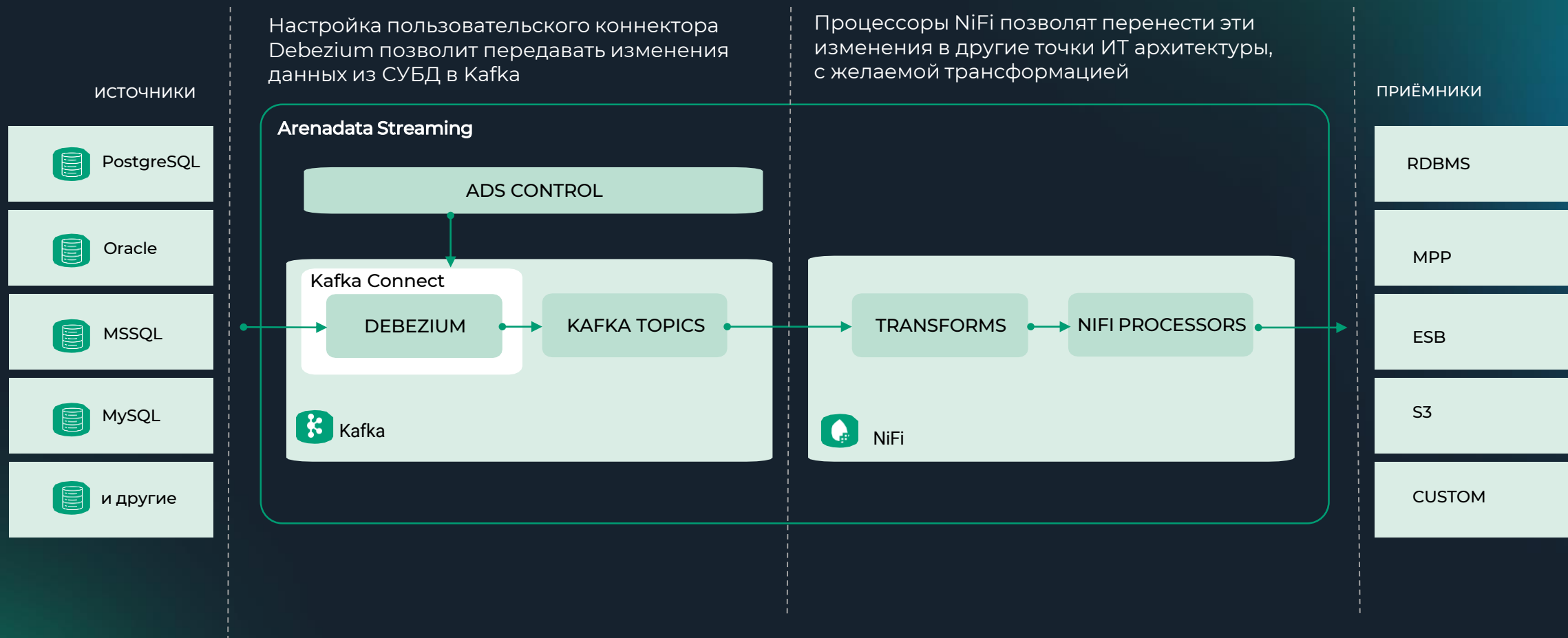
Будущие
версии

ADS 3.9.1.1

ADS 3.9.1.2

Типовые архитектурные шаблоны

Кейс 1. Инфраструктура для CDC



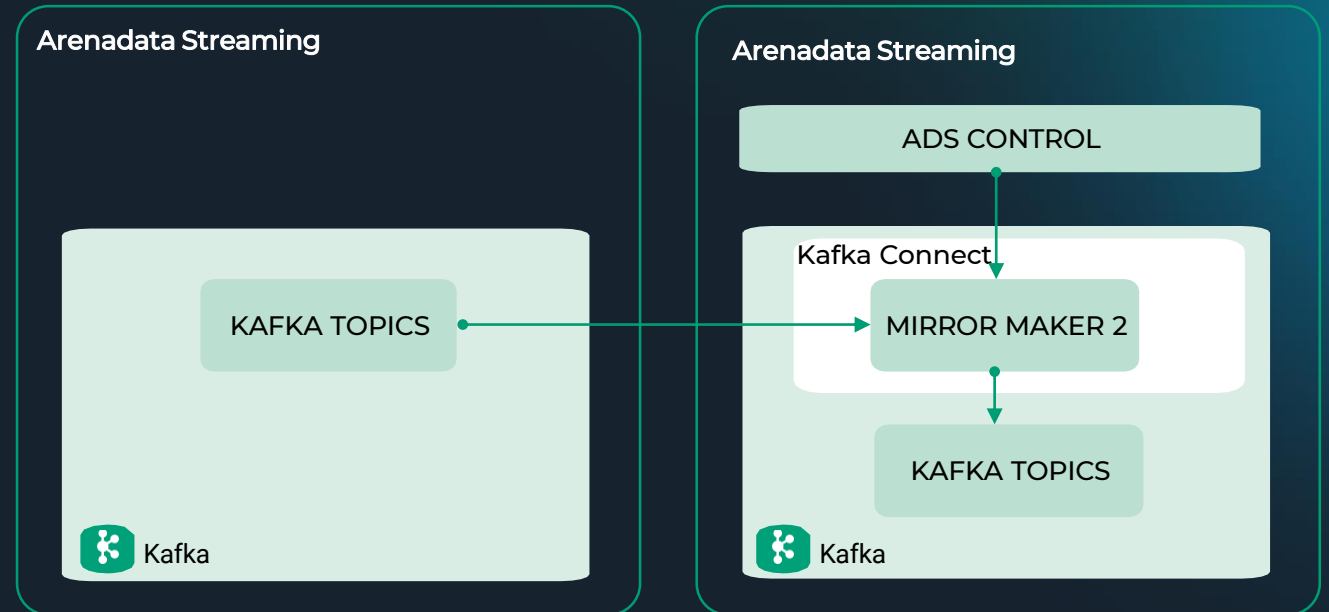
Кейс 2. Репликация и DR Kafka

Mirror Maker 2.0 (MM2) разработан для:

- упрощения зеркалирования
- репликации топиков из одного кластера Kafka в другой

Возможности MM2:

- Обнаруживает изменения в топиках
- Обеспечивает синхронизацию свойств исходного и целевого набора данных, включая смещения и разделы
- Использует фреймворк Kafka Connect для упрощения конфигурации и масштабирования



Arenadata Streaming vs «Ваниль»

Преимущества Arenadata Streaming в сравнении с Open Source Kafka и NiFi

Преимущества Arenadata Streaming в сравнении с Open Source

01

Экосистема сервисов

- Стабильные версии проектов
- Интегрированные проекты экосистемы Kafka и NiFi (Tiered Storage, KRaft, Kafka Cruise Control)
- Собственные разработки: поддержка kTLS в Kafka, NiFi ADB Processor

02

ADS Control

- UI для управления коннекторами
- Работа с топиками, брокерами и группами потребителей
- Disaster Recovery Kafka

03

Информационная безопасность

- ADPS (Ranger, Kerberos, MIT, AD, IPA)
- TLS, интеграция с AD, LDAP/LDAPS
- Плагин kafka-rest-security для авторизации в Kafka REST Proxy

04

Мониторинг всех компонентов

- ADCM
- Сервис Monitoring на базе Prometheus с поддержкой веб-интерфейса Grafana

05

Отечественное ПО

- Продукт в реестре
- Соответствие требованиям регуляторов
- Документация и учебные курсы на русском языке

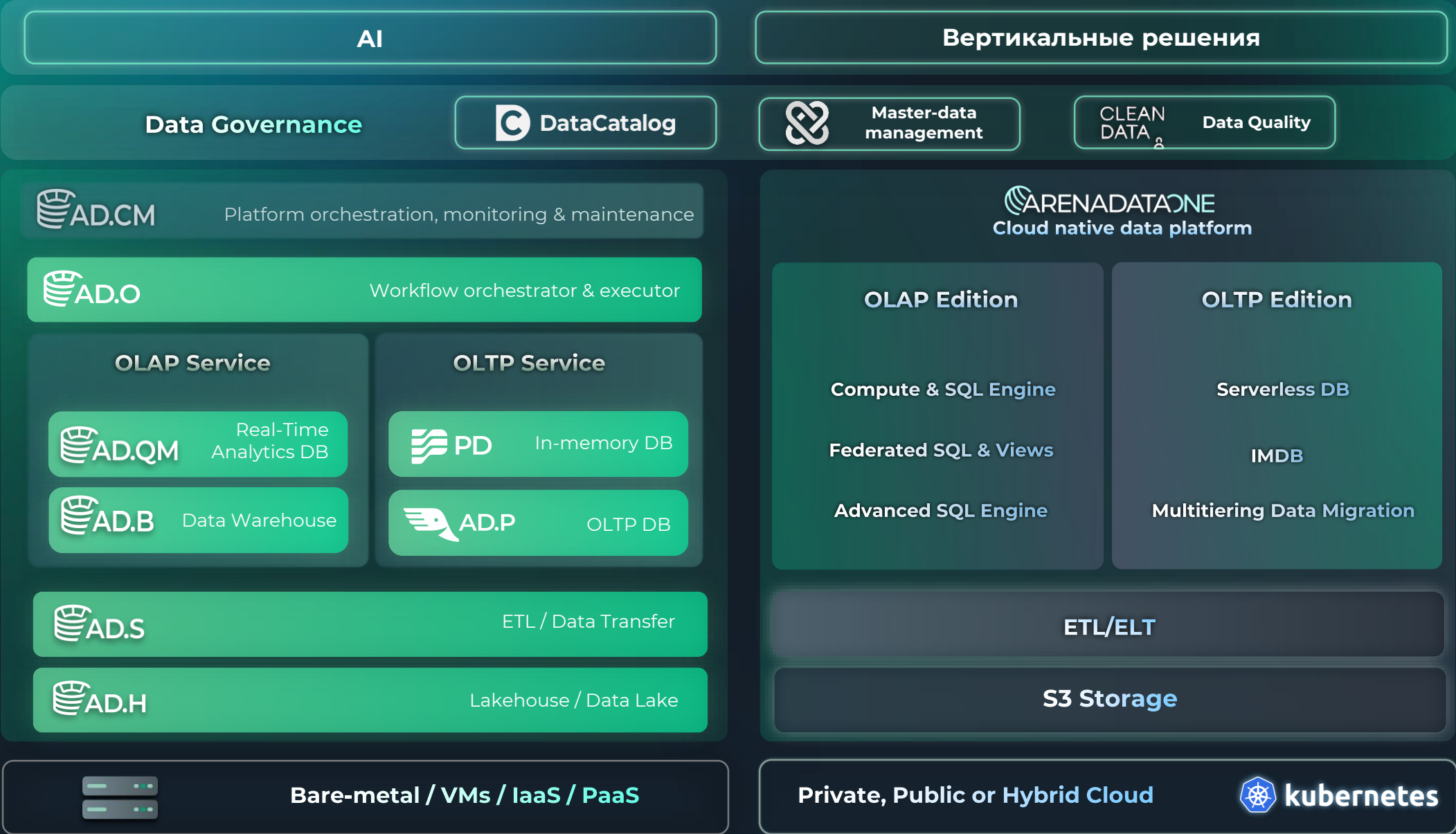
06

Российские ОС (ЕЕ)

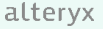
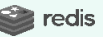
- AltLinux SP 8.4, Astra Linux 1.7 SE «Орёл» и «Воронеж», RED OS 7.3
- а также CentOS 7.x, RHEL 7.x, Ubuntu 22.04.2 LTS

Arenadata EDP

Целевая платформа



Миграция с иностранных решений

Управление данными и качество данных	 Informatica  alteryx  Alation  Collibra  ATAKAMA  IBM	 ADC CLEAN DATA
Базы данных в памяти (резидентные СУБД)	 GridGain  redis  HAZELCAST	 PICODATA
Аналитические СУБД в режиме реального времени (аналитические СУБД)	 Microsoft SQL Server	 AD.QM
Хранилища данных (аналитические СУБД)	 VERTICA  Pivotal Greenplum  teradata  SAP HANA  ORACLE EXADATA	 AD.B
Озёра данных, lakehouse (платформа Hadoop)	 CLUSTERFA  ORACLE BIG DATA APPLIANCE	 AD.H
Потоковая передача (загрузка) данных	 CONFLUENT  kafka  nifi  IBM  sas  Informatica  TIBCO	 AD.S
Транзакционные базы данных (СУБД общего назначения)	 ORACLE DATABASE  IBM DB2  Microsoft SQL Server	 AD.P

Единая платформа, объединяющая востребованные продукты мирового уровня и позволяющая эффективно замещать целый комплекс решений ушедших зарубежных компаний

Каждый добавляемый компонент платформы доработан, чтобы стать её частью

 AD.B	Big Data Warehouse	 Greenplum
 AD.H	Data Lake / Lakehouse	 hadoop
 AD.QM	Fast flat marts	 ClickHouse
 AD.P	Relational DBMS	 PostgreSQL
 PICODATA	In-memory SQL	Tarantool
 AD.S	ESB/MB/MQ	 kafka  nifi
 AD.CM	Data & Cluster Management	 ANSIBLE
 ADC	Data Catalog	 Open Metadata



Обязательный перечень задач для каждого компонента:

Выполнить статистический анализ кода

Выполнить интеграцию с модулем мониторинга

Проверить работоспособность заявленного функционала

Подобрать персонал поддержки, подготовить и обучить архитекторов и разработчиков

Разработать коннекторы для взаимодействия с другими компонентами платформы

Сформировать дорожную карту доработок продукта

Мы осуществляем безопасную разработку



Используем инструменты и методики:

- ✓ Статические анализаторы исходного кода
- ✓ Поиск секретов в исходном коде
- ✓ Безопасность в образах контейнеров
- ✓ Анализатор зависимых компонентов
- ✓ Средства динамического анализа



Разработка в соответствии с ГОСТ Р 56939-2016 «Разработка безопасного программного обеспечения»

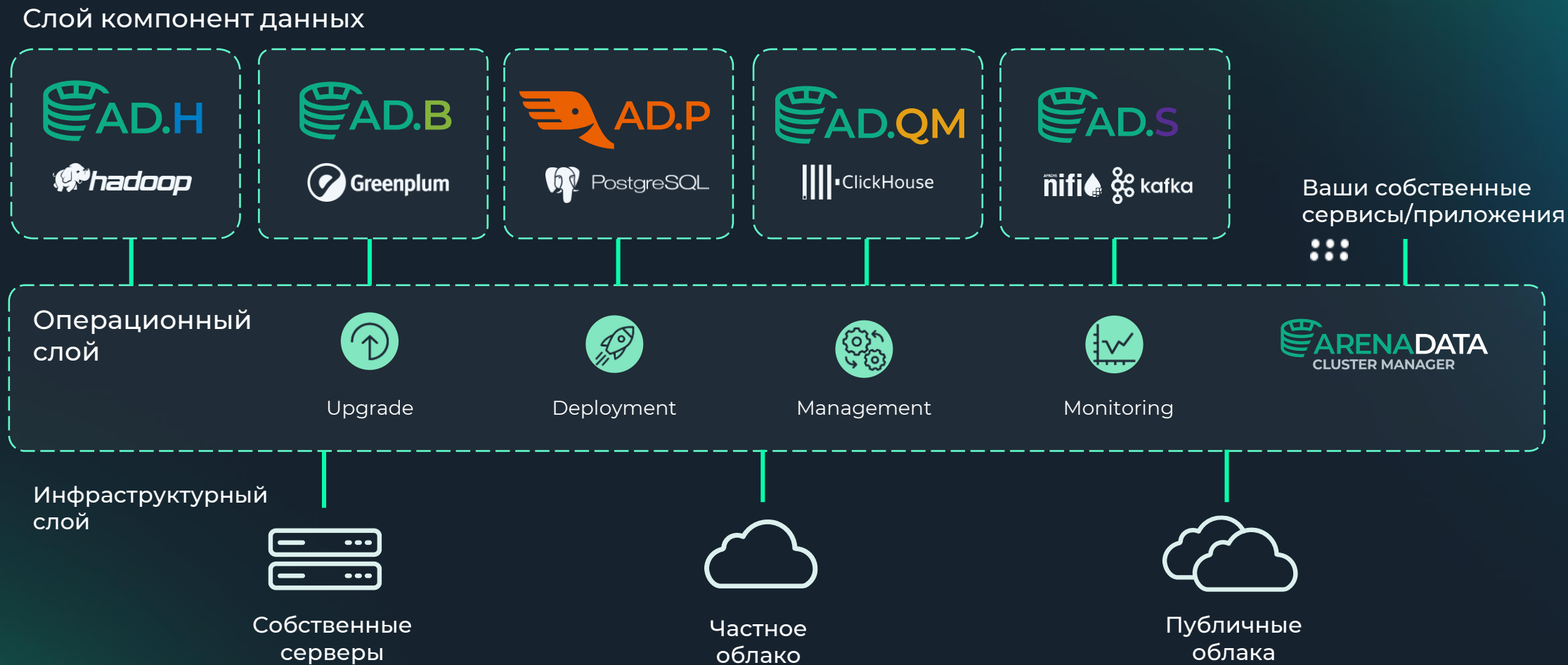


Применяются международные практики OWASP Application Security Verification Standard



Разработаны и внедрены правила для каждого языка программирования

Гибридный корпоративный ландшафт



Enterprise Multi-cloud платформа данных

Виды развёртки любых компонентов платформы Arenadata EDP:

01

в облаке

02

на bare-metal-оборудовании

03

гибридная ИТ-инфраструктура
или multi-clouds



Многоуровневая система технической поддержки

Варианты технической поддержки

24/7 доступность поддержки	600+ кластеров на поддержке	400+ заявок в месяц	99%+ соблюдение SLA в 2024	50+ технических экспертов
Гарантийное обслуживание	Режим работы над обращениями 8x5	Анализ проблем в рамках поступивших инцидентов	Предоставление обновлений ПО	Доступ к базе знаний по решению инцидентов
Базовая техническая поддержка	Режим работы над обращениями 24x7 для Prod	Диагностика и предоставление обходного решения	Предоставление обновлений ПО	Консультации по особенностям ПО
		Анализ проблем в рамках поступивших инцидентов	Решение инцидентов в рамках SLA	Доступ к базе знаний по решению инцидентов
Премиальная техническая поддержка	Услуги, включающие сервисы Базовой технической поддержки и учитывающие индивидуальные потребности Заказчика			

Вендорский консалтинг Arenadata

Минимизация рисков, максимизация отдачи от инвестиций, оперативное устранение проблем

Технический аудит

Комплексный аудит для анализа узких мест кластера. Детальный разбор и рекомендации по изменениям от архитекторов Arenadata позволяют сформировать среднесрочную стратегию эксплуатации

Производительность

Комплекс тактических услуг по оптимизации использования ресурсов, настроек ресурсных групп, настроек резервного копирования, адаптации сайзинга к изменяющейся нагрузке и оптимизации прикладных запросов

Технический аккаунт-менеджмент

ТАМ-команда, состоящая из архитектора-эксперта, руководителя проектов, аналитиков и инженеров. Постоянная команда идёт с заказчиком рука об руку и использует свой опыт для значительного снижения проектных и технических рисков, возникающих в ходе внедрения и эксплуатации сложных систем

Программа приоритетного внимания

Наблюдения истории развития ландшафтов данных наших клиентов позволили выстроить эталонный сервисный путь ландшафта данных для крупных заказчиков с учётом характера и интенсивности потребления услуг

Надёжный старт

Комплекс услуг по развёртыванию ПО Arenadata

Удобная и понятная онлайн-документация мирового уровня



Не копируем, а создаём



Обеспечиваем поддержку разных устройств



Сами разрабатываем примеры



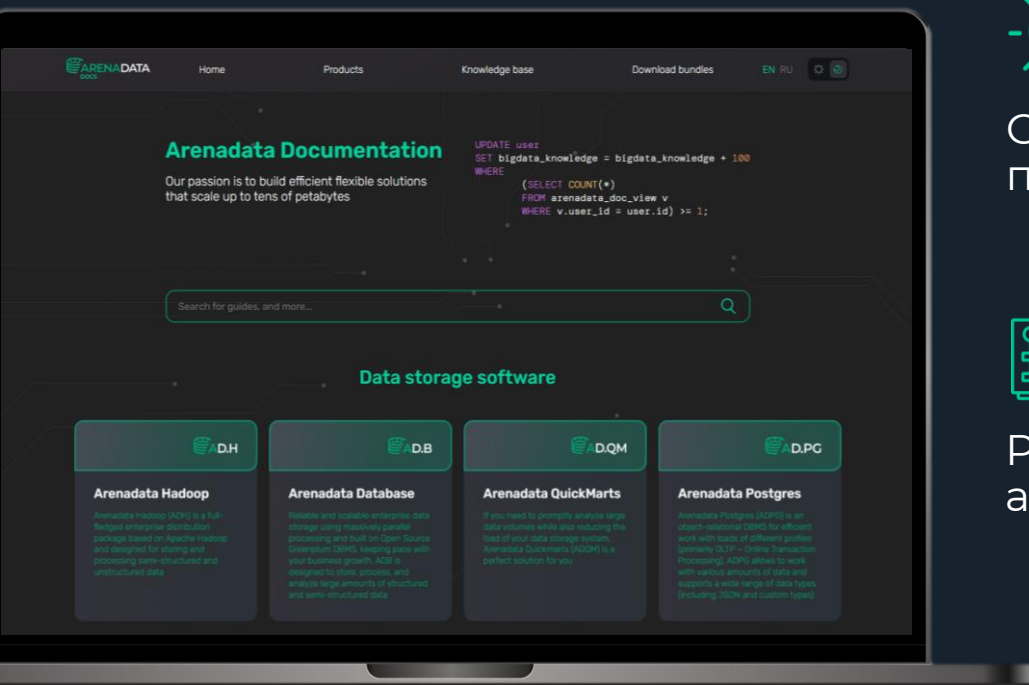
Проводим экспертную проверку архитекторами и аналитиками



Разрабатываем дизайн архитектурных диаграмм



Не переводим, а пишем на русском и английском языках



[Перейти на сайт с документацией](#)

Знания из первых рук:

С сертификацией и присвоением статуса
Arenadata Certified Specialist



Знания из первых рук

Преподаватели — практикующие специалисты, имеющие за плечами богатую экспертизу и значительное количество реализованных бизнес-кейсов

[Узнать подробнее о курсах](#)



Практика от экспертов

Все курсы построены на реальных кейсах и практических примерах, а лабораторные работы проходят на стендах, симулирующих рабочую среду



16 учебных программ по 8 направлениям

- Архитектура платформы данных
- Курсы по Arenadata DB
- Курсы по Arenadata Hyperwave
- Курсы по Arenadata QuickMarts
- Курсы по Arenadata Streaming
- Курсы по Arenadata Postgres
- Курсы по Picodata
- Курсы по Arenadata Catalog

ОСТАВАЙТЕСЬ НА СВЯЗИ



Наши новости
в телеграм-канале

