ARGUMENT

BIG DATA - BIG OPPORTUNITIES

Гибридная платформа данных ANG Platform

Большие данные - большие возможности

2025

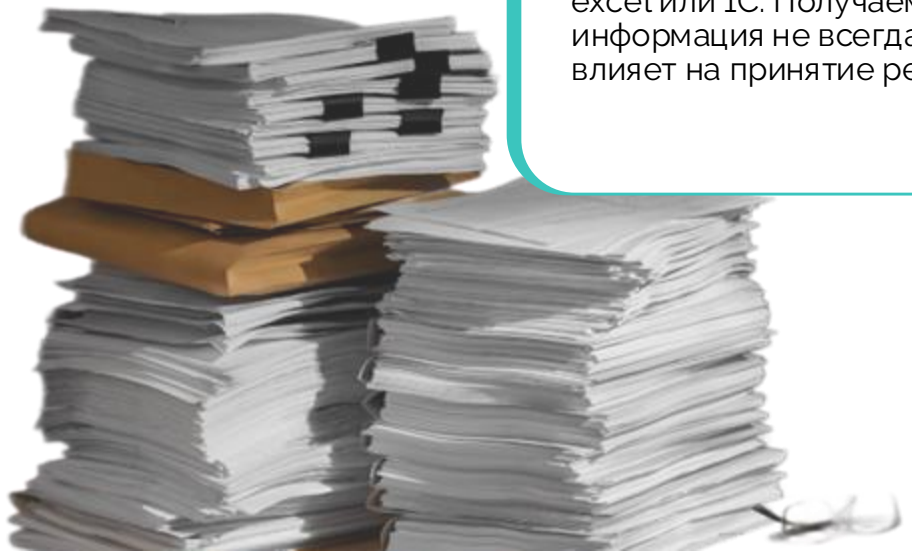
Эволюция цифровизации: путь от интуиции к аналитике

01

Ручное управление

Данные собираются не полностью или не собираются вообще.

Управление происходит на основе опыта или ощущений рынка.



02

Локальная автоматизация

Данных «мало», обновления и актуализация не регулярны, чистота – не высокая.

Для обработки достаточно excel или 1С. Получаемая информация не всегда влияет на принятие решения.

03

Автоматизация сквозных процессов

Данные активно собираются, внедряются системы автоматизации сбора и обработки информации.

Большое внимание уделяется настройке чистоты данных. Данные влияют на приятие операционных и стратегических.

04

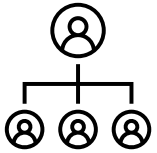
Data driven

Решения принимают только опираясь на данные. Становится возможным построение прогнозных моделей.

Качество, полнота, целостность данных и скорость их получения определяют финансовый результат.
Большие данные – ключевой элемент бизнеса.

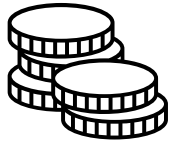
Data Driven:
как данные меняют правила игры в бизнесе

Эффективное управление



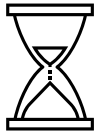
- Стратегические решения на основе данных, а не интуиции.
- Оперативная корректировка действий в реальном времени.

Оптимизация бизнеса



- Улучшение и автоматизация бизнес-процессов на основе аналитики.
- Создание цифрового двойника своего бизнеса

Скорость и точность



- Сокращение времени на принятие решений.
- Прогнозные модели для предсказания трендов и рисков.

Корпоративная культура

- Формирование единого подхода к управлению.
- Развитие Data Driven-мышления на всех уровнях компании.

Процессы DATA Driven

2

Обработка данных

Превращение «сырых» данных в полезную информацию: анализ для нахождения закономерностей, визуализация для ясности и принятие решений.

3

Анализ данных

Исследование и интерпретация данных для выявления закономерностей, трендов и полезной информации.

1

Сбор данных

Получение данных из различных источников: анализ на качество и полноту, визуализация для первичной оценки и принятие решений по оптимизации сбора.

4

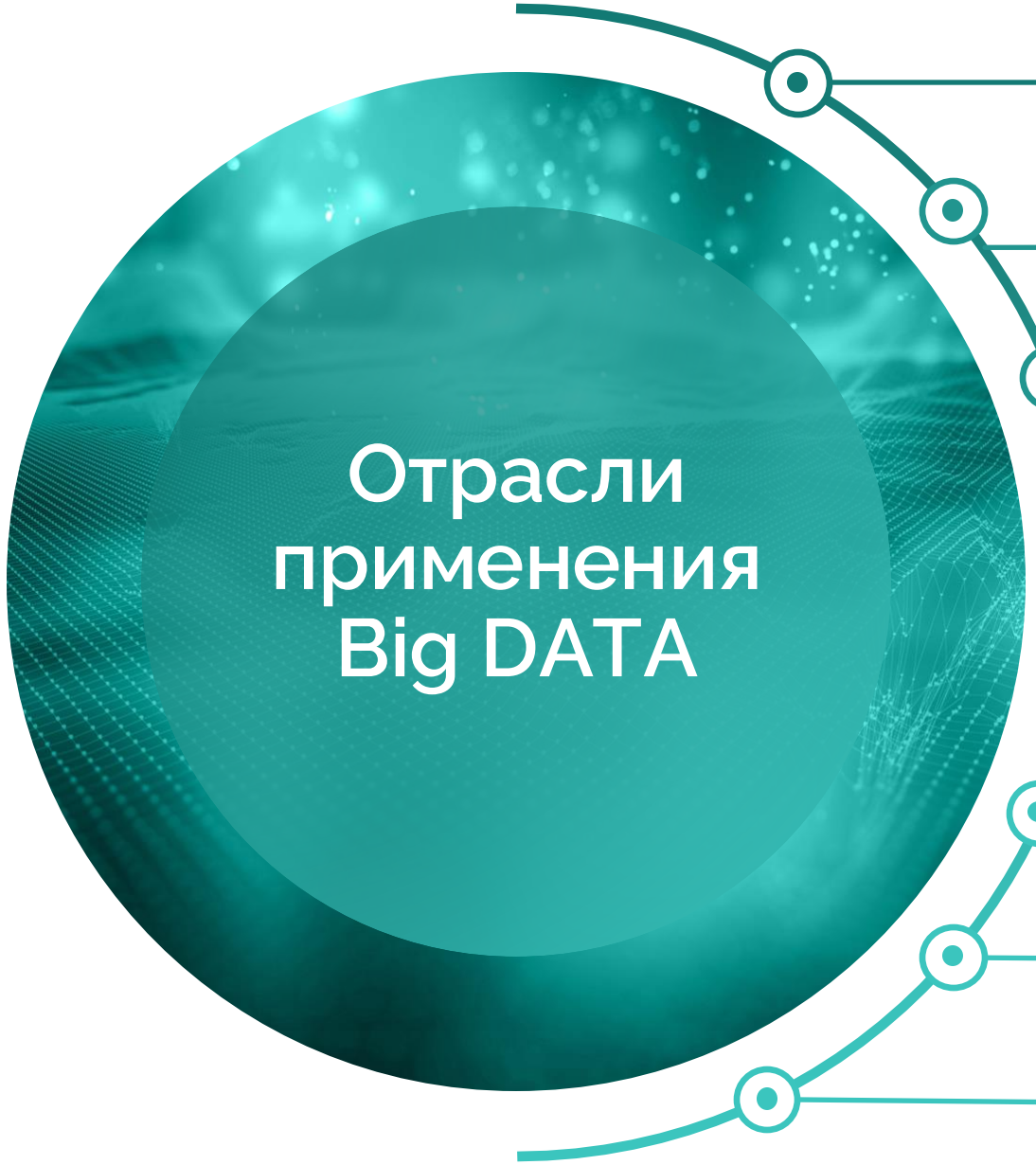
Визуализация данных

Представление данных в графическом виде (диаграммы, дашборды) для упрощения восприятия и анализа.

5

Принятие решений

Использование результатов анализа и визуализации для формирования стратегий и действий.



- **E-Commerce**
- **Страхование**
- **Телекоммуникации**
- **Производство**
- **Госсектор**
- **Банки и Финансы**
- **Медицина**
- **Логистика**

Современные архитектурные парадигмы

Место платформы Argument

DATA WAREHOUSE и MPP СУБД

- Высокое TCO
- Отсутствие потоковой обработки
- Нужен Data Lake для кейсов снижения TCO, ML-кейсов, не табличных данных
- Дорого масштабируется (compute и storage вместе)
- Деградация производительности (> 500Тб, > 200 польз.) -> добавь еще 1н кластер + репликацию

DATA LAKE Экосистема Hadoop

- Низкое TCO
- НЕ поддерживаются транзакции out-of-the-box и НЕ обеспечивается консистентность данных
- Плохо масштабируется (compute и storage вместе)
- Сложность обслуживания – много компонентов
- Производительность сильно зависит от сборки

LAKEHOUSE Функционал MPP DWH с уровнем масштабирования облачного Data Lake

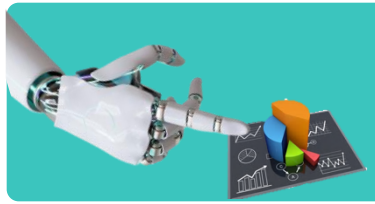
- Минимальное TCO
- Пакетная и потоковая обработка
- Дешево масштабируется (compute и storage раздельно)
- Изоляция ресурсов и задач
- Гибридная архитектура (cloud, on-premise)

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

Почему традиционные подходы к данным больше не работают?

Парадигма "Данные как продукт" (DaaP)

Дата-сеты как самостоятельные продукты, адаптированные под потребности пользователей.



Необходима революция в хранении и обработке данных

Экономические проблемы

- **Дорогое оборудование и высокооплачиваемые специалисты**
  → падает маржинальность.
- **Множественные хранилища**
  → рост расходов на разные типы данных, интеграцию разнородных ИТ-систем, обслуживание данных разными командами для разных групп пользователей.

Риски "перегруза" данными

- **Потеря контроля:**
  усложнение моделей, связей и качества.
- **Данные = риск,**
  если им нельзя доверять.

Запрос на инновации

Цифровизация требует:

- Real-time аналитики
- AI/ML-решений.

ПРОБЛЕМА:
текущие подходы экономически неэффективны.

Cloud Native — новый стандарт корпоративных решений

Kubernetes

- Централизованное управление контейнерами и оркестрация
- Гибкое распределение вычислительных ресурсов
- Автоматическое масштабирование приложений
- Балансировка нагрузки между узлами
- Автоматизация развертывания приложений
- Мониторинг состояния подов и узлов
- Обеспечение отказоустойчивости системы

S3 (Storage)

- Унифицированное хранилище данных
- Масштабируемое хранение информации
- Надежный доступ к данным
- Оптимизация затрат на хранение

Почему Argument?
Ключевые преимущества платформы

Единая платформа для работы с данными

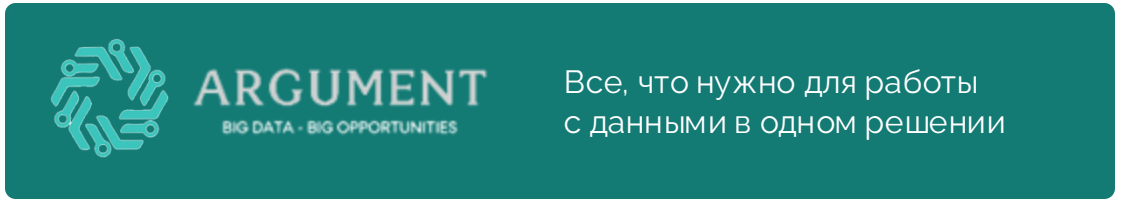
Open Source решение для сбора, обработки и анализа данных.

Быстрый переход на Data Driven

- Готовые инструменты для внедрения управления на основе данных.
- Снижение времени на адаптацию и обучение.

Масштабируемость под любые задачи

- Гибкость инфраструктуры – от стартапов до крупных предприятий.
- Инструменты для роста без ограничений.



Безопасность корпоративного уровня

- Соответствие Enterprise Security стандартам.
- Защита данных на всех этапах.

Полный контроль данных

- Оркестрация всех процессов: от хранения до аналитики.
- Интеграция с существующими ИТ-системами.

All-in-one: простота и эффективность

- Минимизация ручных усилий за счет автоматизации.
- Согласованная работа всех модулей.

Преимущества решения

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

Стоимость владения (TCO)

- Низкая стоимость хранения данных (дешевле, чем HDFS)
- Быстрые вычислительные движки (быстрее, чем Greenplum)
- Гибридная архитектура (PROD on-premise, DEV/TEST/QA on- cloud)
- Нужна одна команда для администрирования всей платформы

Оперативность данных & Time to Market

- Динамическое управление кластерами в зависимости от профиля нагрузки
- Высокая доступность данных из единого защищенного хранилища для любых инструментов системы.

Унификация платформы

- Объединение принципов подходов Lakehouse и Data Mesh
- Одна среда для всех задач Data Engineer и Data Science

Открытые форматы данных / API

- Apache Iceberg – как основной формат организации данных
- Гибкая совместимость с компонентами современной ИТ экосистемы от стриминга/CDC до BI
- Простота миграции данных - как между инструментами/БД, так и в облако

Функциональные возможности ANG

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

- Поддержка гибридной архитектуры (public / private cloud, on-premise) и модели multitenancy – виртуальные compute кластера над одними и теми же данными
- Возможность пакетной и потоковой обработки данных
- Разделение ресурсов для хранения данных и вычислительной среды
- В основе Open Source вычислительные движки Apache Impala, Apache Spark, Trino
- Оркестрация и управление ресурсами средствами Kubernetes
- Поддержка открытого формата хранения табличных данных Iceberg
- Единое хранилище данных для всех инструментов с единой ролевой моделью доступа
- Соответствует требованиям информационной безопасности

Argument: единая платформа для преобразования данных в бизнес-ценность

1

Единый источник истины

Извлечение и преобразование данных из любых источников с подготовкой к аналитике.

2

Централизованное управление данными

Аккумуляция, хранение и готовность данных для глубокого анализа.

3

Автоматизация рабочих процессов

Сокращение времени выполнения задач за счет интеграции и автоматизации.

4

Безопасность на уровне Enterprise

Защита конфиденциальных данных с помощью шифрования и строгой аутентификации.

5

Интеллектуальная аналитика

Аналитика позволяет лучше понимать рынок, выявлять тренды и прогнозировать будущие изменения.

6

Конкурентное преимущество

Адаптация к изменениям рынка и рост стоимости цифровых активов компании.

Argument для IT-руководителя: ключевые преимущества

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

01

RU Безопасность и соответствие

- Отечественная разработка (в реестре Российского ПО)
- Регулярные проверки на уязвимости
- Команда разработчиков в РФ

02

Гибкость интеграций

- Подключение любых источников данных через API
- Миграция с других платформ (включая ушедшие с рынка)
- Работа в облаке или закрытом контуре

03

Производительность

- Высокая скорость работы (современные технологии)
- Единая веб-консоль для управления всеми модулями

04

Экономическая эффективность

- Оптимальная стоимость владения
- Гибкие лицензии (платите только за нужные инструменты)
- Бесшовное масштабирование функционала

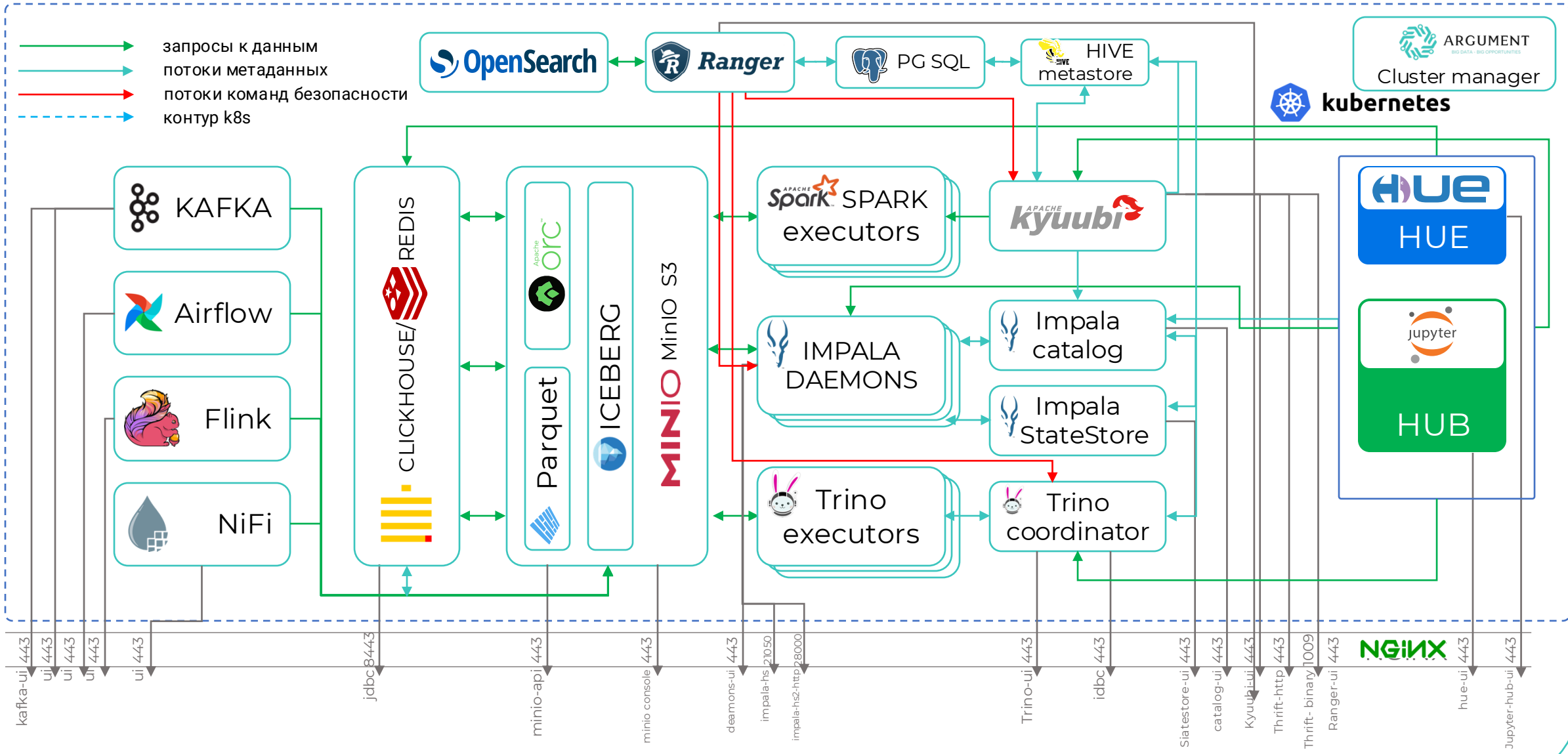
05

Поддержка

Квалифицированная техподдержка доступна 7 дней в неделю 24 часа в сутки

ANG Platform

ANG Platform базовая архитектура



Характеристики платформы

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

Гибкое и эффективное распределение вычислительных мощностей над единым слоем данных

Основные компоненты

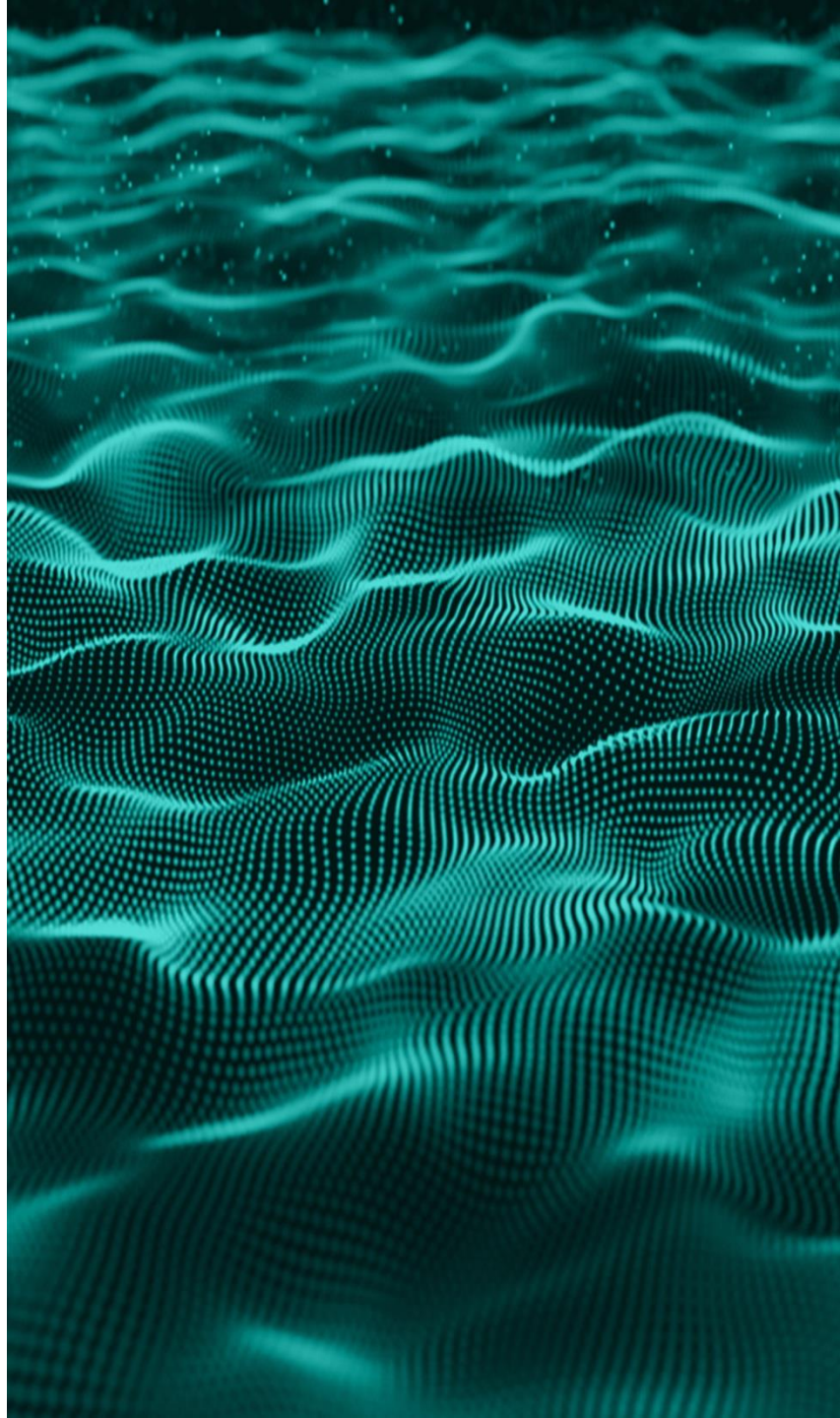
- Data Lake
- Analytical DWH
- Lakehouse

Преимущества

- Гибкое распределение вычислительных мощностей
- Поддержка JupyterHub и SQL web-ноутбуков

Функции

- ETL-движок (витрины данных для моделирования/
- отчетности/приложений)
- SQL-движок для BI-инструментов и self-service анализа данных
- Ad-hoc анализ данных:
  SQL web-ноутбук Hue с широкими функциональными возможностями
  Использование “толстого” SQL-клиента
  через JDBC или ODBC
  JupyterHub



Стратегия развития платформы ANG

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

Road map

Q3 2025

- Airflow для оркестрации процессов на платформе
- Hive Metastore v4.0
- Внедрение механизмов обслуживания Iceberg (Beta)
- Opensearch для полнотекстового поиска

Q4 2025

- ClickHouse для быстрой аналитики и storage для BI
- Redis для горячего in-memory кэша
- Feast (Feature Store для Data Science)
- Kube Flow для ML Ops

Q1 2026

- Kafka (стриминг данных)
- NiFi (low-code ETL)
- Flink (реалтайм аналитика)
- Улучшения механизмов обслуживания Iceberg

Q2 2026

- DBT (Data Build Tool)
- DuckDB для оперативной аналитики
- StarRocks
- OpenWebUI – веб-интерфейс для LLM
- vLLM – инференс для LLM

Q3 2026

- ADBC-коннекторы
- AI-ассистент управления хранилищем (рекомендации ETL)
- Улучшения механизмов обслуживания Iceberg

Q4 2026

- Интеграции с AI-платформами РФ (MTS NLP, AlphaGen, Cloud.ru, Rubbles, Сайбокс)
- Внедрение биллинга запросов пользователей (планирование бюджетов на масштабирование)

Варианты установки

ARGUMENT
BIG DATA - BIG OPPORTUNITIES

Продукт

Argument Next Gen

для компаний со средними и большими объемами данных до десятков петабайт, которые уже используют и развивают Data Driven подход.

Версии размещения

Облачный доступ (Cloud)

Разместите сервер в облачном ЦОД для оптимизации затрат на инфраструктуру и доступа к данным
из любой точки мира.

Локальный доступ (On-Premise)

Установите ПО на свой собственный сервер и работайте абсолютно независимо, даже в закрытом контуре.

ANG Platform – модули

Argument Cluster Console

веб-интерфейс развертки, администрирования
и конфигурирования компонентов системы

Argument Object Storage

система хранения данных, включающая себя компоненты
как для хранения данных в табличном формате,
так и компоненты объектного хранения

Argument Data Processing

контейнеризированная система обработки данных
на основе систем распределенных вычислений Spark, Trino,
Impala, покрывающая все известные подходы
к обработке данных.

Argument Client Tools

набор инструментов для пользовательского доступа
к данным и их визуализации

Argument Monitoring

система мониторинга состояния компонентов
кластеров по огромному количеству метрик

Argument Streaming

Пакет компонентов для обработки потока данных
для аналитики в реальном времени и других
сценариев обработки данных

Argument Data Base

компонент, реализующий функции горячего
слоя хранения на основе ClickHouse, Redis, Opensearch

Argument ML / AI Lab

компоненты для работы с моделями ML и LLM

Argument Security Tools

набор инструментов для реализации аудита действий
пользователей и разграничения доступа к данным

Предпосылки использования S3 object storage и Kubernetes

Проблемы традиционных решений

HDFS

- Объединенные ресурсы для хранения данных и вычислительных мощностей
- Проблемы маленьких файлов и большого количества файлов (Namespace Федерации на практике не используются)
- Избыточный дисковый объем - фактор репликации 3 (ECC на практике не используется)
- Не поддерживается публичными облаками
- Ozone частично решает проблемы, но все еще в beta

MPP

- Объединение ресурсов для хранения данных и вычислительных мощностей
- Проблемы одной узкой части/запроса -> проблемы всего кластера
- Движок ориентирован на full scan (индексы поддерживать «дорого») -> приводит к удорожанию дисковой подсистемы
- RAID-10 SSD/NVMe диски ZTD каждый
- Типовой MPP 6TB на узел vs. HDFS/S3 30-50TB на узел

YARN

- Отсутствие возможности разделения ресурсов для разных профилей нагрузок
- Часть сервисов просто не может работать в YARN
- YuniKorn (эволюция YARN для контейниризации) ушел в сторону работы над k8s

Преимущества S3 и Kubernetes

S3 Object Storage

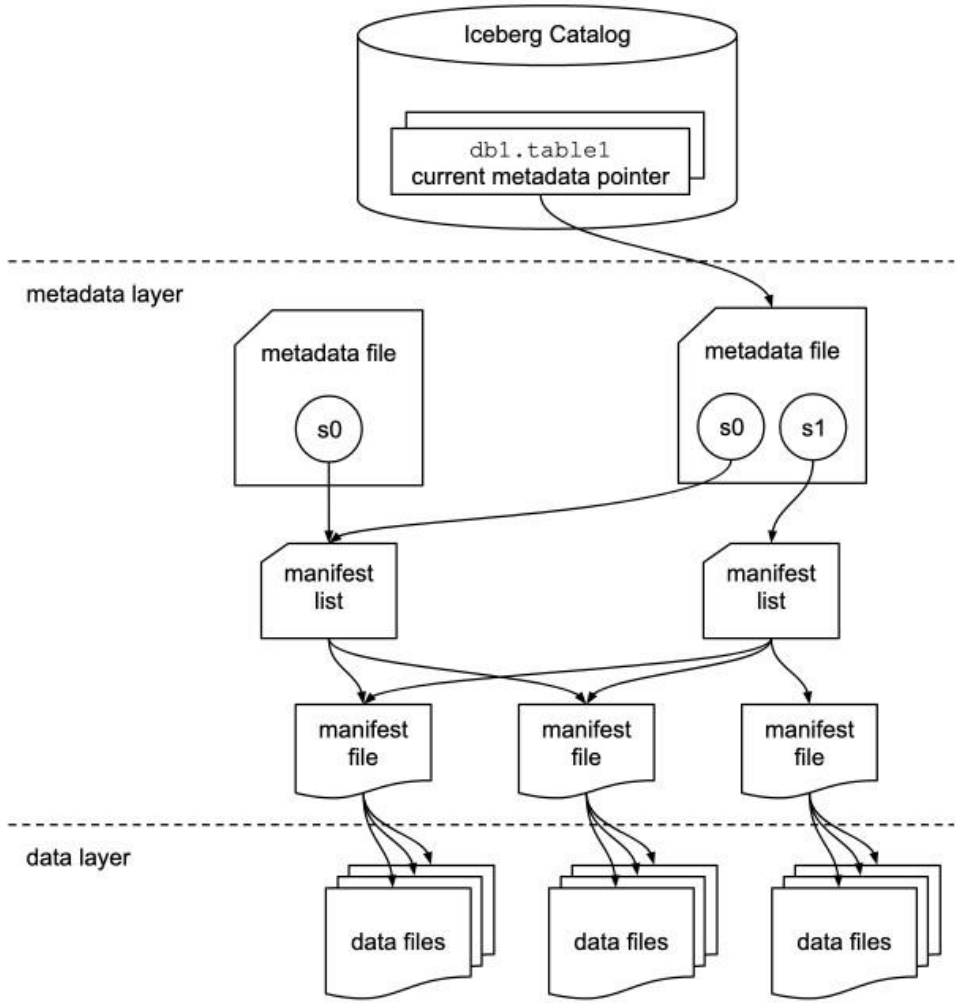
- Индустриальный стандарт, поддерживаемый большинством инструментов и движков
- Текущий мейнстрим
- S3 Select – фильтр данных на уровне хранилища
- Доступен в любом публичном облаке
- On-premise решение: Minio (выбор для аналитических задач, в отличие от CEPH, который ориентирован на универсальные облачные решения)

Kubernetes

- Индустриальный стандарт для управления вычислительными ресурсами
- Подходит для любого облачного окружения
- Развитая экосистема с готовыми решениями (helm charts/CRD)
- Быстрое развертывание виртуальных кластеров (<5 минут)
- Масштабируемость и единые инфраструктурные сервисы
- On-premise решение: Отдельный кластер k8s для аналитики, поддержка нескольких кластеров, возможность использования "ванильного" Kubernetes или других дистрибутивов

Преимущества Iceberg

- Соответствие требованиям ACID - целостность данных и согласованность обновлений
- Эволюция схемы хранения – любые изменения без пересоздания объекта
- Продвинутое партицирование – скрытое партицирование (разбиение на бакеты по выражениям без необходимости создавать вычисляемые поля),
  движки видят и используют скрытые партиции, изменение секционирования «на лету»
- Механизмы просмотра прошлых версий и отката на прошлые версии данных (time travel)
- Автоматическая оптимизация – сжатие, параллелизм UPDATE MERGE DELETE для Spark
- Поддержка потоковой передачи данных и микро-батчей
- Ближайшие альтернативы – Apache Hudi, Delta Lake



Почему ANG ?

Минимальное TCO

- Минимизация стоимости хранения данных за счет разделения compute и storage
- Перевод CapEx в OpEx у облачных провайдеров

Парадигма Lakehouse

- Объединение архитектурных принципов DWH и Data Lake – современный гибкий подход к организации хранения и аналитики

Гибридная архитектура

- Экономия расходов за счет возможности разворачивания платформы в облаках (private cloud) и на ресурсах Заказчика (on-premise) как следствие самая выгодное соотношение стоимости на производительность

Миграция и импортозамещение

- ANG обеспечивает полноценное замещение сервисов Cloudera

ARGUMENT

BIG DATA - BIG OPPORTUNITIES

Контакты

ang-platform.ru

argumentdata.ru

info@argumentdata.ru