

Большие данные

Технологии обработки сверхбольшого
объёма данных («больших данных»)

Инженерия данных

613x-010402D x={1,2,3} осень 2025

Сергей Борисович Попов
sepo@ssau.ru

Материалы лекций:

https://1drv.ms/f/c/5ed33c8b23e26391/EpMzlOhQgt5OqcqHF_ChjwB0NP5AovVgqYTfBokEY1zhA

01.04.02 Прикладная математика и информатика

Большие данные

Лекции: 16 часов (8 лекций)

Среда нечётной недели 15:15 (3-17 недели)

Лабораторные работы: 24 часа (6 занятий по 4 часа)

6131 Четверг чётной недели 15:15 (4-14 недели)

6132 Понедельник чётной недели 15:15 (4-14 недели)

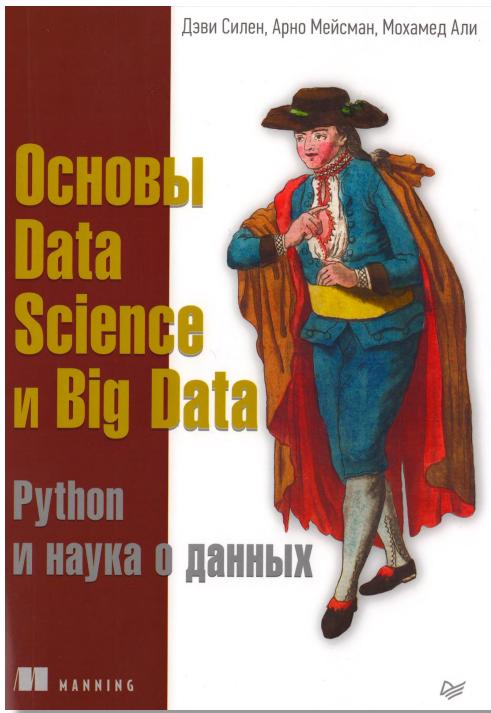
6133(1) Вторник чётной недели 15:15 (4-8, 12-16 недели)

6133(2) Вторник нечётной недели 15:15 (5-15 недели)

Зачёт

КНИГИ

Майер-Шенбергер В., Кукъер К.
Большие данные.
Революция, которая изменит то,
как мы живем, работаем и мыслим.
– Манн, Иванов и Фербер, 2014



Силен Дэви, Мейсман Арно, Али Мохамед
Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336с.: ил. – (Серия «Библиотека программиста»).
ISBN 978-5-496-02517-1

В этой книге много интересных примеров того, как сложнейшие технологии Big Data – методы анализа огромных объемов данных – применяются для решения важных задач из нашей повседневной жизни.

Сергей Мацакий, председатель правления компании iBS

ВИКТОР МАЙЕР-ШЕНБЕРГЕР | КЕННЕТ КУКЬЕР





Дейтэл Пол, Дейтэл Харви

Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»).

ISBN 978-5-4461-1432-0

16

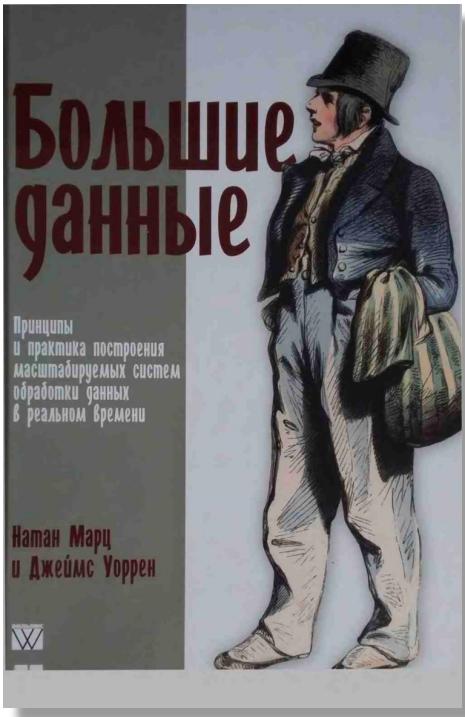
Большие данные: Hadoop, Spark, NoSQL и IoT

В этой главе:

- Концепция больших данных и темпы роста.
- Работа с реляционной базой данных SQLite на языке SQL (Structured Query Language).
- Четыре основные разновидности баз данных NoSQL.
- Сохранение твитов в документной базе данных MongoDB и их визуализация на карте Folium.
- Технология Apache Hadoop и ее применение в приложениях пакетной обработки больших данных.
- Построение MapReduce-приложения на базе Hadoop в облачном сервисе Microsoft Azure HDInsight.
- Применение Apache Spark в высокопроизводительных приложениях, работающих с большими данными в реальном времени.
- Использование потоковой передачи Spark для обработки данных в формате мини-пакетов.
- «Интернет вещей» (IoT) и модель публикации/подписки.
- Публикация сообщений с моделируемого устройства, подключенного к интернету, и их визуализация на информационной панели.
- Подписка на «живой» Twitter PubNub и IoT-потоки, и визуализация данных.

Грас Д.

Data Science. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021. – 416 с.: ил.
ISBN 978-5-9775-6731-2



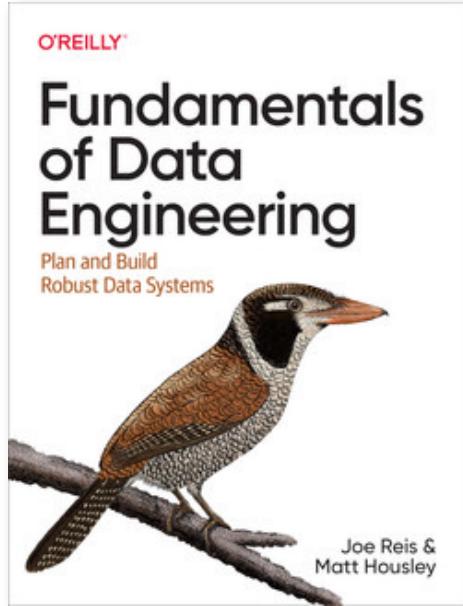
Натан Марц, Джеймс Уоррен

Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени.
— Москва, СПб.: Издательский дом «Вильямс», 2016. — 368 с.
ISBN978-5-8459-2075-1



Железнов, М.М.

Методы и технологии обработки больших данных [Электронный ресурс] : учебно- методическое пособие / М.М. Железнов ; Министерство науки и высшего образования Российской Федерации, Национальный исследовательский Московский государственный строительный университет, кафедра информационных систем, технологий и автоматизации в строительстве. — Электрон. дан. и прогр. (2 Мб). — Москва : Издательство МИСИ – МГСУ, 2020. — Режим доступа: <http://lib.mgsu.ru/> — Загл. с титул экрана.
ISBN 978-5-7264-2193-3



Fundamentals of Data Engineering

by [Joe Reis, Matt Housley](#)

Released June 2022

Publisher(s): O'Reilly Media, Inc.

ISBN: 9781098108304

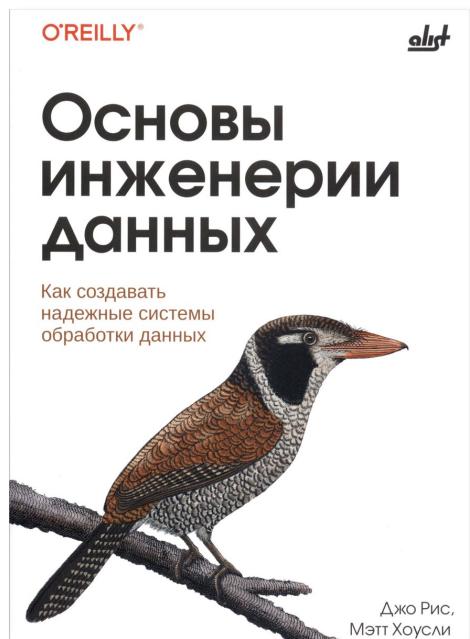


Рис Дж.

Основы инженерии данных: Пер. с англ.

/ Дж. Рис, М. Хоусли – Астана: АЛИСТ, 2024. - 464 с.: ил.

ISBN 978-601-08-4116-1



**Талия Д., Трунфио П., Мароццо Ф., Белькастро Л.,
Кантини Р. и Орсино А.**

П78 Большие данные. Современные фреймворки и разработка приложений /
пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2025. – 272 с.: ил.

ISBN 978-5-93700-358-4

КНИГИ

Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman and
Jeffrey D. Ullman

2019

<http://www.mmds.org>

Mining
of
Massive
Datasets

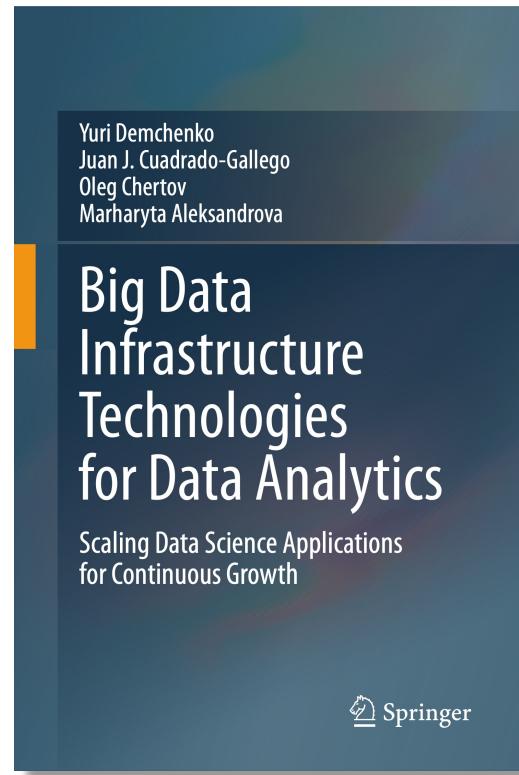
Jure Leskovec
Stanford University

Anand Rajaraman
Rocketship Ventures

Jeffrey D. Ullman
Stanford University

**Data-Intensive Text
Processing with MapReduce**
Jimmy Lin and Chris Dyer
Draft of January 27, 2013





Yuri Demchenko • Juan J. Cuadrado-Gallego
Oleg Chertov • Marharyta Aleksandrova

Big Data Infrastructure Technologies for Data Analytics

Scaling Data Science Applications
for Continuous Growth

ISBN 978-3-031-69365-6
<https://doi.org/10.1007/978-3-031-69366-3>

ISBN 978-3-031-69366-3 (eBook)

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

FIRSTMARK

MAD

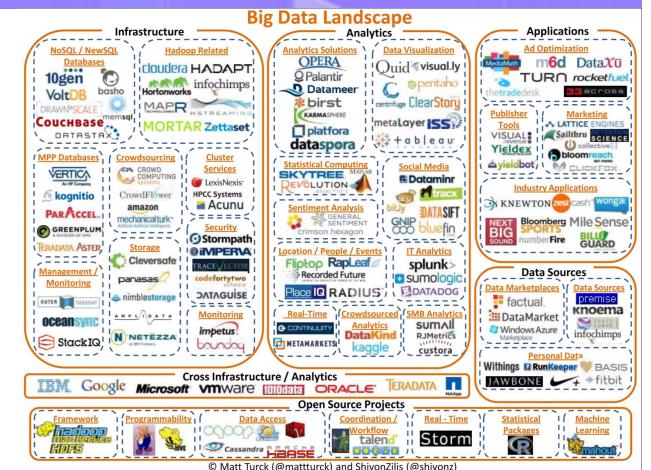
LANDSCAPE

MACHINE LEARNING, AI & DATA

2024

<https://mattturck.com/category/big-data/>

2012, 2014, 2016, 2017, 2018, 2019 (Part I and Part II),
2020, 2021 and 2023 (Part I, Part II, Part III, Part IV).



Как идти в ногу со временем в быстро меняющейся сфере деятельности

Keeping Pace in a Fast-Moving Field

Once a new technology rolls over you, if you're not part of the steamroller, you're part of the road.

—Stewart Brand

How do you keep your skills sharp in a rapidly changing field like data engineering? Should you focus on the latest tools or deep dive into fundamentals? Here's our advice: focus on the fundamentals to understand what's not going to change; pay attention to ongoing developments to know where the field is going. New paradigms and practices are introduced all the time, and it's incumbent on you to stay current. Strive to understand how new technologies will be helpful in the lifecycle.

В момент когда новая технология накатывает на вас как дорожный каток, если вы не часть этого катка, то вы становитесь частью дороги.

- Стюарт Брэнд

Как поддерживать свои навыки на уровне в такой быстро меняющейся области, как инжиниринг данных? Стоит ли сосредоточиться на новейших инструментах или глубоко погрузиться в основы? Совет: **сосредоточьтесь на основах**, чтобы понять, что не изменится; обращайте внимание на текущие разработки, чтобы знать, куда движется область. Новые парадигмы и практики вводятся постоянно, и вам необходимо оставаться в курсе событий. Стремитесь понять, как новые технологии будут полезны в жизненном цикле.

Важные моменты в области больших данных

1997 июнь – первое появление термина "**big data**" в статье исследователей NASA Michael Cox and David Ellsworth (обозначена "**problem of big data**")

2001 февраль – обозначена **проблема 3V (Volume, Velocity, Variety)** в сообщении Doug Laney (META Group)

2003 – **GFS**: The Google file system
19th ACM Symposium on Operating Systems Principles

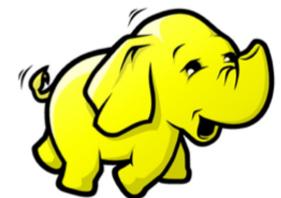
2004 – **MapReduce**: Simplified Data Processing on Large Clusters (OSDI'04)
6th Symposium on Operating System Design and Implementation

2006 – **Bigtable**: A Distributed Storage System for Structured Data (OSDI'06)
7th Symposium on Operating Systems Design and Implementation

2006 – **Hadoop**: HDFS и MapReduce (Doug Cuttig)

2008 – <http://hadoop.apache.org>

2008 сентябрь – спецвыпуск Nature "Big Data: Science in the Petabyte Era"



2010–2025

NoSQL data bases: **Cassandra & Hbase**

SQL-processing: **Hive**

High level Processing: **Pig**

Hadoop compute engine: **Spark**

PostgreSQL на базе МРР: СУБД **Greenplum**

Stream processing: **Apache Airflow, Kafka, Storm, Flink**

Analytic system: **Google BigQuery, Snowflake, Amazon Redshift**

Стандартизация в области Больших данных

- NIST Big Data Interoperability Framework

Цель: эталонная архитектура
больших данных NIST (NBDRA)

<https://www.nist.gov/itl/big-data-nist>



- ISO/IEC JTC 1/SC 2 - Big Data
- ISO/IEC JTC 1/SC 42 - Artificial intelligence



- Технический комитет по стандартизации
ТК 164 «Искусственный интеллект»



NIST Big Data Interoperability Framework

Цель: эталонная архитектура больших данных NIST
(NBDRA – NIST Big Data Reference Architecture)

Разработка проходила в три этапа:

Этап 3: проверка NBDRA путем создания общих приложений для больших данных через общие интерфейсы;

Этап 2: определение общих интерфейсов между компонентами NBDRA;

Этап 1: определение ключевых компонентов эталонной архитектуры больших данных высокого уровня, которые не зависят от технологий, инфраструктуры и поставщиков.

NIST Big Data Interoperability Framework

Том 1, Определения

Том 2, Таксономии

Том 3, Примеры использования
и общие требования

Том 4, Безопасность и
конфиденциальность

Том 5, Обзор официального
документа по
архитектурам

Том 6, Эталонная архитектура

Том 7, Дорожная карта
стандартов

Том 8, Интерфейсы эталонной
архитектуры

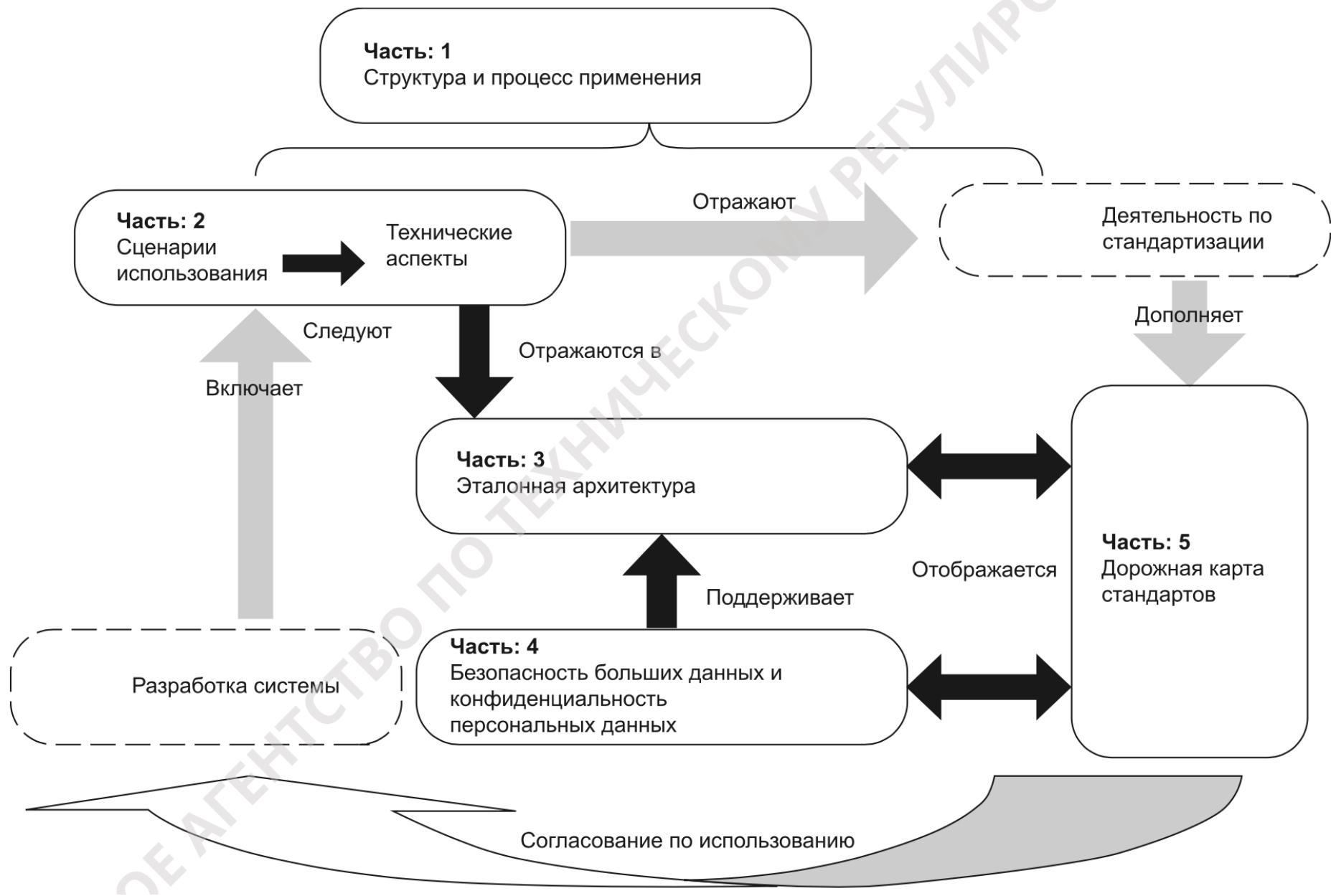
Том 9, Внедрение и
модернизация



ISO-стандарты в области Больших данных

- ISO/IEC JTC 1, Information technology
- ISO/IEC JTC 1/SC 42 - Artificial intelligence (JTC 1/SC 2 - Big Data)
- ISO/IEC 20546:2019 Information technology – Big data – Overview and vocabulary
- ISO/IEC TR 20547-1:2020 Information technology – Big data reference architecture – Part 1: Framework and application process
- ISO/IEC TR 20547-2:2018 Information technology – Big data reference architecture – Part 2: Use cases and derived requirements
- ISO/IEC 20547-3:2020 Information technology – Big data reference architecture – Part 3: Reference architecture
- ISO/IEC 20547-4:2020 Information technology – Big data reference architecture – Part 4: Security and privacy
- ISO/IEC TR 20547-5:2018 Information technology – Big data reference architecture — Part 5: Standards roadmap
- ISO/IEC 24668:2022 Information technology – Artificial intelligence – Process management framework for big data analytics
- ISO/IEC CD 27045 Information technology – Big data security and privacy – Guidelines for managing big data risks

Взаимосвязь частей стандарта ISO/IEC 20547



Российские стандарты Больших данных

- Технический комитет по стандартизации ТК 164 «Искусственный интеллект»
- ПОДГОТОВЛЕНЫ МГУ имени М.В. Ломоносова в лице НОЦ компетенций в области цифровой экономики МГУ и Автономной некоммерческой организацией «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии стандарта
- ГОСТ Р ИСО/МЭК 20546:2019 Информационные технологии – Большие данные – Обзор и словарь
- ГОСТ Р 70466–2022 (ISO/IEC TR 20547-1:2020) Информационные технологии. Эталонная архитектура больших данных. Часть 1. Структура и процесс применения
- ГОСТ Р 59926–2021 (ISO/IEC TR 20547-2:2018) Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования
- ГОСТ Р ИСО/МЭК 20547-3–2024 Информационные технологии. Эталонная архитектура больших данных. Часть 3. Эталонная архитектура

Российские стандарты Больших данных

- ПНСТ 919–2024 (ISO/IEC TR 20547-5:2018) Информационные технологии. Эталонная архитектура больших данных. Часть 5. Направления стандартизации
- ГОСТ Р 59925–2021 Информационные технологии. Большие данные. Техническое задание. Требования к содержанию и оформлению
- ГОСТ Р ИСО/МЭК 24668–2022 Информационные технологии. Искусственный интеллект. Структура управления процессами аналитики больших данных
- ПНСТ 847–2023 Искусственный интеллект. Большие данные. Функциональные требования в отношении происхождения данных
- ПНСТ 848–2023 Искусственный интеллект. Большие данные. Обзор и требования по обеспечению сохранности данных
- ГОСТ Р 71540–2024 (ИСО/МЭК 5392:2024) Искусственный интеллект. Эталонная архитектура инженерии знаний

NIST-определения Больших данных

2 TERMS AND DEFINITIONS

Big Data consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis.

Big Data engineering is the discipline for engineering scalable systems for data-intensive processing.

The ***Big Data Paradigm*** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

3 BIG DATA CHARACTERISTICS

3.1 BIG DATA DEFINITIONS

Big Data refers to the need to parallelize the data handling in data-intensive applications. The characteristics of Big Data that force new architectures are as follows:

- ***Volume*** (i.e., the size of the dataset);
- ***Velocity*** (i.e., rate of flow);
- ***Variety*** (i.e., data from multiple repositories, domains, or types); and
- ***Variability*** (i.e., the change in velocity or structure).

NIST-определения Больших данных

Большие данные состоят из больших наборов данных – в первую очередь по характеристикам объема, разнообразия, скорости и/или изменчивости – которые требуют **масштабируемой** архитектуры для эффективного хранения, обработки и анализа.

Инжиниринг больших данных – это дисциплина разработки **масштабируемых** систем для обработки больших объемов данных.

Парадигма больших данных заключается в распределении систем данных по горизонтально связанным независимым ресурсам для достижения **масштабируемости**, необходимой для эффективной обработки обширных наборов данных.

Под **большими данными** понимается необходимость распараллелить обработку данных в приложениях с интенсивным использованием данных. Следующие характеристики больших данных вынуждают использовать новые архитектуры:

- Объем (*Volume*), т.е. размер набора данных;
- Скорость (*Velocity*), то есть скорость потока;
- Разнообразие (*Variety*), т.е. данные из нескольких репозиториев, доменов или типов; и
- Изменчивость (*Variability*), т.е. изменение скорости или структуры.

Определения Больших данных (ГОСТ и ИСО)

- **большие данные** (big data): Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии **масштабирования** для эффективного хранения, обработки, управления и анализа
- **вариативность данных** (data variability): Изменения в скорости передачи, формате или структуре, семантике или качестве массива данных
- **разнообразие данных** (data variety): Диапазон форматов, логических моделей, временных шкал и семантики массива данных
- **скорость обработки данных** (data velocity): Скорость потока, с которой данные создаются, передаются, сохраняются, анализируются или визуализируются
- **объем данных** (data volume): Количественная характеристика данных, влияющая на выбор ресурсов для вычислений и хранения, а также на управление данными в процессе обработки

Дополнительные определения (ГОСТ и ИСО)

- **изменчивость данных** (data volatility): Характеристика данных, относящаяся к скорости их изменения с течением времени
- **распределенная обработка данных** (distributed data processing): Обработка данных, при которой выполнение операций распределено по узлам вычислительной сети
- **горизонтальное масштабирование** (horizontal scaling): Формирование единого логического целого путем соединения нескольких аппаратно-программных комплексов
- **метаданные** (metadata): Данные о данных или элементах данных, которые могут включать описание, а также сведения о владельце данных, путях доступа к ним, правах доступа и изменчивости данных
- **параллельность** (parallel): Относится к процессу, в котором все события происходят в одном и том же интервале времени, и при этом каждое из них обрабатывается отдельной, но схожей функциональной единицей

Ключевые характеристики больших данных

- **Объем данных.** Объем данных представляет собой определенное количество данных, доступных для анализа с целью извлечения полезной информации. Одним из основных факторов развития технологий обработки больших данных стал рост объемов данных, как следствие интернет-активности
- **Скорость обработки данных.** Скорость обработки данных – это скорость потока создания, хранения, анализа и визуализации данных. Скорость обработки больших данных означает необходимость обработки большого количества данных за короткий промежуток времени. В качестве примеров работы с данными с высокой скоростью обработки обычно приводят средства обработки потоковых данных

Ключевые характеристики больших данных

- **Разнообразие данных.** Свойство разнообразия данных отражает необходимость анализа данных разных типов из различных предметных областей. Как правило, проблема разнообразия данных решалась посредством их преобразования или проведения предварительного анализа с целью выявления свойств, позволяющих объединить их с другими данными. Более широкий диапазон форматов данных, логических моделей, временных шкал и семантики, которые предпочтительно использовать при аналитике, усложняет интеграцию разнообразных данных. В качестве средства, способствующего интеграции, все чаще используются метаданные. Одним из результатов влияния разнообразия на большие данные является необходимость представления семантики данных в машиночитаемом виде

Ключевые характеристики больших данных

- **Вариативность данных.** Вариативность данных означает изменения в скорости передачи данных, их формате/структуре, семантике и/или качестве, которые влияют на поддерживаемое приложение, аналитику или проблему. Влияние вариативности может заключаться в необходимости проведения реорганизации архитектур, интерфейсов, методов обработки/алгоритмов, интеграции/слияния, хранения, применимости или использования данных. В дополнение следует отметить, что вариативность объемов данных подразумевает необходимость увеличения или уменьшения виртуализированных ресурсов для эффективного управления дополнительной нагрузкой на обработку.

Ключевые характеристики обработки данных

- **Наука о данных.** Наука о данных изучает процесс извлечения из них знаний. Используемый научный подход может заключаться либо в проведении исследований, либо в проверке гипотез. Наука о данных изучает полный жизненный цикл аналитики данных, в котором аналитика данных понимается согласно следующему определению.
- **аналитика данных** (data analytics): Составное понятие, охватывающее получение, сбор, проверку и обработку данных, включая их количественную оценку, визуализацию и интерпретацию.

Примечание - Аналитика данных используется для представления объектов, описываемых данными, с целью прогнозирования конкретных ситуаций и формирования пошаговых рекомендаций при решении задач. Закономерности, полученные посредством аналитики, используются в различных целях, таких как принятие решений, проведение исследований, обеспечение устойчивого развития, проектирование, планирование и т. д.

Ключевые характеристики обработки данных

- **Изменчивость данных.** Изменчивость данных связана с ограниченным промежутком времени, в течение которого значения данных остаются актуальными для конкретного анализа, и определяется динамикой изменений.

В тех ситуациях, когда аналитика данных проводится в режиме реального времени, немедленная обработка данных является критически необходимой для принятия решений. Наиболее очевидным образом это проявляется при работе с данными с высокой скоростью генерации, например с данными, связанными с фондовыми рынками или телекоммуникациями. Однако данные, не пригодные для специфического, чувствительного к временным рамкам анализа ввиду устаревания, могут оставаться актуальными для других типов аналитики, не зависимых от времени

Ключевые характеристики обработки данных

- **Достоверность данных.** Достоверность данных определяется их полнотой и точностью, в связи с чем для обозначения качества данных в профессиональном жаргоне длительное время существует выражение «мусор на входе - мусор на выходе». Если аналитика данных направлена на установление причинно-следственных связей, то качество каждого элемента является крайне важным.
Если аналитика осуществляется путем выявления корреляций или трендов в больших массивах данных, то отдельные некорректные элементы могут быть потеряны при общих подсчетах, но тренд может оставаться точным.
- **Выгода** (benefit, value). Выгода определяется степенью достижения системой обработки больших данных целей, для которых эта система создавалась.

Ключевые характеристики обработки данных

- **Визуализация данных.** Под визуализацией данных подразумевается такое их представление, которое позволяет пользователю извлечь из них информацию. Большие данные потребовали новых методов обработки массивов данных больших объемов, включая сбор и обобщение данных для их наибольшей наглядности. Большие данные также требуют более пристального внимания к визуальному представлению для лиц, ответственных за принятие решений. Это необходимо для изложения результатов в доступном для понимания виде, а также для информирования об их сложности, точности и вероятностном интервале ошибок.

Ключевые характеристики обработки данных

- **Структурированные и неструктурированные данные.**
Постоянно увеличиваются как объёмы, так и значение неструктурированных данных. Хотя реляционные базы данных обычно поддерживают эти типы элементов данных, их способность непосредственно анализировать, индексировать и обрабатывать такие типы данных, как правило, ограничена и доступна через нестандартные расширения SQL. Потребность в анализе неструктурированных данных существует уже много лет. Однако переход на парадигму больших данных привел к повышению значимости неструктурированных данных.

Также в отношении неструктурированных данных особое внимание уделяется различным новым методам разработки, которые позволяют проводить анализ таких данных более эффективно.

Ключевые характеристики обработки данных

- **Масштабирование.** Большие данные подразумевают возможность расширения репозиториев данных и их обработку на параллельно работающих ресурсах - аналогичным образом сообщество специалистов, использующих моделирование, требующее ресурсоемких вычислений, массово перешло на параллельную обработку. Благодаря разработке методов взаимодействия между ресурсами, такое же масштабирование теперь доступно для приложений, использующих большое количество данных. Вертикальное масштабирование ограничено физическими возможностями и приводит к росту материальных и временных затрат на реализацию. Альтернативный метод – горизонтальное масштабирование, использующее отдельные распределенные ресурсы, объединяемые для работы в качестве единой системы. Именно горизонтальное масштабирование лежит в основе революции больших данных. Хотя методы достижения эффективной масштабируемости между ресурсами постоянно развиваются, эта смена парадигмы представляет собой единовременное явление.

Ключевые характеристики обработки данных

- **Распределенная обработка данных.** Популярная структура для распределенных вычислений состоит из комбинации уровня хранения и уровня обработки, которая реализует мультиклассовую модель алгоритмического программирования. Недорогие серверы потребительского уровня, поддерживающие распределенную файловую систему хранения данных, могут значительно снизить затраты на хранение вычислений для большого объема данных (например, индексация в сети). При распределенной обработке данных запрос распределен по процессорам, а результаты собираются в центральный процессор. Затем результаты обработки обычно загружаются в аналитическую среду. Для достижения эффективности, надежности, высокой доступности и отказоустойчивости системы несколько узлов (например, клиентские узлы, узлы данных, узлы-реплики) размещаются в виде архитектуры «ведущий-ведомый».

Ключевые характеристики обработки данных

- **Нереляционные базы данных.** В горизонтально масштабируемых системах данные распределяются по узлам кластера, имея при этом единую логическую структуру. Новые парадигмы базы данных нереляционной модели обычно называют NoSQL («не только SQL» или «не SQL»). Проблема с определением парадигмы хранения больших данных как NoSQL заключается, во-первых, в описании хранения данных на теоретико-множественном языке для запросов и извлечения данных и, во-вторых, в расширении возможностей применения языков запросов, похожих на SQL, к новым нереляционным хранилищам данных. В то время как NoSQL используется настолько широко, что будет применяться в новых моделях данных вне рамок реляционной модели, сам термин относится к базам данных, не следующим реляционной модели. Примерами моделей нереляционных баз данных являются столбец, разреженная таблица, ключ-значение, документ-ключ и графические модели.

Масштабируемость

Масштабируемость – это способность системы или технологии (процесса) адаптироваться к увеличению показателей задач и повышению требований (например, увеличение объёмов данных, числа пользователей и т.д.) без потери производительности или функциональности.

Как и где?

Параллелизм vs. распределённость

Масштабируемость:

Как (в части аппаратной), где?

Распределённость

Виды масштабируемости (аппаратной, hardware):

Горизонтальная масштабируемость (Horizontal Scaling) – добавление новых узлов в систему, например, дополнительных серверов или виртуальных машин.

Пример: увеличение количества серверов в веб-кластере.

Вертикальная масштабируемость (Vertical Scaling) – увеличение производительности существующего оборудования, например, добавление оперативной памяти, процессорных мощностей или подключения хранилища на сервер.

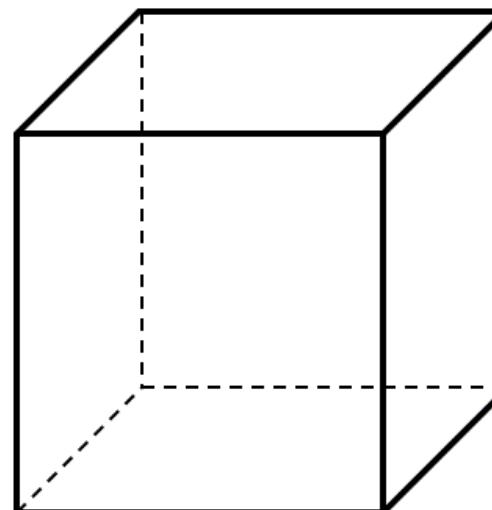
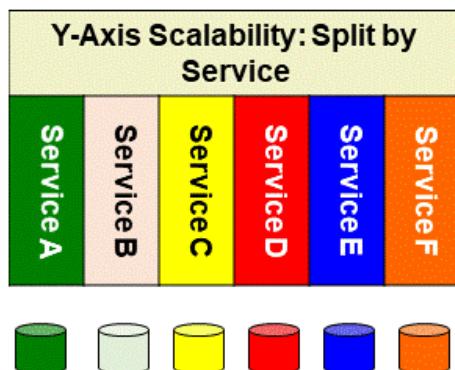
Пример: переход с 8-ядерного процессора на 16-ядерный.

Масштабируемость: Как (в части программной)?

Параллелизм

Направления масштабирования

AKF Scale Cube



Z-Axis Scalability: Segment by Customer

NA	EU		
POD 1	POD 2	POD 3	POD 4

X-Axis Scalability: Replicate & LB

Web Tier	Replicate Web Servers & Load Balance
App Tier	Store Session in browser or separated Object Cache to horizontally scale app tier independent of web tier
DB Tier	Use Read-Replicas for read-only use cases like reporting, search, etc.

Другие определения Big Data

Как характеристика
данных (3V [Gartner,
2001]):

- **Volume** (объём)
- **Velocity** (скорость поступления/наращивания)
- **Variety** (разнообразие)

позднейшие дополнения
(+ 2V):

- **Veracity** (достоверность)
- **Value** (ценность или смысл)

Как характеристика технологии:

- Технологии Big Data – это технологии обработки информации, которые применяются тогда, **когда традиционные технологии** обработки на базе реляционных баз данных **не применимы** для решения стоящих задач

- Большие данные объединяют техники и технологии, которые **извлекают смысл** из данных на экстремальном пределе практичности. [Forrester]
- Технологии Big Data – это технологии преобразования информации в знание

Big Data – технологии получения знаний

Знание = информация +

- связи, зависимости, контекст
 - история изменения (во времени)
 - модель (получения и использования)
-
- Главное отличие знаний от данных состоит в их структурности и активности, появление в базе новых фактов или установление новых связей может стать источником формирования новых знаний, а, следовательно, изменений в принятии решений.
 - Для принятия решения необходимы знания, а не информация.

Value: Big Data Are the Next Great Natural Resource

- Компании, которые принимают решения на "миллиарды долларов" руководствуясь "инстинктами" (gut instincts, «Нутром чувствую!»), а не "интеллектуальным анализом" больших данных, станут «проигравшими» (losers) в постоянно растущей глобальной экономике, основанной на информации.

(генеральный директор IBM
Вирджиния Рометти ("Ginni" Rometty)

<http://www.cfr.org/world/conversation-ginni-rometty/p35497>

Изменить принципы организации компаний (правительства, города)

1. Как вы принимаете решения.
2. Как вы на самом деле создаёте прибавочную стоимость.
3. Как вы поставляете результаты прибавочной стоимости.
 - The first one – it will change how you make decisions.
 - It will change how you in fact create value.
 - And the third is, it will change how you deliver value

Традиционные технологии vs. Big Data

Традиционные технологии:

- **Не создают знания**
- Накопление и интерпретация информации в рамках **существующего знания**
- **Заранее** определяем структуры хранения
- **Заранее** знаем способы использования
- Технология обработки определена **заранее**
- Изменения сопряжены с доработкой (или даже переработкой) ПО, переформатированием данных



Технологии Big Data:

- Хранение и обработка/анализ неструктурированных (сырых, «как есть») данных
- Простой инструмент для создания широкого спектра процедур обработки/анализа
- Итерационный исследовательский подход к анализу хранимых данных:
 - отбор, структурирование и агрегация данных
 - формирование модели (или набора)
 - проверка модели (моделей) на существующих данных
 - сравнительная оценка моделей на вновь поступающих данных
 - обновление моделей



Определения науки о данных

Data Science – деятельность, связанная с анализом данных и поиском лучших решений на их основе.

Дедуктивное обучение. Имеется набор данных (датасет), представляющий собой выборку из генеральной совокупности и описывающий ситуацию в виде прецедентов (пар «объект–ответ»). Известен набор моделей, которые могут в явной форме описать связь между объектами и ответами.

Машинное обучение (обучение по прецедентам). Имеется набор данных (датасет), представляющий собой выборку из генеральной совокупности и описывающий ситуацию в виде прецедентов (пар «объект–ответ»). Зависимость между ответами и объектами неизвестна и, как правило, не выражается аналитически. В этом случае выбирается модельная среда (нейросеть), которая обучается по прецедентам. Для проверки правильности обучения выделяется дополнительный датасет, на котором нейросеть не обучалась. Обученная таким образом нейросеть с весами признается моделью и используется для дальнейших прогнозов развития ситуации по всей генеральной совокупности.

Базовая «тройка» задач DS в классе обучения по прецедентам:
кластеризация, классификация и регрессия

Иерархия потребностей в науке о данных



- Искусственный интеллект, глубокое обучение
- Эксперименты А/В-тестирования, простые алгоритмы машинного обучения
- Аналитика, метрики, сегменты, агрегаты, характеристики, обучающие данные
- Очистка, обнаружение аномалий, подготовка
- Надёжная организация процесса обработки данных, инфраструктура, конвейеры, ETL, хранение структурированных и неструктурных данных
- Инструментарий, журналирование, датчики, внешние данные, содержимое от пользователя