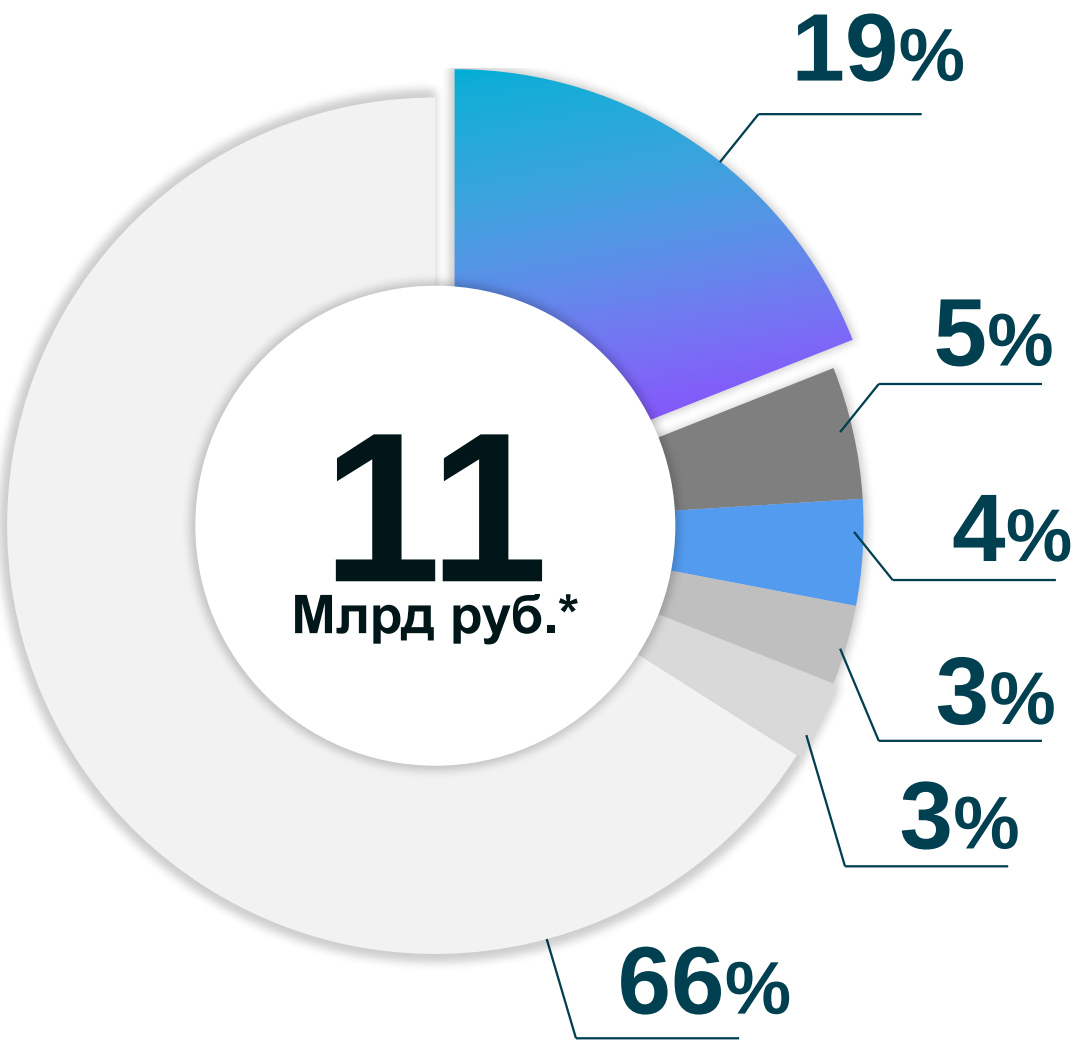
DIS GROUP

Применение Data Lakehouse для повышения эффективности ИИ

Николай Федоткин
Технический менеджер DIS Group

DIS Group – *лидер рынка* по управлению данными



1 место (23% рынка)

Компания DIS Group заняла в категории «Средства управления данными»

DIS Group (19%)

Юниверс Дата (4%)

} 23%

Прочие

*11 млрд руб. - объем рынка в сегменте "Средства управления данными"

Источник: исследование ЦСР «Рынок систем управления и обработки данных», 2024

Текущие *тенденции* в работе с данными

- Современная архитектура данных значительно повышает эффективность принятия решений
- Она обеспечивает четкую структуру и надежную основу для работы с данными
- Это критически важно для каждой data-driven организации

01

Синергетический эффект Data Mesh & Data Fabric

02

Подготовка данных для GenAI

03

Использование GenAI для решения задач управления данными

04

Создание корпоративных Data Lakehouse

05

Рост внимания к наблюдаемости данных

Источник: Data Architecture Trends in 2025, Dataversity, 2025

Управление данными в 2025:
баланс инноваций и рисков

01 Метаданные и ИИ трансформируют стратегию, архитектуру и управление данными

02 **80%** организаций делают ставку на метаданные, а **98%** IT-центров развивают ИИ-инициативы

03 GenAI – ключевой инструмент для многих компаний, но его эффективность напрямую зависит от качества и доступности данных

04 GenAI добавит **15-25%** к стоимости рынка данных и аналитики

05 **75%** компаний внедряют GenAI, но сталкиваются с техническими долгами и регуляторными рисками

06 GenAI – ключевой инструмент для многих компаний, но его эффективность напрямую зависит от подготовки данных, их качества и доступности

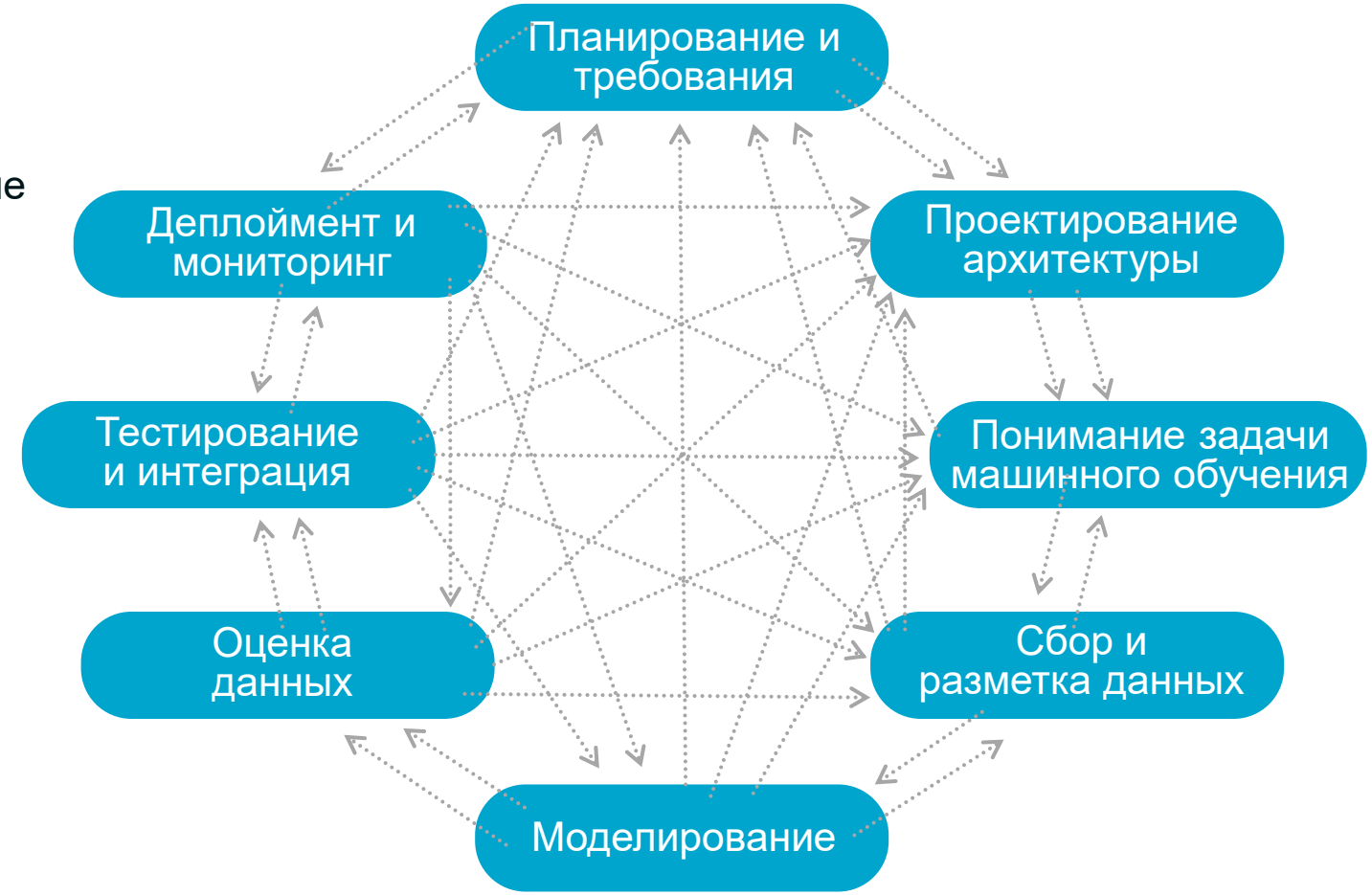
07 GenAI использует неструктурированные данные, что усложняет их управление

Источник: Data Architecture Trends in 2025, Dataversity, 2025

Жизненный цикл *Machine Learning*

Интерпретируемость, воспроизводимость, надежность

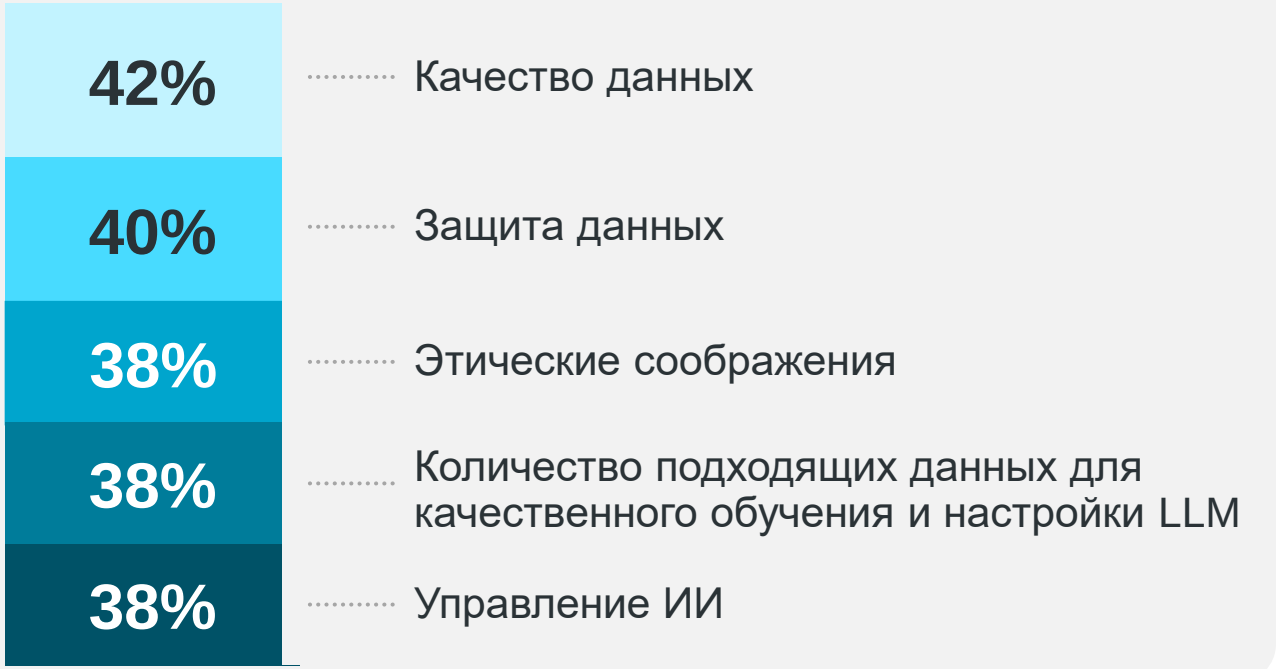
- Данные присутствуют на каждом этапе
- Нужны очищенные и структурированные данные
- Требуются размеченные наборы данных для обучения моделей
- Необходимы размеченные данные в формате, подходящем для машинного обучения
- Описание процесса сбора и обработки данных, включая используемые инструменты и методы
- Организуется отслеживание данных в жизненном цикле модели
- Требуется решить вопрос, как изменяющиеся данные сделать историческими, чтобы их можно было использовать

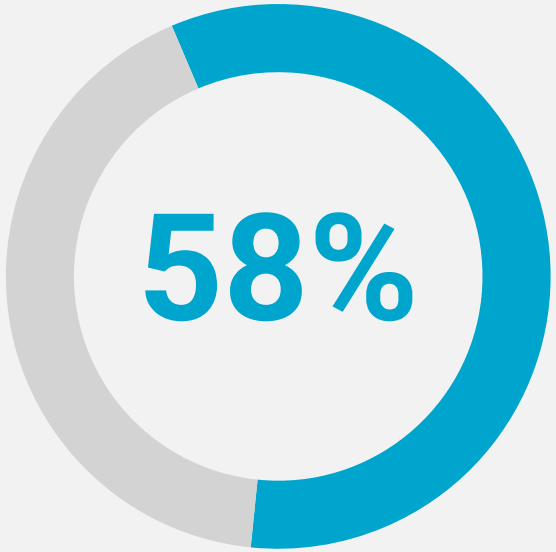


Готовность *достоверных* данных для ML

- Для эффективного принятия решения нужны существенные объемы **достоверных** реальных данных
- Качество данных должно быть **соответствующим** требованиям бизнеса для эффективного принятия решений

Основные сложности ИИ с данными



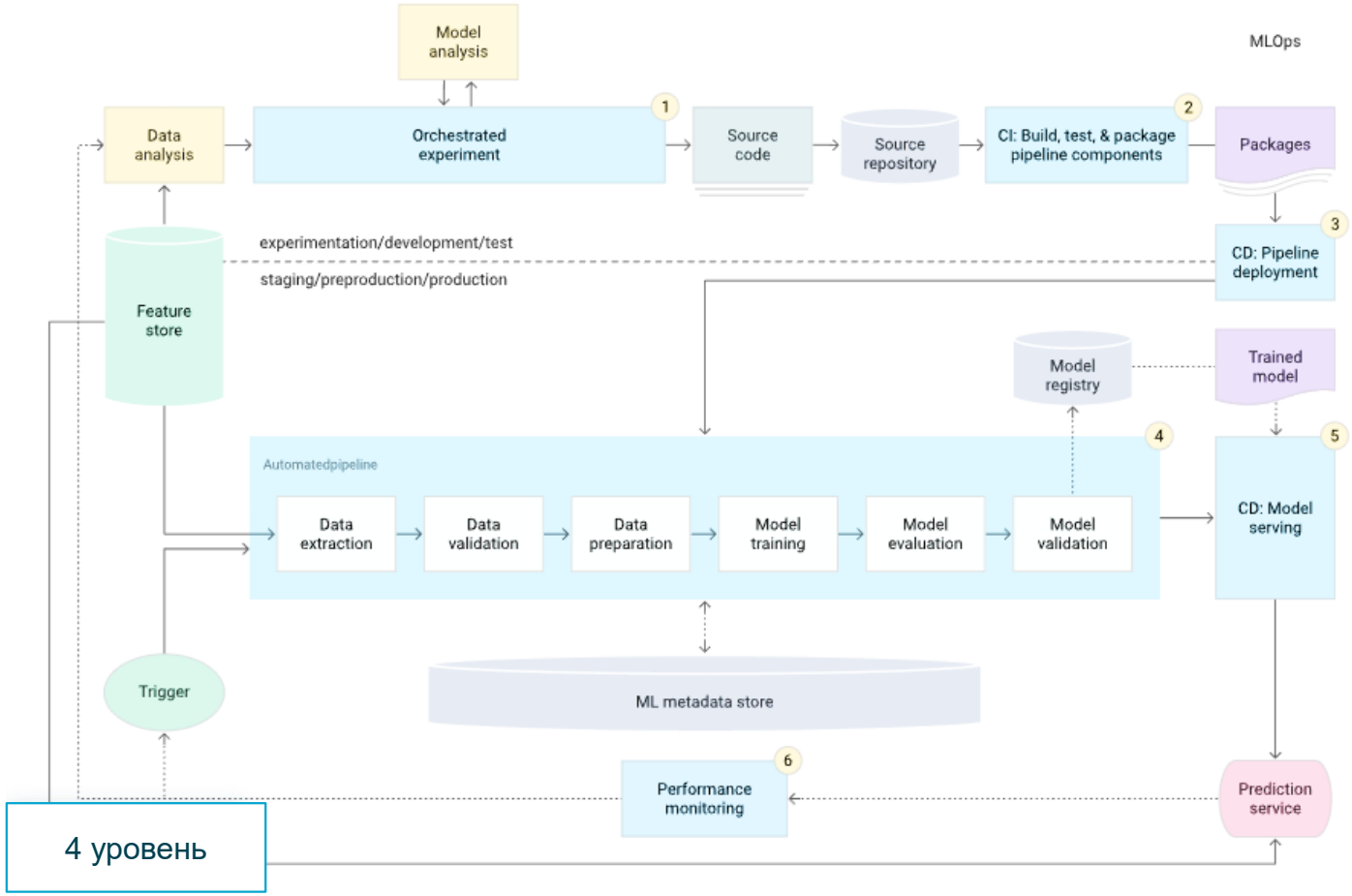


58% респондентов отмечают, что для подготовки достоверных данных для ML потребуется **5 и более** средств управления данными

Источник: CDO Insights 2024: Charting a Course to AI Readiness

Автоматизация на разных уровнях зрелости ML-процессов

- **0 уровень:** в организации не развиваются ML-процессы
- **1 уровень:** ручной, управляемый скриптами процесс
- **2 уровень:** автоматизация подготовки данных
- **3 уровень:** автоматизация пайплайна ML
- **4 уровень:** полная автоматизация конвейера CI/CD, вручную выполняются только этапы анализа данных и анализа модели



Требования к *подготовке* данных для ML

Для работы моделей в производственных средах есть следующий набор требований:

- 95% SLA между рекомендательным сервисом и внутренним сервером
- Высокая пропускная способность при пакетной обработке прогнозов через систему хранения и подготовки данных
- Низкая задержка при обработке запросов
- Отслеживание количества успешных, неудачных и прерванных вызовов
- Система должна контролироваться по потреблению аппаратных ресурсов
- Инфраструктура должна быть независимой от модели и времени выполнения
- Производственная среда не должна требовать частого изменения зависимостей, которые могут привести к сбою конвейеров данных и моделей во время выполнения
- Среда должна быть воспроизводимой, должна быть реализована возможность отката при сбоях
- Обеспечение версионирования каждого инфраструктурного пакета, чтобы конфликты, вызванные изменением зависимостей, можно было легко отладить

Архитектура решения для построения *конвейера* ML

DIS GROUP

Потребители данных

Бизнес-пользователи | Аналитики данных | Сайентисты данных | Чат-боты | Co-pilot | Приложения на основе GPT

2 Модели ИИ
Построение, обучение

- Расширенная генерация результатов поиска (RAG)
- Оперативное проектирование
- Системы агентов ИИ
- Тюнинг
- Языковые модели

Платформа ИИ / GenAI

1 Данные, готовые к использованию
Сбор и подготовка

- Предоставление данных
- Хранение данных: ODS | Озеро данных | Хранилище данных | Data Lakehouse | Master Data Management
- Интеграция и качество данных
- Интеграция приложений и управление процессами
- Исследование и разметка данных

Источники и приемники данных

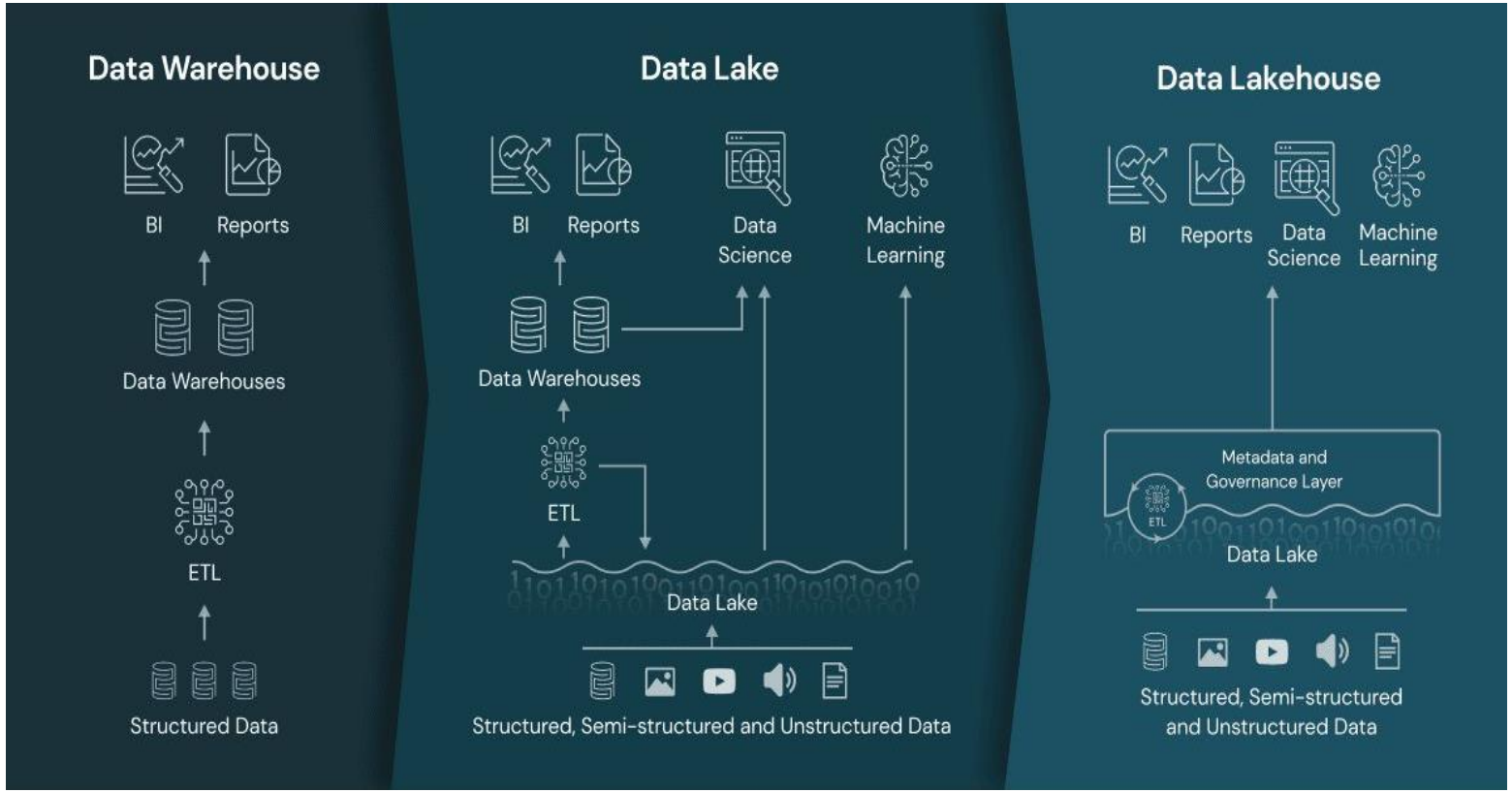
Озера данных | Очереди | Сервисы | Гибридные системы | SaaS | Облака | IoT | Хранилища данных | Машинные данные

3 Операции LLM
Планирование, управление и исполнение

- Внедрение
- Управление
- Наблюдение
- Оркестрация
- FinOps

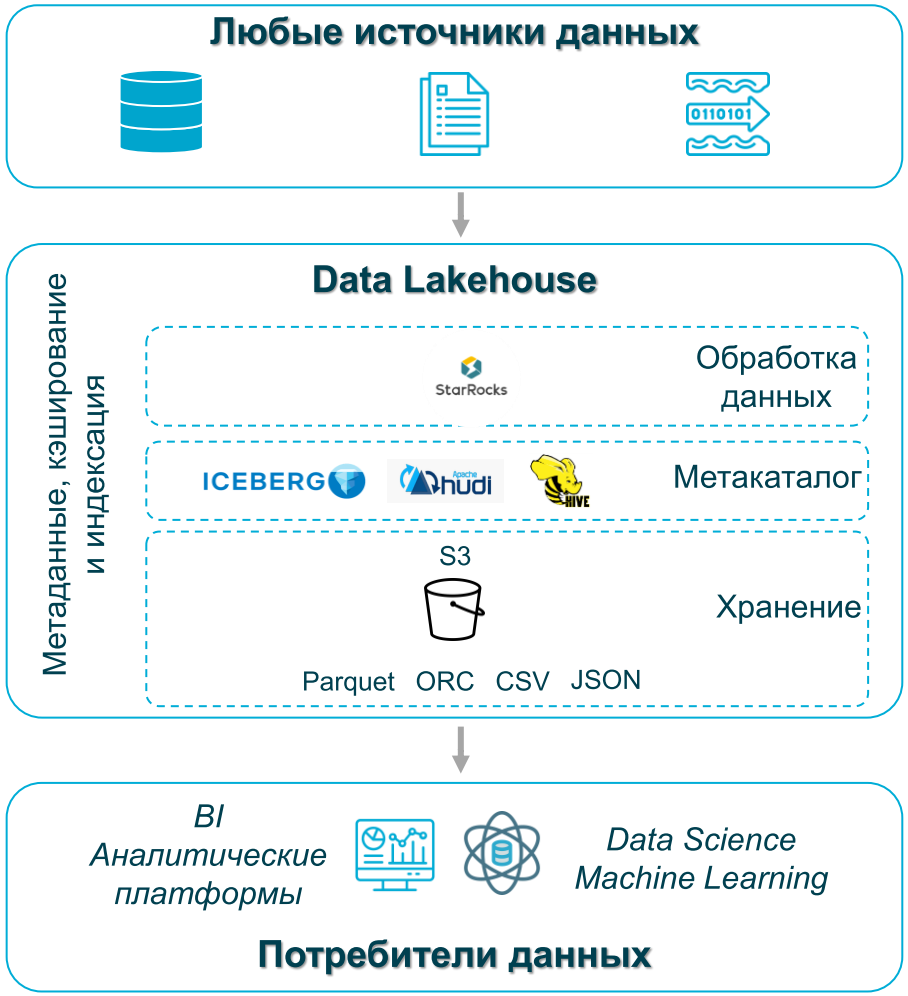
Корпоративное решение Data Lakehouse

- Основная система подготовки и хранения данных для ML
- Решение для решения задач роста объемов данных и повышения эффективности их использования
- Высочайшая производительность и минимальные задержки
- Отказоустойчивость
- Архитектура DLH оптимизирует запросы и процессы обработки данных, что делает работу с большими объемами данных более эффективной
- Data Lakehouse объединяет структурированные и неструктурированные данные в одном хранилище, что расширяет возможности аналитики и подготовки данных для ML
- Решение позволяет федеративно обращаться к внешним данным



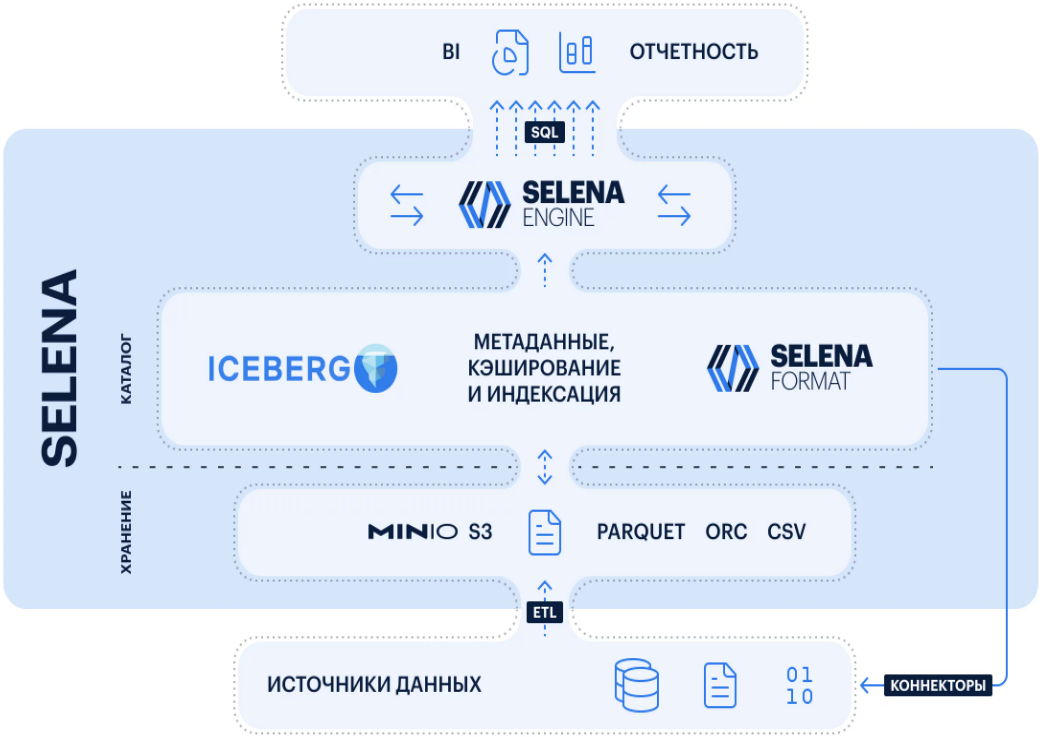
Селена – эволюция аналитических решений с архитектурой Data LakeHouse

- Платформенное решение для обработки и хранения больших объемов данных
- Массивно-параллельная обработка данных:
  - Колоночное хранение
  - Векторная обработка
  - Встроенный оптимизатор запросов
  - Бесшовная интеграция с хранилищем S3
- Объектное хранилище данных S3:
  - Возможность хранения сотен петабайт данных
  - Мощные возможности масштабирования
  - Резервирование данных
- Масштабирование уровня хранения не зависит от масштабирования уровня вычисления



Функциональные *возможности*

- Аналитическая платформа, предназначенная для работы с любыми данными любого объема и типа
- Разделение ресурсов для хранения данных и вычислительной среды
- Возможность инсталляции на виртуальной, физической и конвейеризированной среде
- Поддержка отрытых форматов хранения данных Iceberg, Hudi, DeltaLake, ORC и т.д.
- Совместимость с протоколом MySQL
- Встроенные инструменты управления, масштабирования, мониторинга и резервного копирования
- Синхронные и асинхронных materialized view
- Кэширование запросов и данных
- Федеративный доступ к данным



Освобождение данных
для использования

SELENA | DIS GROUP

Переход от проприетарных форматов данных к открытым форматам

ICEBERG

Open Table Format
(Открытый Формат Таблиц)

Parquet

Open File Format
(Открытый Формат Файлов)

Интероперабельность и открытые стандарты.
Переносимость и совместимость данных как бизнес-стандарт

Независимость от поставщика. Решают проблему "зависимости от поставщика", когда пользователи зависят от одного вендора для обновлений и поддержки

Долгосрочная защита инвестиций. Данные остаются доступными и пригодными к использованию, даже если первоначальные системы управления данными устареют

Экономическая эффективность. Снижают затраты на лицензии и проекты миграции данных, что освобождает ресурсы для инвестирования в другие аспекты бизнеса или инновации

Безопасность и аудируемость. Возможность независимого аудита и повышения уровня защищенности данных согласно внутренним и международным стандартам

Не только *Iceberg*

SELENA | DIS GROUP

SELENA FORMAT

Ускорение работы с данными.
Поддержка синхронных материализованных представлений

Быстродействие.
Максимальная скорость и производительность системы при обработке и загрузке данных

Поддержка и оптимизация.
Максимальная оптимизация функций при работе в нативным форматом хранения, в том числе и с данными в реальном времени

Гибкость реализации.
Возможность реализовать хранение в открытых и в нативном форматах

Унифицированность.
Все данные и метаданные управляются внутри системы и не требует дополнительных слоев или внешних систем для управления данными

Архитектура *платформы* Селена

SELENA | DIS GROUP

Клиенты — BI, Аналитика, Приложения, ML, Data Science

Обработка данных

Мастер - Leader
- Оптимизатор запросов
- Координатор
- Менеджер Каталога

Мастер - Follower
- Оптимизатор запросов
- Координатор
- Менеджер Каталога

Мастер - Observer
- Оптимизатор запросов
- Координатор
- Менеджер Каталога

Вычислительный узел
- Вычисление
- Чтение / Запись
- Кэш данных

Вычислительный узел
- Вычисление
- Чтение / Запись
- Кэш данных

Вычислительный узел
- Вычисление
- Чтение / Запись
- Кэш данных

Компоненты платформы

- **Cluster Manager** — tomcat
- **Мониторинг** — Grafana | Prometheus
- **Оркестратор** — Dolphin
- **AI&ML** — codellama
- **Metastore** — Apache Hudi, Delta Lake, Hive, Iceberg
- **Data Loading** — Trino

Хранение данных

S3 / HDFS

Parquet | ORC | CSV

Векторизованный механизм *выполнения* запросов

01 Вместо обработки данных по одному элементу за раз, **механизм работает с несколькими элементами одновременно**

02 **Улучшения при использовании кэша:** Векторизованные данные обычно хранятся в смежных областях памяти, что повышает локальность кэша и сокращает «промахи» кэша

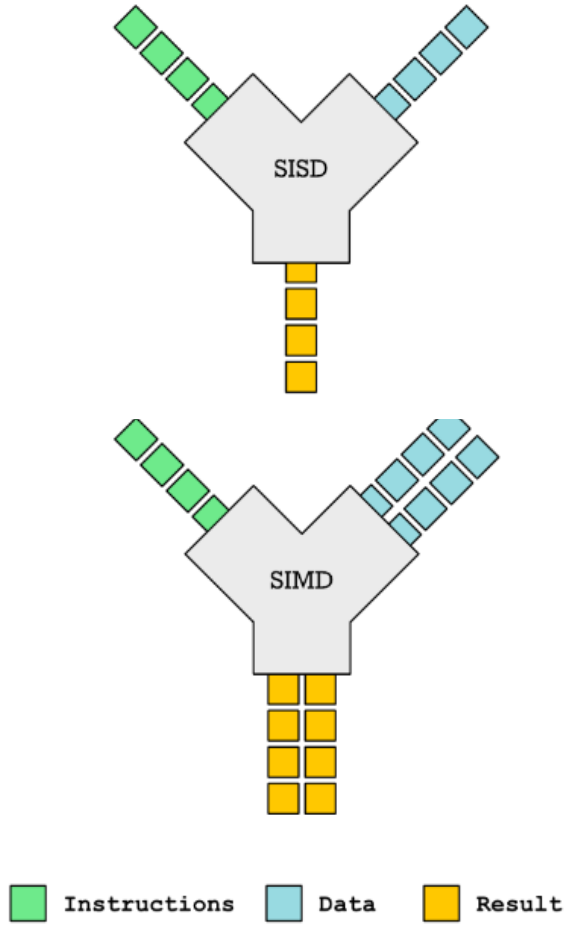
03 **Повышение пропускной способности:**

Одновременная обработка нескольких элементов сокращает количество необходимых инструкций

Такая обработка повышает общую скорость выполнения запросов

04 **Лучшее использование оборудования:**

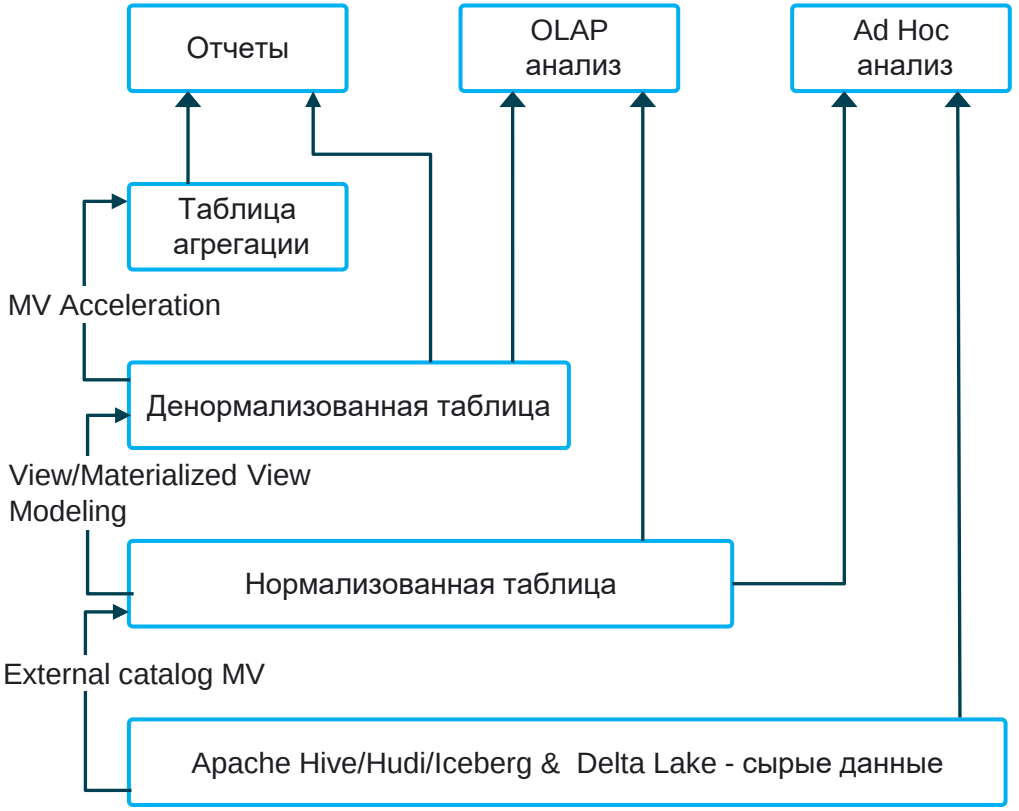
Современные процессоры разработаны для параллельной обработки, и инструкции SIMD используют этот потенциал для повышения эффективности ядра



Интеллектуальное *материализованное* представление

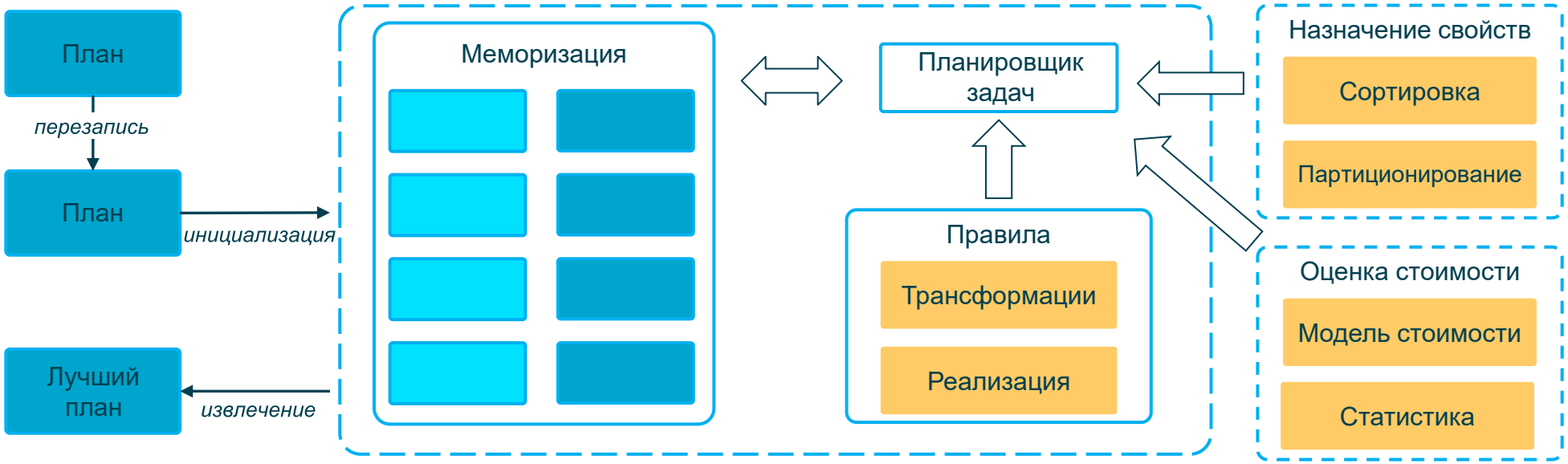
- Материализованные представления автоматически обновляют данные в соответствии с изменениями данных в базовой таблице, не требуя дополнительных операций по настройке
- Выбор материализованных представлений также происходит автоматически
- Механизм позволяет заменить традиционный процесс ETL:

  вместо преобразования данных в приложениях теперь есть возможность преобразовывать данные с помощью MV сразу в хранилище, упрощая конвейер обработки данных



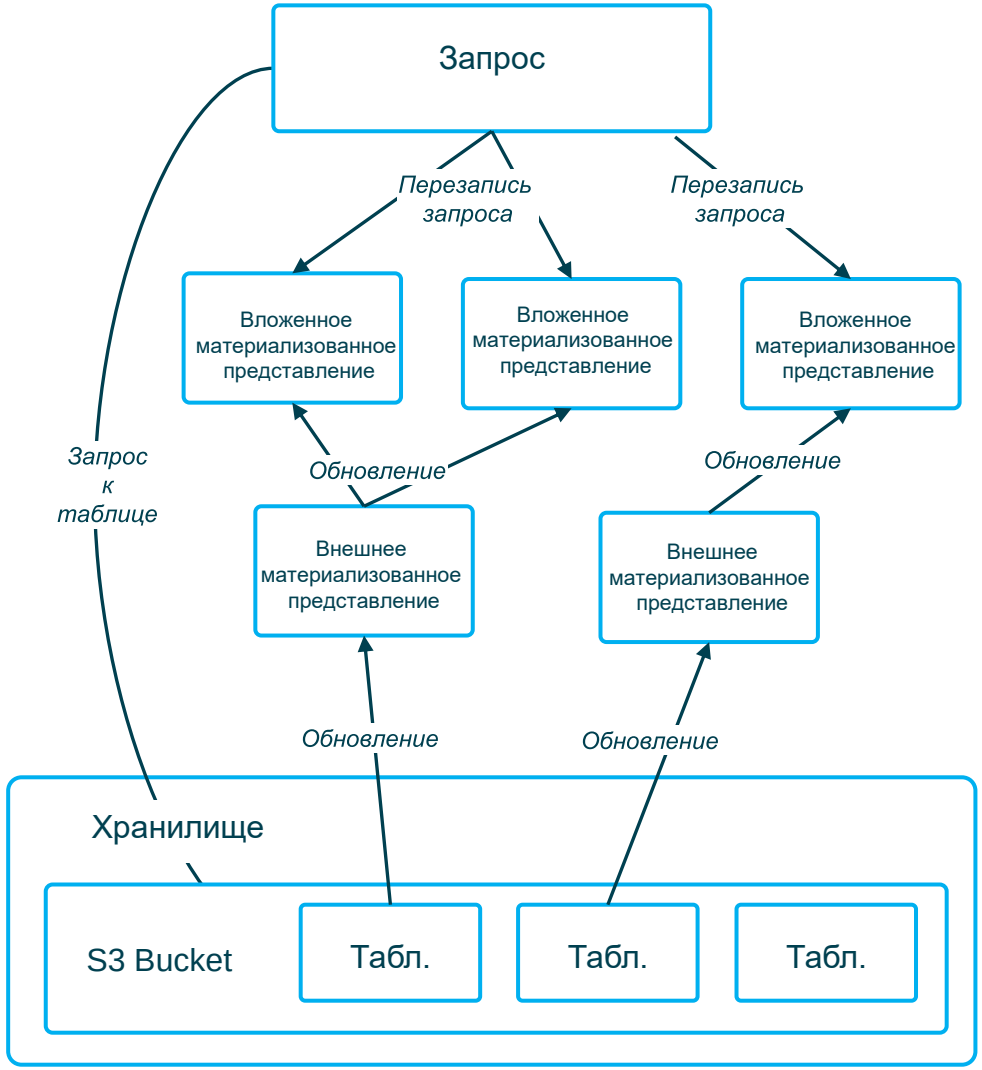
Оптимизатор на основе затрат

- Оптимизатор проработан для векторизованного движка выполнения с рядом оптимизаций
- Оптимизации включают повторное использование общих табличных выражений (CTE), переписывание подзапросов, боковые соединения, переупорядочивание соединений, выбор стратегии для распределенного выполнения соединений и оптимизацию с низкой кардинальностью
- Операторы Hash Join вместе с оптимизатором запросов позволяют обеспечивать лучшую производительность в сложных запросах на многотабличном соединение

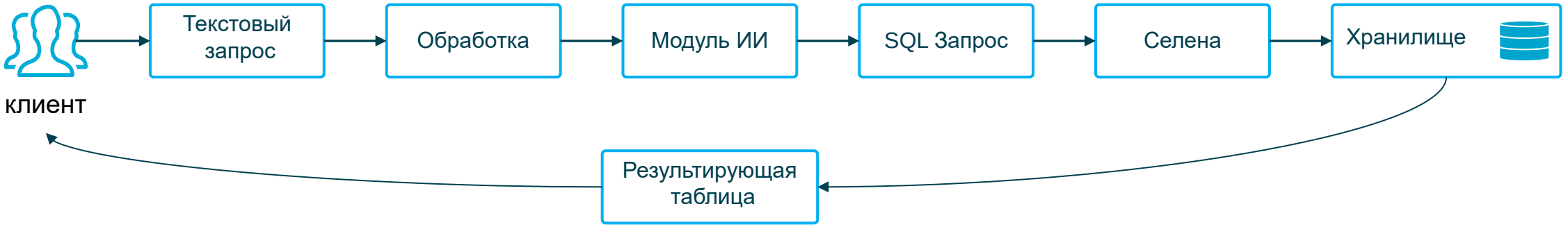


Автоматическое *преобразование* запросов

- Автоматическое преобразование исходного запроса пользователей в альтернативную формулировку с целью достижения того же результата
- Переписывание запросов с таблиц к материализованным представлениям, в которых хранятся предварительно вычисленные данные
- Оптимизатор может переупорядочить порядок операций в запросе
- Оптимизатор может определить и удалить ненужные части запроса, которые не влияют на конечный результат



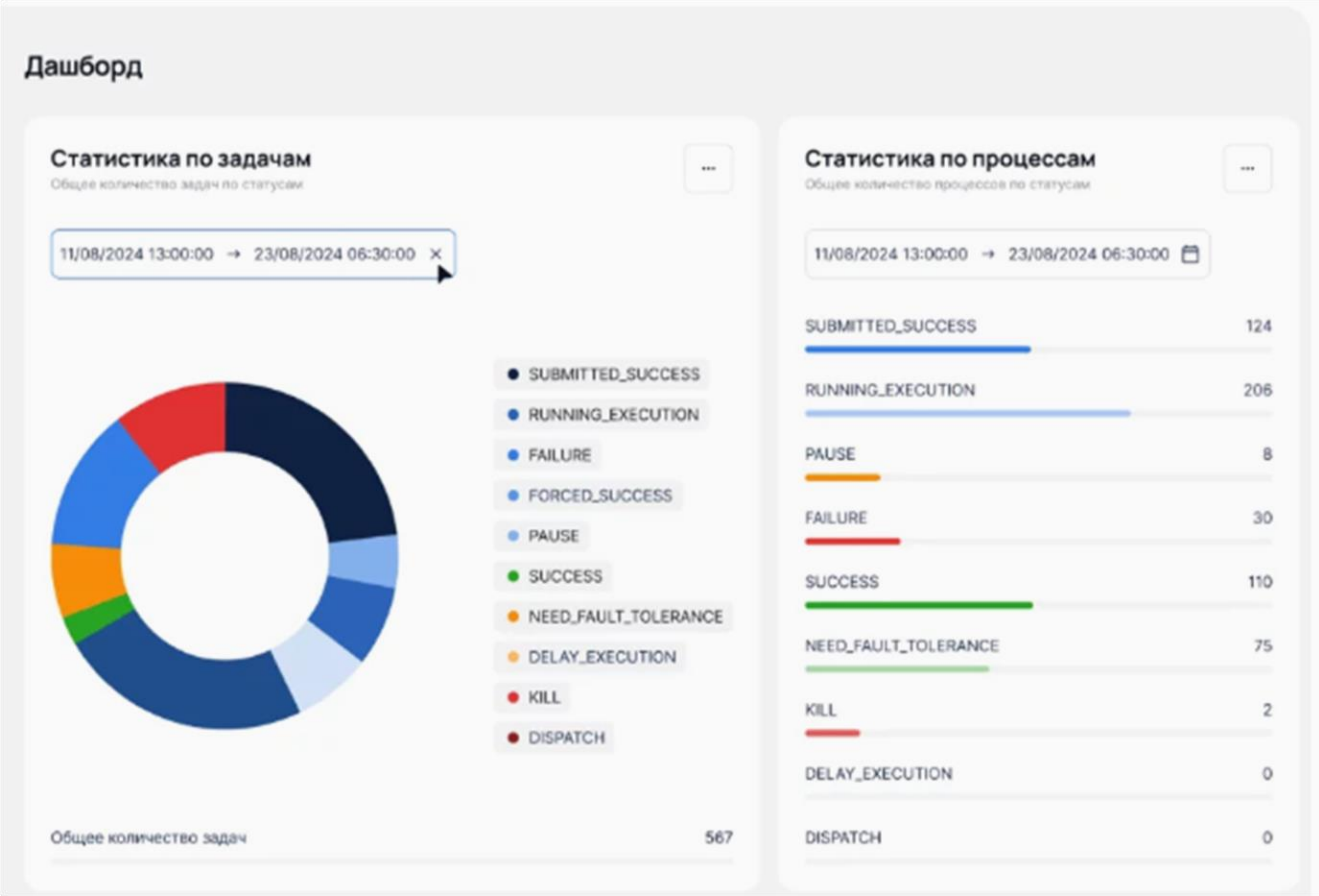
Поддержка *генеративного* ИИ



- Упрощение работы с данными с помощью ИИ
- Поддержка векторного хранилища
- Поддержка построения больших языковых моделей (LLM)
- Готовый модуль, обеспечивающий семантический поиск данных
- Различные сторонние кейсы применения, включая систему рекомендаций, поиск аномалий, чат-боты и т.д.

Высокая производительность

- Высокая скорость обработки запросов благодаря архитектуре, основанной на колоночном хранении данных, векторизации вычислений, оптимизаторе на основе затрат и «умных» материализованных представлениях
- Масштабирование до тысяч узлов без значительной деградации
- Поддержка различных схем данных: плоская, «звезда», «снежинка» и т.д.



Преимущества платформы *Селена*

Единая точка предоставления данных

Высокая скорость обработки большого количества запросов

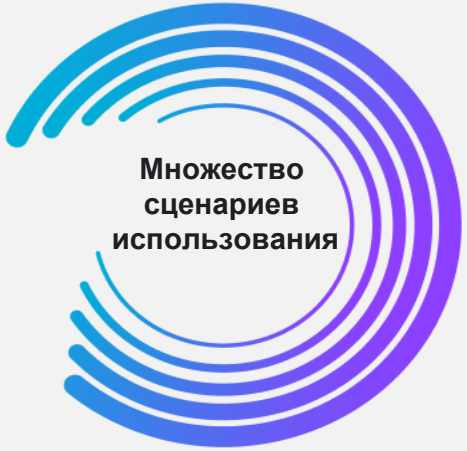
Независимость уровня обработки и уровня хранения данных, что позволяет экономить вычислительные ресурсы при обработке данных по сравнению с Greenplum и Hadoop

MPP (Параллельные вычисления)

Высокая производительность обработки данных при построении витрин данных

Поддержка гибридной инфраструктуры On-prem и Cloud

Высокая надежность и отказоустойчивость системы



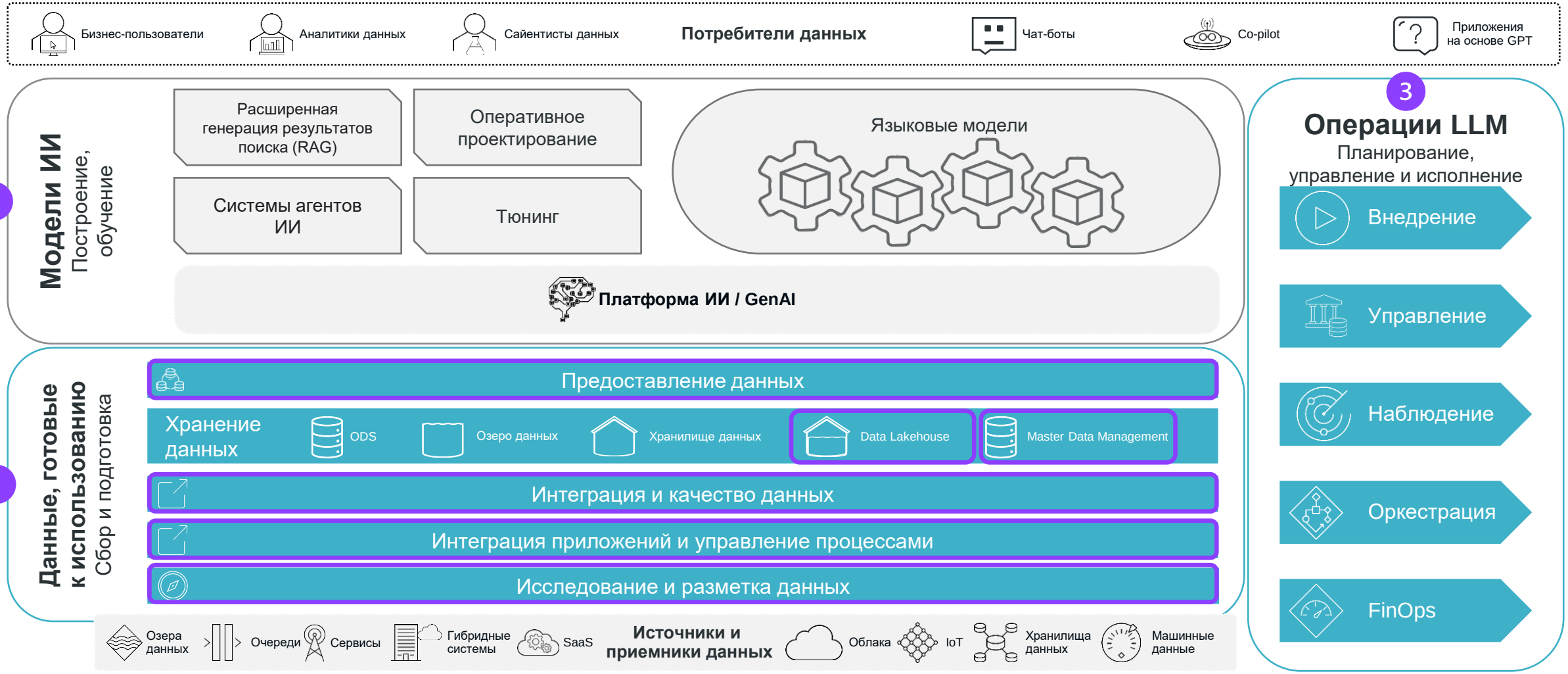
- Многомерная аналитика OLAP
- Аналитика в реальном времени
- High-concurrency аналитика
- Унифицированная аналитика

Гибкое и легко масштабируемое решение

Единое хранилище любого типа данных с открытыми форматами – сочетает в себе возможность классического хранилища данных и озера данных

Низкая стоимость владения - одно решение заменяет набор компонентов Greenplum, Hadoop, ClickHouse, ETL

Архитектура решения с платформой *DIS Group*



Данные как *конкурентное* преимущество

01 Управление метаданными и искусственный интеллект меняют то, как организации извлекают выгоду из своих данных

02 Ключевые направления – освоение инноваций и формирование культуры работы с данными

03 Организации, инвестирующие в развитие технологий, метаданные и культуру данных, будут лидировать в цифровой экономике

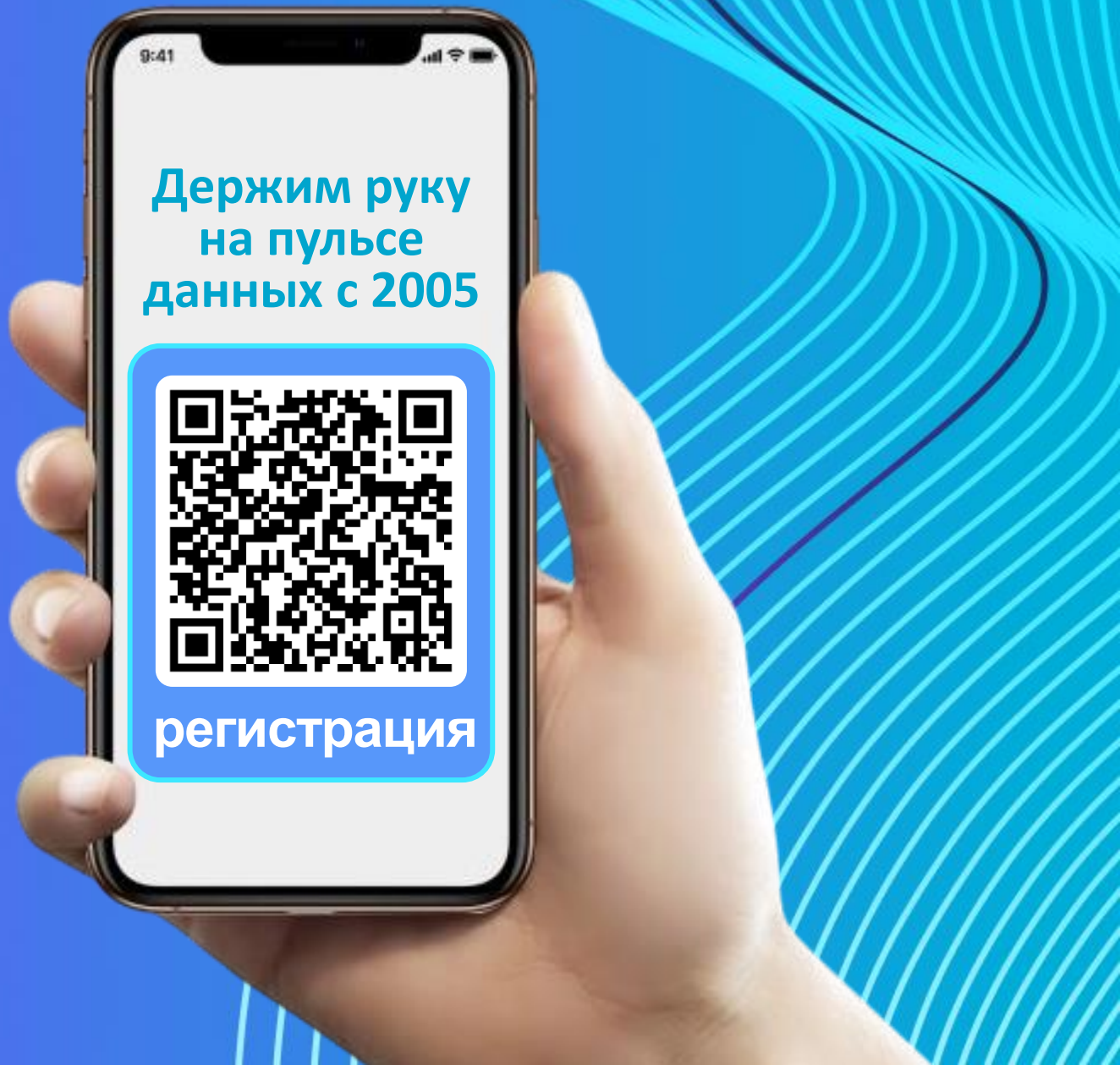
DIS GROUP

ДАТА САММИТ 2025

ПУЛЬС ДАННЫХ

17 апреля

офлайн + онлайн



Спасибо
за внимание!

Будьте с нами – подписывайтесь
на наш телеграм-канал