Яндекс Go

Платформа управления данными вокруг YTsaurus

Владимир Верстов
Наталья Савенкова

Яндекс Go

Yandex Go

Services for Ridetech, Foodtech,
E-commerce and Delivery

40 mln

MAU of Yandex Go

47 mln

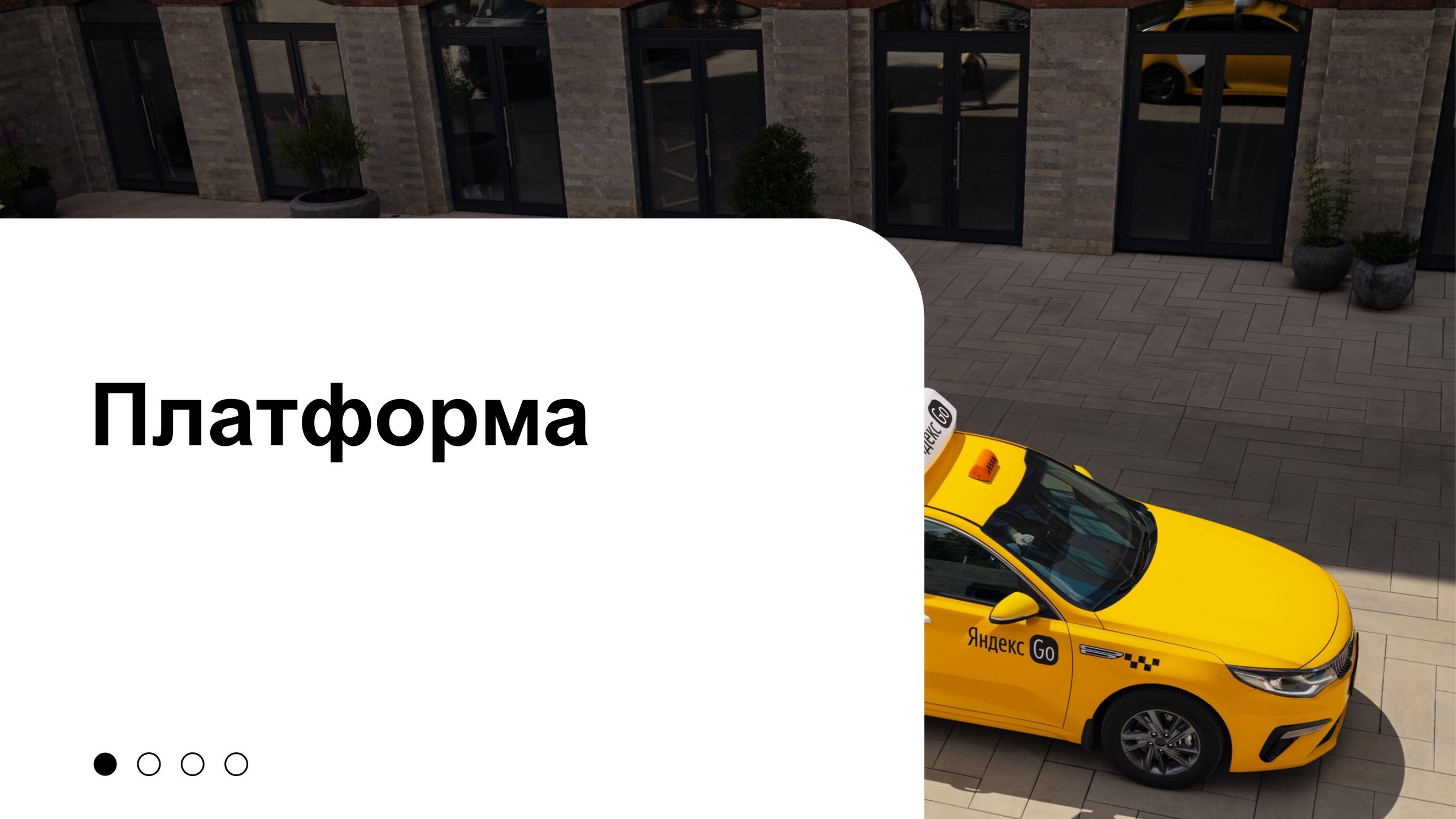
SKU on
MarketPlace

570 k

deliveries
per day

Содержание

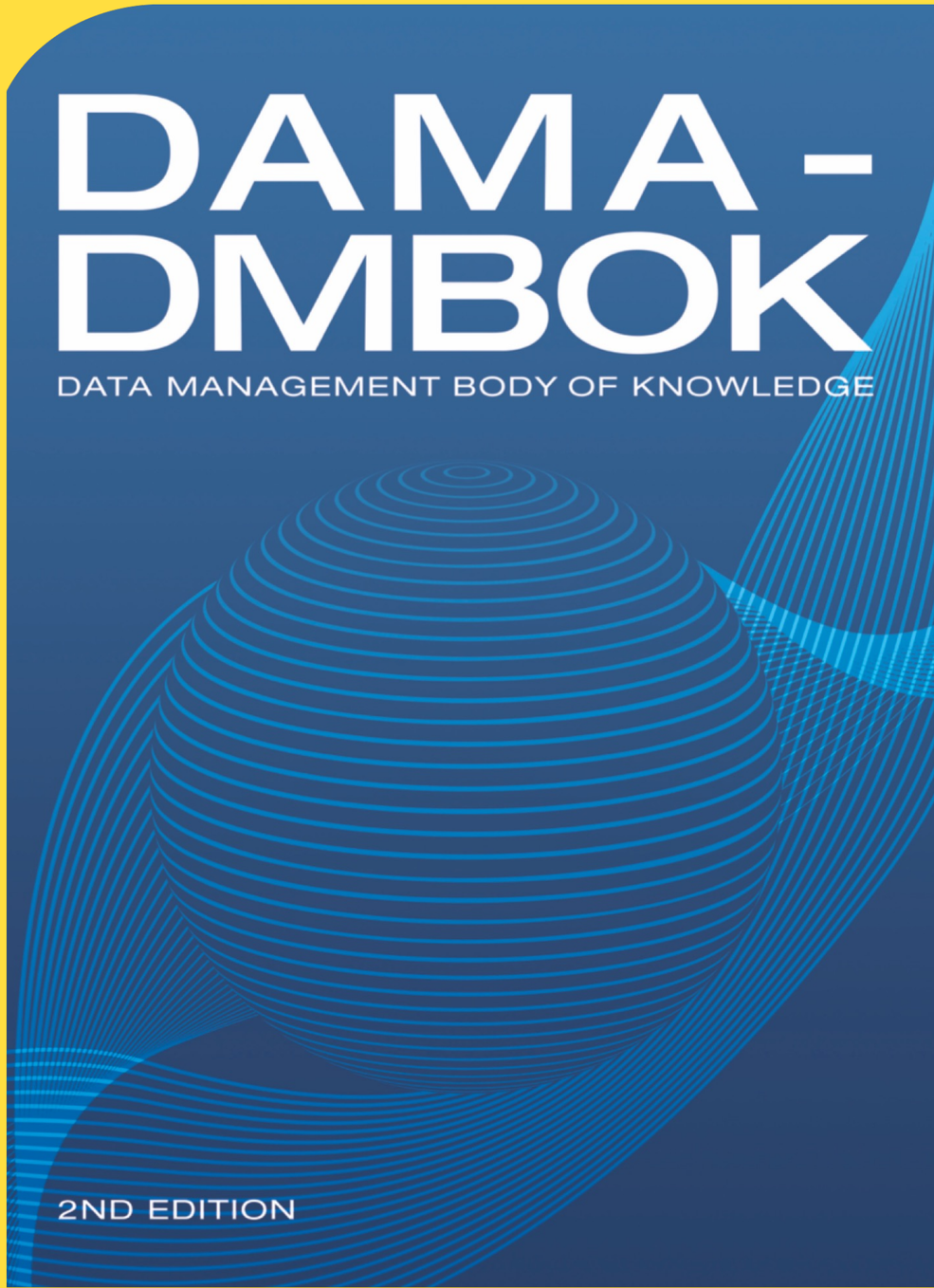
Платформа



Data Management

Data Management is the development, execution, and supervision of plans, policies, programs, and practices that **deliver**, **control**, **protect**, and **enhance the value of data assets throughout their lifecycles**.

The primary driver for data management is to enable organizations **to get value from their data assets**.



Data Management Goals

- Understanding and supporting the information needs of the enterprise and its stakeholders, including customers, employees, and business partners
- Capturing, storing, protecting, and ensuring the integrity of data assets
- Ensuring the quality of data and information
- Ensuring the privacy and confidentiality of stakeholder data
- Preventing unauthorized or inappropriate access, manipulation, or use of data and information
- Ensuring data can be used effectively to add value to the enterprise

Data Lifecycle

Creation and usage are the most critical points:

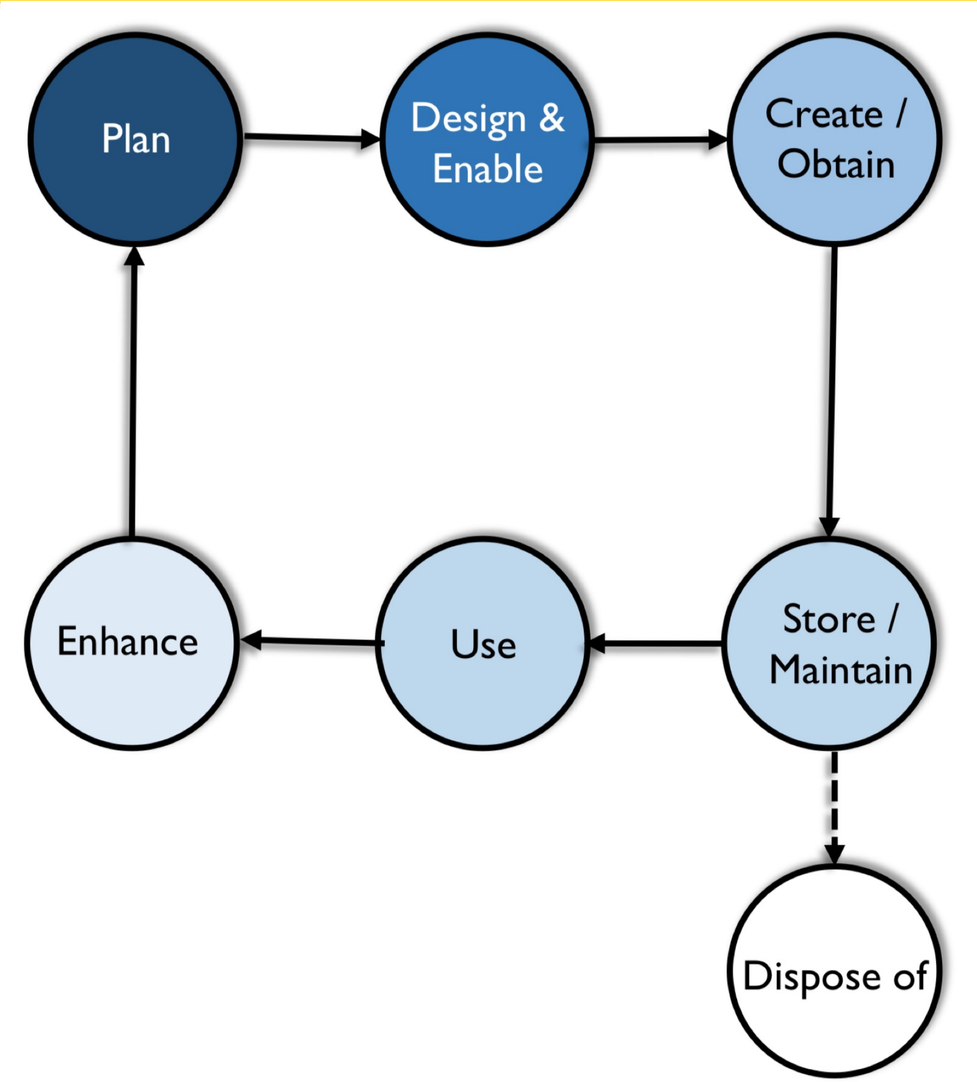
- It costs money to produce data
- Data is valuable only when it is consumed or applied

Management thought lifecycle:

- Data Quality
- Metadata
- Data Security

Efforts should focus on the most critical data:

- Organisations produce a lot of data
- Large portions of data is never actually used
- Manage every piece of data is not possible



Data Lifecycle

Creation and usage are the most critical points:

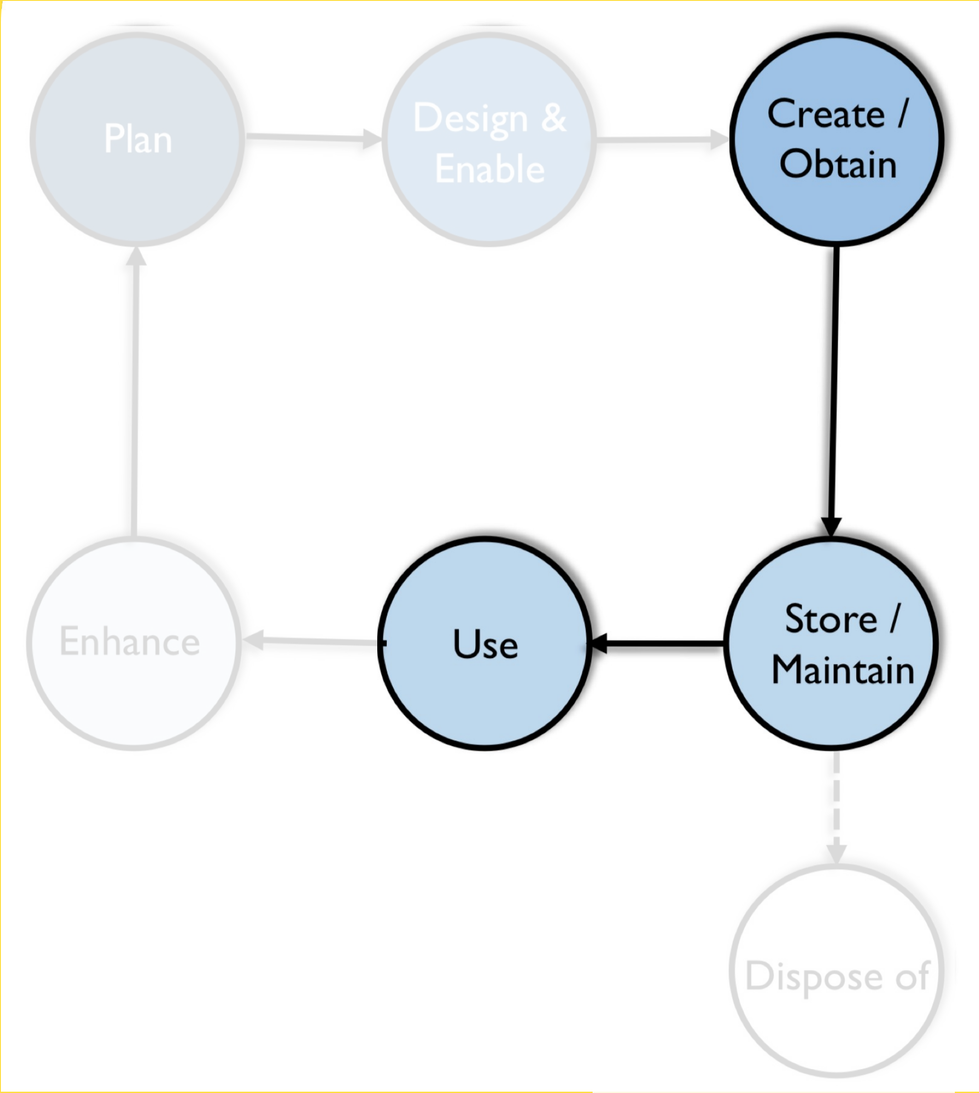
- It costs money to produce data
- Data is valuable only when it is consumed or applied

Management thought lifecycle:

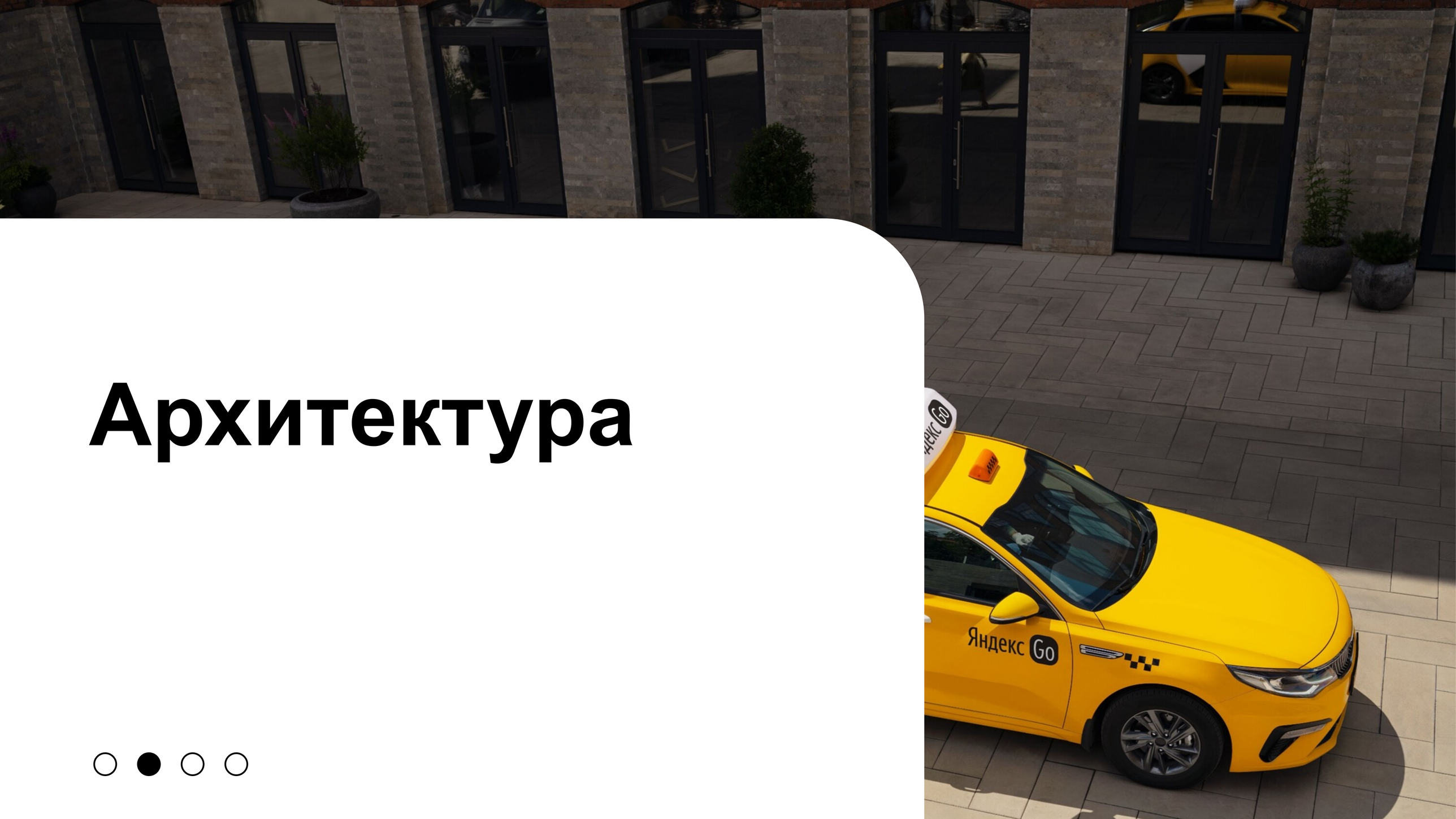
- Data Quality
- Metadata
- Data Security

Efforts should focus on the most critical data:

- Organisations produce a lot of data
- Large portions of data is never actually used
- Manage every piece of data is not possible



Архитектура



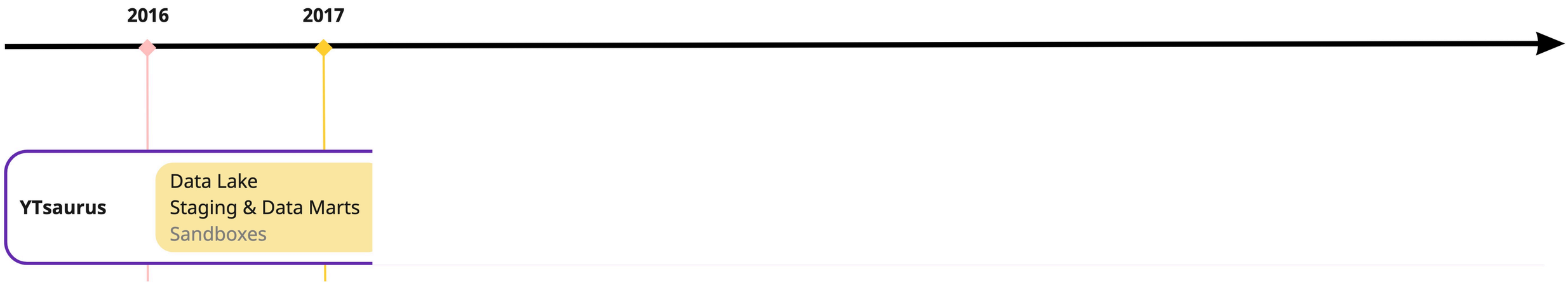
2016

Нужны данные и отчеты

- Реплики баз бекенда
- Витрины по основным сущностям
- Базовые отчеты и дашборды

- Много источников: десятки и их количество непрерывно растет
- Много данных: терабайты и объем растет по экспоненте

- Команда маленькая
- Активно идет найм

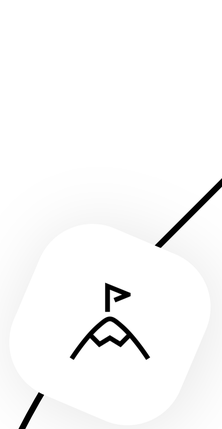

2016
2017
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes


2016
2017
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
BI
Yandex In-house

Бизнес быстро растет

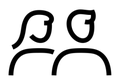
Бэкенд активно развивается

Нужно бежать быстрее

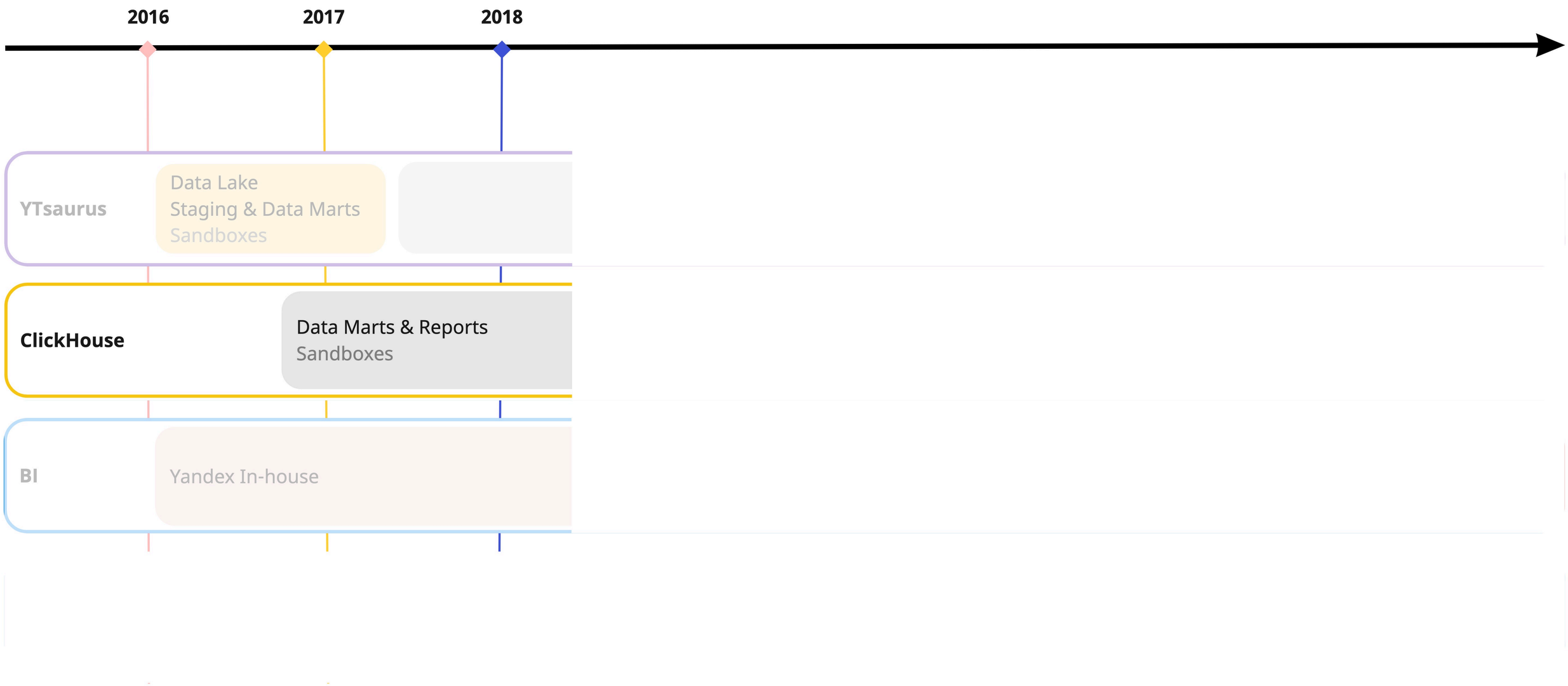


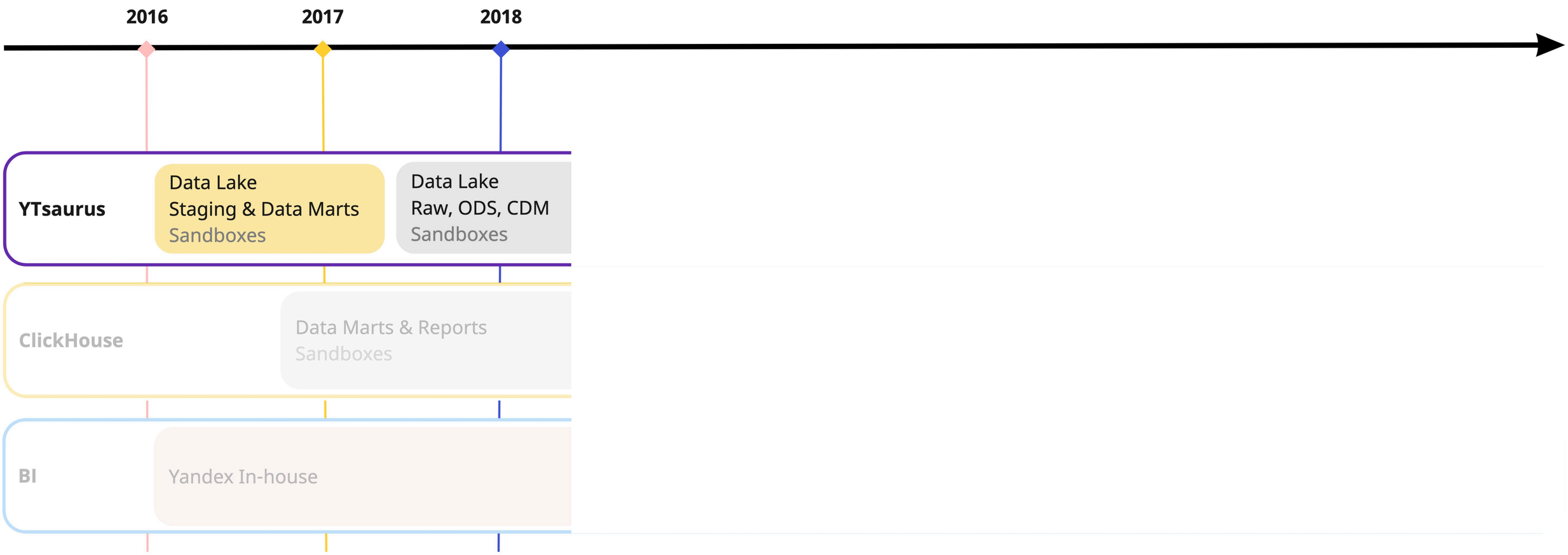
- Быстрый доступ к витринам для ad-hoc и сложные аггрегаты
- Уменьшение latency
- Нужно постоянно пересчитывать историю

- Много источников: десятки и их количество непрерывно растет
- Много данных: терабайты и объем растет по экспоненте

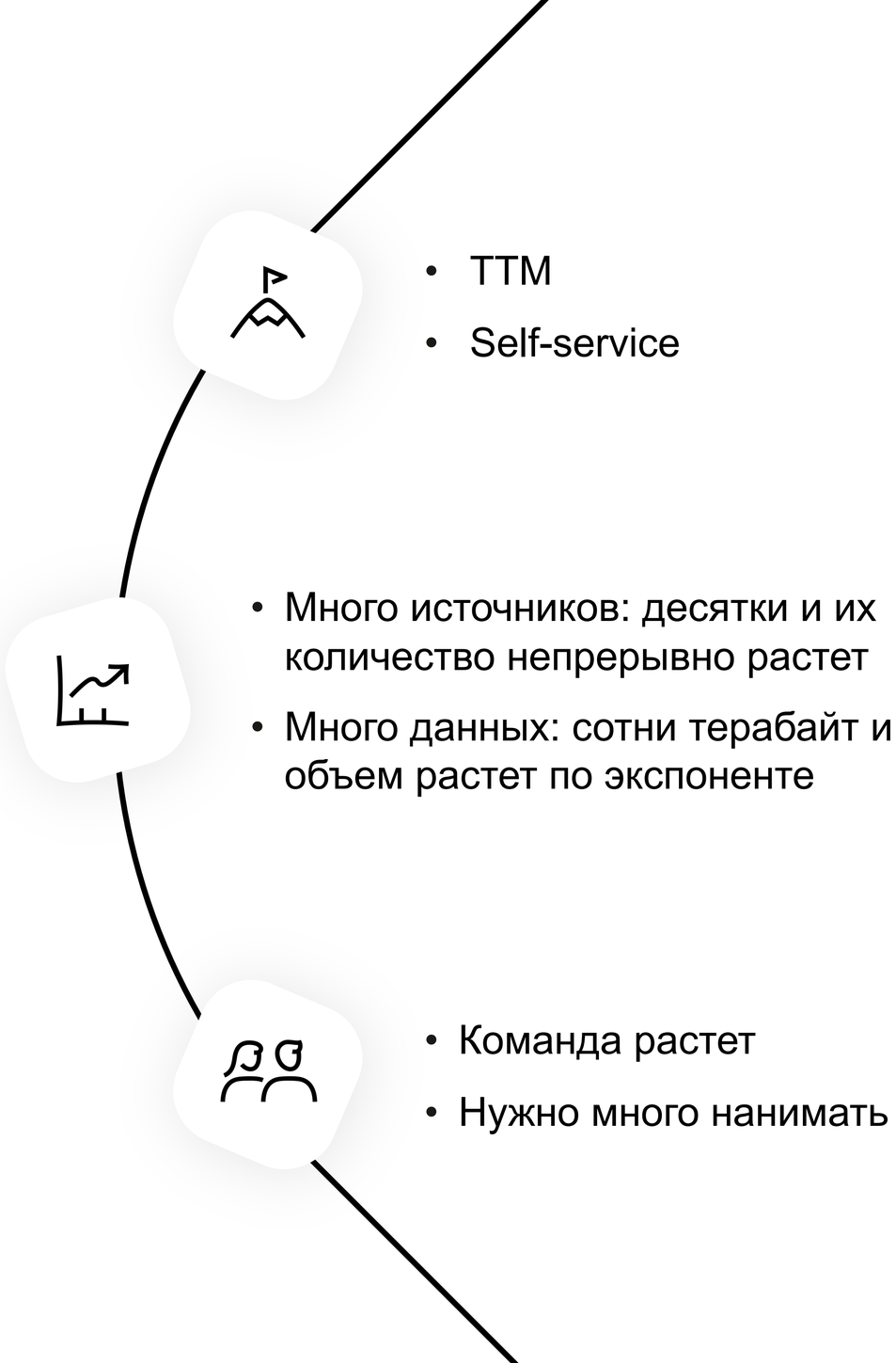


- Команда маленькая
- Активно идет найм


2016
2017
2018
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house


2016
2017
2018
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house

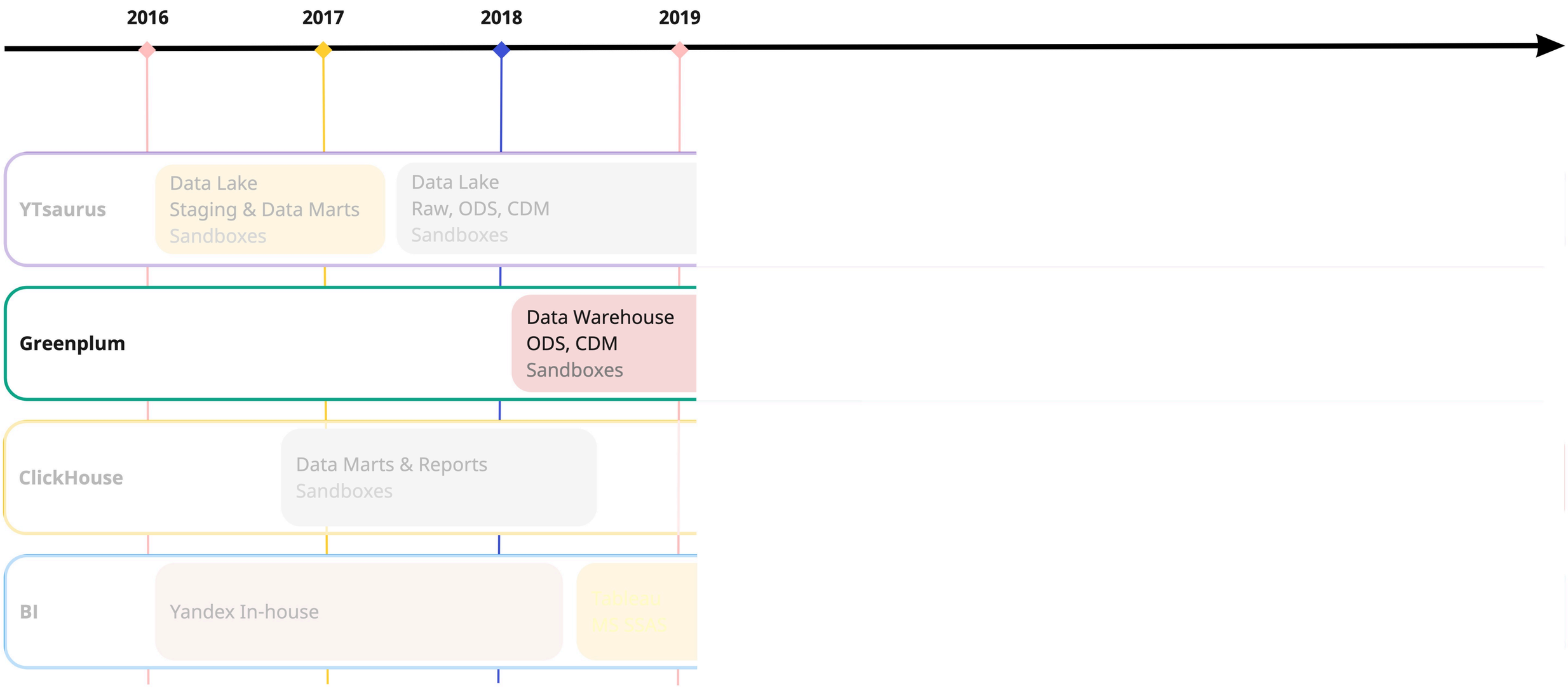
Компания стала большой

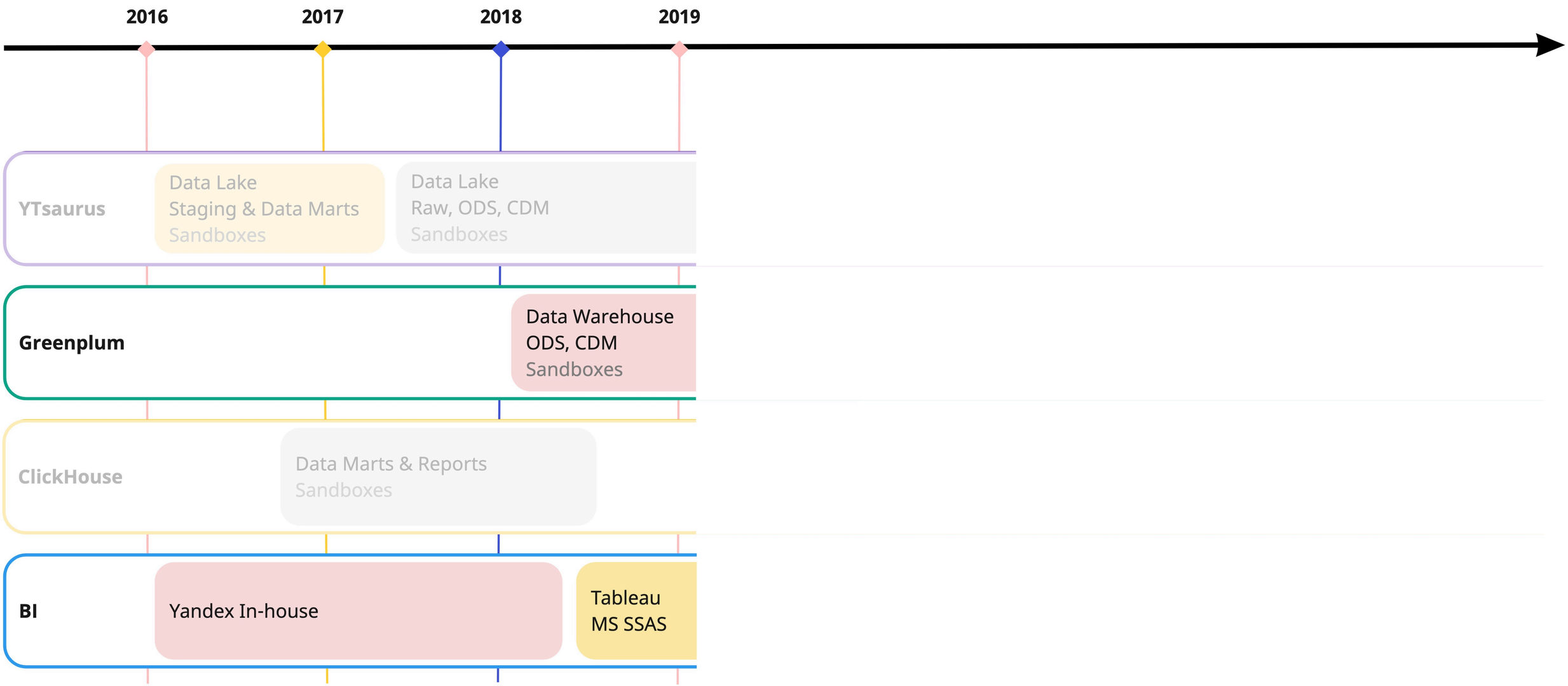


- TTM
- Self-service

- Много источников: десятки и их количество непрерывно растет
- Много данных: сотни терабайт и объем растет по экспоненте

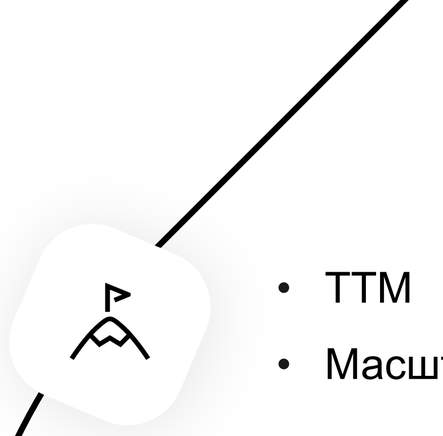
- Команда растет
- Нужно много нанимать


2016
2017
2018
2019
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS


2016
2017
2018
2019
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS

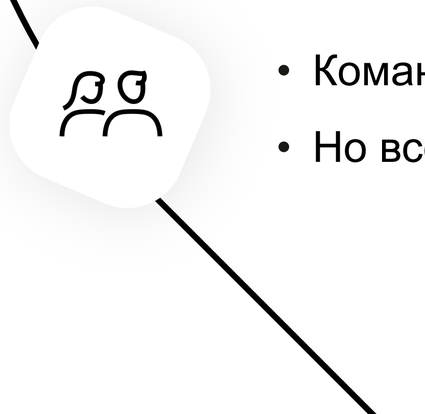
Развиваются новые бизнесы

Растет сложность предметной области

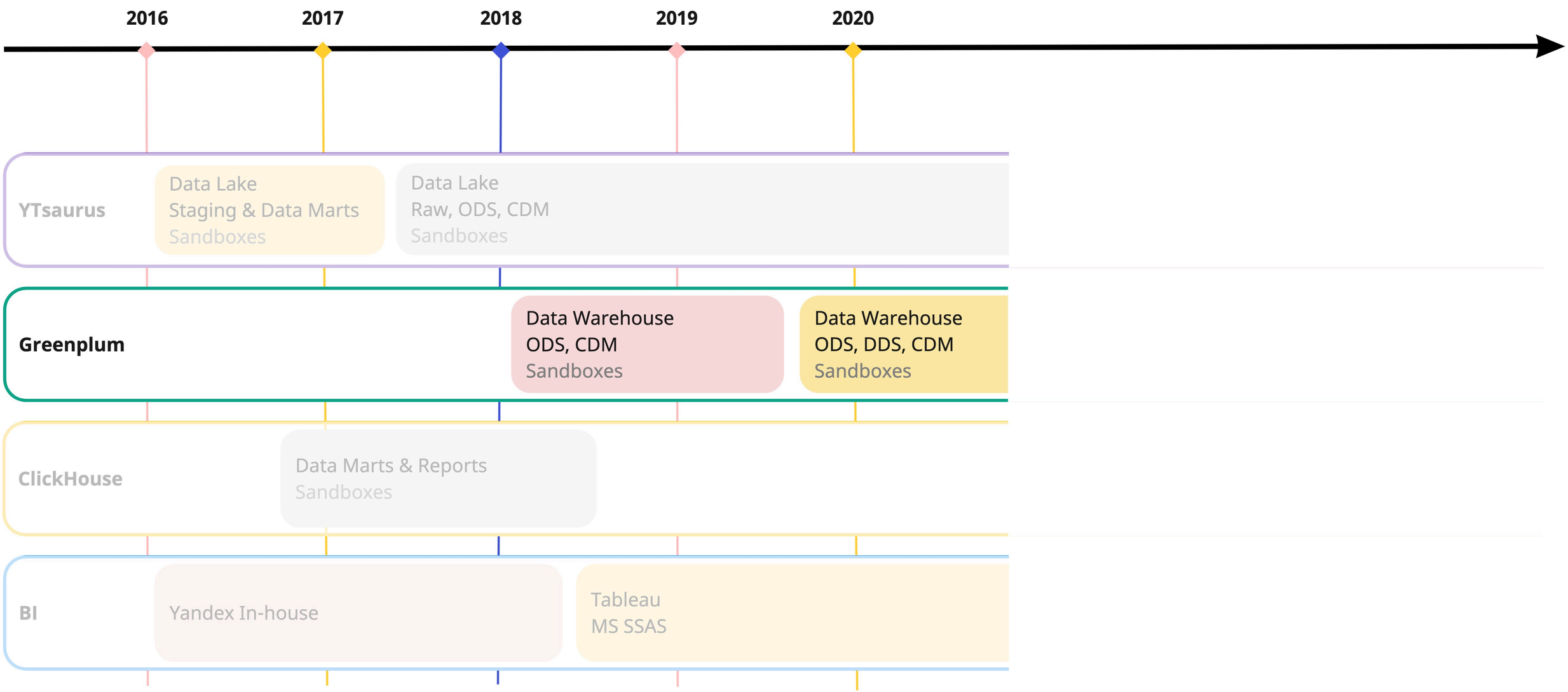


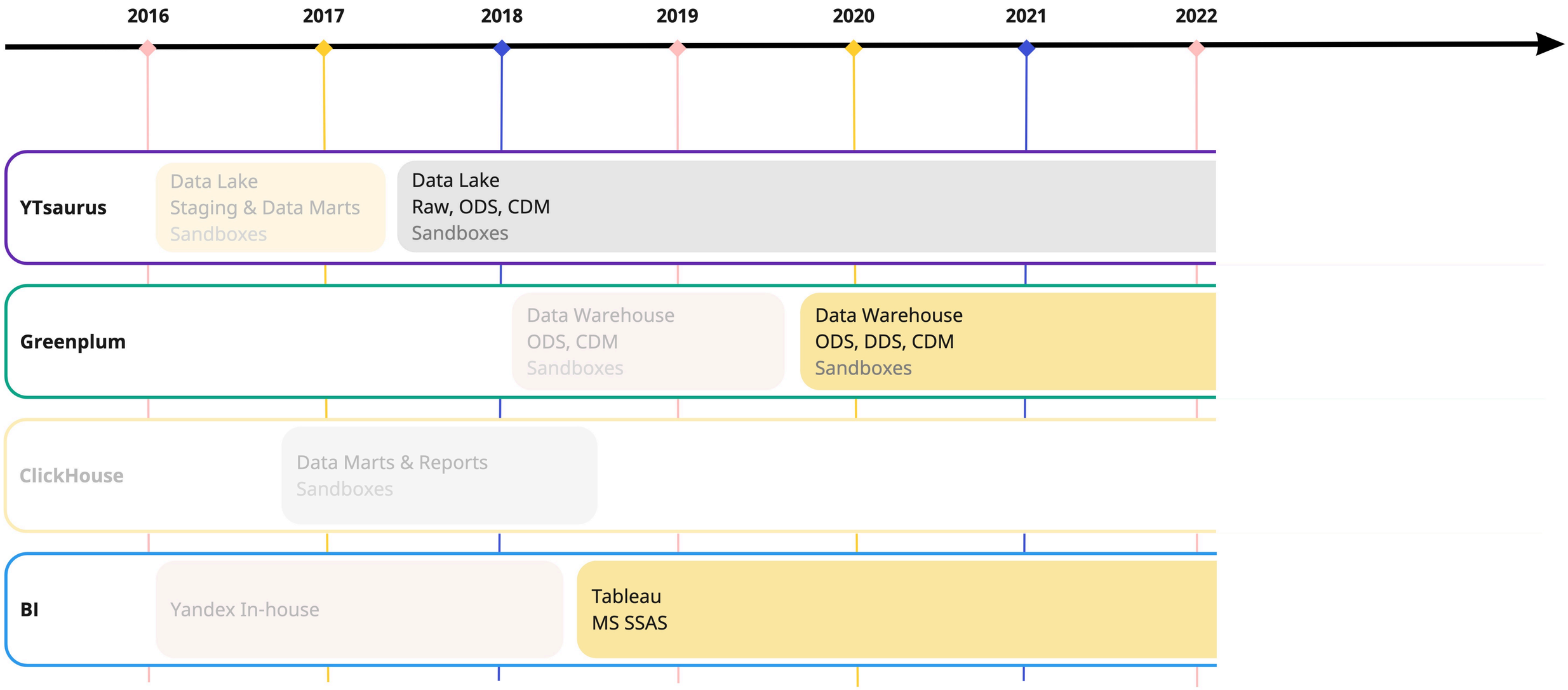
- TTM
- Масштабирование

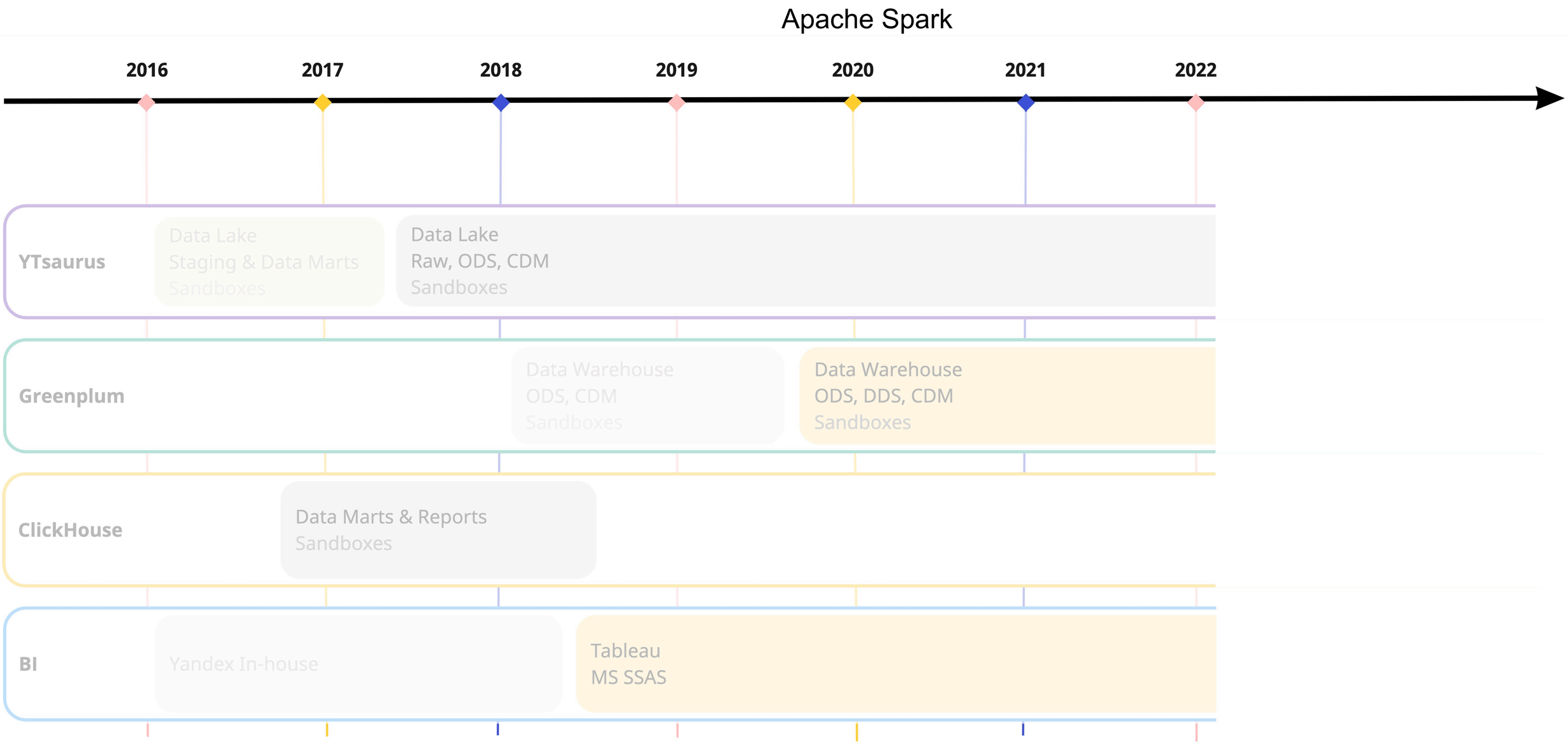
- Много источников: сотни и их количество непрерывно растет
- Много данных: петабайты и объем продолжает расти



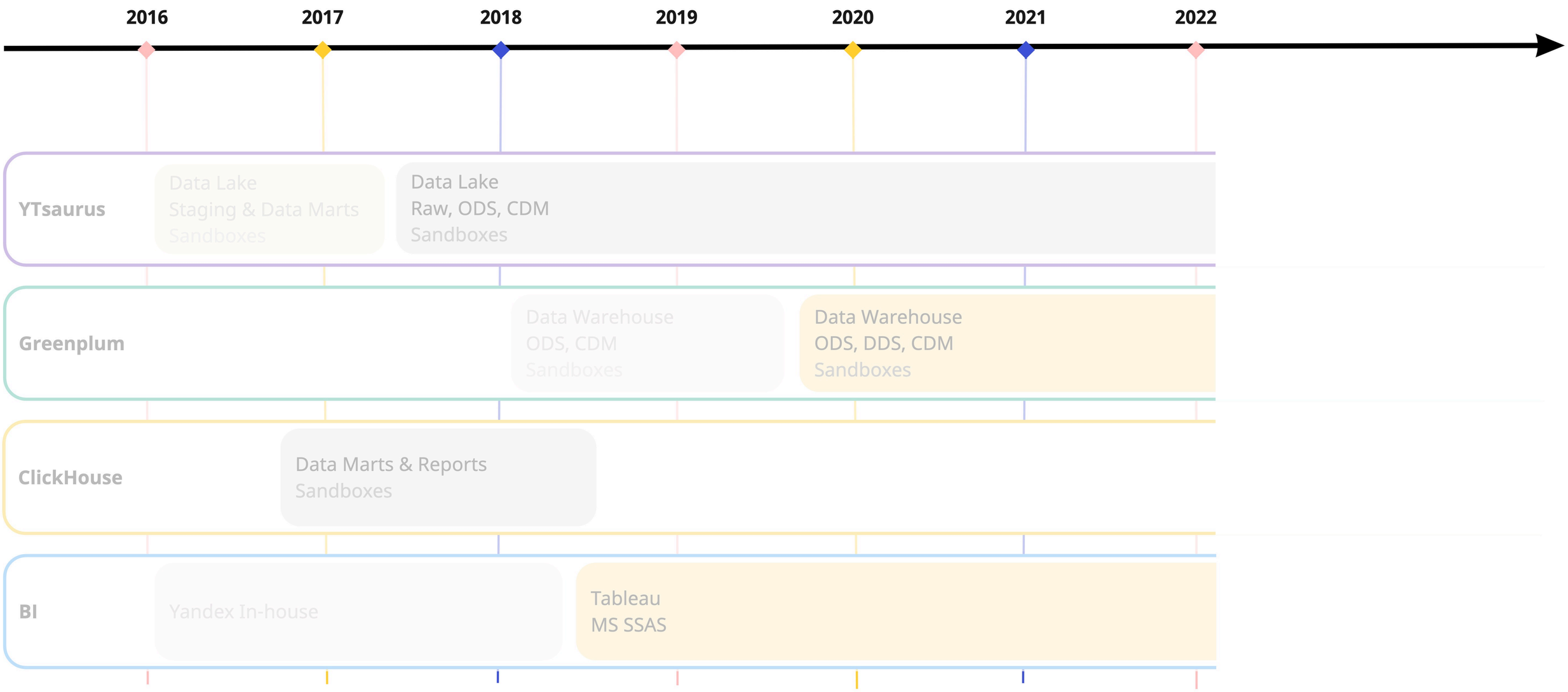
- Команда выросла
- Но все еще нужно много нанимать

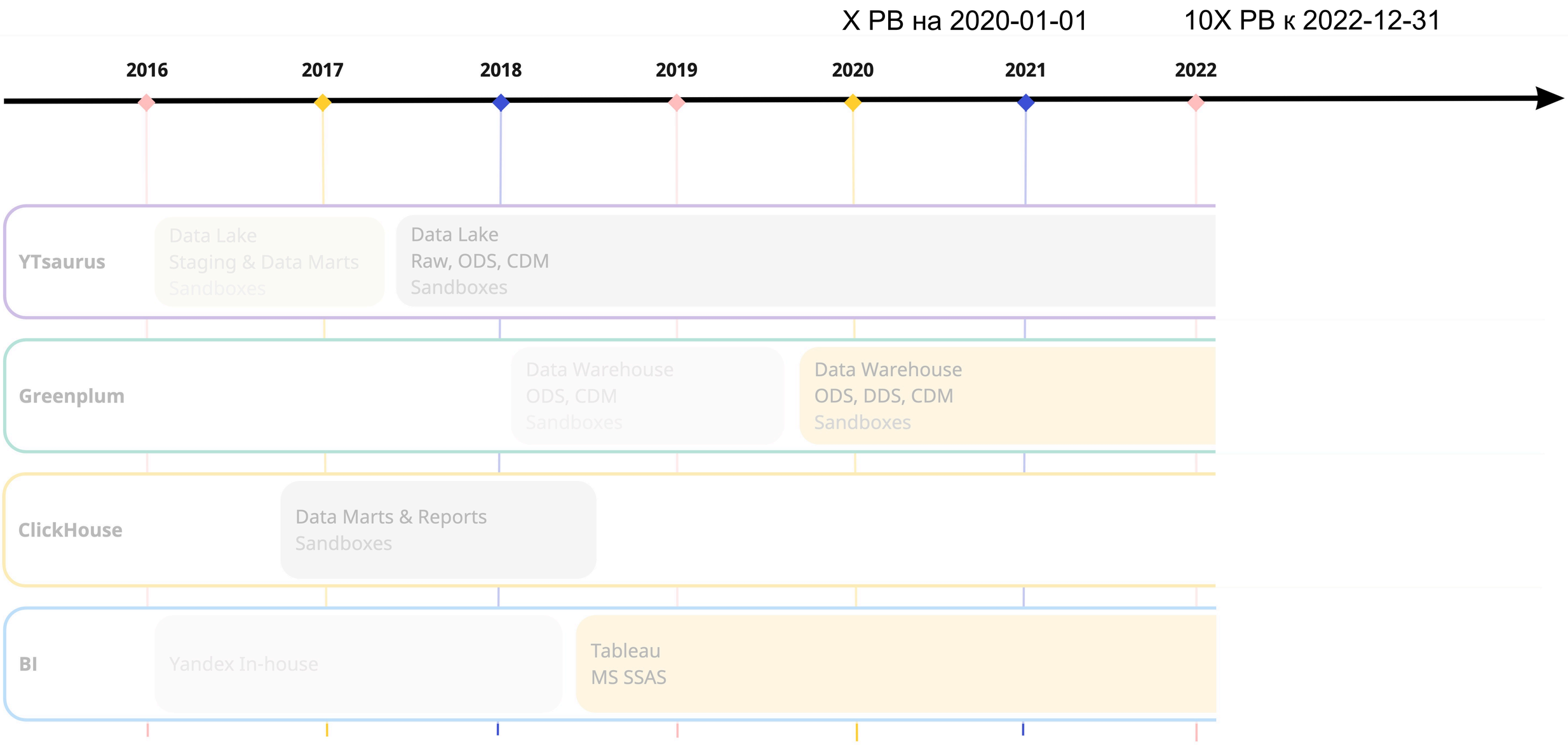

2016
2017
2018
2019
2020
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS


2016
2017
2018
2019
2020
2021
2022
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS


Apache Spark
2016
2017
2018
2019
2020
2021
2022
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS

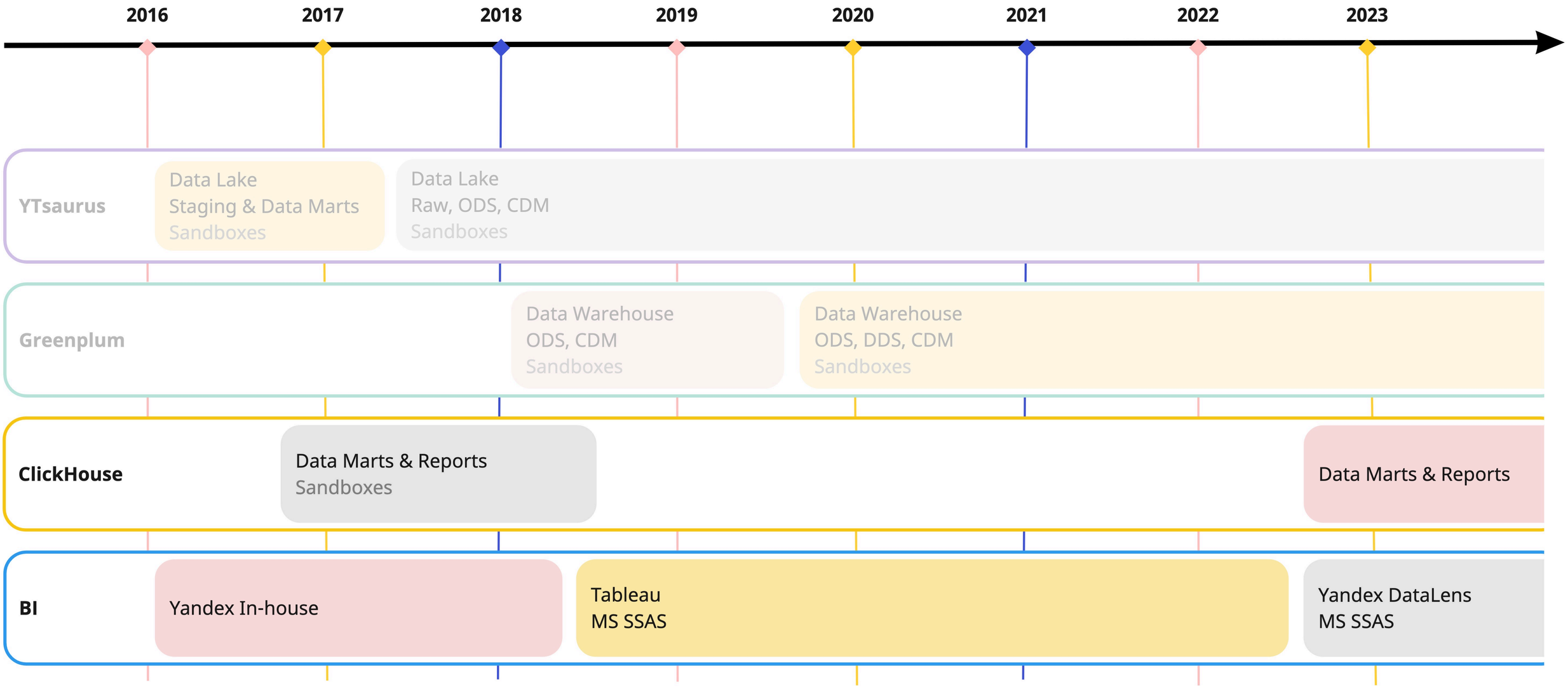
Свой фреймворк

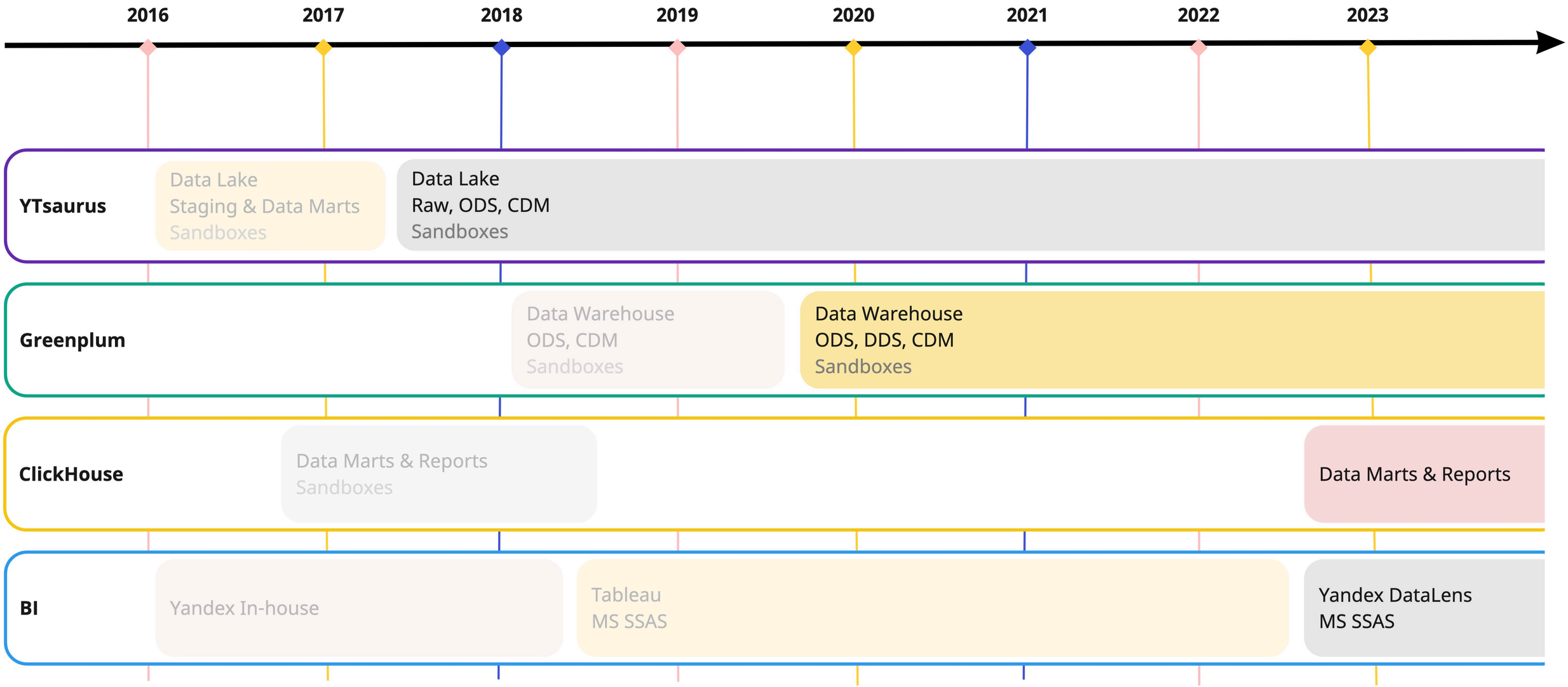


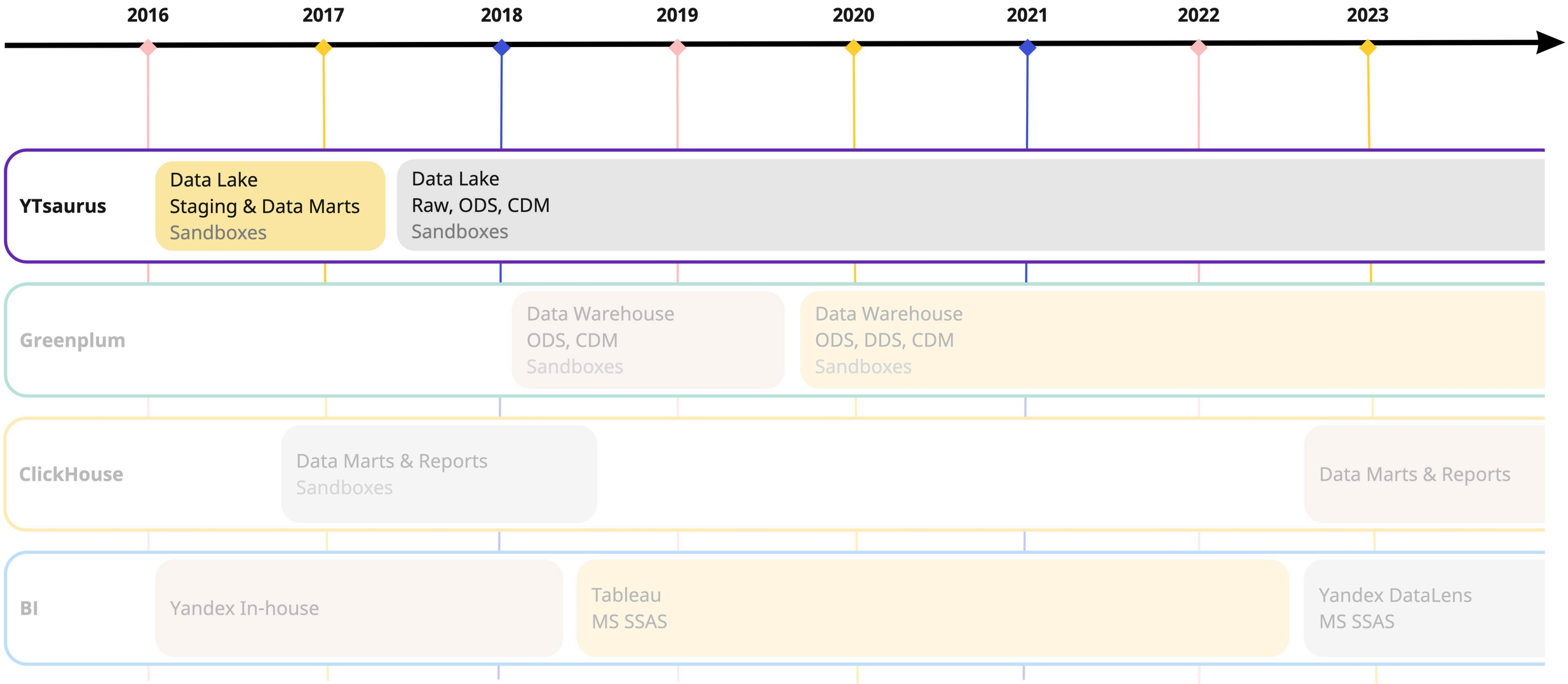

X PB на 2020-01-01
10X PB к 2022-12-31
2016
2017
2018
2019
2020
2021
2022
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
BI
Yandex In-house
Tableau
MS SSAS

Внешние факторы

- Антикризис
- Много источников: тысячи и их количество непрерывно растет
- Много данных: петабайты и объем продолжает расти
- Большая команда
- Но все еще нужно много нанимать


2016
2017
2018
2019
2020
2021
2022
2023
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
Data Marts & Reports
BI
Yandex In-house
Tableau
MS SSAS
Yandex DataLens
MS SSAS

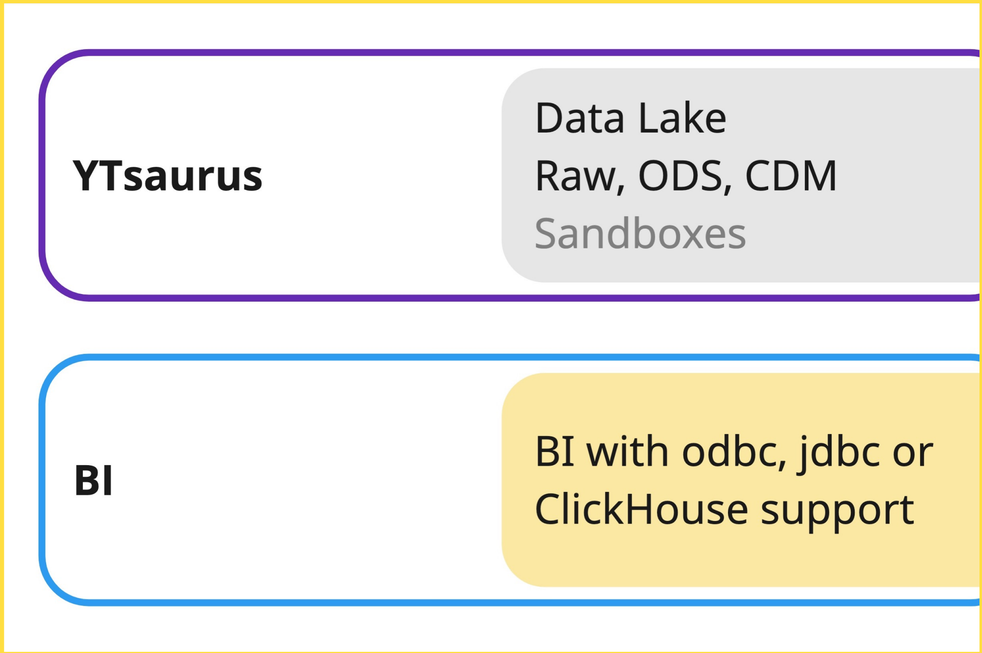

2016
2017
2018
2019
2020
2021
2022
2023
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
Data Marts & Reports
BI
Yandex In-house
Tableau
MS SSAS
Yandex DataLens
MS SSAS


2016
2017
2018
2019
2020
2021
2022
2023
YTsaurus
Data Lake
Staging & Data Marts
Sandboxes
Data Lake
Raw, ODS, CDM
Sandboxes
Greenplum
Data Warehouse
ODS, CDM
Sandboxes
Data Warehouse
ODS, DDS, CDM
Sandboxes
ClickHouse
Data Marts & Reports
Sandboxes
Data Marts & Reports
BI
Yandex In-house
Tableau
MS SSAS
Yandex DataLens
MS SSAS

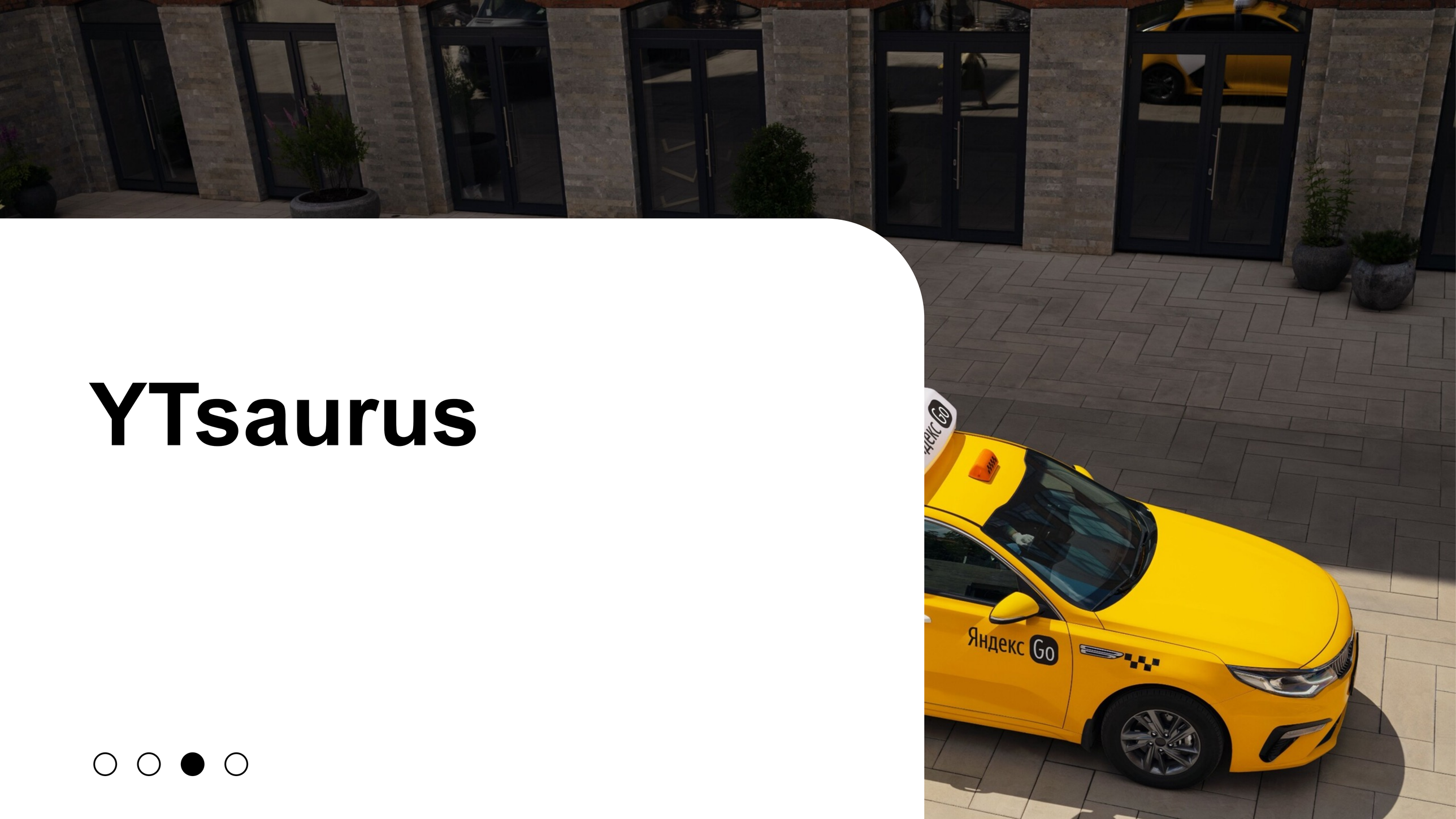
Можно проще

YTsaurus является необходимым и достаточным компонентом нашей платформы для множества сценариев.

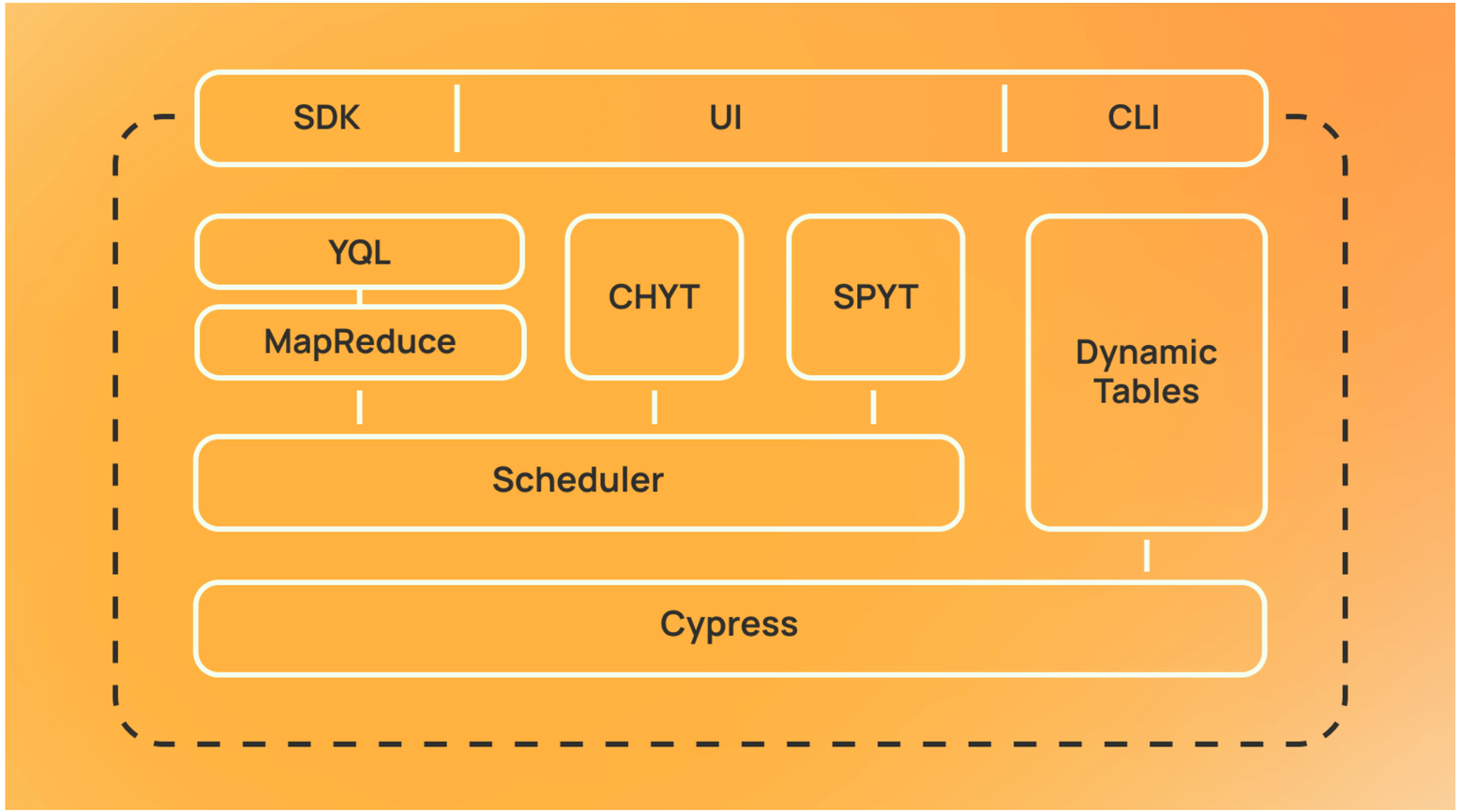
Достаточно дополнить YTsaurus практически любым BI инструментом и это сразу закроет огромное количество пользовательских сценариев в DWH, BI, аналитике и ML.



YTsaurus



Архитектура YTsaurus



https://ytsaurus.tech/ru/platform-overview

1. Cypress

Кипарис - распределенная файловая система и хранилище метаданных. Хранит row- и column-oriented таблицы для batch и OLTP нагрузки

2. Scheduler

Планировщик с поддержкой парадигмы MapReduce и на основе принципа Dominant Resource Fairness

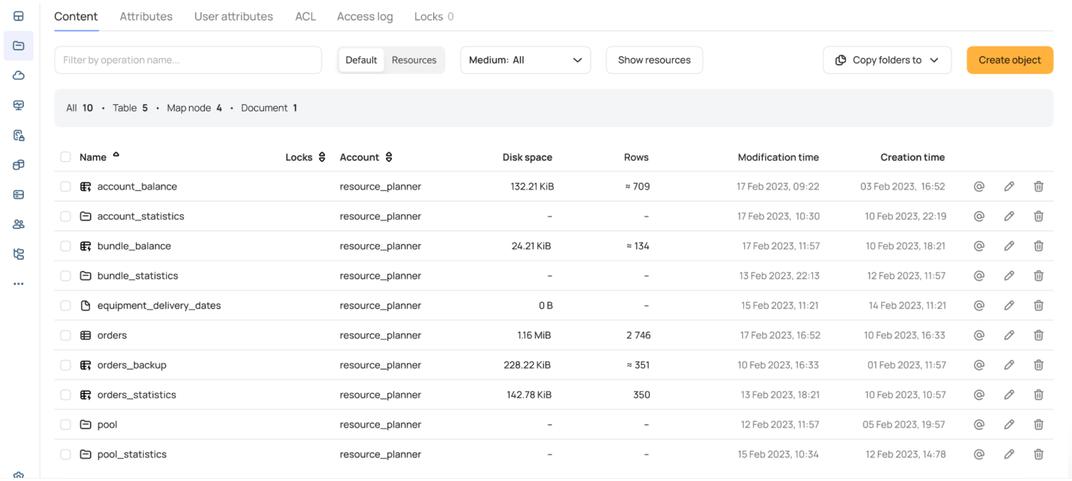
3. Compute

MapReduce и высокоуровневые вычислительные движки:

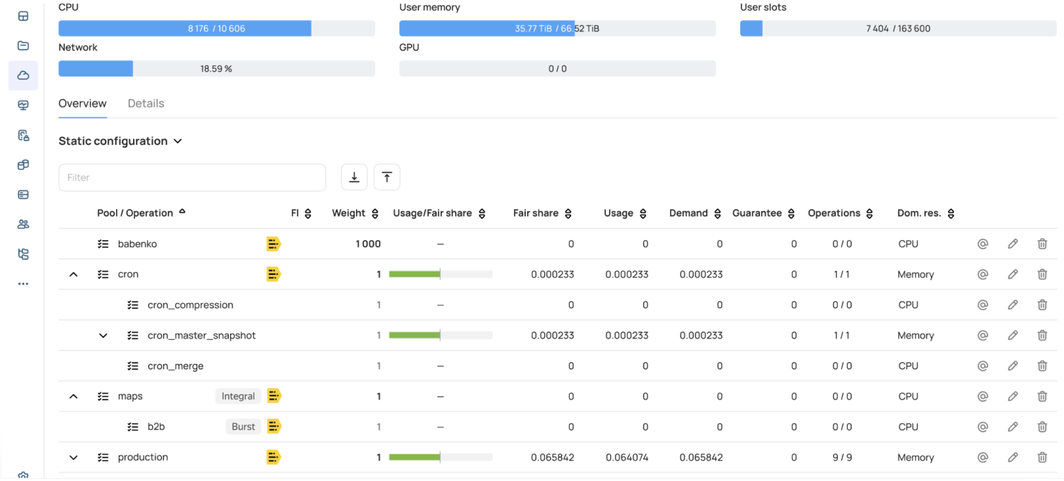
- YQL
- CHYT
- SPYT

Интерфейс

Удобный, понятный и с богатой функциональностью



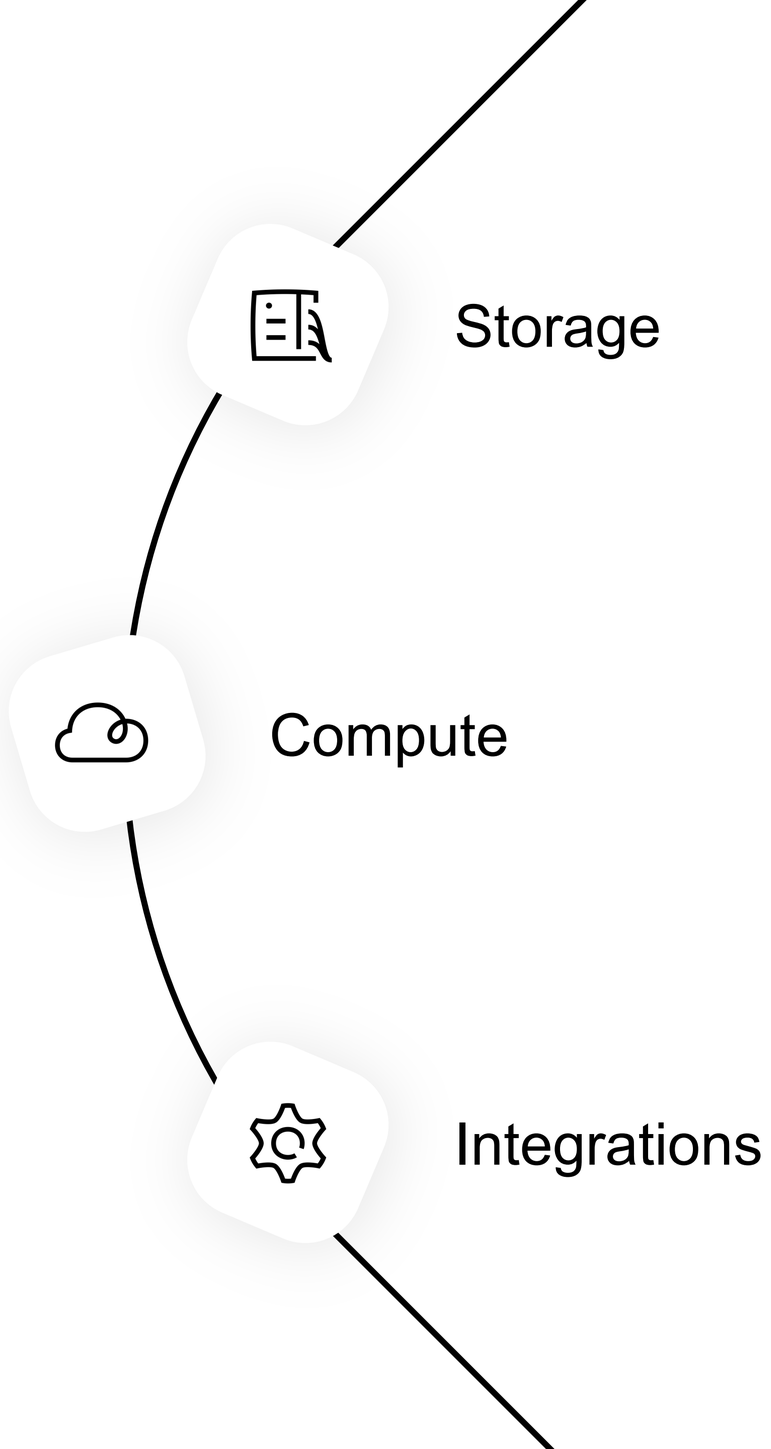
Cypress – распределённая файловая система и дерево метаинформации



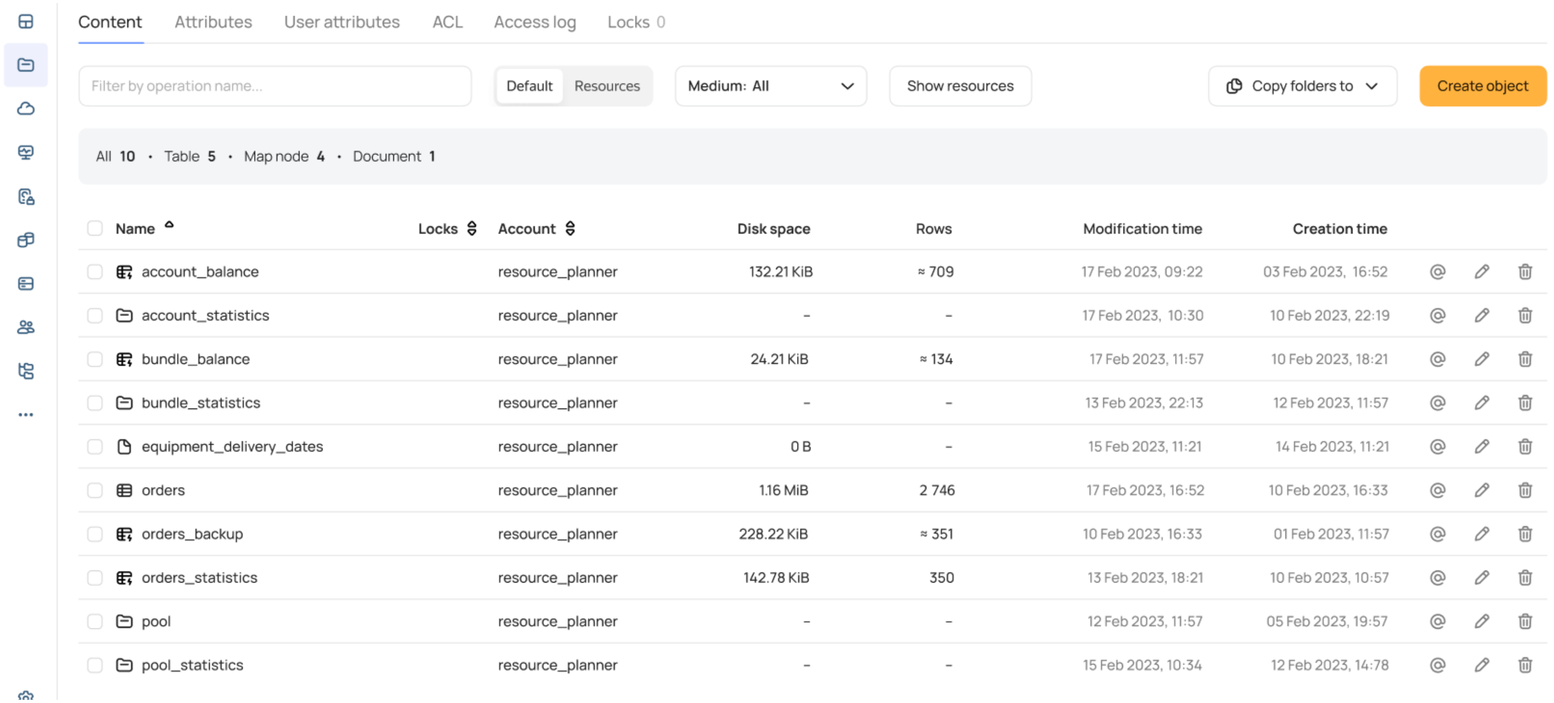
Scheduler – планировщик, который управляет несколькими видами ресурсов (CPU, RAM, GPU) на основе принципа Dominant Resource Fairness

https://ytsaurus.tech/ru/platform-overview

О чем будем говорить



Storage



Name	Locks	Account	Disk space	Rows	Modification time	Creation time
account_balance		resource_planner	132.21 KiB	≈ 709	17 Feb 2023, 09:22	03 Feb 2023, 16:52
account_statistics		resource_planner	–	–	17 Feb 2023, 10:30	10 Feb 2023, 22:19
bundle_balance		resource_planner	24.21 KiB	≈ 134	17 Feb 2023, 11:57	10 Feb 2023, 18:21
bundle_statistics		resource_planner	–	–	13 Feb 2023, 22:13	12 Feb 2023, 11:57
equipment_delivery_dates		resource_planner	0 B	–	15 Feb 2023, 11:21	14 Feb 2023, 11:21
orders		resource_planner	1.16 MiB	2 746	17 Feb 2023, 16:52	10 Feb 2023, 16:33
orders_backup		resource_planner	228.22 KiB	≈ 351	10 Feb 2023, 16:33	01 Feb 2023, 11:57
orders_statistics		resource_planner	142.78 KiB	350	13 Feb 2023, 18:21	10 Feb 2023, 10:57
pool		resource_planner	–	–	12 Feb 2023, 11:57	05 Feb 2023, 19:57
pool_statistics		resource_planner	–	–	15 Feb 2023, 10:34	12 Feb 2023, 14:78

Таблицы

- Статические – основной тип таблиц
- Динамические – key-value хранилище
- Row- & column-oriented
- Богатая система типов

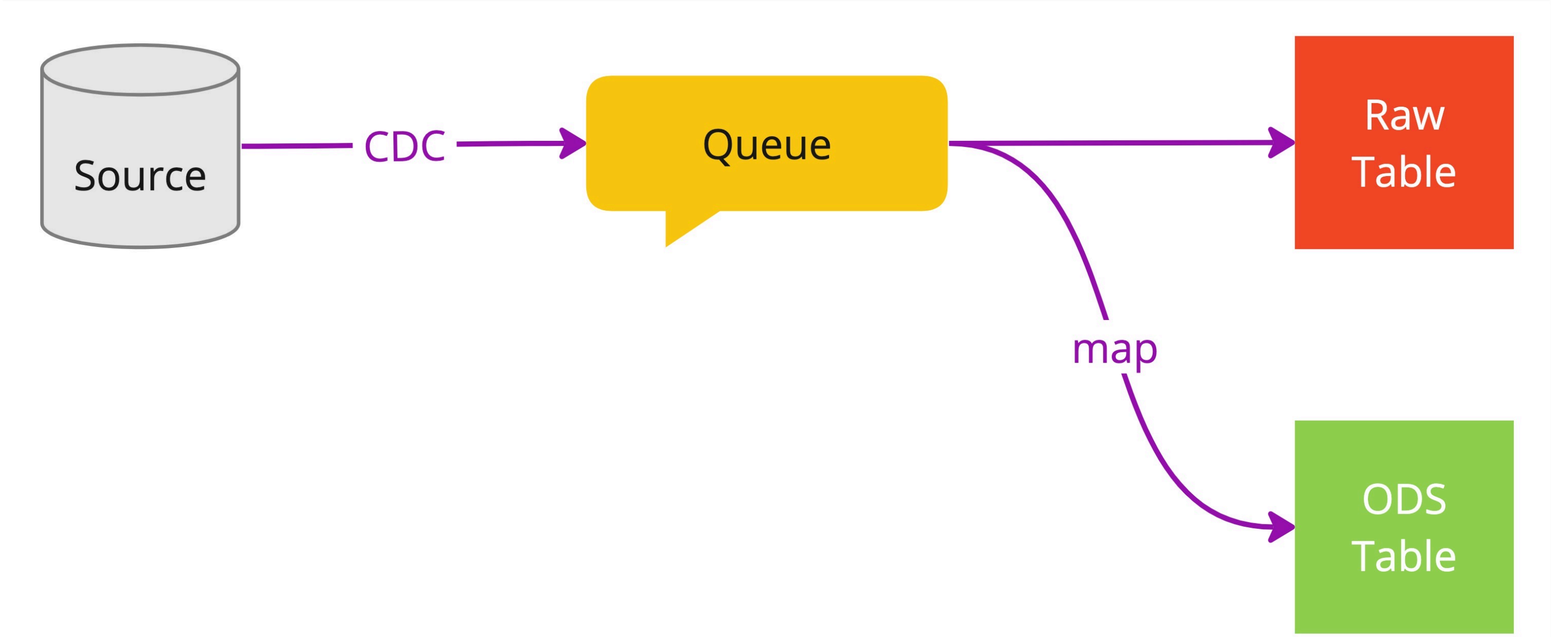
Транзакции

- ACID
- Могут длиться часами и даже днями
- Затрагивают Cypress, а не только таблицы

Квоты

- Диск: HDD, SSD
- Кол-во узлов в Cypress
- Кол-во чанков
- Кол-во таблетов (шардов) в динамических таблицах

ETL в динамические таблицы



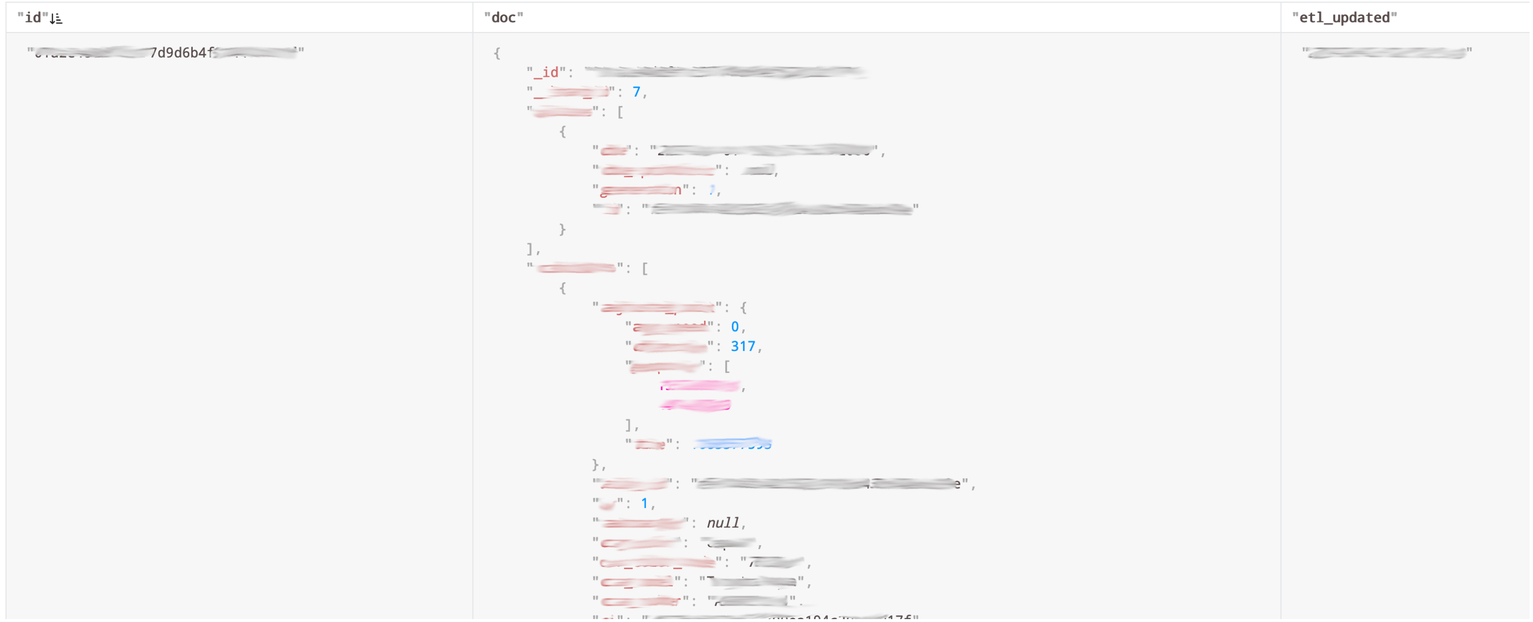
100+ MB/s

- Есть пиковые нагрузки
- Есть особо «жирные» потоки

1000+ сущностей

2+ ПБ

Таблицы в «Raw» слое



Ключи

- Primary key
- Partition key

Doc

- Документ как есть в формате Yson

ETL дата

Таблицы в ODS

"order_id"				
"C	"	4 666666666	6666667	null
"	"	5039999999995	6666667	null
"	null	null	null	null
"	"	4071975816	5262087	null
"	"	809826286	354255	null
"	"	9155188	11555529	null
"	"	3038841215	7338553	"
"	"	65	99999995	null
"	"	974861	287532	null
"	"		333334	null
"	"	309841454	9270184	null
"	null	null	null	null
"	null	null	null	null
"	"	666666667	18	null
"	"	3333333	6666667	null

Column-oriented

- Усложняет вставку
- Упрощает чтение
- Читаем данные мы чаще, чем пишем

Схема

3NF, ну почти

- Разрешаем массивы и словари

Статические таблицы для логов и витрин

Sorting

- Может гарантировать уникальность ключа
- Ускоряет фильтрацию, например, по дате
- Ускоряет join-ы по ключу сортировки
- Life hack: может заметно экономить диск для column-oriented таблиц*

Schema

- Примитивы
- Сложные типы: массивы, словари, структуры и т.п.
- Есть constraint: required (not null)
- Может быть strong, weak или отсутствовать

Partitioning

- Нет нативной поддержки, но есть стандартный подход
- Эмулируется через папки: папку считаем таблицей, а таблице в ней – партициями
- Ускоряет обработку и экономит вычислительные ресурсы

Таблицы: управляем размером

Compression

- Большой выбор кодеков сжатия
- Можно выбирать разные кодеки для «горячих» и «холодных» данных

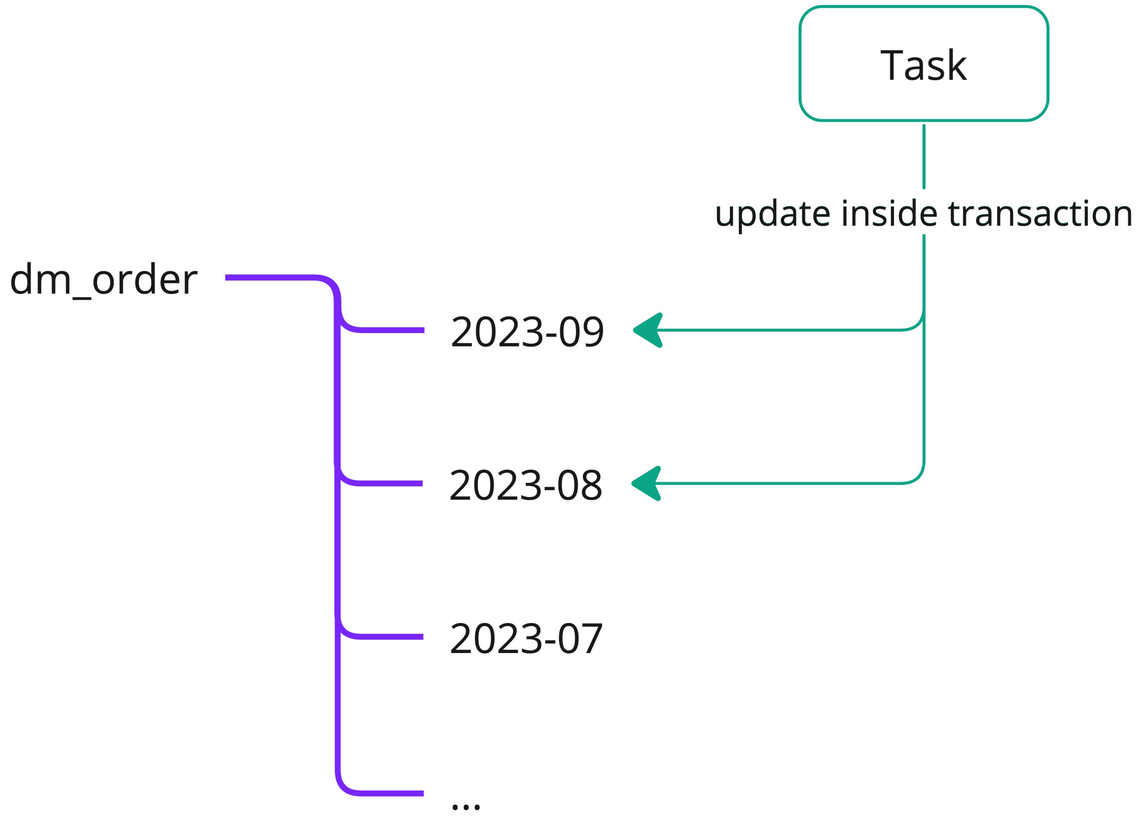
Time To Live

- Expiration Time – момент времени, когда надо удалить узел
- Expiration Timeout – время без обращений к узлу

Redundancy

- Replication: меняем диск на CPU, управляем количеством реплик
- Erasure Coding: меняем CPU на диск, подходит для «холодных» редко используемых данных, объем диска от 1.33x до 1.5x от сжатых данных

Транзакции: расчет за N дней



ACID

Isolation:

- Serializable для таблиц
- Read Commited для Cypress

N узлов Cypress

- Можно пересчитать всю историю и обновить в рамках одной транзакции
- Можно изменять структуру файловой системы

Незаметно для пользователя

- Snapshot Lock на чтение – read-only копия узла
- Exclusive Lock на модификацию узла
- Shared Lock на модификацию части состояния узла

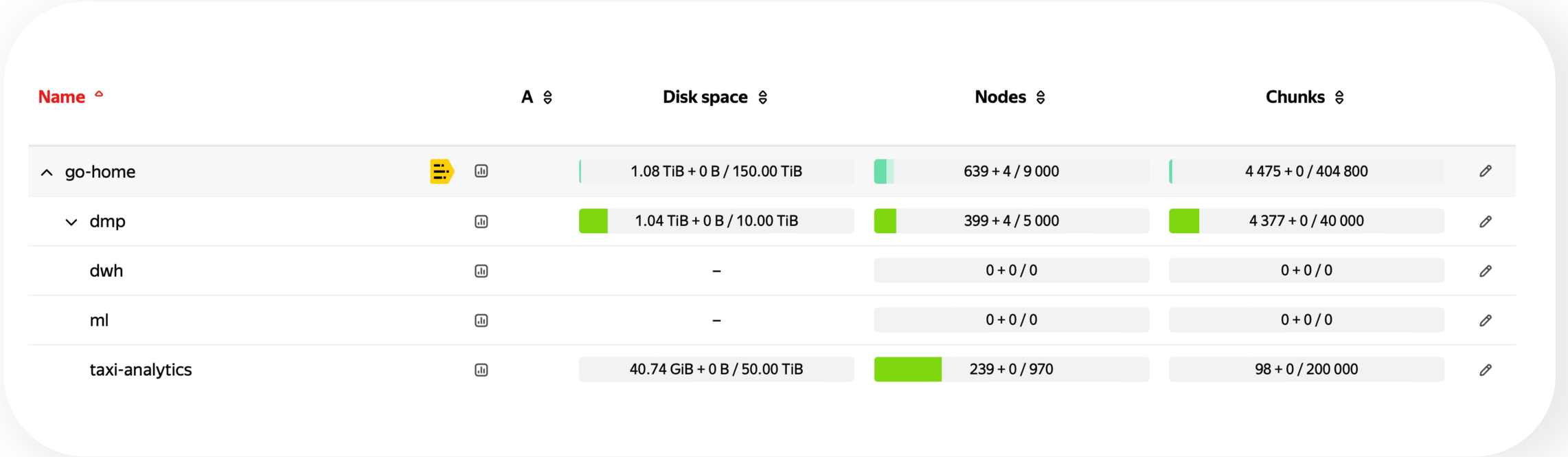
Изоляция: квоты и аккаунты

Квоты:

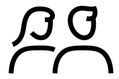
- Размер данных: HDD, SSD
- Кол-во объектов: nodes, chunks

Аккаунты:

- Аккаунты могут быть вложенными, это упрощает управление квотами
- Аккаунт может быть указан для любого узла
- Может быть over commit



Управление доступами



Доступ можно выдать на

- Списки пользователей
- Группы

Группы могут быть созданы внутри самого YTsaurus.

Каждый объект в YTsaurus имеет **Access Control List** (ACL).

Пример:

```
"acl": [
  {
    "action": "allow",
    "subjects": ["user1"],
    "permissions": ["read"],
  },
  ...
]
```

ACL может наследоваться от родительского узла в Cypress.

Это позволяет гибко и удобно управлять доступами и минимизировать их количество.

Можно настраивать **Column Level Security**.

CLS позволяет делать меньше таблиц, но скрывать доступ к чувствительным данным, например, к финансовой информации.

Сравним Storage

YTsaurus

Плюсы:

- Нативная поддержка таблиц
- Нативный Key-Value Storage
- Транзакционность
- Enterprise Level управление квотами и правами
- Удобный и функциональный UI

Минусы:

- В самом начале пути в Open Source
- Нет готовых специалистов на рынке

s3 & hdfs

Плюсы:

- Давно в Open Source: всем известны плюсы, минусы и грабли
- «Много» готовых специалистов
- Огромная экосистема Hadoop & Open Source Big Data Stack
- Доступны в «любом» облаке

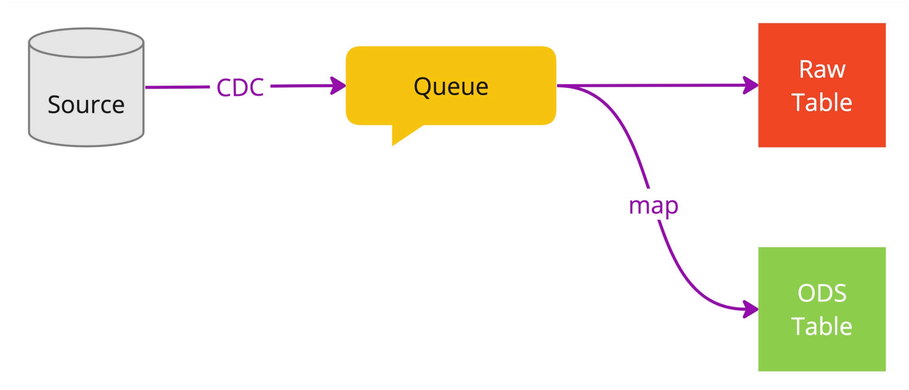
Минусы:

- Нет транзакций
- Для таблиц или key-value нужны отдельные «кубики»

MapReduce в 2023?!..

MapReduce в 2023?!..

Потоковая поставка данных в Raw и ODS

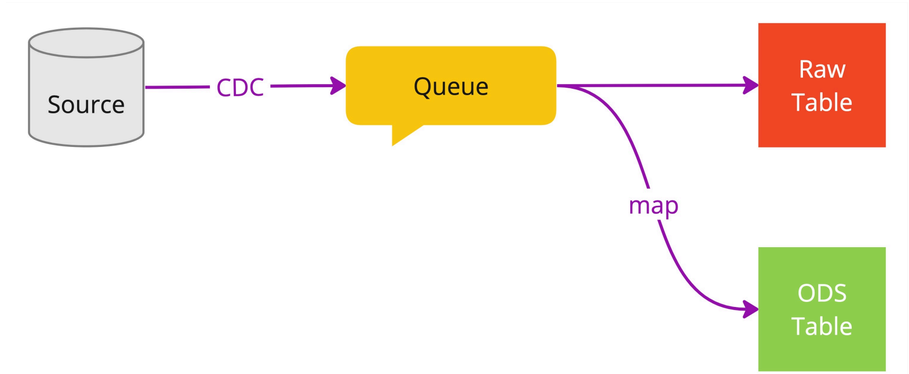


Проблемы пересчета истории:

- Прогонять все через очередь – долго и может тормозить поставку свежих данных
- Источник может не хранить историю
- Могут быть проблемы с огромной пиковой нагрузкой

MapReduce в 2023?!.. Для пересчета истории

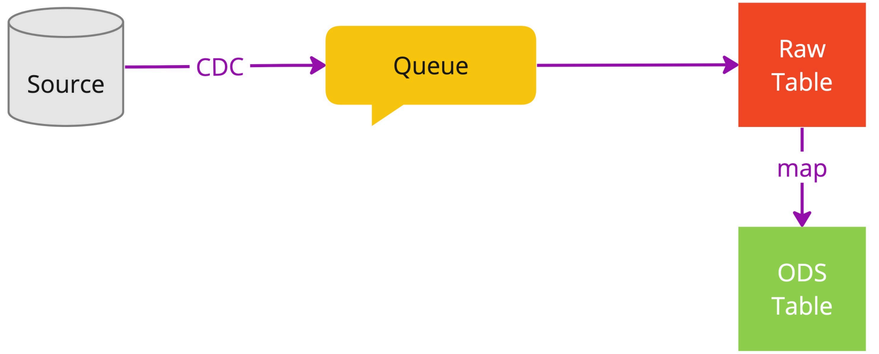
Потоковая поставка данных в Raw и ODS



Проблемы пересчета истории:

- Прогонять все через очередь – долго и может тормозить поставку свежих данных
- Источник может не хранить историю
- Могут быть проблемы с огромной пиковой нагрузкой

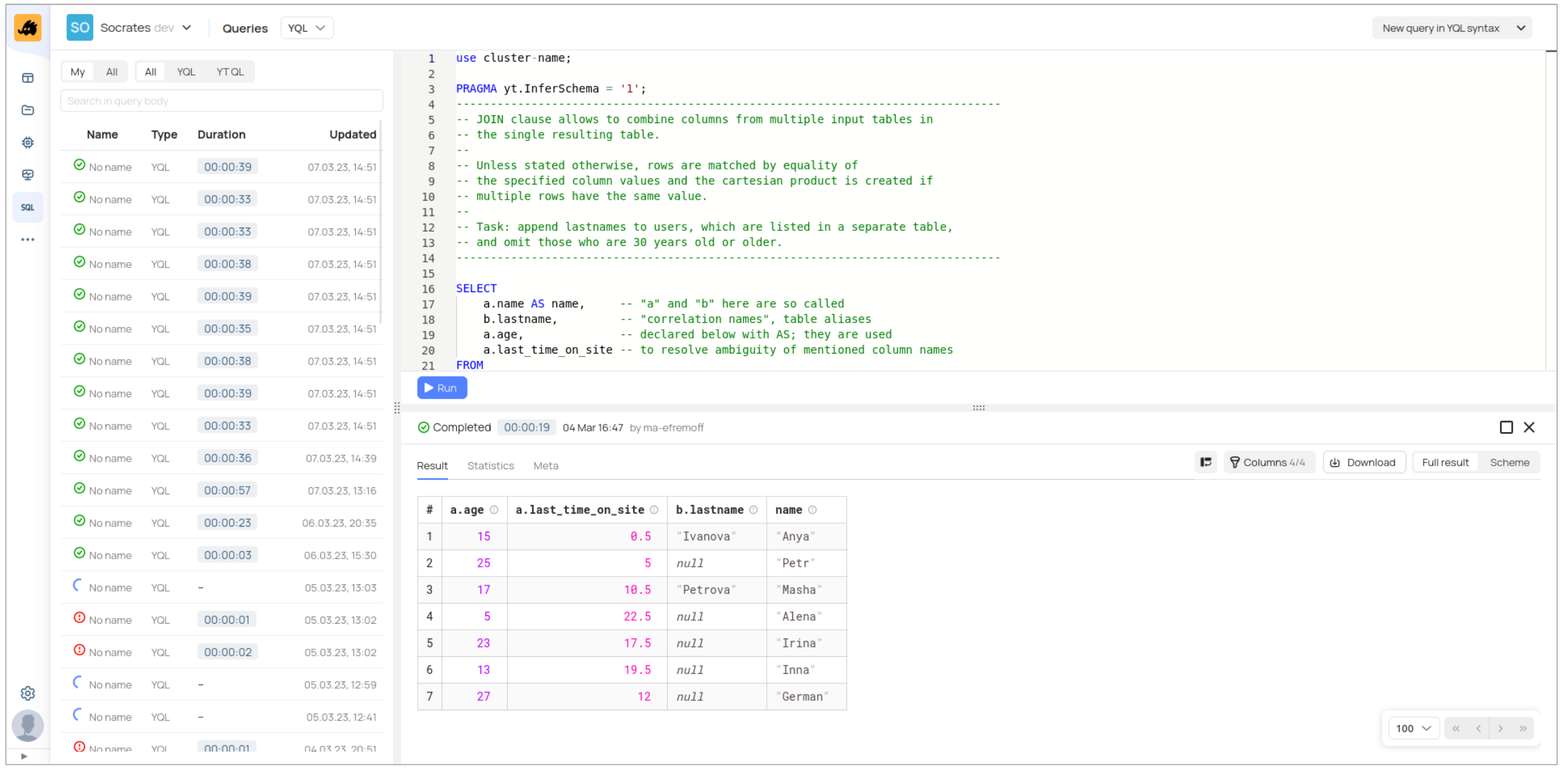
Backfill ODS



Почему Map из Raw данных удобен?

- Нет необходимости останавливать текущую поставку*
- Нет зависимости на источник данных
- Хорошо масштабируется, нативная поддержка Python, Java, Go и C++

YQL



https://ytsaurus.tech/docs/ru/yql/

Over MapReduce

Построение графа вычислений на основе MapReduce операций

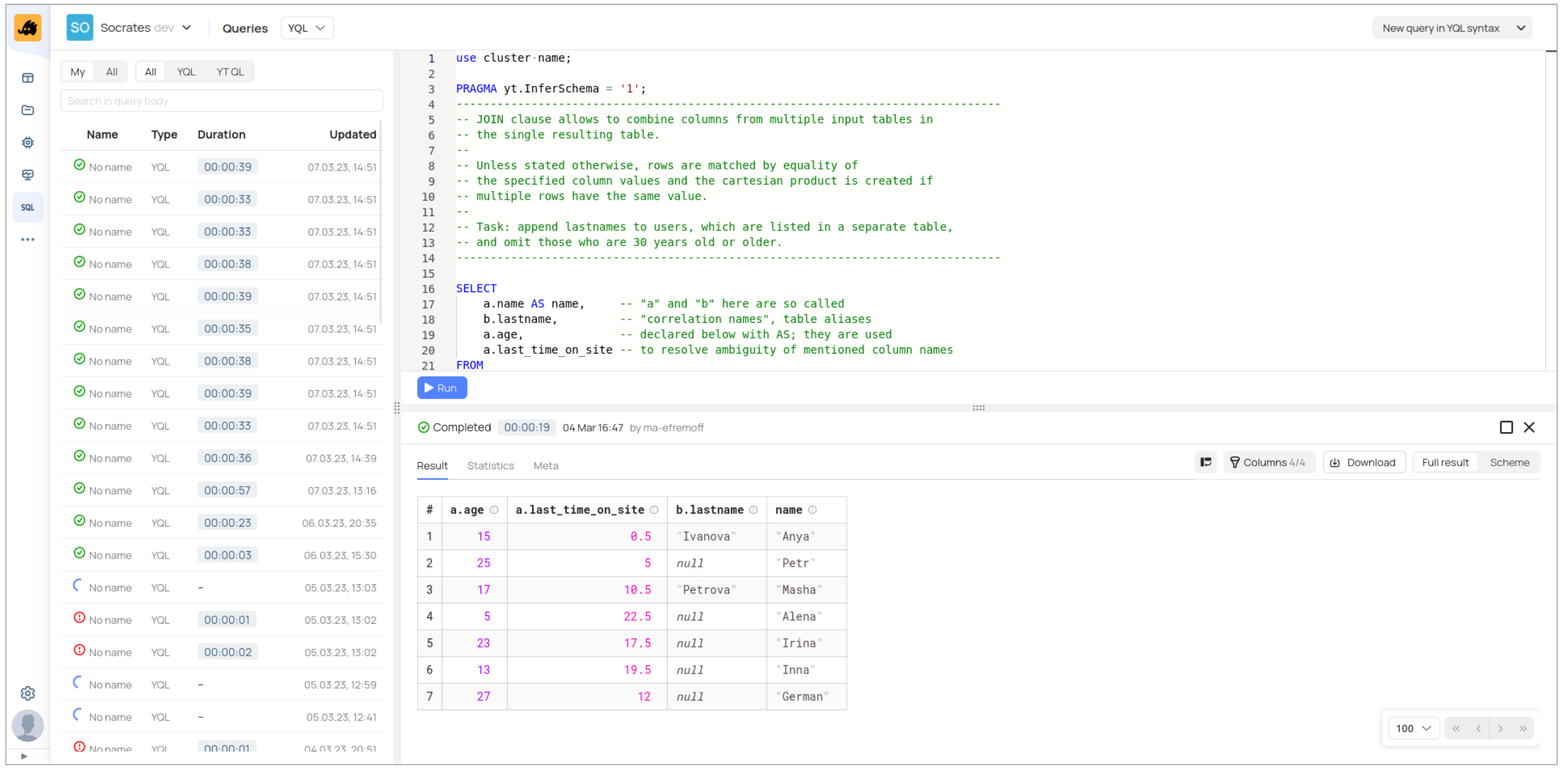
In-memory

При определенных условиях весь граф вычислений или его часть может быть выполнена в памяти без MapReduce операций

UDF

- C++
- Python
- JavaScript

YQL. MapReduce ≠ плохо



Clickstream

- Очень много данных
- На первом этапе происходит простая обработка: фильтрация и очистка

«Большие» логи

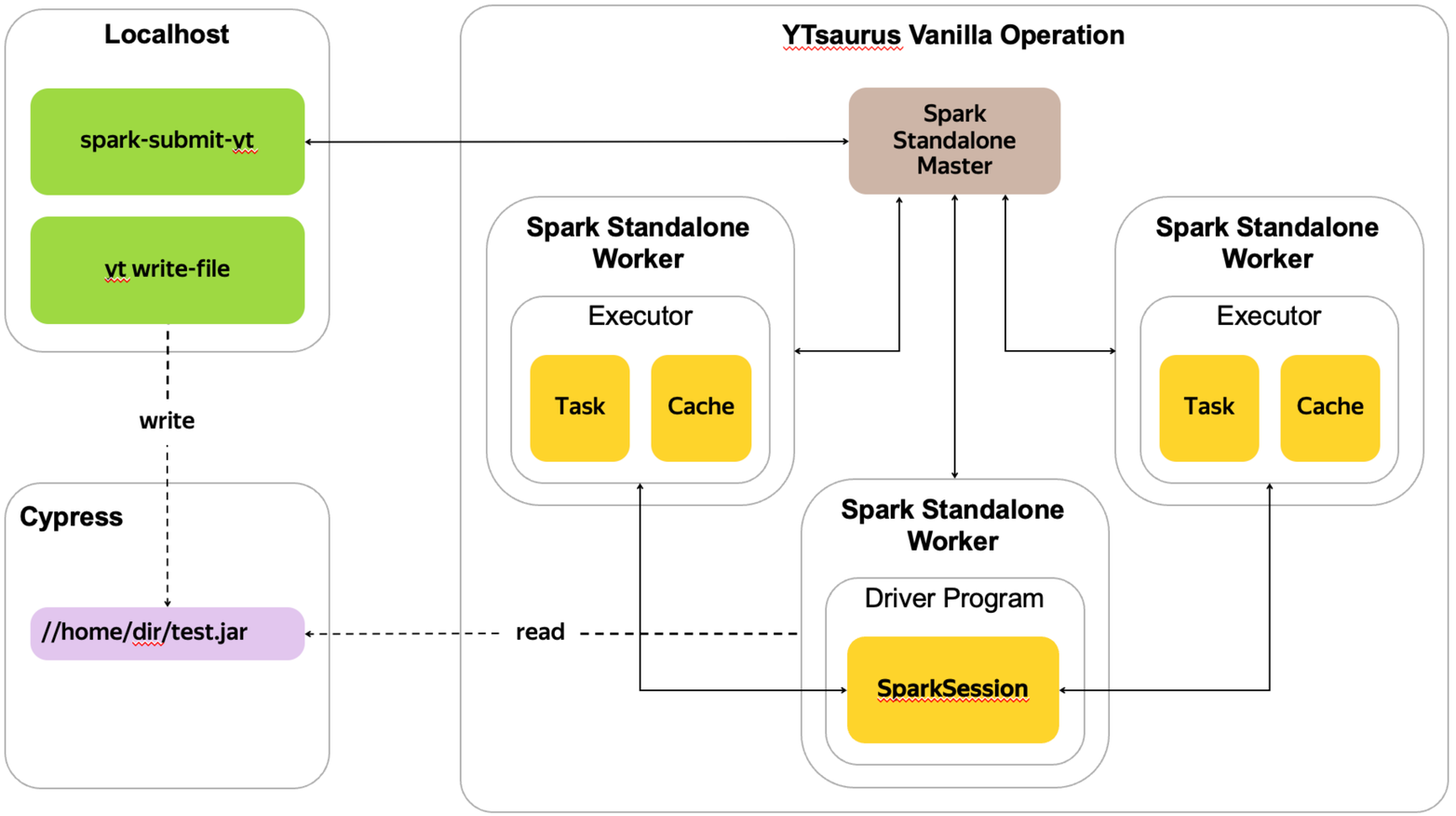
- YQL написан на C++ и весьма эффективен
- YQL пользуется фичами YT. Например, MapReduce вместо Map→Sort→Reduce

Витрины

- Простота разработки
- Хорошо масштабируется
- Часть вычислений может быть выполнена In-Memory

https://ytsaurus.tech/docs/ru/yql/

Spark over YTsaurus



https://ytsaurus.tech/docs/ru/user-guide/data-processing/spyt/overview

Fork, но…

- Все изменения касаются только работы с YT и внутри YT
- Доступна вся функциональность «ванильного» Spark-a

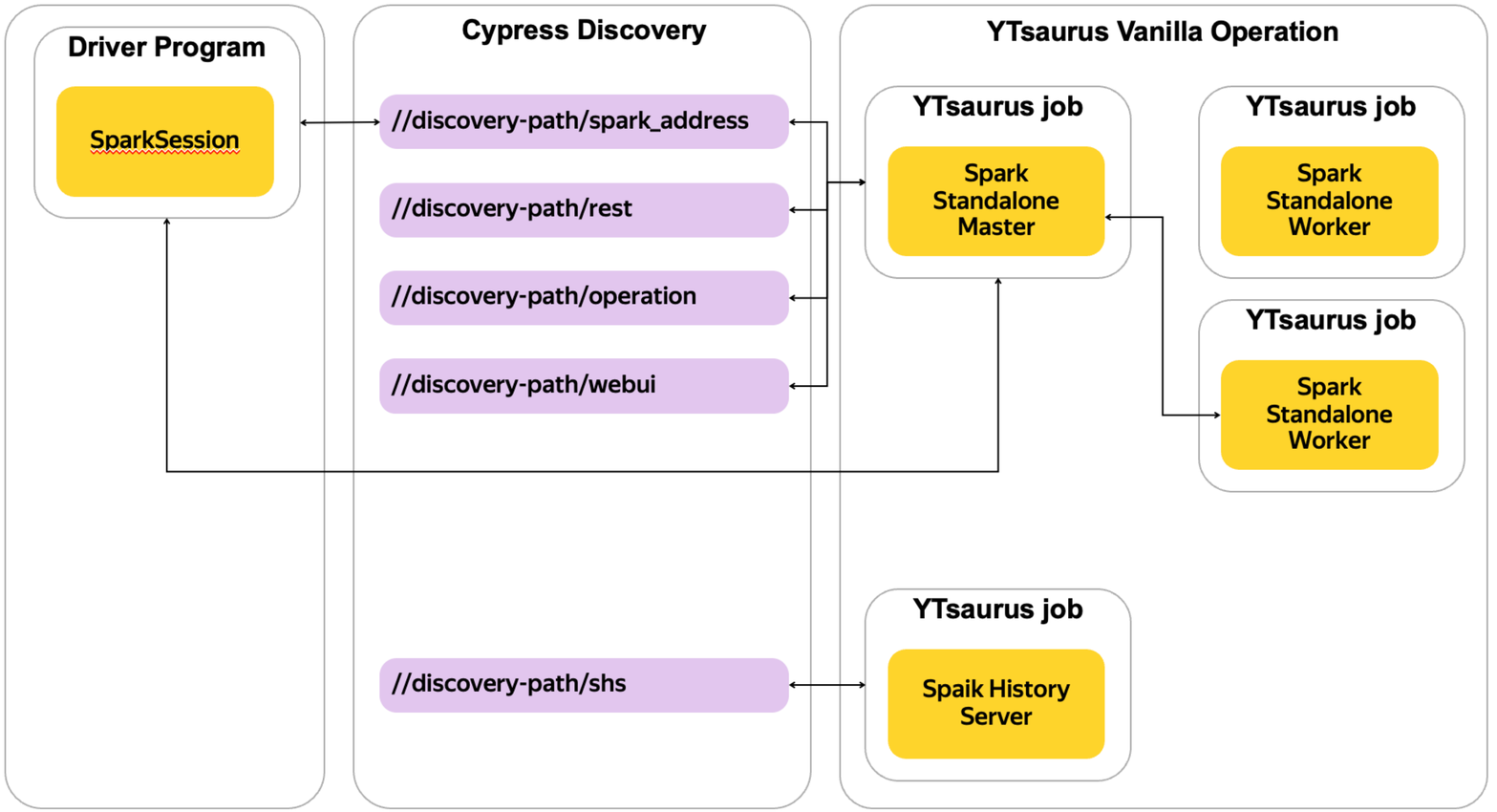
In-Memory

- Чем сложнее расчет, тем больше выгоды ускорения по сравнению с MapReduce
- Spill на SSD или в tmpfs

Внутри YTsaurus

- Легко переиспользуются вычислительные ресурсы между разнымидвижками
- Нет необходимости поддерживать несколько систем

Spark over YTsaurus. Client Mode



Саша Белоусова про подключение Spark к YTsaurus: https://highload.ru/spring/2021/abstracts/7266
Леша Шишкин про оптимизации в SPYT: https://highload.ru/spb/2023/abstracts/10222

Витрины

- Сложная логика: join-ы, аггрегации, UDF и т.п.
- Идеально, когда данные помещаются в память

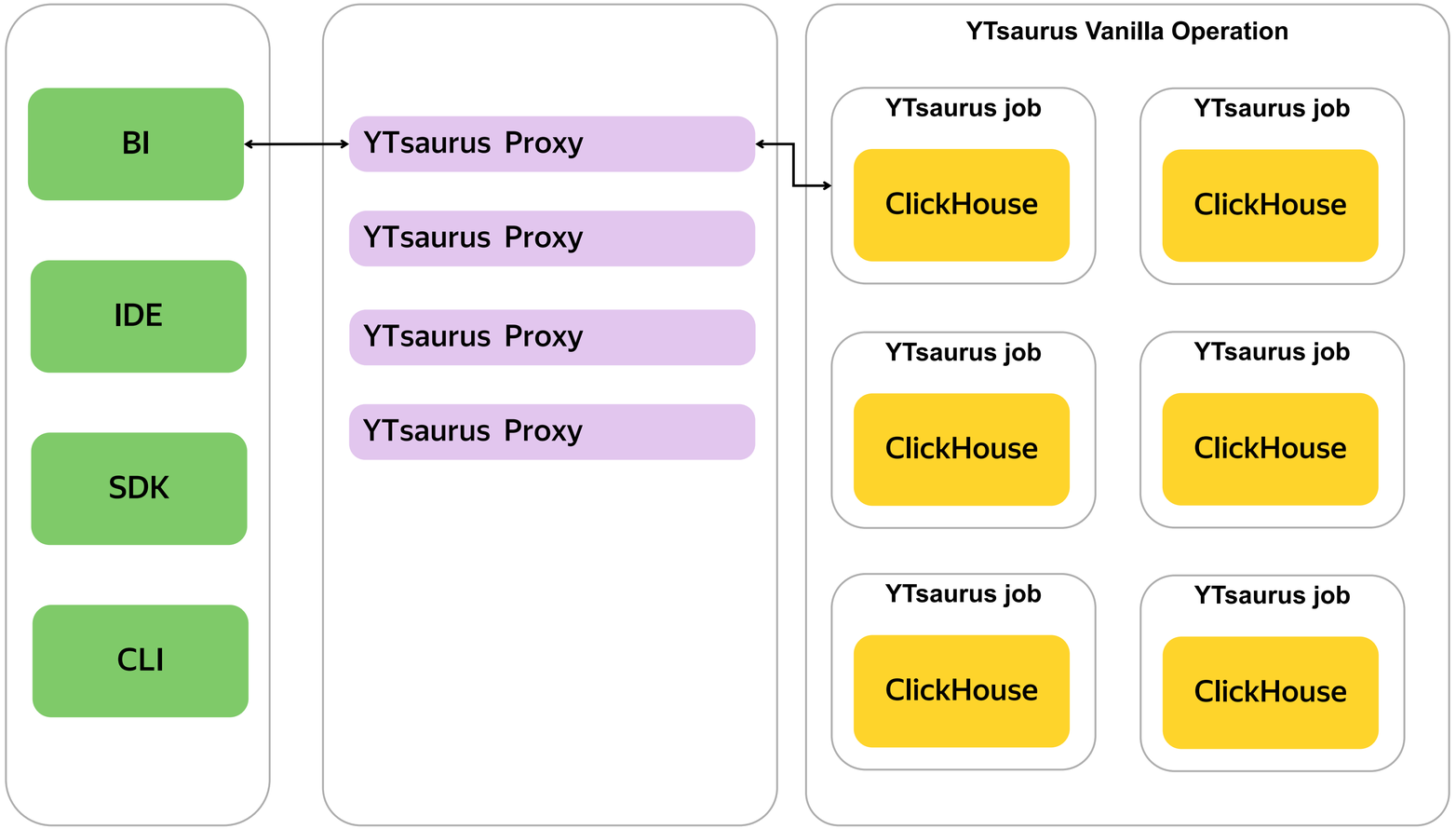
Ad-hoc

- PySpark прекрасен в Jupyter

SQL

BI over YTsaurus?..

ClickHouse over YTsaurus



https://ytsaurus.tech/docs/ru/user-guide/data-processing/chyt/about-chyt

BI

Любой BI с поддержкой коннекторов:

- ClickHouse
- odbc
- jdbc

Ad-hoc

- Скорость работы уступает ClickHouse, но не сильно
- Значительно быстрее YQL и Spark*

Простота

- Не надо копировать данные
- Не надо поддерживать отдельные системы
- Нужно только выделить вычислительные ресурсы

… over YT???!!!111

Планировщик и пулы

Квоты:

- CPU (Strong, Relaxed Integral, Burst Integral)
- RAM

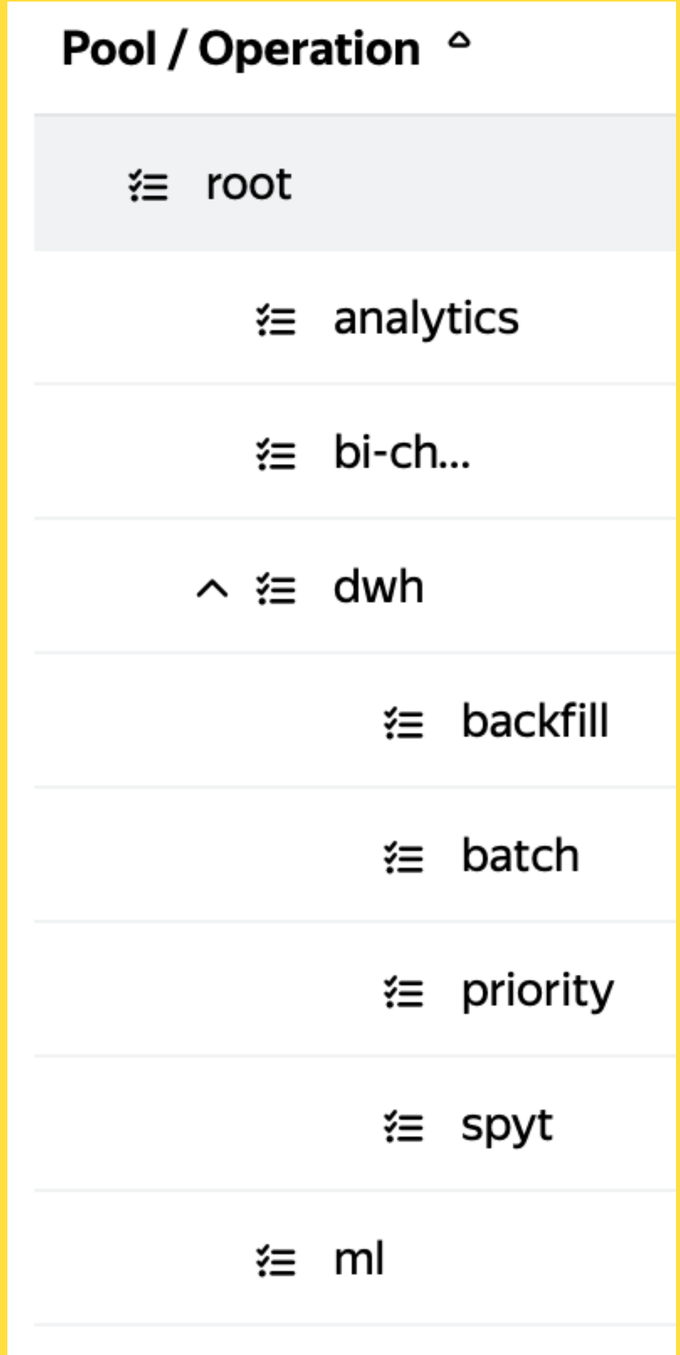
Особенности:

- Fair Share & Preemption внутри операций пула, внутри поддерева пулов и в кластере в целом
- Поддержка MapReduce: Map, Reduce, MapReduce, Sort, Merge, Erase
- Копирование данных между кластерами: RemoteCopy
- Запуск произвольной нагрузки: Vanilla

Параметры:

- Weight
- Max & Running Operation Count
- …

https://ytsaurus.tech/docs/ru/user-guide/data-processing/scheduler/scheduler-and-pools



YTsaurus vs …

- Cypress with static row- & column-oriented tables & ACID transactions
- Sorted Dynamic Tables
- Ordered Dynamic Tables

- hdfs, s3 with Parquet, ORC, Avro

- HBase
- Kafka

- YQL
- Spark over YT
- ClickHouse over YT

- Hive
- Spark
- Impala, Presto, Trino, …

- Fair Share & Preemption планировщик

- YARN или даже Kubernetes

YTsaurus vs …

- Готовое коробочное решение
- Enterprise level: масштабируемость, надежность, ролевая модель, изоляция хранения и вычислений, красивый UI, …

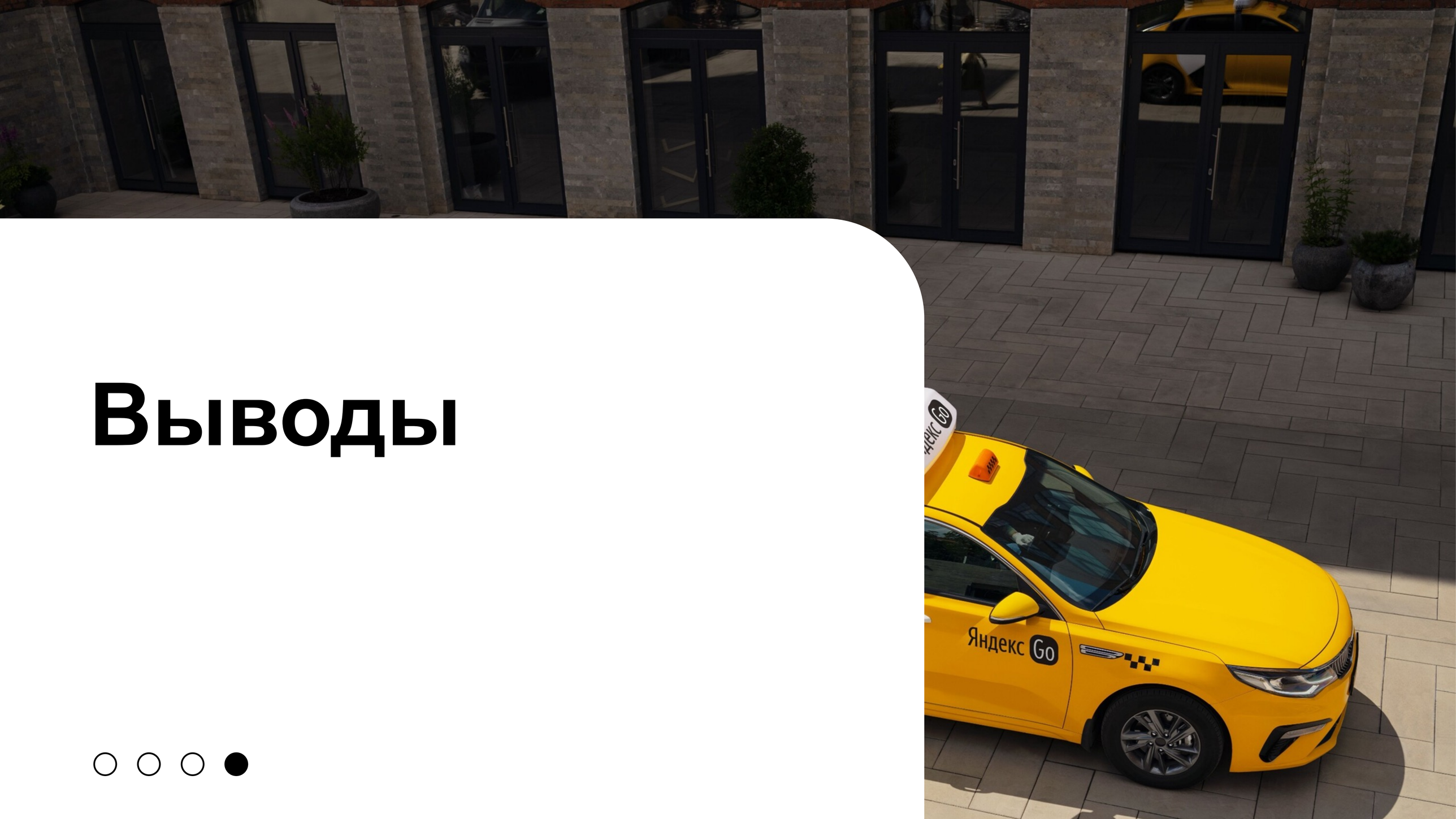
- Конструктор с кучей кубиков
- Нужен напильник ☺ или покупка готовой сборки с батарейками

YTsaurus vs …

- Готовое коробочное решение
- Enterprise level: масштабируемость, надежность, ролевая модель, изоляция хранения и вычислений, красивый UI, …

- Конструктор с кучей кубиков
- Нужен напильник ☺ или покупка готовой сборки с батарейками

Выводы



Все доступно
Все Open Source

1. YTsaurus
   https://ytsaurus.tech/
2. Greenplum
   https://greenplum.org
3. ClickHouse
   https://clickhouse.com

YTsaurus можно попробовать

1. Docker
2. Demo-стенд
3. Kubernetes

https://ytsaurus.tech/docs/ru/overview/try-yt

Как строить платформу

1. Опираться на накопленный опыт и best practices
2. Идти от требований
3. Экспериментировать в поисках лучшего решения

Яндекс Go

Спасибо за внимание!

Вопросы?