

Познакомимся?



РТУ МИРЭА

BigData в реальных проектах

Сысоев Кирилл Романович

Обо мне

4.5+ лет в BigData
HSE University

Senior Data Engineer
1) GlowByte Consulting
2) PochtaTech
3) OneFactor

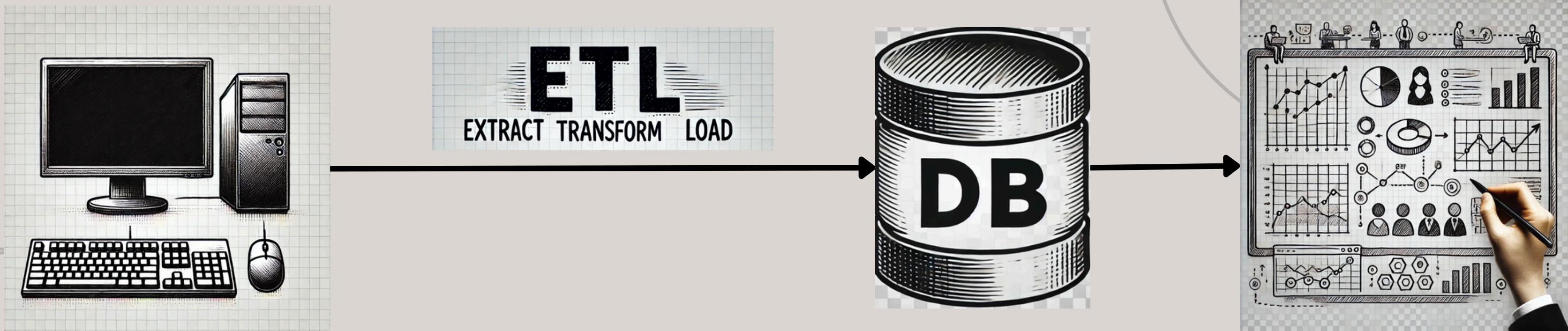
Hadoop, Spark, Kafka, k8s, YandexCloud
Python/Scala, SQL



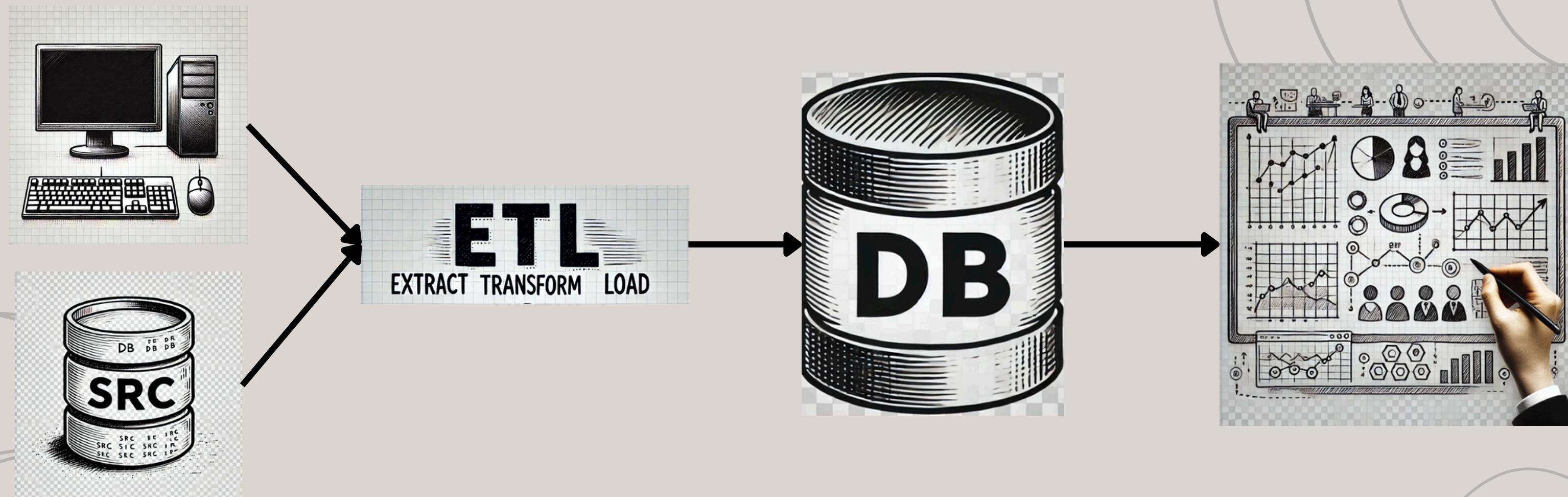
t.me/KRSysoev
ksysoev@hse.ru

Что такое BigData

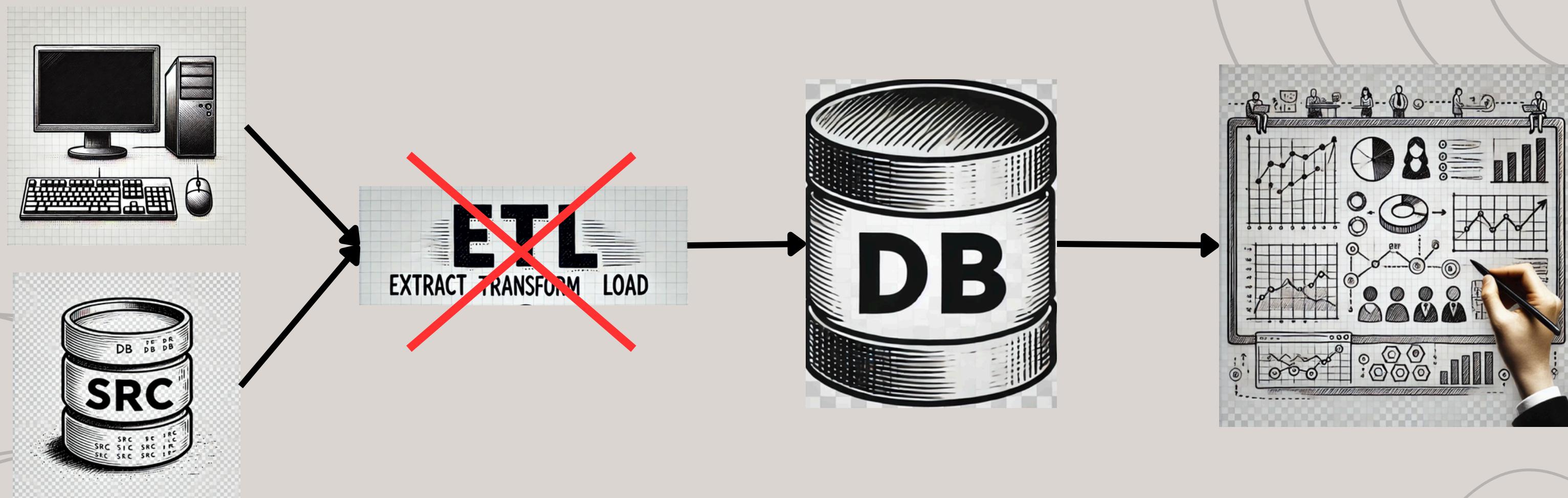
Что такое BigData



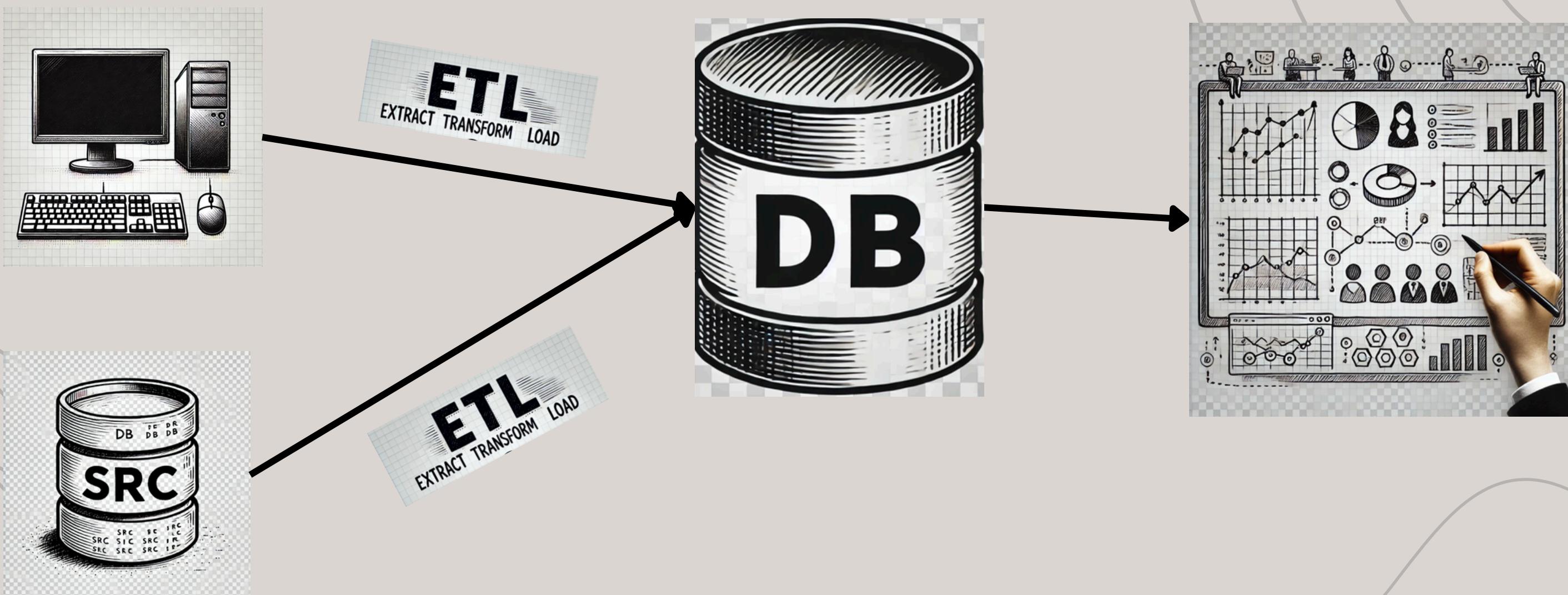
Что такое BigData



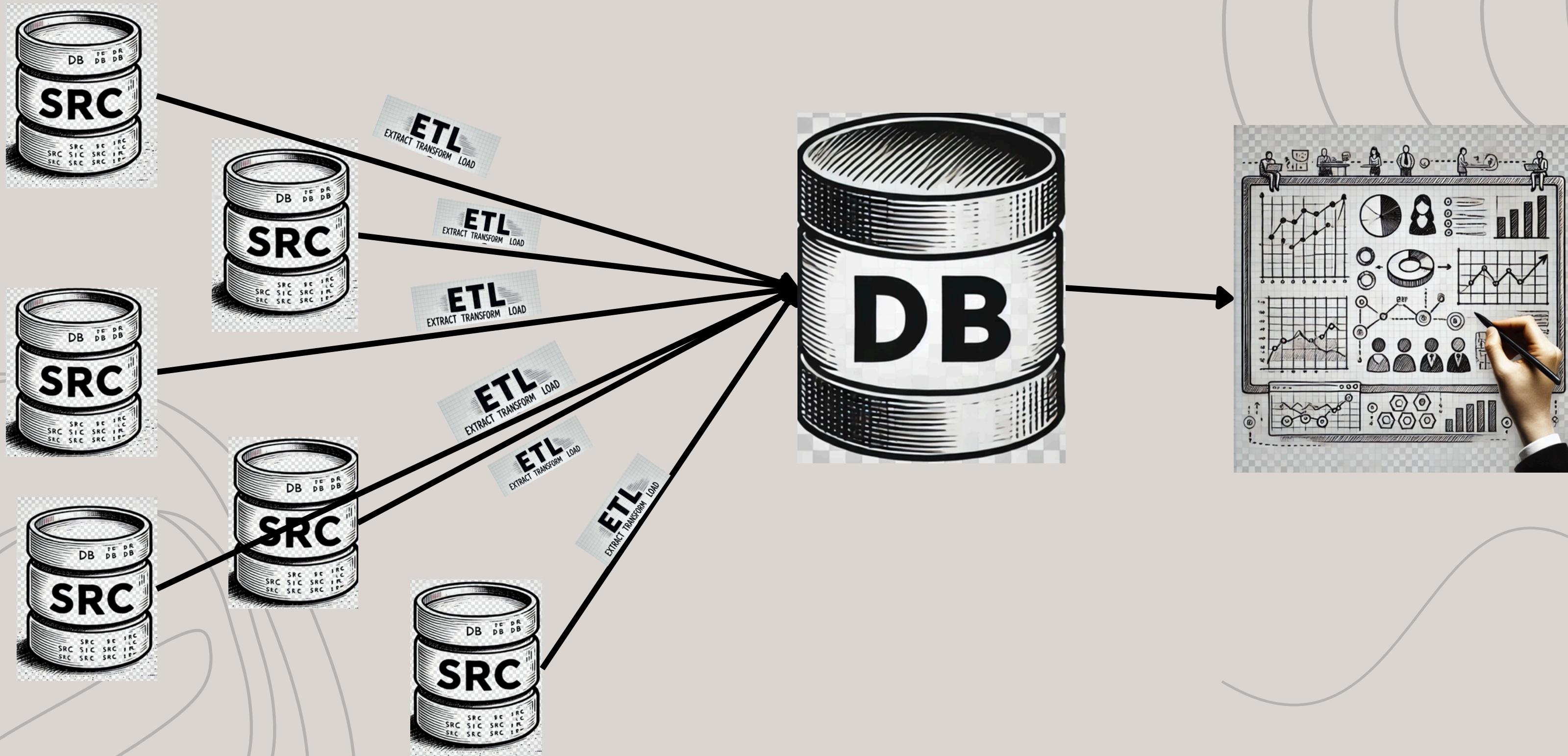
Что такое BigData



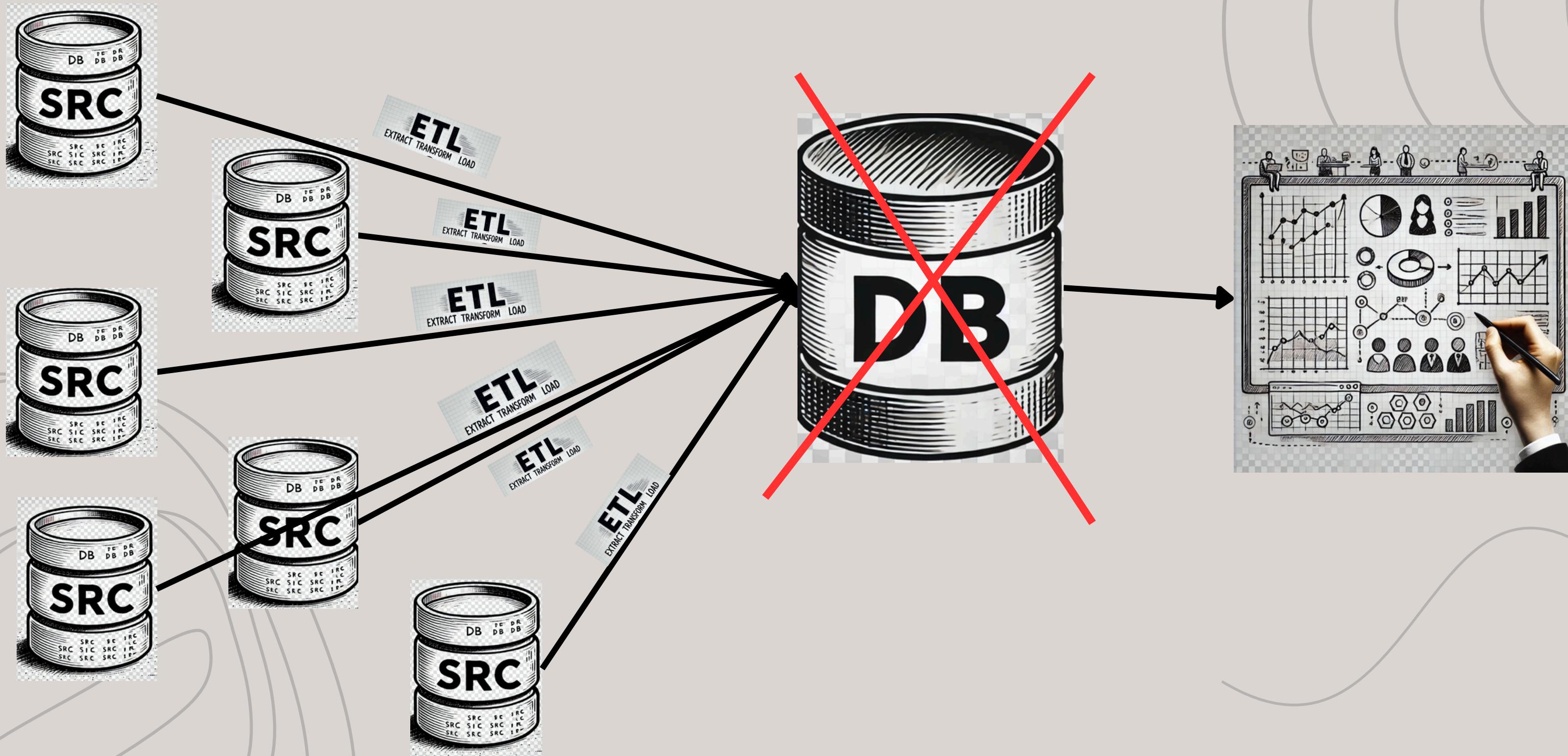
Что такое BigData



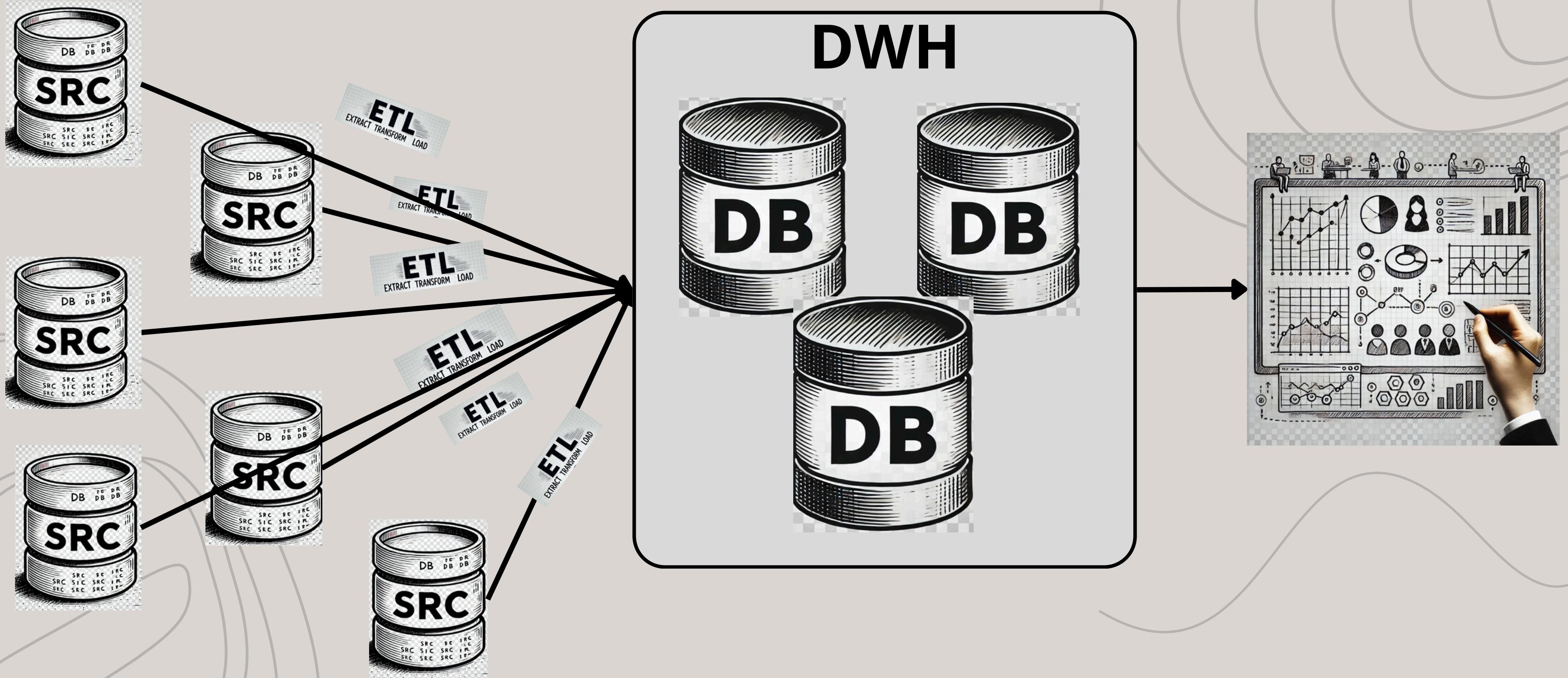
ЧТО ТАКОЕ BigData



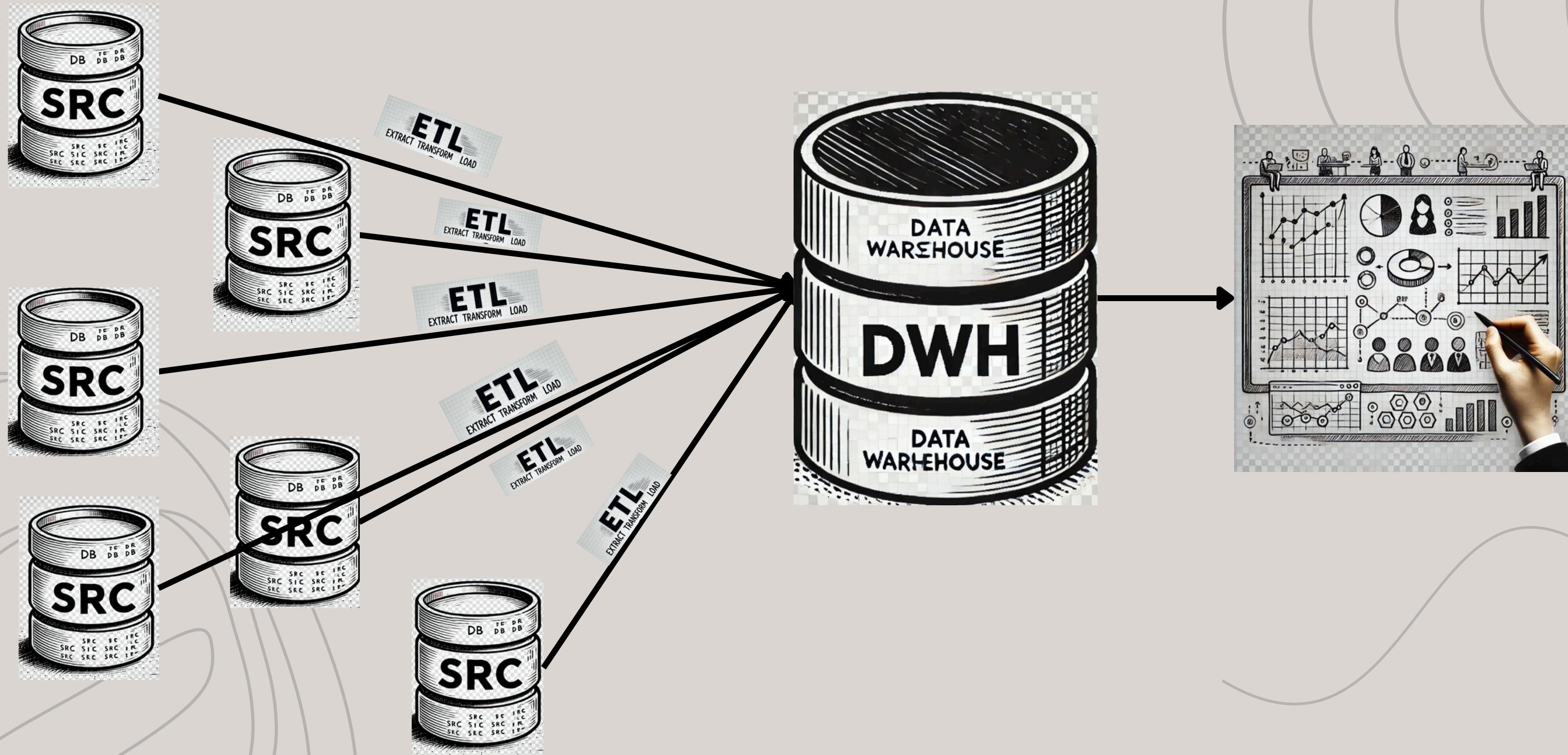
ЧТО ТАКОЕ BigData



ЧТО ТАКОЕ BigData



ЧТО ТАКОЕ BigData

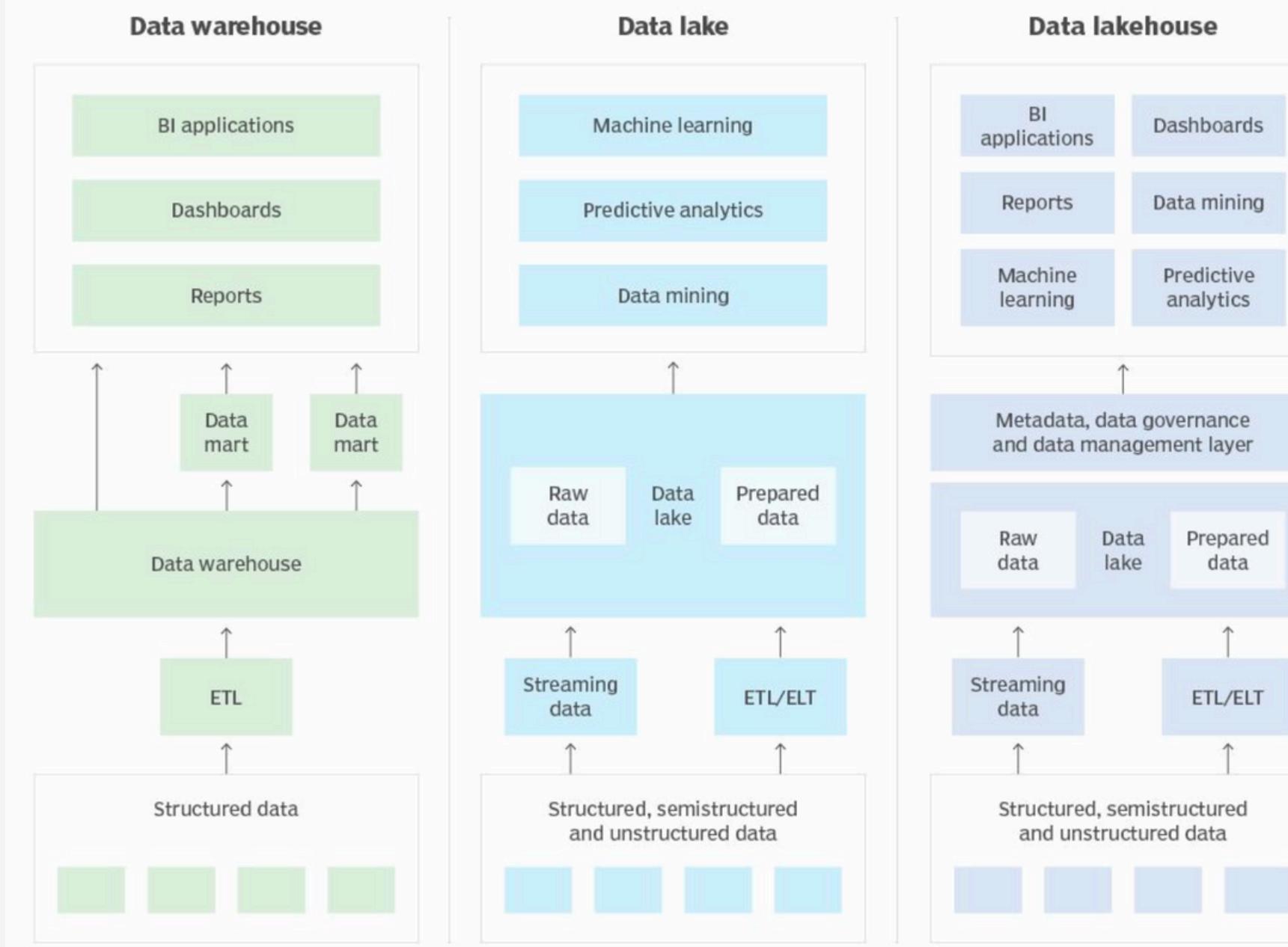


Что такое BigData

DWH (Data Warehouse) – это хранилище данных, которое используется для хранения большого объема структурированной и иногда полуструктурной информации, предназначенный для анализа и принятия решений на основе данных. Основная цель DWH – интеграция данных из различных источников для проведения глубокого анализа и отчетности.

Что такое BigData

Data warehouse vs. data lake vs. data lakehouse

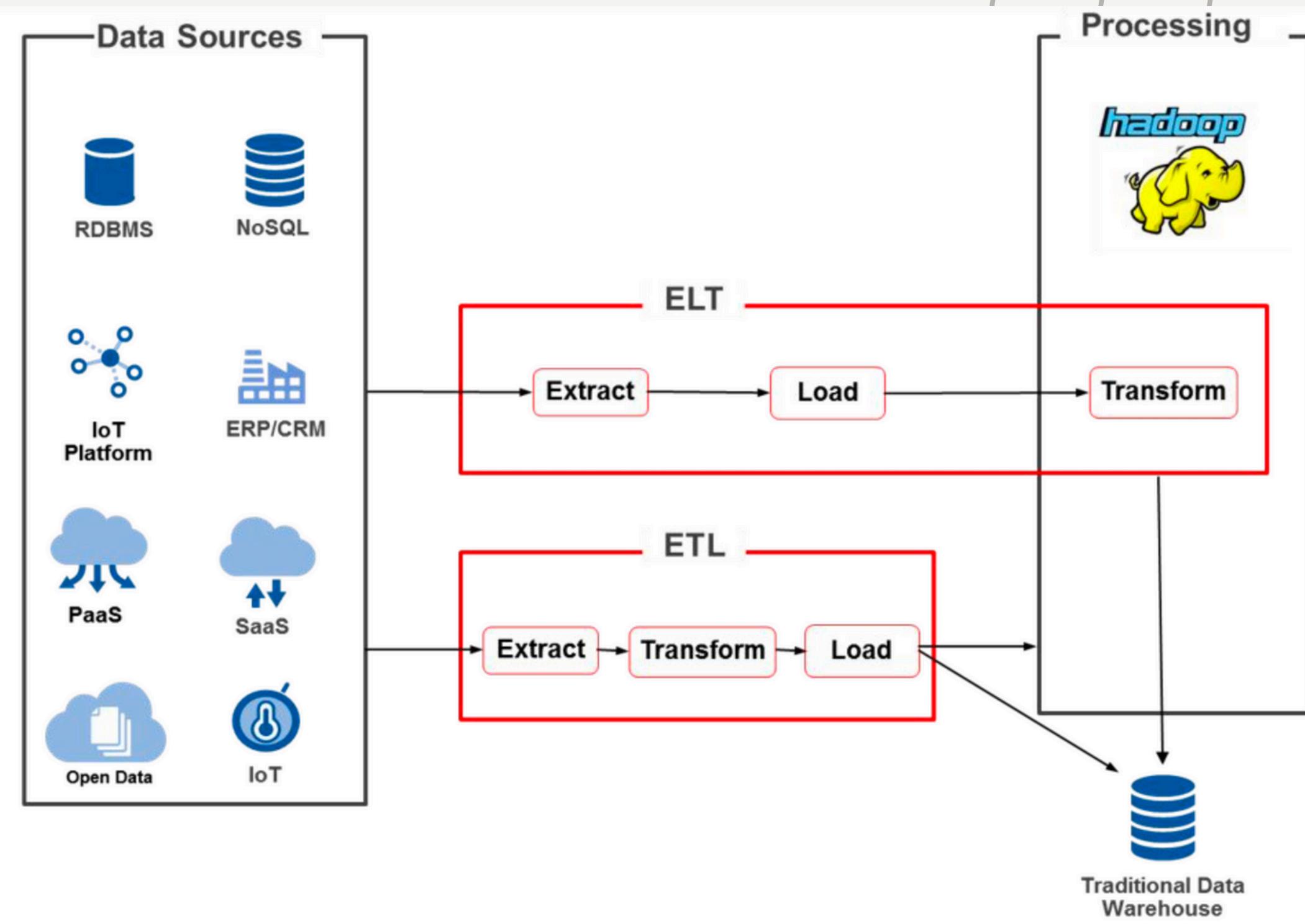


Что такое BigData

ETL

VS.

ELT



Что такое BigData

01

Характеристики (3V):

- Volume (Объем)
- Velocity (Скорость)
- Variety (Разнообразие)

02

Важность для бизнеса (Netflix):

- 1) Сбор данных
- 2) Анализ данных
- 3) Персонализированные рекомендации
- 4) Создание контента

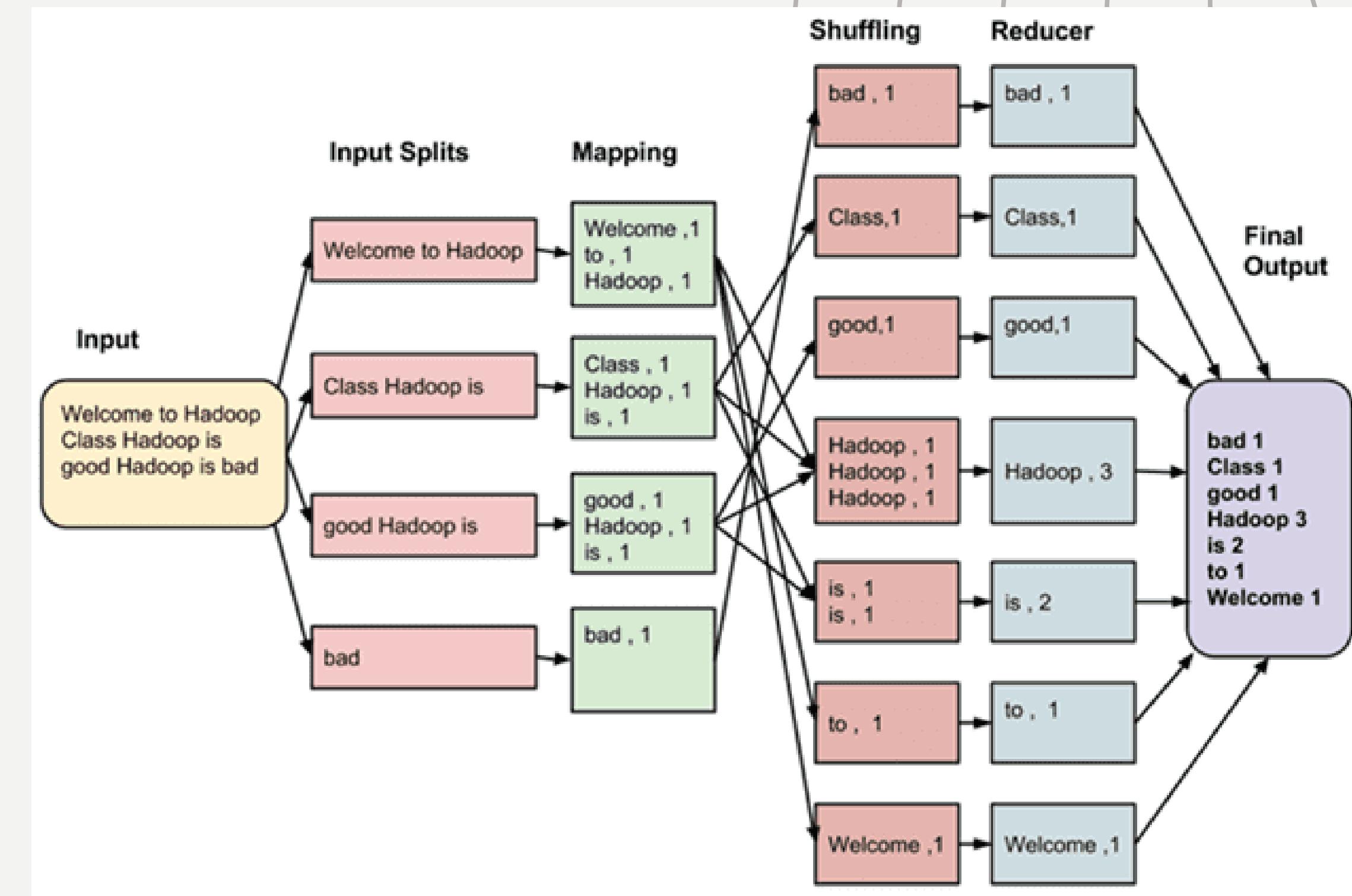
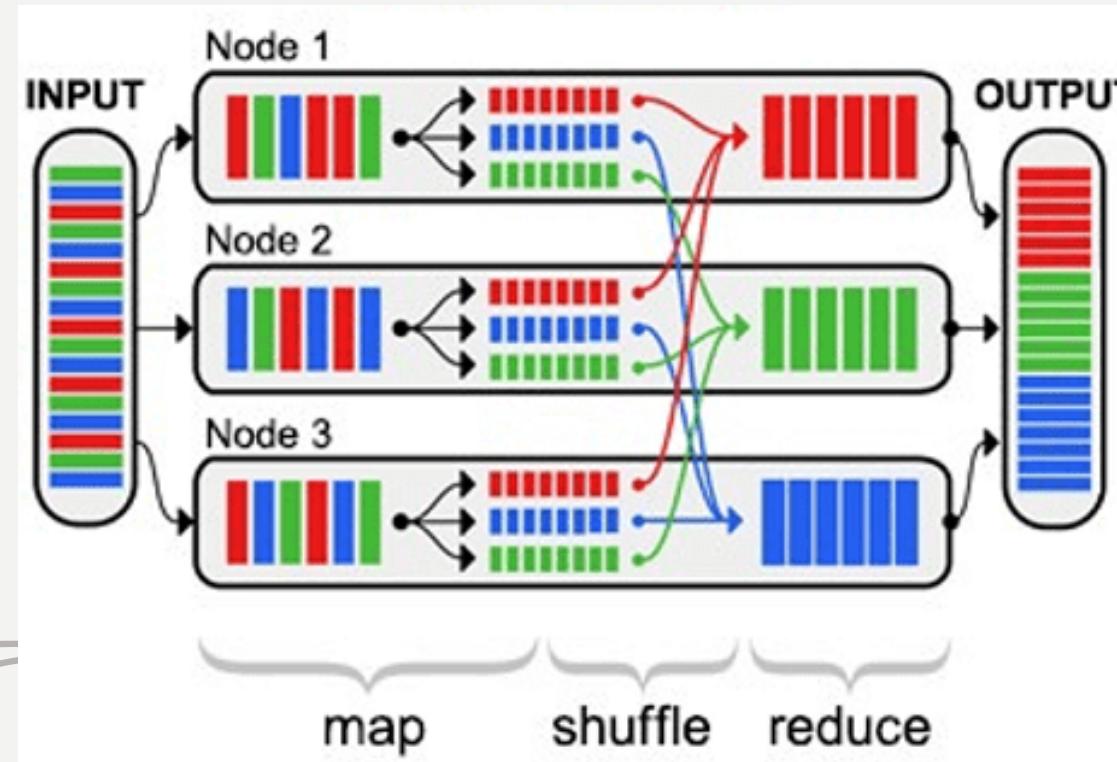
- Повысились удержание пользователей
- Повысилась вовлеченность аудитории
- Экономия на маркетинге и разработке контента

Предпосылки, свойства и примеры

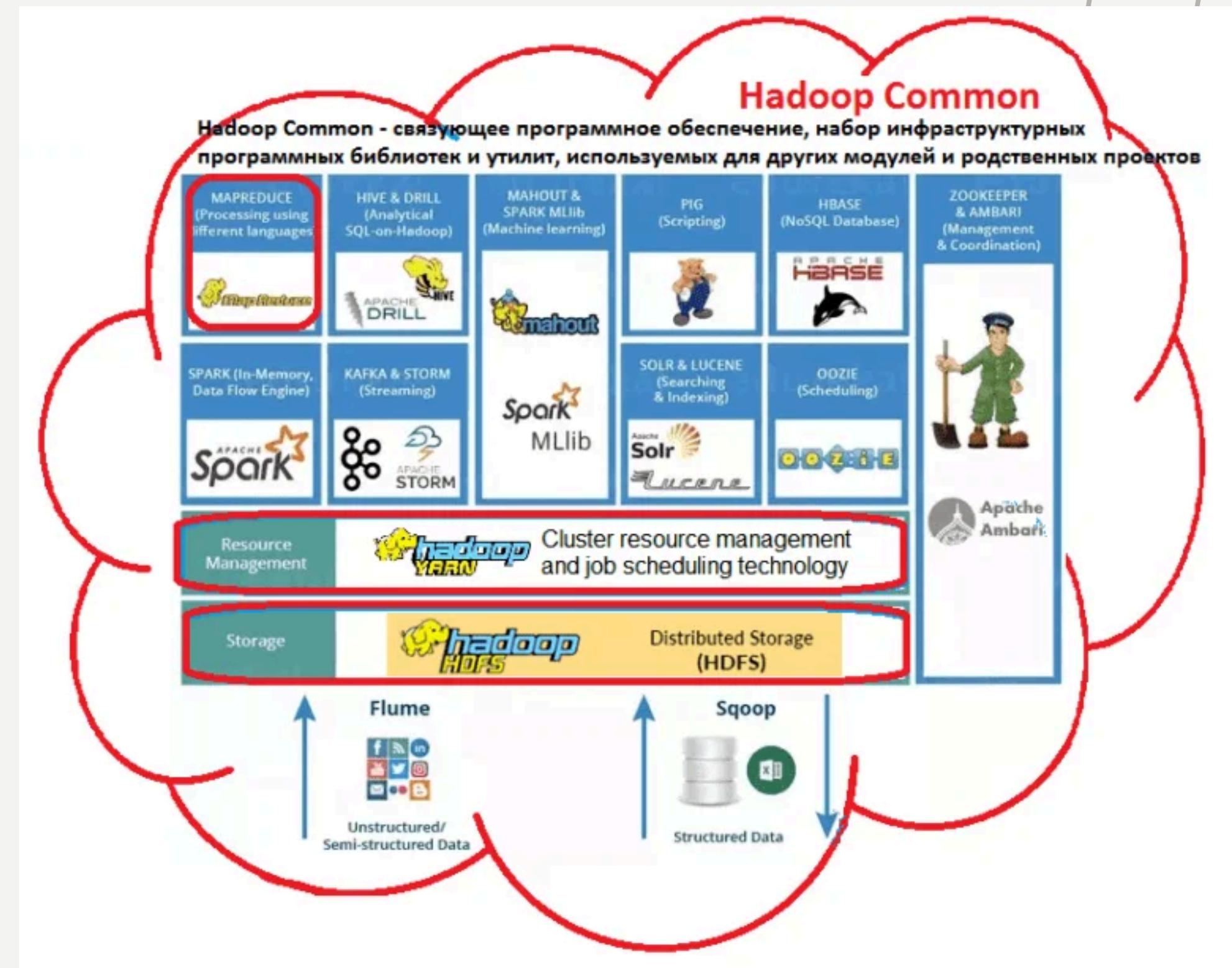
Технологии

BigData

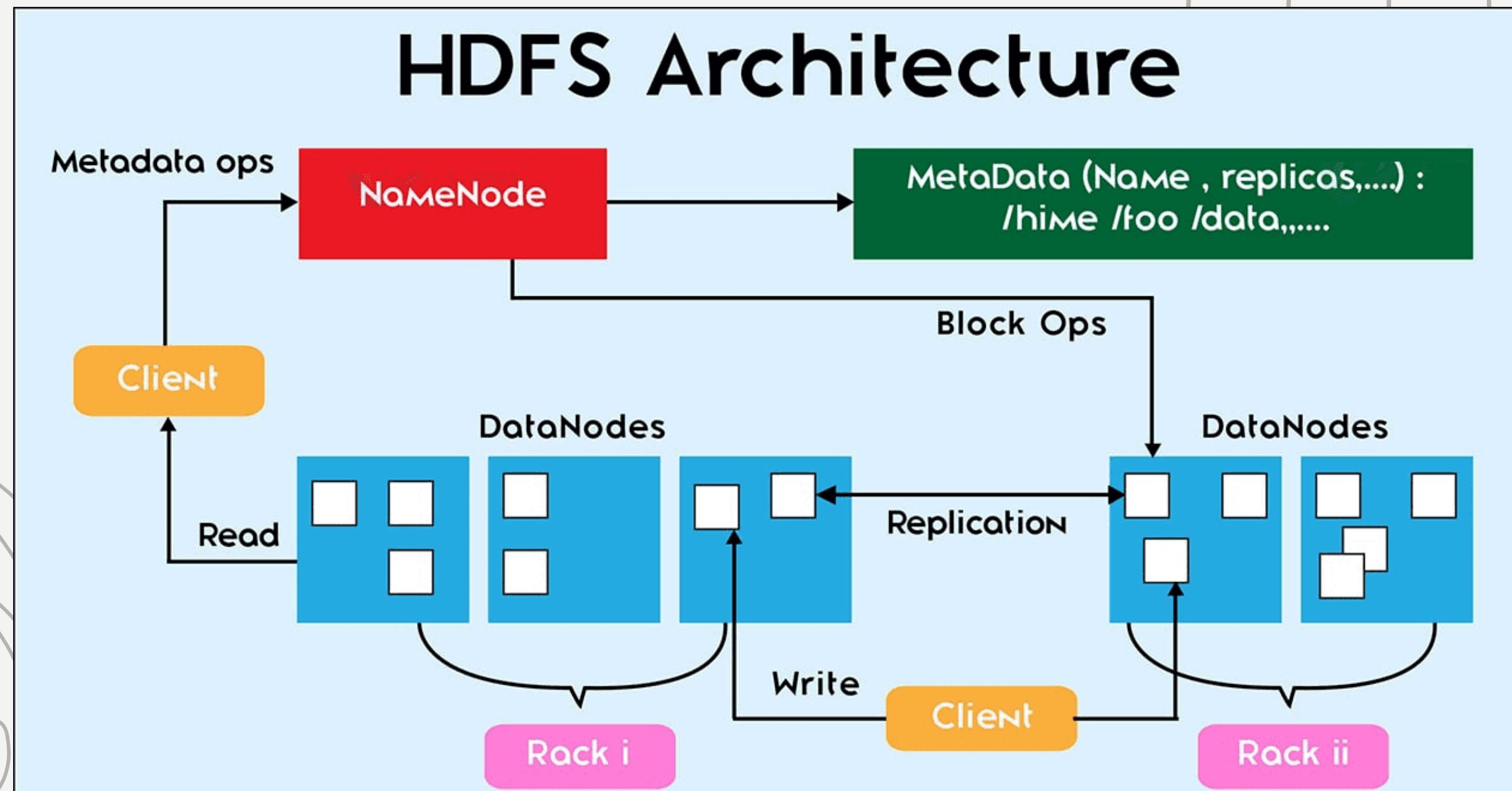
MapReduce



Hadoop



HDFS

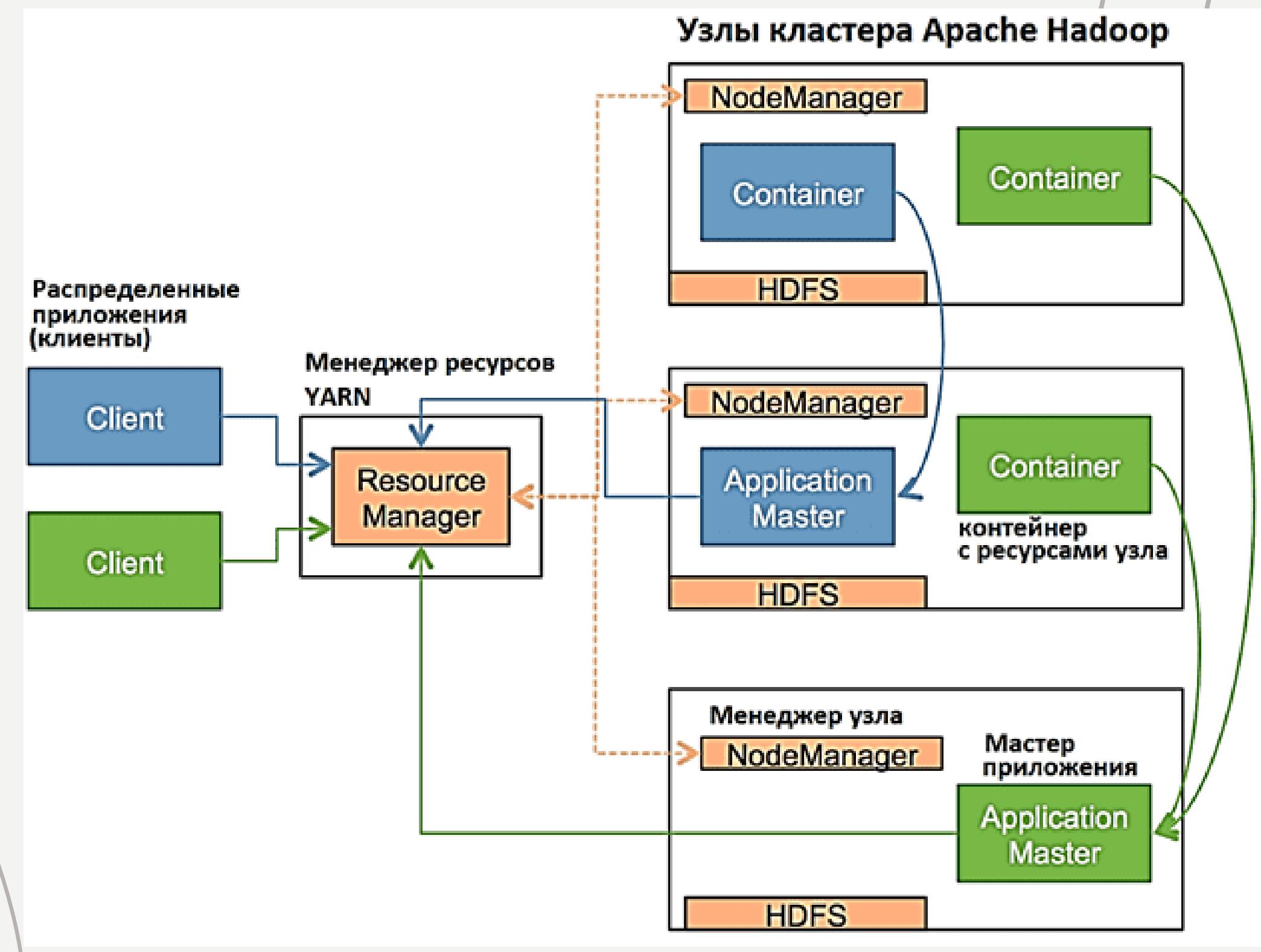


Hive

The screenshot shows the Hue web interface for managing big data. At the top, there's a navigation bar with the Hue logo, a 'Query' dropdown, and a search bar. Below the navigation is a toolbar with icons for Databases, Clusters, Search, and Jobs. A sidebar on the left lists 'Sources' with 'Impala' and 'Hive' selected. The main area is a query editor with tabs for 'Editor' (selected) and 'Scheduler'. A dropdown menu is open over the 'Editor' tab, showing options: 'Editor' (selected), 'Impala' (highlighted with a red box), 'Hive' (selected), 'Pig', 'Java', 'Spark', 'MapReduce', 'Shell', 'Sqoop 1', and 'Distcp'. The code editor displays a series of SQL-like SELECT statements. The URL for this page is https://bigdataschool.ru/wiki/hive.

```
1
2 SELECT
3 from dr
4 where t
5 and t.m
6
7
8 SELECT
9
10 SELECT
11
12 SELECT
13 where m
14 and m.m
15 and m.r
16
17 SELECT * from DRIVER_TEST_OPTIONSIFY_RT_OS_RECVPT_ALL +
```

YARN



YARN (RM UI)

← → C psrlInfa.informatica.com:8088/cluster

 All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decom N
568	0	0	568	0	0 B	32 GB	0 B	0	32	0	1	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved
0	0	0	568	0	0	0	0 B	0 B	0 B

Show 20 ▾ entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores
application_1463379223882_0568	Idmui	InfaSprk0	SPARK	root.Idmui	Mon May 16 13:11:59 -0700 2016	Mon May 16 13:13:03 -0700 2016	FINISHED	SUCCEEDED	N/A	N/A
application_1463379223882_0567	Idmui	InfaSprk0	SPARK	root.Idmui	Mon May 16 13:11:58 -0700 2016	Mon May 16 13:13:01 -0700 2016	FINISHED	SUCCEEDED	N/A	N/A
application_1463379223882_0566	Idmui	InfaSprk0	SPARK	root.Idmui	Mon May 16 13:11:56 -0700 2016	Mon May 16 13:13:00 -0700 2016	FINISHED	SUCCEEDED	N/A	N/A
application_1463379223882_0565	Idmui	InfaSprk0	SPARK	root.Idmui	Mon May 16 13:11:55 -0700 2016	Mon May 16 13:12:59 -0700 2016	FINISHED	SUCCEEDED	N/A	N/A
application_1463379223882_0564	Idmui	InfaSprk0	SPARK	root.Idmui	Mon May 16 13:11:54	Mon May 16 13:12:59	FINISHED	SUCCEEDED	N/A	N/A

YARN (RM UI)

The screenshot shows the Hadoop Resource Manager (RM) User Interface. At the top, there's a logo of a yellow dog with the word "hadoop" next to it. Below the logo, the title "Application application_1484222082388_0001" is displayed. On the left, a sidebar titled "Cluster" lists various cluster metrics like About, Nodes, Node Labels, Applications (with sub-options NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and Scheduler. Below the sidebar is a "Tools" section. The main content area is titled "Kill Application" and displays detailed information about the application:

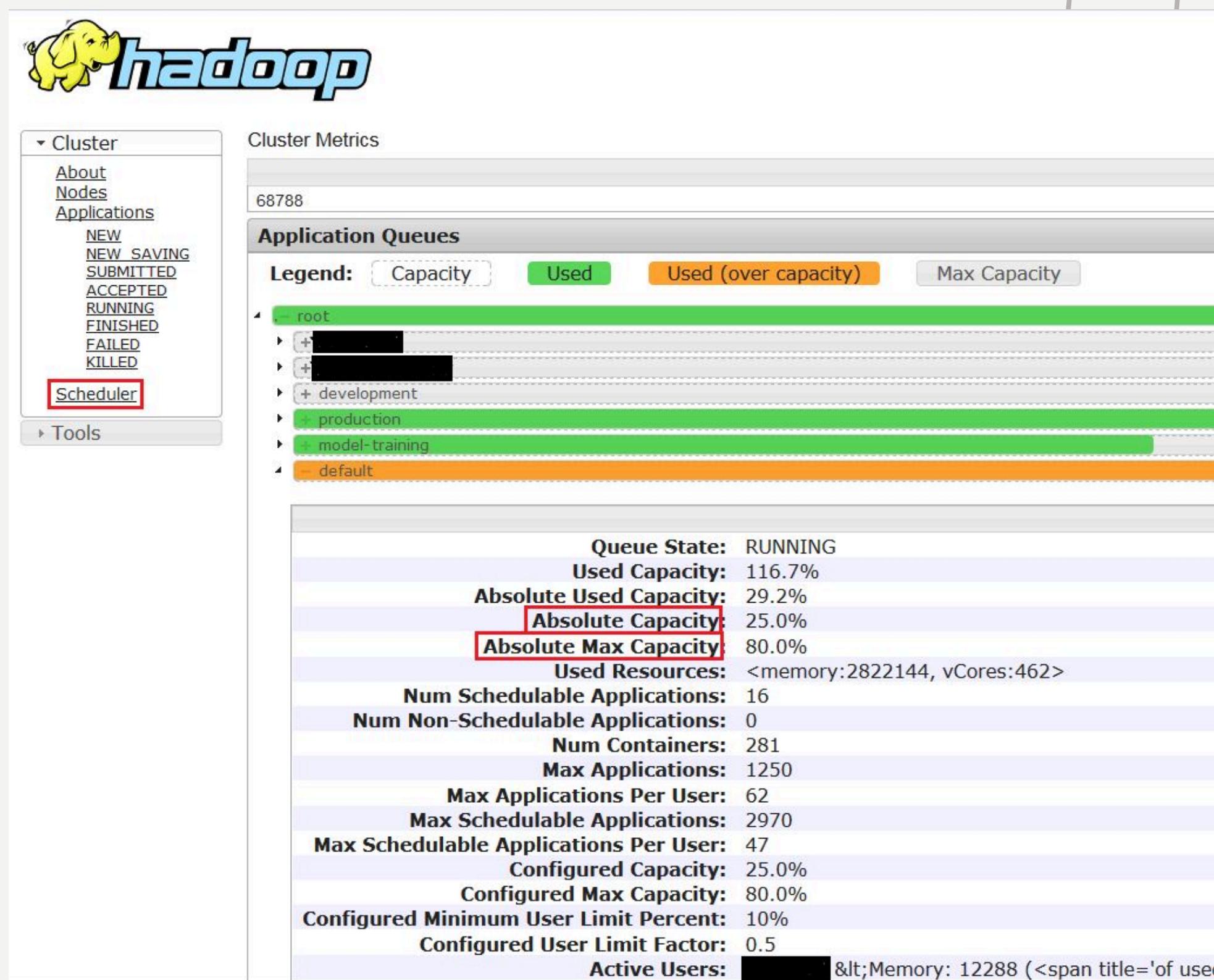
Application	
User:	hdfs
Name:	org.apache.spark.examples.SparkPi
Application Type:	SPARK
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Thu Jan 12 17:48:04 +0530 2017
Elapsed:	27sec
Tracking URL:	History
Log Aggregation Status:	SUCCEEDED
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Below this, there's another section titled "Application" with the following data:

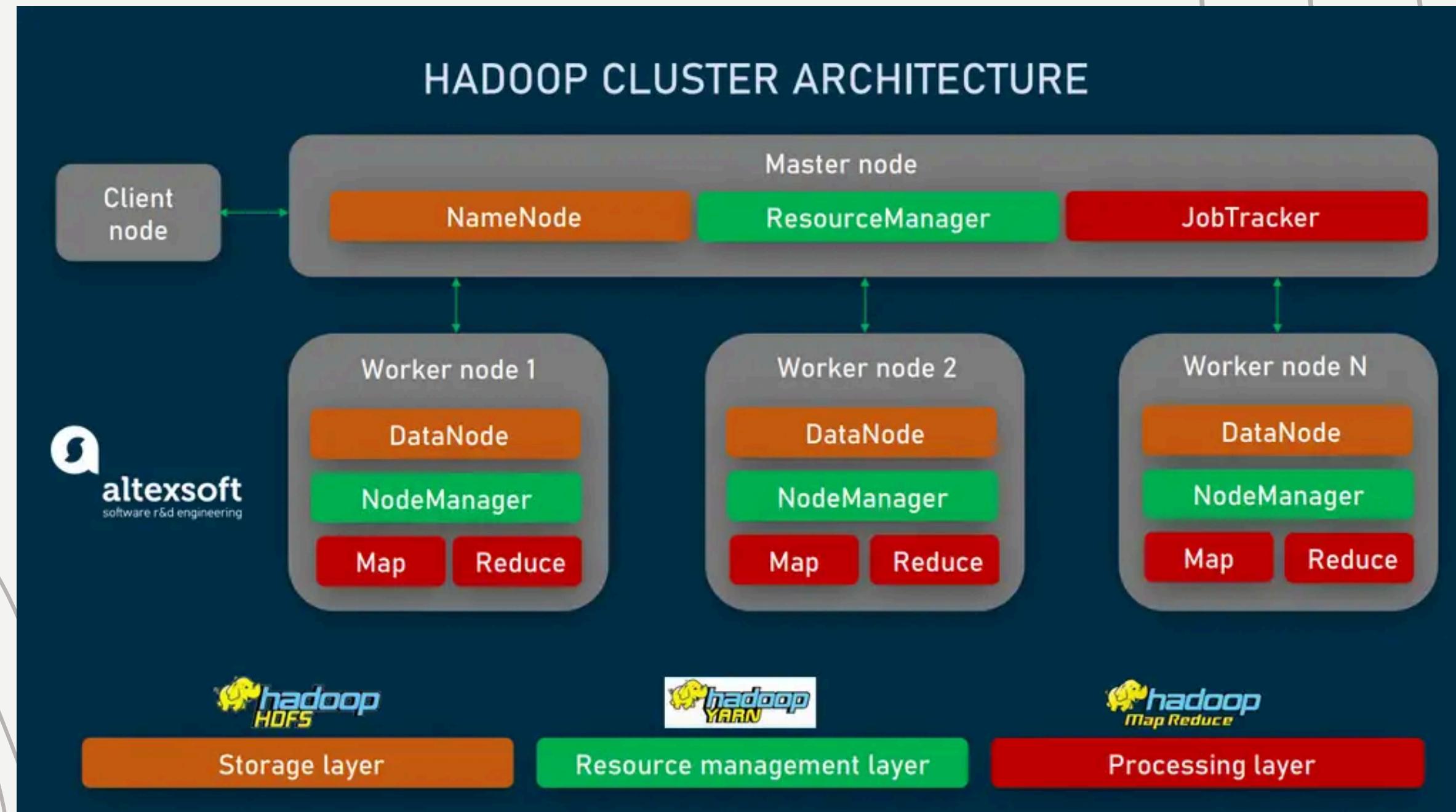
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	257551 MB-seconds, 76 vcore-seconds

At the bottom, there are navigation controls: "Show 20 entries", "Search:", and a table header with columns: Attempt ID, Started, Node, Load, and Blacklisted Node.

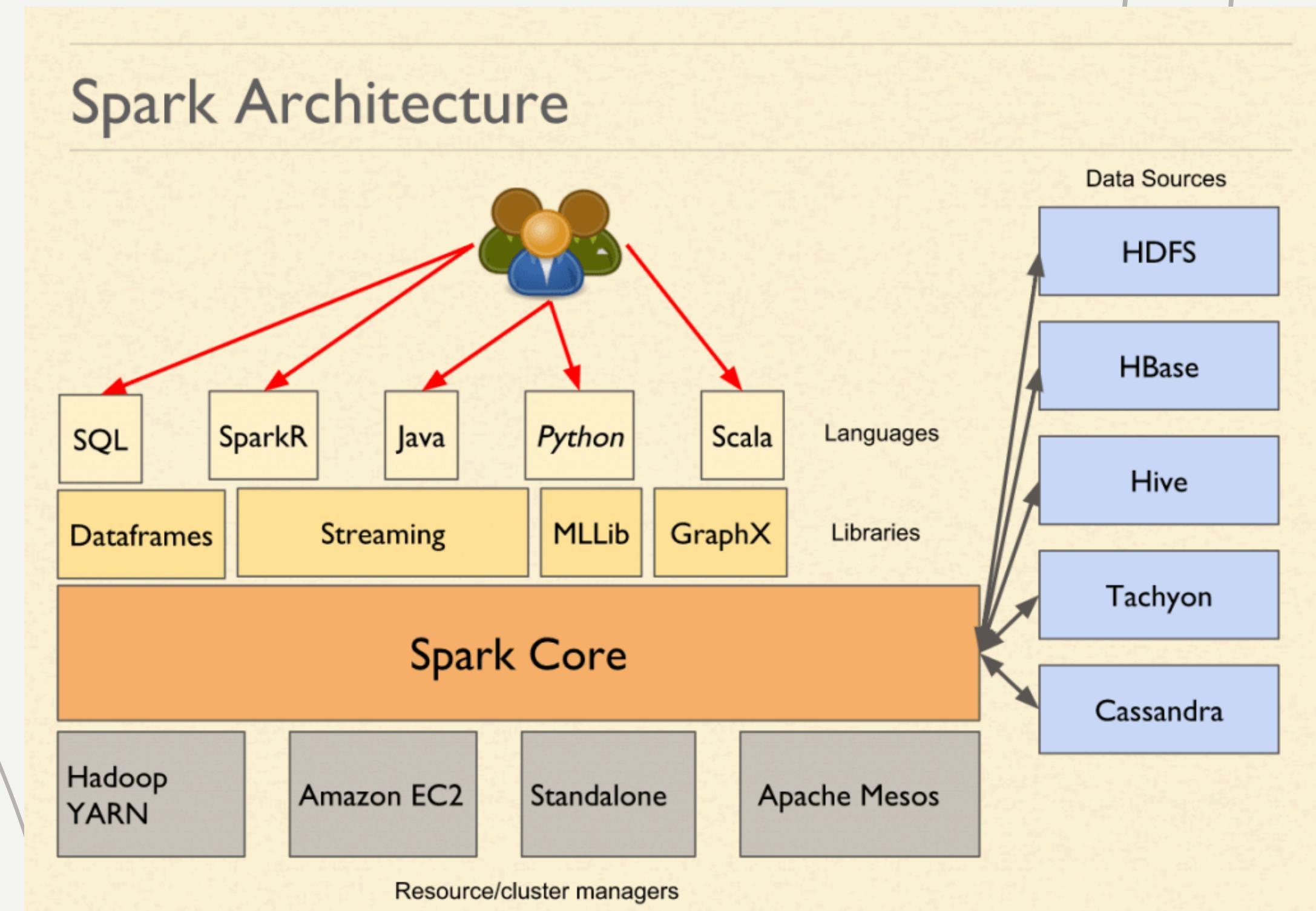
YARN (RM UI)



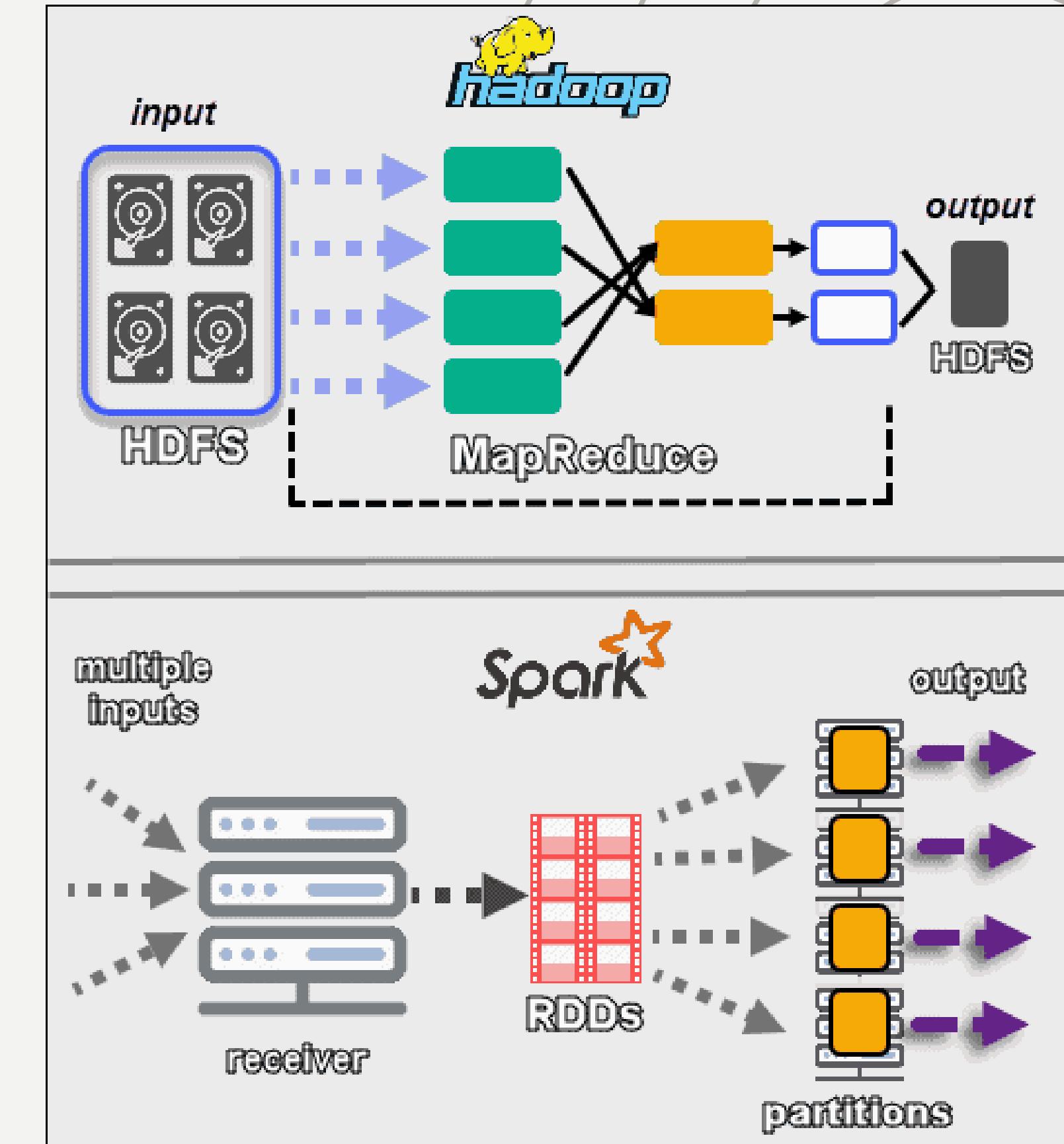
Кластер Hadoop



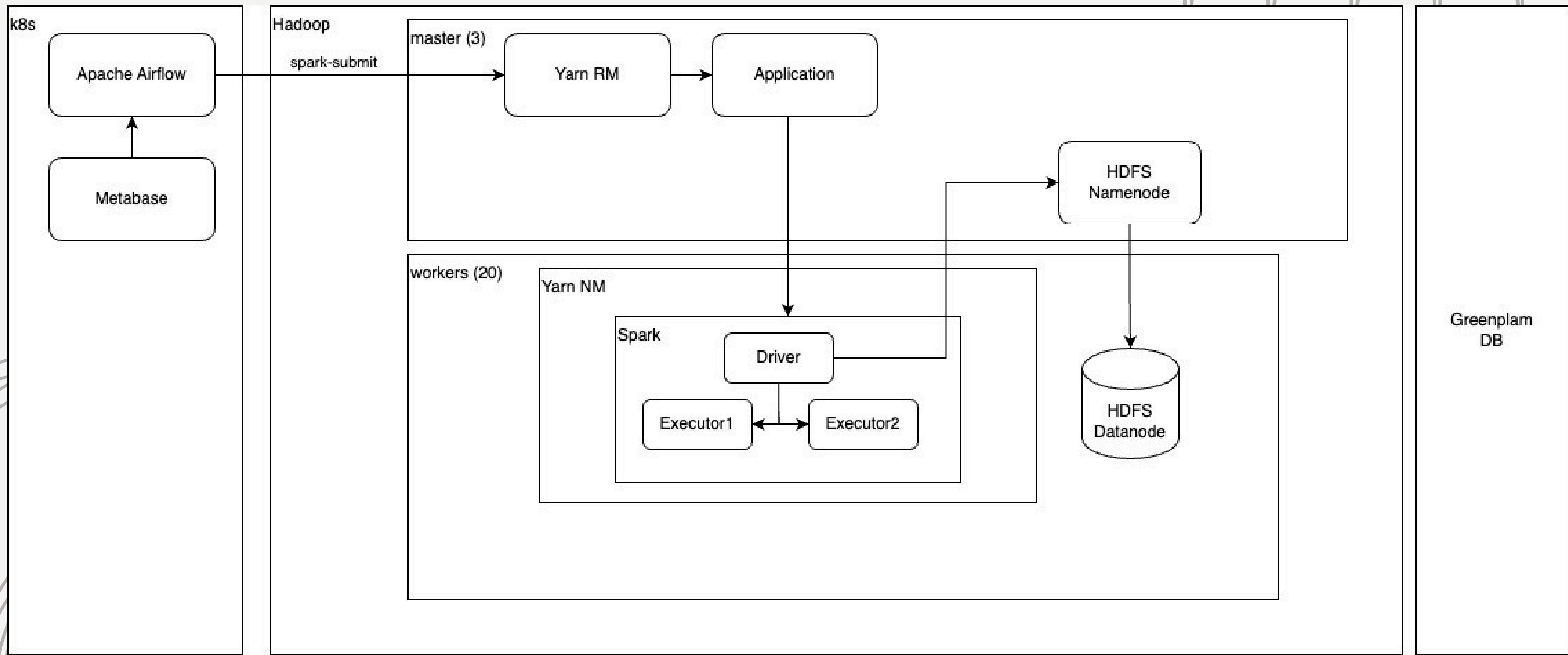
Apache Spark



MapReduce vs Spark



Пример проекта



Литература

O'REILLY®

ВЫСОКО- НАГРУЖЕННЫЕ ПРИЛОЖЕНИЯ

Программирование
масштабирование
поддержка



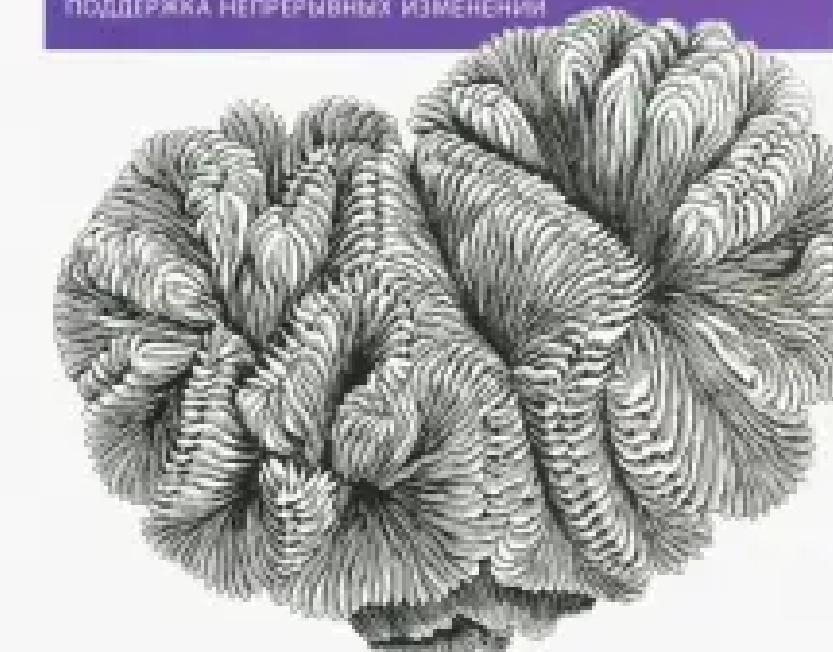
ПИТЕР®

Мартин Клеппман

O'REILLY®

ЭВОЛЮЦИОННАЯ АРХИТЕКТУРА

ПОДДЕРЖКА НЕПРЕРЫВНЫХ ИЗМЕНЕНИЙ



Нил Форд, Ребекка Парсонс, Патрик Куа

ПИТЕР®

O'REILLY®

Data Governance The Definitive Guide

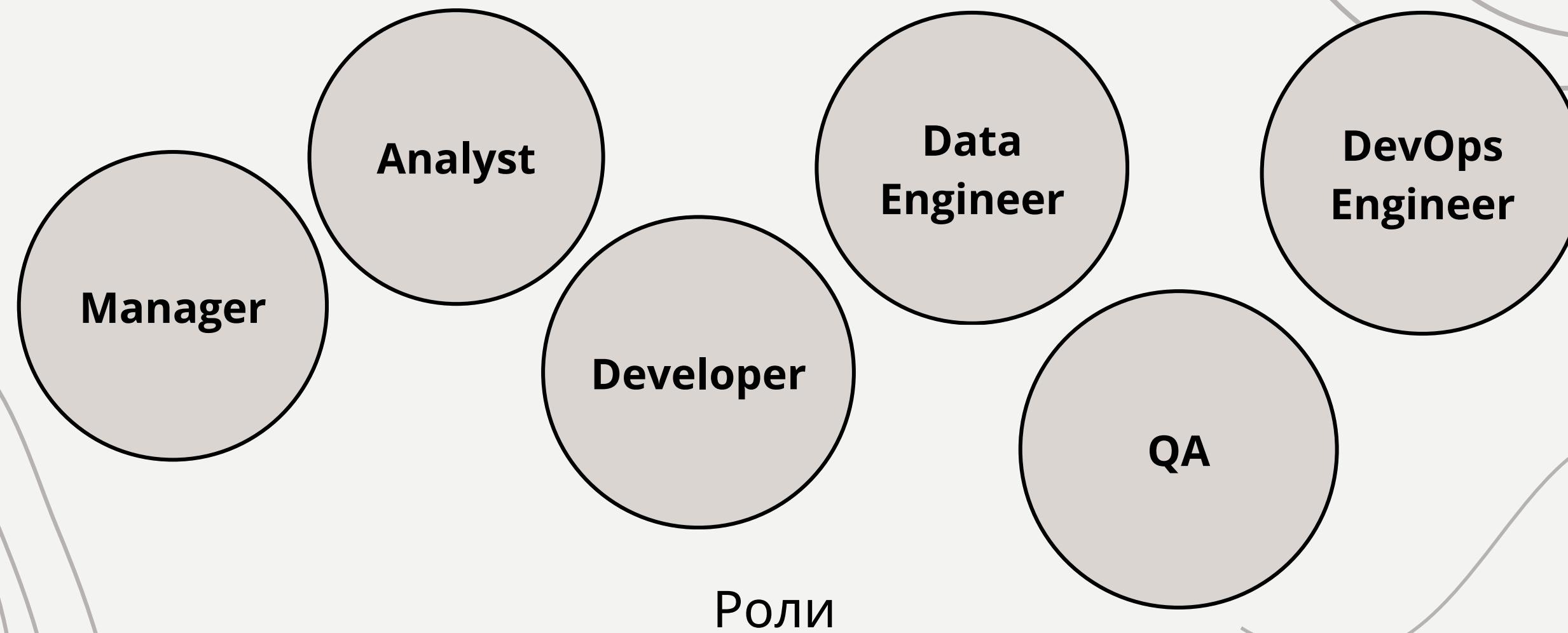
People, Processes, and Tools to Operationalize
Data Trustworthiness



Evren Eryurek, Uri Gilad,
Valliappa Lakshmanan,
Anita Kibunguchy-Grant
& Jessi Ashdown

Особенности работы в команде

Командная работа в проектах BigData



Командная работа в проектах BigData

01 GitHub/GitLab

02 Jira/Trello

03 Confluence

04 Slack/Telegram

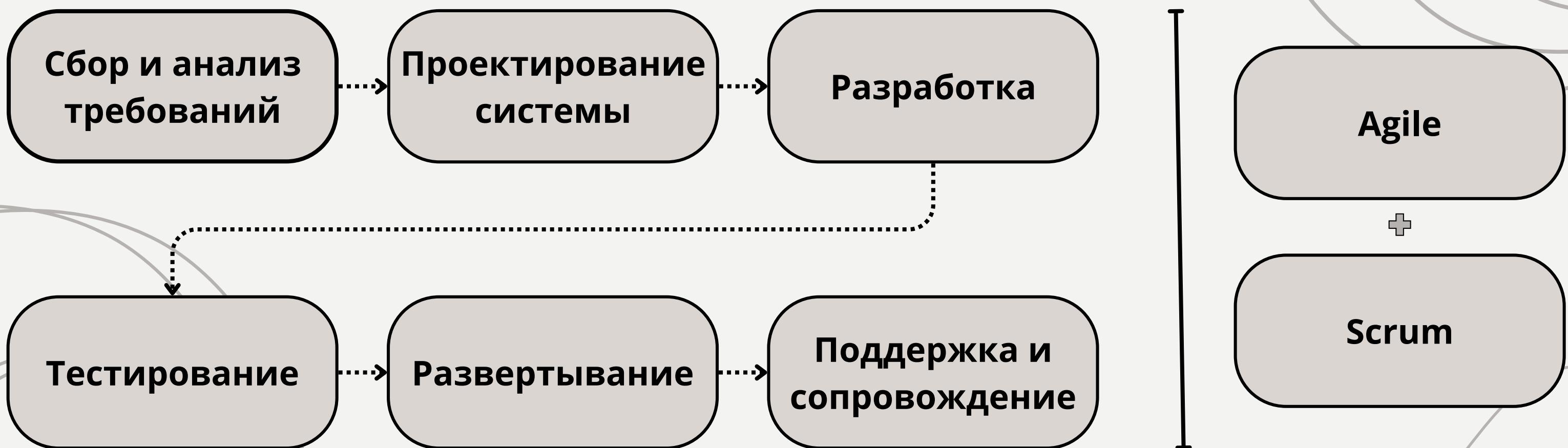
05 Miro/draw.io

06 Tableau/Power BI

07 Docker/Kubernetes

Инструменты

Этапы ЖЦ разработки



Тестирование

- Unit Testing
- Integration Testing
- Functional Testing
- UI Testing
- Performance Testing
- Security Testing
- Usability Testing
- Regressions Testing
- E2E Testing
- Compatibility Testing

Виды тестов

Уровни разработчика

Грейды

Ведущий,
Lead

Старший,
Senior

Средний,
Middle

Джуниор,
Junior

Стажёр,
Intern

Матрица компетенций

Критерий	Junior	Middle
SQL	Написание простых SQL ad-hoc запросов. Знание оконных функций	Написаные сложных оптимизированных SQL запросов
Структуры данных и алгоритмы	Может оценить сложность алгоритма, вычислительную сложность решения и расход памяти	Предыдущий уровень + может реализовать базовые алгоритмы: обход деревьев, сортировки
Автономность	Работает под присмотром более опытных коллег	Самостоятельность в принятии решений, способность вносить предложения в проект

Грейды

Грейд	Задачи	Вилка ~
Стажёр, Intern	Зарабатывает первый опыт в ИТ	до 80к
Джуниор, Junior	Нужна постоянная помощь и контроль	80-120к
Средний, Middle	Поставлено четкое ТЗ Немного знаком с технологиями	150-300к
Старший, Senior	Реализует тех задачки Предлагает арх. решения	200-350к
Ведущий, Lead	Общается с бизнесом Ставит тех задачки	300к +

Первый опыт

- Стажировки

- T-Bank
- Yandex
- VK
- Mail
- Avito
- Ozon
- MTS
- hh/habr-карьера (другое)

- Хакатоны

TG: @RussianHackers_Channel

- Pet-проекты

OpenSource

(<https://journal.tinkoff.ru/open-source/>)

Первый опыт

Fullstack developer (Python)

от 200 000 до 350 000 ₽ на руки

Требуемый опыт работы: 3–6 лет

Полная занятость, удаленная работа

Сейчас эту вакансию смотрят **4 человека**

Откликнуться



Мы ищем в команду опытного фуллстек-разработчика уровня senior+. Фокус в первую очередь на бекенде, но мы делаем фичи от начала до конца, без перекидывания через забор между фронтом и беком, поэтому нужно реализовывать и фронтенд часть задач тоже (в основном, в рамках существующей кодобазы).

Предстоит работать над аналитическим ядром нашей системы, интеграциями и интерфейсом. Кодобаза на Python, Go, JavaScript, Typescript. Используем ClickHouse, ScyllaDB, Dagster, DuckDB, RedPanda, React, GraphQL.

У нас будет идеальный мэтч, если наличествуют

- Отличное владение каким-либо backend-стеком технологий (Python, Java, Go и т.п.),
- Достаточное :) владение react,
- Готовность освоить наш стек технологий, если не все в нем знакомо,
- Внимательность к тестам и качеству того, что уходит на прод,
- Умение работать короткими быстрыми итерациями, дробить задачки на небольшие с понятным результатом,
- Готовность быстро освоить новые технологии,
- Опыт работы с agile практиками,
- Опыт работы в распределенных командах.

Работа удаленная, фултайм. Встречаемся в зуме, слаке и телеграм. Процессы изначально построены под распределенную команду.

Навыки

Python Backend JavaScript

РТУ МИРЭА

**Спасибо за
внимание!**

Вопросы