Введение

Управление ИТ - контентом и анализ больших данных

Юдинцев В. В.

Кафедра математических методов в экономике

19 сентября 2023 г.

САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

Содержание

Содержание курса

В рамках дисциплины Управление ИТ - контентом и анализ больших данных рассматриваются

- способы управления контентом,
- способы обработки и анализа больших данных.

Управление контентом

Контент

https://www.collab.net/solutions/value-stream-management

Контент – это содержимое, информационное наполнение, связанное с ИТ-сервисом (может включать в себя документы, веб-страницы, изображения, мультимедиа и др. файлы и пр.).

Система управления содержимым

Система управления содержимым (англ. Content management system, CMS, система управления контентом) — информационная система или компьютерная программа, используемая для обеспечения и организации совместного процесса создания, редактирования и управления содержимым

Система управления содержимым

- Системы для управления корпоративным контентом (Enterprise Content Management System) — для работы с содержимым внутри какой-либо организации
- Системы для управления веб-содержимым (Web Content Management System) для поддержки работы веб-сайта.

WCMS

- Программный комплекс, предоставляющий функции создания, редактирования, контроля и организации веб-страниц.
- WCMS часто используются для создания блогов, личных страниц и интернет-магазинов и нацелены на пользователей, мало знакомых с программированием.

Структура WCMS = CMA + CDA

- Приложение для управления контентом (CMA) — это пользовательский интерфейс, который позволяет пользователям и создателям контента, проектировать, создавать, изменять и удалять контент с веб-сайта ”без помощи ИТ-отдела”.
- Приложение доставки контента (CDA) предоставляет серверные службы, которые берут контент, созданный пользователями в CMA, и превращают его в веб-сайт, к которому могут получить доступ посетители.

Типы WCMS

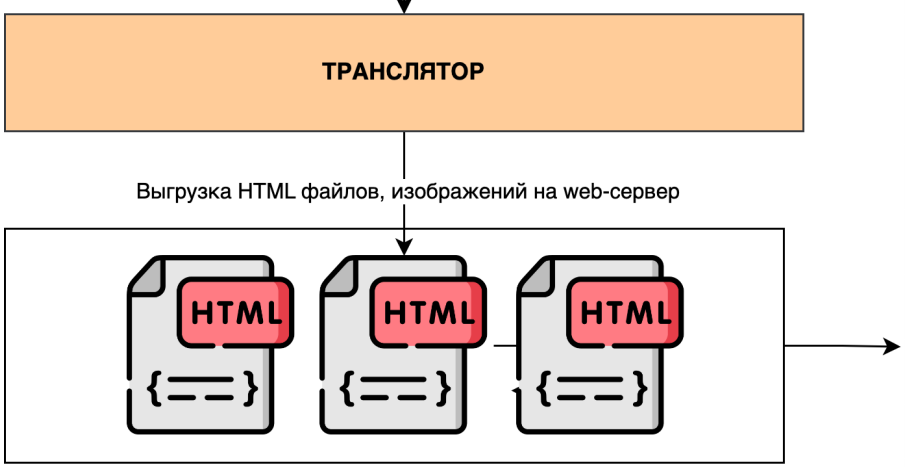
- Асинхронные (”оффлайн”)
- Синхронные (”онлайн”)
- Гибридные

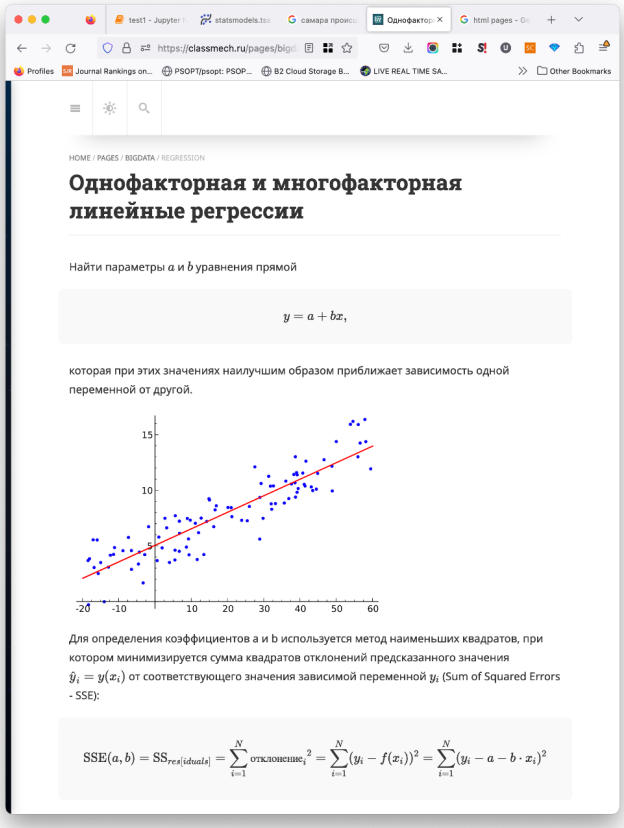
Оффлайн WCMS

- Этот тип WCMS **обрабатывает содержимое перед его публикацией на сервере**.
- Автономные системы обработки позволяют пользователям работать с контентом, когда они не подключены к Интернету.
- Контент, который пользователь загружает в CMS, не публикуется до тех пор, пока автор контента не согласится на его публикацию.
- Примеры: SeaMonkey Composer, статические генераторы сайтов Jekyll, Hugo, Gatsby.

Статические генераторы сайтов

```
---
layout: page
title: Однофакторная и многофакторная линейные регрессии
published: true
---

Найти параметры $$ a $$ и $$ b $$ уравнения прямой

$$
y = a + b x,
$$

которая при этих значениях наилучшим образом приближает зависимость
одной переменной от другой.

![](https://upload.wikimedia.org/wikipedia/commons/3/3a/
Linear_regression.svg)

Для определения коэффициентов a и b используется метод наименьших
квадратов, при котором минимизируется сумма квадратов отклонений
предсказанного значения $$ \hat y_i = y(x_i) $$ от соответствующего
значения зависимой переменной $$ y_i $$ (Sum of Squared Errors – SSE):

$$
\text{SSE}(a,b)=\text{SS}_{\text{res[iduals]}}=\sum_{i=1}^N{\text{отклонение}
_i}^2=\sum_{i=1}^N(y_i-f(x_i))^2=\sum_{i=1}^N(y_i-a-b\cdot x_i)^2
$$
```





HOME / PAGES / BIGDATA / REGRESSION

Однофакторная и многофакторная линейные регрессии

Найти параметры a и b уравнения прямой

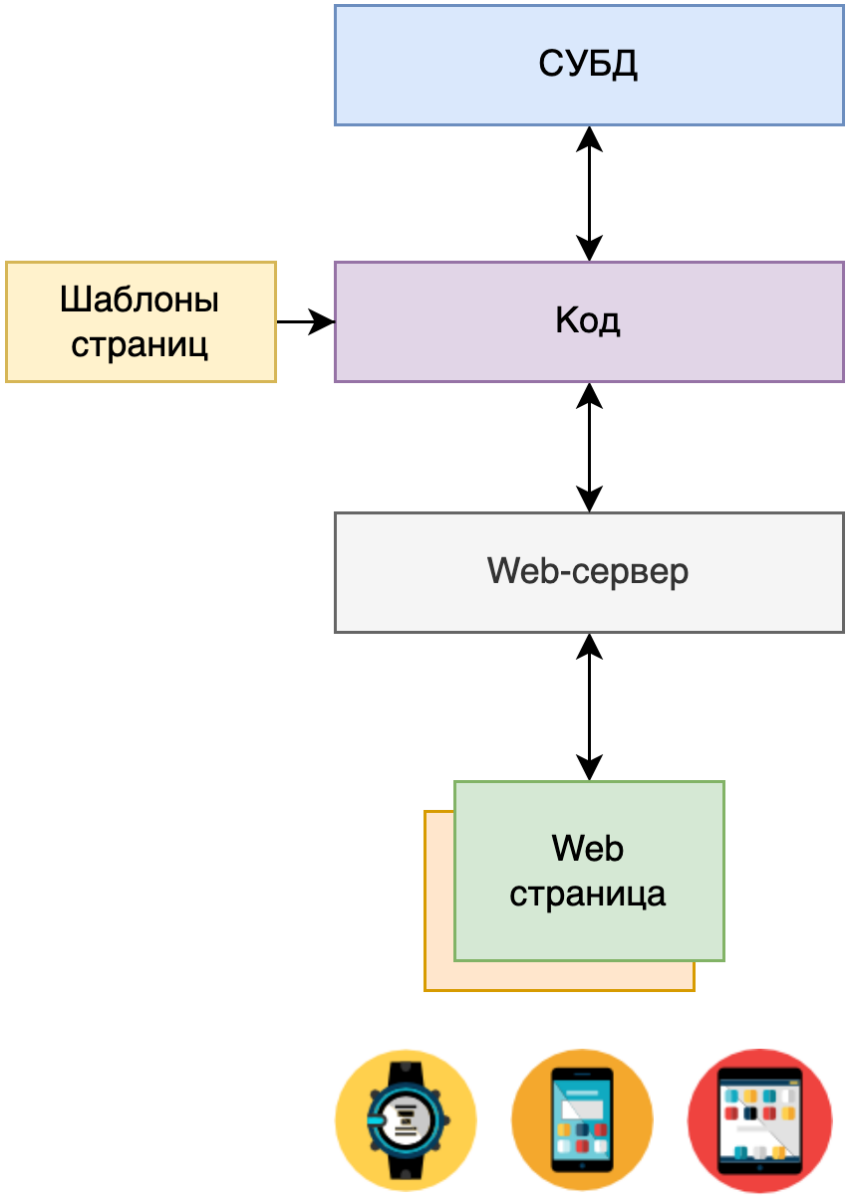
$$y = a + bx,$$

которая при этих значениях наилучшим образом приближает зависимость одной переменной от другой.

Для определения коэффициентов a и b используется метод наименьших квадратов, при котором минимизируется сумма квадратов отклонений предсказанного значения $\hat{y}_i = y(x_i)$ от соответствующего значения зависимой переменной y_i (Sum of Squared Errors - SSE):

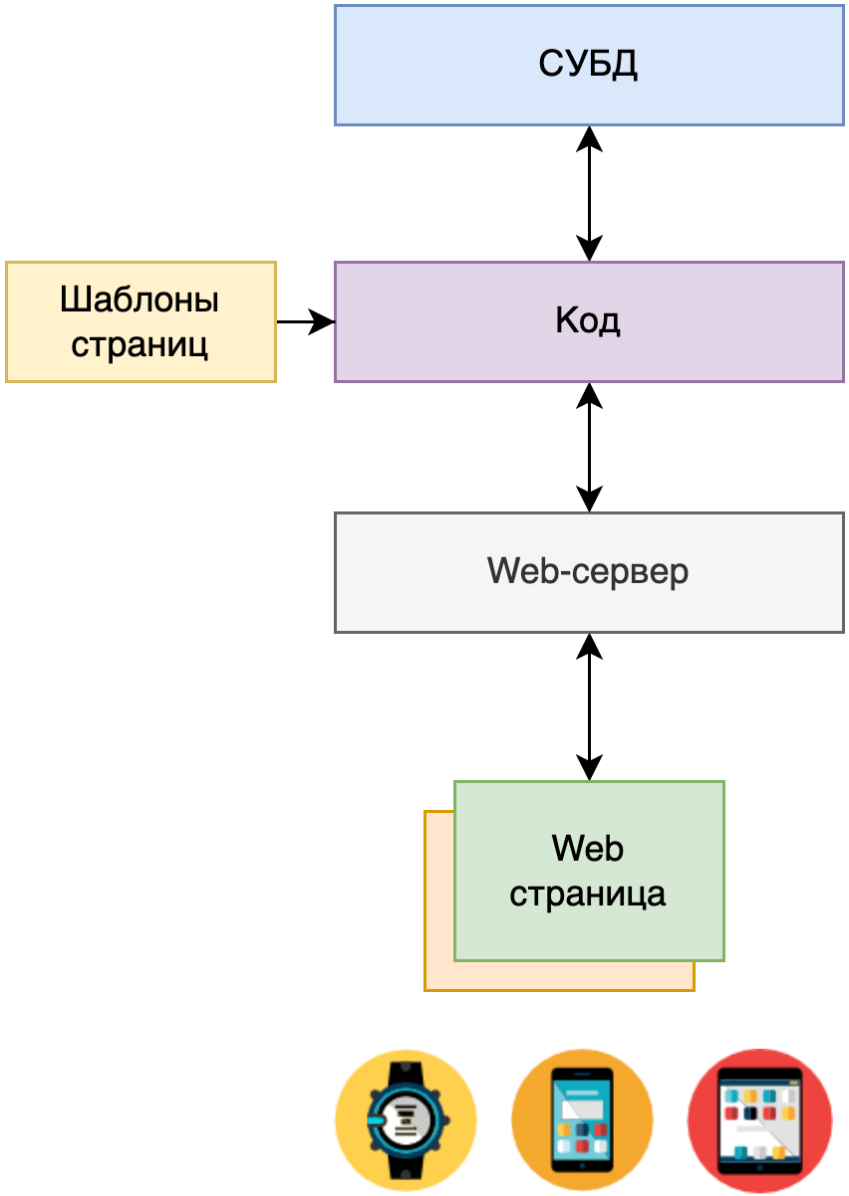
$$\text{SSE}(a,b) = \text{SS}_{res[iduals]} = \sum_{i=1}^{N} \text{отклонение}_i{}^2 = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - a - b \cdot x_i)^2$$

WCMS онлайн обработки



- Системы онлайн-обработки используют шаблоны по запросу и всякий раз, когда пользователь добавляет контент на веб-страницу для публикации. Всякий раз, когда пользователь входит в свою CMS через веб-браузер и получает доступ к веб-странице, генерируется HTML.
- Примеры: **Joomla, Drupal**.

WCMS онлайн обработки



- В автономной WCMS, контент предварительно обрабатывается (применяются шаблоны)
- Онлайн WCMS обрабатывает шаблоны только тогда, когда пользователь запрашивает страницу.

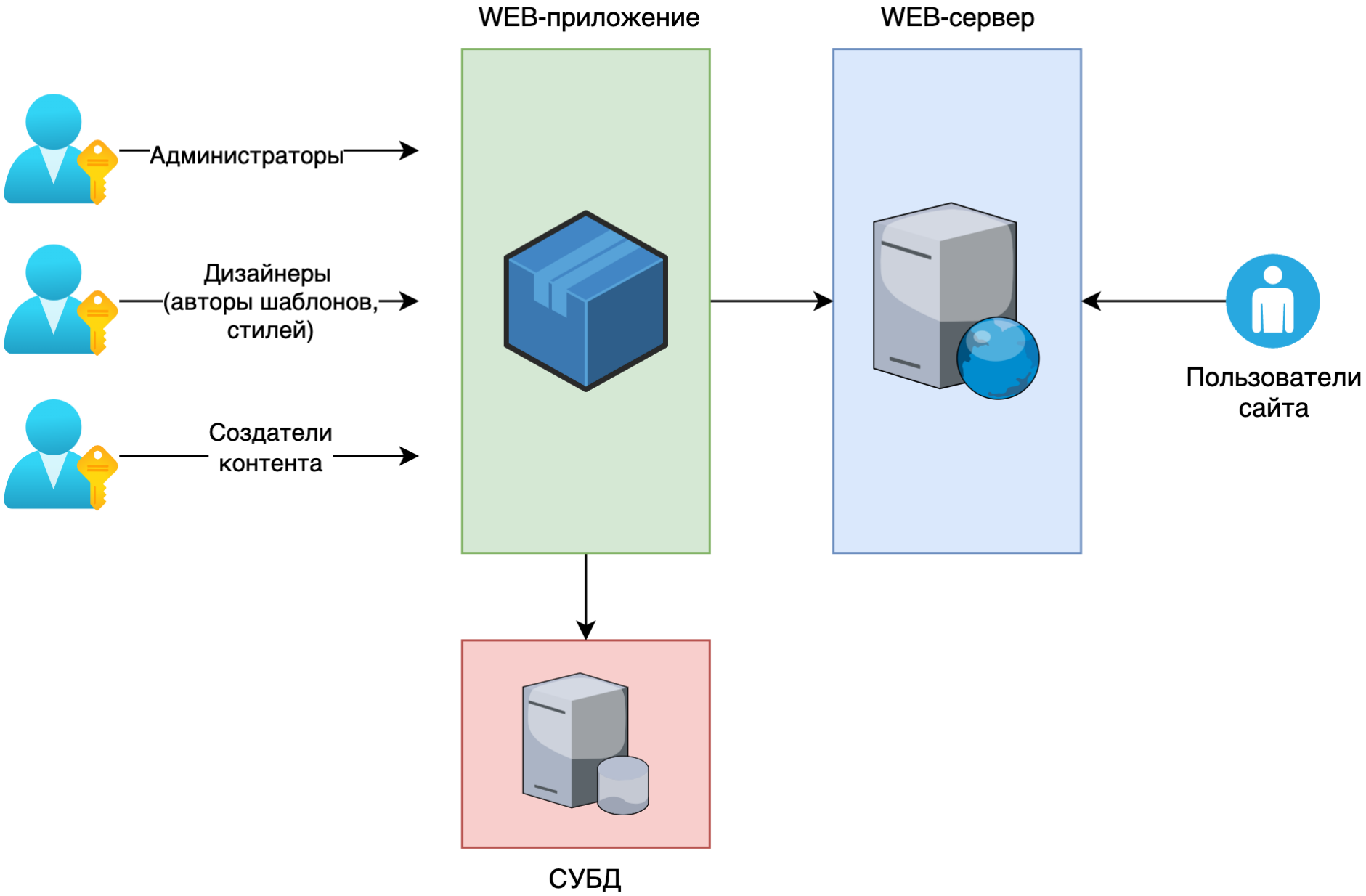
Гибридные системы

- Гибридные системы используют комбинацию автономной и онлайн-обработки.
- Кэширование: модуль представления генерирует страницу один раз, в дальнейшем она в несколько раз быстрее подгружается из кэша.
- Сохранение определённых информационных блоков на этапе редактирования сайта и сборка страницы из этих блоков при запросе соответствующей страницы пользователем.

Возможности WCMS

- Автоматизированные шаблоны.
- Контроль доступа.
- Масштабируемое расширение.
- WYSIWYG – простое редактирование.
- Масштабируемые наборы функций.
- Регулярное обновление.
- Совместная работа.
- Управление рабочим процессом.
- Мультиязычность.
- Различные формы представления контента (HTML, RSS).

Пользователи WCMS



WordPress

- Первый выпуск - 2003 год
- Версия 6.1 - 2022 год
- Веб-сервер: Apache
- PHP
- СУБД: MySQL

Drupal

- Первый выпуск - 2001 год
- Версия 10.0.8 - 2023 год
- Веб-сервер: Apache, Nginx, Lighttpd, ...
- PHP
- СУБД: MySQL, PostgreSQL

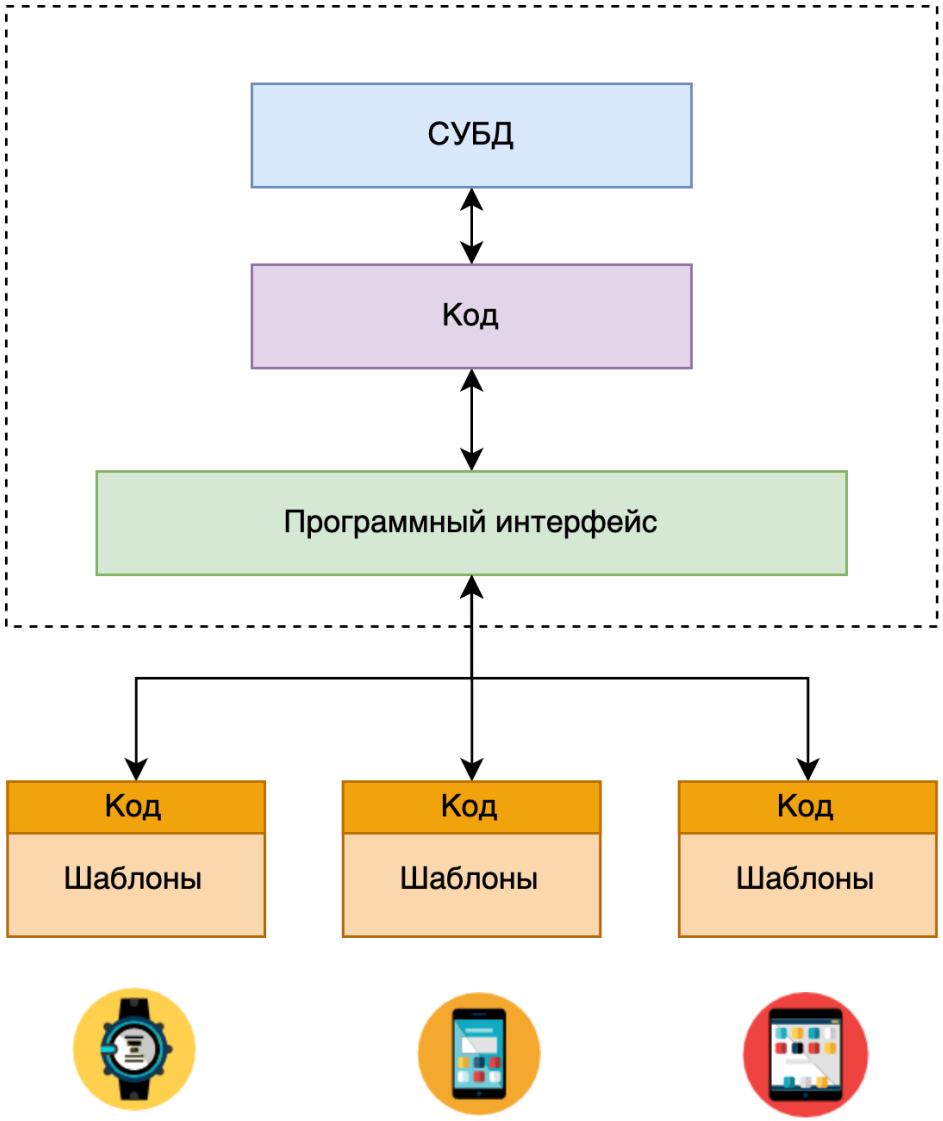
Joomla

- Первый выпуск - 2005 год
- Версия 4.2.9 - 2023 год
- Веб-сервер: Apache
- PHP
- СУБД: MySQL

Headless CMS

- Появление мобильный устройств потребовало перестройки архитектуры WCMS.
- Существующие монолитные CMS не были приспособлена для доставки контента на различные типы устройств, что приводило к необходимости создания различных версий веб-сайтов (обычно урезанных) для мобильных пользователей.
- Появление новых типов устройств с поддержкой Web - смарт-часов, игровых консолей и голосовых помощников только усугубило эту проблему.

Headless CMS



- Система управления контентом, которая изначально проектируется без фронтенда, а только с API (программными интерфейсами для взаимодействия с внешними приложениями и сервисами).
- Headless CMS отделяет «бэкенд», в котором хранится весь контент, базы данных и файлы, от «фронтенда».

Управление контентом предприятия

Информационные ресурсы предприятия

Информационные ресурсы предприятия – это весь объем информации, имеющейся в организации, зафиксированной на материальных носителях и предназначенной для обеспечения внешнеэкономической деятельности и внутренних процессов на предприятии.

Информационный ресурс

Информационный ресурс – информация, обнаруженная, зарегистрированная, оцененная, зафиксированная на материальных носителях, для использования в практической деятельности.

Особенности информационных ресурсов

1. Неисчерпаемость (запас растет с развитем общества)
2. Ценность информационных ресурсов проявляется в соединении с опытом, квалификацией, техникой, энергией
3. Эффективность применения связана с эффектом повторного производства знаний
4. Информацмонный ресурс возникает в результате творческой деятельности
5. Превращение знаний в информационный ресурс определяется возможностями кодирования, распределения и передачи – коммуникационными возможностями.

Классификация информационных ресурсов

1. Специфика
2. Сфера использования
3. Принадлежность
4. Способ доступа
5. Вид носителя
6. Формат представления
7. Способ организации и хранения

Рынок информационных продуктов

- Усиление роли информационных ресурсов в развитии современного общества, возможность их представления в электронном виде с использованием различных форматов и автоматизированной обработки привели к появлению развитого рынка **информационных продуктов** и услуг.
- **Рынок информационных продуктов** – совокупность экономических, правовых и информационных отношений по продаже и покупке информационных ресурсов между поставщиками и потребителями.
- **Информационные продукты** – это информация, полученная в результате преобразования информационных ресурсов, которая может рассматриваться как предмет купли-продажи, хотя она и не является материальным объектом.

Секторы рынка информационных продуктов

- Профессиональный сектор (деловая информация, научная информация)
- Массовая и потребительская информации
- Услуги образования
- Обеспечивающие информационные системы и средства

Байрамукова А. С. Рынок информационных продуктов и услуг: особенности формирования, структура // Пространство экономики. 2008. №2-3.

Информационный процесс

- Предприятие можно рассматривать как информационный центр, в котором обрабатывается информация, содержащаяся как во внешнем, так и во внутреннем потоках, т. е. реализуется информационный процесс.
- Информационный процесс – процесс получения, создания, сбора, обработки, накопления, хранения, поиска, распространения, использования информации.

Для чего нужны ИРП

- Выработка целей
- Разработка программ для достижения целей
- Координация действий подразделений
- Совершенствование системы управления

Задачи

- формирование адекватных информационных ресурсов для системы управления предприятием
- оптимизация информационных потоков путем исключения дублирования информации
- ликвидация разрыва между внедрением информационных технологий и техники и состоянием информационных ресурсов (их формирование и использование)

Корпоративный контент

Корпоративный контент – содержание информационных ресурсов предприятия

- структурированный (базы данных, таблицы)
- неструктурированный (текст, видео)
- веб-контент

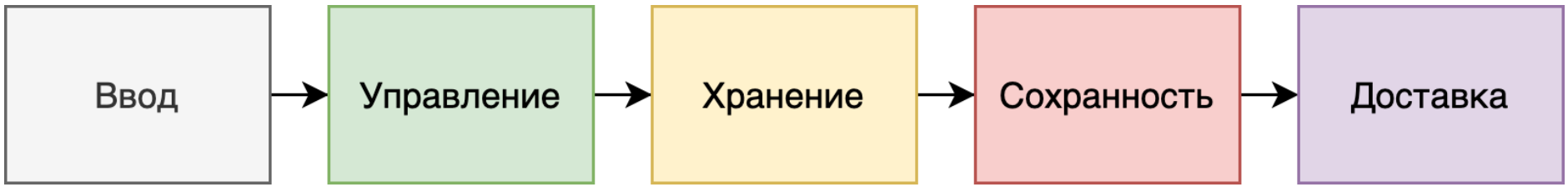
Отличие СЭД от ECM

- В СЭД, где в качестве управляемых данных выступают организационно-распорядительные документы и бизнес-процессы
- ECM системы имеют более гибкий функционал и позволяют работать как со структурированным, так и с неструктурированным контентом.

Основные функции ECM

- Электронный документооборот (СЭД);
- Управление записями (RM) и файлами. Категоризация и упорядочение;
- Управление знаниями (knowledge management). Хранение и предоставление доступа к релевантной для предприятия информации;
- Управление потоками работ (Workflow), автоматизация бизнес-процессов (BPM);
- Управление web-контентом (WCM).

Процессы ECM



Ввод

- Сканирование бумажных документов, ввод электронных писем, мультимедийных объектов, цифровых аудио- и видеозаписей.
- Обработка введенной информации включает процедуры распознавания, категоризации и индексирования информации.

Управление

Управление документами

- хранение документов и метаданных;
- версионность документов;
- разграничение доступа и ведение истории работы с документом;
- контроль целостности документа;
- поиск и навигацию по документам.

Управление

Управление записями

- Управление записями обеспечивает работу с архивами документов длительного хранения, как электронных, так и бумажных.
- Записью или официальным документом называется зафиксированная на материальном носителе идентифицируемая информация, созданная, полученная и сохраняемая организацией или частным лицом в качестве доказательства или подтверждения правовых обязательств либо деловой активности.
- Управляются в соответствии с внешними нормативами (ISO 9000, ГОСТ) или внутренним регламентом организации.

Управление

Управление веб-контентом

- создание/редактирование контента в рамках контролируемого процесса рас- крытия информации;
- автоматическую конвертацию контента в различные форматы представления;
- разграничение прав доступа к информации и выполняемым операциям процесса публикации контента;
- визуализацию данных для представления в Интернет.

Управление

Управление рабочими потоками

Средства автоматизации бизнес-процессов, включая разработку маршрутов, контроль и исполнение:

- инструменты для разработки и отображения рабочего процесса, отображение структур процесса и организации;
- ввод, администрирование, управление версиями, визуализацию и поставку группированной информации со связанными с ней документами или данными;
- средства напоминания, контроля предельных сроков, делегирования задач;
- мониторинг и документирование состояния процесса, маршрутизацию и формирования выхода.

Управление

Поддержка рабочих групп

Средства для обеспечения работы распределенных проектных команд, включая средства интерактивного общения, групповую работу над документами, а также проектно-ориентированные методы взаимодействия:

- средства коммуникаций, включая чаты, программы мгновенного обмена сообщениями, видеоконференции и т. д.;
- совместную обработку информации, включая совместную работу над документами и накопление общей базы информации по проекту;
- средства управления проектами, обеспечивающие планирование, контроль задач и результатов.

Хранение

- Компоненты «Хранение» используются для временного хранения информации, которая не предназначена для архивирования.
- «Хранение» отделено от «Сохранения». Компоненты «Сохранения» ECM обеспечивают долговременное, безопасное хранение и резервное копирование статической, неизменяемой информации.

Доставка

- Компоненты доставки («распространения») ECM используются для представления информации от компонентов «Управления», «Хранения» и «Сохранения».
- Они также содержат функции, используемые для ввода информации в системы (такие, как передача информации на носители или генерация форматированных выходных файлов) или для чтения (например, преобразование или сжатие) информации для компонентов «Хранения» и «Сохранения».

Структура ECMS

- **Репозиторий контента**: это центральное место хранения всего цифрового контента, которым управляет система ECM. Это может быть база данных, файловая система или облачное хранилище.
- **Capture and Ingestion**: этот компонент отвечает за сбор и импорт контента в систему. Он включает в себя такие функции, как сканирование документов, захват электронной почты и возможности массовой загрузки.
- **Управление метаданными**: системы ECM используют метаданные для организации и классификации контента, что упрощает поиск и извлечение. Этот компонент позволяет пользователям определять и управлять полями метаданных для своего контента.

Структура ECMS

- **Поиск и извлечение**. Системы ECM предоставляют мощные возможности поиска, помогающие пользователям быстро находить нужный им контент. Этот компонент включает поисковые фильтры, построители запросов и функции полнотекстового поиска.
- **Управление рабочими процессами и бизнес-процессами**. Этот компонент позволяет организациям автоматизировать бизнес-процессы и рабочие процессы, связанные с их контентом. Он включает в себя такие функции, как назначение задач, утверждения и уведомления.

Структура ECMS

- **Управление записями**: ECM-системы помогают организациям соблюдать правила и управлять своими записями. Этот компонент включает в себя такие функции, как политики хранения, контрольные журналы и юридические удержания.
- **Совместная работа и социальная интеграция**: ECM-системы облегчают совместную работу, позволяя пользователям обмениваться контентом и совместно работать над ним. Этот компонент включает в себя такие функции, как контроль версий, регистрация входа/выхода и комментирование.

Структура ECMS

- **Безопасность и разрешения**. Системы ECM обеспечивают детальный контроль доступа, чтобы гарантировать, что контент доступен только авторизованным пользователям. Этот компонент включает в себя такие функции, как роли пользователей, разрешения и шифрование.
- **Аналитика и отчетность**: системы ECM предоставляют информацию об использовании и производительности контента. Этот компонент включает в себя такие функции, как аналитика использования, журналы аудита и настраиваемые отчеты.

Российские ECM

- ELMA ECM

  `https://www.elma-bpm.ru/product/ecm/`
- ECM-платформа Documino

  `https://www.documino.ru/`
- Docsvision ECM

  `https://docsvision.com/ecm-bpm/docsvision-ecm/`
- ЭЛАР Контекст

  `https://elar-context.ru/`

Зарубежные ECM

- Alfresco (свободная версия - Alfresco Community)
- OpenText (ком)
- Microsoft SharePoint
- IBM FileNet
- Doxis4 iECM

Большие данные

Вопросы

- Что такое большие данные?
- Что такое наука о данных (Data Science) и для чего она нужна?
- Инструменты для работы с большими данными.

Особенности больших данных

Большие данные

Большие данные

Достаточно большие и сложные наборы данных, чтобы их можно было обрабатывать традиционными средствами, используя реляционные системы управления базами данных.

Большие данные

Термин «большие данные» ввёл редактор журнала Nature Клиффорд Линч ещё в 2008 году в спецвыпуске, посвящённом взрывному росту мировых объёмов информации.

Большие данные

Большие данные

достаточно большие и сложные наборы данных, чтобы их можно было обрабатывать традиционными средствами, используя реляционные системы управления базами данных.

Большие данные

Реляционные базы данных

основаны на реляционной модели — табличном способе представления данных. Каждая строка, содержащая в таблице такой БД, представляет собой запись с уникальным идентификатором (ключ). Столбцы таблицы имеют атрибуты данных, а каждая запись обычно содержит значение для каждого атрибута, что дает возможность легко устанавливать взаимосвязь между элементами данных.

В реляционных БД данные структурированы, структура (количество таблиц, столбцов, типов данных) определяется на этапе создания базы данных. Реляционные БД не предназначены для хранения данных, структура которых изменяется (большие данные).

Определение ГОСТ Р ИСО/МЭК 20546-2021

Большие данные

Большие массивы данных, отличающиеся главным образом такими характеристиками, как

- объем,
- разнообразие,
- скорость обработки
- вариативность,

которые требуют использования технологии масштабирования для эффективного хранения, обработки. управления и анализа.

Особенности больших данных

Объем (Volume)

Количественная характеристика данных, влияющая на выбор ресурсов для вычислений и хранения, а также на управление данными в процессе обработки.

Разнообразие (Variety)

Диапазон форматов, логических моделей, временных шкал и семантики массива данных

Особенности больших данных

Скорость обработки данных (Velocity)

Скорость потока, с которой данные создаются, передаются, сохраняются, анализируются или визуализируются.

Вариативность (Variability)

Изменения в скорости передачи, формате или структуре, семантике или качестве массива данных

Источники больших данных

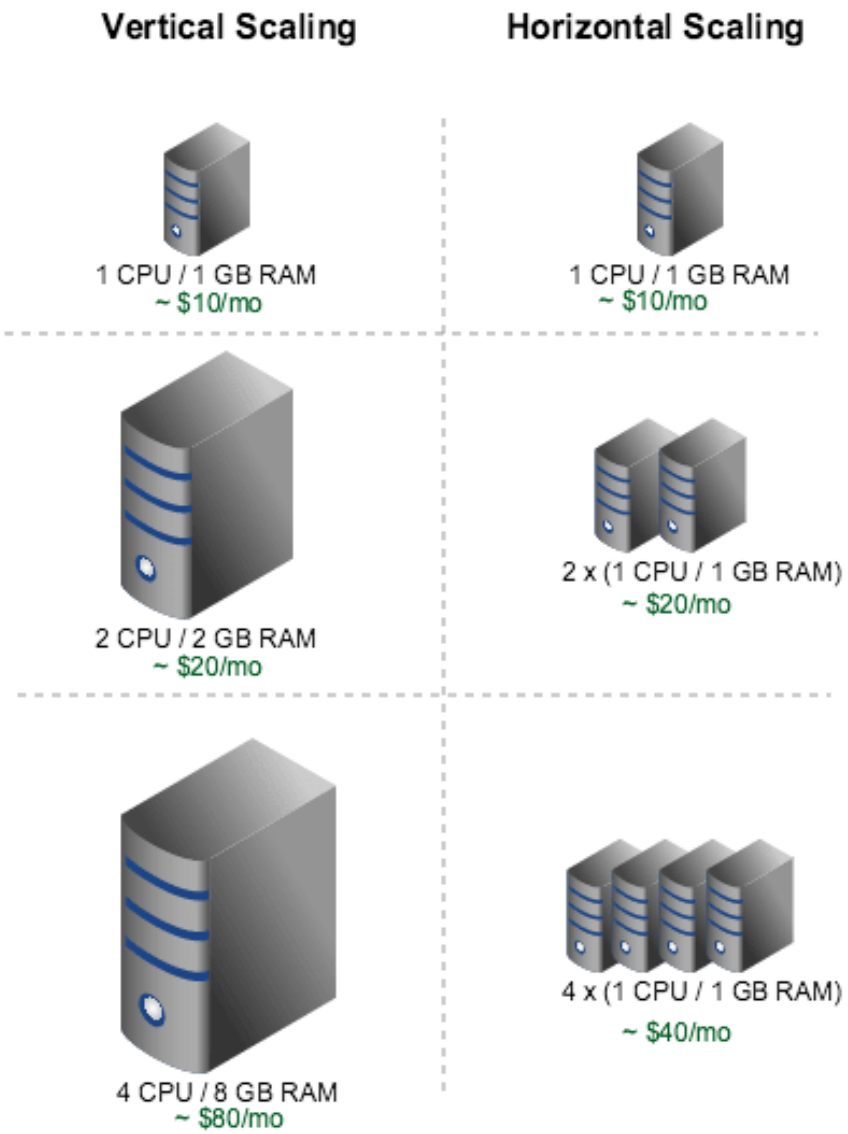
- Интернет вещей
- Соцсети, блоги, СМИ
- Показания приборов
- Статистика (города, государства)
- Медицинские данные

Принцип работы с большими данными

Особенности **BIG DATA** (Volume, Velocity, Variety, Variability) определяют принципы работы с большими данными

- Горизонтальная масштабируемость
- Отказоустойчивость
- Локальность данных

Горизонтальная масштабируемость



- Объем данных постоянно и стремительно растет и информации может быть сколь угодно много.
- Система, которая подразумевает обработку этих данных, должна быть расширяемой.
- Горизонтальное масштабирование (увеличение количества простых серверов) более выгодно чем вертикальное масштабирование.

Отказоустойчивость

- Машин в кластере может быть много (в компании Yahoo кластер насчитывает более 40000 машин).
- Методы работы с большими данными должны учитывать вероятность сбоев и поддерживать работоспособность системы без значимых последствий.

Локальность данных

- В крупных распределённых системах, используемые данные хранятся на большом количестве машин.
- Для снижения затрат ресурсов на передачу данных, данные должны хранится и обрабатываться на одной и той же машине.

Наука о данных

Анализ

- Большие данные мало просто собрать — их нужно как-то использовать, например, чтобы строить прогнозы развития бизнеса или проверять маркетинговые гипотезы. А для использования данные требуется структурировать и анализировать.

Этапы работы с данными

1. чистка данных (data cleaning)
   поиск и исправление ошибок в первичном наборе информации, например, ошибки ручного ввода (опечатки), некорректные значения с измерительных приборов из-за кратковременных сбоев и т.д.;
2. генерация предикторов (feature engineering)
   переменных для построения аналитических моделей, например, образование, стаж работы, пол и возраст потенциального заемщика;
3. построение и обучение аналитической модели (model selection) для предсказания целевой (таргетной) переменной. Так проверяются гипотезы о зависимости таргетной переменной от предикторов.

Data Science (Wiki)

Наука о данных –

- раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.

Наука о данных объединяет:

- методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма,
- статистические методы,
- методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными,
- методы проектирования и разработки баз данных.

Data Science

Наука о данных (data science)

Наука о данных это расширение статистики, способное справляться с огромными объемами данных, производимыми в наши дни.

Наука о данных (data science) ГОСТ

Извлечение практических знаний из данных посредством исследования или создания и проверки гипотез.

Примеры использования

Предиктивная аналитика

- Сейчас на производстве часто внедряют IoT-системы: устанавливают датчики на оборудовании и в помещениях, а потом анализируют собранные ими данные.
- Эти данные и есть big data, их можно использовать для мониторинга состояния оборудования, моделирования производственных процессов, выявления и предотвращения сбоев.

Поиск новых месторождений

- При добыче природных ресурсов месторождения часто приходится искать почти вслепую. Однако с помощью анализа больших данных можно обнаруживать закономерности, изучать состояние почв, наличие подземных пустот, температуру пород — и таким образом эффективно искать перспективные месторождения, сравнивая новые участки с уже известными аналогами.

Планирование грузоперевозок

- В логистике на перевозку товаров влияет много разных факторов: загрузка складов, пробки на дорогах, состояние парка машин, расположение автозаправок. Если собрать все эти факторы вместе, сопоставить их и проанализировать, можно эффективнее планировать маршруты и время доставки, чтобы избежать простоев транспорта.
- Компания ПЭК запустила Центр управления перевозками на базе big data. Это помогло им прогнозировать загрузку 189 складов по всей России на месяц вперед и планировать маршруты грузового транспорта.

Повышение продаж

- Информация о поведении клиентов в магазине или на сайте — это большие данные. На их основе можно предполагать, что именно люди будут покупать, и использовать это для повышения продаж.
- Amazon использует большие данные для системы рекомендаций товаров. Их система основана на машинном обучении — она учитывает поведение других покупателей, ваши предыдущие покупки, время года и десятки других факторов. В итоге 35% всех продаж в Amazon генерируют рекомендации, а 86% пользователей сервиса утверждают, что рекомендации влияют на их решения о покупке.

Оценка платежеспособности

- Оценка платежеспособности. Банкам важно выдавать кредиты только тем, кто точно сможет их вернуть, чтобы не понести убытки. Анализ больших данных помогает анализировать платежеспособность клиентов и оценивать риски.

Медицина

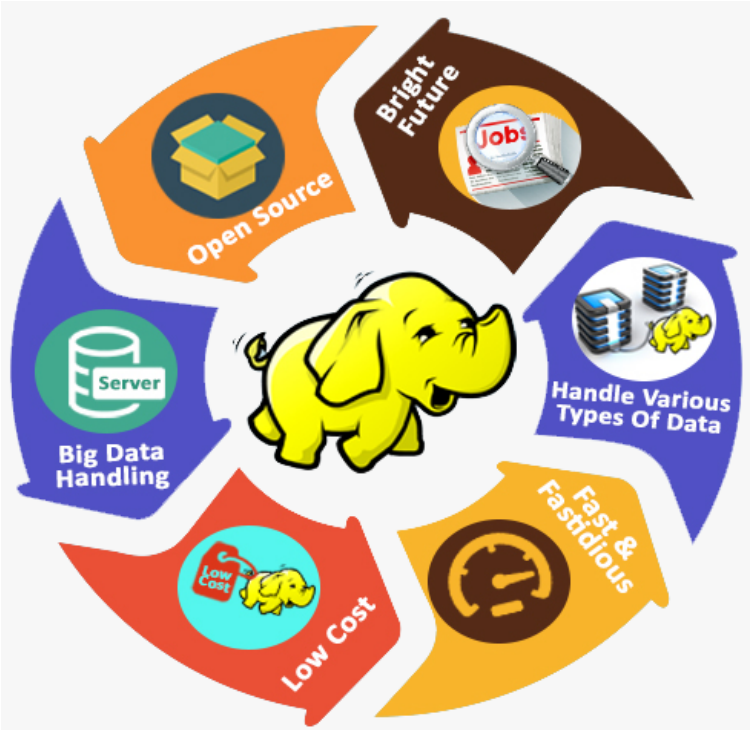
- В медицинской сфере большие данные в перспективе можно использовать для диагностики и лечения, большинство интересных проектов пока находятся на стадии разработки или тестирования, но есть и уже реализованные.

Инструменты

MapReduce

Модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений над очень большими наборами данных в компьютерных кластерах.

Hadoop



- Свободно распространяемый набор утилит, библиотек для разработки и выполнения распределённых программ, работающих на кластерах.
- Разработан на Java в рамках вычислительной парадигмы MapReduce: приложение разделяется на большое количество одинаковых элементарных заданий, выполнимых на узлах кластера.

Hadoop

- HDFS – это распределенная файловая система, предназначенная для работы на стандартном оборудовании.
- MapReduce – модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений.
- YARN – технология, предназначенная для управления кластерами.
- Библиотеки – для работы остальных модулей с HDFS

Cassandra

Распределенная свободная БД Cassandra, предназначена для управления большими объемами данных, раскиданных по серверам. Распространяется бесплатно.

- быстрая обработка огромных объемов данных;
- линейная масштабируемость;
- доступ из облака;
- отсутствие единой точки отказа;
- автоматическая репликация;
- распределение данных между дата-центрами.

Kubernetes

Открытое программное обеспечение для управления контейнеризированными приложениями – автоматизации их развёртывания, масштабирования и координации в условиях кластера.

Spark

Фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop.

Python & DataScience

- Python — это один из самых распространённых языков программирования.
- Существует огромное количество пакетов, которые позволяют решать с помощью этого языка самые разные задачи.
- В наше время весьма востребованы наука о данных (Data Science, DS) и машинное обучение (Machine Learning, ML). И там и там Python показывает себя наилучшим образом.

Библиотеки Python

- **NumPy**
  одномерные и многомерные массивы
- **SciPy**
  численные методы
- **Pandas**
  DataFrame, предназначенный для работы с индексированными массивами.
- **StatsModels**
  Анализ данных, создание статистических моделей, проведение статистических исследований.

Библиотеки Python

- **Matplotlib, Seaborn**
  Визуализация данных
- **Plotly**
  Интерактивные графики, позволяющие исследовать взаимоотношения переменных.
- **Bokeh**
  Графики для web-приложений
- **Scikit-Learn, Keras**
  Пакеты для машинного обучения.

Список использованных источников

- Бараксанов Д. Н. Управление ИТ-сервисами и контентом: учебное пособие / Д. Н. Бараксанов, Ю. П. Ехлаков. — Томск : ФДО, ТУСУР, 2015. — 144 с.
- Топ 10: ECM системы

  `https://www.doc-online.ru/tools/ecm/`
- СЭД (рынок России)

  `https://www.tadviser.ru/index.php/Статья:СЭД_(рынок_России)`