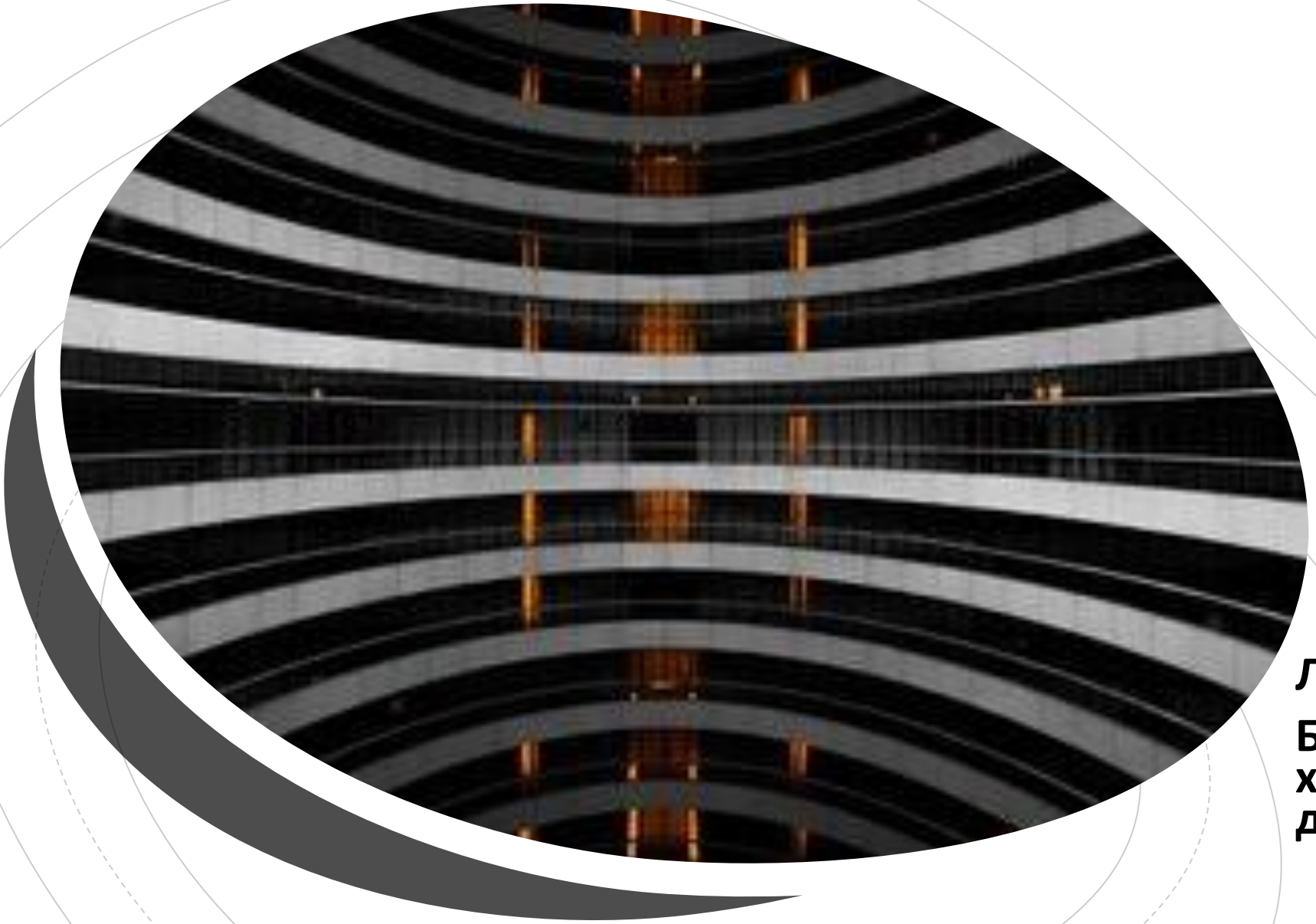


Smart data

Лекция #5

Большие данные и хранилища больших данных

И.А. Куликов

i.a.kulikov@gmail.com

Big Data (Большие данные)

Большие данные (Big data)

По данным Gartner -
«Большие данные» - это объемные, быстро доступные и разнообразные информационные ресурсы, которые требуют рентабельных, инновационных форм обработки информации для лучшего понимания и принятия решений.

Чаще всего о больших данных говорят как о таких данных, *размер которых превышает способность типовых СУБД собирать, хранить, управлять и анализировать их. А также анализ и обработки которых может иметь большую ценность для бизнеса*.

Большие данные (Big data)

2001 год. Дуглас Лэйни из META Group (сейчас Gartner) предложил три измерения (3V) - Volume, Velocity, Variety больших данных. Годами добавлялись новые V:

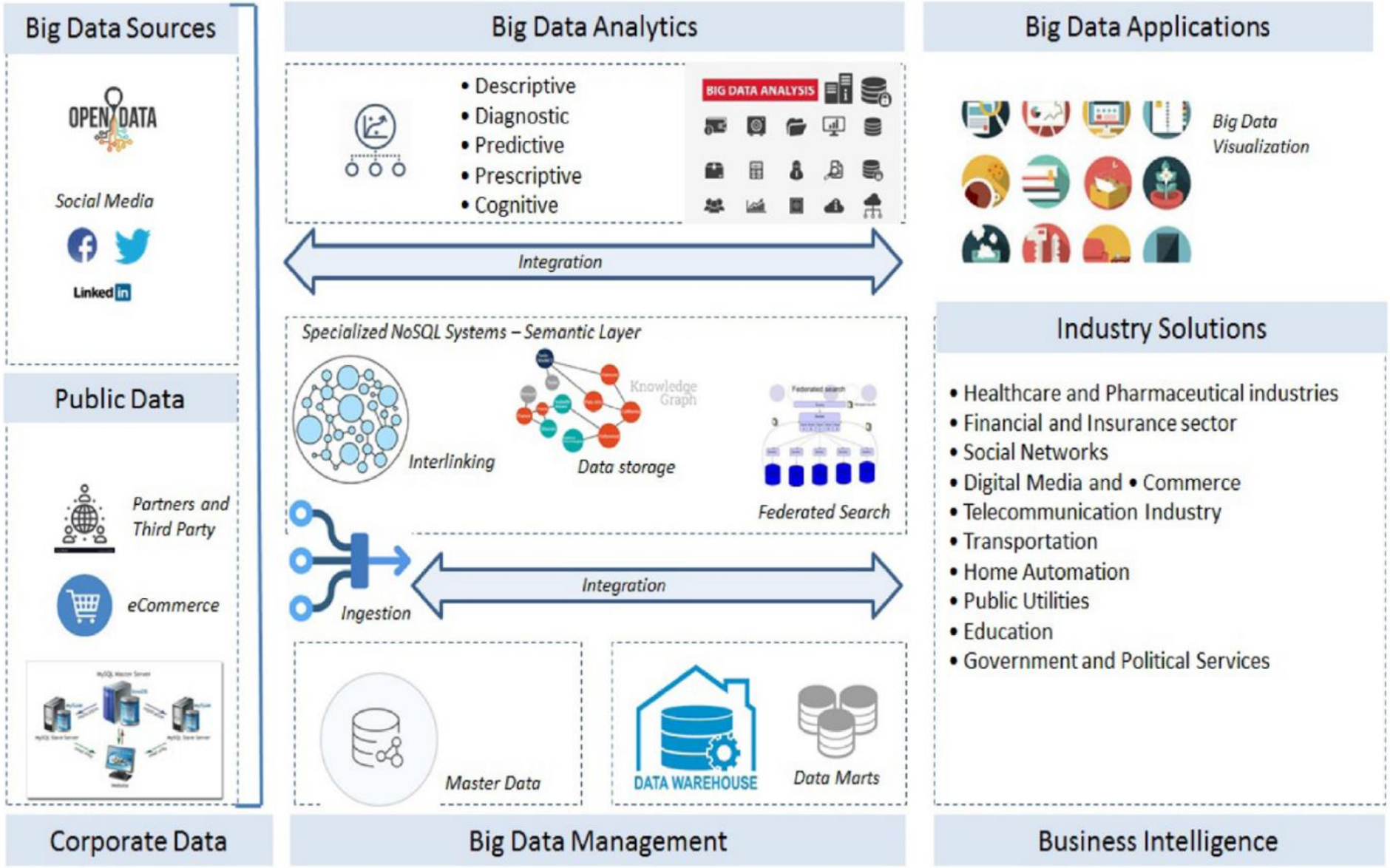
- Veracity – достоверность
- Validity – действительность
- Volatility – изменчивость
- Value – ценность
- Visualization – визуализация
- Vulnerability – уязвимость
- Variability - изменчивость

Примеры больших данных для организаций

- Данные о погоде
- Данные контрактов
- Данные о рабочей силе
- Данные технического обслуживания
- Данные финансовой отчетности
- Данные о соответствии страндартам
- Данные клинических испытаний
- Обработка записей врачей по диагностике и лечению

Источники данных

Источники данных



Типы источников больших данных

- ❑ Направленные данные (Directed data)
- ❑ Автоматизированные данные (Automated data)
- ❑ Данные систем автоматизированного мониторинга (Automated surveillance)
- ❑ Цифровые устройства (Digital devices)
- ❑ Считываемые данные (Sensed data)
- ❑ Данные сканеров (Scan data)
- ❑ Данные взаимодействий (Interaction data)
- ❑ Данные, предоставленные добровольцами (Volunteered data)
- ❑ Данные транзакций (Transactions)
- ❑ Данные социальных сетей (Social media)
- ❑ Данные управления своим здоровьем (Sousveillance)
- ❑ Краудсорсинг (Crowdsourcing)
- ❑ Гражданская наука (Citizen science)

Примеры источников открытых данных для разных доменов

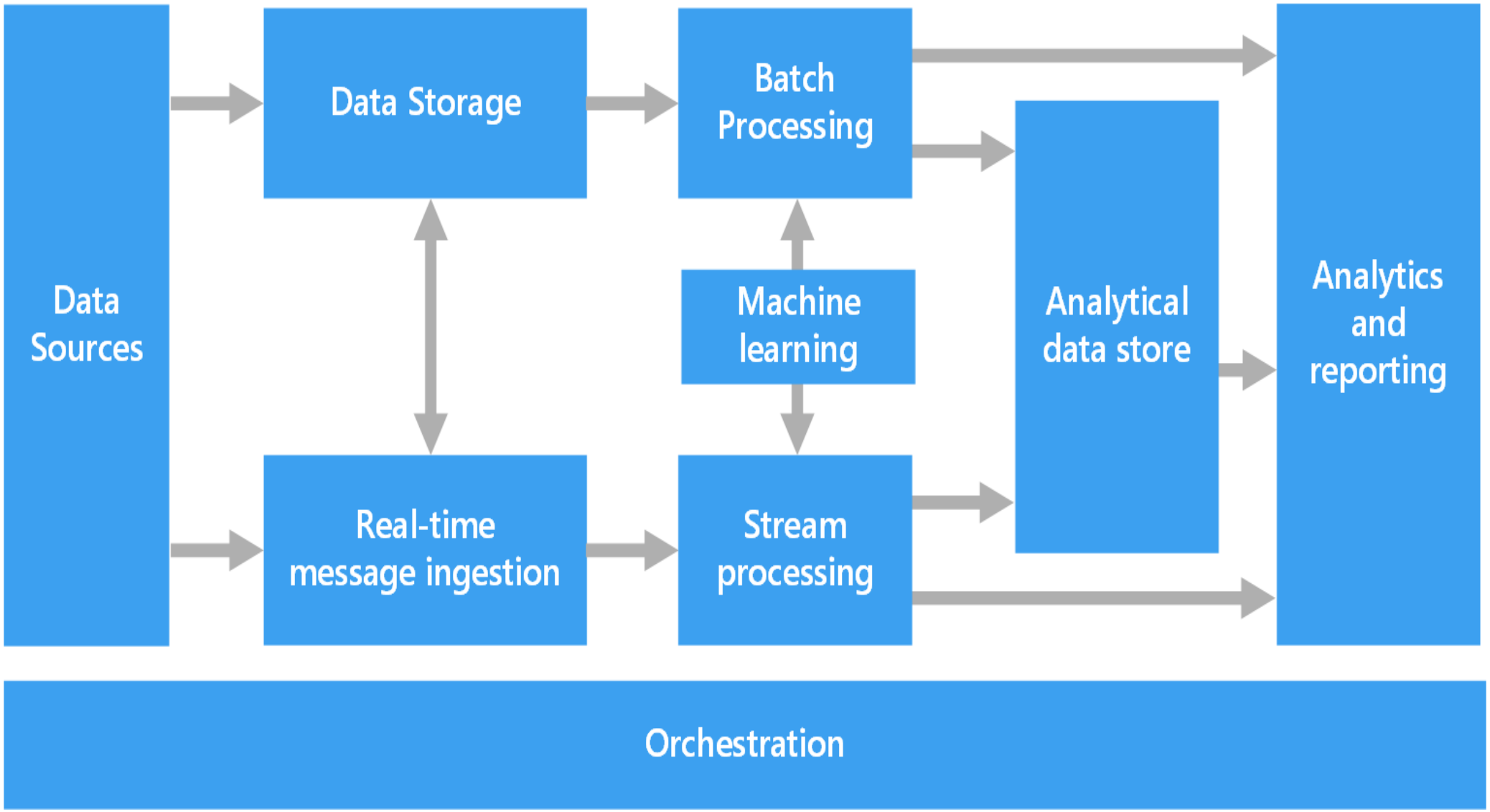
- Facebook API Graph
- Open Corporates
- Global Financial Data
- Open Street Map
- The National Centers for Environmental Information (NCEI)
- DBPedia

Классификация источников данных с точки зрения характера предоставляемых ими данных и методов их обработки

- **Текстовые данные.** Методы обработки: Preprocessing, Named Entity Recognition (NER), Entity Linking (EL), Relation Extraction (RE), Joint tasks
- **Размеченные данные** (например, лог файлы). Методы обработки: Wrapper-based extraction, Web table extraction, Deep Web crawling
- **Структурированные данные** (таблицы, деревья, базы данных, графы знаний). Методы обработки: Mapping from tables, Mapping from trees, Mapping from other knowledge graphs

Архитектура систем обработки больших данных

Компоненты архитектуры для обработки больших данных



Потоковая обработка против пакетной обработки

Пакетная обработка.

Плюсы

- ❑ Наиболее применима при работе с большими объемами.
- ❑ Обработка данных происходит самостоятельно.
- ❑ Это экономичный метод обработки данных.

Минусы

- ❑ Пользователи должны тщательно подготовить входные данные, предназначенные для пакетной обработки, прежде чем запускать их на компьютере.
- ❑ Проблемы с данными, сбои программы и ошибки, возникающие во время пакетной обработки, могут остановить весь процесс. К ним относятся незначительные ошибки в данных, такие как опечатки в датах.

Потоковая обработка.

Плюсы

- ❑ Данные всегда актуальны.
- ❑ Данные обновляются в режиме реального времени.
- ❑ Случаи задержки минимальны.

Минусы

- ❑ Потоковая обработка - это дорого и сложно.
- ❑ Информационный аудит при потоковой обработке является сложной задачей.

Лямбда архитектура

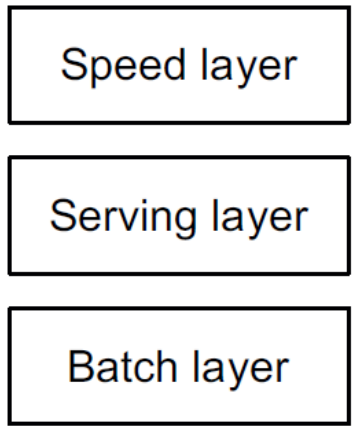


Figure 1.6 Lambda Architecture

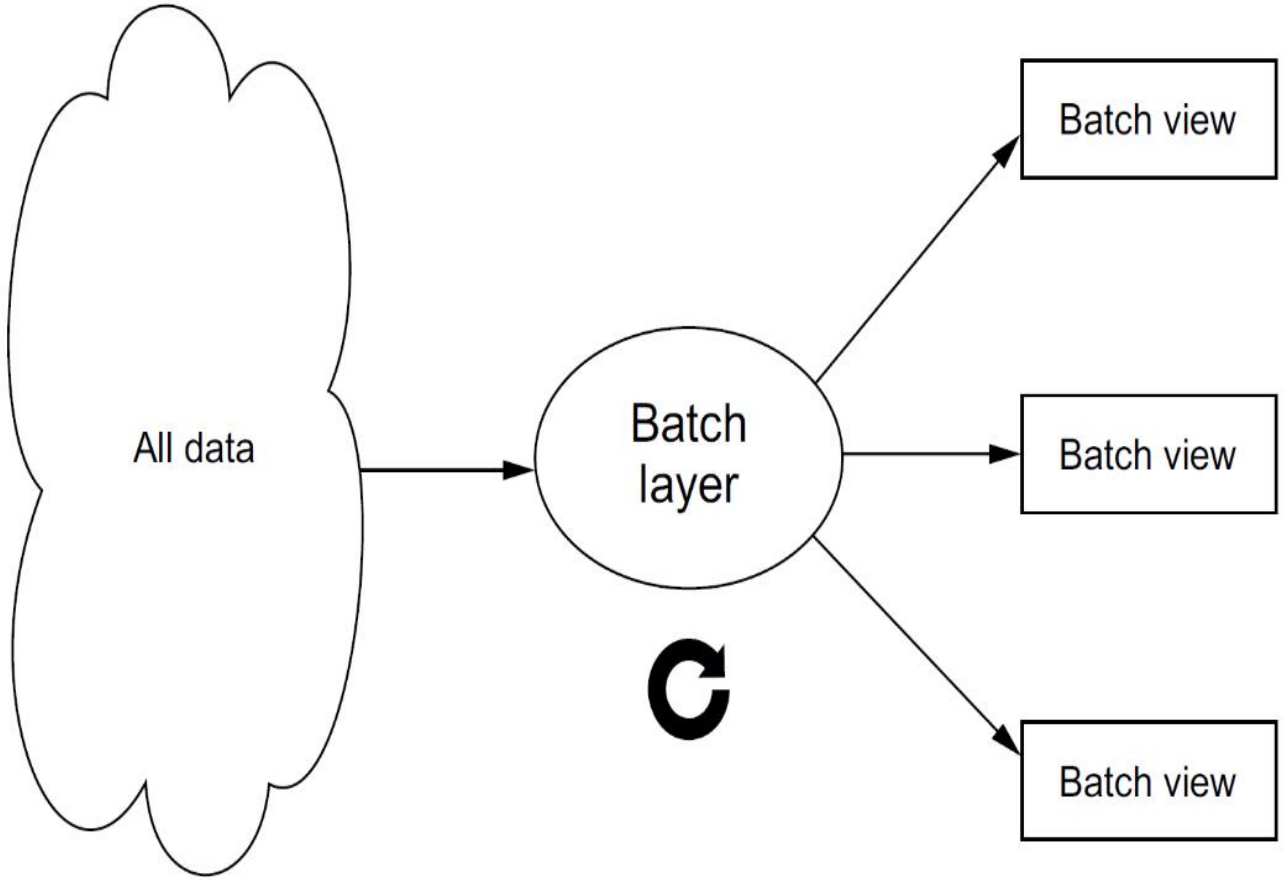


Figure 1.7 Architecture of the batch layer

Лямбда архитектура (продолжение)

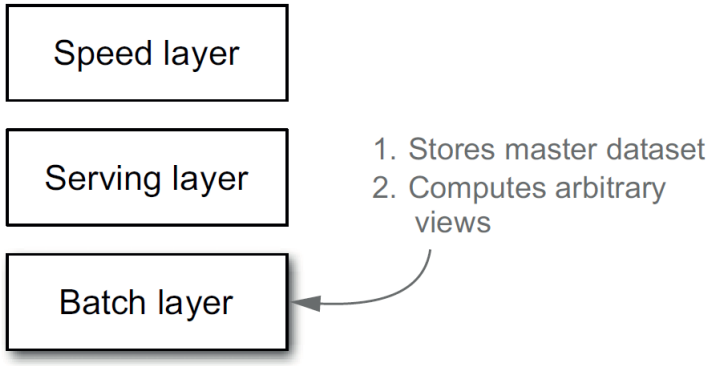


Figure 1.8 **Batch layer**

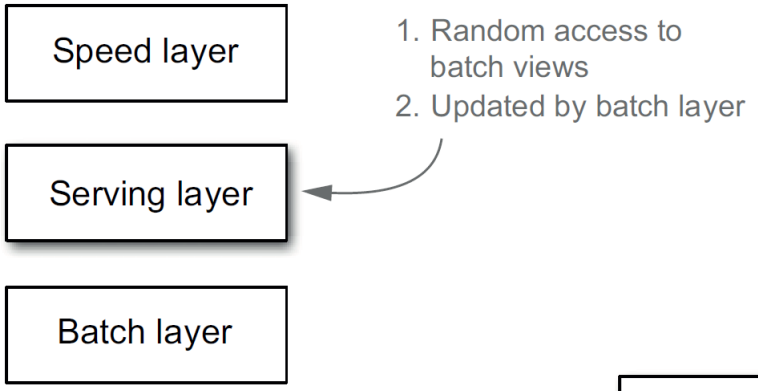


Figure 1.9 **Serving layer**

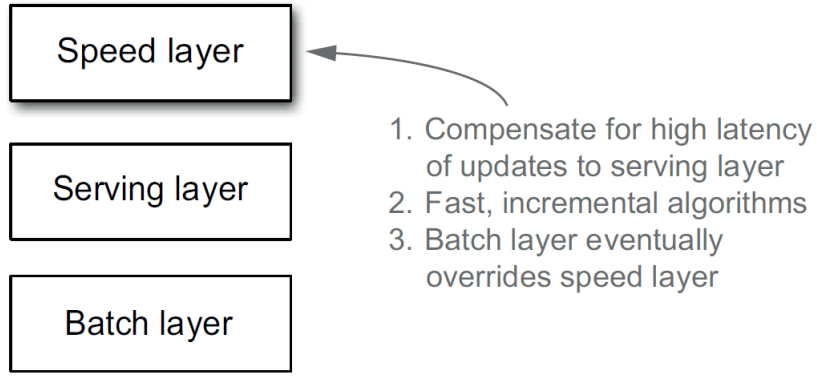


Figure 1.10 **Speed layer**

Лямбда архитектура (продолжение)

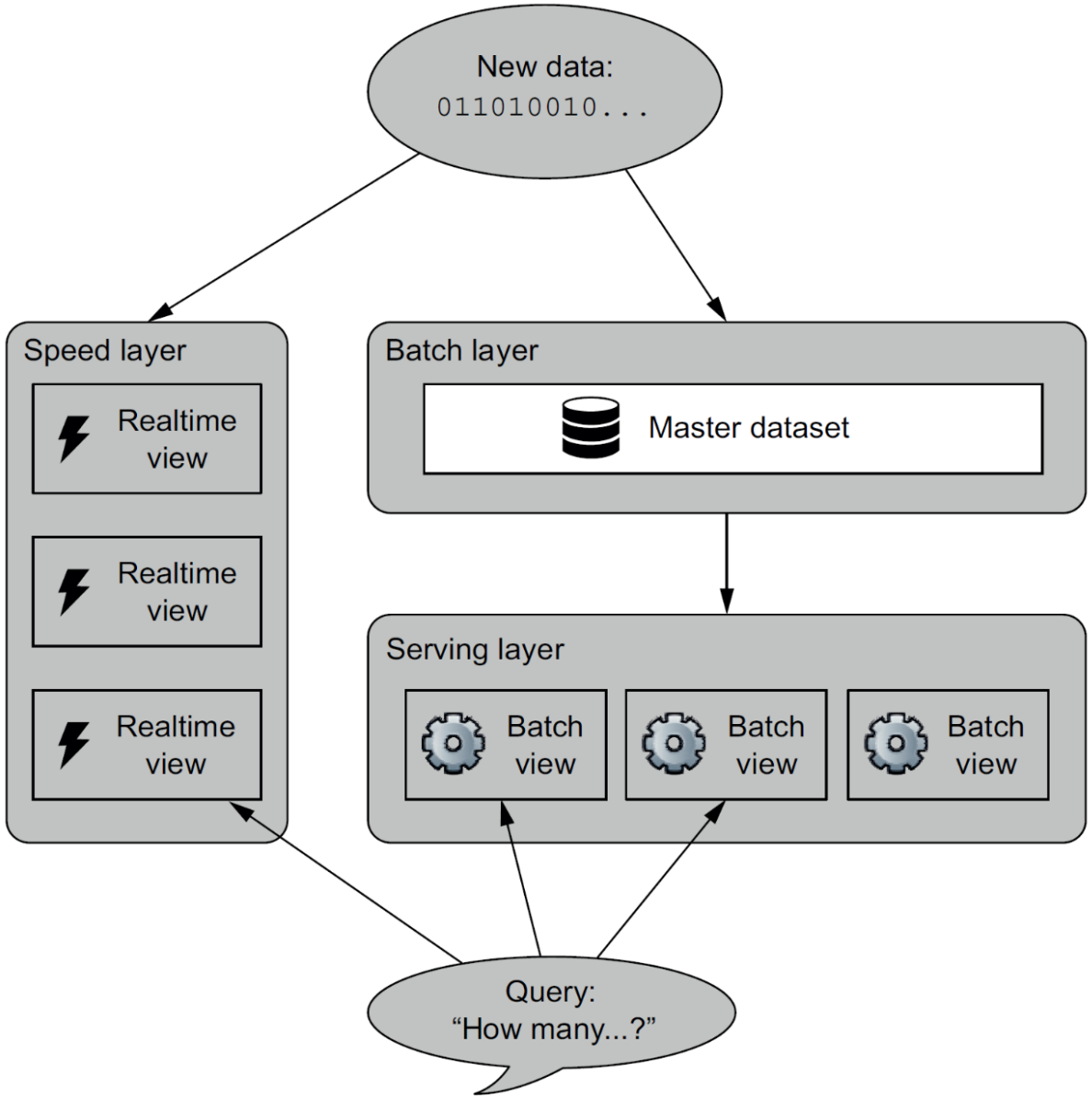
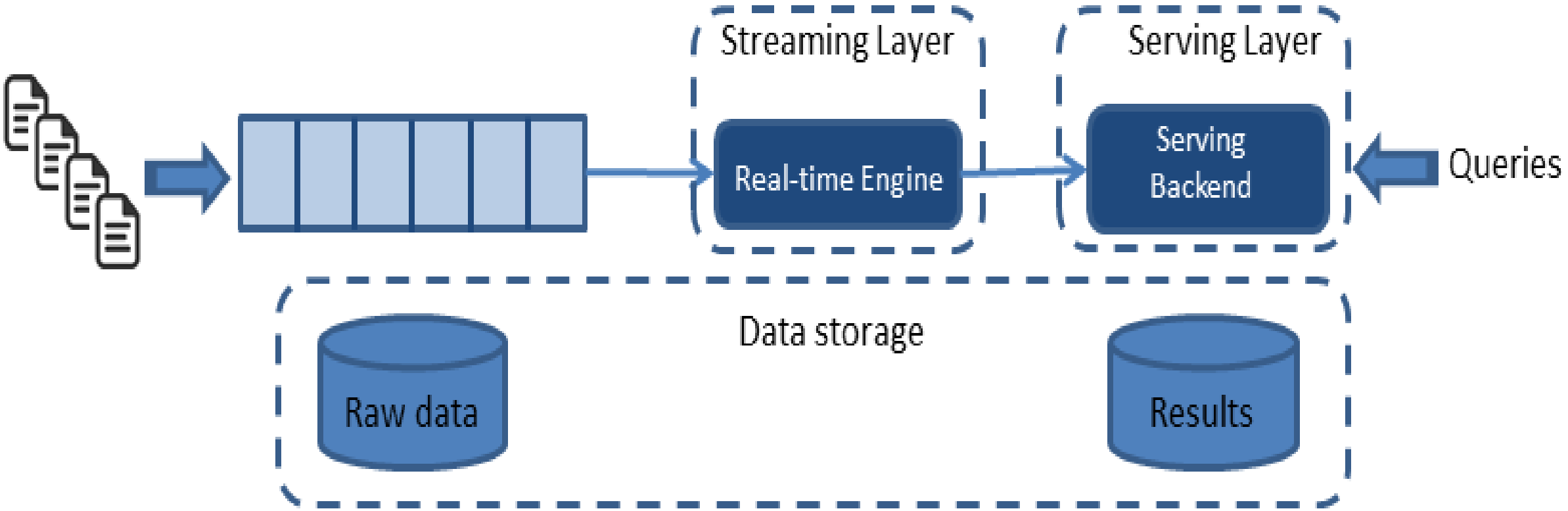


Figure 1.11 **Lambda Architecture diagram**

Каппа архитектура

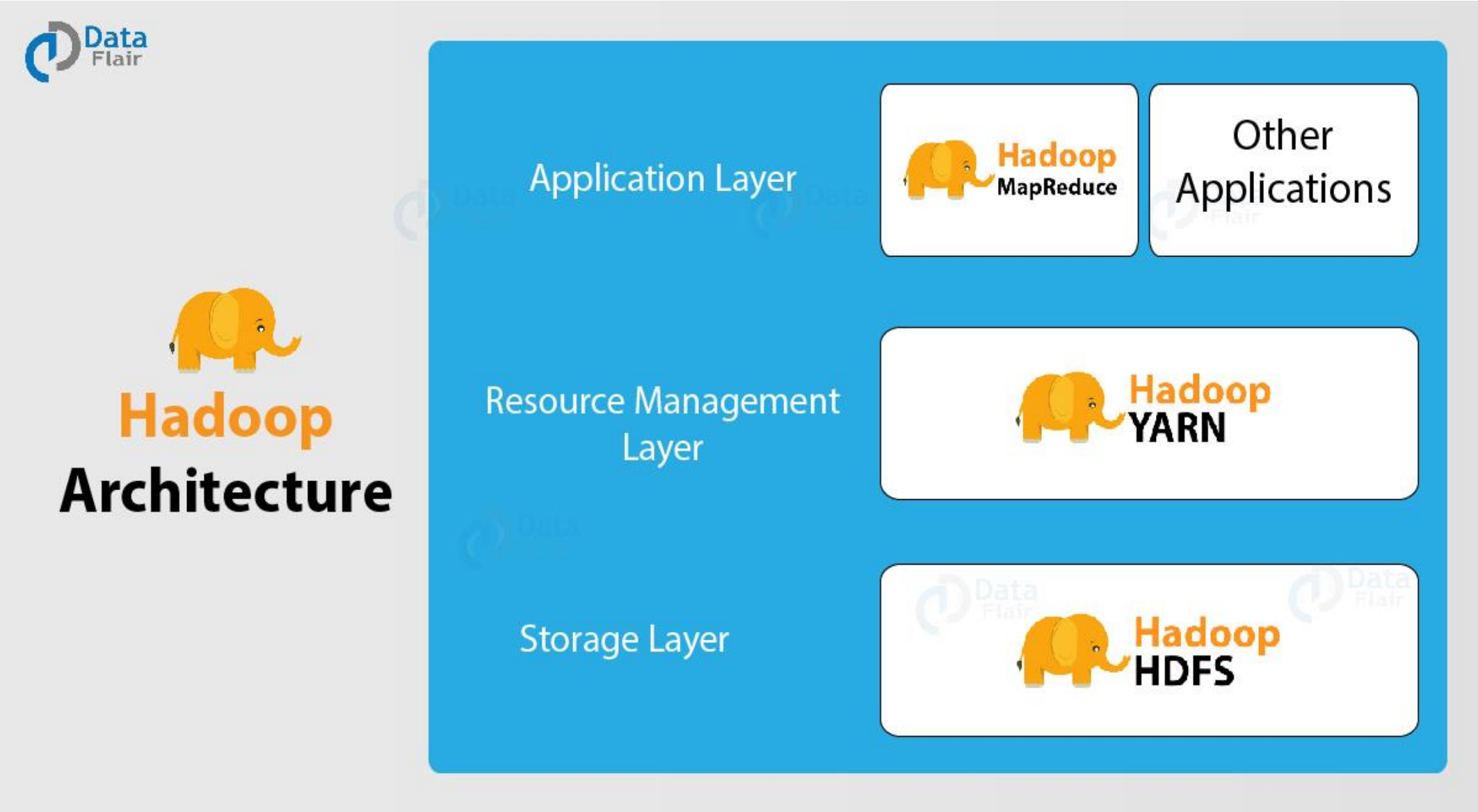


Хранилища данных

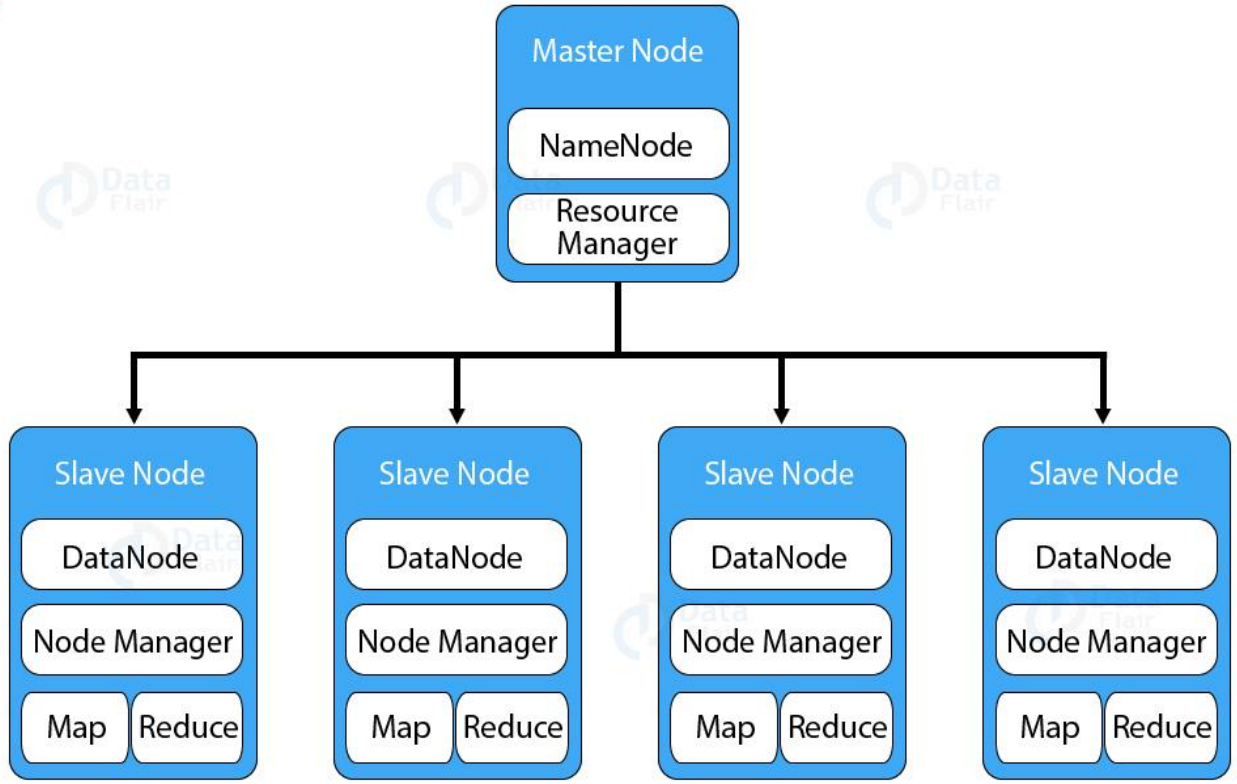
Технологии хранения больших данных

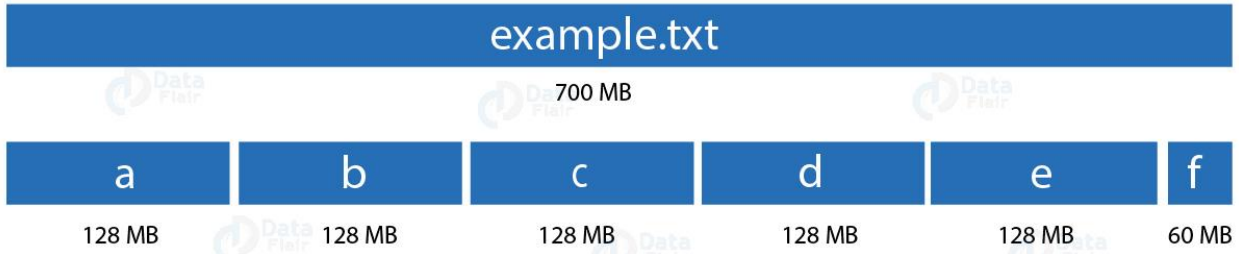
- Распределенные файловые системы
- Базы данных NoSQL
- Базы данных NewSQL
- Платформы запросов к большим данным

Hadoop HDFS


Data Flair
Hadoop
Architecture
Application Layer
Hadoop MapReduce
Other Applications
Resource Management Layer
Hadoop YARN
Storage Layer
Hadoop HDFS

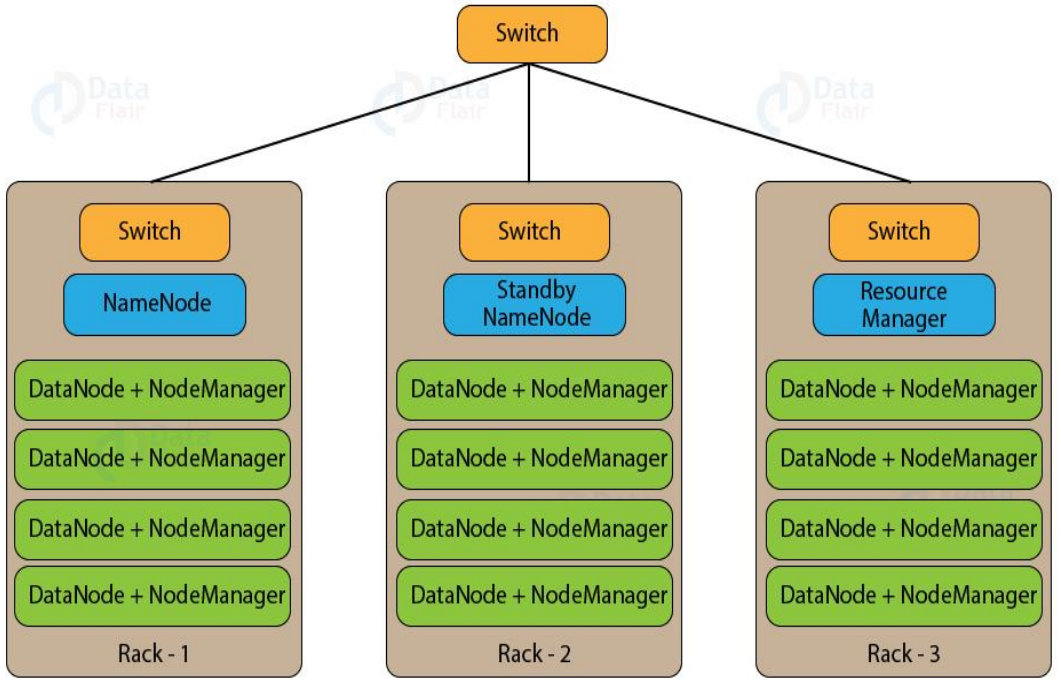
Hadoop HDFS (продолжение)



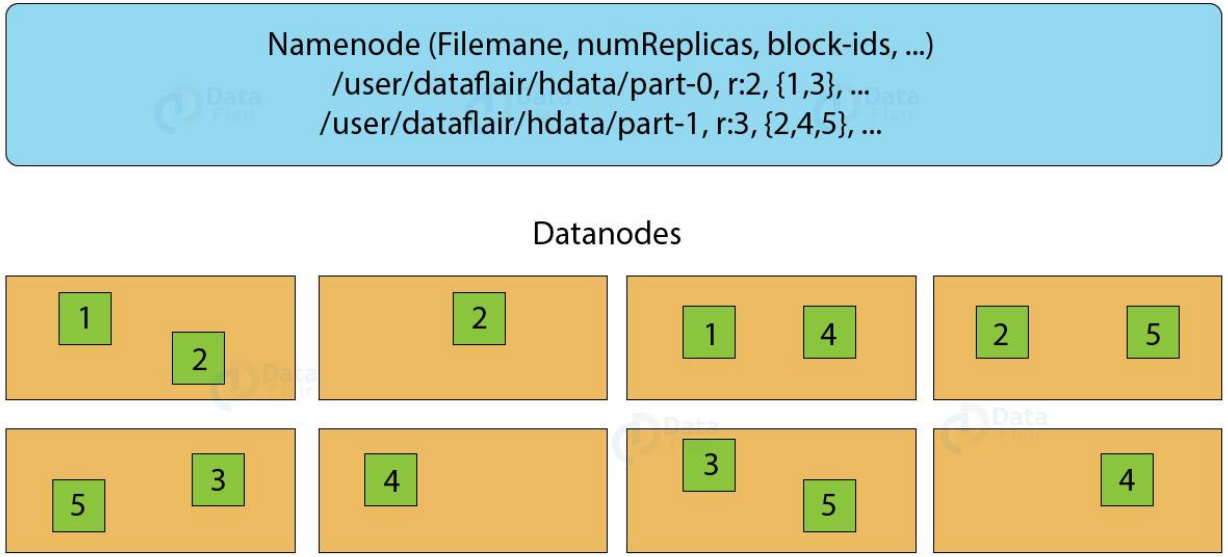


Hadoop HDFS (продолжение)

Rack Awareness

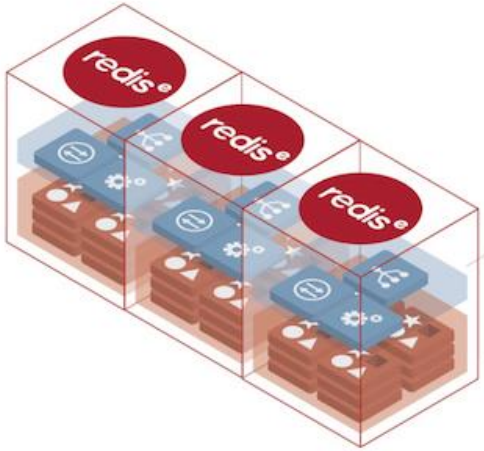


Block Replication



Хранилища данных
NoSQL базы данных

Хранилища "ключ-значение"

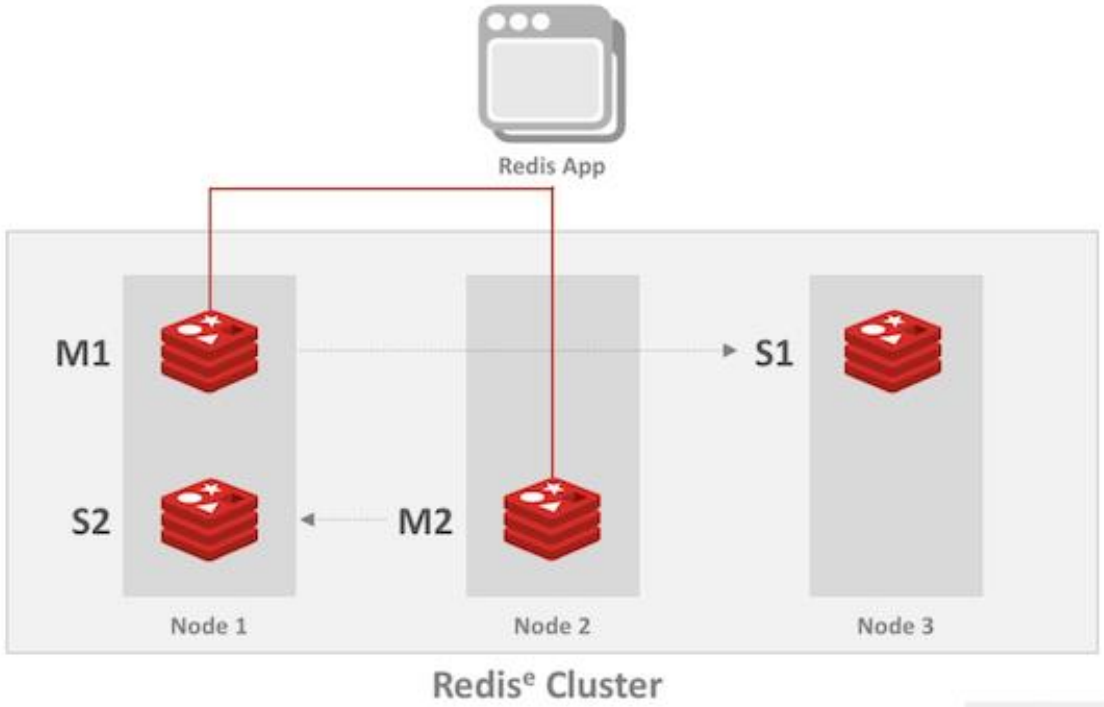


Redis^{e} Cluster Manager

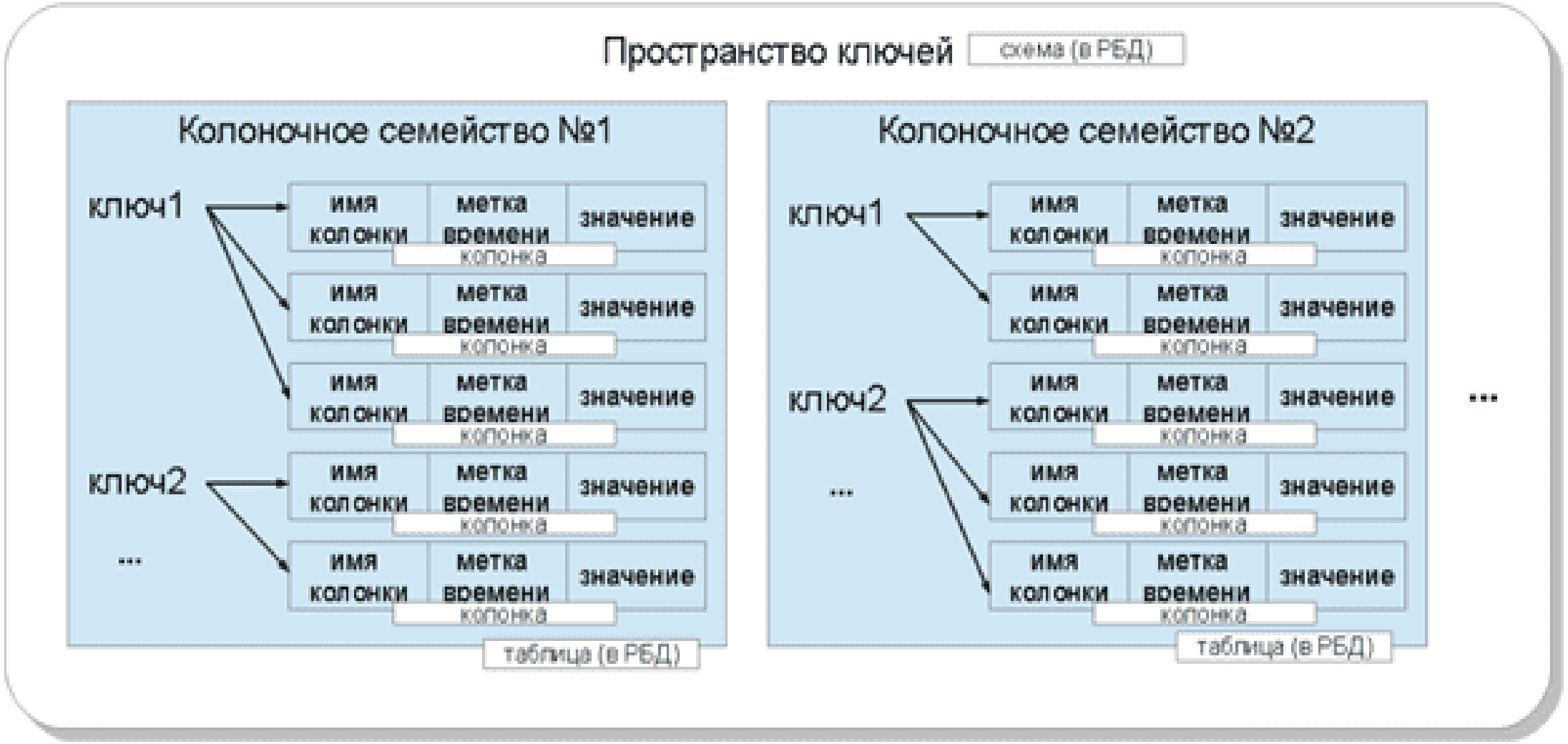
detecting failures and orchestrating auto-failover

Redis Shards

Master and Slave shards maintained using in-memory replication over LAN

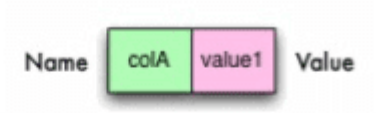


Хранилища столбцов

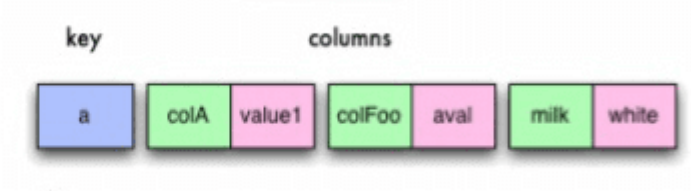


Хранилища столбцов (продолжение)

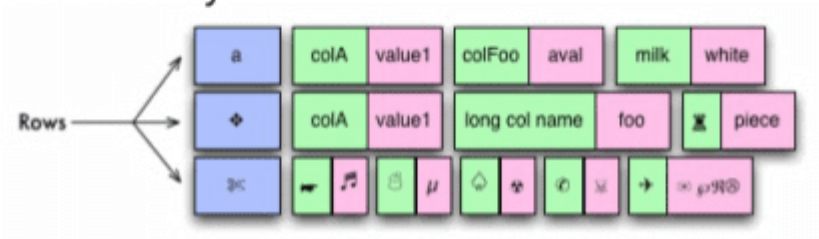
Ячейка

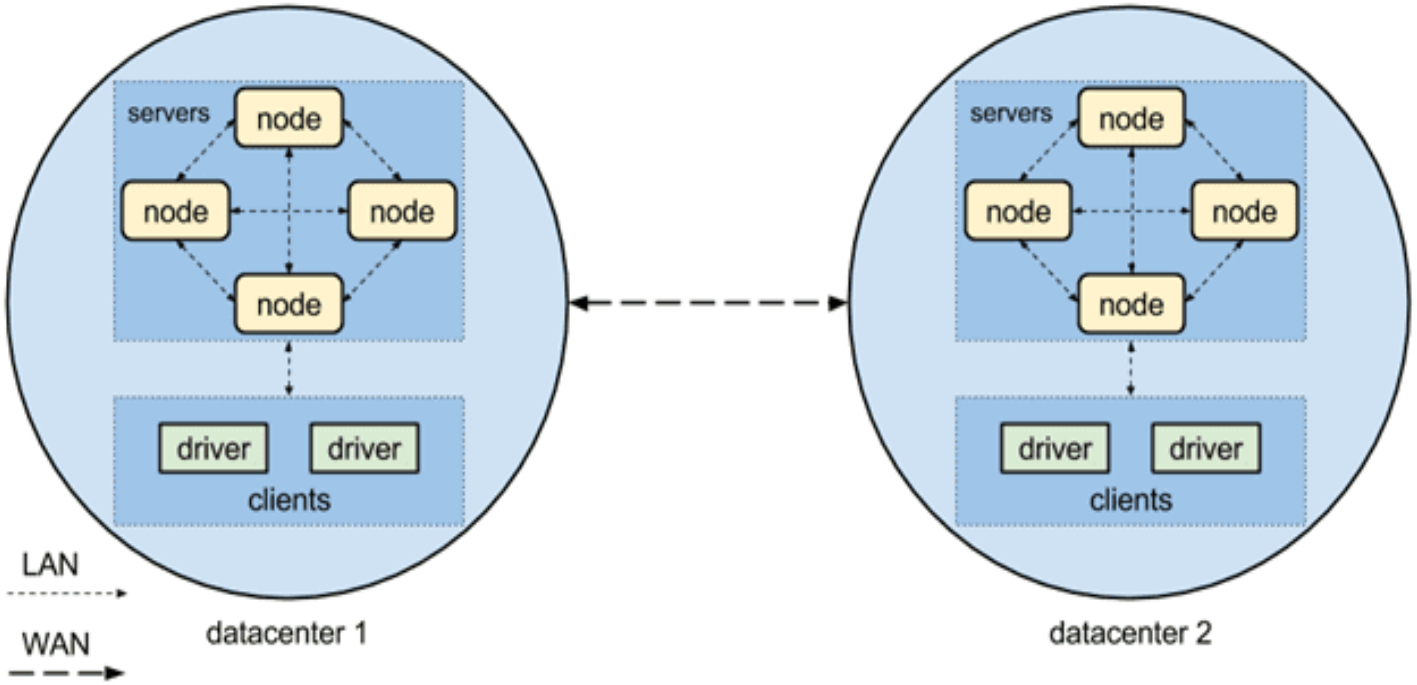


Строка

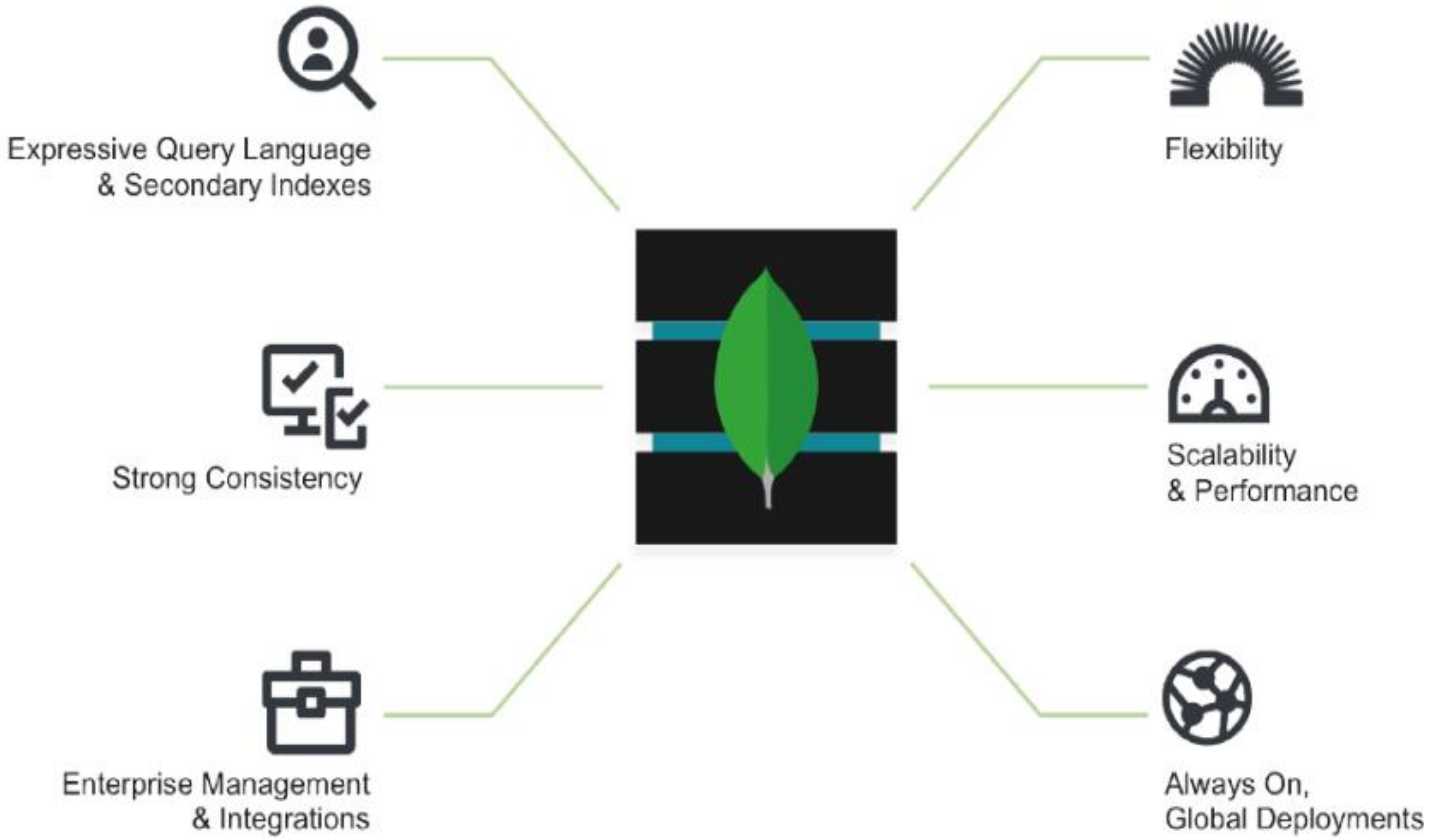


Column Family

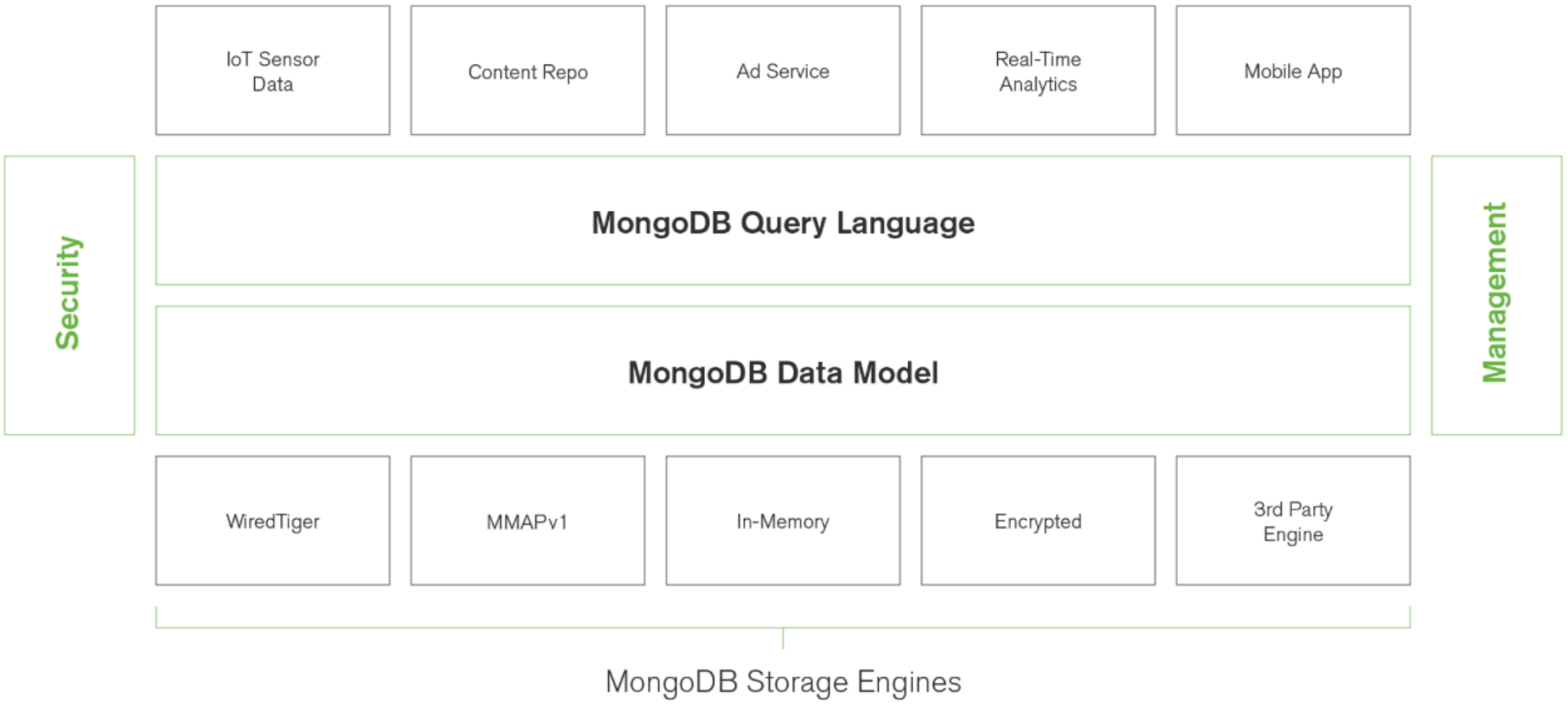




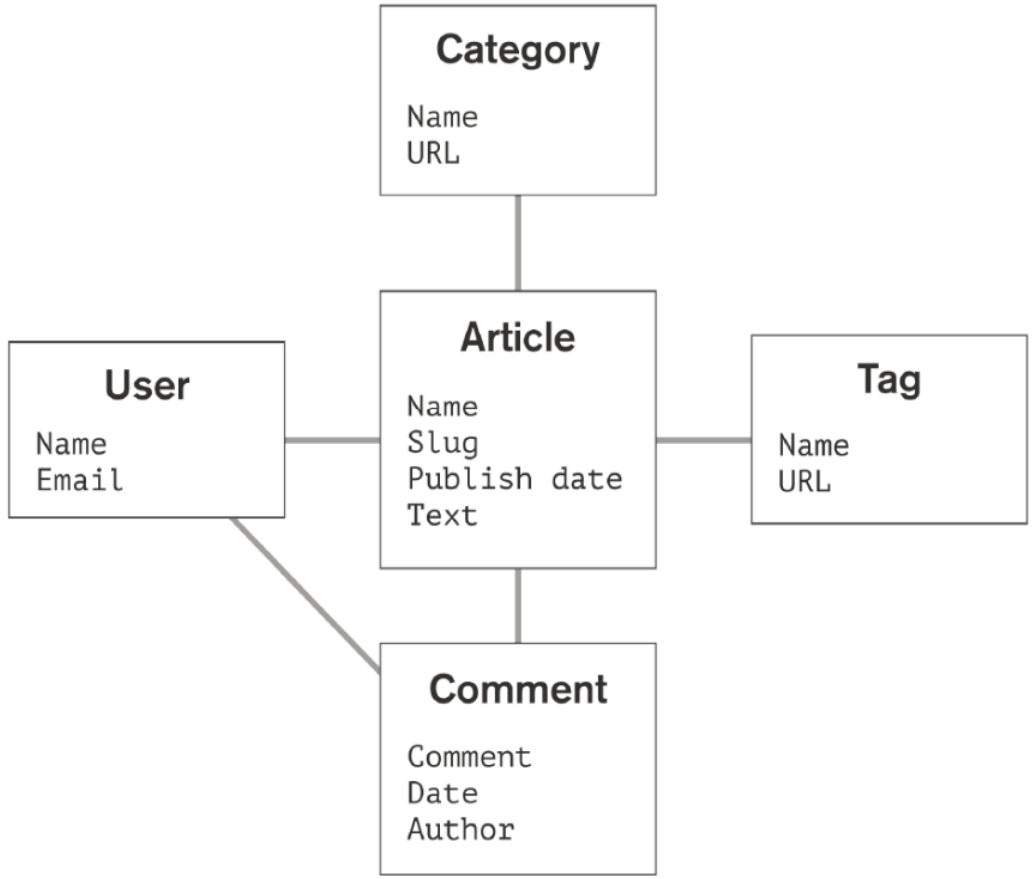
База данных документов

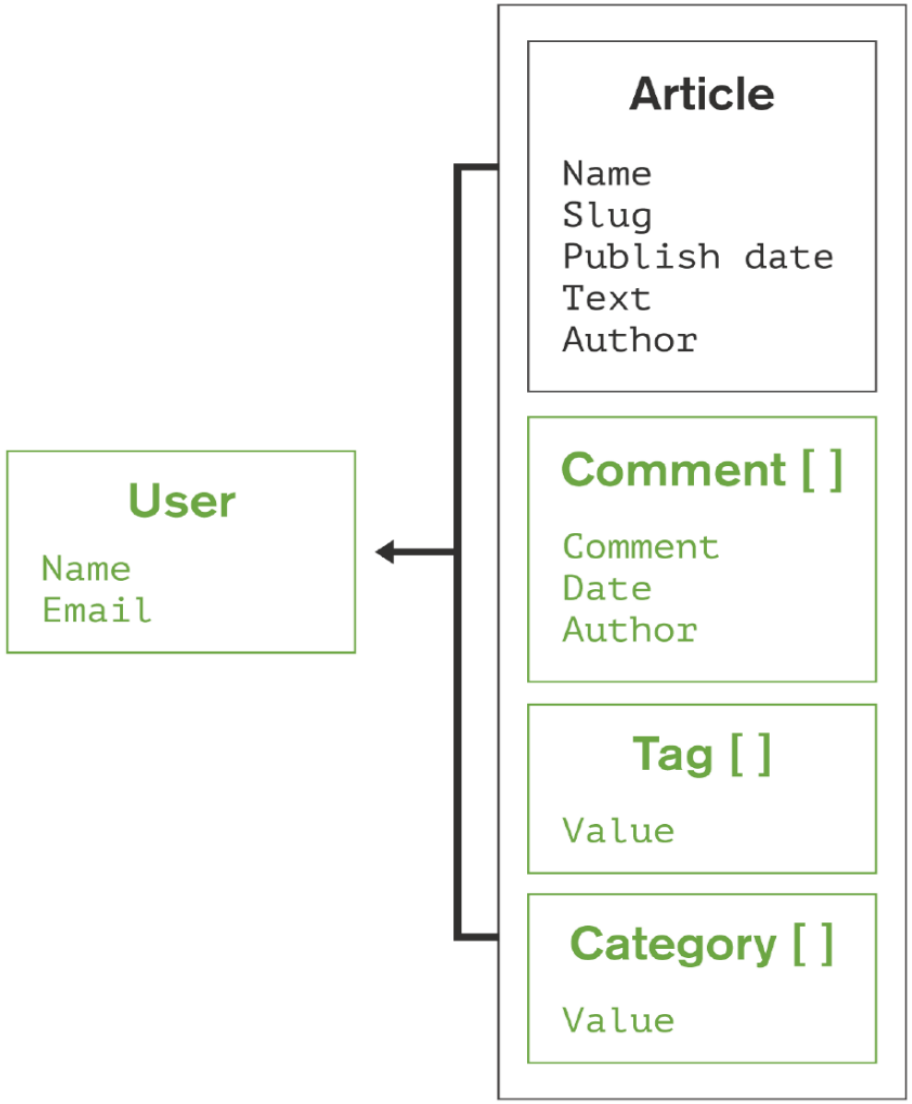


База данных документов (продолжение)

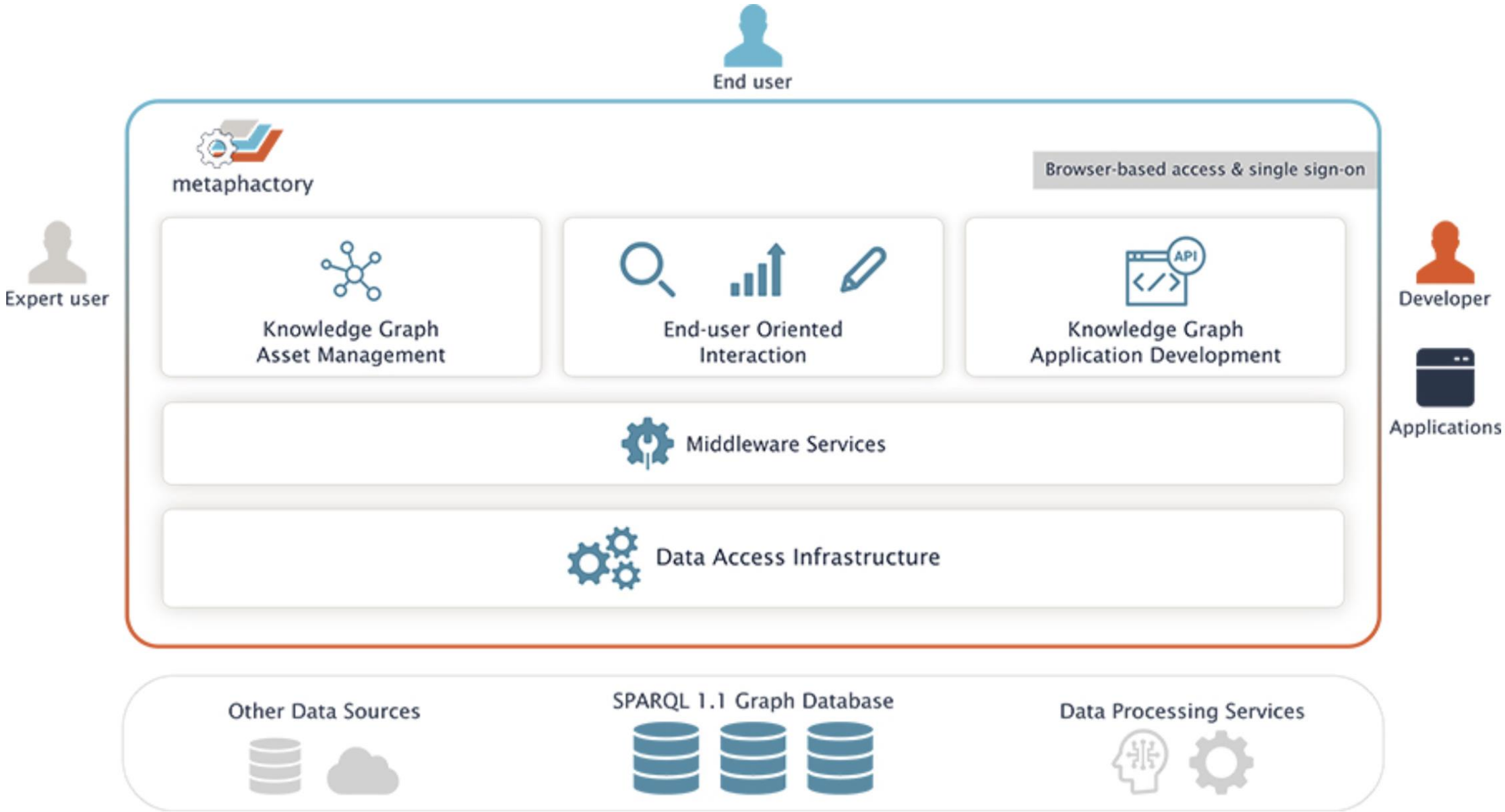


База данных документов (продолжение)

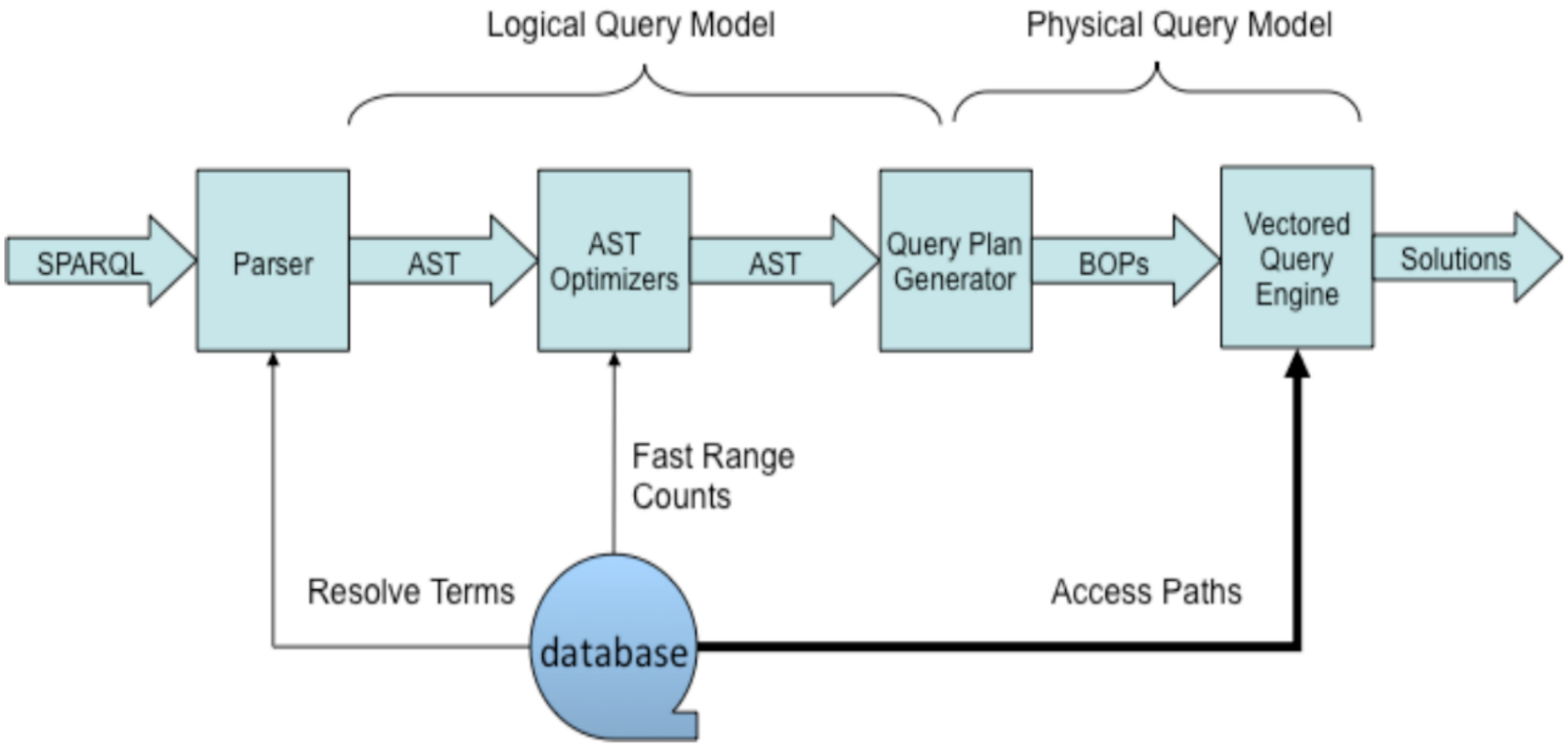




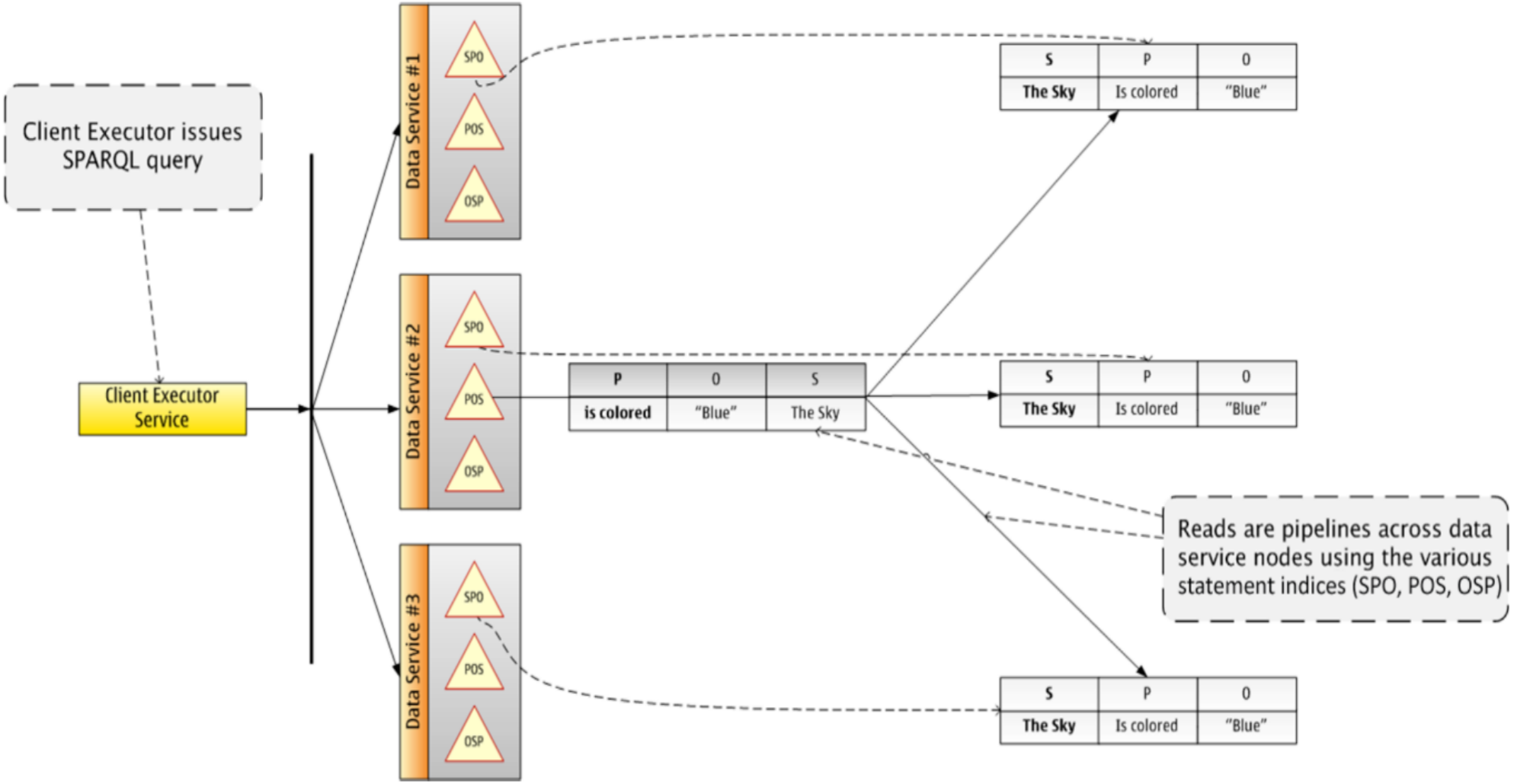
Графовые БД


End user
metaphactory
Browser-based access & single sign-on
Expert user
Knowledge Graph
Asset Management
End-user Oriented
Interaction
API
Knowledge Graph
Application Development
Developer
Applications
Middleware Services
Data Access Infrastructure
Other Data Sources
SPARQL 1.1 Graph Database
Data Processing Services

Графовые БД (продолжение)



Графовые БД (продолжение)



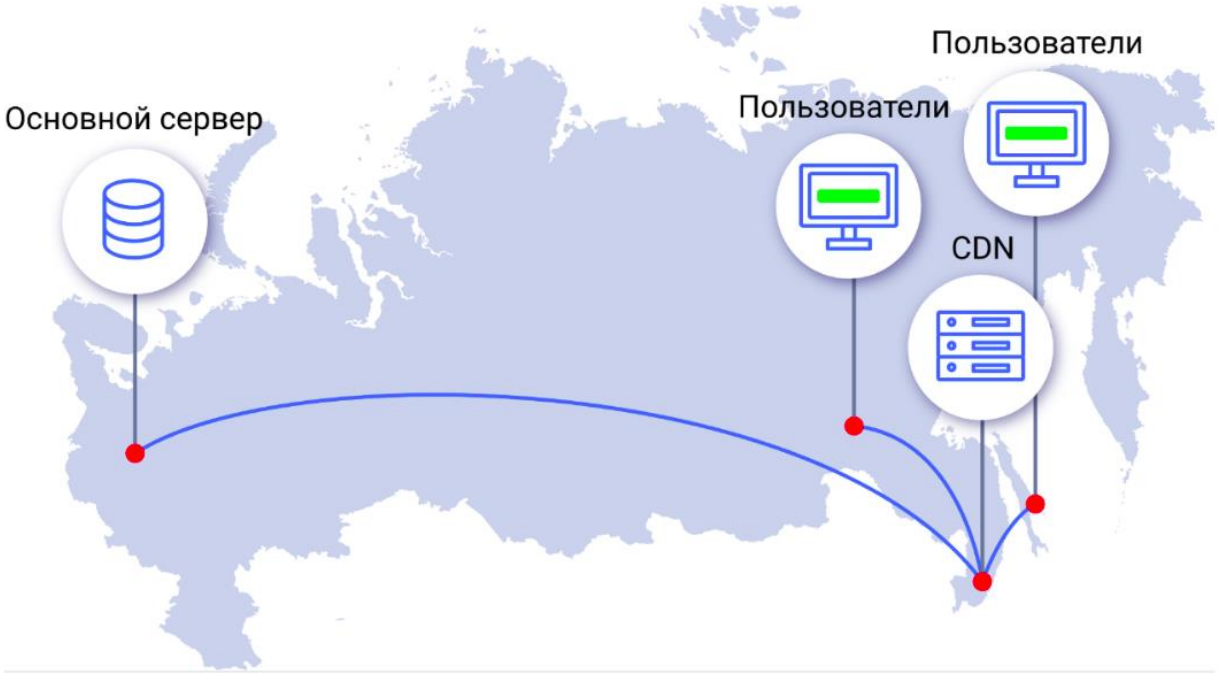
NewSQL базы данных

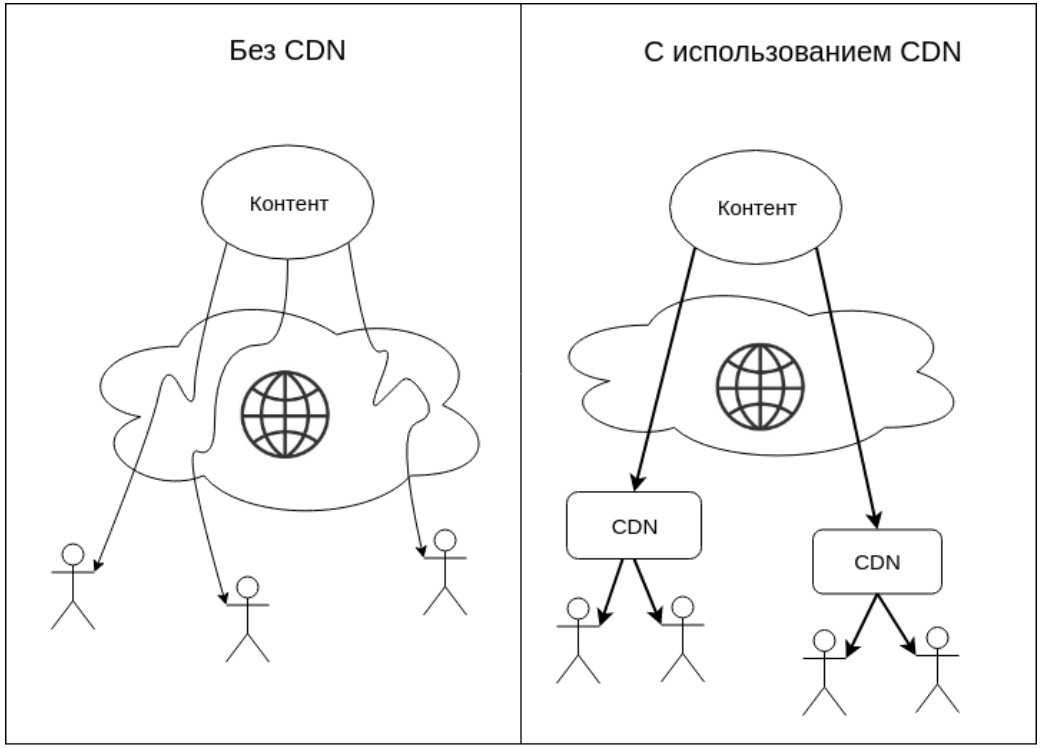
Базы данных NewSQL - современная форма реляционных баз данных, предназначенная для обеспечения масштабируемости, сравнимой с базами данных NoSQL при сохранении транзакционных гарантий, поддерживаемых традиционными системами баз данных.
Такие СУБД обладают следующими характеристиками:

- ❑ SQL - это основной механизм взаимодействия с приложениями.
- ❑ Поддержка ACID для транзакций
- ❑ Механизм управления неблокирующим параллелизмом.
- ❑ Архитектура, обеспечивающая гораздо более высокую производительность на каждом узле.
- ❑ Горизонтально масштабируемая архитектура без совместного использования ресурсов, способная работать на большом количестве узлов без проблем.

Ожидается, что системы NewSQL примерно в 50 раз быстрее традиционных.
СУБД.

CDN – Content Delivery Network





Платформы запросов к большим данным

- Hive
- Impala
- Spark SQL
- Drill

Облачные хранилища

- Amazon cloud
- Microsoft Azure
- Google cloud
- Etc.

Перспективные требования к хранению больших данных

- ❑ Стандартизированные интерфейсы запросов
- ❑ Безопасность и конфиденциальность
- ❑ Отслеживание данных и их происхождения
- ❑ «Песочница» и виртуализация
- ❑ Семантические модели данных

Новые парадигмы для хранения больших данных

- ❑ Расширенное использование баз данных NoSQL
- ❑ Хранение данных в памяти и хранилища столбцов
- ❑ Конвергенция с аналитическими платформами
- ❑ Центр данных (The Data Hub)

Благодарю за внимание