



Databorn

Как выбрать Lakehouse-платформу и не ошибиться



Евгений Вилков

Директор по развитию продукта
Alphyn Lakehouse



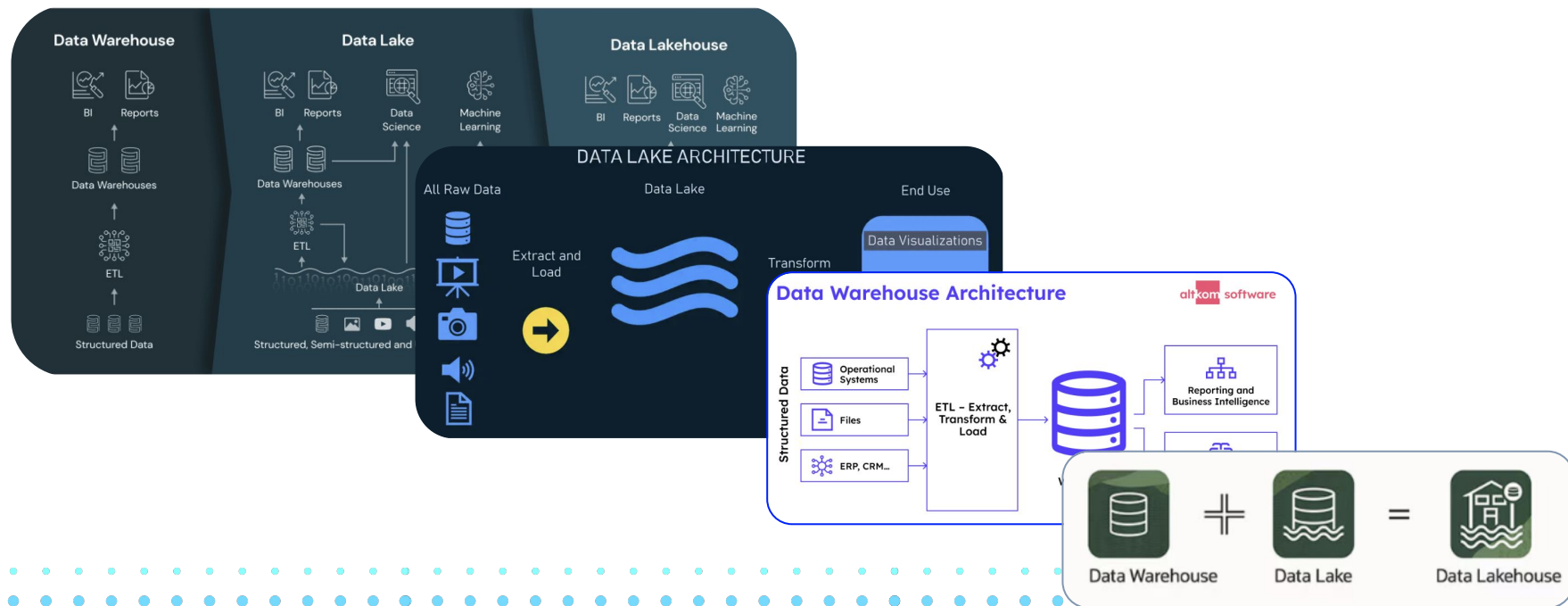
ALPHYN

- **23** года работы с СУБД, MPP, Big Data
- Опыт проектирования и разработки аналитических ХД с 2007 г.
- **50+** проектов в области ХД, интеграции данных, аналитических систем (финтех, телеком, ритейл, производство, от 1 Тб до 5 Пб)

Экспертный проектный
практический опыт

Greenplum • Vertica • SAP IQ • Netezza • Exadata • Teradata
Hadoop • Clickhouse • Cloud • Oracle • MS SQL Server • Postgres

Вам еще не надоело смотреть на эти схемы и изображения в каждом докладе?



Ожидание от универсальной современной платформы данных

Функциональные требования

Замена привычных реляционных СУБД, в том числе MPP:

- Поддержка ANSI SQL и всех DML-операторов данных, включая **Select Update Delete Merge** для простой миграции функционала;
- Желательно поддержка процедурных расширений;

Реализация функции озера данных и ML-лаборатории:

- Spark как главный DS-фреймворк;
- Поддержка неструктурированных данных;

Обработка данных в реальном времени.

Технические требования

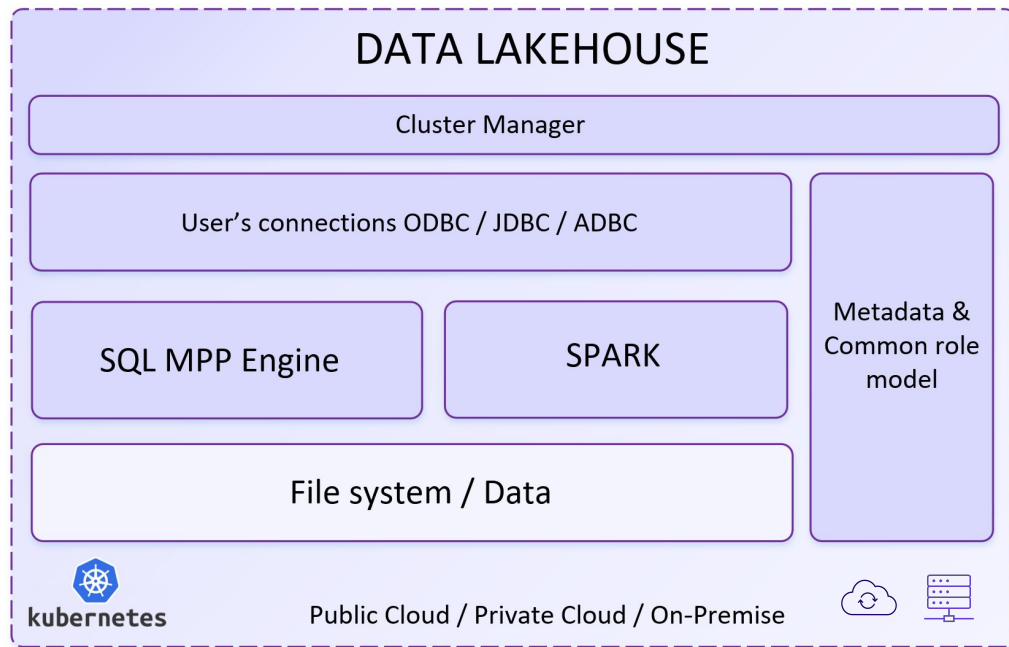
Архитектура по принципу Composable Codex:

- Ориентирование на открытые форматы данных Iceberg, Paimon, Parquet, ORC, а не бинарные форматы;
- Cloud-ready архитектура, в том числе при установке on-prem:
 - Kubernetes deployment;
 - Разделение Storage и Compute;

Высокая производительность.



Верхнеуровневая функциональная архитектура Lakehouse-платформы



Объектное хранилище S3



SQL MPP engine



Spark



Полная поддержка Iceberg table format



Поддержка стандартных подключений ODBC, ADDBC, JDBC



Общие метаданные для всех компонент



Единая ролевая модель



Cluster manager как центр управления

Предложения на рынке Lakehouse-платформ



dremio



Starburst



ONEHOUSE



snowflake



CelerData



А где же Spark?



databricks

А как насчет
on-prem или
private cloud?

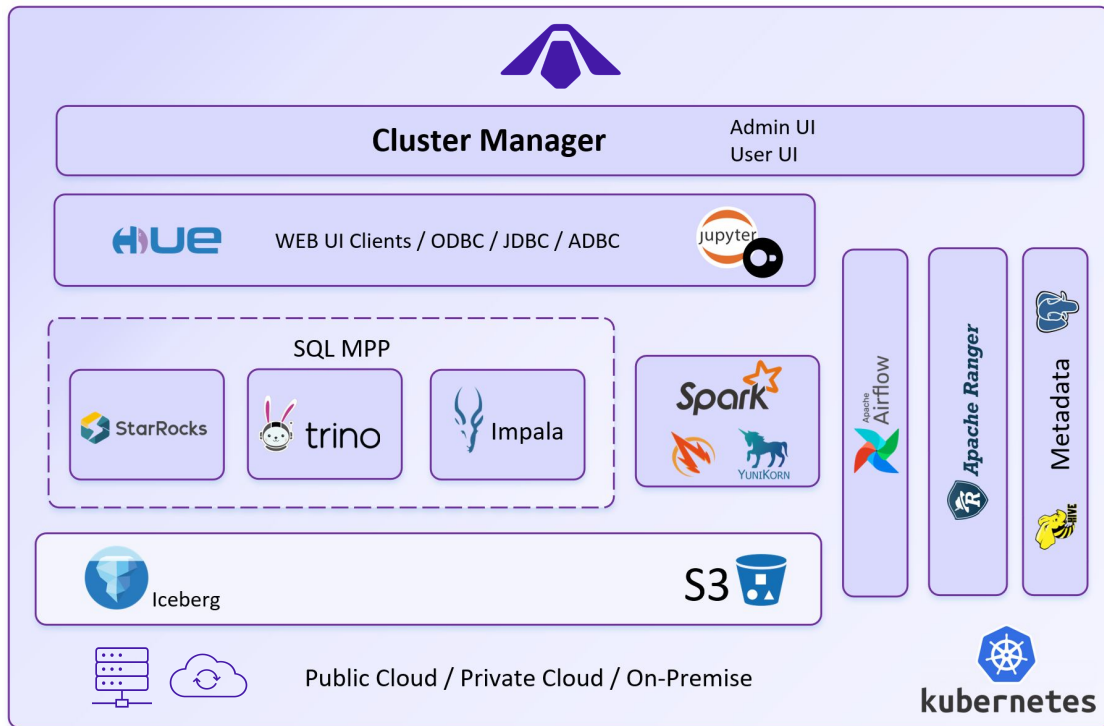
CLUDERA

Компромиссный вариант, в котором есть
и MPP SQL и Spark, но:

- часто меняющаяся ценовая политика;
- неудачное решение в части S3 Storage.

Есть ли
альтернатива?

Alphyn Lakehouse



On-premise or multi-tenant private / public cloud



Раздельные хранение и вычисления



Информационная безопасность enterprise-уровня



Технологии вычислений для любых задач: **Spark, Starocks, Trino, Impala, DuckDB**



Поддержка открытых табличных форматов



Независимое масштабирование



**Как же выбрать SQL-движок?
Может, взять самый быстрый?**

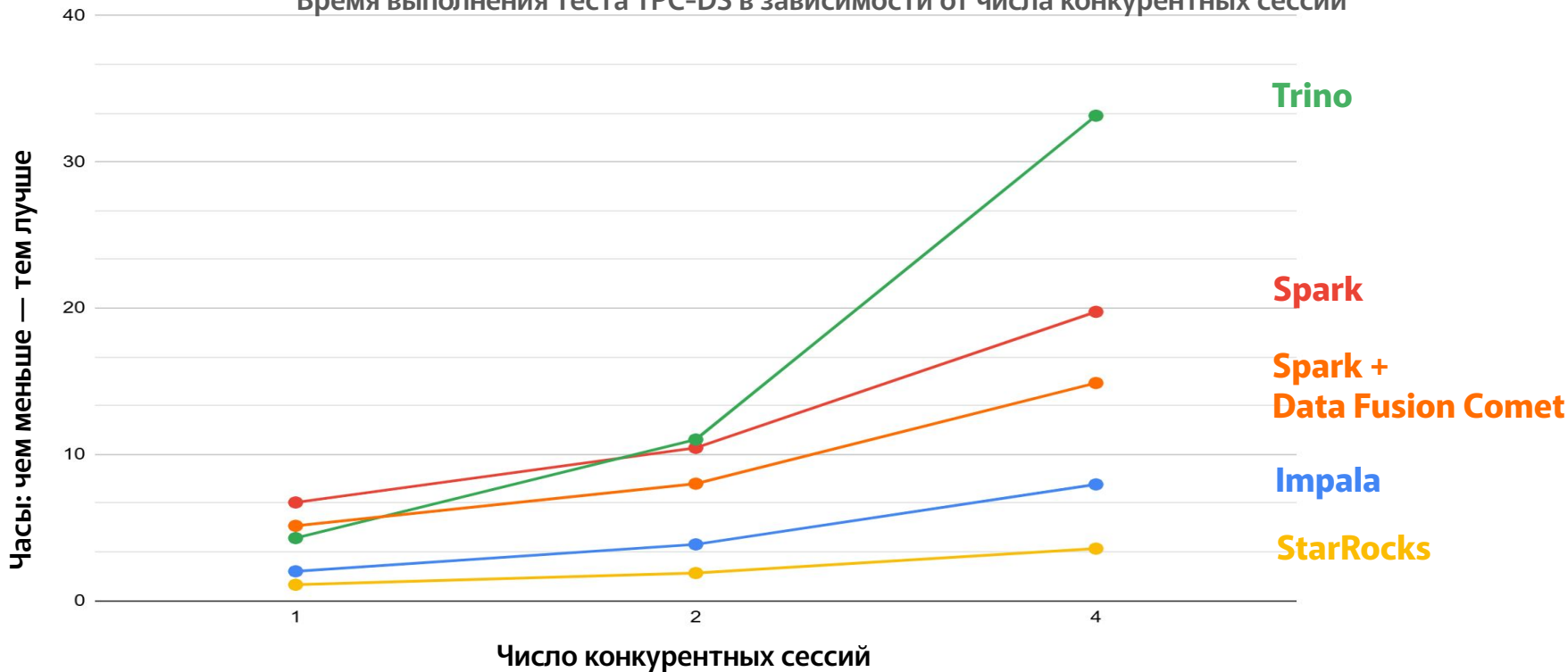
Сравнение производительности движков Alphyn Lakehouse по методике TPC-DS

1

2

3

Время выполнения теста TPC-DS в зависимости от числа конкурентных сессий



Сравнение производительности движков по методике TPC-DS

1

2

3

Конфигурация тестового стенда:

- Public cloud;
- S3 Cloud Object Storage;
- VM для размещения compute-мощностей движков.

Проведение испытаний

- Установлена платформа Alphyn Lakehouse;
- Испытания проводятся со всеми движками по очереди;
- Для каждого движка предварительно выполняется сбор статистики, оптимизация настроек, создание и тюнинг ресурсных очередей для конкурентной работы (workload management);
- Запросы не изменяются, согласно методике;
- Тест считается пройденным, если все запросы завершились без ошибок и падений. Если падения происходили, то тестирование уходило на новую итерацию настроек и запуска.

Сравнение производительности движков по методике TPC-DS

1

2

3

Движок	Impala	StarRocks	Trino	Spark	Spark + DF Comet
Версия движка	4.5 20250429-4.5.0-release	3.4 -20250429	459-20241228	3.5.4	3.5.4 + 0.9.0
Кол-во worker-узлов	4	4	4	4	4
Кол-во vCores на worker-узел	32	32	32	32	32
Оперативная память на worker, Гб	252	270	280	300	300
Всего vCores в кластере	128	128	128	128	128
Всего RAM в кластере, Гб	1008	1080	1120	1200	1200

Число конкурентных сессий	Общее число запросов	Время работы теста				
		Impala	StarRocks	Trino	Spark	Spark + DF Comet
1	99	2 часа	1 час 6 мин	4 часа 30 мин	6 часов 43 мин	5 часов 7 мин
2	198	3 часа 51 мин	1 час 54 мин	11 часов 2 мин	10 часов 27 мин	8 часов 1 мин
4	396	7 часов 57 мин	3 часа 36 мин	33 часа 10 мин	19 часов 45 мин	14 часов 53 ммин



Выбор очевиден?

Использование движков клиентами Alphyn Lakehouse

Клиент / Движок	Impala	Spark	Trino	StarRocks
Банк				
Банк				
Сеть ресторанов быстрого питания				
Топ-ритейлер				
Топ-ритейлер				
Комплексная программа лояльности (банк, телеком, ритейл и др.)				
Страховая компания				
Алкогольная компания				
Банк				
Банк				
Банк				
Страховая компания				
Крупнейший eCom				
Финорганизация				
Итого	12	13	8	3



Ни один клиент не обходится одной compute-технологией!



Ни один клиент не обходится
одной compute-технологией.
Почему?

Open source compute-движки. Сравнение

Движок / Критерий	Trino	StarRocks	Doris	Impala	Spark
Поддержка Iceberg	Поддержка всех операций	Отсутствие Update Delete, Merge, Alter	Отсутствие Update Delete, Merge	Поддержка всех операций	Поддержка всех операций
Обслуживание объектов Iceberg	Все операции, кроме оптимизации манифестов	Отсутствие	Отсутствие	Все операции, кроме оптимизации манифестов	Полный цикл обслуживания, макс оптимально
Работа с Iceberg Meta	Поддерживает	Поддерживает	Поддерживает	Поддерживает	Поддерживает
MatView	Поддерживает	Поддерживает продвинутое	Поддерживает	Не поддерживает	Не поддерживает
Федеративный доступ	Высокие возможности	Средние возможности	Средние возможности	Средние возможности	Средние возможности
Производительность SQL (относительно друг друга)	Средняя	Максимальная	Максимальная	Максимальная	Высокая
Отказоустойчивость	Низкая	Высокая HA	Высокая HA	Высокая HA	Высокая Макс
Конкурентная нагрузка	Низкая, деградация, требует больше аппаратных ресурсов	Максимальная	Максимальная	Максимальная	Средняя, но хорошая перспектива Native Exec Engine
Поддержка процедур	Нет	Нет в OSS, будет в Alhyn в Q1 2026. Собственная разработка	Да	Нет в OSS, будет в Alhyn в Q1 2026. Собственная разработка	Есть начиная с SPARK 4.0



Невозможно спроектировать универсальную платформу данных, представленную одной технологией



Выводы

- В эффективной Lakehouse-платформе данных невозможно обойтись одной compute-технологией;
- Выбирать необходимо не движок, а платформу, которая:
 - спроектирована по принципам открытости к изменениям;
 - поддерживает **2+** движка;
 - обеспечивает единообразие доступа ко всем данным.

Линейка data-продуктов ALPHYN

Alphyn Lakehouse

Универсальная платформа данных

Сценарии использования

NRT ODS

Data Lake

DWH MPP

ROLAP

Ad-Hoc

Alphyn DR (Data Replicator)

Инструмент для пакетной репликации

Alphyn ASM

Решение для онлайн-загрузки и потоковой обработки данных

Ждем вас на нашем демостенде!



Databorn

Благодарю
за внимание



Евгений Вилков

evgeny.vilkov@databorn.ai

alphyn.ai