Кафедра Прикладной математики
Института информационных технологий
РТУ МИРЭА

# Дисциплина «Большие данные»

2023-2024 у.г.

# Наполнение курса

- **Объем курса**
  - 8 лекционных и 8 практических занятий
- **Темы лекционных занятий**

1. Технологии хранения информации и больших объемов данных
2. Технологии сбора информации и больших объемов данных
3. Технологии структурирования данных и табличные данные
4. Технологии обработки данных: преобразование и агрегация
5. Технологии обработки данных: обогащение
6. Технологии аналитики больших данных
7. Технологии визуализации больших данных
8. Технологии обработки больших объемов данных

- **Темы практических занятий**

1. Популярные ОС для Больших данных (Unix/Linux серверные системы)
2. Инструментарий хранения данных (SQL базы данных)
3. Инструментарий анализа данных (Loginom)
4. Инструментарий визуализации данных

# Тематика курса

Курс предназначен для ознакомления с возможностями работы с данными в современных компьютерных системах и получения навыков в рамках обработки и анализа данных

**В результате курса реализуются следующие компетенции:**

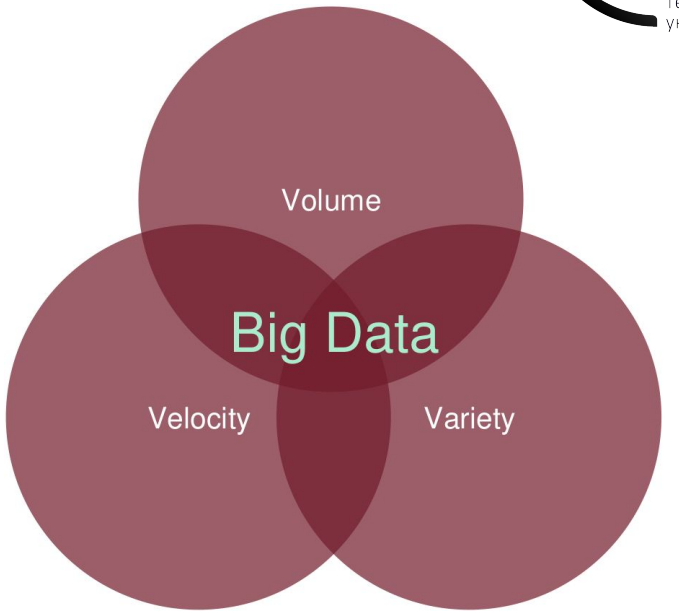
1. Получение первоначальных навыков в инженерии и аналитике данных
2. Знание команд DML языка SQL для извлечения и изменения данных в структурированных СУБД
3. Практическая работа с аналитической Low-code платформой Loginom для построения конвейера обработки больших данных
4. Построение визуализации построенной аналитики больших данных
5. Знание архитектур построения хранилищ данных и обеспечения обработки больших данных

# Лекция 1. Технологии хранения информации и больших объемов данных

# Часть 1. Введение в Большие данные

# Что такое Большие данные?

- **Большие данные** — это **разнообразные данные, которые поступают с постоянно растущей скоростью и объем которых постоянно растет**.
- Три основных свойства больших данных — **разнообразие**, **высокая скорость поступления** и **большой объем**



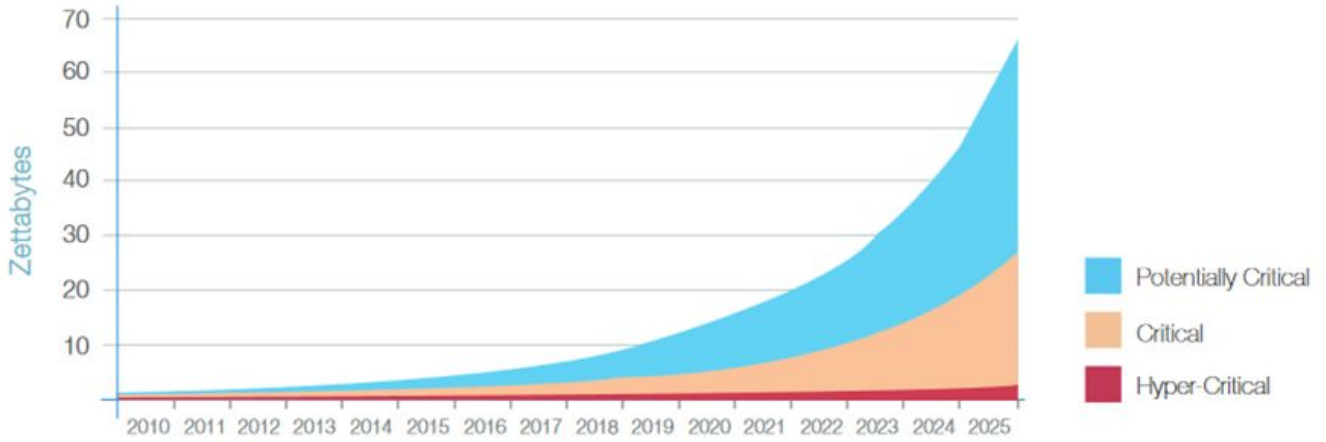

- **Примеры**:

1. Умные устройства
2. Бизнес
3. Здравоохранение
4. Т. д.

# Насколько это необходимо?

**Обзор** (2020) **компании** Data Age Reportпо технологической **цифровизации до 2025 года**

Необходимость RTM обработки, низкая задержка, нерегламентированный характер использования и тяжесть последствий, если данные станут недоступны.

# Задачи обработки больших данных

- Главной задачей обработки больших данных на сегодняшний день является максимимзация пользы от накопленных данных о потреблении ресурсов или услуг.

- Накопленные исторические данные и оперативные данные о потреблении услуг обладают информацией о трендах, тендециях и измечивости вектора предпочтений пользователей

- Обработка больших данных позволяет получить пользу  из исторических данных в сферах бизнеса, здравоохранения, сельского хозяйства, и т.д.

# Задачи в области Больших данных

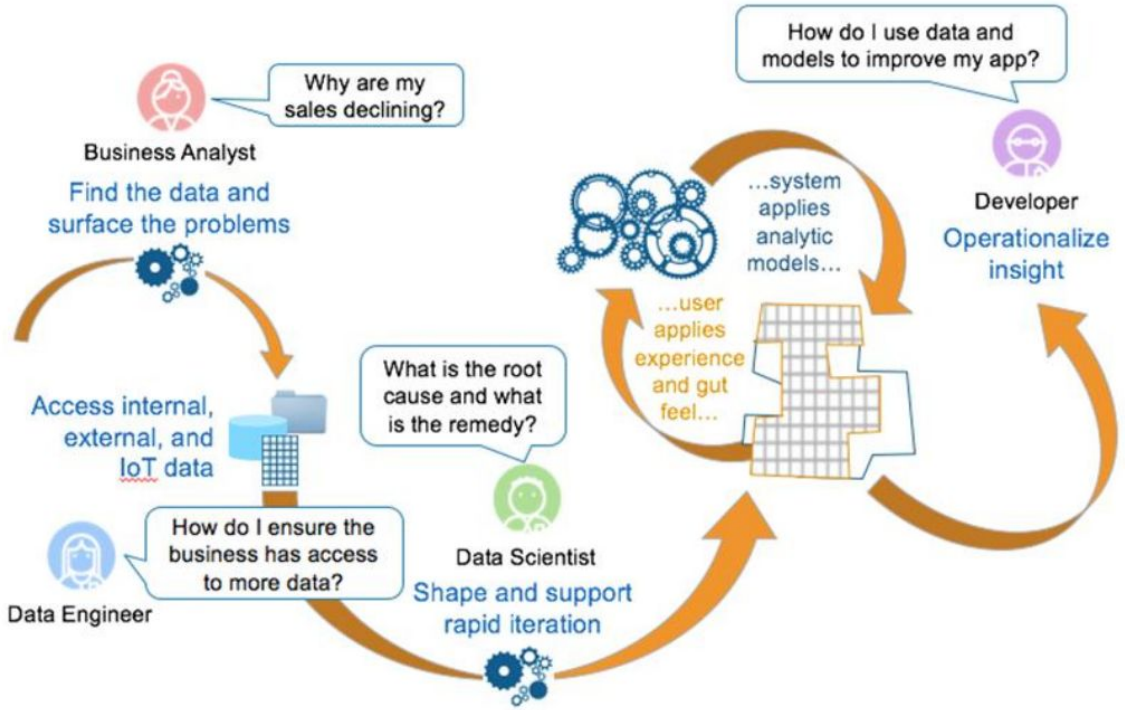
# Задачи в области Больших данных

# Специалисты по работе с данными

- Классификация специалистов, задействованых в работе с данными, на сегодняшний день всё ещё размыта.

**Выделяют следующих специалистов:**

1. Инженер данных
2. Аналитик данных
3. Разработчик
4. DEVops, MLops, ...
5. Управляющий продуктом



# Инженер данных

- **Направления работы инженера данных**

1. Предоставление данных для аналитики (Хранилища данных, Аналитика, Визуализация)
2. Предоставление данных для Machine Learning и Data Science
3. Внедрение моделей машинного обучения в продукт

- **Задачи инженера данных**

1. Сбор данных из различных источников
2. Перемещение данных: потоки данных, ETL
3. Очистка, подготовка, трансформация и обработка данных по бизнес-правилам
4. Анализ, агрегация, разметка данных
5. Изучение данных, оптимизация хранения и обработки данных
6. Построение платформ данных

| Прикладные интрументы | Доля, % |
|---|---|
| SQL | 65 |
| Python | 60 |
| Data Pipelines | 55 |
| Data Warehouse | 50 |
| Hadoop | 45 |
| Hive | 45 |
| ETL | 40 |
| Spark | 40 |
| AWS | 30 |
| Redshift | 30 |
| Java | 25 |
| Kafka | 25 |
| MapReduce | 25 |
| Scala | 25 |
| Vertica | 25 |
| NoSQL | 20 |
| Statistics | 20 |

# Аналитик данных

- **Направления работы аналитика данных**

1. Формулировка бизнес-метрик для построения продуктовых решений на основе данных
2. Построение моделей машинного обучения
3. Построение отчетов для построенных рекомендаций на основе данных

- **Задачи аналитика данных**

1. На основе бизнес-требований строить метрики качества принятия решений
2. Построение аналитических отчетов на основе данных с использованием агрегации разной глубины
3. Построение моделей предиктивной аналитики на основе бизнес-данных
4. Формулировка рекомендаций по данным

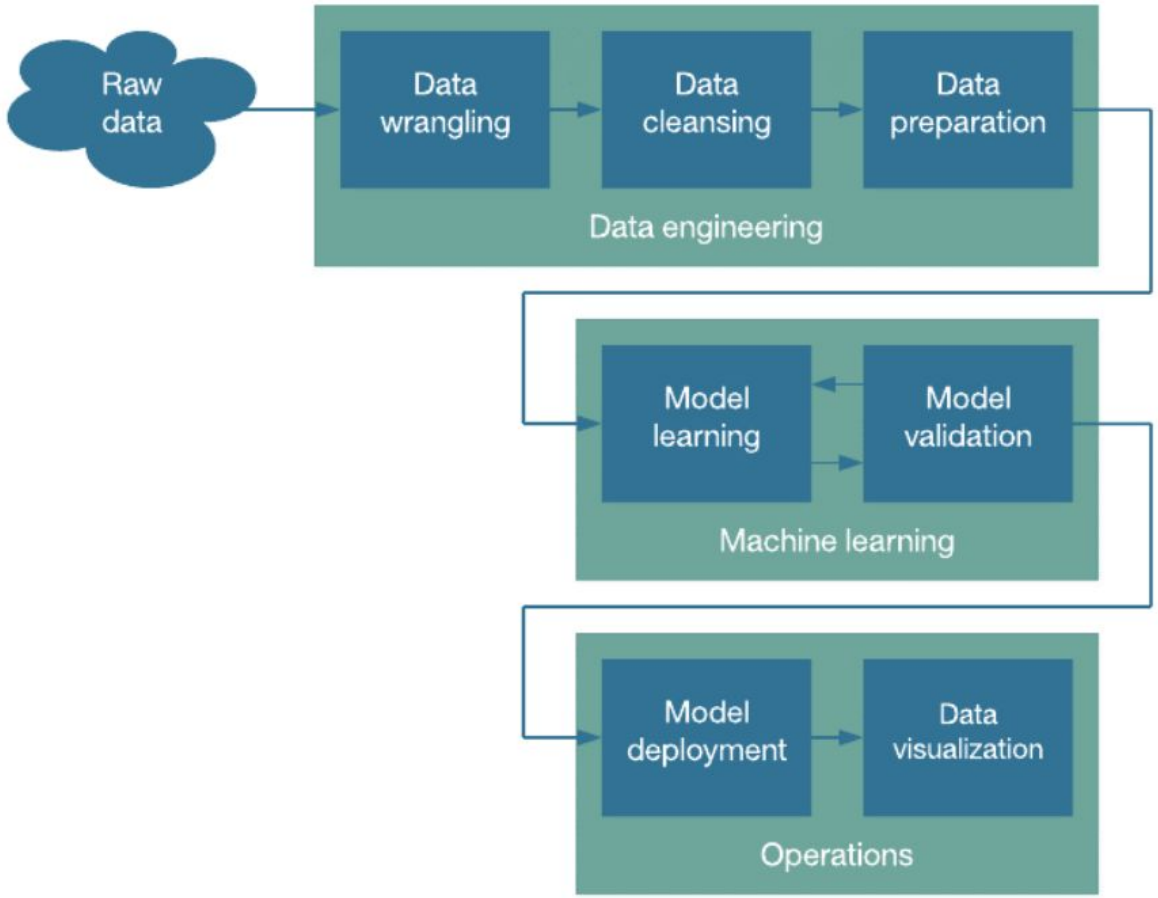
# Конвейер обработки данных

Перед извлечением из данных пользы их необходимо **собрать, очистить, сохранить в нужном виде** и затем работать с ними

В современной индустрии устоялся классический конвейер работы с данными, как основной шаблон по которому компании подстраивают поток работ под себя

Под каждую новую задачу поток работ над данными видоизменяется при неизменной основе

Видоизменения набора задач по обработке данных зависит от **количества источников данных, сложности данных и целей обработки данных**



# Инфраструктура обработки данных

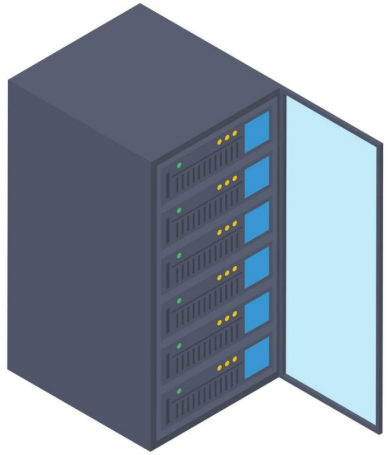
Большие компании хранят, обрабатывают и анализируют данные на серверных вычислительных устройствах или ЦОД (центры обработки данных) разной степени доступности:

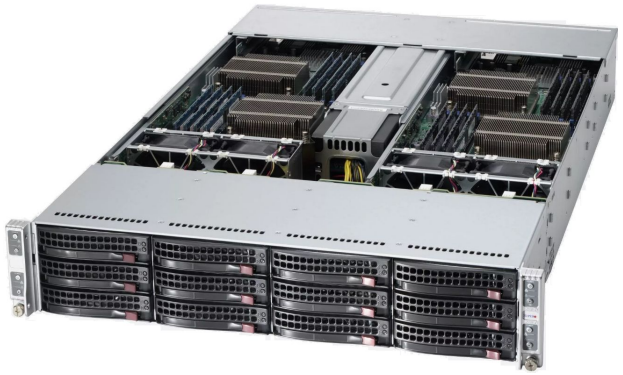
- **Вычислительная инфраструктура:**

1. Локальный вычислительный кластер
2. Частные облачные сервисы
3. Общедоступное облако

- **Популярные серверные ОС:**

  - Linux-серверные системы
  - Debian/CentOS





# Инструменты больших данных

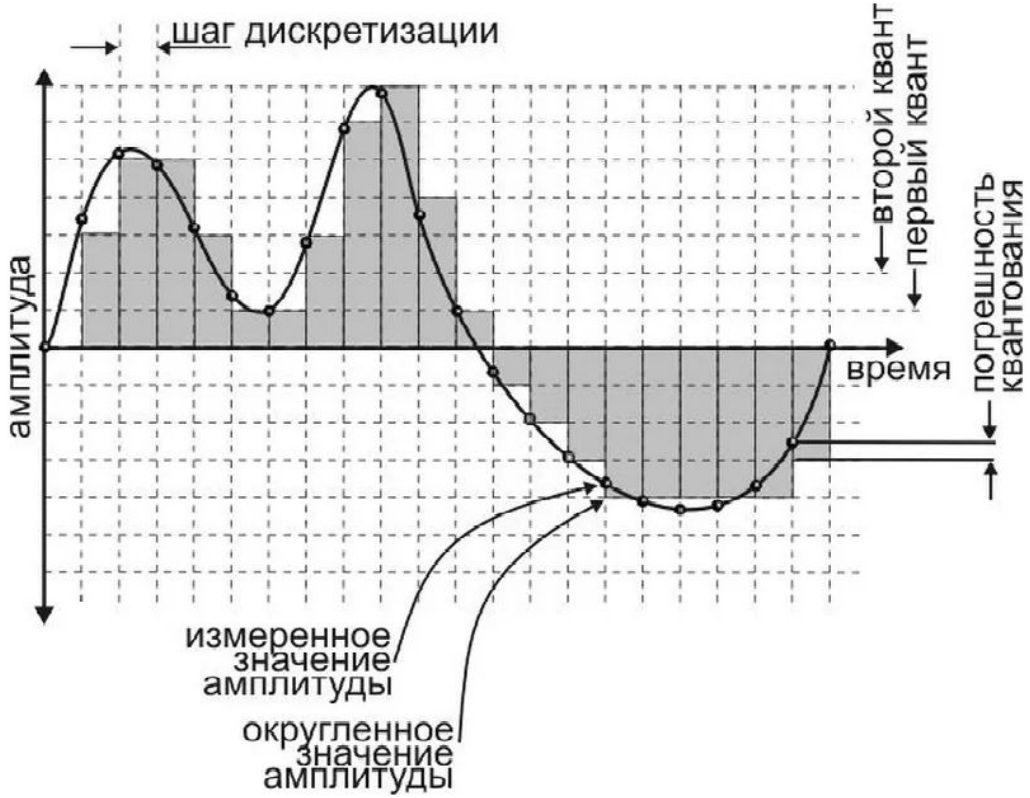
МИРЭА
Российский технологический университет

| Хранение данных | Управление потоками данных | Обработка и анализ данных |
| --- | --- | --- |
| PostgreSQL | Apache Kafka | Python, R |
| Amazon Redshift | Apache NiFi | Loginom |
| Greenplum | Apache Airflow | Hadoop MapReduce |
| Hadoop | | Apache Spark |
| MongoDB | | |

# Часть 2. Информация и данные в вычислительных устройствах

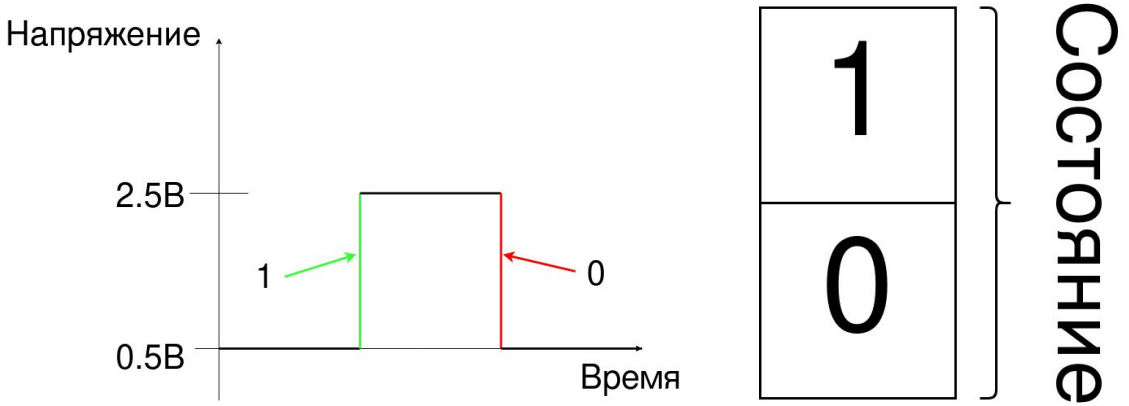
# Информация

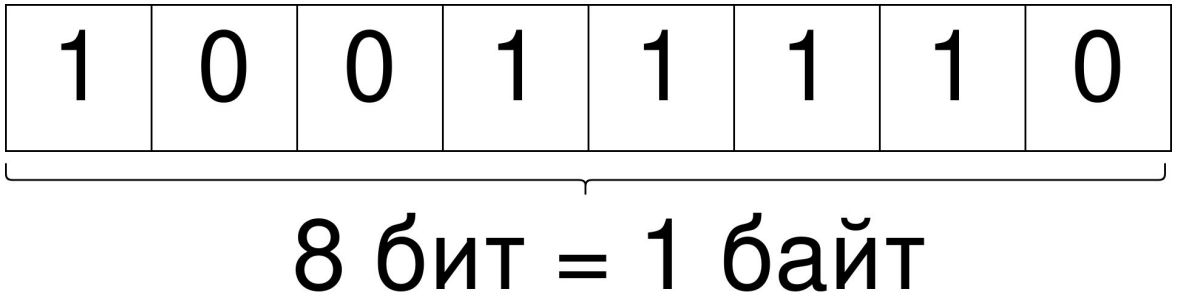
- **Информация** – нематериальная сущность, при помощи которой с любой точностью можно описывать реальные (материальные), виртуальные (возможные) и понятийные (абстрактные) сущности.
- Описываемому объекту (или понятию) ставится в соответствие некоторое число.
- Информация может быть двух видов: **дискретная** информация и **непрерывная** (аналоговая).
- При переводе непрерывной информации в дискретную важна частота дискретизации v, определяющая период (T=1/v).



# Бит, байт

- В современных пользовательских и серверных вычислительных устройствах общего назначения информация, хранимая на носителях и данные использующиеся в памяти представлены в виде набора дискретных состояний – битов
- **Бит** – единица измерения количества информации, использующаяся в компьютерных системах (сигнал)
- **Байт** – набор из 8-ми битов, представляющих собой удобный вид представления информации в вычислительных устройствах



# Данные

- **Пример**. Примитивные типы данных в языке программирования C++

| Ключевое слово | Тип | Диапазон допустимых значений (включительно) | Пример |
|---|---|---|---|
| byte | 8-битное целочисленное значение | от -128 до 127 | 123 |
| short | 16-битное целочисленное значение | от -32768 до 32767 | 12345 |
| int | 32-битное целочисленное значение | от -2147483648 до 2147483647 | 1234567890 |
| long | 64-битное целочисленное значение | от -9223372036854775808 до 9223372036854775807 | 1234567890 |
| float | 32-битное значение с плавающей точкой | приблизительно ±3.40282347E+38F (6-7 значащих десятичных цифр) | 123.45f |
| double | 64-битное значение с плавающей точкой | приблизительно ±1.7976931348623157E+308F (15 значащих десятичных цифр) | 123.456 |
| char | 16-битное значение Unicode | от 0 до 65535 | ‘a’ |
| boolean | истина или ложь | true или false | true |

# Кодовые таблицы символов

- **ASCII7** — первая кодировка, пригодная для работы с текстом. Помимо маленьких букв английского алфавита и служебных символов, содержит большие буквы английского языка, цифры, знаки препинания и другие символы. **(7 бит)**
- **ASCII** — первая кодировка, в которой стало возможно использовать символы национальных алфавитов. **(8 бит)**
- **КОИ8-R** — первая русская кодировка. Символы кириллицы расположены не в алфавитном порядке. **(8 бит)**
- **CP866** — русская кодировка, использовавшаяся на компьютерах IBM в системе DOS. **(8 бит)**
- **Windows-1251** — русская кодировка, использовавшаяся в русскоязычных версиях операционной системы Windows в начале 90-х годов. Кириллические символы идут в алфавитном порядке. **(8 бит)**
- **UTF8** — распространённый стандарт кодирования символов, позволяющий более компактно хранить и передавать символы Юникода, используя переменное количество байт (от 1 до 4), и обеспечивающий полную обратную совместимость с 7-битной кодировкой ASCII. **(8 бит)**

# Кодовые таблицы символов

| UTF-8 | Представленные символы |
|---|---|
| 0xxxxxx | ASCII, в том числе английский алфавит, простейшие знаки препинания и арабские цифры |
| 110xxxx 10xxxxxx | кириллица, расширенная латиница, арабский алфавит, армянский алфавит, греческий алфавит, еврейский алфавит и коптский алфавит; сирийское письмо, тана, нко; Международный фонетический алфавит; некоторые знаки препинания |
| 1110xxxx 10xxxxxx 10xxxxxx | все другие современные формы письменности, в том числе грузинский алфавит, индийское, китайское, корейское и японское письмо; сложные знаки препинания; математические и другие специальные символы |
| 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx | музыкальные символы, редкие китайские иероглифы, вымершие формы письменности |
| 111111xx | служебные символы c, d, e, f |


# Часть 3. Вычислительная инфраструктура и вычислительные устройства

# Вычислительные устройства

- **Основные характеристики вычислительного устройства:**

1. Вычислительная мощность (процессор)
2. Оперативная память (ОЗУ)
3. Хранилище (дисковое пространство)

# Дисковые накопители

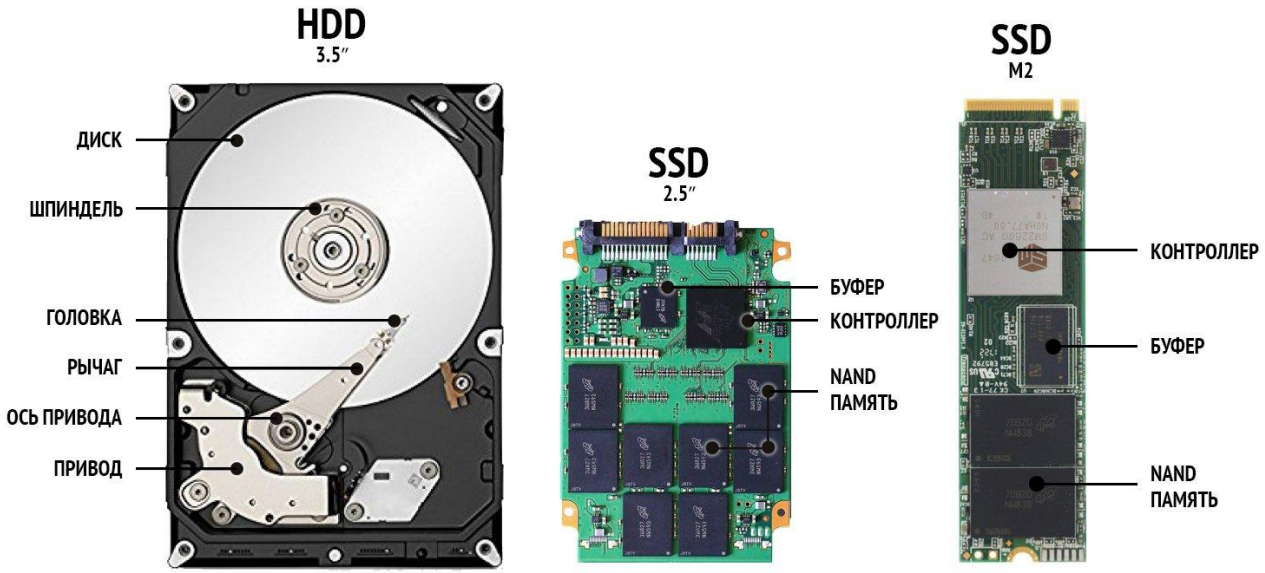
Дисковый накопитель отвечает за долговременное хранение информации пользователя. Это файлы, медиа и данные, которые должны храниться при отсутствии питания от сети.

Дисковый накопитель хранит данные для запуска операционной системы компьютера и данные прикладных программ для работы с ними.

- **От вида накопителя будут зависеть:**

1. Долговечность (надежность),
2. Скорость работы (чтение и запись),
3. Ёмкость (общий размер данных),
4. Стоимость (цена за единицу памяти)

| | HDD | SSD | NVMe |
|---|---|---|---|
| Скорость чтения/записи | ~ 50-100 Мб/с | ~ 300-500 Мб/с | ~ 1500-3000 Мб/с |
| Цена за Гб | низкая | средняя | высокая |
| Надежность | ~ 20 лет | ~ 10 - 15 лет | ~ 5 - 10 лет |
| Поддержка SATA | есть | есть | есть |



# Жесткие диски

- **Жесткий диск (или HDD)** — устройство хранения данных, принцип записи информации в котором заключается в намагничивании областей на поверхности магнитных дисков (пластин).

- Для организации хранения данных магнитный диск разбивается на дорожки и сектора, а совокупность дорожек, расположенных одна над другой (на нескольких пластинах), называется цилиндром.

- В зависимости от объема памяти, внутри корпуса HDD могут находиться до восьми пластин. Пластины крепятся к шпинделю, вращающемуся со скоростью от 4 до 15 тысяч оборотов в минуту (rpm). Запись и чтение информации с пластины осуществляется при помощи магнитной головки.

# Твердотельные накопители

- **Твердотельный накопитель (или SSD)** — устройство, использующее для хранения информации флеш-память.

- **Флеш-память (или flash memory)** — разновидность твердотельной полупроводниковой энергонезависимой перезаписываемой памяти. Она может быть прочитана сколько угодно раз (в пределах срока хранения данных, типично — 10-100 лет), но писать в такую память можно лишь ограниченное число раз (максимально — около миллиона циклов).

# Скорость чтения и записи данных с диска

- **Скорость чтения** измеряет, насколько быстро накопитель может «читать» или получать доступ к файлам, хранящимся на нем. Например, SSD с более высокой скоростью чтения может запустить гигабайтный файл быстрее. Это помогает сократить время загрузки компьютера, так как чтение больших файлов, необходимых для загрузки операционной системы, займет меньше времени.

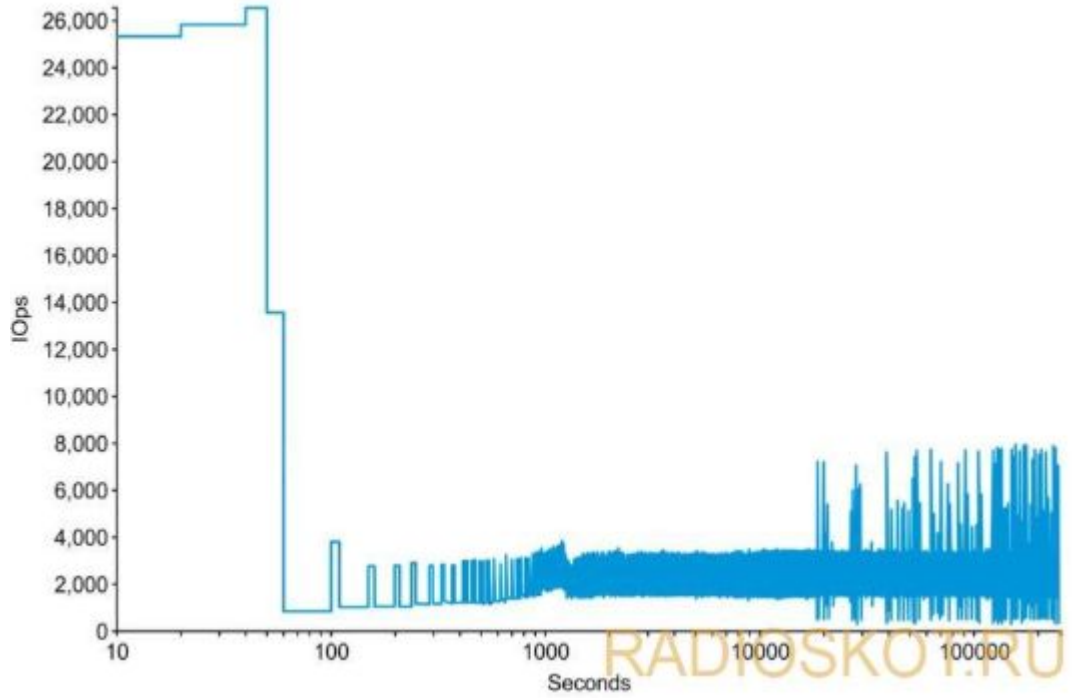
**Скорость записи** измеряет, насколько быстро файл может быть записан на диск. Чаще всего пользователь сталкивается со «скоростью записи», когда пытается скопировать файл из одного места в другое. Чем выше скорость чтения, тем меньше времени потребуется для копирования.


# Долговечность диска

- Мерой эффективности и быстродействия SSD является количество операций ввода/вывода в секунду (IOPS, Input/Output Operations per Second).

- SSD выполняет различные действия в фоновом режиме, связанные с удалением устаревших секторов (сборка мусора), обеспечением равномерного использования всех блоков памяти (выравнивание износа), обновлением сохраненных данных и так далее.
- Некоторые факторы, снижающие производительность диска:

- ошибки чтения(из-за увеличения количества поврежденных областей памяти);
- условия окружающей среды (температура).

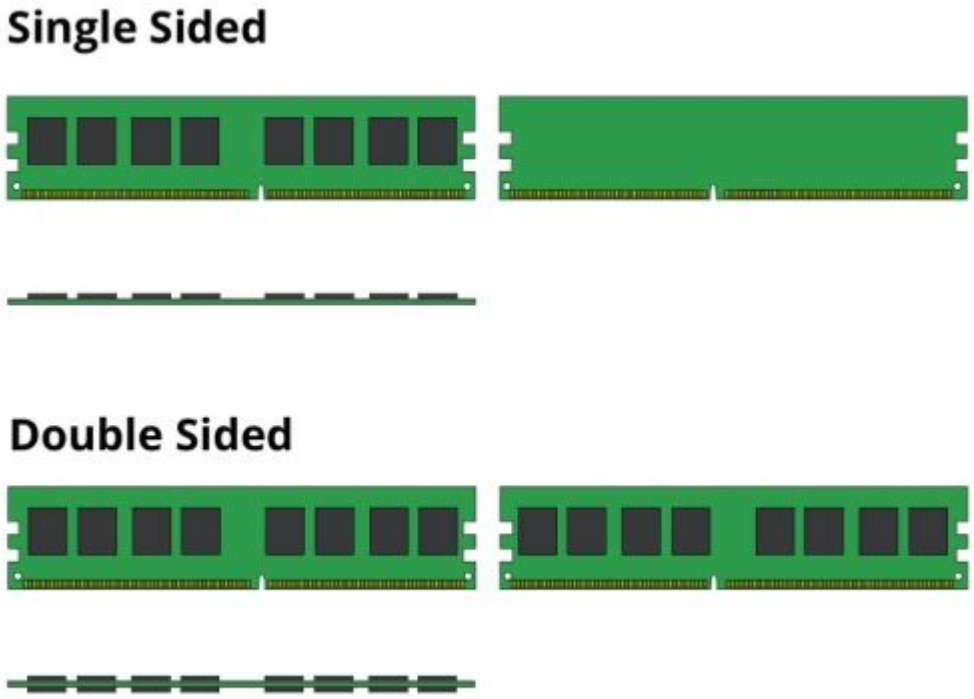


# Оперативная память

**Оперативная память компьютера** – энергозависимая часть системы компьютерной памяти, в которой во время работы компьютера хранится выполняемый машинный код (программы), а также входные, выходные и промежуточные данные, обрабатываемые процессором.

**Оперативное запоминающее устройство** (ОЗУ) – устройство, реализующее функции оперативной памяти.

От стандарта ОЗУ и размера оперативной памяти зависит возможное число запущенных процессов обработки информации в рамках одного вычислительного устройства.



# Современные стандарты ОЗУ

Современные ОЗУ отличаются стандартами хранения.

Более новые версии стандартов отличаются более высокими номерами DDR.

Самый современный стандарт – DDR5 отличается высокой пропускной способностью, максимальным размером памяти, шириной машинного слова, скоростью работы.

Каждый новый стандарт памяти претерпевает значительные инженерные хитрости связанные с изменением задержки постановки данных на шину

| | DDR3 | DDR4 | DDR5 |
|---|---|---|---|
| Напряжение | 1.3/1.5В | 1.2В | 1.1В |
| Максимальная скорость передачи | 1.6 Гб/с | 3.2 Гб/с | 6.4 Гб/с |
| Пропускная способность | 17 Гб/с | 25 Гб/с | 32 Гб/с |
| Максимальный размер 1 планки | 8 Гб | 32 Гб | 128 Гб |
| Кол-во каналов передачи | 1 | 1 | 2 |

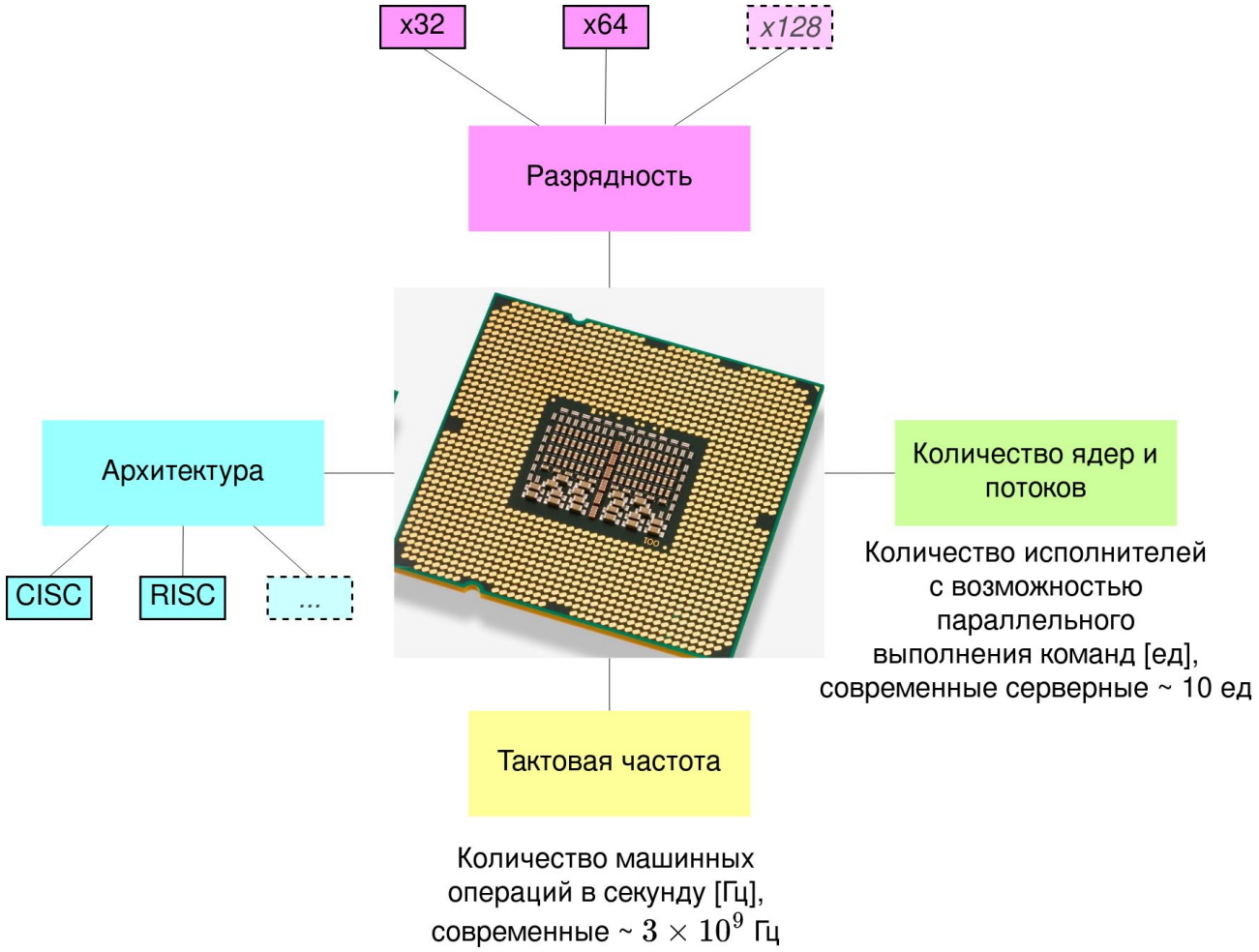
# Процессор

Центральный процессор – интегральная схема, исполняющая машинные инструкции (коды программ).

Машинный код – специфицированный набор битов, обозначающих номер инструкции и поля данных над которыми необходимо произвести инструкции.

- **Процессор с точки зрения обработки данных характеризуется:**

1. Тактовой частотой,
2. Количеством ядер и потоков,
3. Быстродействующей памятью (кэш),
4. Архитектурой,
5. Разрядностью



# Инфраструктура вычислений

На данный момент выделяют следующие виды вычислительных инфраструктур:

1. Персональные компьютеры (терминал доступа к серверу)
2. Локальные вычислительные сервера
3. Частный облачный сервис
4. Общедоступный облачный сервис

# Персональные компьютеры

На данный момент персональные компьютеры используются в роли терминалов доступа к вычислительным серверам


**Типичная конфигурация современного ПК:**

1. Центральный процессор (x64, 4 ядра, ~ 2.6 ГГц)
2. Оперативная память DDR4 8–16 Гб
3. Дисковый накопитель 500–1000Гб (HDD/SSD)
4. Операционная система Windows/Linux/MacOS с GUI

# Локальный вычислительный сервер

**Серверный компьютер** – единица серверной вычислительной инфраструктуры

Производители серверов предлагают устанавливать серверный компьютер в стойки, а стойки в кластер с применением сетевого взаимодействия между устройствами для возможной передачи данных между ними

**Типичная конфигурация одной серверной стойки:**

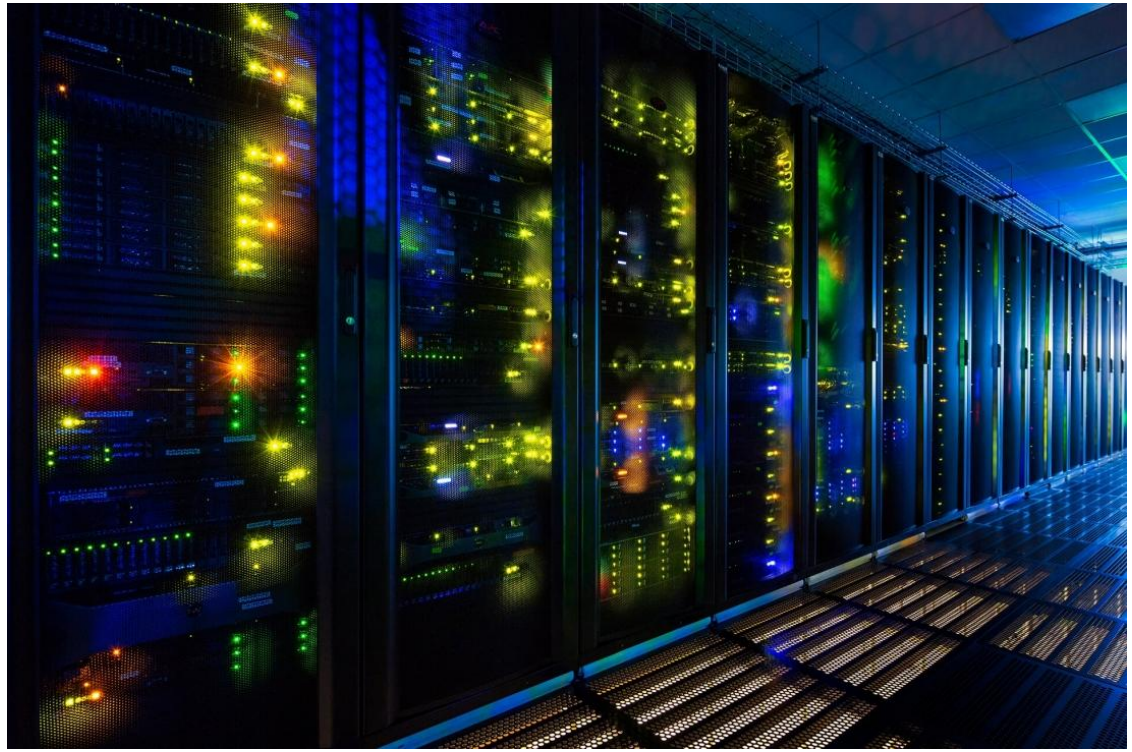
1. Центральный процессор (x64, 20 ядер, ~ 3 ГГц, до 8 процессоров)
2. Оперативная память DDR4 ~ 512-2048 Гб
3. Дисковый накопитель ~ 10-100Тб (HDD/SSD)
4. Операционная система Linux Server CLI (Debian / CentOS / Red Hat)

# Центры обработки данных

**Центры обработки данных (ЦОД)** — это специализированное здание или помещение, в котором компания размещает серверное и сетевое оборудование с последующим подключением клиентов к сети.

Функции ЦОД — обеспечить стабильную и безотказную работу размещённого в нём оборудования. Кроме этого, любой дата-центр предоставляет защищённые каналы связи, по которым происходит обмен данными.

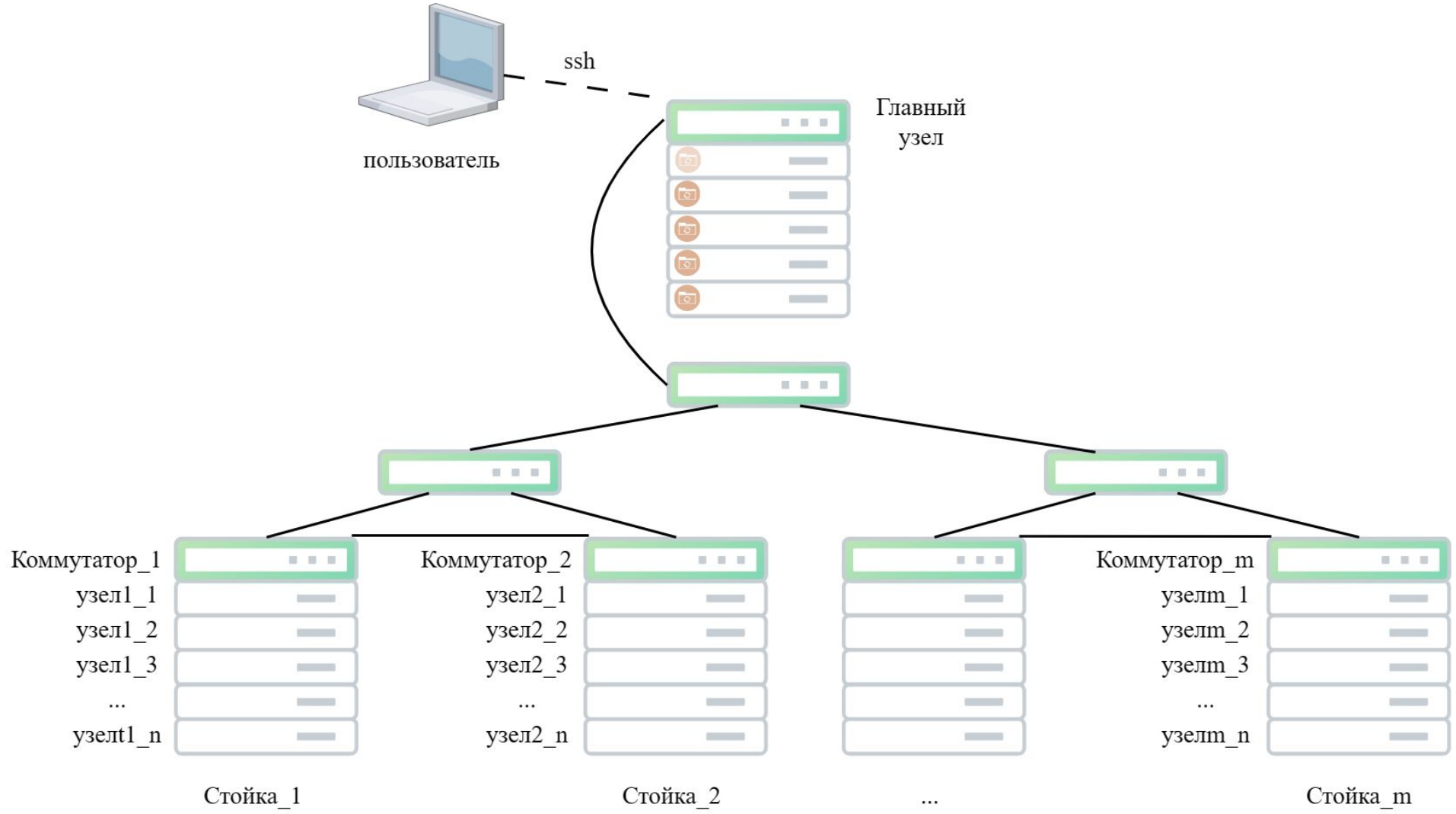
ЦОД обслуживает корпоративных клиентов и обеспечивает их ресурсами для вычислений и организации бизнеса.



# Классы ЦОД

| Уровень | Особенности | Отказоустойчивость |
|---|---|---|
| Tier 1 | В случае отказов работа ЦОД прерывается. Не предусмотрено обязательное использование источников бесперебойного питания и возможность резервирования данных. | 99,67% |
| Tier 2 | Предусматривает улучшенные условия размещения оборудования: резервные источники электроснабжения, фальшполы, избыточные системные ресурсы, улучшенные системы охлаждения. | 99,75% |
| Tier 3 | Не нужно останавливать для ремонта и профилактических работ. Для соответствия Tier 3 ЦОД должен иметь полное резервирование всех систем жизнеобеспечения. | 99,98% |
| Tier 4 | В требования этого стандарта входит двойное резервирование и полное дублирование всей системы. | 99,99% |

# Архитектура ЦОД (упрощенная схема)



# Облачные сервисы

Операторы дата-центров и облачные сервисы на коммерческой основе предоставляют ресурсы для развертывания вычислений или платформ для обработки данных

Дата-центры предлагают до тысяч стоек для нужд бизнеса и других отраслей экономики

В РФ функционируют 4-5 крупнейших оператора дата-центров и до десятка крупнейших облачных сервисов у которых напрямую можно развернуть облачные сервисы вычислений

DataPro

Selectel

Yandex Cloud

truenetwork

SberCloud

#CloudMTS

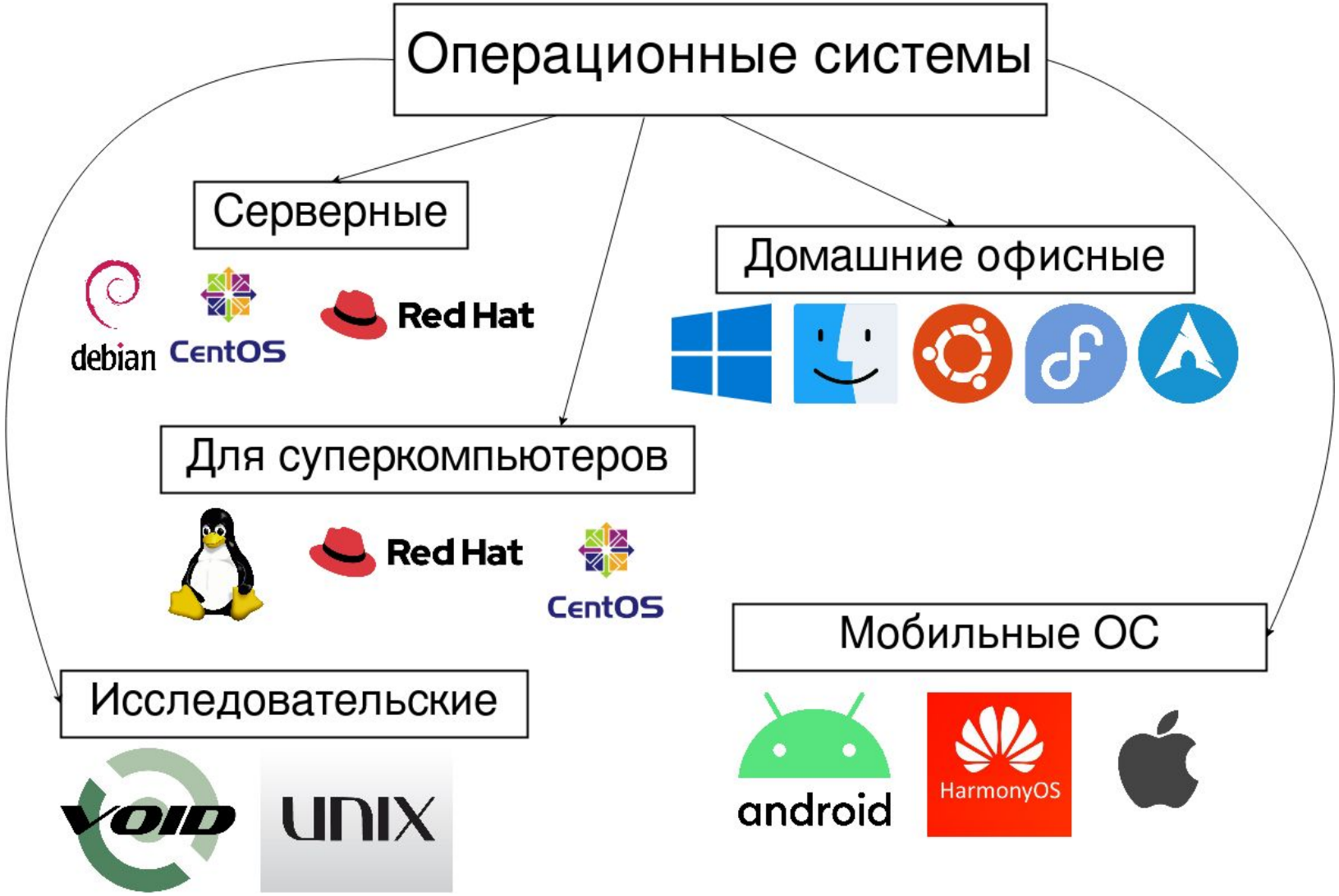
1cloud

DataSpace

CERTIFIED TIER III DATA CENTERS

# Часть 4. Операционные системы для работы с данными

# Классификация операционных систем



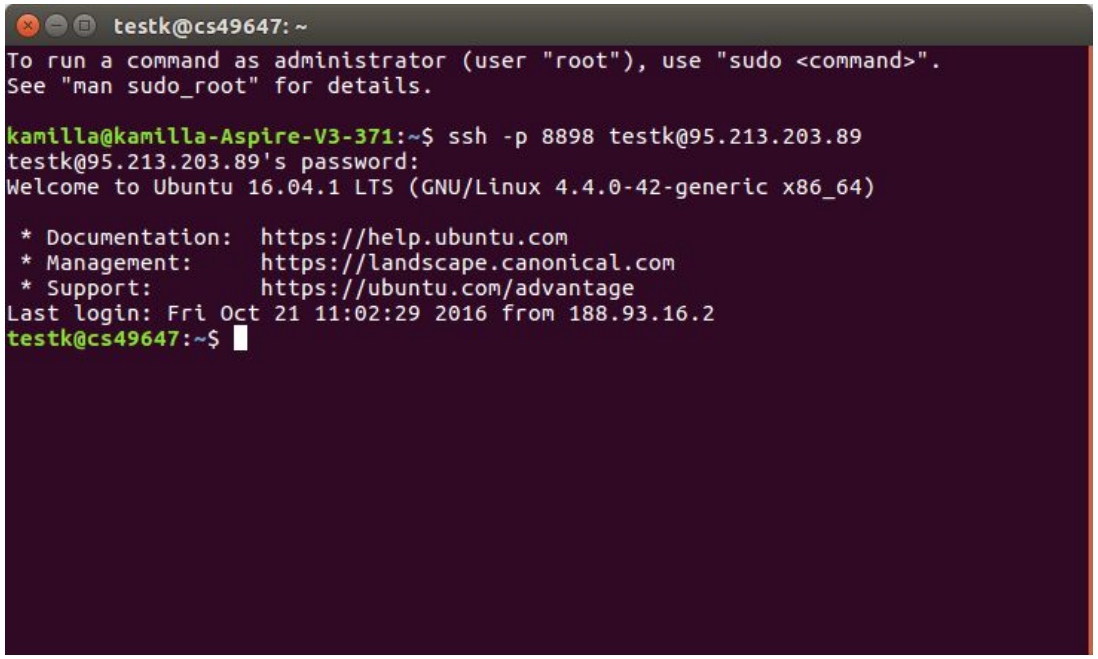
# Серверные операционные системы

**Серверные ОС** – предназначены для управления программным обеспечением, которое в свою очередь обслуживает всех пользователей сети, как внутренней, так и внешней

Серверные ОС не предоставляют специализированный графический инструментарий управления системой и управляются напрямую с использованием командной строки

Серверные ОС являются более предпочтительными с точки зрения экономии ресурсов и гибкости использования управления системой на основе команд

Наибольшее распространение получили Linux системы с UNIX-подобными утилитами командной строки

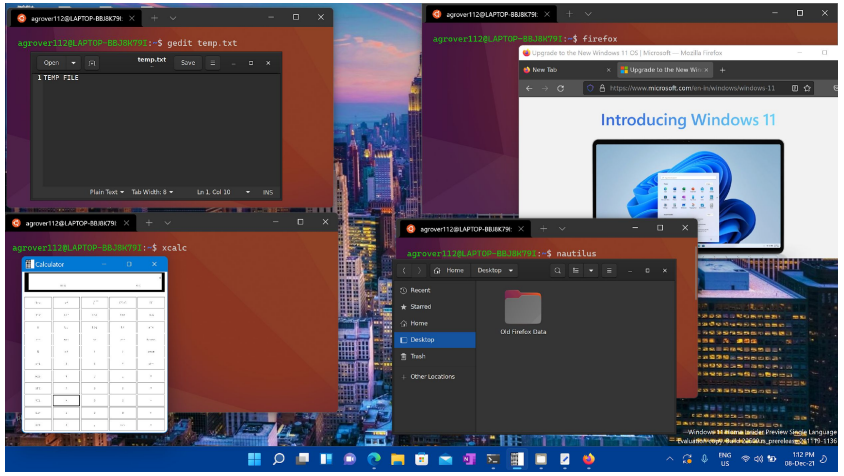


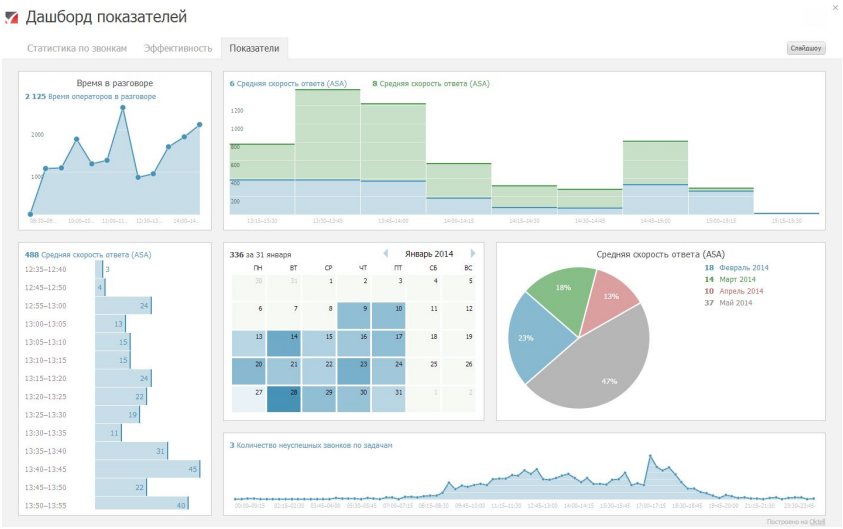

# Офисные операционные системы

**Офисные/пользовательские ОС** снабжены графической оболочкой (интерфейсом), удобной для взаимодействия с компьютером посредством координатного устройства ввода-вывода (мышь, графический планшет, и т.д.)

Пользовательские ОС предоставляют возможность пользователю взаимодействовать с компьютером посредством визуальной ориентации

Пользовательские ОС позволяют визуализировать результаты вычислений, анализа и предлагают возможность пользователю воспринимать мультимедийную информацию визуального характера



# Взаимодействие ОС

Вычисления

Визуализация

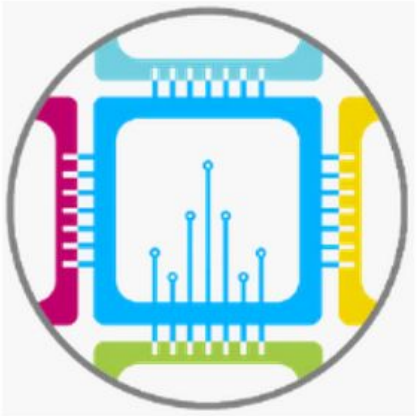
Серверная ОС

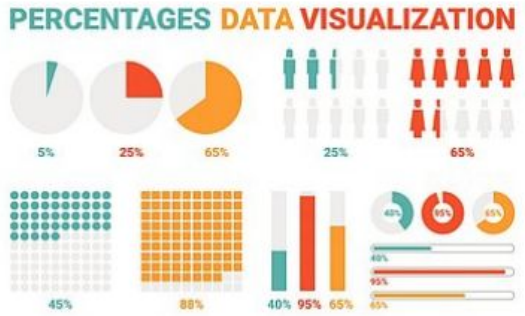
Файлы данных

Протокол передачи данных

Задачи на вычисления

Пользовательская ОС





# Часть 5. Форматы, файлы и введение в файловые системы

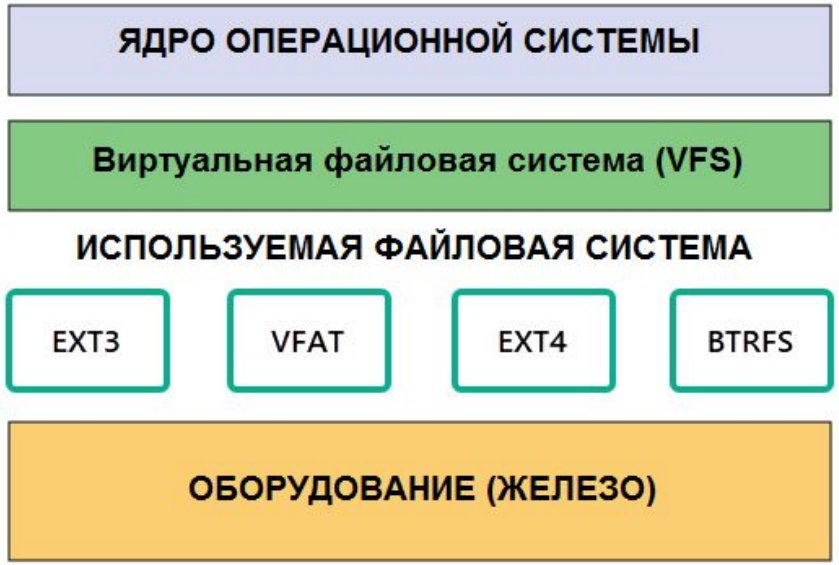
# Файловые системы

**Файловая система** определяет формат содержимого и способ физического хранения информации, которую принято группировать в виде файлов.

Конкретная файловая система определяет **размер имен файлов** (и каталогов), **максимальный возможный размер файла** и **раздела**, **набор атрибутов файла.**

Некоторые файловые системы предоставляют сервисные возможности, например, разграничение доступа или шифрование файлов.

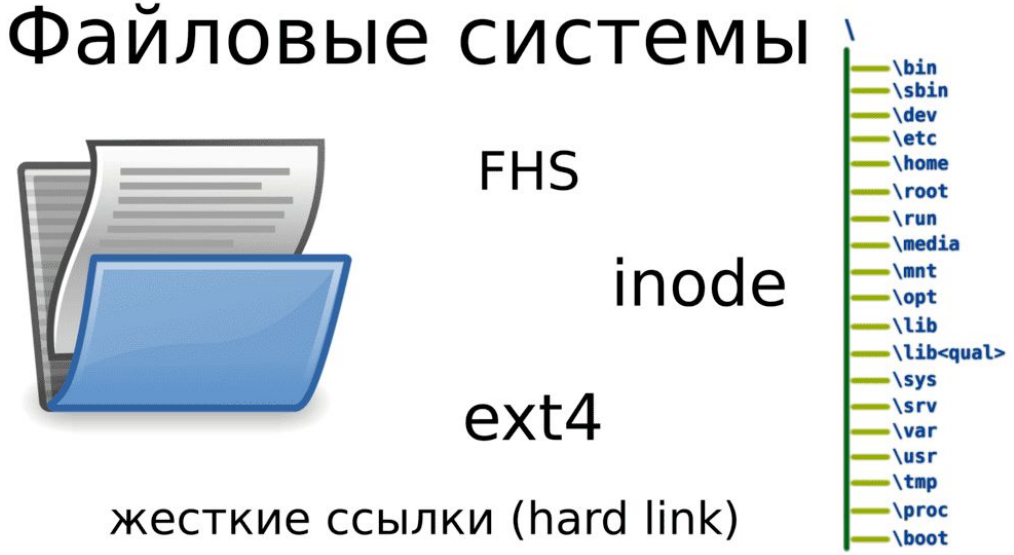
Файловая система связывает **носитель информации** с одной стороны и **набор прикладных команд** для доступа к файлам — с другой



# Файловые системы

**Файловая система** – это инструмент, позволяющий операционной системе и программам обращаться к нужным файлам и работать с ними. При этом программы оперируют только названием файла, его размером и датой создания. Все остальные функции по поиску необходимого файла в хранилище и работе с ним берет на себя файловая система накопителя.

**Файловая система** устанавливает правила на эксплуатацию и организацию данных на накопителе, и тем самым экономит ресурсы операционной системы и рабочих программ. К тому же наличие файловой системы позволяет использовать накопитель на разных компьютерах без каких-либо предварительных настроек и оптимизации




**Windows**

FAT32

NTFS

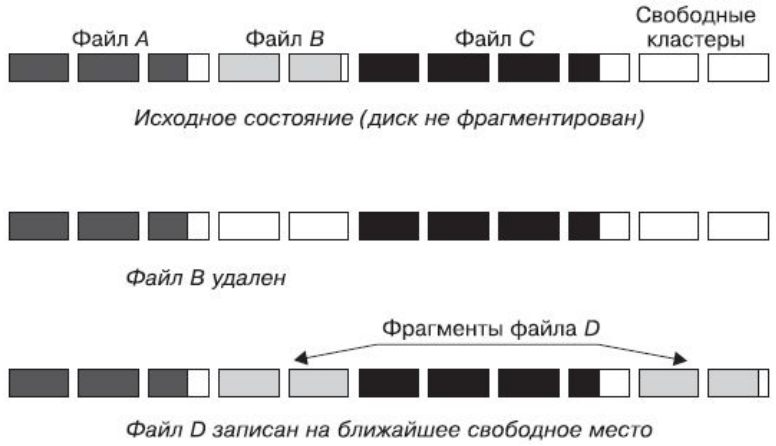
**Linux**

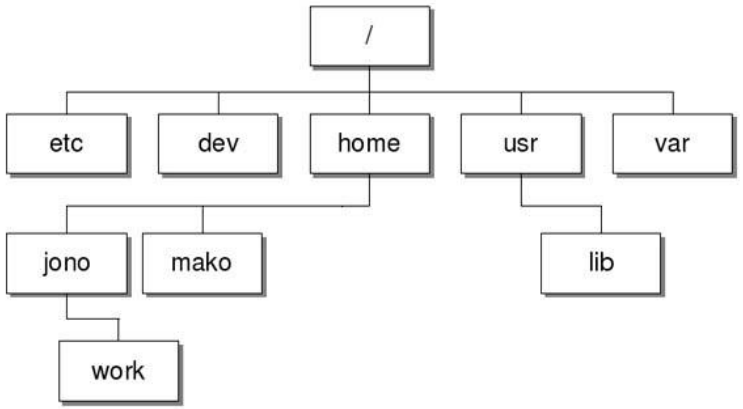
Extended File System

**(ext2fs,ext3fs)**

# Функции файловой системы

- Фрагментация файлов и их распределение на носителе.
- Поиск файла при запросе программ.
- Участие в создании, чтении и удалении файлов.
- Работа с атрибутами файлов: изменение названия, размера, времени последнего изменения, доступ к файлу и многое другое.
- Каталогизация и организация файлов.
- Защита файлов от несанкционированного доступа и сбоев системы.
- Определение права доступа к файлам.
- Восстановление информации в случае сбоев.



# Файл, формат файла

**Файл** — именованная область данных на носителе информации, используемая как базовый объект взаимодействия с данными в операционных системах

Обычно выделяют **исполняемые файлы** (программы) и собственно **файлы данных** (например, текстовые файлы или медиа)

**Формат файла** — способ организации данных внутри файла, позволяющий записывать в него информацию в соответствии с её смыслом и интерпретировать записанное.

# Полное имя файла

Путь к файлу

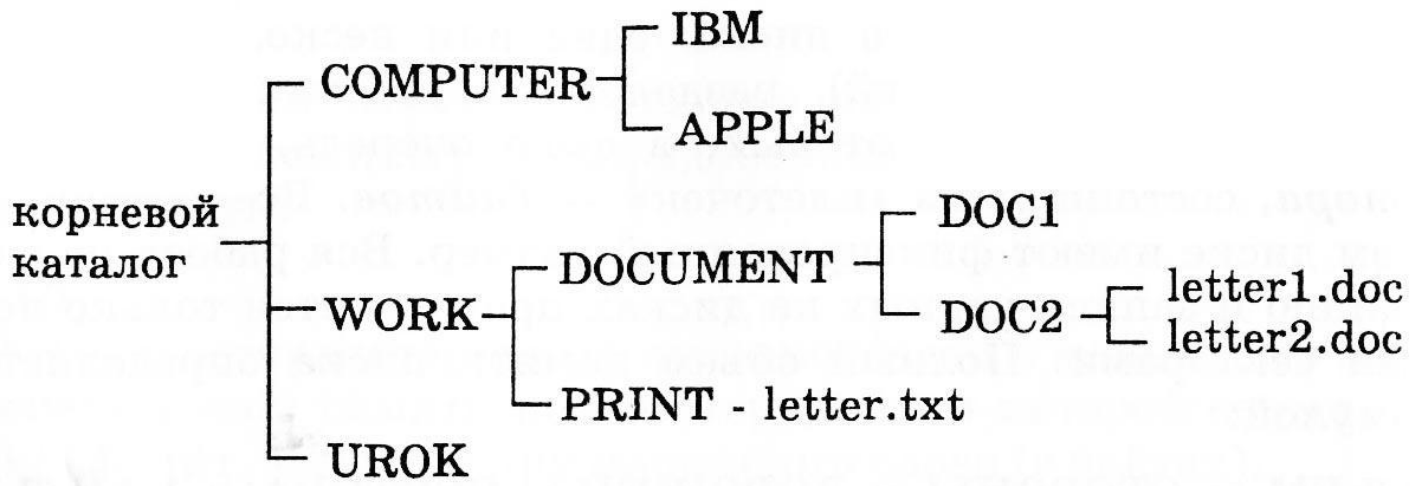
C:\User\Folder\AnotherFolder\file1.txt

Имя файла

Путь к файлу

/home/user/Images/Screen11.png

Имя файла



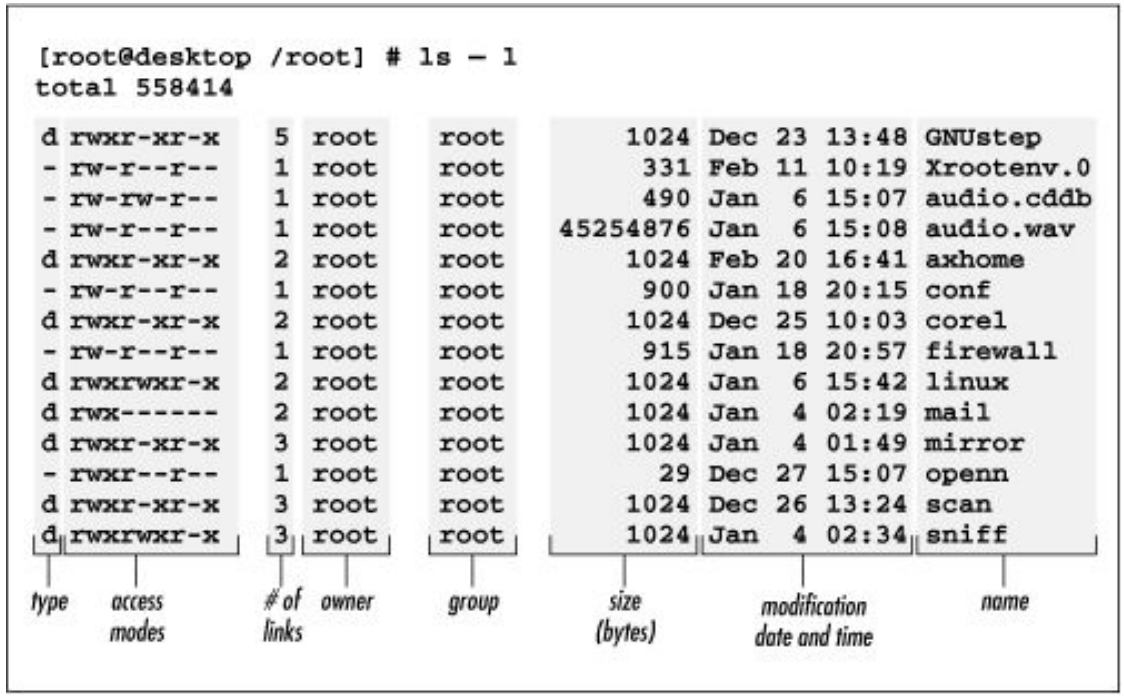
# Дополнительные атрибуты файла

**Расширение имени файла:** позволяет системе определить, каким приложением следует открывать данный файл. Обычно, часть имени, отделённая самой правой точкой в имени

**Время:** для файла могут быть определены временные метки создания, последней модификации, последнего доступа и другие

**Владелец и группа файла:** В некоторых файловых системах предусмотрено указание на владельца файла и группу-владельца

**Права доступа:** В некоторых файловых системах предусмотрена возможность для ограничения доступа пользователей к содержимому файла. Каждое право задаётся раздельно для владельца, для группы и для всех остальных.



# Права доступа в Linux

## Права доступа к файлам (числовая нотация)

*Примеры записи прав доступа в **двоичной** форме:*

110 110 110 (все могут читать и изменять )
111 000 000 (полный доступ имеет владелец файла)
110 100 100 (все могут читать, владелец также изменять)
111 101 101 (все могут читать и исполнять, владелец также изменять)

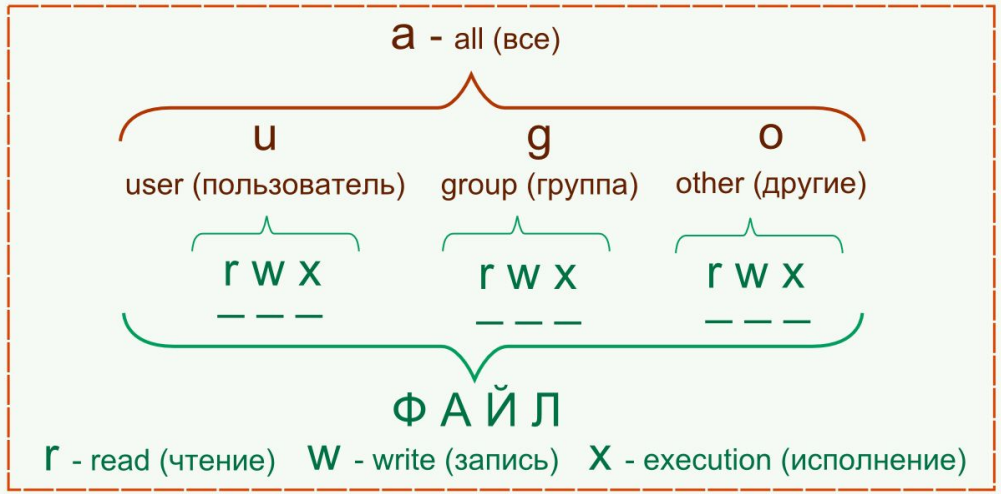
*Перевод представления прав доступа к восьмеричной форме:*

| rwx-представление | Двоичное число | Восьмеричное число | Значение |
|---|---|---|---|
| - - - | 0 0 0 | 0 | Все запрещено |
| - - x | 0 0 1 | 1 | |
| - w - | 0 1 0 | 2 | |
| - w x | 0 1 1 | 3 | |
| r - - | 1 0 0 | 4 | Только чтение |
| r - x | 1 0 1 | 5 | Чтение и исполнение |
| r w - | 1 1 0 | 6 | Чтение и запись |
| r w x | 1 1 1 | 7 | Все разрешено |

*Примеры записи прав доступа в **восьмеричной** форме:*

666 (все могут читать и изменять )
700 (полный доступ имеет владелец файла)
644 (все могут читать, владелец также изменять)
755 (все могут читать и исполнять, владелец также изменять)

## Права доступа к файлам (rwx)

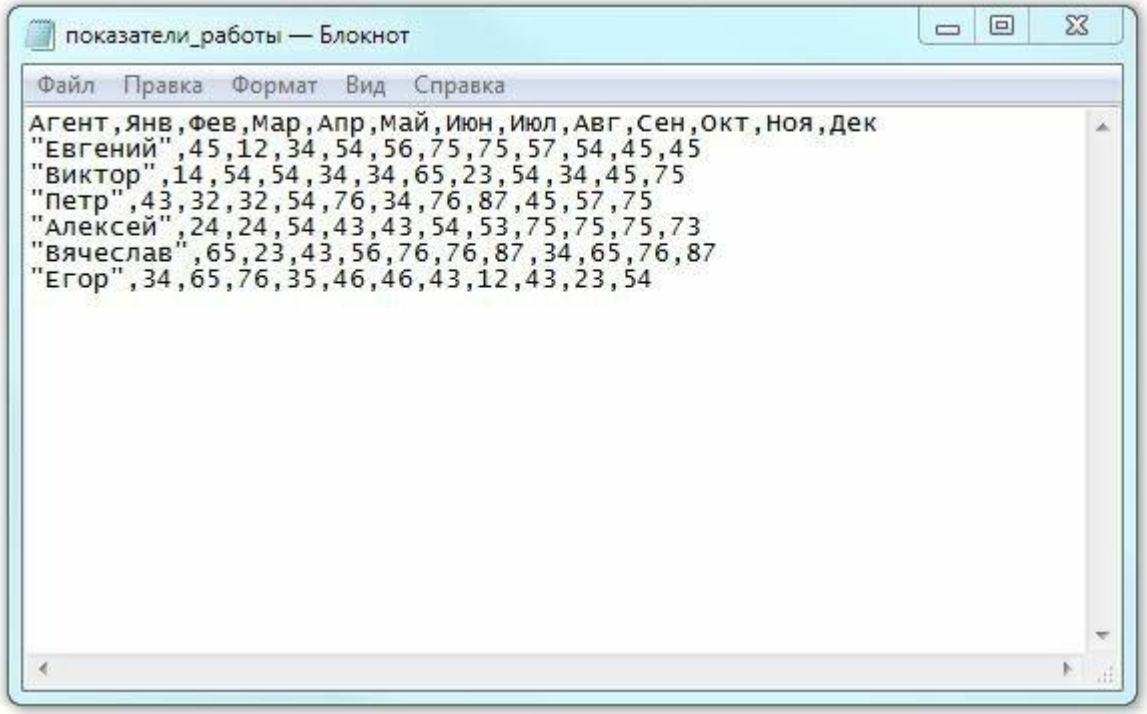



*Примеры:*

rw-rw-rw- (все могут читать и изменять )
rwx------ (полный доступ имеет владелец файла)
rw-r--r-- (все могут читать, владелец также изменять)
rwxr-xr-x (все могут читать и исполнять, владелец также изменять)

# Операции с файлами

- **Открытие файла** – возможность обращения к файлу для последующих циклов чтения или записи данных
- **Закрытие файла** – завершение процесса чтения или записи в файл
- **Запись** – процесс помещения информации в файл из памяти или устройств ввода-вывода.
- **Чтение** – получение данных из файла в терминал или в область памяти компьютера.
- **Перемещение указателя** — указатель перемещается на указанное число байт вперёд или назад или перемещается по указанному смещению относительно начала или конца.

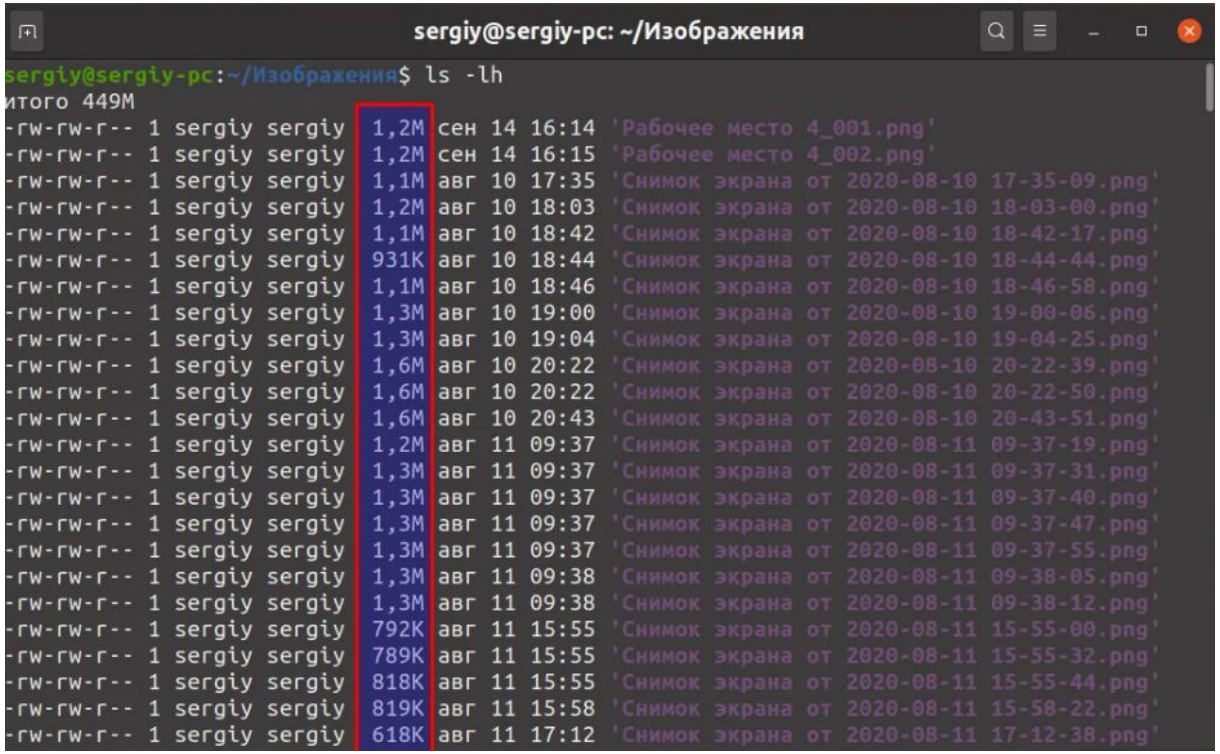


# Размер файла

- **Размер файла** – это показатель того, сколько данных содержит компьютерный файл или, наоборот, сколько места он занимает.

- Обычно размер файла выражается в единицах измерения, основанных на байтах. По соглашению, единицы измерения размера файла используют  метрический префикс (например, мегабайт и гигабайт).

- Максимальный размер файла, поддерживаемый файловой системой, зависит не только от емкости файловой системы, но и от количества битов, зарезервированных для хранения информации о размере файла.



# Типы файлов

- По способу организации файлы делятся на файлы с произвольным доступом и файлы с последовательным доступом.

1. **«Обыкновенный файл»** — файл, позволяющий операции чтения, записи, позиционирования внутри файла, изменения размера, иногда работу с атрибутами.
2. **Каталог** или **директория** (также «папка») — файл, содержащий записи о входящих в него файлах. Каталоги могут содержать записи о других каталогах, образуя древовидную структуру, а при наличии ссылок — сетевую структуру.
3. **Жёсткая ссылка** — одна и та же область информации может иметь несколько имён. Такие имена называют жёсткими ссылками (хардлинками). После создания жёсткой ссылки сказать, где «настоящий» файл, а где жёсткая ссылка, невозможно, так как имена равноправны.
4. **Символьная ссылка** — файл, содержащий в себе ссылку на имя нужного файла любого типа. Может ссылаться на любой элемент файловой системы, в том числе, и расположенный на другом физическом носителе.

# Источники информации


1. Андрей Найдич «Большие данные: насколько они большие?» – https://compress.ru/article.aspx?id=23469
2. Семенов Ю.А. (ИТЭФ-МФТИ) «Обзор компании IDC по проблемам цифровизации и ситуация в РФ» – http://book.itep.ru/4/7/digi_world.htm
3. Pro Hi-Tech в ЦОД Tier III. Дизельные ИБП, продвинутое охлаждение Schneider и многое другое – https://www.youtube.com/watch?v=ZINMxB7Jd2g
4. DataPro – Крупнейший независимый оператор дата-центров в России – https://datapro.ru/about