
DATA
meetUp
Data
Under
Control
click
online
offline

Data Under Control

# На пути от КХД и озер к современным Data Lakehouse: рекомендации по выбору компонентов Data Lakehouse платформы

Data Under Control

- Развитие КХД и Озер данных
- Переход к современным арх-рам на базе Data Lakehouse и какие выгоды приобретает компания
- Основные типы компонентов платформы
- Основные критерии выбора для компонентов
- Технологии для реализации
- Основные тренды
- Рекомендации при планировании и проектировании компонентов

# АЛЕКСАНДР СУЛЕЙКИН

## Founder DUC Technologies

Кандидат технических наук

**Telegram:**
**@dataundercontrol**

### 10+ лет
**опыта в проектах**
внедрения и проектирования систем больших данных, озер и хранилищ

### 5+ лет
**опыта преподавания**
в НИТУ МИСиС и НИУ ВШЭ
«Архитектура Big Data систем»
Доцент НИТУ МИСиС

### 34+
**научных публикации**,
индексируемых в WoS & Scopus, в т. ч. Q1

### Закончил
**НИУ ВШЭ**
по специальности «Big Data Systems» (англ.)
**Аспирантуру ИПУ РАН**
по специальности «Автоматизация и управление технологическими процессами и производствами»

### Защитил
**кандидатскую диссертацию**
по теме «Методы анализа и синтез архитектуры цифровых производственных экосистем»

### Опыт
**внедрения решений**
более чем в 6 предметных областях

### Победитель
**стипендиальных**
программ и конкурсов

Data Under Control

# 2025 State of the Data Lakehouse

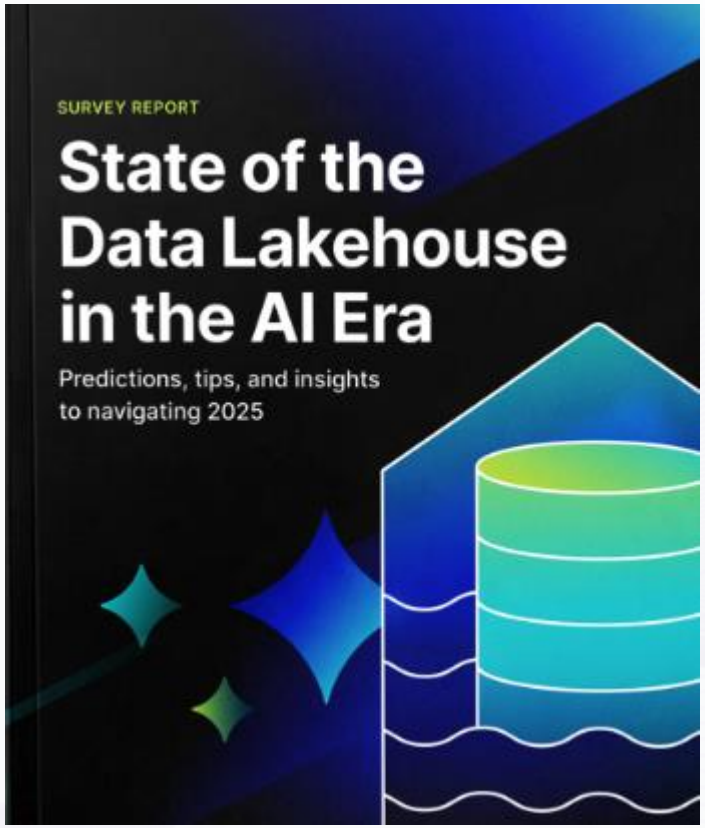
## Ключевые моменты:

**Тенденции внедрения:** использование Data Lakehouse достигло критической массы и, по прогнозам, вырастет до 67% к 2028 году.

**Данные, готовые к использованию с помощью ИИ:** основные характеристики данных, готовых к использованию с помощью ИИ.

**Движущие силы предприятия:** экономическая эффективность, единый доступ к данным и self-service-аналитика в качестве главных приоритетов.

**Ключевые проблемы:** управление, подготовка данных и инфраструктурные препятствия на пути масштабирования ИИ.




**Также из отчета:**

**11%** активно планируют внедрить этот подход в ближайшем будущем.
**36%** называют управление данными и безопасность существенными препятствиями для внедрения.
**33%** выделяют стоимость и сложность, связанные с подготовкой данных, как основные проблемы.

**Стратегические последствия:** Организации, которые отдают приоритет разработке экосистем данных, готовых к ИИ, делая упор на доступность, надежное управление и неизменное качество данных, имеют больше возможностей для:

— Ускорения циклов инноваций.
— Улучшения процесса принятия решений с помощью аналитических данных на основе данных.
— Оптимизации операционной эффективности и сокращения затрат.
— Достижения значительных конкурентных преимуществ.

Data Under Control

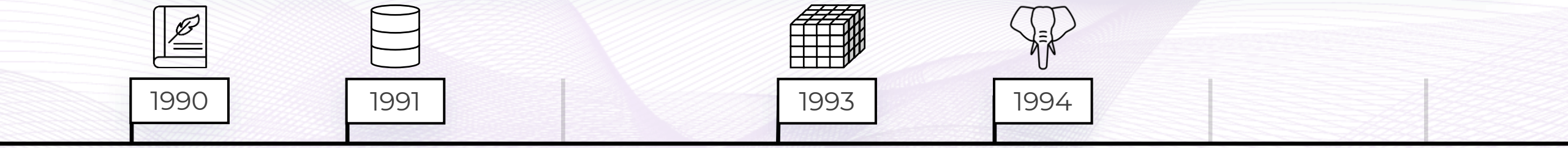
# Когда появилось понятие DWH и Datalake?

Data Under Control

# Немного истории - Появление понятия DWH

## 1990 – 2000

- **В 1991 году** американский ученый Билл Инмон публикует книгу «Building the Data Warehouse» и вводит в обиход концепцию DWH как централизованного хранилища всех данных компании, разработка которого начинается с создания нормализованной модели данных.
  Антагонист - Ральф Кимбалл, использующий подход «снизу вверх»
- **В 1993 году** уже известный Эдгар Кодд предлагает «12 правил аналитической обработки в реальном времени», так появляется OLAP (OnLine Analytical Processing)
- **В 1990-1991** Microsoft выпускает первые релизы реляционной SQL Server, которые сразу занимают лидирующие позиции на рынке СУБД.
- **В 1994 году** на основе некоммерческой СУБД Postgres с добавлением интерпретатора языка SQL появляется Postgres95 и начинает свой путь как Open-Source проект.

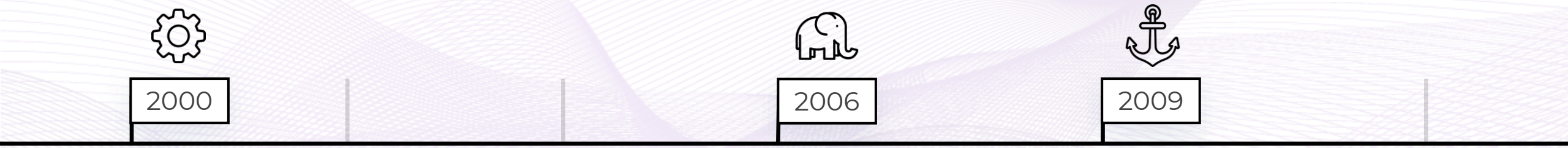


Data Under Control

# Немного истории - Появление Big Data

## 2000 – 2010

- Рост неструктурированных данных в социальных сетях
- Понятие Big Data
- Появление NoSQL СУБД
- MapReduce от Google
- **В 2006 году** на ее базе Дугом Каттингом создана основополагающая технология больших данных — распределенная ФС Hadoop (HDFS).
- В то же время **в 2000 году** Дэном Линстедтом представлена модель проектирования корпоративных хранилищ данных Data Vault, **а в 2009** Ларсом Рённбеком и Олле Регардтом — якорное моделирование (Anchor Modeling).



Data Under Control

# Немного истории – Datalake и Hadoop-экосистема

## 2010 – 2020

- Развитие проектов экосистемы Hadoop: Spark, Hive, Pig, Airflow, Kafka, Storm и др.
- **В 2011–2014** появляется несколько облачных проприетарных распределённых - Google BigQuery, Amazon Redshift, и Snowflake
- Тренд на облака
- Существенное увеличение кол-ва данных
- Развитие инструментов безопасности
- Развитие решение по DG
- Методологии управления данными и лучшие практики (DAMA и др.)
- Сосуществование DWH и DataLake – разные паттерны использования

# 2020 – настоящее время

Использование DWH и Datalake:

- Дублирование данных и инфопотоков
- Сложная и дорогая поддержка
- Отсутствие единой точки правды (куда идти за Золотой записью клиента?)
- Трудозатратное масштабирование
- Простои железа, когда не работают пользователи и ETL – дорого
- Развитие Cloud Native решений и технологий гибкого увеличения ресурсов кластера (k8s)
- Сложно и дорого управлять

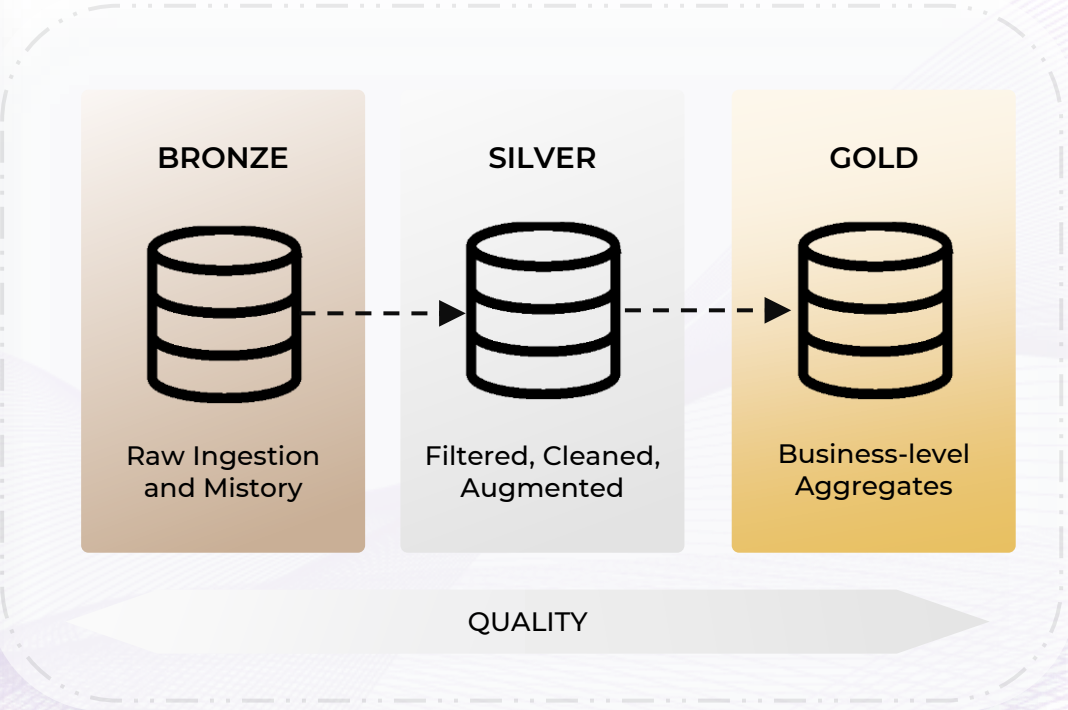
В связи с этим появились новые гибридные концепции хранилищ: **Data Lakehouse**, совмещающая преимущества озера данных и хранилища данных, и децентрализованная **Data Mesh**.

Data Under Control



# 2020 – настоящее время

- **История Lakehouse** начинается с предложенной компанией Databricks многоуровневой архитектуры Medallion Lakehouse, которая описывает преобразование данных от исходной формы к структурированной для бизнес-аналитики
- Lakehouse совмещает гибкость озер с четкой структурой классических хранилищ. Поверх Data Lake развертывают дополнительный слой для управления метаданными и реализации транзакций
- Поддерживает ACID-транзакции и SQL-запросы
- Поддерживает ML и библиотеки Python/R
- Дает доступ к разным типам данных, в том числе к изображениям, видео, аудио, ввод данных в реальном времени
- Позволяет использовать BI-инструменты без дополнительных обработок данных



Data Under Control

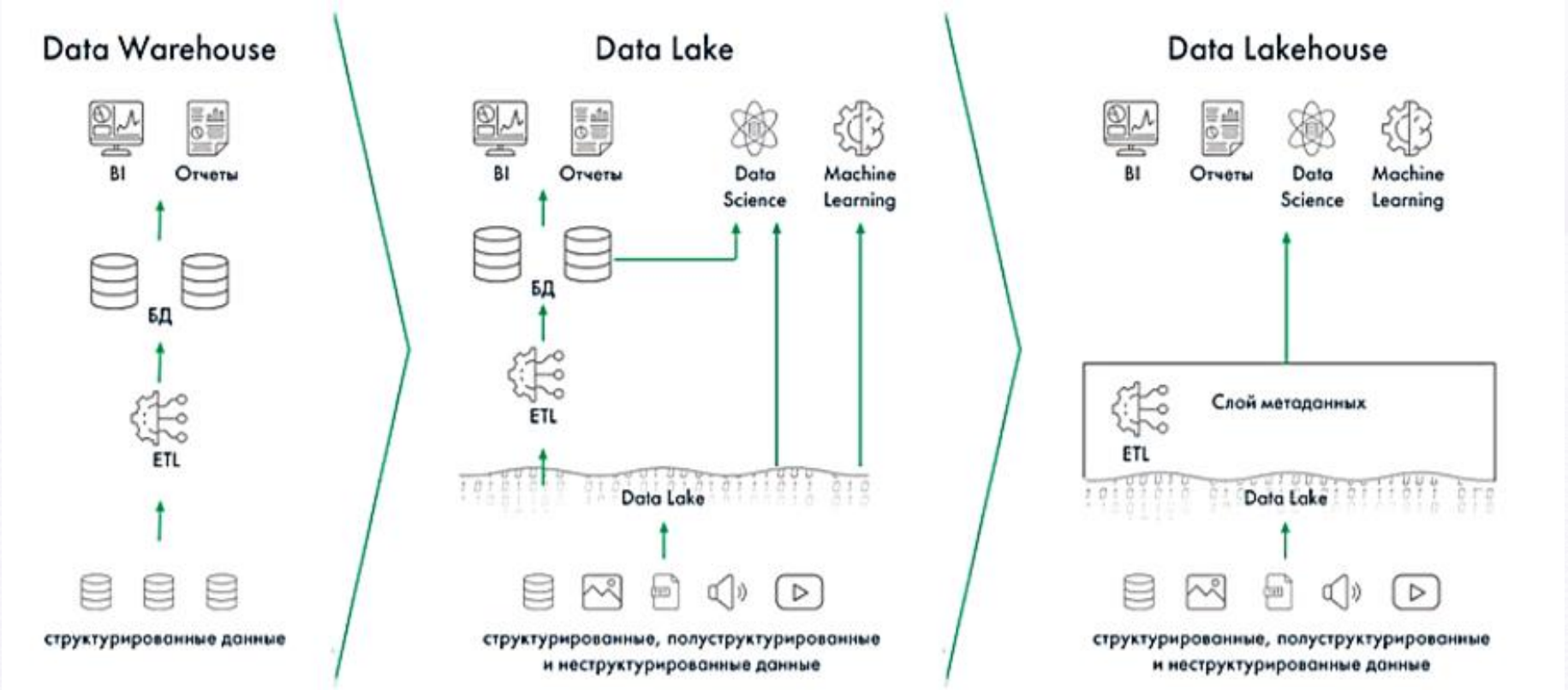
# Как совместить DWH, Data Lake и Data Lakehouse воедино или теперь нам нужен третий КХД?

Data Under Control



# 2020 – настоящее время

Такая концепция позволяет **использовать только один репозиторий, где не только располагаются все типы данных**: структурированные, полуструктурированные и неструктурированные — **но и выполняются все запросы и отчеты**.



Data Under Control

# Data Lakehouse – выгоды

- Упрощение хранения, управления и анализа данных
- Уход от дублирования данных и лишних инфопотоков

- Удешевление стоимости поддержки в силу упрощения арх-ры и простаивающего «железа»
- Эффективное использование вычислительных ресурсов – Cloud Native

- Единая точка правды и доступа для задач ИИ, аналитиков и построения регламентных отчетов
- Доступ к данным из разных движков

Data Under Control

# Переход к Data Lakehouse — как не запутаться в сравнениях форматов

Data Under Control

# Data Lakehouse – основные типы компонентов

- Интеграционный слой
- **Хранение**
- **Метаданные (формат открытых таблиц)**
- **Формат хранения**
- **ETL-движок**
- Потоковый движок
- **Ad-hoc движок**
- Вспомогательные сервисы: мониторинг, логирование, аудит ИБ

- Авторизация
- DG (Каталогизация, качество, глоссарий)
- Визуализация
- Self-Service
- ML-компоненты (инференс, трекинг экспериментов, хранение признаков)
- Оркестрация
- Шаблонизация ETL
- Слой доступа
- Роль ИИ

Data Under Control

# Data Lakehouse – основные критерии выбора компонентов

- Паттерн загрузки
- Наличие перегрузок исторических данных (апдейты)
- Наличие потовой загрузки и аналитики реального времени
- Конфликты при блокировках
- Необходимость поддержки определенных движков
- Объемы данных

- Эволюция схемы
- Time-travel
- ACID
- Необходимость Self-service
- Преобладающие запросы
- Зрелость инструмента (популярность, релизы, кол-во звезд и планы)

Data Under Control

# Data Lakehouse – технологии для реализации

**Core-часть**

Выбор формата открытых таблиц и хранения:

- Hudi **vs** Iceberg **vs** Deltalake **vs** Paimon
- S3-хранение (Ozone, S3 Minio, S3 Ceph, облака)

**Выбор технологий интеграции данных/шины:**

- Нужен ли брокер? (Kafka **vs** RabbitMQ)
- CDC - Debezium
- Поток - Flink vs Spark Streaming
- Батч – Airbite, NIFI, Python+Airflow
- ESB – Jboss Fuse, ServiceMix (ActiveMQ, Camel, CXF, and Karaf ), NIFI

Data Under Control

# Data Lakehouse – технологии для реализации

часть 2

**ETL-движки для регламентных задач:**
Поток – Flink, Spark Streaming

**Аналитические / Ad-hoc движки:**
Trino, Pinot, StarRock

**Безопасность:**
Авторизация – Apache Ranger (+Solr), Sentry
Аутентификация – интеграция с MS AD / Key Cloak

**Шаблозизация ETL:**
DBT, Meltano

**DG:**
OMD, GE, Atlas

Data
Under
Control

# Начните с минимального!

Data Under Control

# Data Lakehouse – каталоги

**2 типа каталогов:**

- **Каталог управления данными:** информационный, помогает централизованно определять политики управления в различных базах данных и метаданных с возможностью поиска.
- **Каталог объектов базы данных:** операционный, используется непосредственно платформами данных и механизмами запросов для чтения и записи данных – метастор. Каталоги хранят состояние таблиц и обслуживают метаданные.

Каталоги стали критически важными при работе с этими расширенными форматами таблиц

**Hudi:** свой метасервер на базе RocksDB, поддержка синка с HMS
**Delta Lake:** Delta Lake Catalog

| Тип каталога Iceberg | Описание |
|---|---|
| **REST** | Каталог REST — это реализация RESTful спецификации каталога Iceberg. Клиент отправляет запросы REST в каталог на стороне сервера, применяя коммиты и обновляет указатели снимков. |
| **Hive Metastore** | Hive Metastore является неотъемлемой частью многих систем озер данных |
| **JDBC** | Каталог JDBC позволяет запрашивать данные из источников данных с поддержкой JDBC без необходимости их загрузки |
| **Nessie** | Nessie поддерживает все функции, доступные каждому клиенту Iceberg, поскольку он реализован как настраиваемый каталог Iceberg. |
| **lakeFS** | Можно использовать реализацию каталога Iceberg в lakeFS для добавления функций управления версиями данных lakeFS в таблицы Iceberg. Интеграция позволяет запрашивать таблицы Iceberg, используя ссылки lakeFS, включая ветки, теги и хэши коммитов |
| **Polaris** | Полнофункциональный каталог с открытым исходным кодом для Apache Iceberg. Он реализует REST API Iceberg, обеспечивая бесшовную совместимость с разными движками. |

Data Under Control

# Как выбрать верный каталог

Требования к развертыванию — оцените требования к развертыванию.
Нужны ли образы Helm или образы Docker для упрощения управления развертываниями?

Ведение записей — проверьте, содержит ли документация подробные инструкции по использованию каталога. Существуют ли блоги, руководства и примеры использования, которые организация может использовать для начала работы?

Управление и безопасность — проанализируйте характеристики безопасности таблиц. В какой степени правила, разработанные для защиты каталога, применимы к различным инструментам и пользователям?

Масштабируемость — проанализируйте, как каталог обрабатывает распределенную функциональность в ситуациях с крупномасштабными хранилищами данных.

Специальные функции — нужны ли вам какие-либо специальные продукты или услуги?
Например, lakeFS допускает семантику, подобную Git, в таблицах, что позволяет выполнять ветвление, слияние, маркировку, откат изменений и другие функции.

Поддержка каталога REST — хотя все эти каталоги соответствуют определению каталога REST, некоторые операции, такие как создание или регистрация таблицы, могут не поддерживаться.

Необходимость использования каталога для других форматов хранения – например, если будет использован Iceberg-Hudi, то каталоги можно связать через HMS

Data Under Control

# Data Lakehouse – основные тренды

**Унификация аналитических и транзакционных workloads**
**Глубокая интеграция с генеративным ИИ - AI-ассистенты для управления данными**:
Инструменты на базе LLM (например, ChatGPT) будут встроены в платформы Lakehouse для:

- Автоматической классификации данных
- Генерации SQL-запросов на естественном языке
- Оптимизации производительности

**Рост популярности open-source решений и унификация форматов. Борьба за экосистему**:
Apache Iceberg, Apache Hudi и Delta Lake станут де-факто стандартами,
вытесняя проприетарные аналоги

- *Тренд*: Увеличение совместимости между форматами
  (например, UniForm в Delta Lake 2.4 для работы с Iceberg и Hudi)
- *Пример*: Snowflake и BigQuery добавляют нативную поддержку Iceberg

**Усиление роли streaming-аналитики – реальное время как default**
**Фокус на экономическую эффективность. Оптимизация стоимости хранения**:

- Развитие форматов с автоматическим tiering (горячее/холодное хранение)
- Сжатие и дедупликация на уровне метаданных
- *Пример:* Iceberg’s metadata pruning для сокращения затрат на S3

Data Under Control

**Когда, через 6 мес.,
ты выбрал оптимальное решение, а тут появился новый, более современный аналог… Делаем миграцию!**

Data Under Control

# Data Lakehouse – основные тренды

часть 2

**Управление данными через метаданные. Data Catalog как центр управления**:

- Интеллектуальные каталоги (например, Databricks Unity Catalog) будут предлагать:
  - Автоматическое тегирование данных
  - Поиск данных через NLP
  - Контроль доступа на основе политик
- *Пример*: Google Dataplex с AI-классификацией

**Multi-cloud и гибридные развертывания. Свобода от вендор-локинга**:

- Лейкхаусы станут агностичными к облачным провайдерам

**Повышение требований к безопасности**

**Edge-аналитика с Lakehouse**

Data Under Control

# Рекомендации по проектированию Data Lakehouse

- Начните с целей и метрикам по направлениям данных/продуктов – от 1 до 5 лет
- Определите возможности размещения инфраструктуры: облако, он-прем или гибрид
- Определите внутренние компетенции и возможности найма специалистов
- Определите паттерны загрузки (много ли апдейтов данных?)
- Есть ли задачи потоковой загрузки данных
- Определите движки обработки
- Выберите хранение и формат хранения, соответствующий каталог
- Объемы данных, SLA на загрузки данных
- Кол-во пользователей платформы (DE, Аналитики, бизнес-пользователи, смежные системы)
- Делайте бенчмаркинги и скоринги решений на базе ваших паттернов загрузки и сценариев и официальной документации (много противоречивой информации!)
- Не бойтесь реальных пилотов

Data Under Control

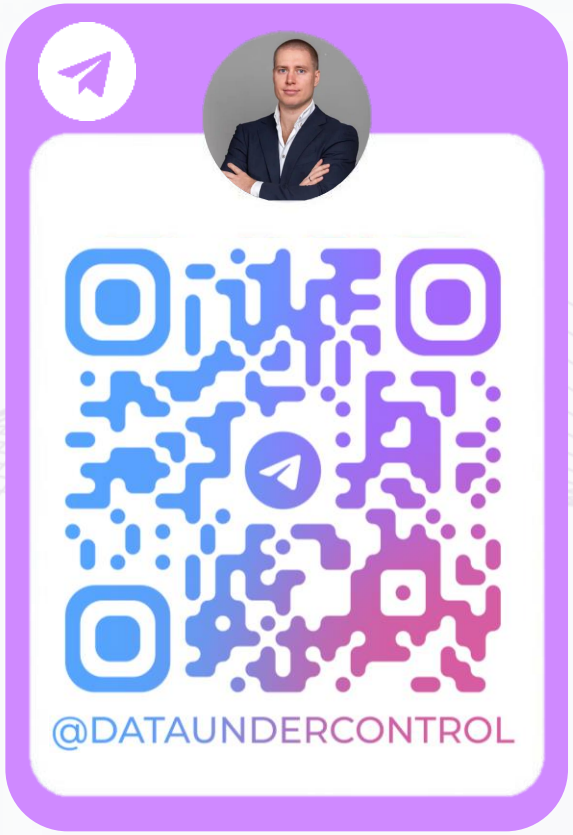
# Открытый микрофон

Если у вас есть вопросы по теме

или вам хочется уточнить какие-то моменты,

сейчас самое время!

Data Under Control




Авторский канал