**Министерство науки и высшего образования Российской Федерации**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»**

**(Университет ИТМО)**

Факультет **Инфокоммуникационных технологий**

Образовательная программа **Интеллектуальные системы в гуманитарной сфере**

Направление подготовки **45.03.04 Интеллектуальные системы в гуманитарной сфере**

**К У Р С О В О Й П Р О Е К Т**

**по дисциплине «Инфокоммуникационные системы и технологии»**

на тему: «Система агрегации медицинских услуг»

Обучающийся: Тихонова Марина Сергеевна, группа К3144

**Работа сдана**

Дата 20.12.2023

Санкт-Петербург 2023

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_heading=h.30j0zll)

[1 Содержание проекта 5](#_heading=h.1fob9te)

[2 Процессы работы над проектом 6](#_heading=h.tyjcwt)

[3 Задачи, которые стояли передо мной 8](#_heading=h.3dy6vkm)

[4 Результаты работы и анализ 13](#_heading=h.17dp8vu)

[5 Оценка работы 15](#_heading=h.35nkun2)

[ЗАКЛЮЧЕНИЕ 16](#_heading=h.1ksv4uv)

**ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ**

В данном отчете применяют следующие сокращения и обозначения:

1. Парсер – это программа для сбора и систематизации информации, размещенной на различных сайтах.
2. База данных – это упорядоченный набор структурированной информации или данных, которые обычно хранятся в электронном виде в компьютерной системе. База данных обычно управляется системой управления базами данных (СУБД). Данные вместе с СУБД, а также приложения, которые с ними связаны, называются системой баз данных, или, для краткости, просто базой данных.
3. ETL (Extract, Transform, Load) – процесс преобразования сырых данных в обработанные. Включает в себя выгрузку данных из STG слоя, обработку, загрузку обработанных данных в DDS слой.
4. STG (Staging) – слой базы данных, в котором хранятся сырые данные.
5. DDS (Data Detail Store) – слой базы данных, в котором хранятся данные, прошедшую ETL обработку.
6. Docker – это программная платформа для быстрой разработки, тестирования и развертывания приложений.

# ВВЕДЕНИЕ

Во время выполнения моей курсовой работы я приняла непосредственное участие в проекте «Система агрегации медицинских услуг»под руководством Гарифуллина Марата.

Здоровье – очень важная сфера жизни любого человека, но жителям Петербурга в связи с климатическими условиями региона и уровнями загрязненности города на нее приходится обращать внимание чаще многих других. Для поддержания здоровья в совокупности с грамотным планированием личных финансов часто оказывается необходимым просмотр множества различных организаций, предлагающих одну и ту же услугу. Проект, над которым работала наша команда, существенно упрощает данный процесс.

Основная цель проекта заключалась в реализации интерактивного дашборда, с помощью которого можно просматривать историю цен, самое выгодное предложение искомой услуги, а также адреса и контактные данные представленных клиник Санкт-Петербурга. Данный результат может далее использоваться как подсистема в продукте заказчика.

# 1 Содержание проекта

Проект “Система агрегации медицинских услуг” заключается в разработке полного ETL-процесса, итогом работы которого станет интерактивный дашборд, соответствующий ряду требований, которые были перечислены выше.

Задачи проекта включают:

1. Установка необходимых инструментов и компонентов на персональные ЭВМ
2. Разработка архитектуры базы данных
3. Написание парсеров и интеграция полученных данных в разработанную базу данных
4. Написание ETL скрипта, который будет обрабатывать данные из STG-слоя и записывать их в DDS
5. Развертывание всех необходимых компонентов из предыдущих шагов в Docker контейнере
6. Передача данных в Yandex DataLens и их визуализация в виде интерактивного дашборда

В результате работы над перечисленными шагами был создан агрегатор медицинских услуг, собирающий данные с сайтов клиник Петербурга и представляющий их в доступном формате на дашборде. Его легко развернуть с помощью Docker, его процессы автоматизированы, а результат работы очень удобен в использовании конечным пользователем.

# 2 Процессы работы над проектом

Перед тем, как приступить к заданиям технического характера, вместе с нашим руководителем был составлен следующий план действий:

1. Установка необходимых инструментов и компонентов на персональные ЭВМ:

Данный этап включает в себя установку на персональные компьютеры Python, а также среду программирования, которой пользовалась команда, - PyCharm, их настройка.

1. Проектирование базы данных:

База данных реализуется через два слоя: STG (исходные данные, полученные от парсеров) и DDS (обработанные скриптом данные, используемые для непосредственной визуализации). Каждый из указанных слоев должен содержать две таблицы: **analyzes** (данные о медицинских услугах)и **addresses** (данные о адресах и контактных данных клиник).

Таблица **analyzes** обязательно включает в себя следующие столбцы:

* Код
* Группа
* Наименование
* Стоимость услуги
* Наименование лаборатории
* Дата

Таблица **addresses** состоит их следующих столбцов:

* Город
* Адрес
* Контактные номера
* Часы работы
* Станция метро
* Наименование клиники

Помимо этого, необходимо было обеспечить создание слоев базы данных и работу процессов в контейнере.

1. Написание парсеров.

Данный этап подразумевает написание веб-парсеров. Каждому из членов команды был дан адрес сайта одной из исследуемых клиник, для которого он должен был подготовить программу, выполняющую функцию парсера, на языке Python.

1. Написание ETL - скрипта.

Данный этап подразумевает создание скрипта на Python, который чистит, упорядочивает, обрабатывает данные, полученные парсингом, и «укладывает» их в соответствующий слой базы данных.

1. Создание основного файла запуска всех написанных частей сбора и обработки данных.
2. Выгрузка данных и визуализация.

Данный этап предполагает выгрузку обработанных данных из DDS слоя базы данных и загрузку их в Yandex DataLens для последующей визуализации в виде интерактивного дашборда.

1. Контейнеризация

Данный этап подразумевает загрузку всех парсеров в Docker контейнер и запуск базы данных и ETL-скрипта в рамках контейнера.

# 3 Задачи, которые стояли передо мной

Передо мной стояли четыре основные задачи: установка и настройка необходимых для работы компонентов на персональную ЭВМ, разработка веб-парсера сайта Lab4u, визуализация данных в Yandex Datalens и подготовка материалов для презентации проекта на онлайн-защите.

В рамках выполнения первой задачи мне не пришлось устанавливать никакие новые компоненты, я лишь проверила их работу и убедилась в верности их установленных версий на своих устройствах. Никаких трудностей или потребности в изучении новых тем на данном этапе работы не выявилось.

Интенсивная и действительно интересная работа началась на второй моей задаче – написании парсера для сайта Lab4u. Мной было внимательно изучено устройство исходного кода сайта, над которым мне предстояло работать, чтобы понять структуру страниц для парсинга. Убедившись в том, что мне понятен принцип организации данных внутри html-разметки необходимых страниц, я приступила к написанию кода на Python с использованием ранее мною изученных библиотек requests и BeautifulSoup.  
В процессе написания парсера я столкнулась с трудностью в виде динамической страницы, содержимое которой генерировалось кодом на JavaScript, что делало изначальный подход неприменимым при обработке данных оттуда. Для решения данной проблемы я изучила и использовала библиотеку selenium, с помощью которой я написала функцию (см. рисунок 1), результатом работы которой является html-разметка необходимого состояния страницы. Написанная Python программа собирает все необходимые данные по клинике и создает две таблицы (услуги и адреса). После проверки и одобрения написанного парсера руководителем проекта и небольшого ожидания завершения работы участников, занимающихся БД и ETL, я приступила к следующему этапу работы над проектом – визуализации данных.

Изображение выглядит как текст, снимок экрана, документ, Шрифт

Автоматически созданное описание

Рисунок 1 – функция обработки динамической страницы

Визуализация полученных данных создавалась в формате интерактивного дашборда внутри облачного сервиса Yandex Datalens. Прежде всего, я трансформировала таблицы услуг и адресов в распознаваемые сервисом объекты – датасеты. Далее, на основании двух получившихся датасетов были созданы чарты, отображающие различные аспекты собранных данных. Среди них были следующие: график истории цен (минимальная + средняя), индикатор минимальной цены по услуге, индикатор клиники, ее предоставляющей, таблица адресов и контактных данных клиник. Для интуитивного и удобного расположения всей необходимой информации по медицинским услугам исследуемых клиник Петербурга был разработан макет расположения основных блоков данных дашборда, представленный на рисунке 2. После начала работы я несколько модифицировала разметку дашборда, разделив его на 2 вкладки по характеру представленной информации (цены и контакты), а также добавив в раздел контактов поле со ссылками на сайты всех представленных клиник. В качестве интерактивной составляющей я добавила селекторы множественного выбора в обе вкладки дашборда. Для вкладки услуг этими селекторами стали выбор конкретной услуги, выбор категории услуг и выбор клиник. Для вкладки с адресами были добавлены селекторы по ближайшей станции метро к клинике и по наименованию клиники. Таким образом, мной был создан оформленный интерактивный дашборд со всем функционалом из требований к проекту (см. рисунок 3).

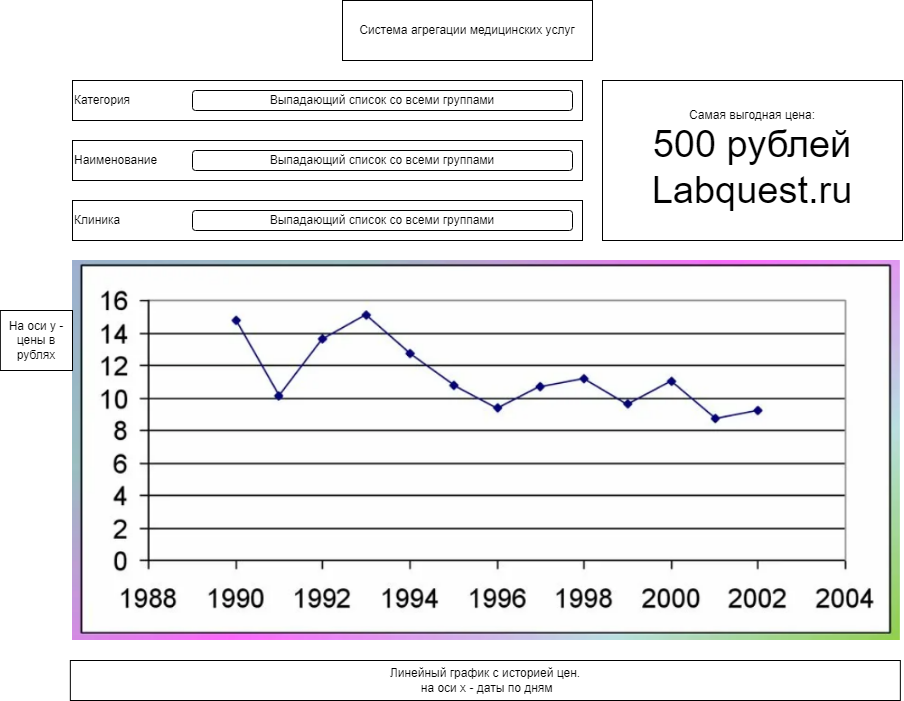


Рисунок 2 – предварительный макет дашборда

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Рисунок 3 – скриншот дашборда на первой вкладке

Под конец работы над проектом на мои плечи легла задача подготовки материалов к выступлению команды на онлайн-защите проекта. Я обсудила содержание и структуру будущей презентации с руководителем, получила исходник презентации-рекламы проекта и шаблон с символикой университета от него. После этого я составила смысловой скелет презентации, нашла визуальные элементы для сопровождения каждой из освещенных тем, поработала над эстетической складностью и простотой восприятия информации зрителем, побыв в роли графического дизайнера. Следующим шагом выполнения данной задачи стало индивидуальное общение с каждым из участников проекта для определения содержания слайдов, закрепленных за определёнными личностями. Завершающим этапом подготовки материалов к выступлению стало создание документа в общем доступе, предназначенного для написания нашей речи. На каждый слайд был выделен блок текста с подписанным оратором, который должен был прописать свои слова, укладываясь в заранее обсужденный тайминг. Итогом моей работы над данной задачей стали презентация проекта (см. рисунок 4) и текст выступления к ней.

Изображение выглядит как текст, снимок экрана, программное обеспечение, веб-страница

Автоматически созданное описание

Рисунок 4 – один из слайдов презентации

# 4 Результаты работы и анализ

Среди результатов проделанной работы можно выделить следующее: созданный на GitHub репозиторий, содержащий исходные py файлы с парсерами каждой из клиник, а также примеры таблиц, генерируемых ими, база данных, дашборд. Проект запускается внутри Dockerконтейнера, что упрощает работу с ним и процесс развертывания. Созданная база данных легко восстанавливается в случае непредвиденных ситуаций. Визуализация данных в виде интерактивного дашборда наглядно и интуитивно демонстрирует всю необходимую информацию и помогает в выборе организации для осуществления медицинских услуг.

В процессе работы над поставленными передо мной задачами мне удалось впервые применить на практике свои знания по Python-библиотекам requests и BeautifulSoup. Написание парсеров также заставило меня познакомиться с принципом работы динамических веб-страниц, изучить и успешно применить библиотеку selenium для эффективной обработки данных с одной из них. Помимо этого, я вновь попрактиковалась в работе с Git и GitHub, научилась просить о помощи и рассказывать о трудностях или неудачах, когда это представляется необходимым. На этапе визуализации я впервые попробовала работать в сервисе Yandex Datalens. Данный сервис был выбран для повышения скорости разработки, а также для упрощения процесса входа в визуализацию для участников, не знакомых с данной областью. На собственном опыте я убедилась в преимуществах тех инструментов, которыми пользовалась ранее при выполнении подобных задач. Я повысила свою способность эффективно и быстро искать нужную для решения новых для меня проблем информацию в интернете и среди знакомых и друзей, разбирающихся в конкретной области.

Если смотреть на личные итоги более глобально, то все мы (участники проекта) приобрели ценнейший опыт командной работы, развили навыки эффективной коммуникации, познакомились с устройством ETL-процессом и СУБД, научились планировать время и разбивать большие задачи на более маленькие подзадачи. Уверена, все ошибки и победы, с которыми мы сталкивались в процессе разработки, помогут нам стать более компетентными специалистами в своих областях в будущем. Нам всем удалось в некоторой мере сблизиться в процессе совместного выполнения задачи и подготовки к выступлению. Мы не только стали лучше как специалисты, но и обзавелись новыми приятными знакомыми и воспоминаниями.

Что касается зон роста и планов на сам проект, нашей команде хотелось бы перенести визуализацию данных на другой BI-сервис (к примеру, Tableau или Power BI), что позволило бы получить доступ к большей гибкости в визуальной составляющей дашборда, а также разнообразить имеющиеся чарты более удобными и выгодными для пользовательского восприятия чартами. Помимо этого, мы обсуждали возможность автоматизации всех процессов, входящих в итоговый проект с помощью AirFlow.

# 5 Оценка работы

Нашей команде удалось в полной мере реализовать все этапы и функционал проекта, над которым мы работали. Каждый из нас внес существенный вклад в данный процесс, поэтому такой успех можно приписать нашей хорошей командной работе. Мы не стеснялись задавать друг другу вопросы, давать советы, собираться с помощью различных средств коммуникаций (Discord, Telegram и другие) и обсуждать проделанную работу и планы на будущее. Постоянный контакт между участниками сыграл огромную роль в успешном и своевременном продвижении по этапам проекта.

Также стоит выразить благодарность нашему руководителю проекта – Гарифуллину Марату Рустамовичу. Он поддерживал уютную и спокойную атмосферу в команде, помогал нам в планировании и реализации всех этапов проекта. С ним было крайне легко связаться, Марат отвечал на все заданные ему вопросы и помогал в случае затруднений в какой-либо из задач одного из членов команды. Он организовывал созвоны, разработал грамотное техническое задание, вовремя добавлял нам задачи с подробным описанием требований к реализации каждого из них.

# ЗАКЛЮЧЕНИЕ

В рамках данной работы был разработан агрегатор медицинских услуг среди медицинских центров и лабораторий города Санкт-Петербург. Помимо этого, агрегатор обладает возможностью удобной для конечного пользователя визуализации данных с возможностью просмотра истории цен и контактных данных организаций, поиска самого выгодного предложения из ныне имеющихся. Проект является готовым продуктом, соответствующим требованиям из изначального технического задания. Тем не менее, наша команда видит большой потенциал для развития и роста проекта.

Все участники проекта успешно выполнили поставленные перед ними задачи и получили бесценный опыт разработки, командной работы и развили свои навыки во многих сферах. К примеру, теперь каждый из нас может сказать, что имеет опыт написания веб-парсеров на языке Python.

Работа была интересной и насыщенной для каждого из членов команды. Мы справились с поставленной целью и многому научились.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Максим Ожегов Визуализация данных: применение в работе, основные принципы, способы и инструменты для использования. URL: <https://practicum.yandex.ru/blog/vizualizaciya-dannyh/> (дата обращения: 08.12.2023).
2. Абрамова Татьяна Алфиевна Разработка парсинг-системы для получения скрытых ссылок со страниц социальных сетей // Вестник ПензГУ. 2016. №3 (15). URL: <https://cyberleninka.ru/article/n/razrabotka-parsing-sistemy-dlya-polucheniya-skrytyh-ssylok-so-stranits-sotsialnyh-setey> (дата обращения: 12.11.2023).
3. Алексей Баранцев Что такое Selenium? URL: <https://habr.com/ru/articles/152653/> (дата обращения: 15.11.2023).