beeline cloud

Облако или on-premise? Честное сравнение подходов обработки данных

Максим Еремин
Руководитель направления развития продуктов beeline cloud

Алексей Струченко
Технический экперт Arenadata

Agenda

Особенности работы с данными: проблемы и пути их решения

Концепция работы с данными: о чем стоит знать

Выгоды и преимущества облачного подхода к КХД — посчитаем вместе с вами TCO

Обзор Cloud-инструментов для работы с данными

Демо по управлению кластером Greenplum в облаке

Вопросы и ответы

01

Мы в цифрах

100+
Облачных сервисов и решений

2000+
Клиентов из разных отраслей бизнеса

99,95%
Облачный SLA

6
ЦОДов в РФ уровня Tier III

5
Независимых интернет-провайдеров

Secure by design

УЗ-1 152-ФЗ
Публичное облако

ГИС К1, ИСПДн УЗ-1
Защищенный сегмент

PCI DSS 3.2.1
Соответствие стандарту

ГОСТ Р 57580.1-2017
Соответствие требованиям

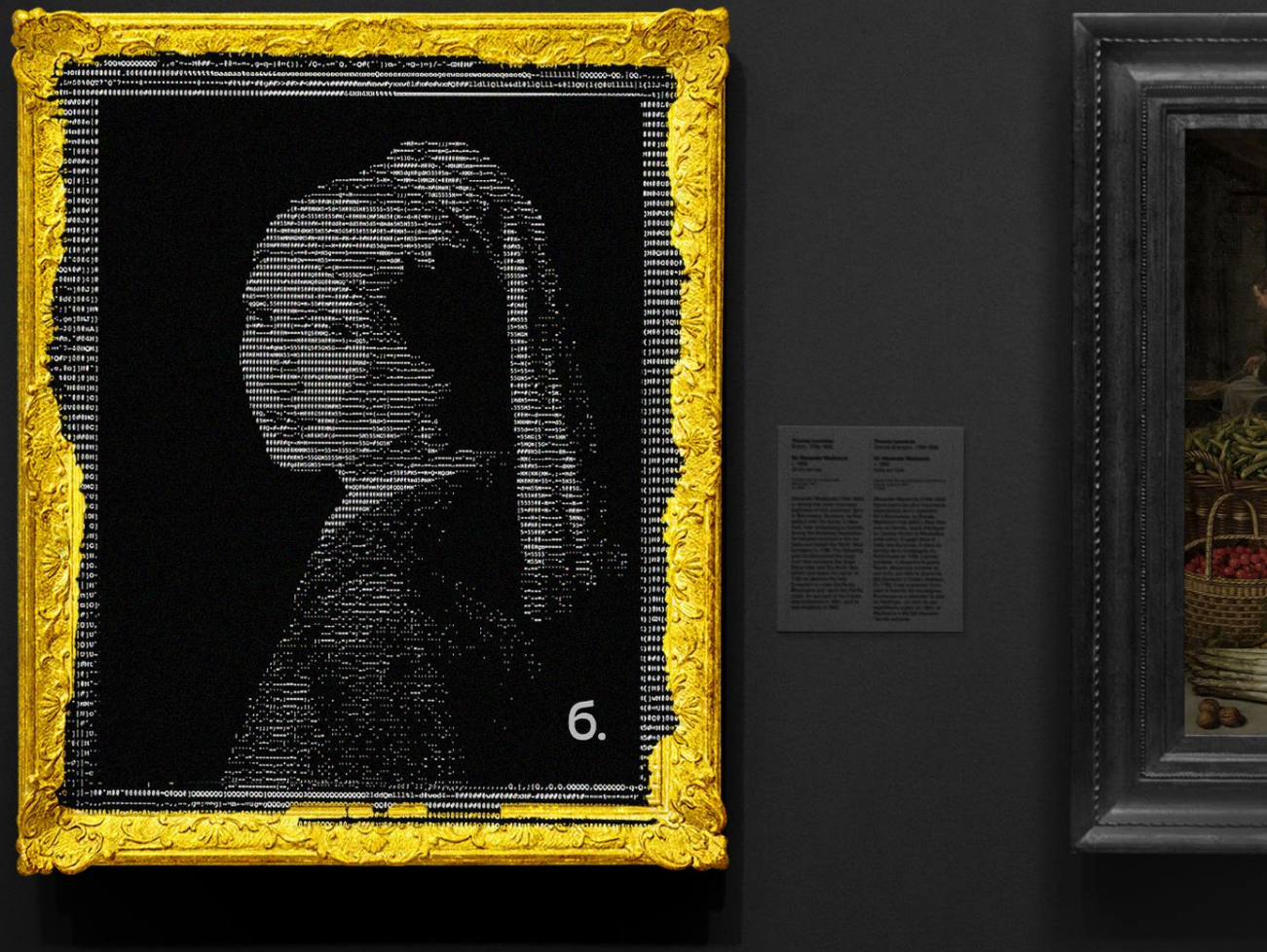
ФСБ, ФСТЭК
Лицензии

Партнеры

positive technologies · kaspersky · MULTIFACTOR · UserGate · ARENADATA · SANGFOR · radware · COMMUNIGATE SYSTEMS

beeline cloud

Особенности работы с данными



Ситуация

- Данные в компании хранятся во множестве систем и источниках
- Данные не выделены в определенные предметные области и разрознены
- Выгрузка данных занимает немалое время
- Для подготовки отчетов приходится брать «грязные» данные и формировать новые таблицы вручную
- Слишком большой поток данных генерируется источниками

Проблемы

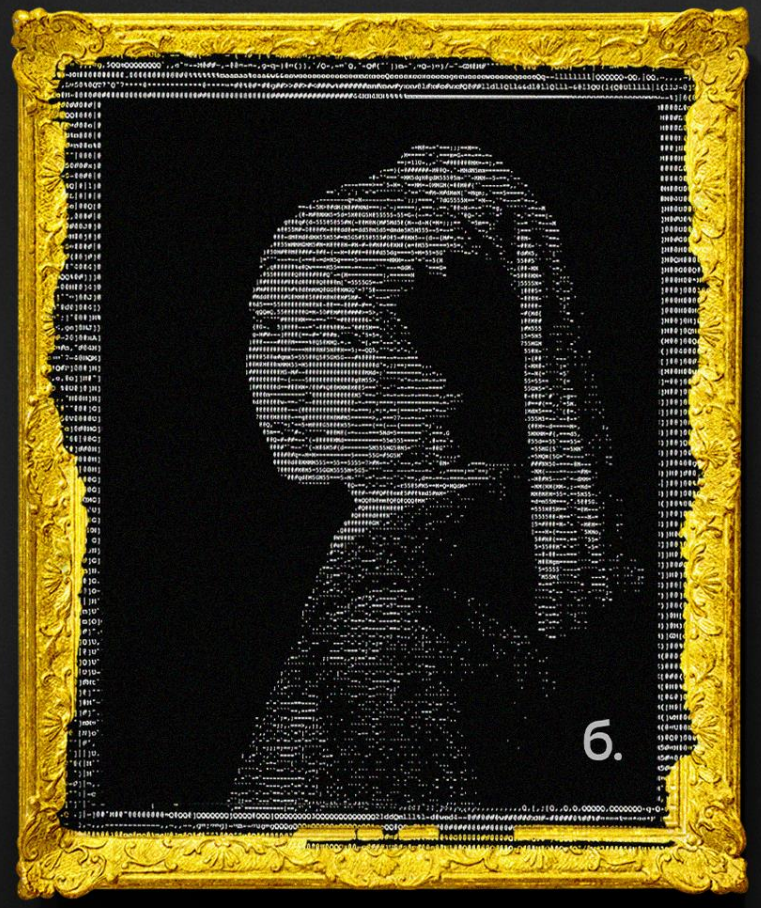
- Данные не централизованы, сложно ими управлять, увеличиваются затраты
- Нет прямых связок для построения отчетов, невозможно соотнести предметные области для анализа
- Подготовка отчетов и использование данных тормозит процессы, основанные на этих данных
- «Грязные» данные пагубно влияют на результаты отчетов, не отражают реальную суть процессов в компании
- Нет свободных физических вычислительных ресурсов для расширения хранилищ

Решение

- Организовать централизованное корпоративное хранилище данных
- Организовать процесс очистки и препроцессинга данных
- Подключить источники данных с помощью ETL-инструментов к КХД
- В КХД организовать предметно-ориентированные таблицы представления (витрины данных)
- Использовать СУБД, способную работать с параллельными запросами с данными объемом свыше 5 ТБ

beeline cloud

Концепция работы с данными



Сбор данных

Основная идея

Предоставить простой и понятный интерфейс для сбора и вытягивания данных

- В качестве источника может быть база данных, любое другое хранилище или потоковая передача данных в реальном времени
- В качестве потокового сбора данных используется Apache Kafka, для вытягивания данных можно использовать Apache NiFi

Препроцессинг данных

Основная идея

Предоставить набор утилит и решений для трансформации данных и ETL/ELT

- Для таких задач может использоваться Apache Nifi, Apache Spark, Apache Airflow
- Источником может быть точка входа данных из других источников или корпоративное хранилище данных, куда уже попали данные из источников

Хранение данных

Основная идея

Создание хранилища данных, разного уровня скорости доступа к данным для создания централизованной точки хранения и подключения к данным

- Хранилища могут делиться на 3 разных типа: холодные, теплые и горячие
- Уровень типа зависит от скорости доступа и записи данных

Холодное хранение — Data Lake

Основная идея

Создание холодного хранилища данных с редким запросом на чтение

- Задача холодного хранилища: централизовать выгрузку данных из всех источников без изменения, чтобы можно было работать с сырыми данными в разных сценариях
- Используется Hadoop или S3-Storage
- Данные в таком хранилище неструктурированные, сырые и «грязные»

Теплое хранение — КХД

Основная идея

Создание теплого хранилища со структурированными и очищенными данными

- Возможность параллельно исполнять запросы сразу несколькими аналитиками
- Данные в хранилище уже являются очищенными и готовыми к использованию для OLAP-запросов
- Под теплым хранилищем подразумевается быстрый доступ к данным, но не быстрее, чем горячий, поскольку есть вероятность создания новых связей и таблиц на базе уже существующих
- В качестве корпоративного хранилища можно использовать PostgreSQL (до 3 ТБ) и Greenplum (свыше 3 ТБ)

Горячее хранение — витрины данных

Основная идея

Создание горячего хранилища со структурированными и очищенными данными, полностью готовыми для аналитических выводов и анализа ситуации на базе одной таблицы

- Основной подход: колоночное хранение данных для быстрой выгрузки
- Используется Clickhouse в качестве витрин данных
- Подходит для интеграции с сервисами визуализации и Business Intelligence

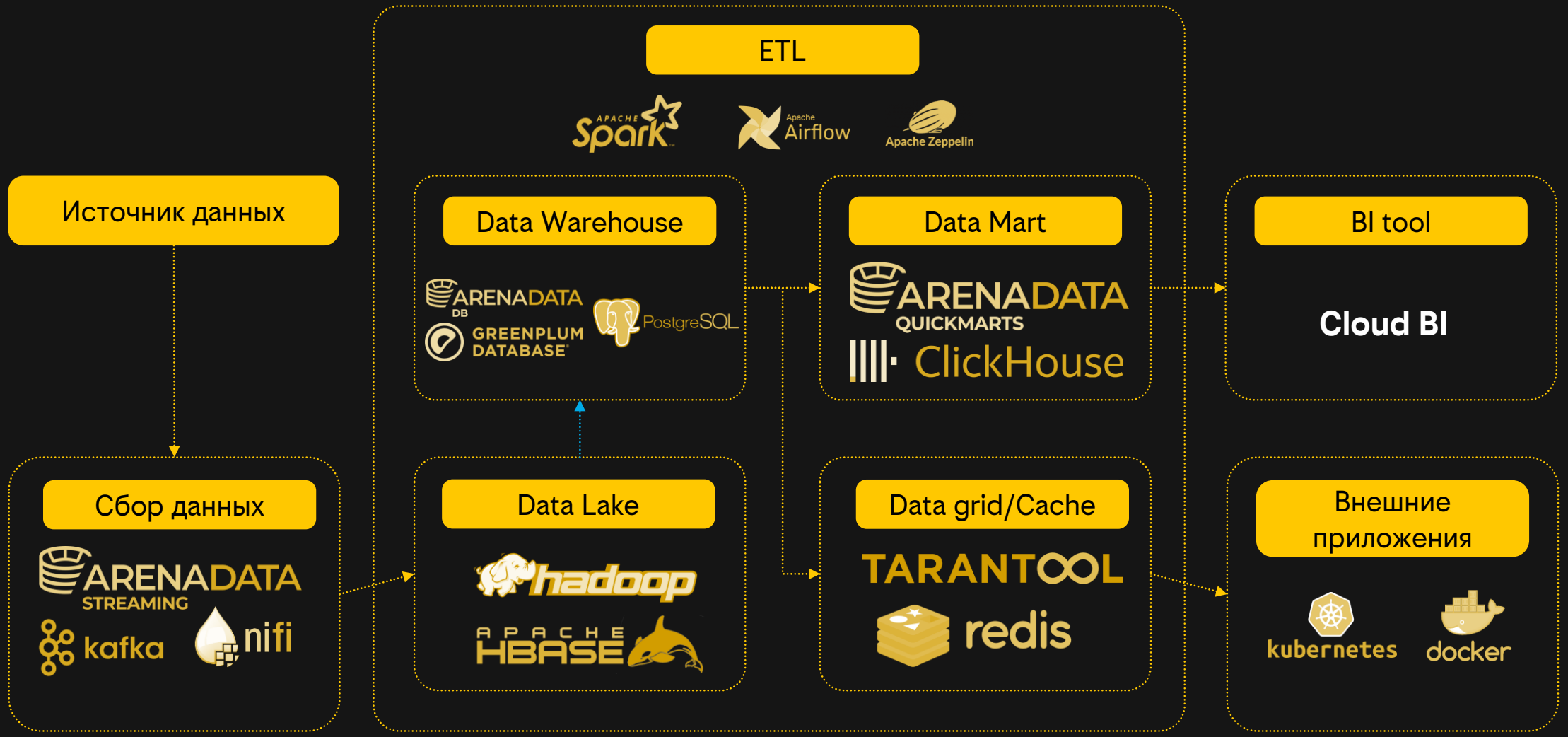
Визуализация и аналитика

Основная идея

Предоставить возможность визуализации данных для конечного анализа и принятия решений

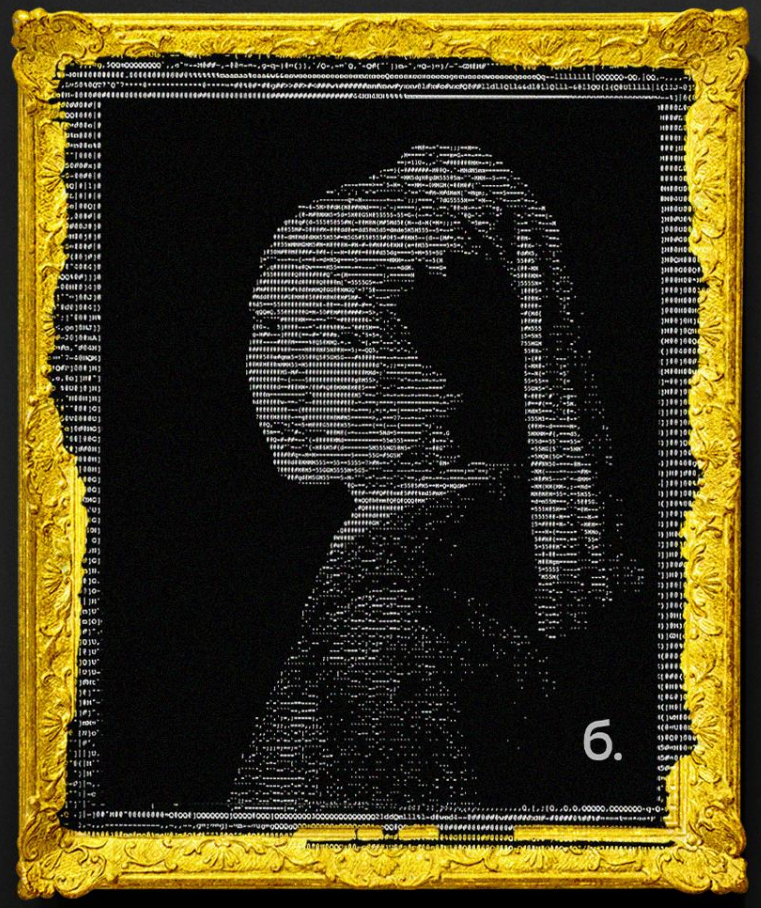
- Задача: построение real-time графиков и дашбордов для предоставления их главному потребителю в компании — топ-менеджменту
- Помогает анализировать процессы в компании на основе собранных данных

Платформа данных



beeline cloud

Выгоды и преимущества работы с данными в облаке



Отличие on-premise и cloud

- Снижение затрат на физические серверы
- Перевод CAPEX в OPEX (потребление ресурсов помесячно)
- Гибкое масштабирование СУБД в облаке — быстро расширяем, быстро сужаем
- Сокращение TCO

TCO

50 Человек

Работает в компании

35 Тыс. руб. в месяц

Средняя зарплата сотрудников

1,75 Млн. руб

Зарплатный фонд

<30%

ЗФ не превышает от прибыли

~10-15%

Процент маржи в месяц

~5,8

млн. руб.

Прибыль компании в месяц

~870

тыс. руб

Выручка компании в месяц

~250

тыс. руб

Стоимость простоя в день

TCO: Open Source vs Vendor

Opensource	Y1	Y2	Y3	Y4	Y5	5Y TCO
Команда разработки Greenplum	263,7	306,7	235,1	262,4	293,9	1 362
Закупка серверов	247,8			82,6		
	511,15	**306,7**	**235,1**	**345,0**	**293,9**	**1692**
DR = 15%						
Discounted cash flow	511,5	266,7	177,8	226,9	168,1	1351

Cloud Arenadata DB	Y1	Y2	Y3	Y4	Y5	5Y TCO		
Стоимость облачной услуги	172,2	185,9	201,0	217,5	235,7	1012		
Поддержка	28,2	30,3	32,6	35,2	38,0	164		
	200,4	**216,2**	**233,6**	**252,7**	**273,8**	**1177**	**30%**	Difference
DR = 15%								
Discounted cash flow	200,4	188,0	176,6	166,2	156,5	888	**34%**	Difference

TCO – Команда разработки Greenplum (Open Source)

№	Задача	Month net salary, rub in Y1	Churn risk rate for customer	Y1, Team FTE	Y1, Cost	Y2, Team FTE	Y2, Cost	Y3, Team FTE	Y3, Cost	Y4, Team FTE	Y4, Cost	Y5, Team FTE	Y5, Cost
					0		1		1		3		4
1	Разработка и поддержка доп. функционала в ядро системы	560 000	30%	4,0	55 910 400	4,0	62 619 648	2,0	31 309 824	2,0	39 275 043	2,0	43 988 048
2	Разработка и поддержка системы управления, мониторинга, деплоймента, интеграция	320 000	30%	4,0	31 948 800	4,0	35 782 656	3,0	26 836 992	2,0	22 442 882	2,0	25 136 028
3	Сборка, релизы, обеспечение тиражируемости продукта для разных сред и проектов	340 000	25%	2,0	16 320 000	2,0	18 278 400	2,0	18 278 400	2,0	22 928 425	2,0	25 679 836
4	Техническая поддержка 1 и 2 линии 24x7. Компетенции решения аналогичных проблем	230 000	15%	4,0	20 313 600	6,0	34 126 848	6,0	34 126 848	8,0	57 078 291	8,0	63 927 686
5	Связь команды разработчиков и бизнес заказчика - команда бизнес-партнеров	520 000	40%	4,0	55 910 400	4,0	62 619 648	2,0	31 309 824	2,0	39 275 043	2,0	43 988 048

ARENADATA

Платформа Arenadata: обзор продуктов

Алексей Струченко

PARTNER TECHNICAL ADVISER

Кто мы

ARENADATA — один из крупнейших в стране разработчиков системного программного обеспечения для хранения и обработки больших данных, построенного на базе технологий с открытым кодом.

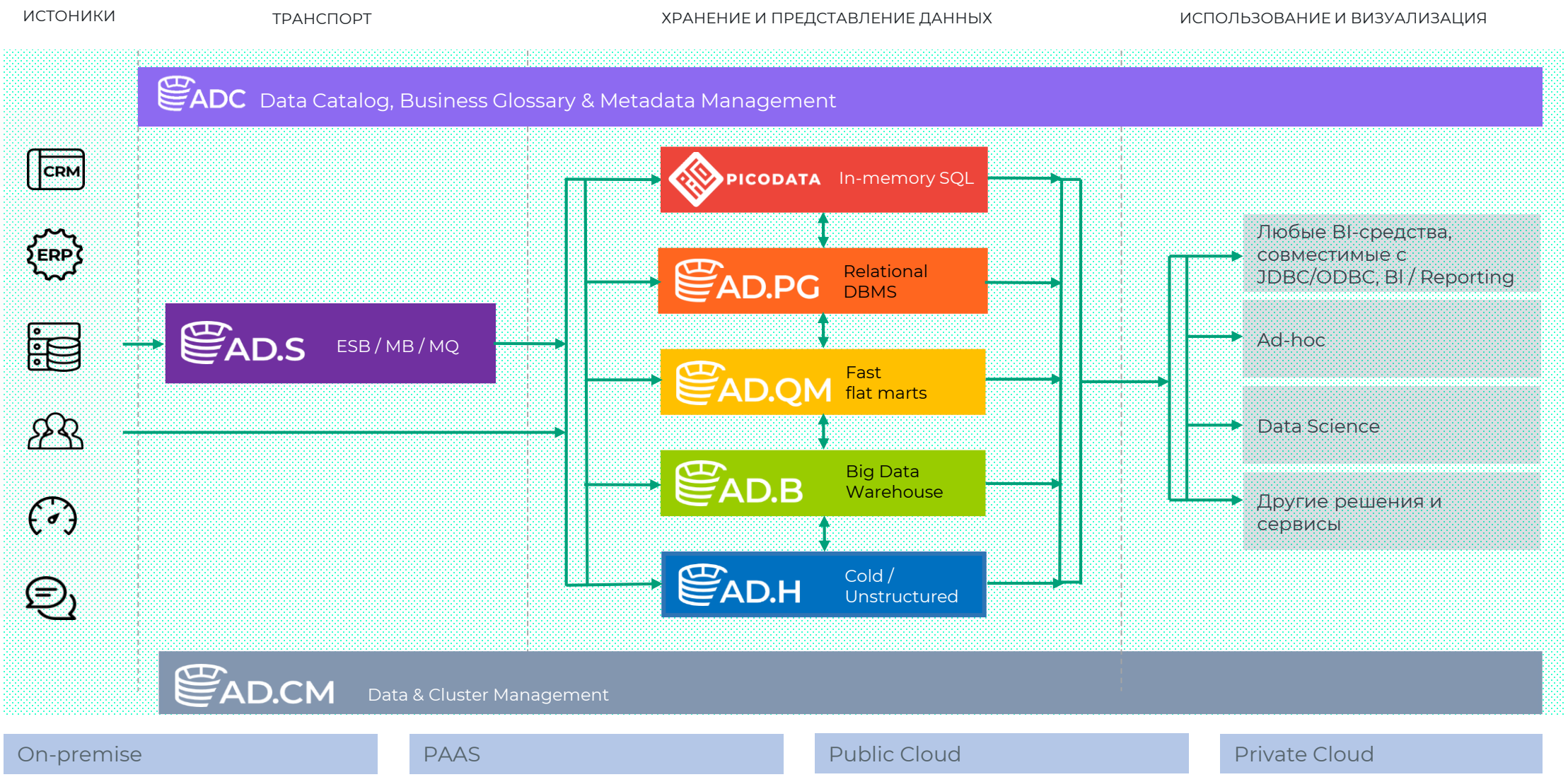
Как один из наиболее активных участников сообщества свободного ПО в России, Arenadata вносит вклад в развитие нескольких международных проектов. Все программные продукты компании объединены в многофункциональную платформу данных, которая позволяет строить надёжные гибко масштабируемые хранилища и озёра данных. Качество программного обеспечения Arenadata подтверждено государственными сертификатами и неоднократно проверено крупнейшими организациями России и зарубежья.

Наши решения востребованы при реализации задач по:

- Хранению больших данных
- Разгрузке legacy-систем
- ML & AI
- Озера и фабрики данных

ARENADATA

Arenadata Enterprise Data Platform



Streaming (ESB / MB / MQ)

Arenadata Streaming (ADS)	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. Может использоваться как: • корпоративная шина, • среда управления большими потоками данных, • фреймворк для разработки потоковых приложений.

Хранилище данных (MPP-СУБД)

Arenadata DB (ADB)

AD.B

Arenadata DB (ADB) — аналитическая, распределённая СУБД с открытым исходным кодом, использующая концепцию MPP (massively parallel processing), построенная на базе СУБД Greenplum.

ADB реализована на кластере из множества серверов и предназначена для хранения и обработки больших объёмов данных — до десятков петабайт.

Витрины данных

Arenadata QuickMarts (ADQM)	Arenadata QuickMarts (ADQM) — кластерная колоночная система управления базами данных, созданная на основе ClickHouse. С помощью ADQM можно генерировать аналитические отчёты в режиме реального времени, используя большие объёмы информации. Используется под быстрые витрины, в том числе в связке с Arenadata DB.

ARENADATA CLUSTER MANAGER

Arenadata Cluster Manager (ADCM) - универсальный оркестратор гибридного ландшафта. Позволяет быстро устанавливать, настраивать все data-сервисы компании и управлять ими независимо от инфраструктуры.

Возможности работы с сервисами:

- Установка и настройка
- Обновление
- Управление
- Мониторинг
- Настройка прав доступа
- Интеграция с другими сервисами

Возможности работы с инфраструктурой:

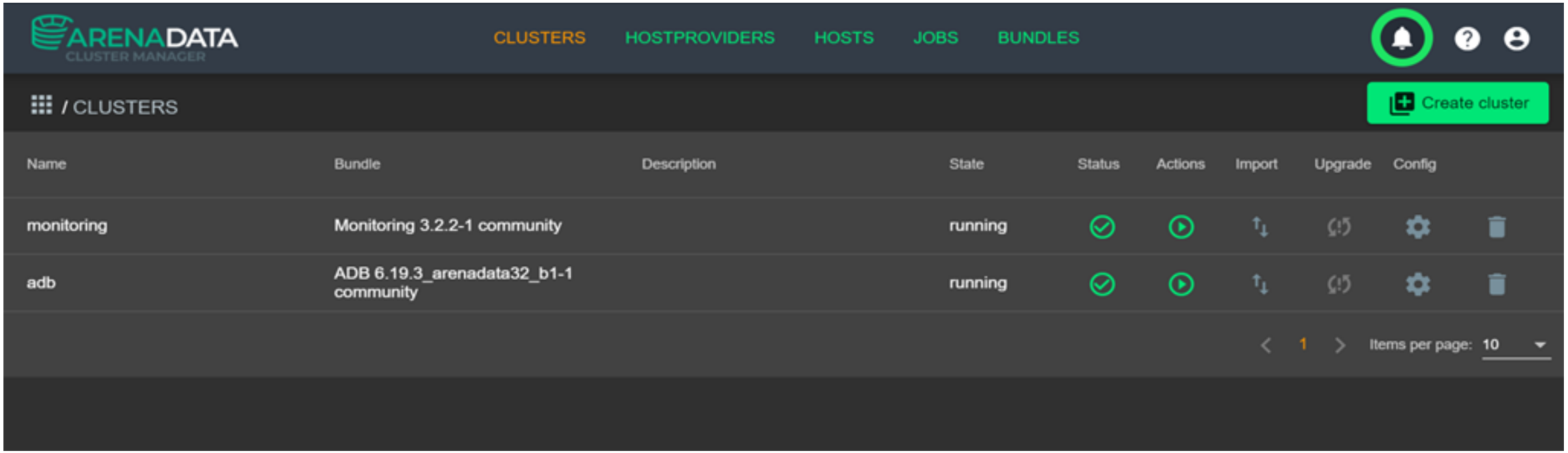
- Создание и удаление виртуальных машин
- Конфигурирование ОС
- Мониторинг
- Управление пользователями
- Настройка прав доступа

Дополнительные возможности:

- Настройка прав безопасности в ADCM
- Открытый ADCM API
- Открытый формат для создания новых бандлов - добавьте ваши собственные сервисы/инфраструктуру

в 10-ки раз быстрее вы сможете развернуть комплексный кластер из компонент Arenadata с помощью Arenadata Cluster Manager

Пример интерфейса ADCM



Вопросы и ответы

Терминология

1. **SLA** (Service Level Agreement) – соглашение об уровне предоставления услуги.
2. **ЦОД** – центр обработки данных (дата-центр).
3. **Tier III** - показатель надежности центра обработки данных. Ключевое отличие уровня Tier III — возможность ремонта и модернизации без отключения оборудования и остановки работы дата-центра.
4. **PCI DSS** (Payment Card Industry Data Security Standard) - это стандарт безопасности данных платёжных карт, учреждённый международными платёжными системами.
5. **VDI (**Virtual Desktop Infrastructure) — инфраструктура виртуальных рабочих столов.
6. **TC** (Thin Client) – тонкий клиент.
7. **ПО** – программное обеспечение.
8. **SDS** – Software Defined Storage – программно-определяемый слой хранения данных.
9. **SDN** – Software Defined Networking – программно-определяемая сеть.
10. **SDC** – Software Defined Computing – программно-определяемые вычислительные ресурсы.
11. **Open source** – открытое программное обеспечение.
12. **Framework** – фреймворк – программная платформа, определяющая структуру программной системы.
13. **API** (Application Programming Interface) – описание способов взаимодействия одной компьютерной программы с другими.
14. **OS** (Operating System) – операционная система.
15. **Management** – управление.
16. **vDC** (Virtual Datacenter) – виртуальный дата-центр.
17. **VM** (Virtual Machine) – виртуальная машина (ВМ).
18. **NIC** (Network Interface Controller) – сетевая плата.
19. **IP Pool –** это набор IP-адресов, доступных для распределения пользователям.
20. **Legacy** – устаревшее оборудование.
21. **CPU overhead** (Central Processing Unit) – перегрузка центрального процессора.
22. **vCPU** (Virtual Central Processing Unit) – виртуальный процессор.
23. **CPU overcommit** – использование большего количества ресурсов центрального процессора, нежели имеющиеся в наличии.
24. **commodity hardware** – это стандартные доступные по цене устройства, совместимые с другими подобными устройствами.
25. **SPA** (Single Page Application) – приложение одной страницы – это тип web-приложений, в которых загрузка необходимого кода происходит на одну страницу.