Лекция 3. Сквозная цифровая технология Большие данные

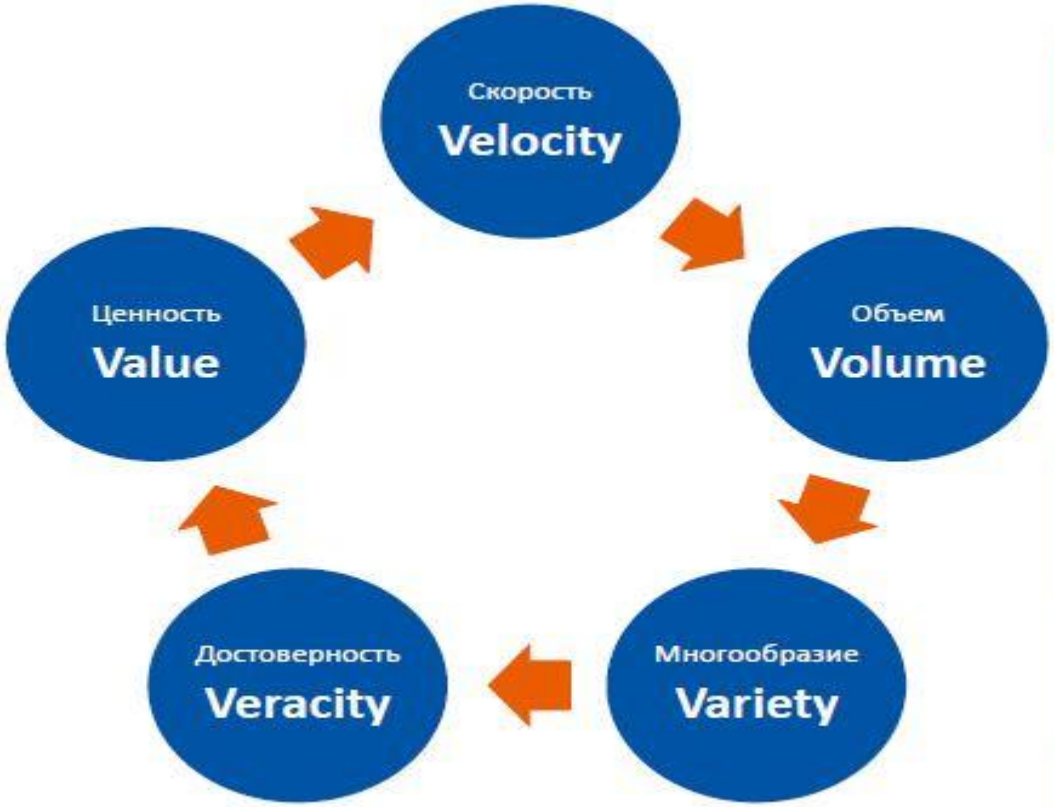
Д.т.н., профессор Гусева А.И.

2025 г.

Сквозная цифровая технология Большие данные

Большие данные – новое поколению технологий, предназначенных для экономически эффективного извлечения полезной информации из очень больших объемов разнообразных данных путем высокой скорости их сбора, обработки и анализа

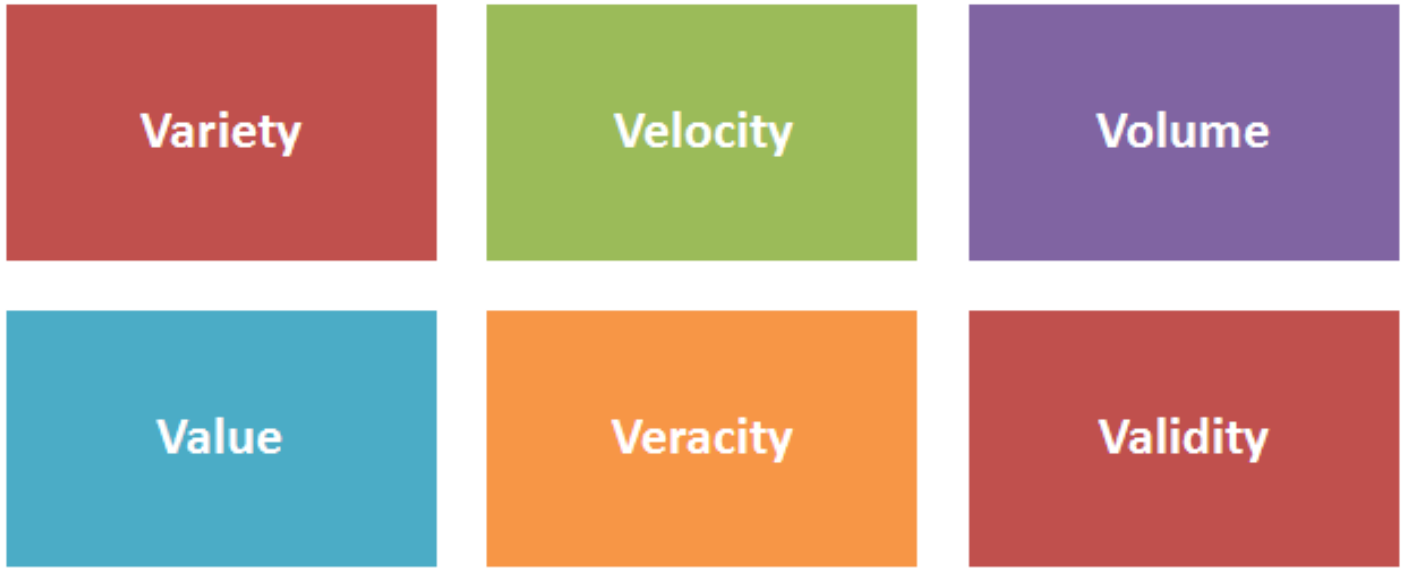
V-модель больших данных

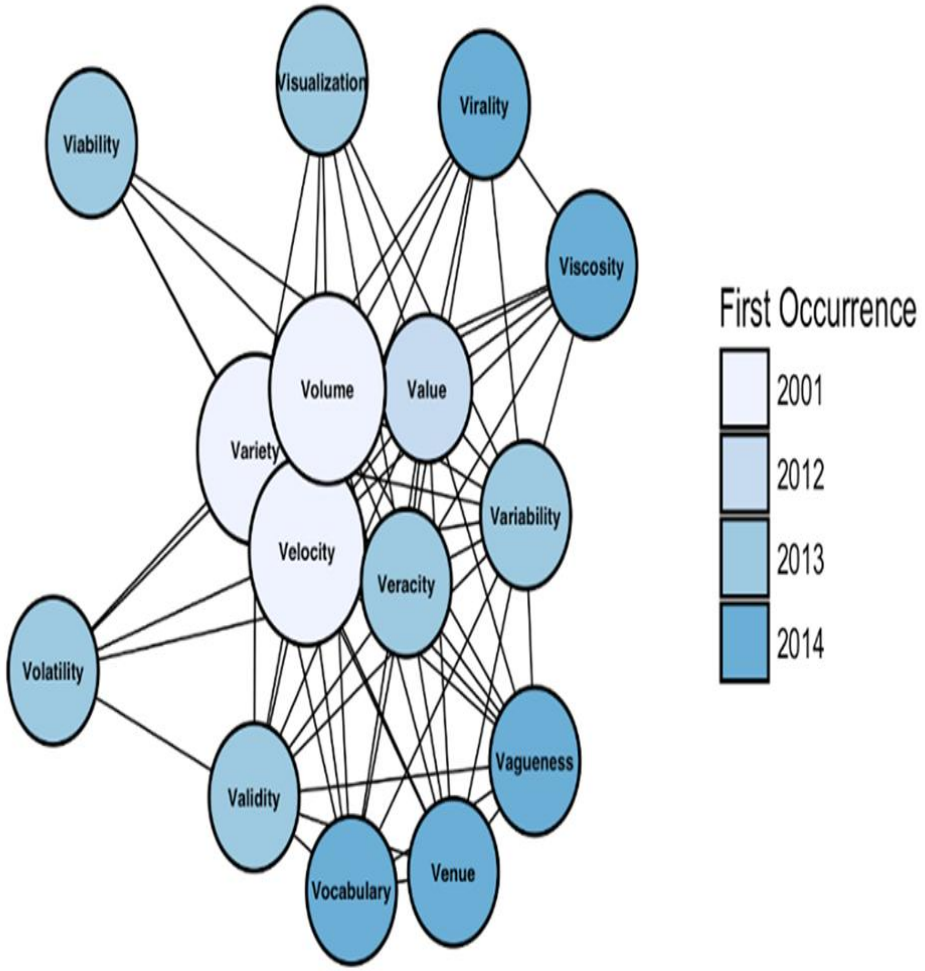


Характеристика	Традиционная база данных	База Больших Данных
Объем информации	От гигабайт до терабайт	От петабайт до эксабайт
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурирована	Полуструктурирована или неструктурирована
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Сквозная цифровая технология Большие данные

V-модель больших данных





Variety **(разнообразие)**

Velocity **(скорость обработки)**

Volume **(объём хранения)**

Value **(ценность данных)**

Validity **(надёжность данных)**

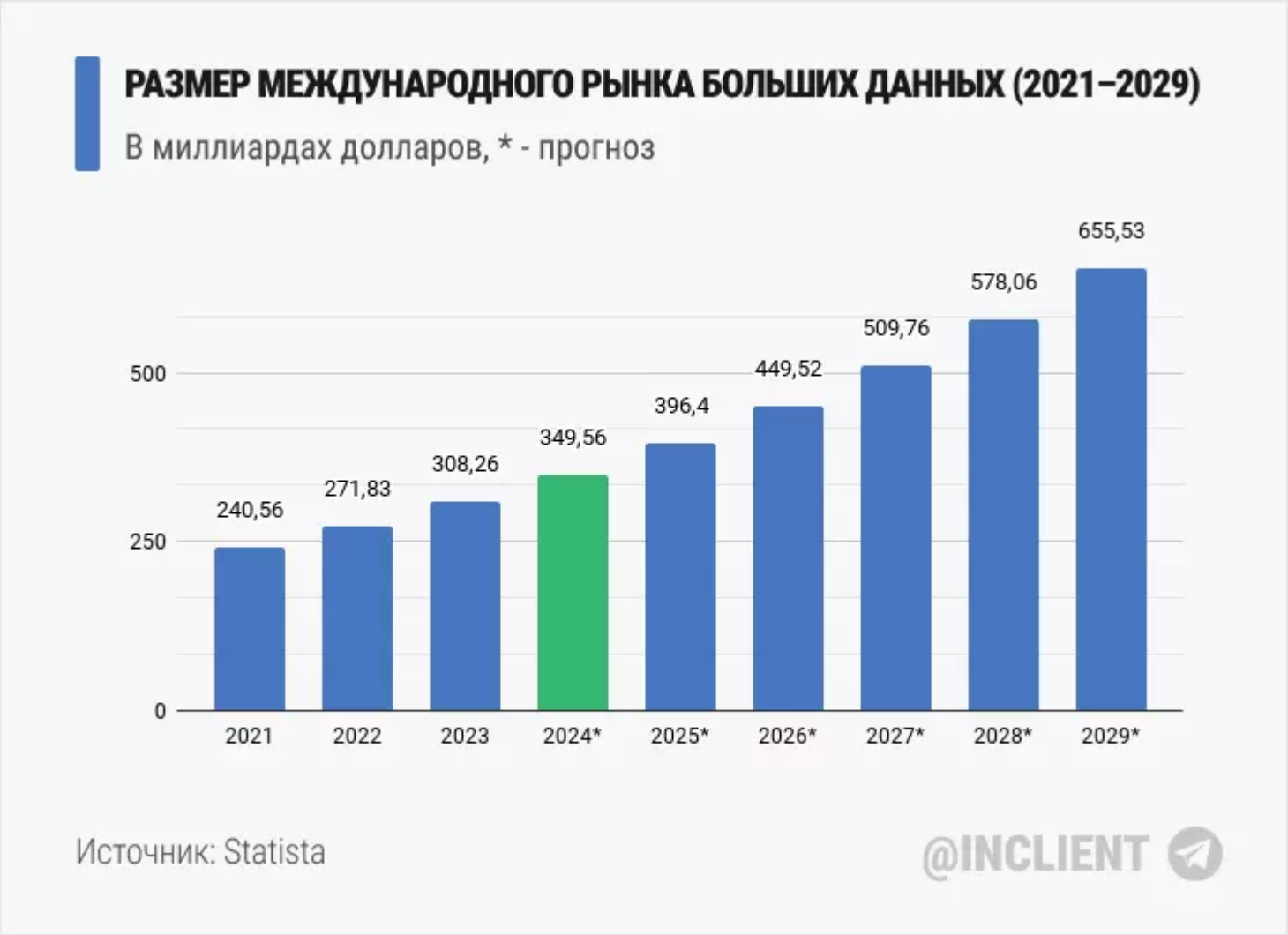
Veracity **(точность данных)**

и т.д.

Аналитика больших данных

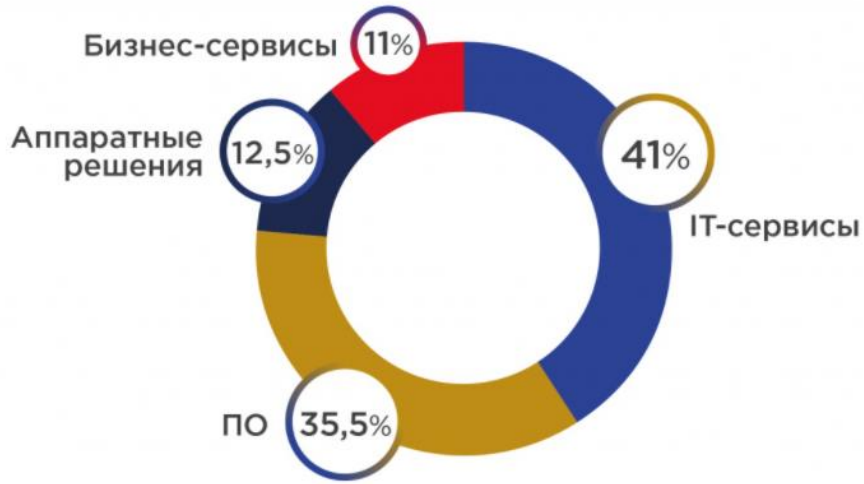
Источник: https://rubda.ru

Динамика мирового рынка больших данных



https://inclient.ru/data-create-stats/

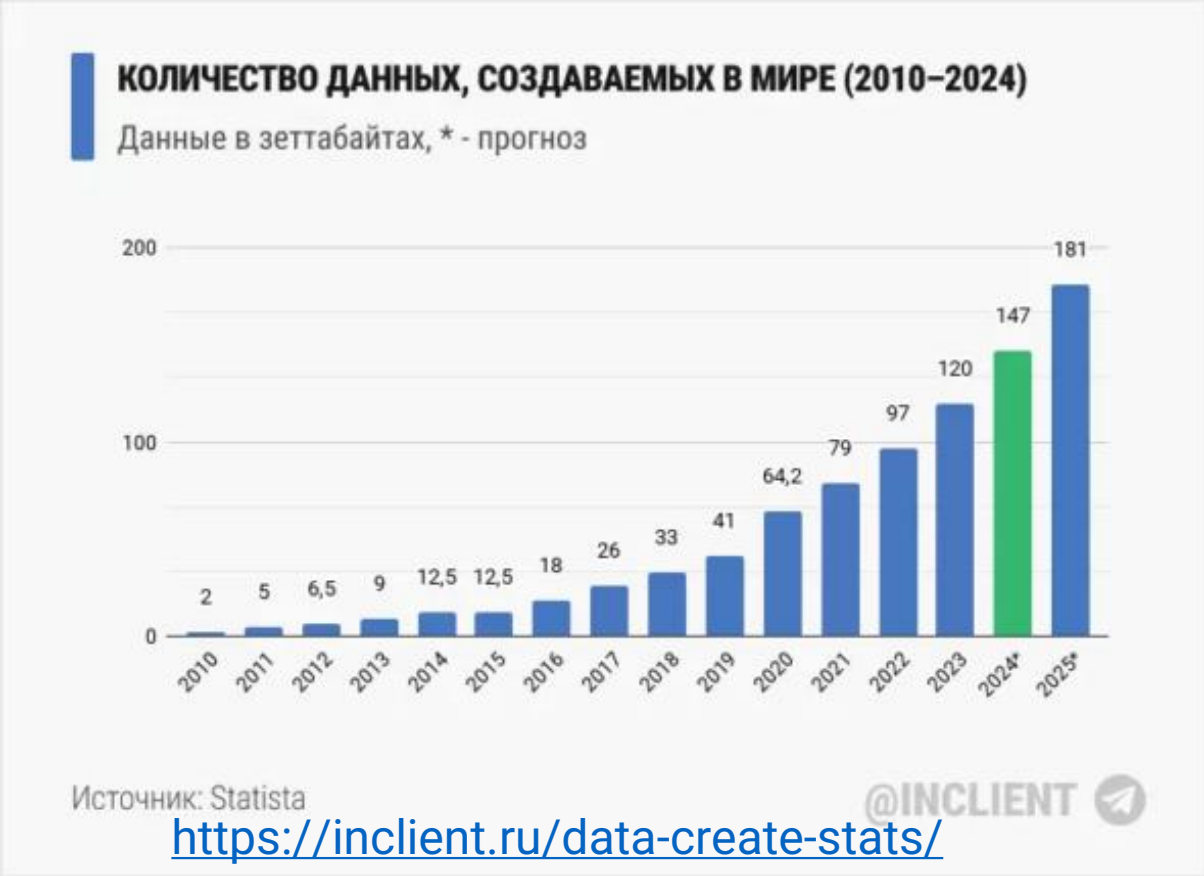
Доля сегментов рынка в общем объеме выручки, %



Grand View Research:

к 2025 году глобальный рынок Big Data как услуги (global big data as a service, BDaaS) достигнет 51,9 млрд долл., при этом CAGR составит 38,7% в период 2019-2025 гг.

Динамика мирового рынка больших данных

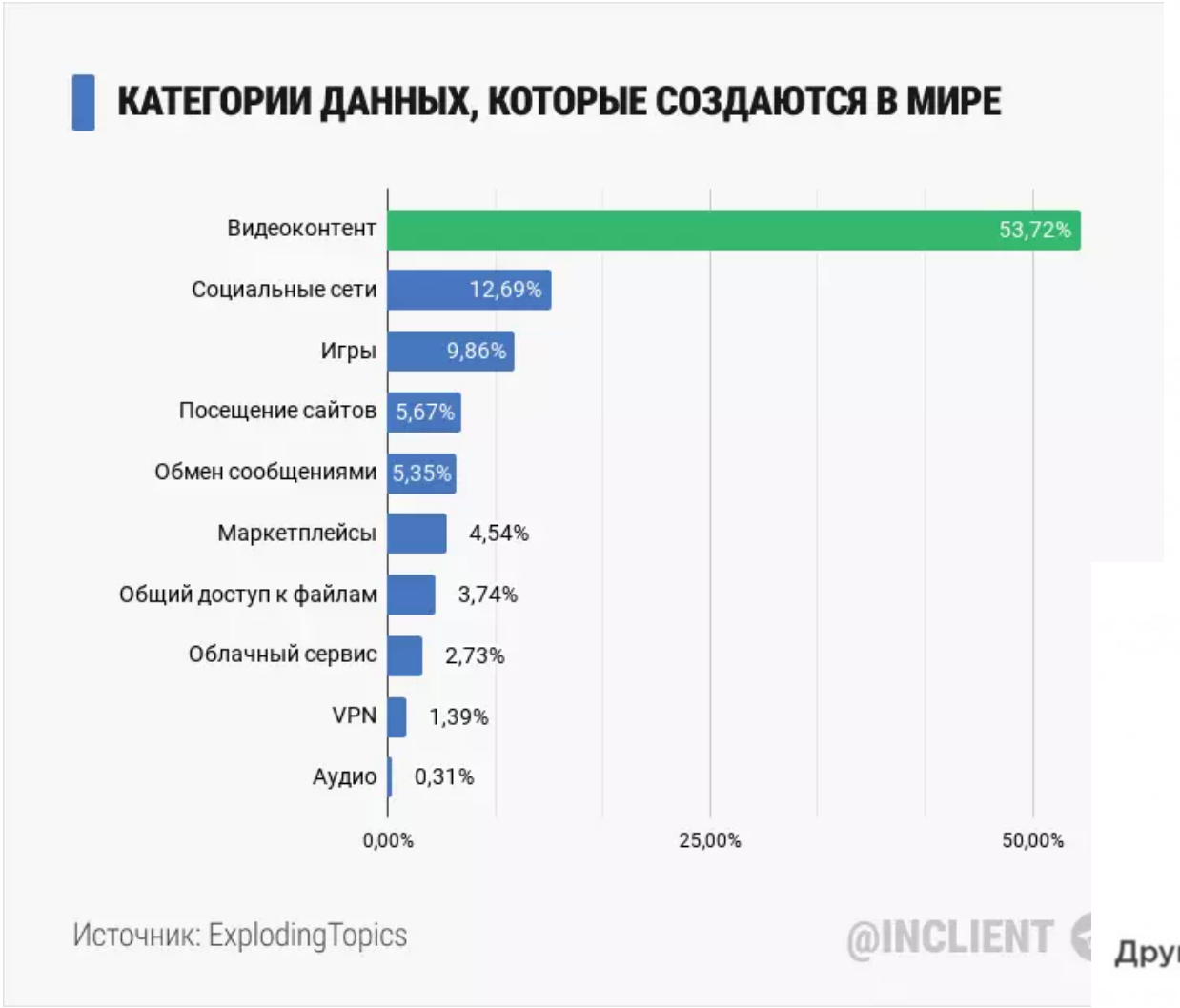


https://inclient.ru/data-create-stats/

- **Ежедневно в мире создаётся 328,77 миллиона терабайт данных**
- **Объем данных растет в геометрической прогрессии: за последние три года было создано 90% всех мировых данных**
- **Видеоконтент занимает доминирующую роль, на его долю приходится 53,72% мирового трафика данных**
- **60% компаний в мире уже используют анализ больших данных**
- **Мировой рынок больших данных в 2023 году достиг $349,56 млрд**

С 2025 по 2033 год мировой рынок больших данных будет расти со среднегодовым темпом 12,44%. По прогнозам, к 2033 году объём рынка достигнет 573,47 млрд долларов США, начав с 224,46 млрд долларов США в 2025 году

Динамика мирового рынка больших данных



https://inclient.ru/data-create-stats/

Ключевые технологические тренды больших данных (Gartner)

1. «Расширенная» (дополненная) аналитика (Augmented analytics)

Совершенствование процесса анализа за счет автоматизации процесса поиска, обработки данных с использованием технологий машинного обучения (Machine Learning, ML) и искусственного интеллекта (Artificial Intelligence, AI)

2. «Расширенное» (дополненное) управление данными (Augmented data management)

Технологии позволяют осуществлять автоматизацию и самонастройку процесса управления корпоративными данными, включая управление метаданными, качеством данных, интеграцию данных и баз данных

3. Технологии обработки естественного языка (Natural language processing, NLP and conversational analytics)

Технология обработки естественного языка позволяет компьютерам понимать человека. Как результат, рядовые бизнес-пользователи смогут делать запросы к сложным массивам данных обычными словами и фразами – голосом или вводом с клавиатуры и получать такие же легко понимаемые результаты бизнес-анализа

4. Аналитика графов (Graph analytics)

Применение методов обработки графической информации и графических баз данных на структурированных и неструктурированных данных, часто из нескольких приложений и источников

Ключевые технологические тренды больших данных (Gartner)

5.Коммерческие инструменты искусственного интеллекта и машинного обучения (Commercial AI and machine learning)

Переход от использования платформ с открытым исходным кодом к применению специально разработанных коннекторов, подключающихся к open-source экосистеме, позволит реализовать функции управления моделями, проектами, а также предоставит возможность для преобразования и многократного использования данных, обеспечит интеграцию и прозрачность, недоступные в рамках open-source платформ

6. Фабрика данных (Data fabric)

Подходы к интеграции данных в виде логически организованной структуры для облегченного доступа и обмена в распределенной среде данных

7. Объясняемый искусственный интеллект (Explainable AI)

Возможность формирования описательной модели на естественном языке, позволяющей обосновать автоматически сгенерированные решения и результаты, полученные на базе технологий AI

8. Блокчейн в области данных и аналитики

Реализует взаимосвязь транзакций, активов, обеспечивает прозрачность и гарантии в сложных сетях взаимодействия участников

Ключевые технологические тренды больших данных (Gartner)

9. Непрерывная интеллектуальная обработка данных (Continuous Intelligence)

Подход, при котором результаты аналитики в реальном времени интегрируются в бизнес-операции, происходит обработка потоковой контекстной информации, поступающей с датчиков IoT, и исторических данных, позволяющий моментально реагировать на изменения и предписывать поведение моделей

10. Серверы «постоянной» памяти (Persistent memory servers)

Технологии сохранения данных при отключении питания позволяет решить проблему ограниченности объемов памяти при возрастающем количестве данных; предоставляет возможность анализировать больше данных в оперативной памяти и в режиме реального времени; повышает энергоэффективность, операции с данными становятся более рациональными за счет уменьшения дублирования

11. Ужесточение регулирования в сфере обращения с данными

Нормативное регулирование ставит компании и организации перед необходимостью внедрить строгий контроль за данными, обеспечением их защищенности и конфиденциальности. Все это окажет влияние на практику сбора, обработки, хранения и использования данных компаниями, и, в первую очередь, это касается данных потребителей.

Российский рынок больших данных

В 2024 г. российский рынок Больших данных составил 300 млрд руб. при среднем темпе роста в 21% в год, ожидается сохранения темпов в следующие четыре года

Основные тренды

- **Внедрение искусственного интеллекта (ИИ) и машинного обучения**

ИИ автоматизирует сбор и проверку данных, прогнозирует поведение рынка и предпочтения клиентов. Модели на основе машинного обучения адаптируются к новым данным, что делает прогнозы актуальными

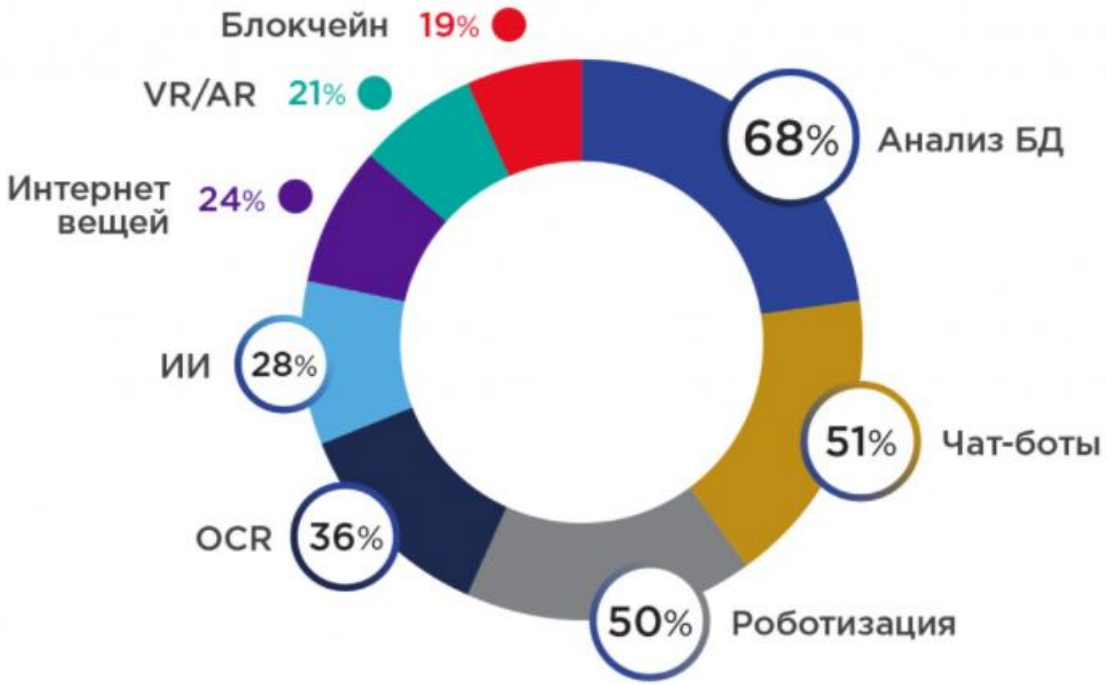
- **Распространение модели «Данные как услуга» (DaaS)**

Облачный сервис, предоставляющий доступ к структурированным и неструктурированным наборам данных по запросу

- **Рост популярности дополненной аналитики**

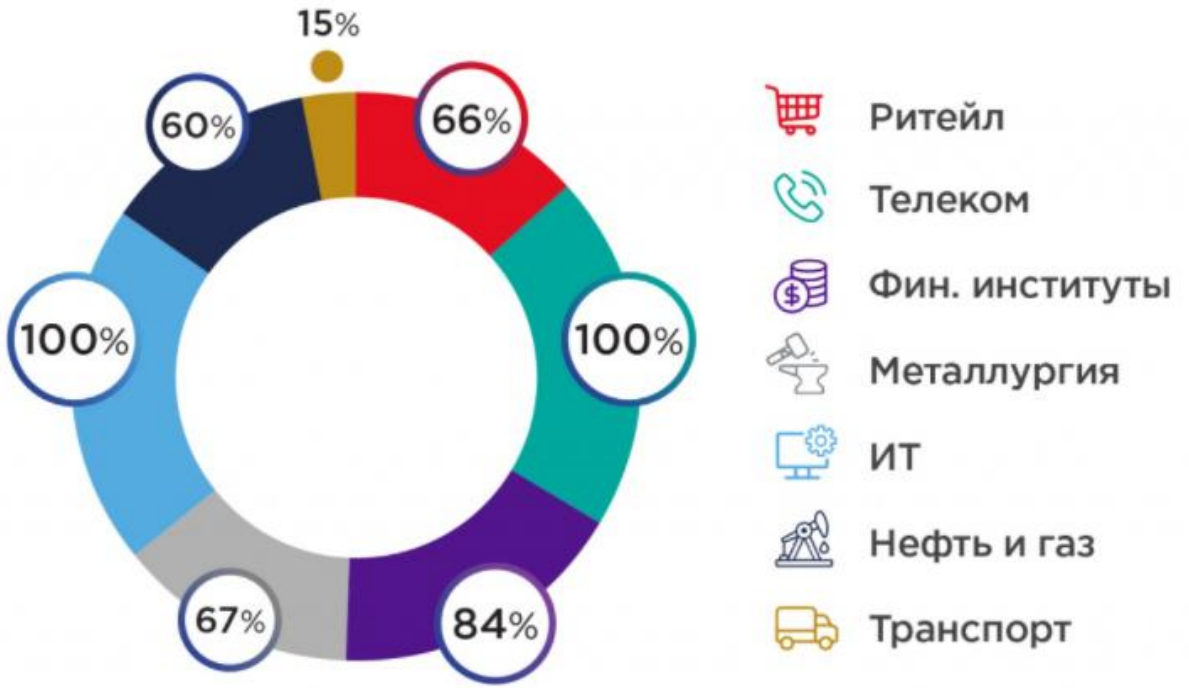
Автоматизирует процесс обработки, обнаружения и визуализации данных, позволяя компаниям быстрее извлекать из них полезную информацию

Российский рынок больших данных



Технологии, используемые среди российских компаний, %

Индустрии использования больших данных в России, %



Ассоциация больших данных: прогноз до 2024 г.

- **По данным BCG, на 2019 г. рынок больших данных в России оценивается в 45 млрд руб., а его ежегодный темп прироста с 2015 г. составляет 12%**
- **Ассоциация больших данных представила стратегию развития рынка Big Data на пять лет. Согласно подсчетам ассоциации и The Boston Consulting Group, ее реализация при самом пессимистичном сценарии - барьеры со стороны регуляторов на использование данных и отсутствие адресной поддержки - даст прирост ВВП в размере 20 млрд руб. к 2024 г. (0,3%) по сравнению с текущим годом**
- **Членами Ассоциации больших данных являются компании "Яндекс", Mail.Ru Group, Сбербанк, Газпромбанк, Тинькофф Банк, "МегаФон", "Ростелеком", "Билайн", oneFactor, QIWI, Аналитический центр при правительстве РФ**

Ассоциация больших данных: прогноз до 2024 г.

2018 Boston Consulting Group для Ассоциации Больших Данных

Ассоциация больших данных: прогноз до 2024 г.

Ассоциация больших данных: отчет 2023 г.

Ассоциация больших данных: отчет 2023 г.

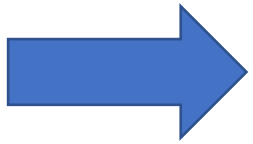
Ассоциация больших данных: отчет 2023 г.

Ассоциация больших данных: отчет 2023 г.

Большие данные - проект дорожной карты

- **Проект дорожной карты развития технологии «Большие данные» был подготовлен «Национальным центром информатизации» (НЦИ, «дочка» госкорпорации «Ростех») вместе с входящей в «ИКС-Холдинг» компанией «Форпост». Документ был разработан в рамках реализации мероприятий федерального проекта «Цифровые технологии» национальной программы «Цифровая экономика»**

- **Субтехнология сбора данных**
- **Субтехнология хранения данных**
- **Субтехнология обработки и управления данными**
- **Субтехнология вывода данных**



- **Нейротехнологии и искусственный интеллект**

Решение наблюдательного совета АНО «Цифровая экономика» под руководством помощника Президента Российской Федерации Андрея Белоусова и заместителя Председателя Правительства Российской Федерации Максима Акимова, май 2019

Проект закона о регулировании рынка больших данных

- **Минкомсвязи в феврале 2020 года разработало проект закона, направленный на регулирование рынка больших данных (big data). В документе министерство вводит определения понятий: большие данные, оператор больших данных и обработка больших данных. Контролировать оборот big data будет Роскомнадзор. Для этого ведомство создаст реестр операторов больших данных. Игроки рынка называют законопроект "сырым" и непродуманным**

- **Согласно законопроекту: "Большие данные - совокупность неперсонифицированных данных, классифицирующая по групповым признакам, в том числе информационные и статистические сообщения, сведения о местоположении движимых и недвижимых объектов, количественные и качественные характеристики видов деятельности, поведенческие аспекты движимых и недвижимых объектов, полученных от различных владельцев данных либо из различных структуированных или неструктуированных источников данных, посредством сбора с использованием технологий, методов обработки данных, технических средств, обеспечивающих объединение указанной совокупности данных, ее повторное использованием, систематическое обновление, форма представления которых не предполагает их отнесение к конкретному физическому лицу"**

Проект закона о регулировании рынка больших данных

- **В проекте поправок в ФЗ "Об информации, информационных технологиях и о защите информации" говорится, что под большими данными подразумеваются все данные, которые можно получить от владельцев структурированных и неструктурированных источников, используя любые технологии и средства**
- **Операторами больших данных могут быть госорганы, муниципальные органы, юрлица или физлица, саморегулируемые организации или общественные объединения (НКО и иностранные агенты тоже), которые организуют или сами обрабатывают big data. Определяют цели обработки больших данных, их состав и алгоритм действий с ними**
- **Под обработкой больших данных подразумевается действие или совокупность действий, которую совершают операторы больших данных с помощью средств автоматизации или без их использования. Речь идет о сборе, записи, систематизации, накоплении, хранении, обновлении, изменении, а также об извлечении, использовании, передаче, удалении, уничтожении и анализе таких данных**

Стандарт «Информационные технологии. Большие данные. Обзор и словарь»

- **В мае 2020 был представлен стандарт «Информационные технологии. Большие данные. Обзор и словарь» устанавливает термины и определения основных понятий в области технологий работы с большими данными. Соответствующий проект представили Национальный центр цифровой экономики МГУ имени М.В. Ломоносова и Институт развития информационного общества**

- **Национальный стандарт входит в серию национальных стандартов, гармонизирующих международные документы в области больших данных, и идентичен положениям действующего международного стандарта ISO/IEC 20546:2019.Information technology – Big data – Overview and vocabulary**

- **15 июня 2020 года Минкомсвязи сообщило об отзыве законопроекта о регулировании рынка Big Data. Речь идёт о поправках в Закон об информации, информационных технологиях и защите информации, вводящих новые правила обращения с большими данными**

Защита персональных данных

- **Федеральный закон от 8 августа 2024 г. N 233-ФЗ "О внесении изменений в Федеральный закон "О персональных данных" и Федеральный закон "О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации - городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона "О персональных данных"**

- **- в приоритете вопросы обеспечения защиты прав граждан при обработке больших данных и применении технологий ИИ;**

  - согласованы процедуры обезличивания больших данных и чётко регламентирован порядок формирования необходимых датасетов для машинного обучения, а также условия доступа к ним разработчиков нейросетей;

  - предусмотрен прямой и однозначный запрет на обработку больших данных в случае, если это может привести к риску причинения вреда жизни, здоровью, безопасности и имуществу граждан;

  - возможен доступ внешних разработчиков к большим данным только в случае исключения любых сведений, которые позволят идентифицировать конкретного гражданина;

  - запрещён внешний доступ к сформированным датасетам иностранным лицам и российским юрлицам с преимущественным иностранным участием.

Субтехнология сбора данных

- **Субтехнология сбора данных представляет собой технологии, обеспечивающие прослеживаемость и интероперабельность данных**

- **В нее входят стандарты, протоколы и системы сбора данных из различных источников, обеспечивающих прослеживаемость данных от источника до потребителя, включая интероперабельность данных**

Физический уровень состоит из конечных устройств промежуточных шлюзов

На сетевом уровне выполняются задачи по организации сетей и транспортировке информации

На платформенном уровне осуществляется управление устройствами, SIM-картами, интеграцией в сеть оператора связи или виртуального оператора интернета вещей, а также обеспечивается работа приложений

На уровне управления данными собираются и хранятся данные из различных источников

Физический уровень сбора данных

- **Этот уровень отвечает за отделение шума от соответствующей информации, а также регулирование объема, скорости и разнообразия данных. Он должен иметь возможность проверять, очищать, преобразовывать, уменьшать и интегрировать данные в стек технологий больших данных для дальнейшей обработки**
- **Это новое программное обеспечение, которое должно быть масштабируемым, устойчивым, отзывчивым и регулирующим в архитектуре больших данных. Если детальная архитектура этого слоя не спланирована должным образом, весь стек технологий может оказаться хрупким и нестабильным**
- **Источниками больших данных являются разнообразные информационные системы, как внутренние, так и открытые, как государственные, так и частные и корпоративные, бизнес-ориентированные или научные**
- **Сами по себе данные представляют собой структурированные и неструктурированные файлы, цифровое видео, изображения, данные датчиков, log-файлы и вообще любые данные, не содержащие в записях специальных поисковых полей**
- **Появляются новые источники больших данных, такие как машинное генерирование (например, log-файлы или данные сенсорных сетей), мобильные устройства (видео, фотографии и текстовые сообщения) и системы машина-машина, когда "Интернет вещей" сообщает о состоянии в целях планирования обслуживания парка автомобилей или самолетов либо в целях телеметрического мониторинга**

Физический уровень сбора данных

- **На данном уровне располагается физическая инфраструктура, необходимая для функционирования и масштабируемости архитектуры больших данных**
- **Для поддержки непредвиденного или непредсказуемого объема, скорости или разнообразия данных физическая инфраструктура для больших данных должна отличаться от инфраструктуры для традиционных данных**
- **Это служит источником развития таких распределенных архитектур, где в роли узлов выступают не просто компьютеры общего назначения, а специализированные серверы**
- **Еще одним способом для решения возникших проблем является горизонтальное масштабирование, когда вычисления выполняются параллельно на нескольких физических серверах**

Физический уровень сбора данных

- **Симметричная многопроцессорная архитектура (symmetric multiprocessing, SMP) – представляет собой архитектуру с общей физической памятью, с которой могут взаимодействовать сразу несколько процессов**
- **Память может использоваться для передачи сообщений между процессами**
- **При обращении к общей памяти все процессы имеют идентичные права и идентичную адресацию для всех ячеек памяти, отчасти поэтому данная архитектура называется симметричной**
- **Использование SMP-архитектуры является довольно простой и универсальной с точки зрения проектирования и разработки**
- **Использование такой архитектуры способствует решению проблемы недостатка вычислительных ресурсов, но имеет ограничения по масштабированию при работе с «большими данными»**

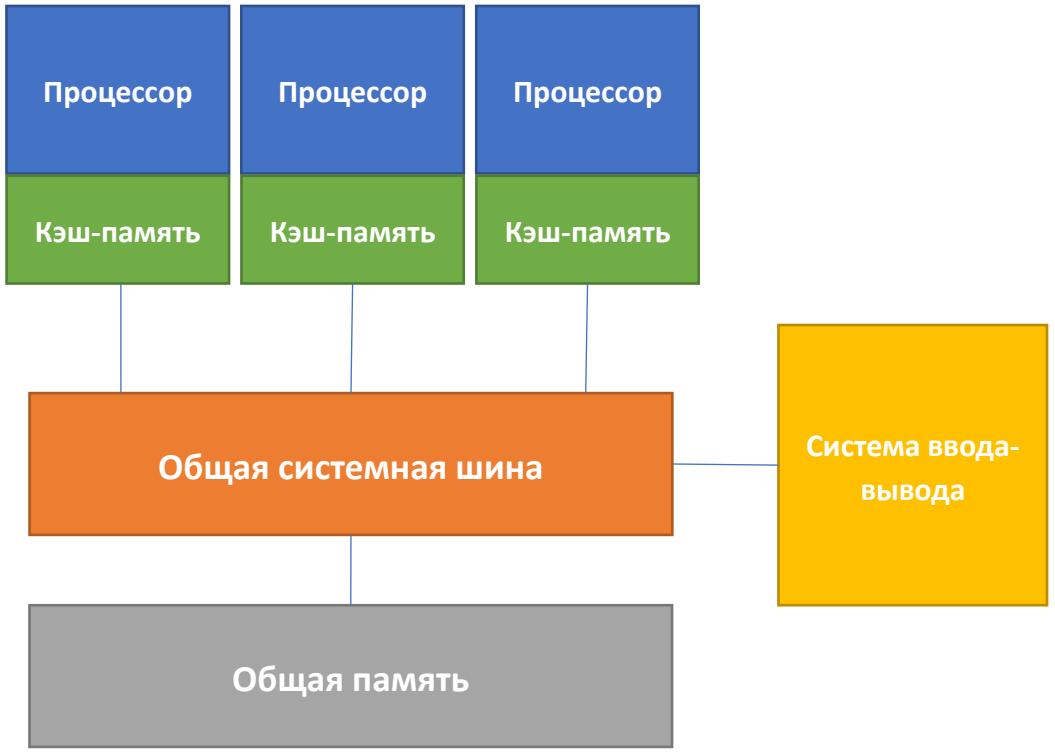


Схема симметричной многопроцессорной архитектуры

Физический уровень сбора данных

- **Массивно-параллельная архитектура (Massive Parallel Processing, MPP) – это класс параллельных вычислительных систем, состоящих из множества узлов, где каждый узел представляет собой автономную, независимую от других единицу (каждый узел имеет отдельный ЦП, память и диски для локальной обработки данных) – принцип «shared nothing».**

- **Если применить это определение к области хранилищ данных, то лучше всего его смысл будет отражать термин «распределённые базы данных». Каждый узел в распределенной базе данных представляет собой полноценную СУБД, работающую независимо от других**

- **Сама же распределенная база данных – это совокупность независимых, автономных узлов, связанных высокоскоростной коммуникационной сетью**

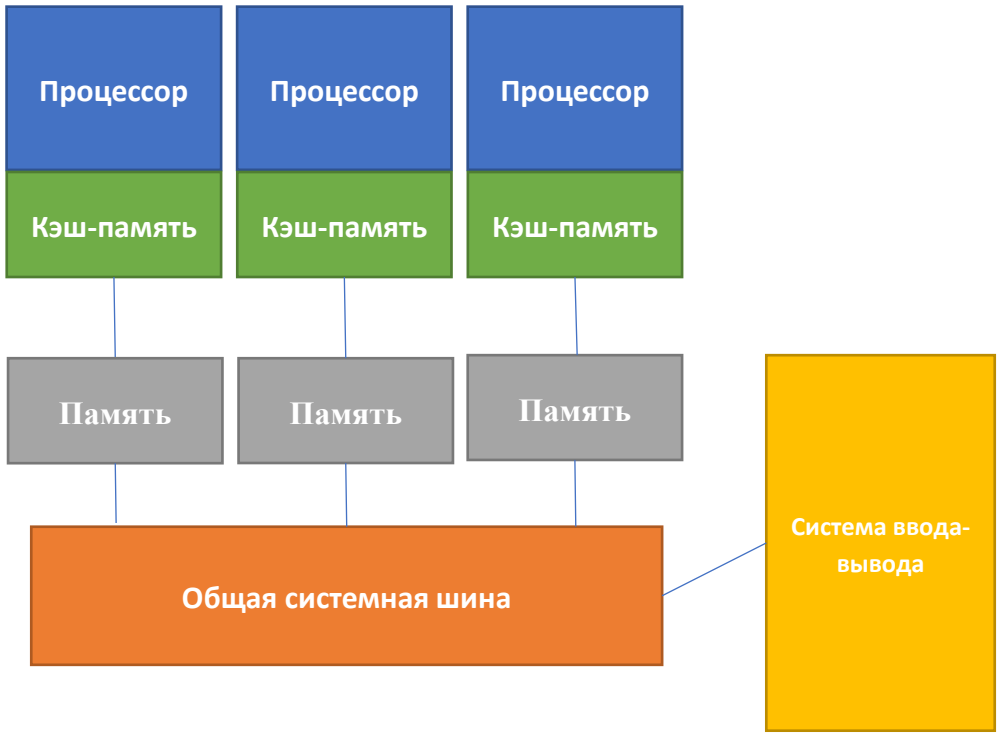


Схема массивно-параллельной архитектуры

Физический уровень сбора данных

- **Самым популярным и часто используемым является экосистема Hadoop, в которую входит компоненты распределенной файловой системы HDFS, реализация модели MapReduce, а также множество инструментов для хранения, обработки и манипулирования данными. Данное решение хорошо подходит для использования в качестве пакетного обработчика**

- **Уровень физической инфраструктуры Hadoop (HPIL) также основан на модели распределенных вычислений. Принцип «share-nothing» подразумевает, что данные могут физически храниться в разных местах и связываться между собой через сети и распределенную файловую систему**

- **Как и в традиционной модели клиент-сервера, данные больше не нужно передавать на монолитный сервер, где функции SQL применяются для его завершения. Резервирование также встроено в эту инфраструктуру.**

- **Hadoop-это платформа с открытым исходным кодом, которая позволяет нам хранить огромные объемы данных распределенным способом на машинах с низкой стоимостью. Она обеспечивает разъединение между разработкой программного обеспечения распределенных вычислений и фактической логикой приложения, которую вы хотите выполнить**

- **Hadoop позволяет взаимодействовать с логическим кластером узлов обработки и хранения, а не взаимодействовать с операционной системой (OS) и процессором**

Субтехнология сбора данных

- **В данной сфере планируется обеспечить интероперабельность устройств и платформ интернета вещей, разработать стандарт поддержки до 10 протоколов обмена данными для устройств интернета вещей и разработать стандарт обмена данными между платформами интернета вещей**
- **Также ожидается поддержка решения по обеспечению обмена данных между устройствами и платформами интернета вещей, разработка универсального программного шлюза для работы устройств в критичных отраслях с поддержкой до 10 тыс. разных видов устройств**
- **Есть планы разработать универсальный программный модуль обмена данными для платформ интернета вещей в критичных отраслях**
- **Платформы будут поддерживать интероперабельность на 80%**

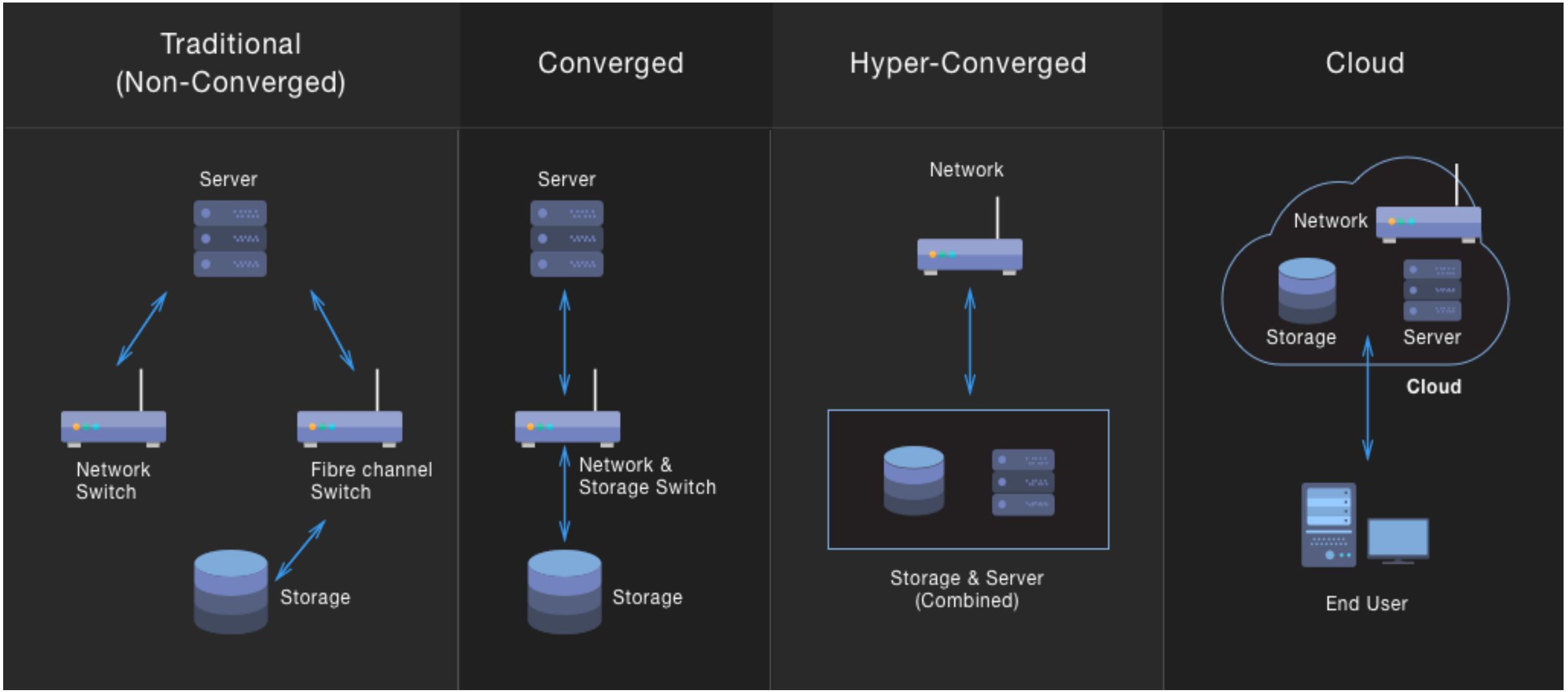
Субтехнология хранения данных

- **Субтехнология хранения данных представляет собой программно-определяемые хранилища данных (SDS)**
- **SDS включает в себя пулы виртуализированных хранилищ с характеристиками, которые могут быть заданы через управляющий интерфейс, облачные хранилища и конвергентные программно-определяемые хранилища**
- **Ожидается, что к 2020 г. 90% организаций адаптируют свою гибридную инфраструктуру к облачному хранению данных, а вместимость центров хранения данных вырастет до 1,8 Збайт. К этому времени рынок облачных хранилищ вырастет на 30% и достигнет $92 млрд, при этом 83% данных компаний будет хранится в облаке**
- **Мировой рынок соответствующих технологий с $442 млн в 2018 г вырастит до $15,7 млрд в 2024 г. Доля российских разработчиков на этом рынке за аналогичны период вырастит с 2,2% до 3,4%**
- **Запланирована поддержка разработки российских программно-определяемых хранилищ, увеличение надежности хранения на 50% и увеличение скорости развертывания хранилища до 6 часов**

Субтехнология хранения данных

- **Программно-определяемая хранилище (software-defined storage, SDS) — программное решение, обеспечивающее создание сети хранения данных на неспециализированном оборудовании массового класса, как правило, группе серверных узлов архитектуры x86-64 под управлением операционных систем общего назначения (Linux, Windows, FreeBSD). Основная отличительная возможность — виртуализация функции хранения, отделяющая аппаратное обеспечение от программного, которое управляет инфраструктурой хранения; в этом смысле является развитием концепции программно-определяемой сети, специализированном для систем хранения.**
- **Аппаратное обеспечение в такой сети хранения обычно без какой-либо аппаратной агрегации или защиты предоставляет доступные накопители в программную часть, которая, как правило, объединяет их в пулы, и уже в рамках агрегированных пулов реализуются необходимые функции выделения томов, их презентации, ведения ограничений, управления производительностью, отработки отказов. Среди возможных функций программного уровня кэширование, редупликацию, репликация, мгновенные снимки, резервное копирование, тонк ое резервирование.**
- **Центральную роль программно-определяемые сети хранения играют в гиперконвергентных системах, где обеспечивают отказоустойчивую среду хранения томов виртуальных машин в предконфигурированных системах на базе серверного оборудования массового класса, выполняющих функции одновременно узлов сети хранения и узлов виртуализации вычислительных ресурсов. Среди серийно используемых продуктов для построения программно-определяемых сетей хранения — vStorage (Virtuozzo), vSAN (VMware), GlusterFS и Ceph (обе — Red Hat), Storage Spaces Direct (Microsoft).**

Субтехнология хранения данных



Субтехнология хранения данных

Конвергентные хранилища данных (конвергентная инфраструктура) — это интегрированные решения, которые объединяют компоненты ИТ-инфраструктуры: серверы, системы хранения данных (СХД), сетевое оборудование и средства виртуализации серверов. В отличие от традиционных систем, где каждый компонент работает независимо, конвергентные решения объединяют все элементы в одну платформу

Преимущества:

- упрощение управления инфраструктурой через один интерфейс
- повышение масштабируемости и гибкости, возможность быстро расширять инфраструктуру
- сокращение затрат за счёт консолидации ресурсов

CI (Converged Infrastructure) *— на основе аппаратных компонентов. Каждый компонент остаётся физически отдельным и может быть заменён независимо от других. Например, в CI все компоненты находятся в одном физическом устройстве на большой стоечной платформе*
HCI (Hyperconverged Infrastructure) *— на основе программного обеспечения. Все компоненты работают как единое целое и управляются через единую консоль. В HCI общая платформа объединяет не разнородные компоненты, а унифицированные аппаратные узлы, например, серверы форм-фактора 2U на базе архитектуры x86.*

Субтехнология хранения данных

Гиперконвергентные хранилища данных — это хранилища, в которых вычислительные мощности, серверы и сети объединены в единое пространство с помощью программных средств

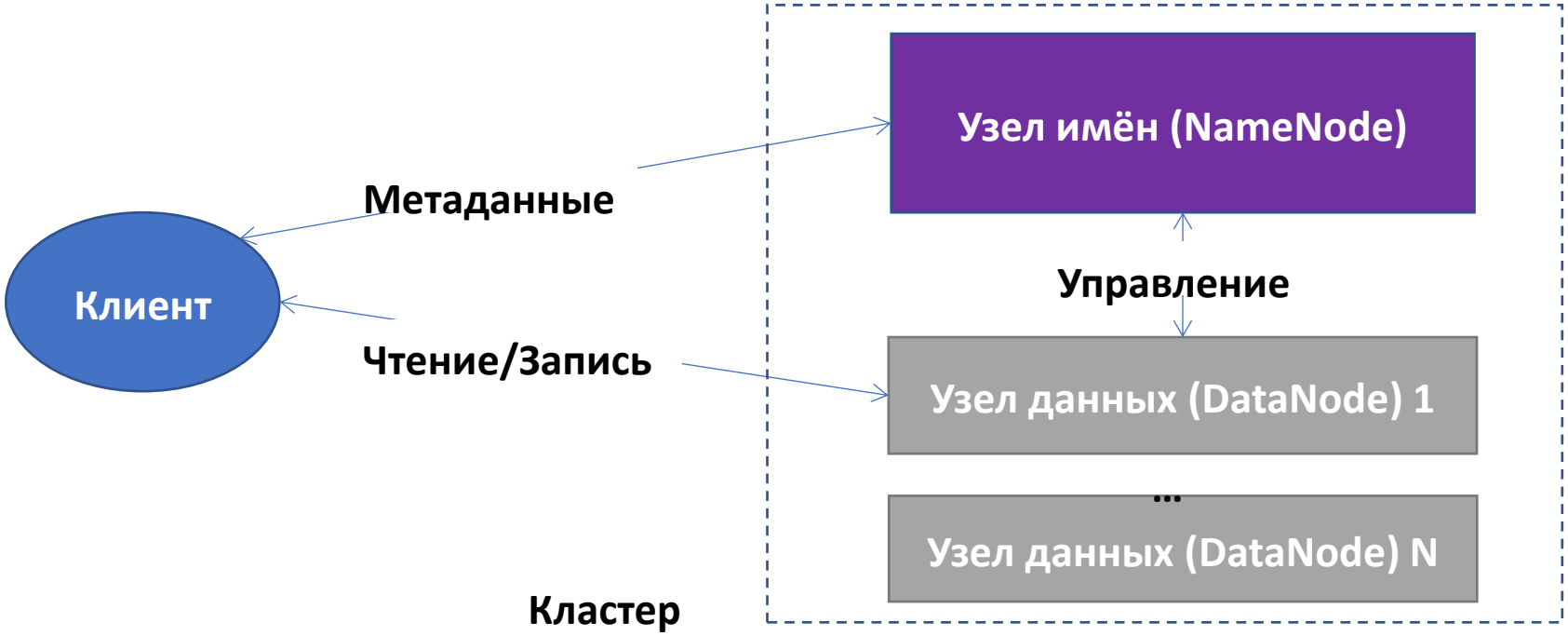
В гиперконвергентной инфраструктуре нет выделенных систем хранения, данные хранятся на внутренних дисках серверов

Некоторые особенности гиперконвергентных хранилищ данных:

- **Объединение механизмов распределённого хранения. В гиперконвергентной инфраструктуре уже присутствует встроенный программно-определяемый слой хранения, поэтому использование внешней физической СХД изначально не требуется**
- **Возможность подключения внешних систем. К слою хранения можно подключать другие платформы виртуализации, отдельные сервера и т. п.**
- **Высокая отказоустойчивость и защита данных. Встроенные механизмы репликации, восстановления и распределения нагрузки гарантируют стабильную работу, даже если что-то упадёт**
- **Упрощение управления. Все аппаратные и программные элементы объединены в одну платформу, что снижает сложность администрирования**

Субтехнология хранения данных: HDFS

- **HDFS проекта Hadoop (Hadoop Distributed File System - HDFS) - это файловая система, предназначенная для хранения очень большого объема информации (терабайт или петабайт) на большом количестве компьютеров в кластере. Он надежно хранит данные, использует блоки для хранения файла или частей файла и поддерживает модель доступа к данным с возможностью записи после чтения**



- **Для этого используется два типа сервера: NameNode и DataNode. На сервере NameNode хранятся метаданные файлов, к которым могут относится имя файла, его размер, время модификации и т.д, а также фиксирует все изменения в данных. Сами данные хранятся на серверах DataNode**

Субтехнология хранения данных: HDFS

- **Один сервер NameNode может обслуживать один кластер, куда могут входить сотни и тысячи серверов**

- **Непосредственно сами данные хранятся на серверах DataNode в виде последовательности блоков. Гарантией сохранности данных выступает то, что каждый блок храниться на нескольких серверах одновременно, а при выходе из строя одного сервера всегда можно будет восстановить утерянные блоки с других серверов с помощью операции репликации. HDFS поддерживает стандартные операции по чтению, записи и удалению файлов**

- **При выполнении любой операции взаимодействие клиента строится через сервер NameNode, который определяет серверы DataNode, на которых уже содержатся требуемые блоки файлов или на которые только нужно записать блоки файлов.Чтение и запись происходит непосредственно на сервер DataNode, минуя сервер NameNode**

- **Главной особенностью работы HDFS является то, что даже если размер одного файла будет превосходить размер жесткого диска на одном из серверов, а файл будет разнесен по разным серверам, то использование HDFS позволит работать с файлом как с целым**

- **HDFS требует сложных программ чтения / записи файлов, которые будут написаны квалифицированными разработчиками. Он недоступен в качестве логической структуры данных для удобства работы с данными. Для этого необходимо использовать новые распределенные нереляционные хранилища данных, распространенные в мире больших данных, включая пары "ключ-значение", базы данных "документ", "график", "столбчатый" и "геопространственный". В совокупности они называются nosql, а не только SQL, базами данных**

Субтехнология хранения данных: NoSQL

- **NoSQL - (англ. not only SQL, не только SQL), в информатике — термин, обозначающий ряд подходов, направленных на реализацию хранилищ баз данных, имеющих существенные отличия от моделей, используемых в традиционных реляционных СУБД с доступом к данным средствами языка SQL**
- **Применяется к базам данных, в которых делается попытка решить проблемы масштабируемости (англ. scalability) и доступности (англ. availability) за счёт атомарности (англ. atomicity) и согласованности данных (англ. consistency)**
- **Под термином NoSQL скрывается большое количество продуктов с абсолютно разными дизайнами и, иногда, при обсуждении разговор может идти о разных системах**
- **NoSQL основаны на принципах BASE, данный термин был предложен Эриком Брюером:**
  - **Basic Availability - базовая доступность — каждый запрос гарантированно завершается (успешно или безуспешно).**
  - **Soft State - гибкое состояние — состояние системы может изменяться со временем, даже без ввода новых данных, для достижения согласования данных.**
  - **Eventual Consistency - согласованность в конечном счёте — данные могут быть некоторое время рассогласованы, но приходят к согласованию через некоторое время**
- **Примерами могут являться решения Oracle NoSQL Database, Redis, Dynamo**

Субтехнология хранения данных: документо-ориентированные базы данных

- **Документо-ориентированные базы данных служат для хранения иерархических структур данных. Основной идеей является введение понятия "документ"**
- **Хотя все базы данных имеют чем-то отличающиеся определения, во всех них предполагается, что документ хранит и инкапсулирует данные в каких-либо стандартных форматах, например XML или JSON**
- **Каждому документу присваивается свой уникальный ключ, для каждой базы данных такого типа есть свой язык запросов или API для доступа к данным. Обычно в базах данных такого типа присутствует богатая структура документа**
- **Примерами могут являться решения CouchDB, Couchbase Server, MarkLogic, MongoDB**

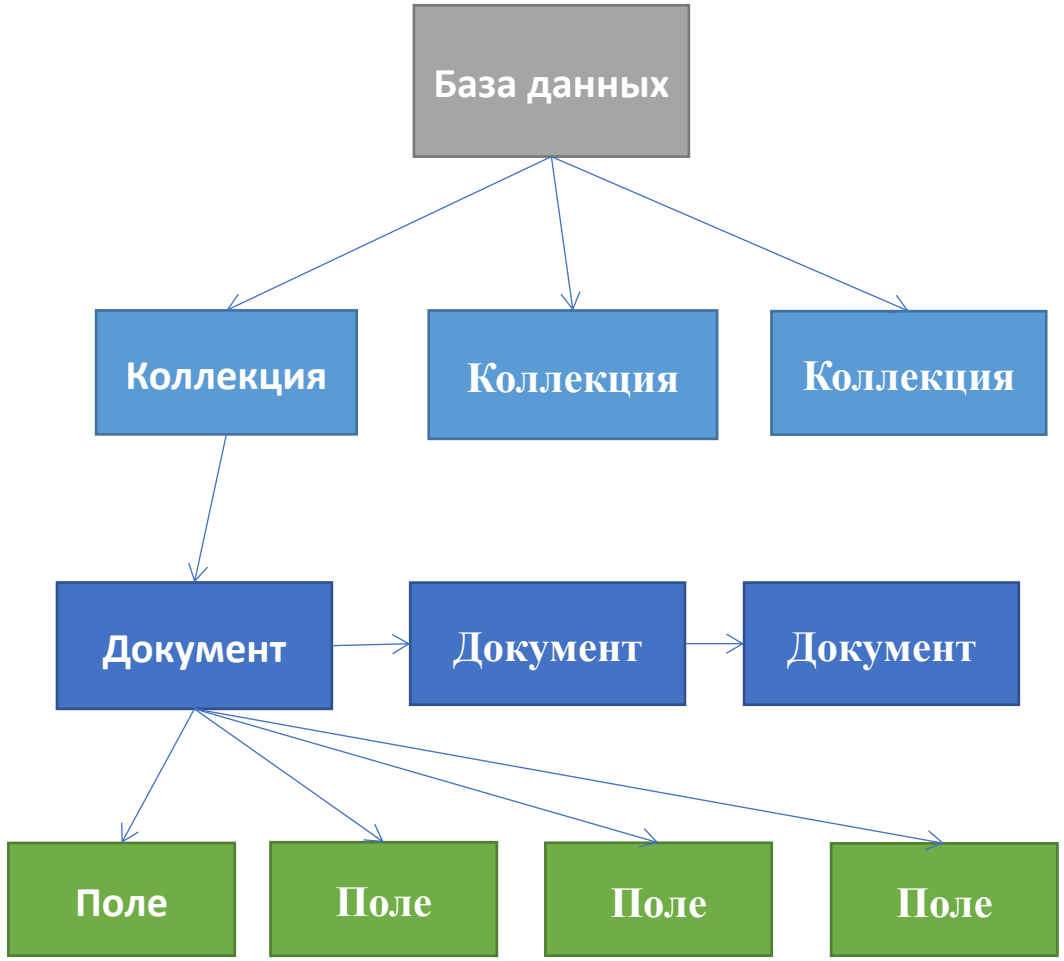


Схема данных MongoDb

Субтехнология хранения данных: графовые базы данных

- **Графовая база данных предназначена для данных, отношения которых хорошо представлены как граф, состоящий из элементов, связанных с конечным числом отношений между ними.**

- **Типом данных могут быть социальные сети, общественные транспортные связи, дорожные карты или сетевые топологии**

- **Примерами могут являться решения AllegroGraph, ArangoDB, Neo4j**

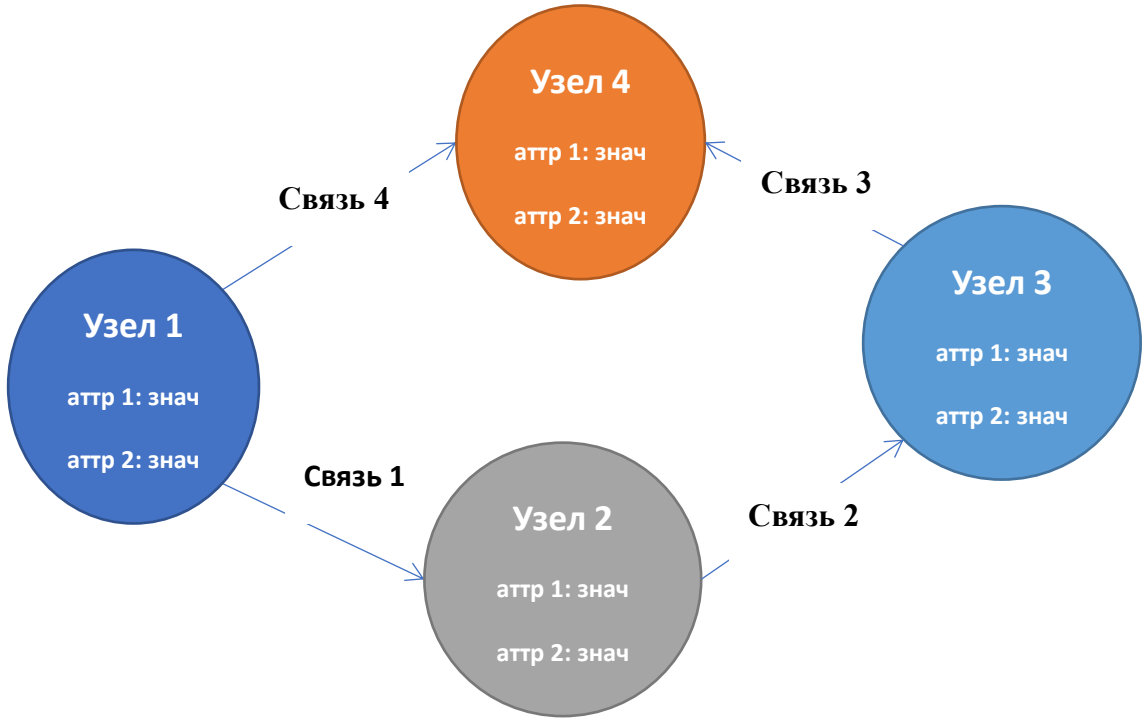


Схема данных Neo4j

Субтехнология хранения данных: изменение парадигмы

Хранилище данных Data Warehouse (DWH), 1990

Компании столкнулись с проблемой данных, разбросанных в разных системах (CRM, ERP, бухгалтерия и так далее)
Подход DWH позволил централизовать данные для отчетности и аналитики на основе различных реляционных баз данных
В итоге создавалось высококачественное структурированное хранилище, в которое данные загружались через процессы ETL (Extract, Transform, Load), проходили очистку и оптимизировались для аналитики

«Озеро данных» Data Lake , 2010

Компании решали проблему хранения неструктурированных данных (логами, видео, сенсорными данными)
«Озера данных» позволяют хранить любые данные в сыром виде (структурированные, неструктурированные, полуструктурированные)
При этом данные загружались в «озеро» и обрабатывались по мере необходимости. Данный подход позволил аналитикам использовать для анализа данных не только традиционные методы, но и технологии ИИ

Гибридная архитектура «озеро-хранилище» (Data Lakehouse) 2017-2018

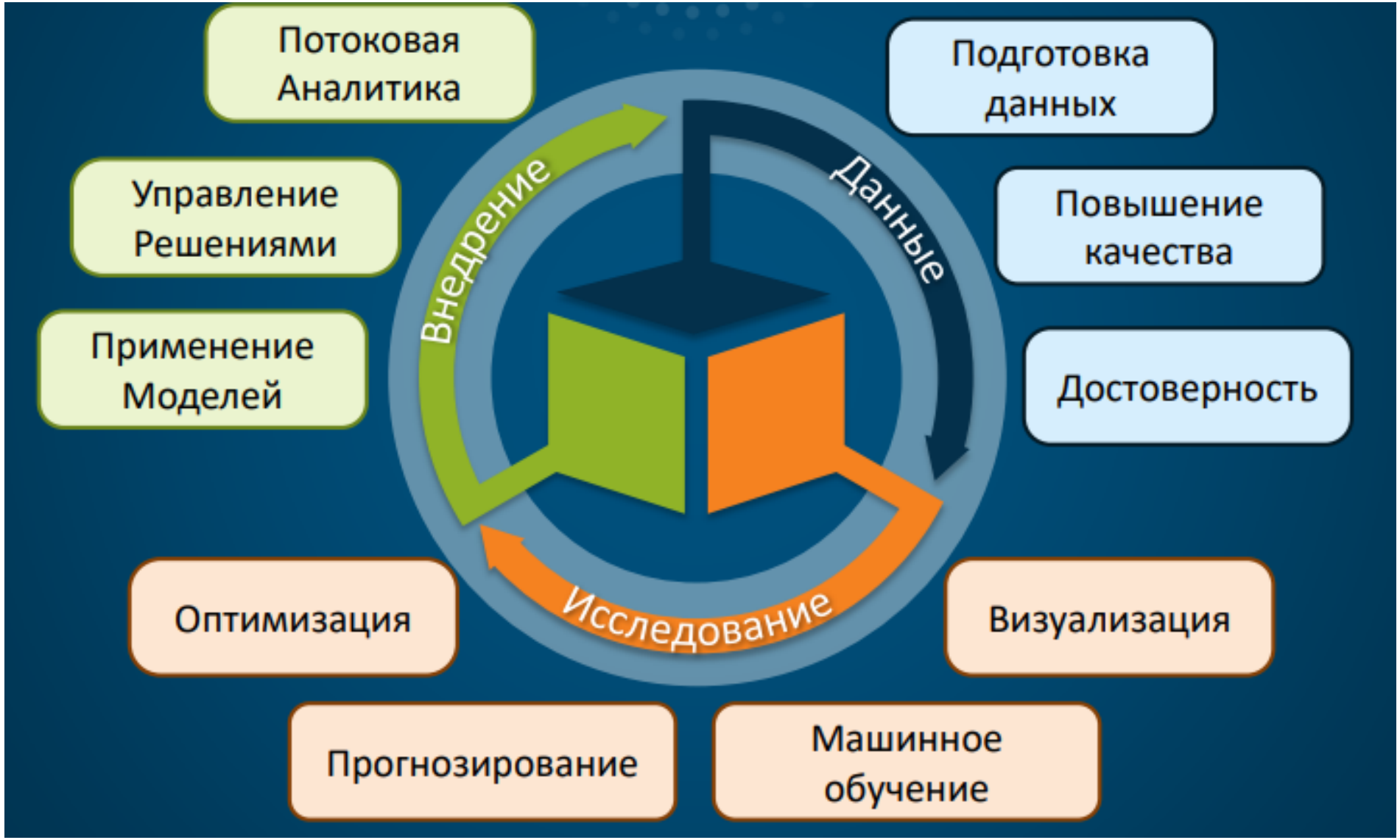
Объединяет преимущества подходов Data Lake (гибкость и дешевизна хранения больших объемов данных) и Data Warehouse (структурированность и высокая производительность запросов).
Причинами растущей популярности у компаний «озер-хранилищ» данных являются

- экономическая эффективность, возможность унифицированного доступа к данным
- повышенная простота их использования

Субтехнология обработки и управления данными

- **Субтехнология обработки и управления данными** включает в себя технологии обработки и утилизации данных с использованием искусственного интеллекта (AI) и машинного обучения (ML), а также технологии обогащения данных
- **Машинное обучение** представляют собой класс методов искусственного интеллекта, характерной чертой которых является не прямое решение зада, а обучение в процессе применения решений множества сходных задач
- Должны быть реализованы следующие решения: системы управления базами структурированных и неструктурированных данных, системы кластерной параллельной обработки потоковых данных, системы создания ансамблей (оркестов) алгоритмов и глубокого обучения, «озеро данных» для последующей обработки неструктуированных и нестабильных данных и технологии разметки, аннотации и создания data set для ML
- **Обогащение данных** — процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи

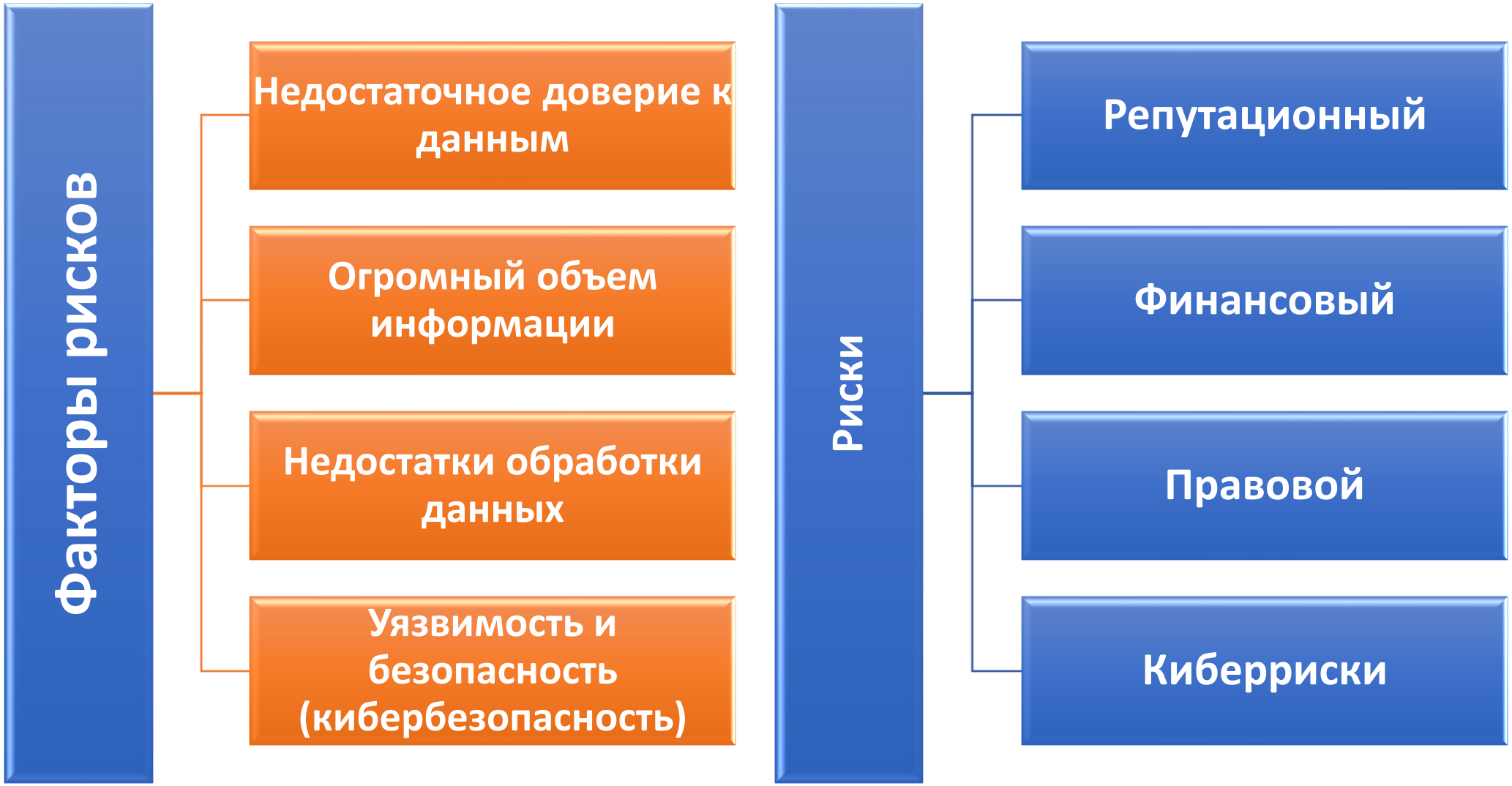
Субтехнология обработки и управления данными: обогащение данных



Субтехнология вывода данных

- **В нее входят технологии, обеспечивающие использование доверенных данных для BI**
- **Частью субтехнологии вывода данных является предиктивная аналитика, как финальное и самое ответственное звено в извлечении пользы из больших данных для бизнеса и государства**
- **В данной сфере запланировано поддержание решений обработки и утилизации данных с использованием предиктивной аналитики и разработка СПО (свободно распространяемое программное обеспечение), позволяющего предоставлять валидированыне прогнозные данные, сформированные на основе больших данных**
- **Должны быть созданы технологии создания предиктивных признаков, технологии анализа данных, технологии обучения моделей для подготовки прогнозов.**
- **Благодаря предлагаемым мерам скорость обработки данных для принятия решения сократится до нескольких минут, а точность прогноза повыситься до 65 – 85%**

Риски технологии Большие данные



Риски кибербезопасности технологии Большие данные

риск конфиденциальности

риск потери данных

риск переполнения хранилища

риск снижения эффективности больших данных

риск формирования неэффективного набора данных

риск мошенничества

риск неготовности к переменам

риск внешнего консультанта

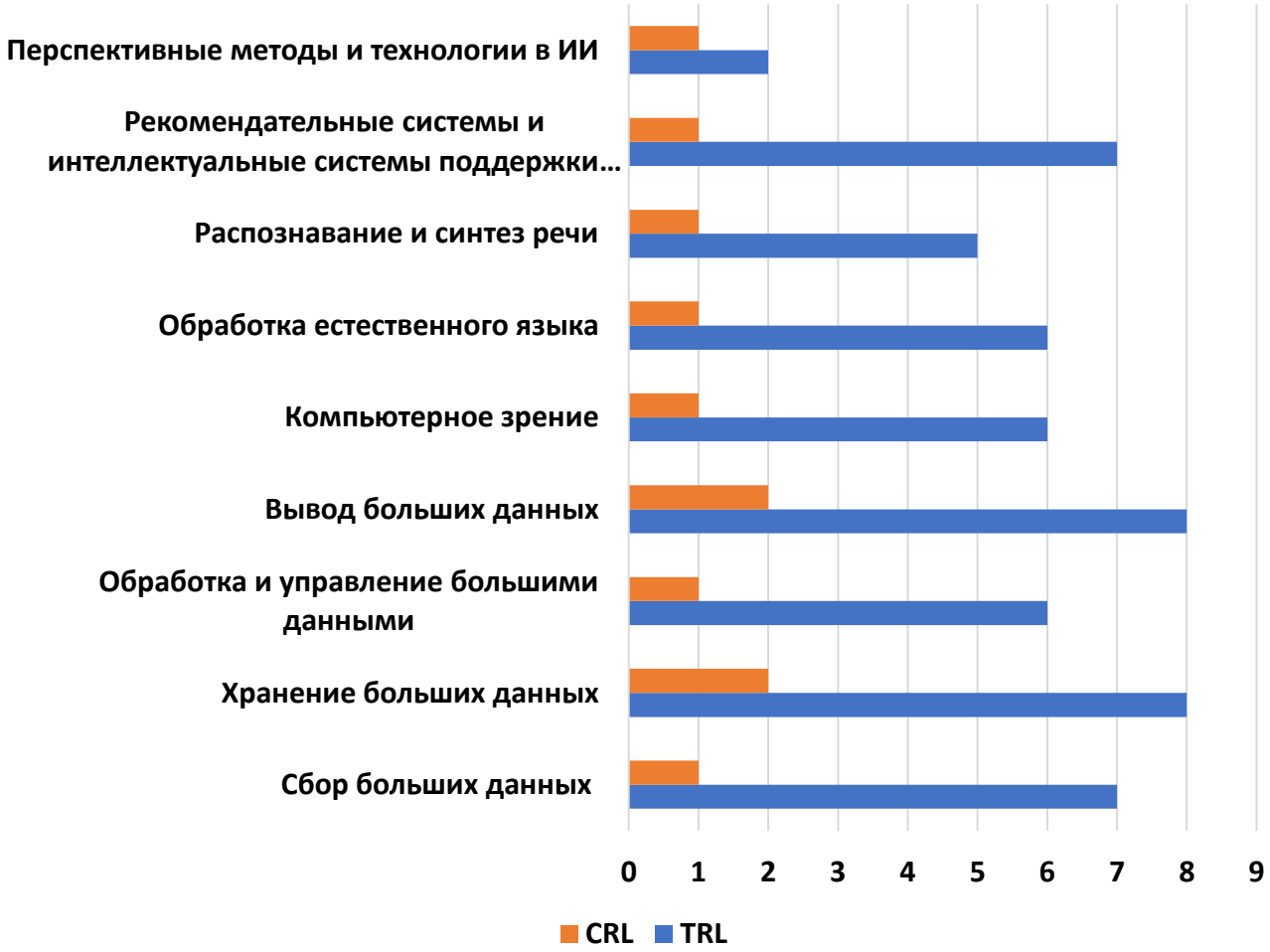
риск экономической нецелесообразности

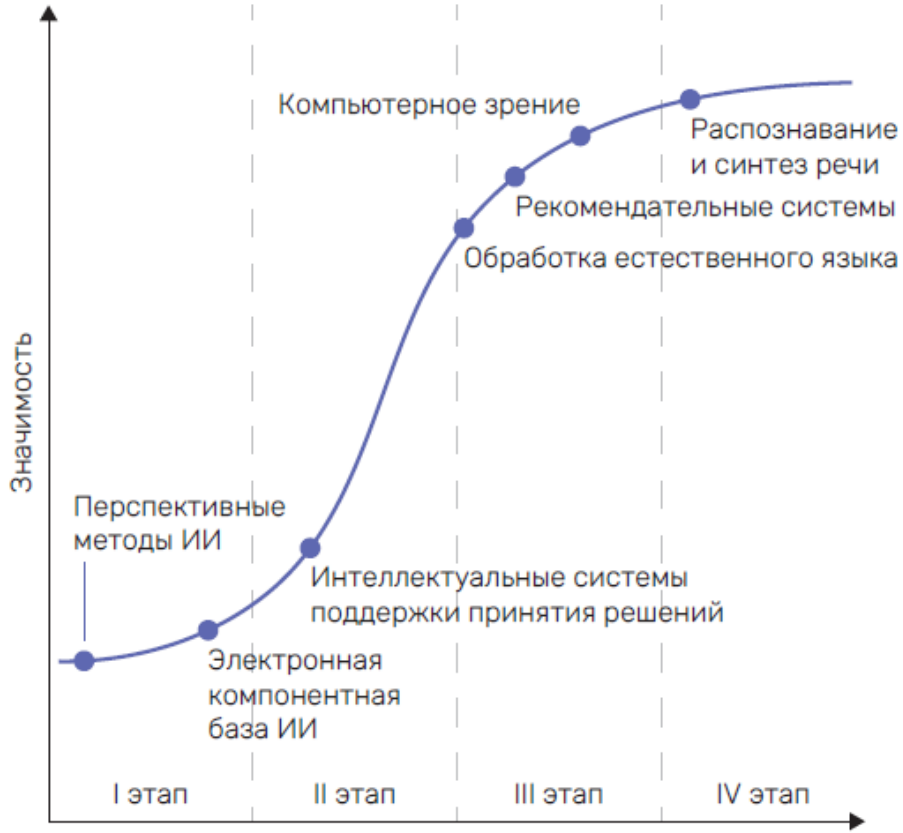
риск ошибок бизнес-модели

риск ошибок больших данных

Сравнительный анализ оценок (Дорожная карта, 2019 г.)

Уровни готовности сквозных цифровых технологий





Национальный проект Экономика данных и цифровая трансформация государства»

- **Основная цель — внедрение принципов управления на основе данных во всех сферах общественной жизни**
- **Планируется достичь качественно нового уровня в логистике, телемедицине, онлайн-образовании и предоставлении государственных услуг**

Некоторые задачи проекта:

- **предоставить стабильный, безопасный доступ в интернет всем жителям государства независимо от их статуса, места проживания, возраста и других критериев**
- **стимулировать развитие отечественного ПО и оборудования, которое нужно для цифровой трансформации**
- **подготовить кадры, участвующие в цифровой трансформации, а также профильных специалистов, способных эффективно работать в новом формате**
- **модернизировать отрасли деятельности, от социальной сферы до бизнеса, используя современные IT-решения**

Национальный проект Экономика данных и цифровая трансформация государства»

Национальный проект состоит из девяти федеральных проектов:

- **«Доступ в интернет»**
- **«Цифровые платформы в отраслях социальной сферы»**
- **«Цифровое государственное управление»**
- **«Отечественные решения»**
- **«Прикладные исследования и перспективные разработки»**
- **«Инфраструктура кибербезопасности»**
- **«Кадры для цифровой трансформации»**
- **«Государственная статистика»**
- **«Искусственный интеллект»**

Ожидаемые результаты проекта к 2030 году

- **100%-ное покрытие интернетом всей территории России и мира**. Это позволит подключить к сети даже самые удалённые регионы страны
- **Создание отраслевых платформ**, таких как «Моя школа», «Университеты», «Наука», «Безопасная среда» и «Умный город». Все школы и колледжи будут оснащены ИТ-инфраструктурой и Wi-Fi, а 634 тысячи учителей получат отечественные планшеты
- **Полная цифровизация госуправления** и переход на 100%-ный безбумажный документооборот. Это позволит упростить процессы и повысить эффективность работы госорганов
- **Производство в России 100% оборудования сотовых сетей и программного обеспечения**, что укрепит технологическую независимость страны
- **Предоставление персонализированных госуслуг** по принципу «жизненных ситуаций». Гражданам и бизнесу больше не придётся заполнять заявления или посещать ведомства — не менее 100 услуг будут оказываться проактивно, на основе анализа данных и предпочтений пользователей
- **Оценка защищённости 100% ключевых государственных информационных систем**
- **Увеличение мощности квантового компьютера** с 50 до 300 кубитов
- **Создание цифровой аналитической платформы** (ГИС «ЦАП») для сбора, обработки и анализа больших объёмов данных в режиме реального времени. Это позволит на 100% автоматизировать предоставление официальной статистики
- **Обучение не менее 250 тысяч студентов** при участии бизнеса, а общее число работников ИТ-отрасли вырастет до 1,4 миллиона человек