# **Приветствие**

Добрый день уважаемая комиссия, меня зовут Семенов Евгений выпускник группы БПЦ 21-01 и сейчас в течении 7 минут вам будет представлена моя выпускная квалификационная работа на тему «Разработка программного обеспечения для анализа рынка строительных материалов».

# **Актуальность**

Строительная отрасль требует **оперативной информации** о стройматериалах, **ручной сбор** занимает много времени и приводит к ошибкам и обработке устаревшей информации из-за человеческого фактора. **Объем** продаж стройматериалов **растет**, как и растет число компаний, которые предлагают свой товар. Стоит отметить так же **стремительный рост цен** на стройматериалы, что делает ручной сбор информации затруднительным.

Исходя из этого выделим ряд проблем.

# **Проблемы**

1. **Данные** о стройматериалах **разрознены:**
   1. Нет единого стандарта
   2. 100+ источников в разных форматах
2. **Ручной сбор неэффективен:**
   1. Десятки часов на мониторинг
   2. Риск устаревших/ошибочных данных
3. **Рынок меняется стремительно:**
   1. Десятки часов на мониторинг
   2. Риск устаревших/ошибочных данных

Результаты таких тенденций для бизнеса плачевны, а это переплаты, простои, потеря позиции на рынке.

# **Цели и задачи**

Поэтому была разработана система анализа рынка строительных материалов, в этом и заключалась основная цель данной работы.

Чтобы цель была достигнута выделим основные задачи:

1. **Проанализировать основные подходы** и инструменты анализа данных
2. Определить **ключевые требования**.
3. Разработать **архитектуру систему**.
4. Реализовать основные **алгоритмы**.
5. Провести **тестирование**.
6. Оценить **эффективность системы**.

# **Сравнительный анализ решений, представленных на рынке**

В данной таблице представлены основные подходы для анализа данных строительных материалов, как видно все они имеют свои **недостатки,** которые мы попытаемся избежать в нашей системе.

# **Возможности системы**

На этапе проектирования были описаны основные возможности системы, которые представлены на данном слайде, в частности это:

1. **Информация о системе –** описание для кого система разработана и какие возможности предоставляет**.**
2. **Регистрация.**
3. **Вход.**
4. **Получение данных** – имеет подфункции загрузка данных(то есть сам парсинг), сохранение данных(сохранение полученых данных после парсинга в базу данных), скачать данные(загрузка данных в виде файла json).
5. **Анализ данных –** имеет подфункции найти данные в базе данных, построить диаграмму, скачать в формате csv.
6. **Руководство пользователя –** подробное руководство о том как пользоваться системой, от парсинга данных до визаулизации.

# **Модель данных**

Для работы системы была определена структуру хранимых данных, она представлена в виде **ER-диаграммы** на данном слайде и включает **2 сущности (Пользователь и Данные**), а также **связь** между ними.

# **Логика работы системы**

Логика работы системы описана в нотации BPMN, и в нее включены основные функции, которые были сформулированы на этапе описания возможностей системы, так же есть отработка различных вариантов поведения системы в зависимости от действий пользователя. То есть это карта маршрута пользователя по системе.

# **Средства разработки системы**

Стоит отметить стек технологий для разработки системы, в частности это:

1. На стороне клиента использована платформа **Node.js** и её фреймворк **React.js** с дополнительными библиотеками.
2. На стороне сервера язык программирования **Python** и его библиотеки, в частности **Scrapy, Pandas, JWT.**
3. За хранение данных была выбрана **СУБД MySQL**
4. Проект был подвергнут контейнеризации при помощи **Docker и Docker Compose**
5. Операционной системой для развертывания была выбрана **Linux/Ubuntu 24.04 LTS.**

Минимальные аппаратные требования для развертывания проекта:

1. 4 ядра CPU, 8+ ГБ RAM, SSD 512ГБ
2. Стабильный интернет (10+ Мбит/с)

# **Интерфейс**

Интерфейс и стиль оформления страниц представлены на данных слайдах, в частности это интерфейс: **стартовой страницы, блоки управления доступом, руководство пользователя, сбор данных, поиск данных, визуализация в виде линейного графика с точками.**

# **Алгоритм получения данных**

Стоит затронуть тему алгоритма получения данных, суть заключается в том, что **система получает входные данные** для сбора, из них **получает домен** стартовой страницы, на ней **собирает все возможные ссылки** и начинает **фильтрацию** их на предмет наличия искомых данных на страницах на которые ведут ссылки, если данные найдены, то **система их сохраняет** во временное хранилище. Вне зависимости от того получила ли система данные со страницы, она **сохраняет ссылку на страницу** в категорию пройденные, что **не допускает** её **повторный проход**, а также **собирает все возможные ссылки** с этой страницы **и добавляет их в очередь** для фильтрации уже их. Если данные не найдены, **то страница будет пропущена**. Парсинг продолжается, либо до **момента остановки пользователем**, либо **до прохода всех возможных ссылок** на домене.

# **Тестирование системы**

На этапе тестирования системы были написаны **автотесты** для проверки работоспособности и корректности работы конечных **точек API**. **Вручную** были протестированы остальные аспекты системы, такие как **сценарии парсинга**, **валидация данных, интеграция с БД, сохранение данных**.

# **Экономическая эффективность разработки системы**

В тему экономической эффективности стоит отметить, что **первоначальные затраты** на разработку составляют **1 200 000 руб**. **Годовая доходность проекта - ~1 848 000 руб./год** после вычета **налогов** и оплаты **хостинга**. Соответственно проект окупиться за **8 месяцев**.

# **Обоснование юридической правомерности**

Так как веб-скрапинг всего считался серой стороной разработки веб-приложений стоит объяснить юридическую правомерность использования данного продукта.

После анализа **ФЗ 149** и судебной практике по делу «**ВК против Дабл»** были выделены дополнительные требования к системе:

1. Собираем только **общедоступные данные**.
2. Не требовать **обхода авторизации или CAPTCHA**.
3. Соблюдать ограничения **robots.txt.**
4. Не собирать **персональные данные** (ФИО, телефоны, email).
5. Не **взламывать защиту** сайтов (обход блокировок, DDoS-атаки).
6. Не **нарушать пользовательские соглашения** (ToS).
7. Частотность запросов — **без перегрузки**.
8. Открытые источники — **только публичные страницы**.
9. Структурирование данных — **без искажения оригинальной информации**.

# **Результаты работы**

В результате проделанной работы была достигнута цель это разработка ПО для автоматизированного анализа рынка стройматериалов, а также цели:

1. **Анализ решений.**
2. **Определение требований.**
3. **Разработка архитектуры системы.**
4. **Реализация системы.**
5. **Тестирование.**
6. **Оценка эффективности.**

# **Перспективы развития системы**

Перспективы развития системы включают следующие направления:

1. **Внедрение ИИ** и машинного обучения – позволит проводить авто классификацию товаров, прогнозирование спроса.
2. **Самообучающийся парсер** – позволит автоматически искать необходимые сайты, распознавать структуру данных.
3. **Упрощенный интерфейс** - упрощенный ввод входных данных.
4. **Подключение к бизнес-системам** – интеграцию с последующей выгрузкой/загрузкой данных
5. **Расширенные форматы выгрузки** – позволит выдавать больше форматов данных.
6. **Трекер изменений в реальном времени** – позволит получать уведомление о изменении цены.