Балаковский инженерно-технологический институт - филиал
федерального государственного автономного
образовательного учреждения высшего образования
*«Национальный исследовательский ядерный университет
«МИФИ»*

«Большие данные. Анализ и обработка»

Выполнил: студент группы *ИФСТ-11*

Купцов Даниил Олегович

Проверила: *доцент кафедры ИСТ Очкур Галина Викторовна*

Балаково 2024 г.

Введение

- **Цель:** закрепление знаний и умений, приобретённых в результате освоения теоретических знаний в области информационных систем и технологий.
- **Задачи учебной практики:**
- практическое применение знаний путем разработки презентации и сайта по теме «Большие данные. Анализ и обработка»;
- совершенствование навыков решения информационных задач на конкретном рабочем месте.

- **Актуальность:**
- особенность современной жизни человека, в частности с сильным внедрением цифровых технологий, приводит к необходимости уметь хранить большой массив данных и уметь работать с ним.

Большие данные. Определение

- **Большие данные** - это огромные массивы данных или потоки, которые содержат информацию, требующие быстрой интерактивной обработки с целью исследования или проверки гипотез.
- Для реализации данных задач требуется *высокий уровень параллелизма* и *большой объём оперативной памяти.*
- Отличительными характеристиками больших данных являются **три параметра:** *объем* (от 150 гигабайт в сутки), *скорость* (большое поступление данных в малые сроки), *разнообразие* (данные неоднородны).

Обработка данных: приём и сбор данных

- Для обработки **больших данных** используются *сложные системы*, в которых можно выделить несколько **этапов:** приём, сбор, анализ данных и представление результатов.
- Первым этапом является приём данных. Его **задача** заключается в *базовой подготовке сведений* с целью приведения данных к единому формату представления.

- **Сбор данных** – второй компонент обработки данных, характеризующийся взаимодействием с системами хранения данных.
- В зависимости от разной степени структурированности данные обрабатываются по-разному:
- *данные структурированы:* преобразование по алгоритмам;
- *данные полуструктурированы:* интерпретация данных в общий формат;
- *данные не структурированы:* создание специального софта.

Обработка данных: анализ и представление результатов

- На этапе анализа ценность имеет именно сама **информация**, содержащаяся в данных. *Анализ данных* является самой трудоёмкой частью в процессе обработки данных.
- **Анализ данных** – различные *аналитические механизмы*, которые применяют *аналитические алгоритмы* для управления моделями и идентификации сущностями для получения новой содержательной информации, являющаяся искомой в рамках поставленной задачи.

- Полученная информация представляется на уровне потребления пользователем.
- **Конкретные практические задачи:**
- *мониторинг данных;*
- *генерация отчётов и запросов;*
- *трансформация данных.*

Облачные технологии

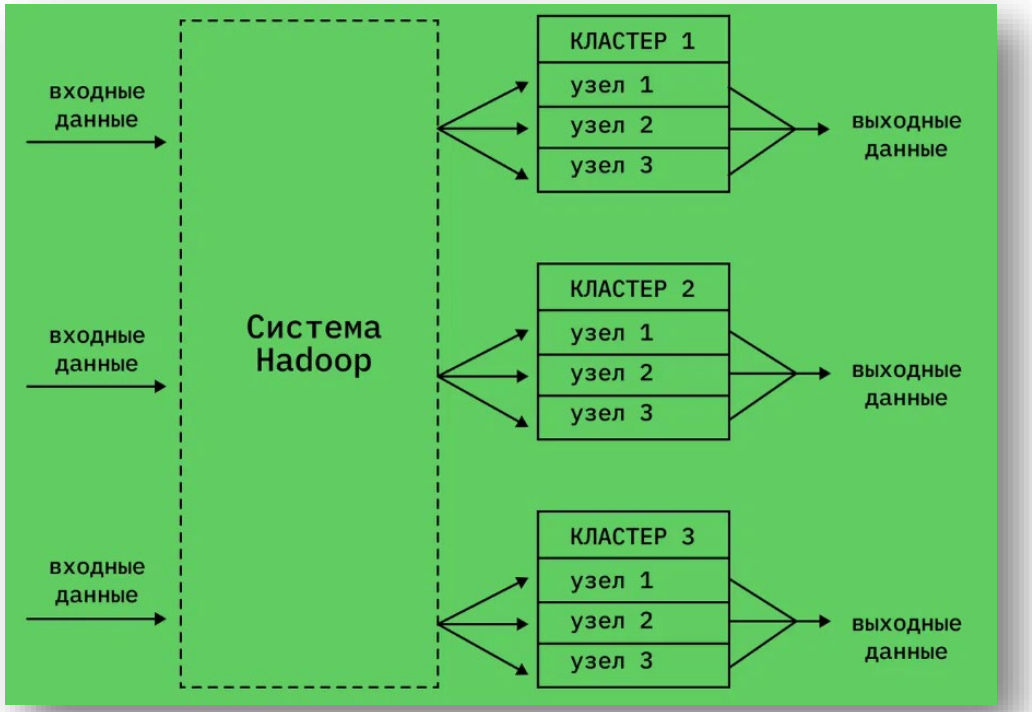
- **Облачные технологии** позволили пользователям использовать вычислительные ресурсы ровно на столько, на сколько им нужно:
- плата начисляется за **фактическое время** использование ресурсов;
- предоставление и освобождение ресурсов производятся пользователем через *веб-портал*;
- **Эластичные виртуальные машины** могут быть использованы не только в качестве хранилища, но и для создания на них собственных информационных систем.

Data Mining

- При наличии **большого массива неструктурированных данных,** анализ информации может *не представляться возможным,* поскольку вычислительных мощностей может просто не хватать для выполнения задачи в разумное время.
- Отсюда вытекает *необходимость систематизации данных,* что обеспечит их анализ. Один из распространённых способов – Data Mining.
- **Основу дата-майнинга** составляют всевозможные *методы классификации* и *моделирования.*

Hadoop

- Для работы с Big Data требуется **собственный инструментарий**.
- **Hadoop** – свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов.
- При поступлении на кластер обширных задач, **Hadoop** делит её на много мелких подзадач и выполняет каждую на своём узле.
- *Данные методы* позволяет параллельно решать несколько задач и в несколько раз быстрее дать конечный результат.



NoSQL и среда разработки R

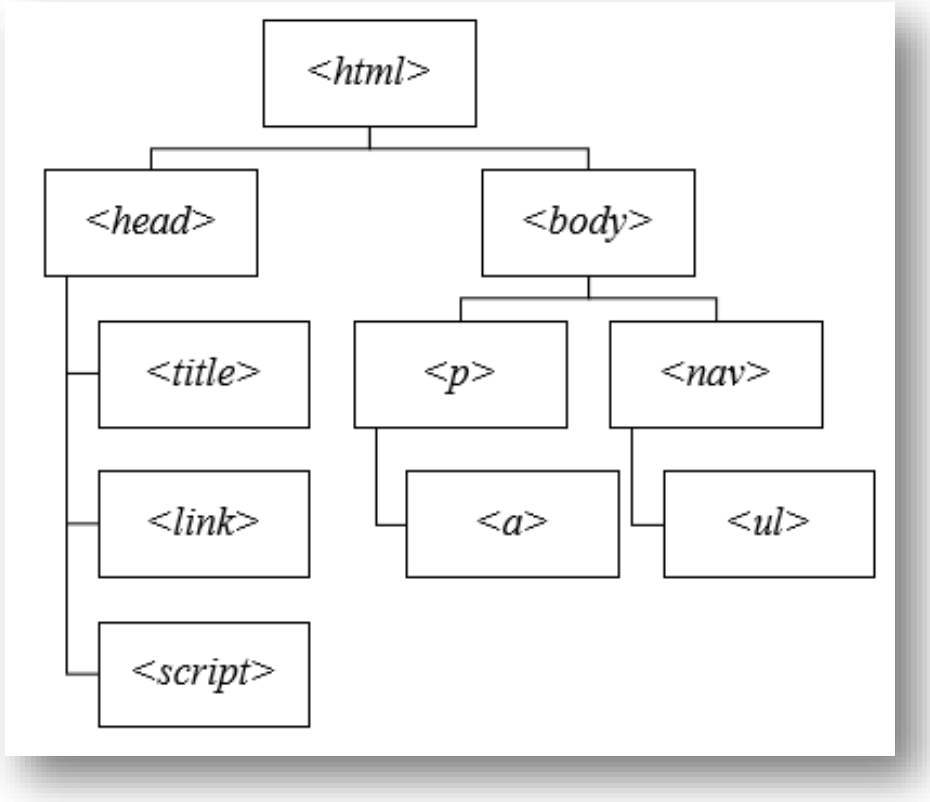
- **NoSQL** - название *разнородных систем* управления базами данных, отличных от реляционных таблиц с доступ к данным через средства языка **SQL**.
- Данную технологию стали применять ввиду того, что при больших объёмах хранилищ *базы данных тяжело масштабируются*.

- **Среда R** – это *язык программирования* и среда статистических вычислений и графического анализа.
- Предназначен для *математиков* и *статистов*, а также для коммерческих специальностей: *аналитиков данных* и *дата-сайентистов.*

- Были рассмотрены *основные особенности* **больших данных**, *методы обработки* больших данных, а также *технологии их анализа*, таких как *Hadoop*, софт *NoSQL* и среду *R*.
- На данный момент в среде дата-специалистов свою популярность приобрёл язык программирования *Python*, так как также отлично справляется с обработкой Big Data.

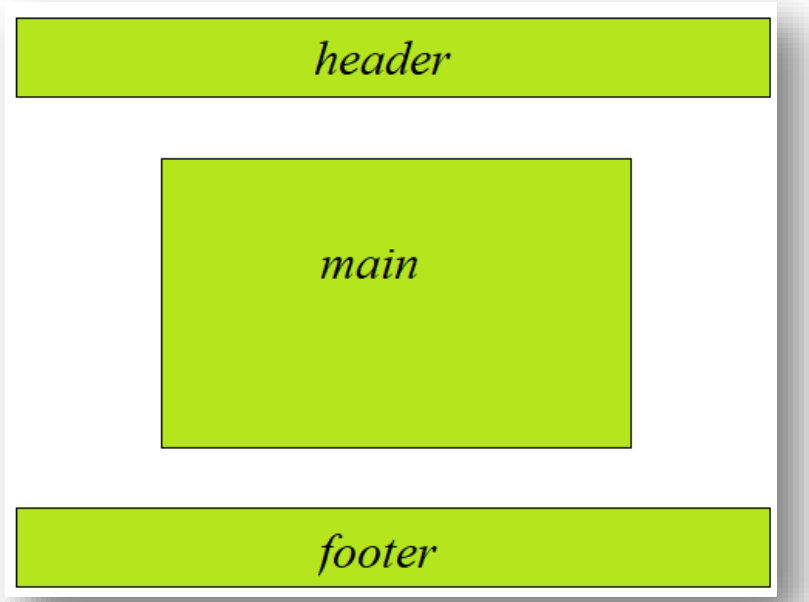
Создание веб-сайта. Софт

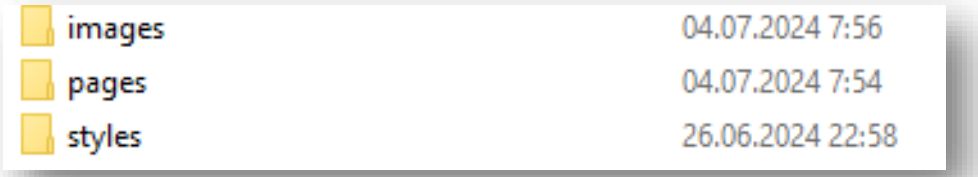
- Для реализации веб-страницы использован **HyperText Markup Language** (HTML) – *языка гипертекстовой разметки.*
- Для реализации графического дизайна сайта будет использован язык каскадных таблиц стилей CSS – **Cascading Style Sheets.**
- **Cascading Style Sheets** (CSS) – язык таблицы стилей, который позволяет прикрепить различные стили, например, к **HTML**-документу, такие как шрифт, цвет, особенности верстки (flex-, или grid-системы) и так далее.



Создание веб-сайта. Макет

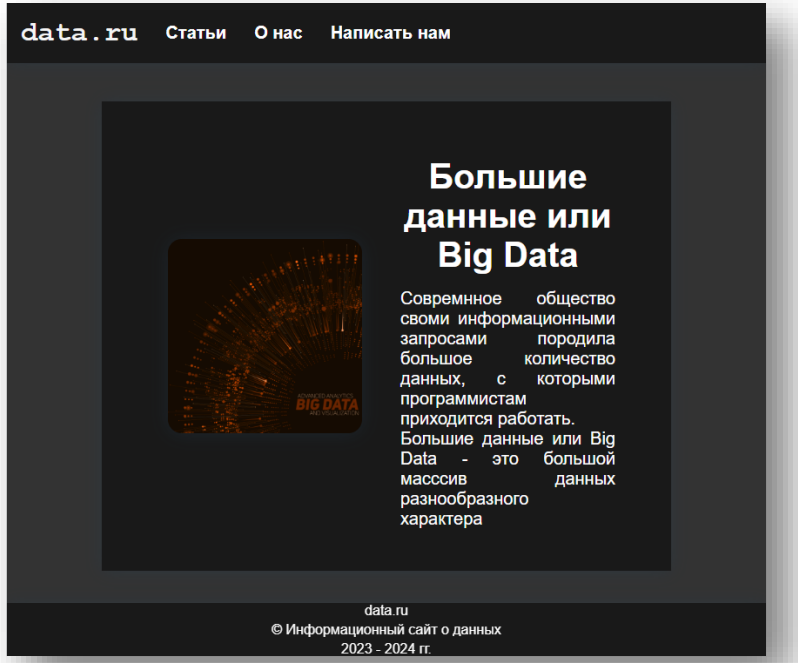
- Для всех страниц будет использован **одинаковый макет.**
- Для страницы **«Написать нам»** разделения на два логических раздела *не будет,* поскольку будет единая форма для заполнения.
- Для всех веб-страниц будет использован единая **«темная тема».**
- Цвет всех текстов будет белым.

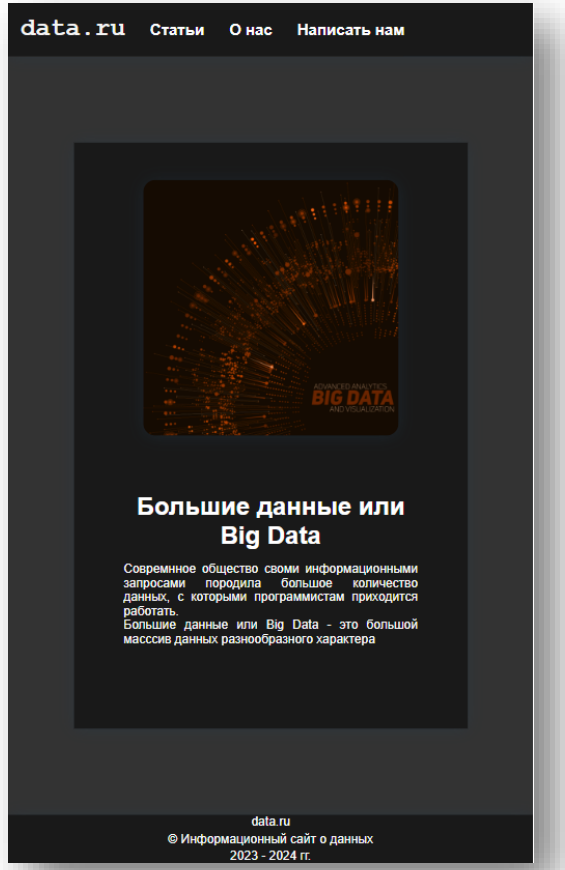




Создание веб-сайта. Верстка

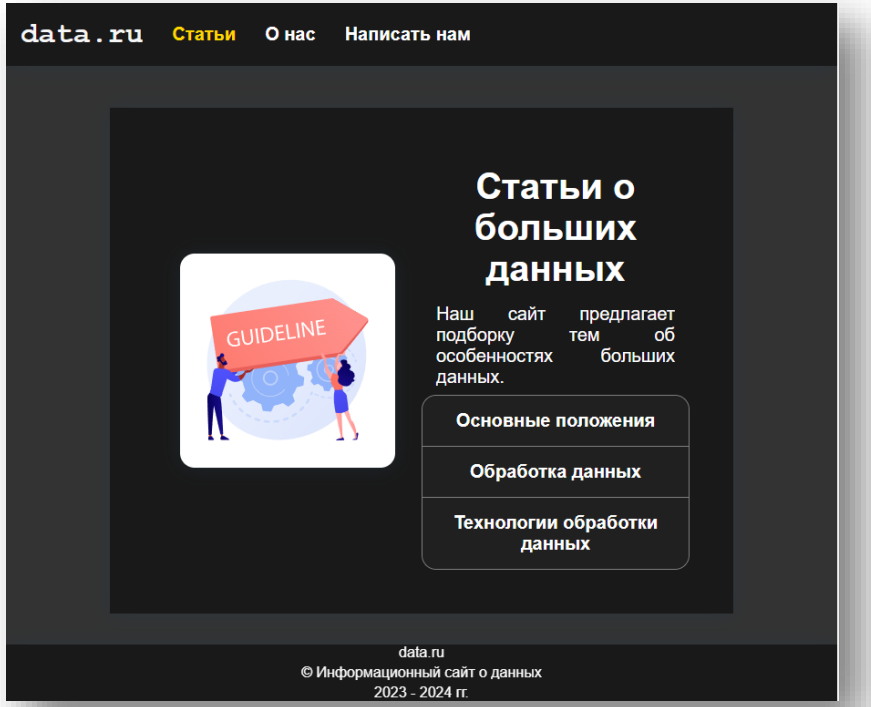
- Применим *flex-* и *grid-верстку* для реализации верстки веб-страницы, а также *позиционирование* в зависимости от ширины экрана.
- Подобные действия обеспечат *адаптивность* страницы.
- В **«голове»** сайта реализовано меню: линейное и интерактивное.

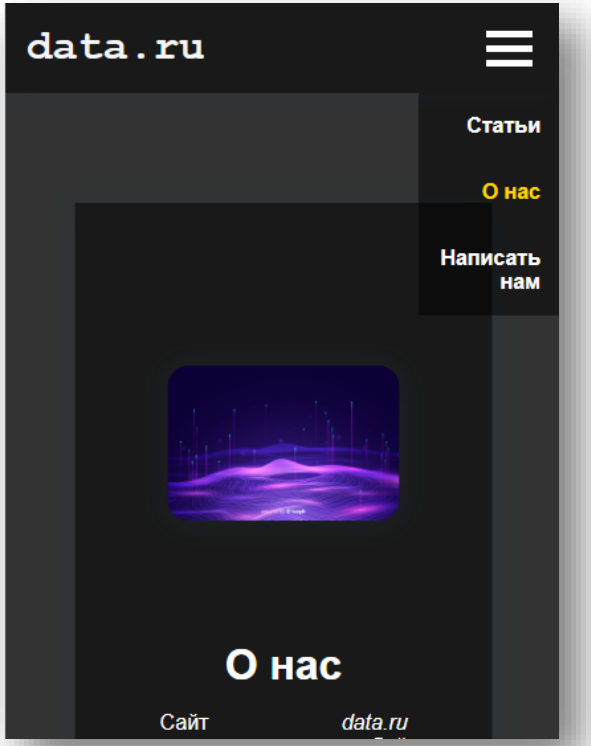




Создание сайта. Тестирование

- Под *тестированием* следует понимать *качество отображения* веб-страниц на различных устройствах
- Рассмотрены *следующие размеры ширины* для тестирования адаптивности:1280px, 992px, 768px, 576px, 420px.





Заключение

- Изучена *предметная область* **«Большие данные. Анализ и обработка».**
- Выявлены *основные особенности* данной сферы информационных технологий: этапы обработки данных и то, какие технологии существует для эффективного анализа.
- Был разработан **веб-сайт** с учётом требований к его адаптивности, интерактивности, и содержательности.
- Данный **веб-сайт** может быть полезен людям, которые заинтересовались наукой о данных.
- Также **веб-сайт** может быть размещён под собственным именем *«data.ru»* или продан дата-компаниям.

Спасибо за внимание!

Список использованной литературы

- 1. Большие данные в информатике / Большая российская энциклопедия : [сайт]. – 2024. URL: https://bigenc.ru/c/bol-shie-dannye-v-informatike-e4a22a (дата обращения: 03.07.2024г.)
- 2. Ланских Ю. В. Введение в большие данные : учебное пособие / Ю. В. Ланских, В. Г. Ланских, К. В. Родионов. – Киров : ВятГУ, 2023. – 172 с. – Текст: электронный // Лань: электронно-библиотечная система. – URL: https://reader.lanbook.com/book/408566 (дата обращения: 03.07.2024г.)
- 3. Что такое Big Data и как они устроены / Яндекс. Практикум – сервис онлайн образования : [сайт]. – 2024. URL: https://practicum.yandex.ru/blog/chto-takoe-big-data/ (дата обращения: 03.07.2024г.)
- 4. Что такое «Big Data» / Хабр – блоги об информационных технологиях : [сайт] – 2024. URL: https://habr.com/ru/companies/productstar/articles/503580/ (дата обращения: 04.07.2024 г.)

Список использованной литературы

- 5. Что такое Hadoop и почему аналитику данных полезно уметь с ним работать / Яндекс. Практикум – сервис онлайн образования : [сайт]. – 2024. URL: https://practicum.yandex.ru/blog/gde-i-zachem-ispolzuetsya-hadoop/ (дата обращения: 04.07.2024г.)

- 6. Hadoop: что, где и зачем / Хабр – блоги об информационных технологиях : [сайт] – 2024. URL: https://habr.com/ru/articles/240405/ (дата обращения: 04.07.2024 г.)

- 7. NoSQL: что это за базы данных, для чего они нужны и как работают / SkillBox – образовательная платформа : [сайт] – 2024. URL: https://skillbox.ru/media/code/nosql-chto-eto-za-bazy-dannykh-dlya-chego-oni-nuzhny-i-kak-rabotayut/?ysclid=ly7gjf4mhu311547234#stk-4 (дата обращения: 04.07.2024 г.)