State of Data

результаты второго ежегодного опроса

Олег Кочергин, 2025

Для чего мы это делаем?

Для чего мы это делаем?

- Формируем общую картину состояния отрасли

Для чего мы это делаем?

- Формируем общую картину состояния отрасли
- Отслеживаем эволюцию подходов и технологий

Для чего мы это делаем?

- Формируем общую картину состояния отрасли
- Отслеживаем эволюцию подходов и технологий
- Даем возможность сравнить себя, команду и компанию с рынком

Для чего мы это делаем?

- Формируем общую картину состояния отрасли
- Отслеживаем эволюцию подходов и технологий
- Даем возможность сравнить себя, команду и компанию с рынком
- Увидеть популярные(и не очень) технологии и инструменты

Для чего мы это делаем?

- Формируем общую картину состояния отрасли
- Отслеживаем эволюцию подходов и технологий
- Даем возможность сравнить себя, команду и компанию с рынком
- Увидеть популярные(и не очень) технологии и инструменты
- Определить тренды для собственного обучения

Про что опрос

- Маленький блок соцдема (единственный с обязательными вопросами)
- Пачка вопросов вокруг тулинга и прочего касательно DE
- Пачка вопросов вокруг тулинга и прочего касательно BI
- Пачка вопросов вокруг тулинга и прочего касательно MLOps

170+ человек

прошло опрос

56%
дата инженеров

vs 64% в прошлом году

24%
дата аналитиков

vs 21% в прошлом году

10% руководителей

vs 6% в прошлом году

3%
DS / ML

vs 4% в прошлом году

Industry vs data team size

	до 10 человек	10-50 человек	50-100 человек	100+ человек
Fintech / Финансы	2.94%	14.71%	11.76%	70.59%
E-commerce / Ритейл	12.12%	21.21%	15.15%	51.52%
IT / Big Tech	11.76%	23.53%	5.88%	58.82%
EdTech	66.67%	16.67%	16.67%	0.00%
Прочее	23.81%	36.51%	17.46%	20.63%

Org size vs data team size

	до 10 человек	10-50 человек	50-100 человек	100+ человек
10-25 сотрудников	100.00%	0.00%	0.00%	0.00%
25-100 сотрудников	44.44%	55.56%	0.00%	0.00%
100-500 сотрудников	44.00%	44.00%	12.00%	0.00%
500-1000 сотрудников	16.67%	43.33%	33.33%	6.67%
1000+ сотрудников	5.68%	12.50%	10.23%	70.45%

Data team setup vs data team size

	до 10 человек	10-50 человек	50-100 человек	100+ человек
Централизованная (все в одном департаменте/отделе)	28.57%	45.24%	11.90%	11.90%
Децентрализованная (специалисты встроены в продуктовые/бизнес-команды)	14.29%	19.05%	14.29%	52.38%
Гибридная (есть центральная платформа/центр компетенций и есть спецы внутри продукта/бизнес-линий)	12.36%	17.98%	15.73%	53.93%

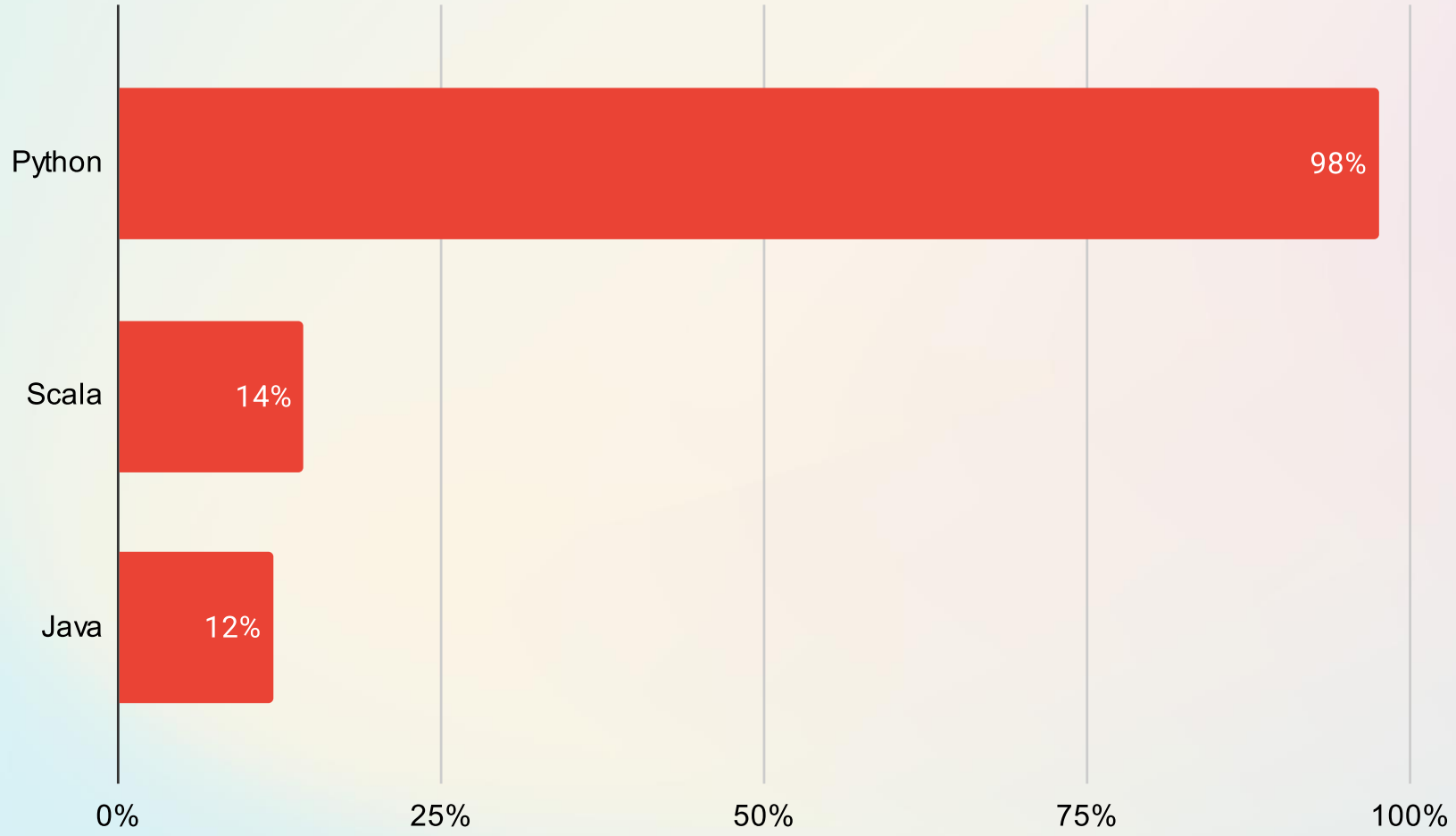
Buy vs Build

	Ванильный Open Source	Свой форк Open Source	Вендорское решение
Оркестрация и трансформация (ETL/ELT)	60.69%	28.97%	10.34%
ML платформы	55.74%	30.33%	13.93%
Хранилища данных (DWH/Data Lake)	47.26%	23.97%	28.77%
Потоковая обработка (Kafka, Flink)	72.18%	14.29%	13.53%
BI и визуализация	33.79%	22.07%	44.14%

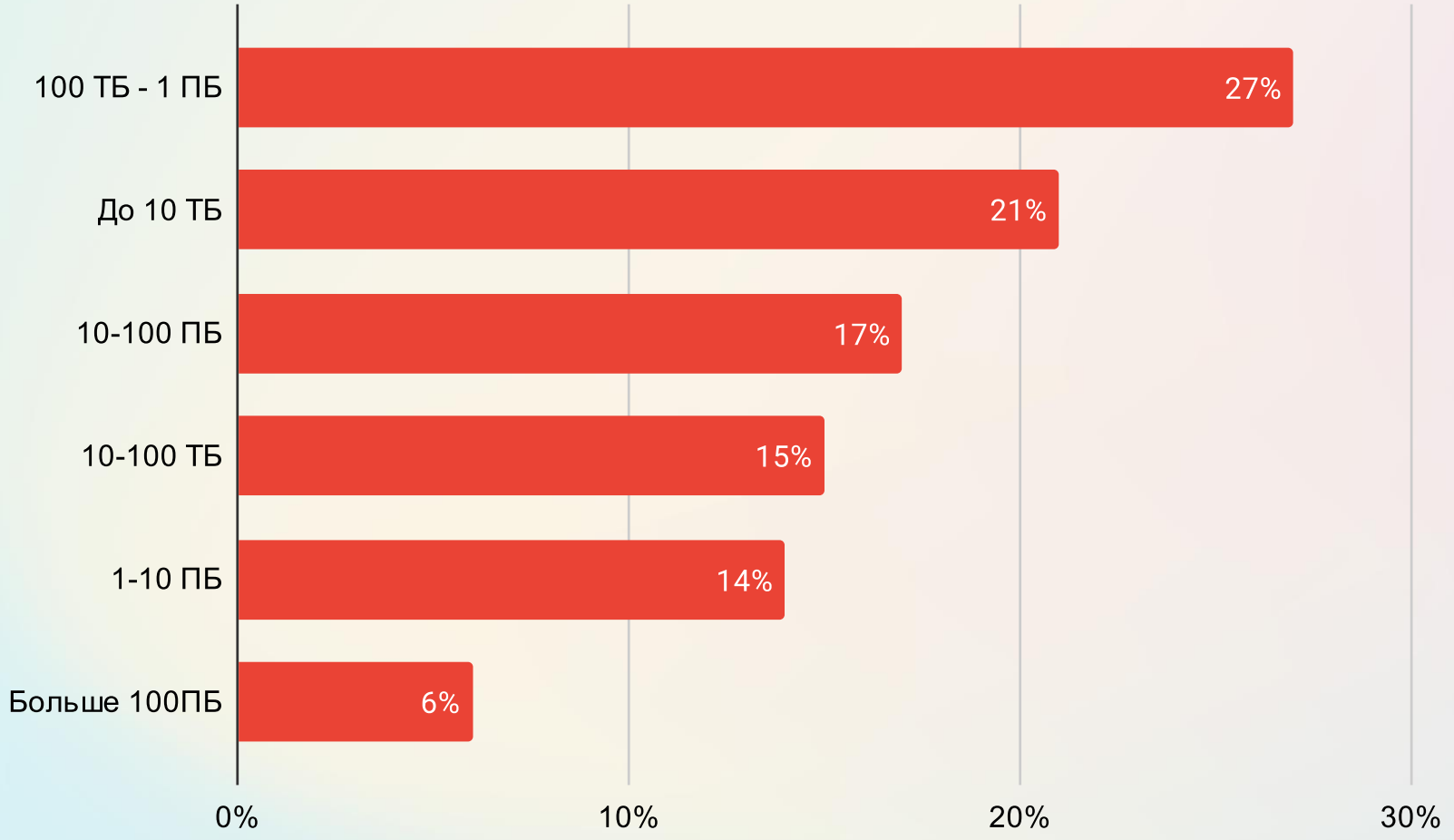
Основной ЯП дата-инженеров

 Scala?

Основной ЯП дата-инженеров

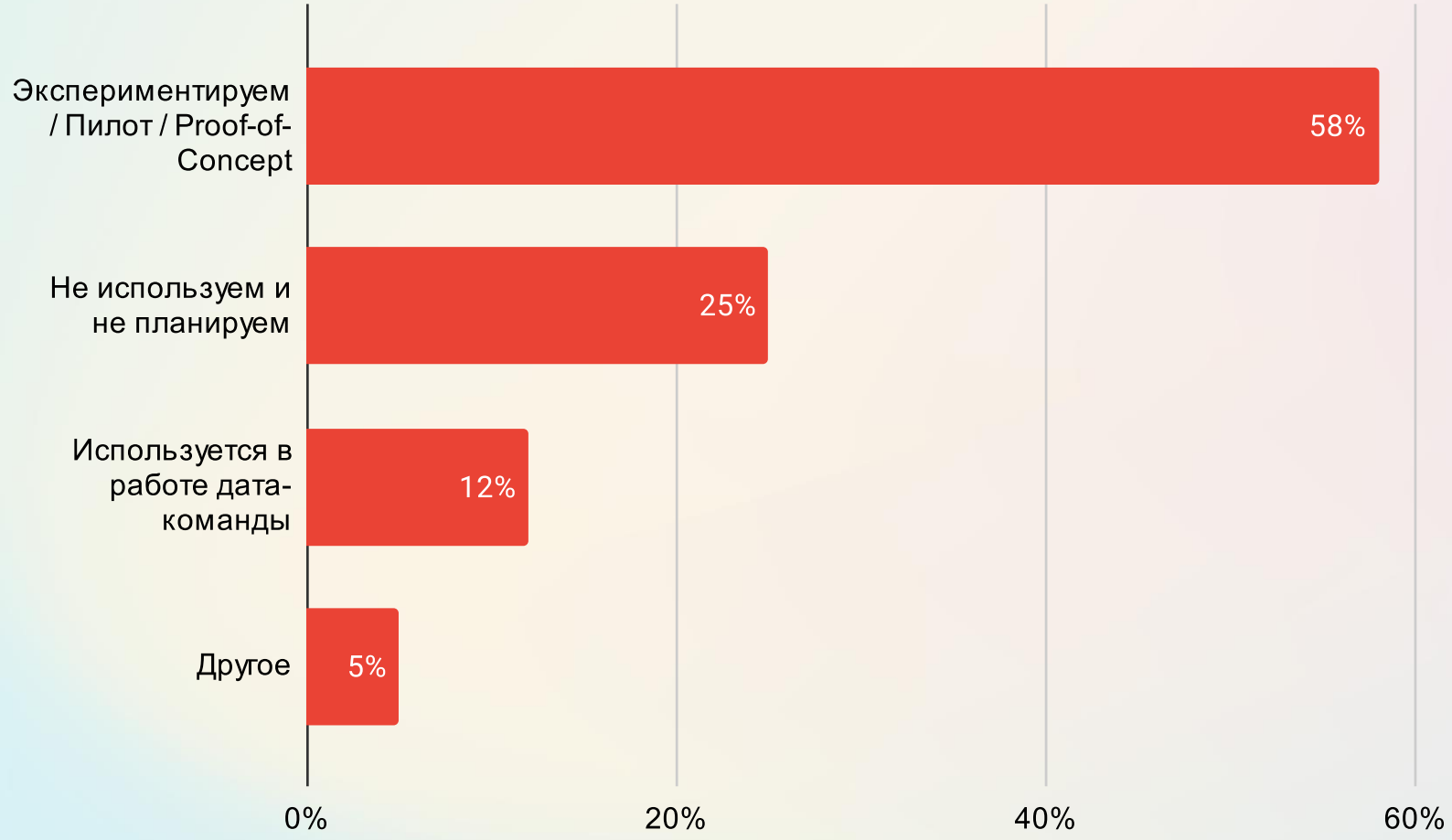


Насколько ваша бигдата большая?



Text2SQL anyone?

Text2SQL anyone?



Gainers & Loosers

- Будем сравнивать сумму используем+планируем 2024 vs 2025

Основной формат хранения данных

	Даже не слышал о таком	Не интересует	Использовали в прошлом	Планируем попробовать	Используем
Parquet	○	○	○	○	○
ORC	○	○	○	○	○
Hudi	○	○	○	○	○
Iceberg	○	○	○	○	○
Delta Lake	○	○	○	○	○
Avro	○	○	○	○	○
Json	○	○	○	○	○

Основной стек для хранения и обработки данных

Presto/Trino + K8s + S3

Hadoop

Greenplum

Starrocks

Основной стек для хранения и обработки данных

Presto/Trino + K8s + S3
+17%

Hadoop

Greenplum

Starrocks

Основной стек для хранения и обработки данных

Presto/Trino + K8s + S3
+17%

Hadoop
-4.3%

Greenplum

Starrocks

Основной стек для хранения и обработки данных

Presto/Trino + K8s + S3
+17%

Hadoop
-4.3%

Greenplum

Starrocks
+14%

Основной стек для хранения и обработки данных

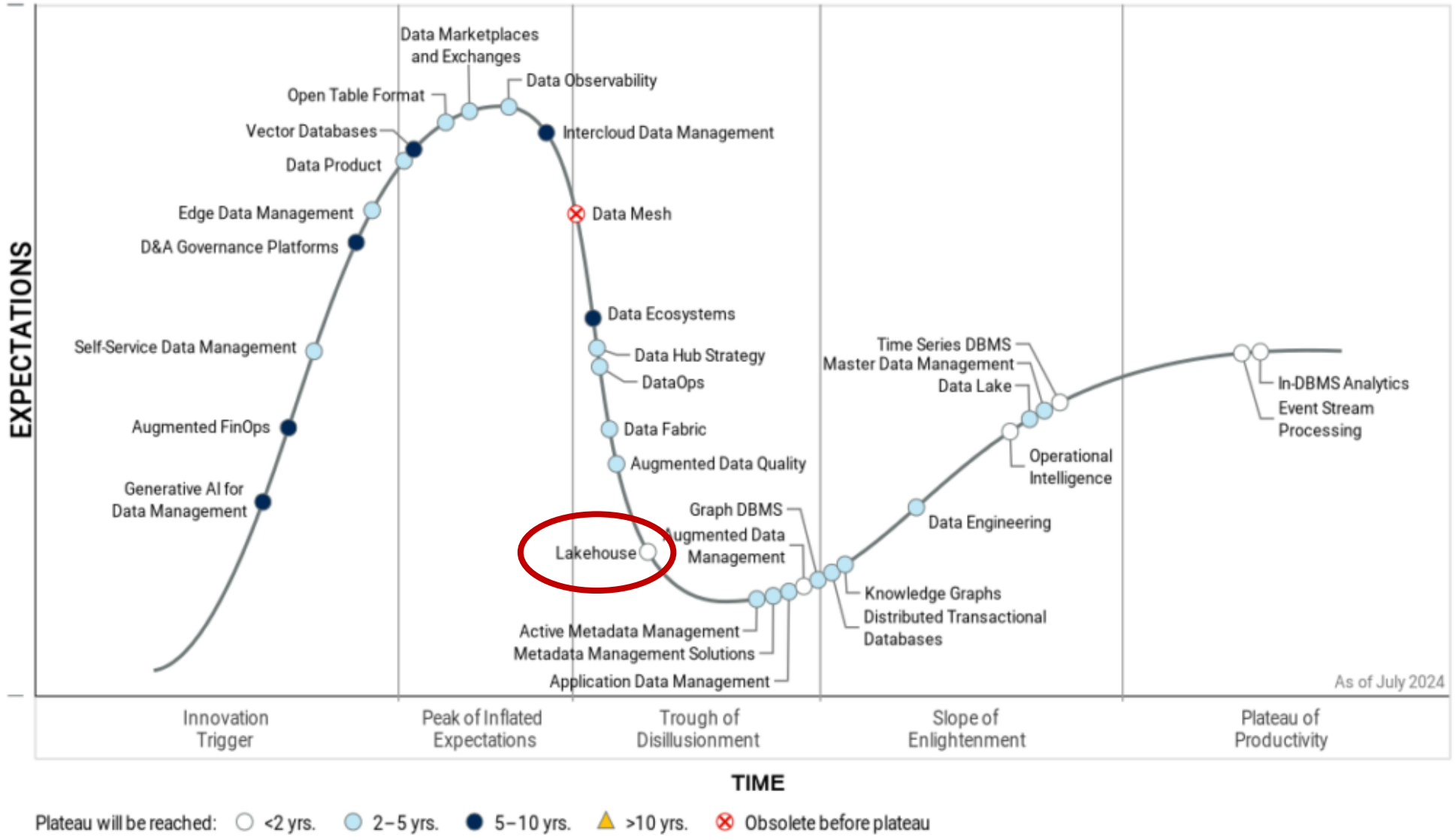
Presto/Trino + K8s + S3
+17%

Hadoop
-4.3%

Greenplum
-4%

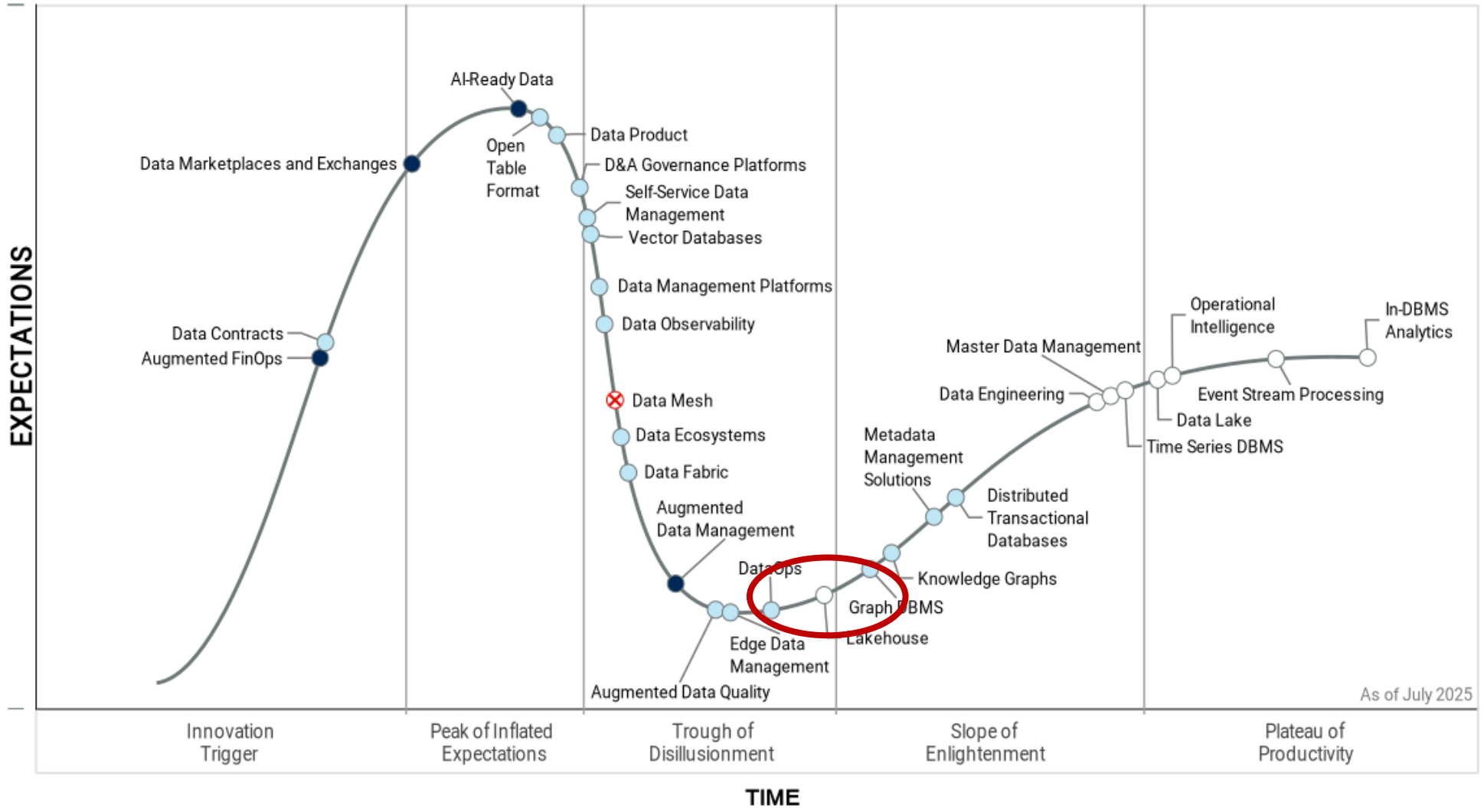
Starrocks
+14%

Hype Cycle for Data Management, 2024



Gartner

Hype Cycle for Data Management, 2025



Plateau will be reached: ○ <2 yrs. ● 2–5 yrs. ● 5–10 yrs. ▲ >10 yrs. ⊗ Obsolete before plateau

Gartner

Доклад

Как мы строили lakehouse на Ozone

Как мы переходили с платформы на базе Vertica, HDFS к новой архитектуре Dota 2 (второй версии нашей внутренней аналитической платформы), основанной на Apache Ozone (S3), Trino, Spark и Iceberg. Поделюсь опытом выбора хранилища, объясню, почему отказались от HDFS и почему выбрали Apache Ozone как on-prem реализацию S3.

Виталий Моисеев
Островок!

RU

Architecture of Data Platforms

Доклад

StarRocks — реальность современной платформы данных

Платформа данных в нашей компании существует уже более 5 лет, за это время она вобрала множество модных (и не очень) решений. Расскажу, как мы пытались выбрать наше будущее среди ClickHouse, Greenplum и Trino, а нашли StarRocks.

Станислав Лысиков

RU

Data Tools

Доклад

Hadoop Is Not Dead — Just Secure!

История о том, как небольшая команда инженеров внедрила Hadoop с полноценной безопасностью на базе Kerberos и Ranger без остановки бизнес-процессов.

Антон Александров
Детский мир

RU

Use Cases

Обсуждение 18:45 – 19:45

Круглый стол «Hadoop мертв, да здравствует Hadoop?!»

Михаил Марюфич Т-Банк
Алексей Белозерский VK Tech, VK Cloud
Виталий Моисеев Островок!
Игорь Дмитриев Wildberries & Russ
Дмитрий Зуев Positive Technologies

Зал 2 RU Offline only ~~REC~~

Основной формат хранения данных

Iceberg

Parquet

Hudi

Json

Основной формат хранения данных

Iceberg
+30.7%

Parquet

Hudi

Json

Основной формат хранения данных

Iceberg
+30.7%

Parquet

Hudi
-5.7%

Json

Основной формат хранения данных

Iceberg
+30.7%

Parquet
+9.8

Hudi
-5.7%

Json

Основной формат хранения данных

Iceberg
+30.7%

Parquet
+9.8

Hudi
-5.7%

Json
-4.5%

Доклад

Перспективы развития Apache Iceberg

Обсудим ключевые вызовы, стоящие перед Apache Iceberg, а также перспективы развития технологии.

Владимир Озеров
CedrusData

RU

Data Tools

Инструменты загрузки данных

Apache Flink

Debezium

Airbyte

Apache NiFi

Инструменты загрузки данных

Apache Flink
+22%

Debezium

Airbyte

Apache NiFi

Инструменты загрузки данных

Apache Flink
+22%

Debezium
+20.6%

Airbyte

Apache NiFi

Инструменты загрузки данных

Apache Flink
+22%

Debezium
+20.6%

Airbyte

Apache NiFi
-11.7%

Инструменты загрузки данных

Apache Flink
+22%

Debezium
+20.6%

Airbyte
+10%

Apache NiFi
-11.7%

Доклад

Debezium и PostgreSQL после happy-path: какие проблемы ждут в проде и как их решать

Получение событий изменений из источников — довольно частая задача, которая может решаться разными способами. Одно из таких решений — Debezium. Но так ли с ним все просто и всегда ли именно это решение наилучшее? Постараюсь ответить на эти вопросы и рассмотреть Debezium с точки зрения сложностей, которые возникают на пути решения задачи захвата изменений.

Никита Рьянов

RU

Data Tools

Новинки в инструментах загрузки

- dlt (dlthub.com)
- Apache Seatunnel
- Apache Hop
- Bruin (getbruin.com)

Инструменты трансформации данных

Apache Flink

DuckDB

Pandas

Polars

Инструменты трансформации данных

Apache Flink
+22%

DuckDB

Pandas

Polars

Инструменты трансформации данных

Apache Flink
+22%

DuckDB
+19%

Pandas

Polars

Инструменты трансформации данных

Apache Flink
+22%

DuckDB
+19%

Pandas

Polars
+19%

Инструменты трансформации данных

Apache Flink
+22%

DuckDB
+19%

Pandas

Polars
+19%

Инструменты трансформации данных

Apache Flink
+22%

DuckDB
+19%

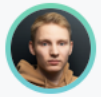
Pandas
-9.5%

Polars
+19%

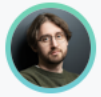
Доклад

От бакета в S3 к Data Lakehouse: эволюция платформы данных в гонке за автономией

Как Data Lakehouse стал нашим спасательным кругом: обеспечил безболезненную миграцию при непрерывном потоке более 150 ТБ в день.



Никита Бандурко
Navio



Георгий Попов
Navio

RU

Architecture of Data Platforms

Оркестрация и DQ

Kestra
+6%

NiFi
-14%

Soda Core
+12.7%

Свой DQ велосипед
66%

Доклад

Как мы улучшили процессы по работе с данными в Airflow: практические кейсы

Расскажу, как мы используем Airflow на практике: от болей с sensors до удобства с datasets, от стандартных возможностей — к собственным кастомным решениям. Доклад не оставит равнодушными тех, кто сталкивался с реальной эксплуатацией Airflow.

Дмитрий Морозов
Инновационный центр «Безопасный транспорт»

RU 

Use Cases

Доклад

Как мы искали инструменты для DQ и к чему в итоге пришли

Обзор и сравнение существующих библиотек Python и самописного инструмента профилирования для анализа качества данных. Описание функционала инструмента.

Павел Павлюков
Газпромбанк.Тех

Александр Свяжин
Газпромбанк.Тех

RU 

DQ

Доклад

Data Quality as a Service — инструмент самообслуживания в большой компании

Как внедрить Data Quality-инструмент с распределенной архитектурой, обеспечивающий бесперебойную работу для большого числа команд и являющийся единой точкой правды о качестве данных в системах компании.

Андрей Азеев
MTC Web Services

Богдан Петров
MTC Web Services

RU 

DQ

Доклад

Мониторинг DWH: от метаданных до DataOps

Практический кейс внедрения мониторинга DWH от Skyeng: от архитектуры метаданных до автоматизированных проверок качества данных и перехода к DataOps-практикам.

Данил Захаров
Skyeng

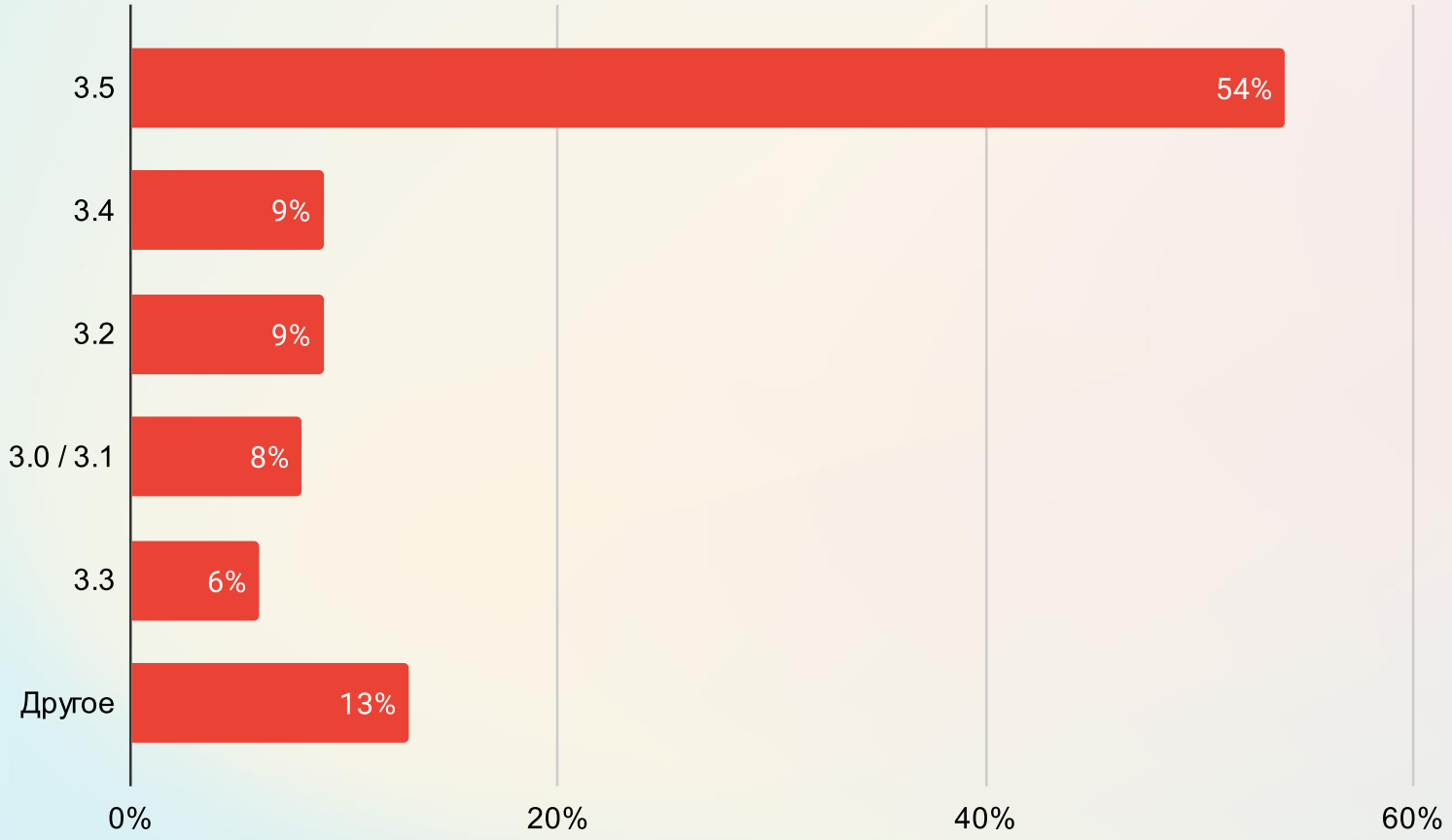
RU 

Data Management

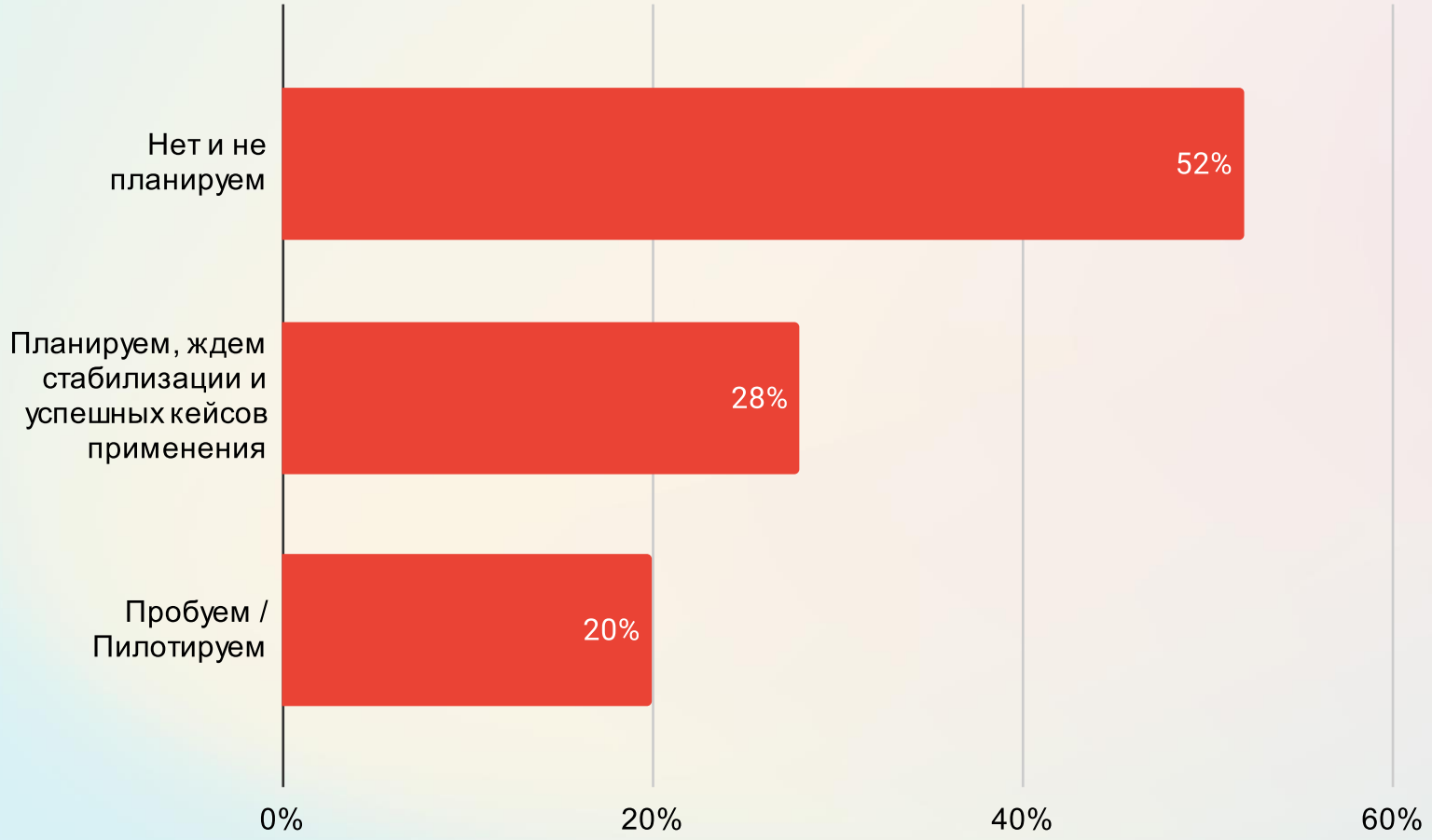
Давайте поговорим про Spark

- Используемые версии
- Использование native execution акселераторов

Apache Spark: какая основная версия используется?



Используете ли вы native execution акселераторы?



Доклад

Spark — ВСЁ!

Поговорим про Spark. Что он дал дата-инженерам? Почему многие из нас используют именно его?

Spark уже более 15 лет. С какими проблемами мы сталкиваемся при его использовании? Появилось ли что-то получше? Можно ли его уже чем-то заменить?

Почему %SQLEngineName% тормозит? Как это исправить? Бенчмарки, open source и тому подобное.

Евгений Глотов
Navio

RU

Data Tools

Доклад

Сторонние движки исполнения для Apache Spark: опыт использования

Опыт применения движков исполнения Comet и Gluten (Velox) — от введения и особенностей сборки до результатов тестирования на реальных ETL. Расскажу о подводных камнях и неочевидных моментах, покажу результаты работы и рассмотрю кейсы, когда эти движки полезны, а когда вообще не работают.

Никита Благодарный
Честный знак

RU

Data Tools

Доклад

Spark Connect: новый подход для работы с Apache Spark

Расскажу о Spark Connect — новом подходе для работы с Apache Spark, который позволяет разрабатывать клиентскую часть приложения на любом языке и не зависеть от JVM. Поговорим об архитектуре Spark Connect и ее отличиях от классического Spark. Вы узнаете о проекте, в котором мы использовали Spark Connect API для C++.

Александр Токарев
Яндекс

RU

Data Tools

Доклад

Apache Spark SQL. Расширяй и управляй

Как настроить и доработать Apache Spark под свои задачи без переписывания фреймворка. Расскажу о подходах к расширению функциональности Spark SQL без вмешательства в исходный код платформы. Вы узнаете о создании собственных источников данных, разработке пользовательских функций для специализированной обработки и внедрении правил оптимизации, адаптирующихся под различные запросы.

Дмитрий Вертлиб
Честный знак

RU

Data Tools

Технические вызовы прямо сейчас?

- Миграция и модернизация стека
- Инфраструктура, производительность и масштабирование
- Внедрение новых технологий и архитектурные задачи
- Качество данных, управление и процессы (Data Governance)

Миграция и модернизация стека

Отказ от Legacy

Переход на Lakehouse

Импортозамещение
и Open Source

Обновление инструментов

Инфраструктура, производительность и масштабирование

Нехватка вычислительных ресурсов

Дорогие облака

Эффективное масштабирование

Внедрение новых технологий и архитектурные задачи

Real-time обработка данных

Внедрение LLM

Построение платформ с нуля

Self-service аналитика

Качество данных, управление и процессы

Качество данных

Data Governance

Тестирование
и документация

Интеграция и источники

Бонус!

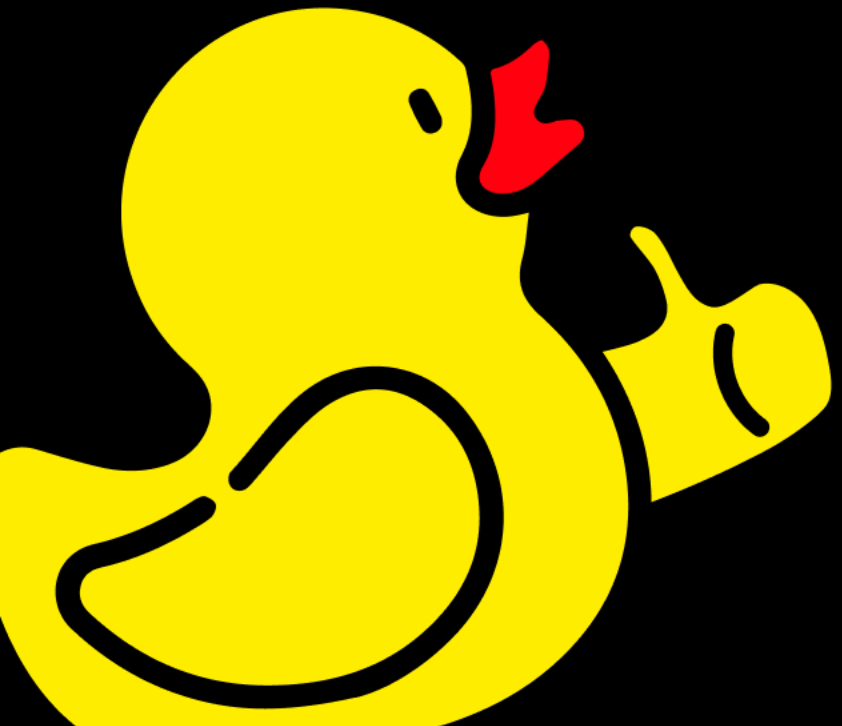
Сырые данные опросов
2024 и 2025

Спасибо

и до встречи в следущем году

↑ опрос проходим кто не прошел

Оценивайте выступление, переходите в чат доклада,
скачивайте презентацию — по QR-коду



11:15–12:00

Зал 1

Перспективы развития Apache Iceberg

Владимир Озеров CedrusData

Оценить выступление

Обсудить / задать вопрос

Смотреть презентацию

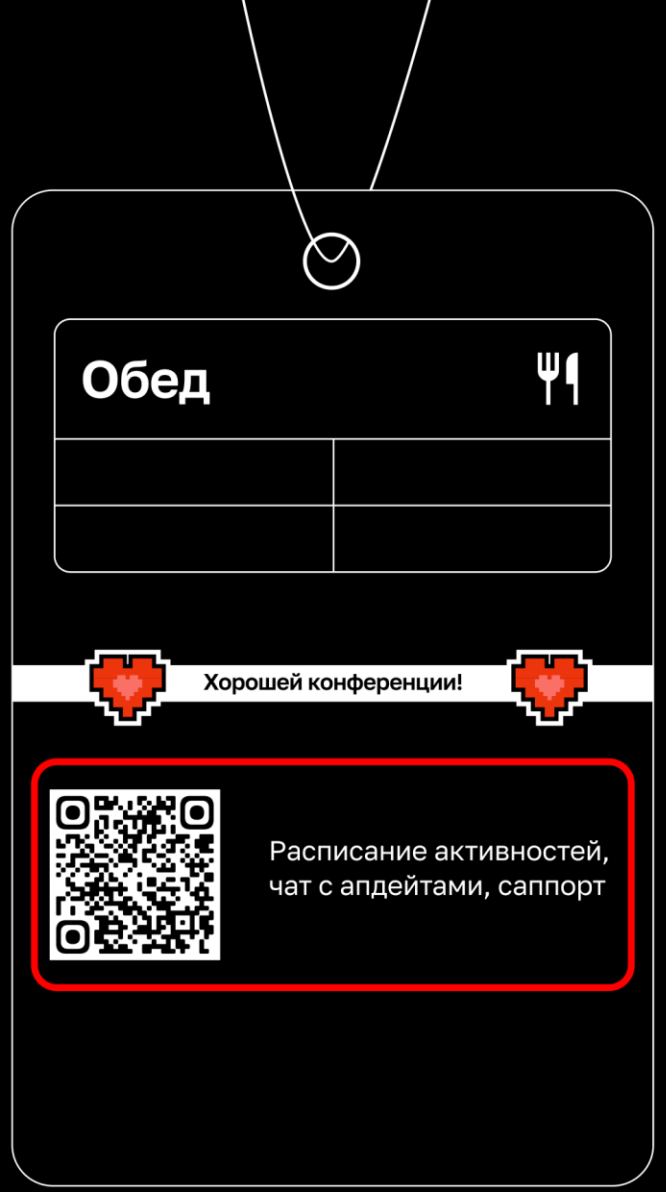
QR-Код да Винчи

#smartdata

- **Фотографии**
- **Чат конференции** Telegram
- **Расписание** на сайте
- **Расписание + схема площадки** PDF
- **Эксперты** на стендах
- **Квест по выставке**
- **Помощь / вопрос** Telegram

Также вы можете задать вопрос на стойке регистрации.

Для комфорта участников конференции действует Code of Conduct.




#smartdata
18:15
Вечеринка

#smartdata

18:45 Зал 2

Круглый стол «Hadoop мертв, да здравствует Hadoop?!»

#smartdata

19:00 Зал 3

Lightning Talks

#smartdata

Партнеры

Ведущий партнер

MAGNIT TECH

Ведущий партнер

cedrusdata

Ведущий партнер

X5 Tech

Ведущий партнер

tech

Партнер

КРИПТОНИТ

Партнер

mindbox

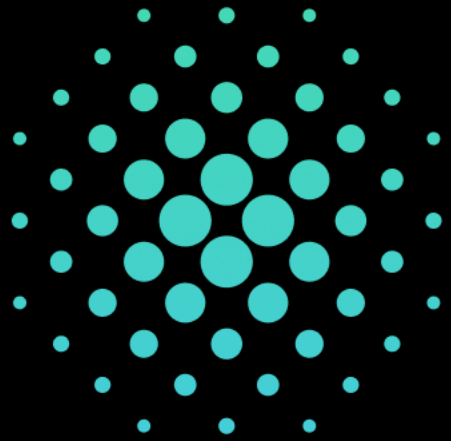
#smartdata

Квест по выставке

Среди всех, кто соберет 4 чекина и оставит 4 комментария
с обратной связью мы выберем рандомно 1 победителя
и подарим ему сумку с мерчем от партнеров и билет
на следующую конференцию

Support

#smartdata



SmartData

2025