

Санкт-Петербургский государственный университет

Кафедра системного программирования

Группа 23.M07-мм

Логистический портал Cargotime.ru:  
оптимизация инструментов для оценки  
сроков прибытия контейнеров на основе  
исторических данных и текущего  
местоположения судна

*Счеревский Виктор Максимович*

Отчёт по преддипломной практике  
в форме «Производственное задание»

Научный руководитель:  
доцент кафедры информатики СПбГУ, к. т. н., Абрамов М.В.

Консультант:  
м. н. с. лаб. ПИИ СПб ФИЦ РАН, Есин М.С.

Санкт-Петербург  
2025

# Оглавление

<b>Термины и определения</b>	<b>3</b>
<b>Введение</b>	<b>4</b>
<b>1. Постановка задачи</b>	<b>6</b>
<b>2. Анализ релевантных работ</b>	<b>7</b>
2.1. Оценка сроков прибытия . . . . .	7
2.2. Сравнительный анализ аналогов . . . . .	9
<b>3. Обзор сервиса</b>	<b>13</b>
3.1. Основные компоненты и используемые технологии . . . .	13
3.2. Статусы перевозки . . . . .	17
3.3. Локации перевозки . . . . .	20
3.4. Данные о судах . . . . .	23
3.5. Общий сценарий работы сервиса . . . . .	25
<b>4. Унификация геоданных перевозки</b>	<b>27</b>
4.1. Подход к унификации . . . . .	27
4.2. Архитектура хранилища . . . . .	30
4.3. Методология оценки покрытия локаций . . . . .	33
4.4. Результаты . . . . .	35
<b>5. Инструменты унификации статусов перевозки</b>	<b>37</b>
5.1. Визуализация переходов между статусами . . . . .	37
5.2. Система управления унификацией статусов . . . . .	41
<b>6. Оптимизация парсинга данных о судах</b>	<b>43</b>
6.1. Фоновая актуализация и кеширование . . . . .	43
6.2. Многопоточная обработка запросов . . . . .	44
6.3. Результаты . . . . .	47
<b>Заключение</b>	<b>51</b>
<b>Список литературы</b>	<b>52</b>

# Термины и определения

- **Мультимодальный** — относящийся к перевозке грузов несколькими видами транспорта.
- **Унификация** — приведение элементов к единой системе или форме.
- **Пайплайн** — последовательность этапов обработки данных.
- **Сырые данные** — необработанные данные, полученные напрямую из источников.
- **AIS-данные (Automatic Identification System data)** — данные об идентификации и навигации судов (координаты, скорость, курс и др.), автоматически передаваемые через систему AIS для мониторинга движения судов.
- **ETA (Estimated Time of Arrival)** — расчетное время прибытия судна или груза в пункт назначения.
- **Парсинг (веб-скрейпинг)** — автоматическое извлечение данных из веб-страниц.
- **Трекинг** — отслеживание местоположения объектов.
- **Трек** — последовательность событий, отражающих этапы перемещения груза.
- **Микросервис** — автономный компонент системы, выполняющий отдельную функцию.
- **Микросервисная архитектура** — построение системы как набора независимых микросервисов.
- **Валидация** — проверка данных на соответствие заданным требованиям.

# Введение

Логистика – это ключевой элемент современной экономики. Она охватывает широкий спектр задач: от управления запасами до доставки товаров конечным потребителям. В условиях глобализации и растущей сложности цепочек поставок мониторинг перевозок, в частности, отслеживание перемещения контейнеров морских линий, приобретает критическую значимость [33, 16]. Эта возможность делает процесс доставки более прозрачным и предсказуемым для клиентов, а также позволяет оперативно реагировать на такие проблемы, как повреждение или кража контейнеров, что, в свою очередь, повышает надежность перевозки [37].

При этом одной из ключевых проблем остается низкая точность прогнозирования сроков доставки. По данным исследований [15], менее 60% контейнерных перевозок в мире завершаются в указанные компанией сроки, что приводит к сбоям в планировании, увеличению затрат и снижению эффективности логистических процессов. Данные за июнь 2024 года показывают, что среднее время задержки контейнерных перевозок составляет более 5 суток [15].

Существующие системы мониторинга и оценки сроков доставки, в том числе и крупные западные аналоги [12, 22, 5], характеризуются значительными ограничениями. Они либо не используют данные о текущем местоположении судна, либо игнорируют исторические сведения о задержках и иных особенностях маршрутов. Это приводит к невозможности предсказать отклонения, вызванные, например, сезонной перегруженностью портов или изменениями в инфраструктуре. Вследствие этого системы предоставляют лишь частичные решения, не позволяя эффективно управлять мультимодальными цепочками поставок.

На фоне данных ограничений портал Cargotime.ru [40] выступает как активно развивающийся цифровой инструмент, цель которого – предоставить клиентам грузовых компаний простой и доступный способ получения информации о перевозке грузов. Автоматизированные инструменты оценки сроков прибытия контейнеров – среди ключевых

разрабатываемых сервисом продуктов. Эти инструменты востребованы среди логистов благодаря возможности получить альтернативный прогноз по срокам доставки, не зависящий от предоставленных перевозчиками данных, что особенно важно для минимизации рисков в логистике.

Однако, чтобы обеспечить высокую точность и надёжность таких инструментов, требуется предварительная работа по устранению ряда ограничений в логистических данных. Локации маршрутов и статусы перевозки, передаваемые транспортными компаниями, часто представлены в свободной форме и требуют приведения к единому формату. Координаты портов и станций во многих случаях отсутствуют, что затрудняет построение маршрутов и требует дополнительной географической привязки. Отдельную задачу представляет объединение морских и железнодорожных локаций в рамках единой модели мультимодальной перевозки. Кроме того, для поддержки интерактивных сервисов и аналитических инструментов, ориентированных на текущее положение судов, необходимо обеспечить более быстрый доступ к актуальной информации.

Настоящая работа направлена на решение указанных проблем как на необходимый этап оптимизации инструментов оценки сроков прибытия контейнеров. Повышение согласованности, полноты и доступности логистической информации призвано не только повысить точность аналитических сценариев, реализуемых в рамках существующих сервисов, но и создать надёжную основу для построения предсказательной модели, способной автоматически оценивать ключевые параметры перевозки, включая сроки доставки, на базе унифицированных и оперативно обновляемых данных.

# 1. Постановка задачи

Целью данной работы является повышение качества предоставляемой информации о контейнерных перевозках и скорости доступа к ней для дальнейшего построения предиктивной модели параметров перевозки путем разработки и внедрения пайплайна автоматизированной обработки, стандартизации и обогащения логистических данных из различных источников. Для выполнения цели были поставлены следующие задачи:

- разработать механизм стандартизации и обогащения сырых данных о локациях контейнерной перевозки, использующий локальное хранилище данных;
- разработать интерактивный инструмент для экспертного сопоставления разнородных статусов перевозки с унифицированной системой кодов;
- обеспечить высокоскоростной доступ к данным о судах путем внедрения фоновой актуализации и многопоточной обработки запросов;
- внедрить разработанный пайплайн обработки и стандартизации сырых данных в существующий продукт.

## 2. Анализ релевантных работ

Глава посвящена анализу требований к процессу обработки, унификации и обогащения, необходимых для использования полученных данных в работе аналитических систем, в том числе для предсказания сроков прибытия. В главе рассматривается, как разнородность и неполнота логистических данных, предоставляемых различными операторами, затрудняют их автоматизированную обработку. Особое внимание уделено необходимости нормализации локаций и логистических статусов, а также обогащения информации источниками, включая AIS-данные о судах. На основе обзора существующих коммерческих решений показана реализуемость подхода к построению пайплайна, обеспечивающего стандартизированное и ускоренное предоставление данных, необходимое для дальнейшего применения в прикладных задачах, включая прогнозирование ЕТА.

### 2.1. Оценка сроков прибытия

Согласно ежегодному докладу международной организации UNCTAD о состоянии мирового рынка морских перевозок за 2024 год [31], морская контейнерная логистика обеспечивает не менее 80% мирового товарооборота. На фоне устойчивого роста объемов перевозок возрастает значимость эффективного планирования мультимодальных логистических цепочек и сопряженных с ними портовых операций. Один из ключевых факторов такого планирования — высокая точность прогнозирования ЕТА. Низкая достоверность ЕТА может стать причиной значительных операционных и экономических потерь для всех участников цепочки поставок: от вынужденного простаивания портовой инфраструктуры до сбоев в расписаниях перевозок, включая баржи, железнодорожный и автотранспорт. Как правило, предварительное ЕТА формируется самой морской линией и публикуется через её информационные сервисы, однако, согласно результатам исследований [15], более 40% контейнерных перевозок в мире прибывают позже заявленного морской линией срока, при этом среднее отклонение от первоначально-

го прогноза ЕТА превышает пять суток.

Точное прогнозирование ЕТА контейнера невозможно без учёта маршрута судна и его сопоставления с актуальными AIS-данными о местоположении судна [23]. На итоговое время прибытия также существенно влияют характеристики самого судна — текущие скорость и курс [24], — а также принадлежность к конкретной морской линии [18], поскольку операционные стратегии перевозчика, включая политику по расходу топлива и выбор типовых маршрутов, напрямую отражаются на балансе скорости и надёжности доставки.

Главным препятствием для построения независимого прогноза ЕТА, не опирающегося на расчёты самих морских линий, является фрагментарность данных о контейнерных перевозках. Каждая транспортная компания использует собственные структуры описания маршрутов, портов захода, этапов доставки и даже базовой информации о судне и рейсе. Даже такие критически важные параметры, как ЕТА и текущий статус контейнера (например, «в пути», «в порту», «на погрузке»), могут быть представлены в различных форматах, с разной степенью детализации, частотой обновления и трактовкой значений. Однако, фрагментарность данных проявляется не только в разнообразии форматов, но и в неполноте самих данных. Некоторые провайдеры предоставляют сведения о номере рейса и названии судна — информации, которая позволяет существенно уточнить прогноз ЕТА, — но при этом не публикуют ключевые характеристики судна: его текущее местоположение, курс и скорость. Возникает необходимость дополнительно обогащать данные из внешних источников, чтобы обеспечить модели достаточным контекстом для построения точного предсказания.

В условиях подобной фрагментарности данных использование универсальной модели прогнозирования ЕТА для разных компаний становится невозможным без промежуточного этапа приведения данных к единому виду. Разнородность и неполнота исходных представлений препятствуют масштабированию аналитических и предиктивных решений и затрудняют интеграцию модели в существующий продукт. Поэтому цель данной работы — разработка слоя нормализации и обогаще-



ния данных, обеспечивающего единое представление полной информации о перевозке, — является актуальной в контексте развития сервиса Cargotime.ru.

## 2.2. Сравнительный анализ аналогов

В рамках исследования была поставлена задача построения системы, способной нормализовать логистические данные и обогащать их сведениями из внешних источников. Чтобы оценить реализуемость такого подхода в условиях реального рынка, был проведён сравнительный анализ существующих решений, представленных как на международном, так и на российском логистическом рынке. Цель данного анализа — определить, насколько практика унификации и обогащения логистических данных уже применяется на рынке контейнерных перевозок и какова степень зрелости таких решений.

Существующие аналоги сравнивались по приведённым ниже критериям, которые были выбраны с учётом ключевых требований к системе, способной обеспечивать нормализацию логистических данных, независимое предсказание ЕТА и адаптацию к российским реалиям.

1. **Наличие механизмов унификации географических локаций.** Оценивается на основе способности платформы приводить названия портов и других географических объектов к единому виду. Если при анализе нескольких перевозок от разных операторов одна и та же локация представлена единообразно, то считается, что механизм унификации присутствует.
2. **Наличие механизмов унификации логистических статусов.** Аналогично первому критерию, оценивается способность сервиса приводить разнородные текстовые статусы к единому перечню событий. Если на уровне интерфейса или API платформы статусы от разных операторов отображаются в одинаковой классификации, это трактуется как наличие механизма унификации.

3. **Интеграция с системами автоматической идентификации судов (AIS).** Для проверки данного критерия анализировалась перевозка, находящаяся в активной морской фазе (контейнер на борту судна в пути). Если в ответе сервиса присутствует информация о текущем местоположении судна, то считается, что интеграция с AIS имеется.
4. **Построение собственных моделей предсказания ЕТА.** Платформа имеет собственную модель прогнозирования ЕТА, если публично заявляет о внутреннем расчёте ЕТА, не основанном исключительно на ЕТА перевозчика, и по крайней мере в одной из проверенных перевозок отображается ЕТА, отличающееся от заявленного перевозчиком срока.
5. **Интеграция с операторами, работающими с Россией.** Оценивается численно: для каждой платформы подсчитывается количество поддерживаемых перевозчиков из фиксированного списка из 12 компаний, включающего российских и ориентированных на РФ операторов, которые явно указывают российские порты и терминалы в истории перевозки.
6. **Поддержка отслеживания российских железнодорожных отправок.** Поддержка осуществляется в случае, если аналог позволяет отслеживать отправления в крупных российских железнодорожных операторах.

Анализ был проведён на основе открытых данных из веб-интерфейсов и публичных API сервисов отслеживания контейнерных перевозок. В выборку вошли платформы, занимающие высокие позиции в поисковой выдаче по релевантным запросам ("*container tracking*", "*отслеживание контейнеров*"). Рассматривались ресурсы, стабильно появляющиеся на первых страницах поисковиков, включая международные агрегаторы, сервисы морских линий, а также локальные решения, ориентированные на российский рынок. При этом не рассматривались ресурсы, работоспособность которых не удалось подтвердить. Результаты

анализа приведены в таблице 1.

Аналог	Кол-во операторов	Кол-во операторов для РФ	Поддержка отслеживания ЖД	Унификация локаций	Унификация статусов	Данные о судне (AIS)	Прогноз ЕТА
CARGOTIME	81	12	+				-
GoCOMET	512	9	-	-	+	+	+
SEARATES	171	6	-	+	+	+	-
SHIPSGo	144	6	-	+	+	+	+
SAFECUBE	180	5	-	+	+	-	-
TRACKTRACE	168	3	-	-	-	-	-
BESTDEALS.RU	13	2	-	-	-	-	-
CREDOTRANS.RU	25	2	-	-	-	-	-
VISIWISE	36	1	-	+	+	+	-
VISIONAPI	35	1	-	+	+	+	-

Таблица 1: Сравнительный анализ аналогов CARGOTIME.RU

Проведённый анализ показал, что практика нормализации логистических данных и построения независимого прогноза ЕТА уже частично реализуется в существующих решениях и технологически осуществима. Отдельно стоит отметить, что ни один из рассмотренных сервисов не поддерживает отслеживание железнодорожных отправок, что существенно снижает их применимость в российских условиях. С учётом высокой доли железнодорожных перевозок во внутреннем грузообороте на территории РФ<sup>1</sup>, отсутствие поддержки ЖД-направления ограничивает возможности построения сквозной модели ЕТА и подчёркивает необходимость разработки решения, адаптированного под специфику российской внутренней логистики.

Таким образом, для построения достоверной и универсальной модели оценки ЕТА необходимыми являются наличие механизмов унификации локаций и логистических статусов, а также интеграция с AIS-данными, обеспечивающая объективную информацию о текущем положении и динамике движения судов. Проведённый анализ подтверждает техническую реализуемость такого подхода: ключевые элементы — нормализация разнородных данных и их обогащение внешними источниками — успешно применяются в ряде существующих коммерческих

<sup>1</sup>Внутренний грузооборот в России в 2023 году на 47,5% состоит из железнодорожных перевозок [<https://rosstat.gov.ru/storage/mediabank/Gruzooborot-02-2025.xlsx>].

решений, что демонстрирует жизнеспособность предлагаемого подхода и обоснованность разработки автоматизированного пайплайна для стандартизации и обогащения логистических данных.

## **3. Обзор сервиса**

Глава кратко описывает текущую архитектуру сервиса, используемые технологии и принципы работы с сырыми данными на уровне, достаточном для понимания ограничений системы. Внимание уделяется используемым подходам к стандартизации разнородных статусов перевозок и географических локаций контейнерных маршрутов. Рассматривается взаимодействие с сервисом судов как ключевым источником информации о текущем местоположении и технических параметрах контейнеровозов. Целью данной главы является фиксация исходного состояния системы, что в дальнейшем позволит объективно оценить эффективность предлагаемых улучшений.

### **3.1. Основные компоненты и используемые технологии**

#### **3.1.1. Архитектура сервиса трекинга**

Все компоненты портала CARGOTIME.RU, включая сервис отслеживания контейнеров, реализованы с использованием микросервисной архитектуры. Такой подход обеспечивает модульность, гибкость и масштабируемость системы, что позволяет балансировать нагрузку на отдельные микросервисы, а также обеспечивает отказоустойчивость всего сервиса в случае выхода из строя какой-то из его частей.

Часть архитектуры сервиса трекинга контейнеров, связанная с инструментами для оценки сроков доставки, выглядит следующим образом:

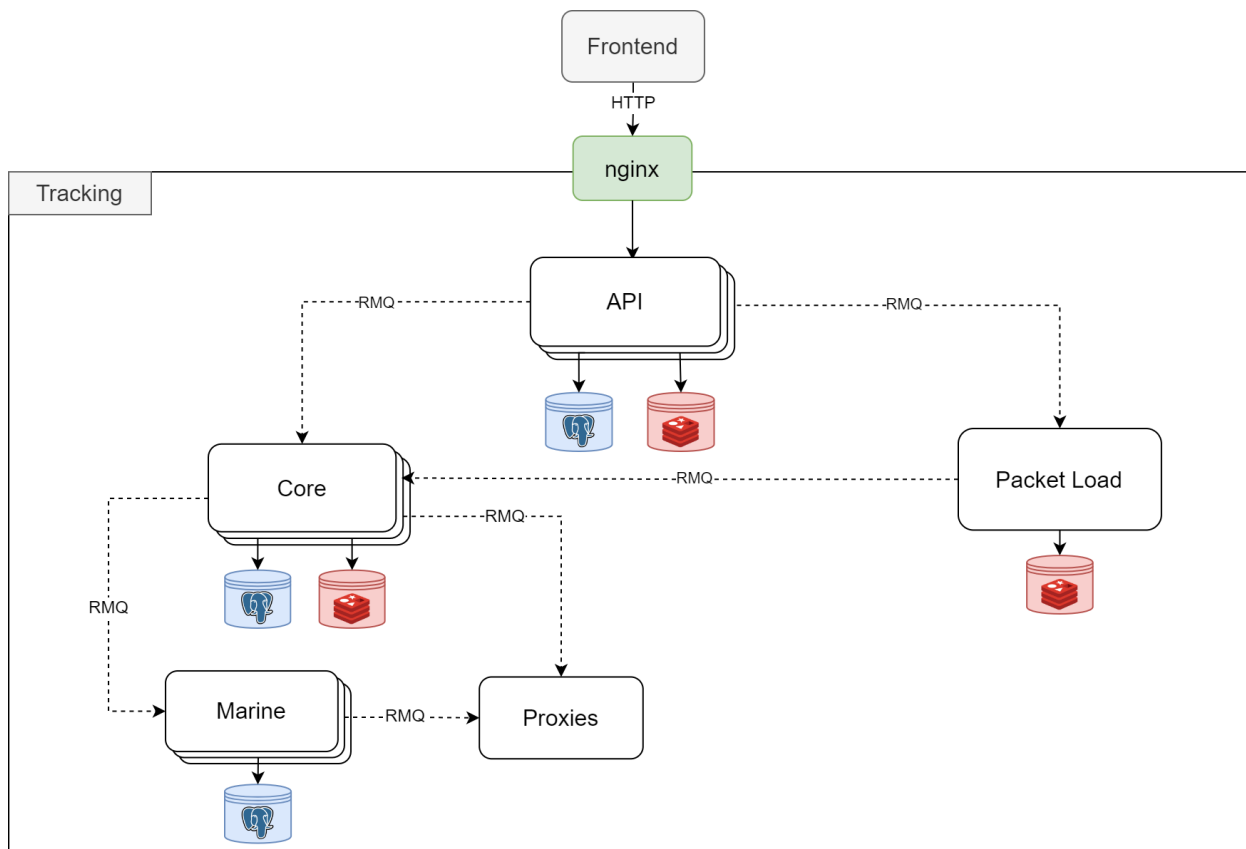


Рис. 1: Часть архитектуры сервиса трекинга

На схеме сплошными стрелками с меткой “HTTP” обозначено взаимодействие по протоколу HTTP, пунктирными стрелками – взаимодействие с помощью брокера сообщений RABBITMQ по протоколу AMQP. Эта технология играет ключевую роль в обеспечении отказоустойчивости системы, поскольку предоставляет гарантии доставки сообщений, а также помогает распределять нагрузку между репликами микросервисов. Синим и красным цветами обозначены база данных PostgreSQL и хранилище REDIS соответственно. Более подробное описание этих технологий будет приведено в следующем разделе.

Описание компонентов приведено ниже.

- **Core** – ядро парсеров, в котором обрабатываются все запросы к источникам. Данный сервис реплицируется.
- **API** – промежуточный сервис между веб-интерфейсом и парсерами. В нём агрегируется и кэшируется метаданная о компаниях и направляется клиенту вместе с ответами **Core**.

- **Packet Load** – сервис, необходимый для обработки больших групп трек-номеров, которые клиенты могут загрузить через специальный интерфейс на сайте.
- **Proxies** – оптимально распределяет динамические и статические порты для парсинга.
- **Marine** – сервис судов, включающий инструменты для парсинга информации о судах и портах из открытых источников.
- **Frontend** – веб-сайт, предоставляющий клиентский интерфейс для взаимодействия с инструментами портала, такими, как интерактивная карта с маршрутом контейнера, и другими.

Данная архитектура создаёт прочный фундамент для дальнейшего развития функциональности портала. Модульная организация системы, использование микросервисного подхода и поддержка масштабируемости обеспечивают возможность гибкой интеграции новых компонентов без риска нарушения работоспособности существующих сервисов. В частности, данная архитектура позволит эффективно внедрить инструменты прогнозирования сроков доставки.

### 3.1.2. Используемые технологии

Ниже перечислены основные программные технологии, используемые в сервисе трекинга контейнеров.

- **NODE.JS [21]** – программная платформа с открытым исходным кодом, предназначенная для разработки серверных приложений на языке JavaScript.
- **TYPESCRIPT [34]** – строго типизированный язык программирования с открытым исходным кодом, компилирующийся в JAVASCRIPT и расширяющий его возможности. Повсеместно применяется в проекте для построения серверных приложений.

- NESTJS [19] – современный фреймворк с открытым исходным кодом для создания производительных и масштабируемых серверных приложений на NODE.JS. Поддерживает разработку на языках JavaScript и TypeScript.
- NEXT.JS [20] – фреймворк для создания веб-интерфейсов на основе REACT [29]. Используется для разработки фронтенд-сервисов портала.
- RABBITMQ [28] – открытая система обмена сообщениями (брокер сообщений), реализующая протокол AMQP. С помощью RABBITMQ производится обмен данными между ядром парсеров, API трекинга и другими сервисами.
- REDIS [30] – высокопроизводительная in-memory (в оперативной памяти) база данных с доступным исходным кодом (source-available). Используется для кэширования данных в сервисе ядра парсеров, API трекинга и сервиса пакетной загрузки.
- ELK (ELASTICSEARCH, LOGSTASH, KIBANA) [10] – стек технологий для поиска, анализа и визуализации логов. Используется для поиска информации о новых локациях среди ответов парсеров.
- POSTGRESQL [26] – свободная реляционная СУБД. Используется в сервисе ядра парсеров для хранения информации о компаниях, портах и др.
- DOCKER [9] – открытая платформа для разработки, доставки и запуска приложений в контейнерах. Также используется DOCKER COMPOSE для автоматизированного развертывания контейнеров в единой сети.
- PLAYWRIGHT [25] – библиотека для автоматизированной работы с веб-сайтами. Поддерживает эмуляцию различных браузеров.
- PUPPETEER [27] – ещё один инструмент для эмуляции работы браузера. Ориентирован преимущественно на CHROMIUM [4].



## 3.2. Статусы перевозки

### 3.2.1. Система унифицированных статус-кодов

Для корректной работы сервиса по отслеживанию контейнеров, а также для дальнейшего построения модели прогнозирования ETA критически важно корректно определять текущее положение контейнера в рамках логистического процесса. Источником этой информации являются статусы перевозки от транспортных компаний. Однако в отрасли отсутствует единый стандарт описания событий перевозки, которым пользуются все компании: каждая использует собственные наборы статусов, отличающиеся по структуре, семантике и уровню детализации. В результате модель не может обрабатывать такие данные напрямую: ей необходима единообразная система статусов, в которой каждое событие интерпретируется однозначно.

В связи с этим командой CARGOTIME ранее была разработана система унифицированных статус-кодов. За основу был взят мировой стандарт DCSA (Digital Container Shipping Association) [7], которому уже следуют некоторые крупные перевозчики [6]. Система включает 18 статус-кодов, охватывающих ключевые этапы перевозки — от выдачи пустого контейнера до его возврата после доставки. Например, CLL обозначает погрузку в порту отправления, VDL — выход судна из порта, CDD — выгрузку в порту назначения, CDC — доставку контейнера получателю. Переход на иную систему статус-кодов в текущей архитектуре продукта представляется нецелесообразным, поскольку используемая система статус-кодов интегрирована в несколько ключевых компонентов платформы: публичный API, пользовательский интерфейс, систему уведомлений и внутренние модели маршрутизации. Изменение этой системы потребует значительной переработки всей инфраструктуры, включая миграцию существующих данных и адаптацию клиентских интеграций.

Эта классификация позволит модели точно определить, на каком этапе находится перевозка, и использовать соответствующие исторические данные для прогнозирования оставшегося времени до завершения

маршрута независимо от формата данных, предоставляемых конкретной транспортной компанией. Вместе с тем корректное сопоставление исходных статусов компаний с унифицированной системой статус-кодов представляет собой нетривиальную задачу, требующую дополнительных подходов и инструментов.

### **3.2.2. Ограничения первичной интерпретации статусов**

Процесс построения сопоставления статусов перевозки с унифицированной системой кодов не может быть полностью автоматизирован. Это связано с тем, что подходы, основанные исключительно на интерпретации текстовой информации, недостаточны для надёжного определения значения статусов, предоставляемых различными транспортными компаниями. Хотя между статусами существуют формальные логические связи — например, выход судна из порта возможен только после его погрузки — такие закономерности не зафиксированы в самих формулировках статусов. Корректная интерпретация требует анализа взаимосвязей между статусами, частоты их переходов, а также учёта возможных ошибок или искажений со стороны поставщика данных. Этот анализ может выполняться как живым экспертом, так и в связке с языковой моделью, где модель предлагает интерпретацию, а эксперт её уточняет и утверждает.

Одной из распространённых проблем при построении сопоставления является неоднозначность наименований. Например, компания MAERSK использует статус GATE-OUT для обозначения сразу нескольких различных операций: отправки пустого контейнера отправителю из порта отправления, выдачи груженого контейнера получателю в порту назначения, а также перемещений контейнера по суше между промежуточными терминалами. Без информации о предыдущих и последующих событиях корректная интерпретация такого статуса затруднена и нередко приводит к ошибкам.

Дополнительную сложность представляет определение завершения перевозки. Разные компании фиксируют "финальные" события по-разному, и в ряде случаев статус, который воспринимается как окон-

чение маршрута, фактически является промежуточным. Без анализа всего маршрута и понимания контекста сложно точно определить момент завершения логистической цепочки.

Как только соответствие между исходными статусами и унифицированными кодами построено, процесс интерпретации становится технически простым: во время обработки клиентского запроса сервис просто подставляет соответствующий код либо применяет несложные правила сопоставления на уровне кода без дополнительных валидации и уточнений со стороны эксперта или языковой модели. Таким образом, основная сложность заключается не в применении сопоставлений, а в их первоначальном построении.

В результате процесс требует участия эксперта, который на основе собственного опыта и анализа информации вручную устанавливает соответствие между исходными статусами и унифицированными статус-кодами. На начальном этапе такой анализ выполнялся исключительно по текстовым описаниям, без учёта их порядка и взаимосвязей, что существенно ограничило точность первичных сопоставлений.

Таким образом, задача построения сопоставления статусов и унифицированных кодов требует учёта дополнительных факторов, включая порядок следования событий, частоту переходов между статусами, структуру маршрута и специфику конкретного перевозчика. Кроме того, со временем могут подключаться новые компании, а также изменяться статусные данные у уже интегрированных. Для решения этих задач необходим инструмент, который позволит редактировать как сам набор исходных статусов, поступающих от компаний, так и их соответствие унифицированным статус-кодам, а также предоставит эксперту весь необходимый контекст для принятия обоснованных решений при сопоставлении. Подходы к решению этих задач будут рассмотрены в одной из следующих глав.

### 3.3. Локации перевозки

#### 3.3.1. Источник геоданных

В рамках текущей реализации системы для получения геоданных о локациях маршрута используется сторонний картографический сервис BINGMAPS [1]. Ядро парсеров при обработке полученной от компаний информации о перевозке контейнера отправляет в BINGMAPS название каждого географического объекта и использует первый из возвращённых сервисом результатов, который считается наиболее релевантным. Реализованный ранее подход [43] обладает рядом существенных ограничений, которые негативно влияют на качество и надёжность работы системы.

Во-первых, BINGMAPS, как и любой другой универсальный картографический сервис, не ориентирован на специфические задачи морской и железнодорожной отраслей логистики. Это приводит к частым случаям некорректной идентификации объектов — вместо искомого порта или станции система может получить координаты центра одноимённого города. Для целей данной работы особенно важно обеспечить корректную унификацию именно морских портов и железнодорожных станций, поскольку они составляют ключевые точки маршрута и играют критическую роль в формировании точного и воспроизводимого представления о перевозке. Описанный выше подход, предполагающий сохранение ответов BINGMAPS без какой-либо валидации, не подходит для целей унификации, поскольку не различает транспортную и административную семантику названий. Для решения задачи требуется специализированный перечень объектов, ориентированный на логистическую составляющую локаций.

Во-вторых, BINGMAPS является внешним сервисом, скорость отклика, стабильность и пропускную способность которого нельзя контролировать силами разработчиков. Это снижает надёжность и предсказуемость работы решения, особенно при масштабной загрузке данных или необходимости обработки большого количества запросов в реальном времени. Несмотря на наличие механизма кэширования, он не устра-

няет зависимости от внешней инфраструктуры и не покрывает случаи массовой загрузки новых или ранее неиспользуемых локаций, что делает текущее решение недостаточно устойчивым для промышленных нагрузок и требует дальнейшей доработки.

Таким образом, использование внешнего сервиса в качестве основного источника геоданных имеет критические ограничения, связанные с точностью, полнотой информации и производительностью. В условиях специализированных требований морской и железнодорожной логистики более надёжным и масштабируемым подходом является разработка собственной локальной базы портовых и железнодорожных объектов, обеспечивающей устойчивую работу системы и контроль над качеством данных.

### **3.3.2. Необходимость унификации локаций**

В текущей версии сервиса реализована частичная географическая привязка локаций: каждой точке маршрута морского контейнера присваиваются координаты через BINGMAPS, что позволяет визуализировать перемещение контейнера на карте. Визуализация строится на основе сопоставления статусов с унифицированными кодами, выделения ключевых точек маршрута (портов и терминалов), построения пути по заранее рассчитанной сетке переходов и встраивания текущего положения судна по AIS-данным [43]. Это даёт возможность отобразить маршрут с учётом текущей геопозиции контейнера.

Визуализация маршрута — лишь одна из второстепенных задач, в которой используются локации перевозки. Более важной задачей в свете дальнейшего построения модели предсказания сроков прибытия и достижения цели работы является аналитическая обработка маршрутов: расчёт времени в пути для типовых направлений, агрегация статистики по портам и узлам и выявление паттернов движения груза. Все эти задачи требуют того, чтобы один и тот же объект представлялся в системе одинаково — независимо от различий в написании, языке или структуре входных данных, зависящих от транспортной компании. Без унификации локаций, как и без унификации статусов перевозки, ста-

новится невозможным сравнивать маршруты между собой, объединять данные от разных перевозчиков и строить обобщённые модели поведения.

Однако полноценной унификации географических объектов не осуществляется. Во избежание искажений данные сохраняются в том виде, в каком они были переданы в ответе транспортной компании. Это относится к написанию стран, регионов, а также наименований портов и терминалов. Такое решение на предыдущем этапе было принято осознанно, поскольку BINGMAPS не гарантирует достаточную точность данных — в ряде случаев сервис может возвращать координаты объекта, который не имеет отношения к перевозке и вообще находится в соседней стране. В условиях использования стороннего картографического сервиса попытка унифицировать локации может привести к потере смысла или подмене информации, исходно переданной перевозчиком.

Дополнительным ограничением текущего решения является отсутствие обработки железнодорожных станций и терминалов России, Китая и стран СНГ, несмотря на то, что весомая часть клиентских перевозок (около 35% от всех запросов, по внутренним данным за апрель 2025 года), проходит именно через железнодорожные станции. В текущей версии системы такие локации не обогащаются координатами и не нормализуются, поскольку визуализация маршрута по земле ещё не реализована. Это снижает полноту и сопоставимость данных для маршрутов, включающих железнодорожное плечо.

Таким образом, несмотря на наличие базовой системы распознавания и географической привязки объектов, текущий подход к обработке локаций перевозки обладает рядом существенных ограничений, связанных с отсутствием полноценной унификации и слабой поддержкой железнодорожной логистики. Для повышения точности, воспроизводимости и аналитической ценности маршрутов требуется внедрение локального справочника портов, терминалов и станций, обеспечивающего их надёжную интерпретацию.

### 3.4. Данные о судах

#### 3.4.1. Источник данных о судах

В рамках текущей реализации AIS-данные о судне используются исключительно для построения маршрутов — визуального отображения положения контейнера на карте в морской части перевозки. Однако, как было установлено ранее, такие данные играют ключевую роль и для построения предсказательной модели ETA: фактическое местоположение судна, его скорость, курс и другие технические параметры напрямую влияют на точность прогноза времени прибытия.

AIS-данные извлекаются для каждого контейнера, если в ответе транспортной компании указано название судна или его код IMO (уникальный идентификационный номер судна, присваиваемый Международной морской организацией). Эти данные уже стандартизированы: коды IMO и наименования грузовых судов являются уникальными, что позволяет однозначно идентифицировать судно и избежать ошибок при получении информации из сторонних источников.

В качестве источника данных используется сервис MARINETRAFFIC [17], который отличается самым полным покрытием контейнерных судов среди аналогов (например, VESSELFINDER [32] содержит меньшую часть флота). Согласно внутренней статистике работы сервиса за период с октября 2024 года, в 77,1% случаев данные о позициях судов поступают не позднее часа с момента их фиксации первоисточником (спутником или наземным радаром). По тем же данным сервис позволяет находить судно из ответа транспортной компании в 97,7% случаев. В рамках работы считается, что подобный подход к сбору AIS-данных о судах, не использующий платные API-интерфейсы, обеспечивает достаточные полноту и точность данных.

#### 3.4.2. Необходимость ускорения доступа к AIS-данным

Вместе с тем, скорость получения данных остаётся достаточно низкой из-за зависимости технологий эмуляции браузера: по внутренним измерениям, проведённым в октябре 2024 года, запрос к

MARINETRAFFIC обрабатывается в среднем 8,22 секунд с учётом извлечения запроса из кэша в 25,9% случаев. Для сравнения, среднее время сбора данных о контейнере в компаниях без учёта сбора AIS-данных за тот же период составляет 11,76 секунд. Таким образом, время запроса к MARINETRAFFIC составляет значительную долю общего времени сбора данных, что делает оптимизацию механизма получения информации о судах приоритетной задачей для повышения скорости получения прогноза и общей производительности сервиса.

В рамках работы был проведён анализ распределения запросов к сервису MARINE, собирающему AIS-данные по судам. Установлено, что 18,2% судов формируют 80,0% всех обращений к сервису, что свидетельствует о высокой концентрации трафика на ограниченном числе судов. Это позволяет реализовать стратегию фоновое обновления данных по наиболее востребованным судам, повышая долю кэшированных ответов и, как следствие, снижая среднее время отклика сервиса.

Кроме того, для регулярного обновления прогнозов ETA, которое будет заложено как функциональность личного кабинета, потребуются данные о большом количестве судов сразу, поэтому актуальна задача быстрой обработки пакетных запросов. Потенциальным направлением оптимизации является использование асинхронной обработки с задействованием возможностей многопоточности браузерного движка CHROMIUM и инструментария PLAYWRIGHT, что позволяет реализовать параллельный парсинг, несмотря на однопоточность среды NODE.JS.

AIS-данные играют ключевую роль не только в визуализации маршрутов судов, но и в повышении точности прогнозов ETA за счёт учёта фактической навигационной информации. Несмотря на высокую полноту и точность данных, текущий механизм получения данных через эмуляцию браузера остаётся узким местом при обработке запросов к сервису.

Предложенные оптимизации должны существенно ускорить отклик сервиса как при обработке одиночных запросов на получение AIS-данных, так и при обработке пакетных запросов обновления.



### 3.5. Общий сценарий работы сервиса

Ниже приведён сценарий работы сервиса по отслеживанию контейнеров с учётом предложенных путей его оптимизации.

1. **Получение запроса от пользователя или по расписанию.** Пользователь или система инициирует отслеживание контейнера по номеру (рис. 2).
2. **Запрос к API транспортной компании.** Ядро делает запрос к публичному API (или его аналогу) перевозчика и получает данные о локациях перевозки, событиях со статусами и судне, которое перевозит контейнер.
3. **Нормализация статусов перевозки.** Полученные «сырые» статусы интерпретируются с помощью заранее построенного соответствия между исходными значениями и системой унифицированных кодов, сформированного с использованием разработанных инструментов экспертной разметки.
4. **Нормализация локаций перевозки.** Полученные локации перевозки унифицируются и обогащаются данными и координатами. Используются разработанные подходы к унификации и обогащению локаций.
5. **Построение маршрута.** На основе унифицированных данных о локациях и выделенных по статус-кодам этапах перевозки строится маршрут (рис. 3).
6. **Добавление информации о судне.** Если в ответе компании было указано название судна или его IMO код, то у оптимизированного в рамках данной работы сервиса MARINE запрашиваются AIS-данные. Местоположение судна (если оно актуально для текущей перевозки) встраивается в построенный маршрут.

Простой трекинг

Пакетная загрузка номеров контейнеров

Номер контейнера состоит из 4 буквенных символов, идентифицирующих компанию перевозчика, и 7 цифровых символов.

Например, **ABCD1234567**. ABCD — это префикс морской линии, 1234567 — семизначный цифровой идентификатор.

Укажите номер контейнера, коносамента или букинга!

Тип поиска:

по контейнеру

Номер контейнера:

MRKU3496752

Компания:

Maersk

Временно недоступные компании

Наиболее вероятные варианты:

Maersk

ТрансКонтейнер

Найти контейнер

Рис. 2: Ввод данных для трекинга

Простой трекинг

Пакетная загрузка номеров контейнеров

MRKU3496752

Искать снова

Последняя перевозка контейнера:

Текущее событие

Перевозка

Maersk

Детали перевозки

Текущий статус:

DISCHARG

Тип контейнера:

40 DRY

Последняя дата:

16.12.2024 | 01:20

Просмотреть результаты на сайте компании

Прибытие контейнера:

26.01.2025 | 03:00

Статус:

CONTAINER ARRIVAL

Судно:

COSCO SAO PAULO

Рейс:

503W

Порт выгрузки:

MALAYSIA TANJUNG PELEPAS

Отметка о событии:

Заход судна в порт назначения

Рис. 3: Результат поиска информации о контейнере

Таким образом, предложенный сценарий демонстрирует, как интеграция механизмов нормализации, кэширования, фонового обновления и пакетной обработки запросов позволяет существенно повысить устойчивость и производительность сервиса при сохранении высокого качества логистических данных.

## 4. Унификация геоданных перевозки

Глава описывает новый подход к унификации и обогащению геоданных локаций маршрута контейнера, подразумевающий замену внешнего источника данных на внутреннее хранилище. Глава также включает в себя описание структуры сущностей и механизм взаимодействия системы с базой данных, предназначенной для хранения унифицированных данных о портах и станциях, и описание методологии оценки эффективности выбранного подхода.

### 4.1. Подход к унификации

#### 4.1.1. Решаемые проблемы

В процессе обработки логистических данных, получаемых от различных транспортных компаний, выявлен ряд системных проблем, затрудняющих корректную интерпретацию и визуализацию маршрута контейнера. Ниже перечислены основные из них.

1. **Нестандартизированное представление локаций.** Наименования локаций в ответах компаний часто указаны в произвольной форме, на разных языках и с различной степенью детализации. Это затрудняет как автоматическую обработку данных, так и ручной анализ маршрута аналитиками. Кроме того, отсутствие единообразия мешает корректной локализации данных в русской и английской версиях портала.
2. **Ограниченность координатных данных.** Лишь небольшое число перевозчиков (3–4 из более чем 80) указывают координаты локаций. В большинстве случаев это не позволяет сразу отобразить маршрут контейнера на карте и требует дополнительной обработки.
3. **Нечёткая географическая привязка.** В отдельных случаях исходная информация недостаточна для уверенного определения

страны или региона, в котором находится контейнер. Это затрудняет отображение маршрута и ограничивает возможности географического анализа.

Унификация локаций позволит устранить эти проблемы и облегчить построение маршрута на основе собранных данных.

#### **4.1.2. Унификация портов**

Для морских и сухопутных портов в качестве основного идентификатора используется международный стандарт UN/LOCODE [35], включающий двухбуквенный код страны (по ISO 3166-1 ALPHA-2) и трёхбуквенное обозначение порта. Многие транспортные компании указывают коды портов или их названия в унифицированной форме, что позволяет выполнять автоматическое сопоставление.

Однако на практике встречаются расхождения. Один и тот же порт может обозначаться разными кодами — например, для Шанхая используются как CNSHA, так и CNSHG. Кроме того, некоторые компании, такие как FESCO, применяют обозначения, визуально похожие на UN/LOCODE, но не соответствующие официальному стандарту.

Благодаря наличию общепринятого международного стандарта UN/LOCODE унификацию портовых локаций можно реализовать без существенных затруднений. Основные усилия будут направлены на обработку оставшихся исключений — нестандартизированных кодов и неоднозначных обозначений, используемых отдельными перевозчиками.

#### **4.1.3. Унификация железнодорожных станций**

В процессе анализа логистических данных выяснилось, что значительная часть локаций, поступающих от российских транспортных компаний (в частности, FESCO и ТРАНСКОНТЕЙНЕРА), не относится к портам. Эти локации представляют собой железнодорожные станции, для которых, в отличие от портов, не существует общепринятого международного стандарта, подобного UN/LOCODE. Это создало необходи-

мость внедрения отдельного подхода к унификации такого рода объектов.

Для станций, расположенных в России, странах СНГ и Балтии, используется система ЕСР (Единая Сетевая Разметка) [38] — исторически сложившийся стандарт, введенный ещё во времена СССР. Каждая станция в этой системе имеет уникальный пятизначный цифровой код, что позволяет однозначно её идентифицировать. Благодаря единообразию структуры и широкой распространённости, коды ЕСР рассматриваются в качестве основы для унификации железнодорожных локаций на территории постсоветского пространства. Их использование позволит обеспечить точную географическую привязку и сопоставимость данных при построении маршрутов.

Унификация китайских железнодорожных станций требует дополнительной работы, поскольку в Китае используются две независимые системы идентификации, представленные ниже.

1. **Трёхбуквенные коды**, основанные на латинице. Эти обозначения формируются на базе названия станции, однако покрытие таких кодов ограничено. Основным доступный источник представлен в виде карты [3], не содержащей структурированных координат, что затрудняет точную географическую привязку станций.
2. **Пятизначные цифровые коды**, основанные на китайском государственном стандарте GB/T 10302-2010 (также известном как TMIS) [11]. Эта система поддерживается организацией сотрудничества железных дорог (ОСЖД) [41] в перечне грузовых станций Китайских железных дорог [42]. Благодаря формализованной структуре и более широкому покрытию по сравнению с трёхбуквенными кодами, пятизначные коды были выбраны в качестве основного идентификатора китайских станций. Также присутствует возможность расширения охвата за счёт дополнительных справочников, в частности списка, опубликованного компанией БЕСТ-ТРАНСЛИНК [39], включающего более 800 станций. Каждая запись в этом списке содержит русское и английское наименования, а

также координаты, что может позволить выполнить сопоставление между языковыми версиями и включить эти данные в общее внутреннее хранилище.

Таким образом, внедрение унификации железнодорожных станций на основе национальных систем кодирования позволит расширить охват маршрутов, включающих наземные участки, и обеспечить корректную географическую привязку локаций в рамках общего подхода к обработке логистических данных.

## **4.2. Архитектура хранилища**

Для поддержки унификации и дальнейшей обработки геоданных была разработана структура внутреннего хранилища (рис. 4), обеспечивающая хранение нормализованной информации о локациях, связанных с контейнерными перевозками. Хранилище реализовано в виде реляционной базы данных и включает несколько взаимосвязанных таблиц, основными из которых являются `locations`, `countries` и `regions`.

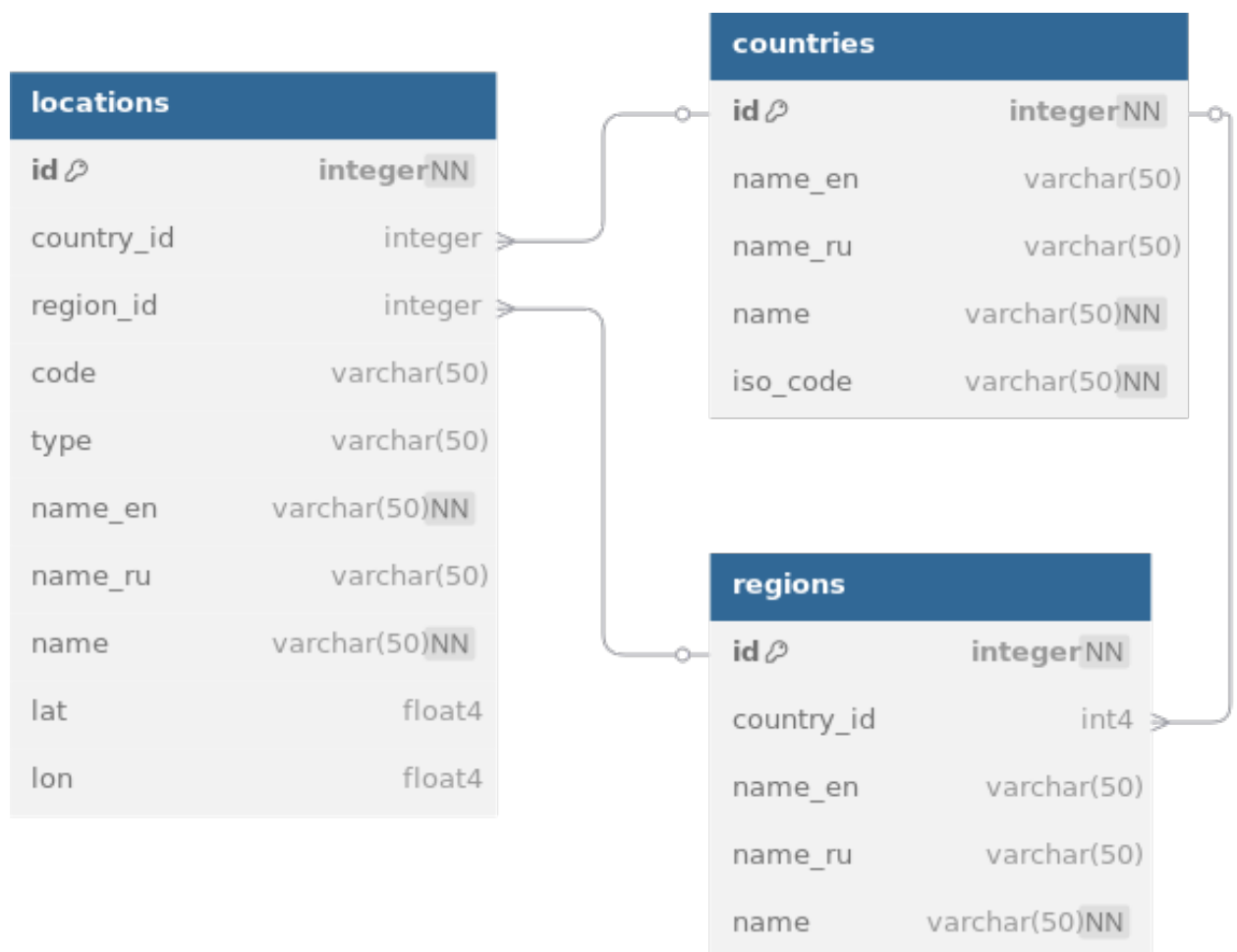


Рис. 4: Архитектура хранилища локаций

Центральную роль в архитектуре играет таблица `locations`, содержащая сведения о конкретных пунктах маршрута: портах и железнодорожных станциях. Каждая запись может содержать поля из перечня, представленного ниже.

- `code` — специальный код локации. Для портов это UN/LOCODE, для станций на территории СНГ и Балтии — код ЕСП, для станций в Китае — пятизначный код по стандарту GB/T 10302-2010.
- `type` — тип локации, определяющий её категорию. На текущем этапе поддерживаются два значения: `port` и `station`.
- `lat`, `lon` — координаты (широта и долгота).
- Внешние ключи `country_id` и `region_id` указывают на соответствующие записи в таблицах `countries` и `regions`. Хранение обо-

их полей приводит к некоторой избыточности: страну можно определить через регион. Однако такое решение допустимо в рамках модели, поскольку упрощает выполнение запросов и позволяет корректно обрабатывать случаи, когда регион не указан (без `country_id` страна в таком случае не может быть установлена). Согласованность между значениями поддерживается на уровне бизнес-логики приложения.

Во всех трёх таблицах (`locations`, `countries`, `regions`) используется одинаковая схема для хранения наименований: `name` — основное название, `name_ru` — русская версия, `name_en` — английская. Поле `name` обязательно и используется по умолчанию при отсутствии перевода.

Таблица `countries` также содержит поле `iso_code` — трёхбуквенный код страны по стандарту ISO 3166-1 ALPHA-3, используемый для унифицированной идентификации страны в маршрутах и фильтрации данных.

В дальнейшем, для поддержки локализации интерфейса на другие языки, структура хранилища может быть расширена за счёт выноса переводов в отдельные таблицы. Вместо хранения полей `name`, `name_ru`, `name_en` в основной структуре, планируется использование сущностей `*_translations`, содержащих переводы на разные языки в табличной форме. Каждая запись в таблицах перевода содержит локализованное значение (`value`), код языка (`lang`) и, при необходимости, указание на вариант по умолчанию (`is_default`). Такой подход упростит расширение языковой поддержки, избежав усложнения схемы при добавлении новых языков. Предполагаемая структура локализационных таблиц показана на диаграмме (рис. 5).



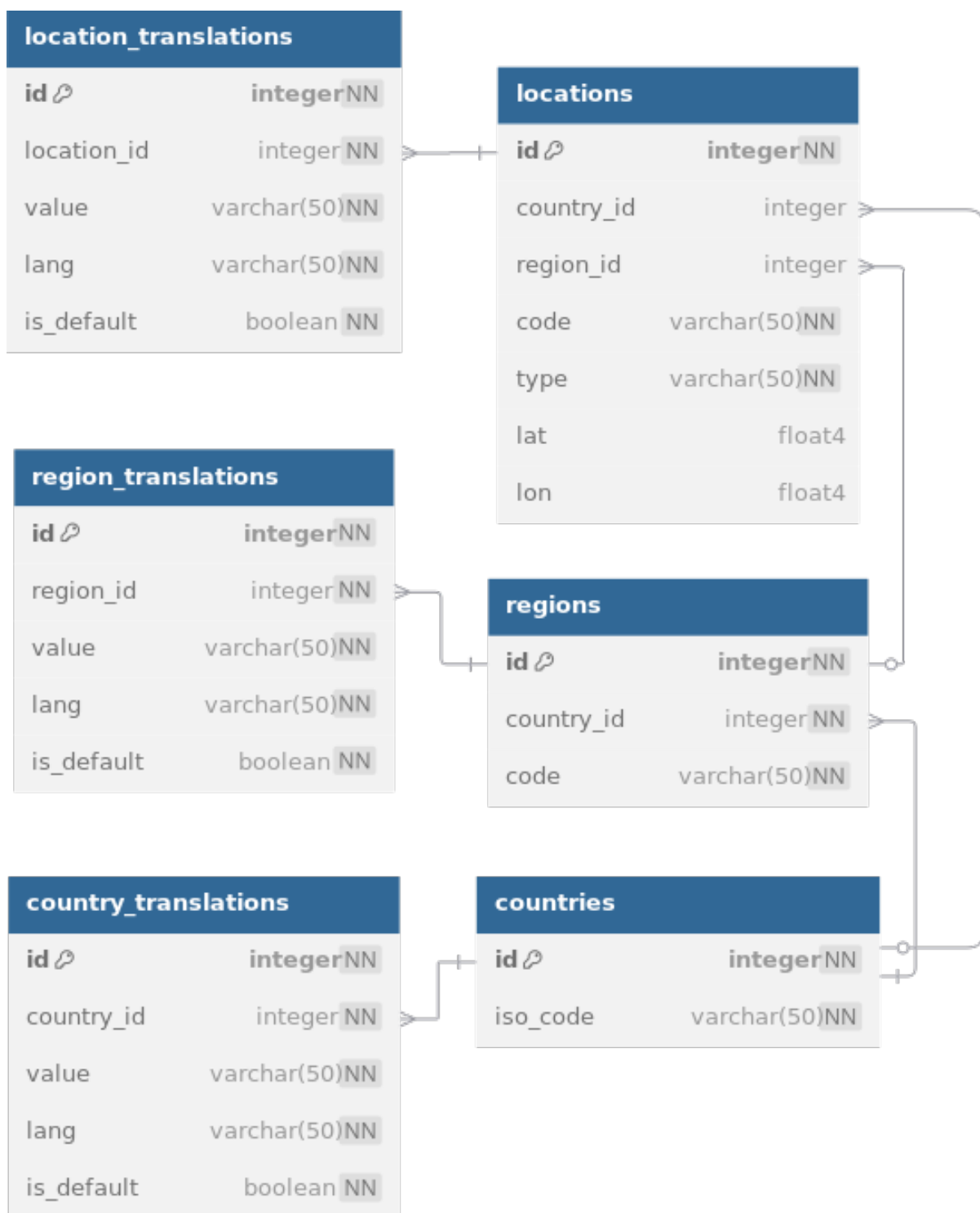


Рис. 5: Архитектура хранилища локаций (локализация)

### 4.3. Методология оценки покрытия локаций

Одними из ключевых показателей качества предоставляемой информации о локациях контейнерной перевозки, являются полнота и кор-

ректность географической информации, необходимой для унификации маршрутов контейнерных перевозок и последующего использования данных в аналитических и предиктивных задачах. Для их количественной оценки используются метрики покрытия — доли определенных локаций в ответе компании, которые удалось унифицировать. В работе используются две специализированные метрики покрытия географических данных, различающиеся по принципу включения локаций в расчет показателей в зависимости от их роли в маршруте. Расчёт показателей осуществляется на основе логов, формируемых парсерами при обработке ответов от транспортных компаний. В процессе работы каждая локация проверяется на наличие координат: если координаты успешно найдены, локация считается покрытой (унифицированной).

**Покрытие summary-локаций** рассчитывается как доля перевозок, в которых все начальные и конечные точки маршрута (например, пункт отправления, порт погрузки, порт выгрузки, пункт назначения), указанные в ответе транспортной компании, унифицированы. Эта метрика позволяет оценить полноту информации о критических узлах маршрута, без которых невозможно построение базовой структуры перевозки. В ряде перевозок начальная и конечная точки могут быть единственными известными локациями, поэтому обеспечение их покрытия особенно важно для поддержания минимального качества данных.

**Покрытие port-локаций** определяется аналогичным образом: как доля перевозок, в которых все локации, связанные с событиями в морских портах (погрузка, выгрузка, прибытие в промежуточный порт), унифицированы. В отличие от summary-метрики, сюда включаются все локации, относящиеся к событиям движения через порты, вне зависимости от их положения в маршруте (начало, середина или конец). Эта метрика особенно важна в случаях, когда транспортная компания не предоставляет отдельные поля для summary-локаций: тогда маршрут приходится восстанавливать по статусам событий. Кроме того, данная метрика позволяет оценить полноту данных о морской части перевозки, что критично для построения траектории движения контейнера по

моря.

Выбор метрик покрытия был обусловлен спецификой задач, решаемых в рамках проекта. Поскольку разработка пайплайна стандартизации и обогащения направлена на обеспечение качества данных для последующего построения аналитических и предсказательных решений, особое внимание уделено ключевым точкам маршрута: пункту отправления, портам погрузки, портам выгрузки и пункту назначения, а также промежуточным морским портам. Именно эти локации формируют основу структуры маршрута и оказывают прямое влияние на расчет времени доставки и интерпретацию статусов.

Для апробации предложенного подхода к унификации локаций расчет метрик проводился на уникальных 100 перевозках для каждой компании по отдельности. Для получения итоговых значений метрик использовалось средневзвешенное значение, где в качестве весов компаний использовалась доля трафика в общем объеме запросов. Такой подход позволяет объективно учесть реальное влияние каждой компании на полноту данных в системе и дает более репрезентативную оценку качества унификации с учетом фактического распределения трафика между перевозчиками.

## 4.4. Результаты

Проведённая апробация показала, что переход от подхода с использованием внешнего картографического сервиса BINGMAPS (исходный показатель) к внутреннему справочнику локаций (полученный показатель) привёл к повышению степени полноты географических данных (см. таблицу 2).

Метрика	Исходный показатель	Полученный показатель	Прирост
Покрытие summary-локаций	76,94%	92,49%	+15,55%
Покрытие port-локаций	78,03%	88,73%	+10,70%

Таблица 2: Прирост показателей покрытия локаций

Анализ полученных результатов по отдельным компаниям показал,

что при наличии корректно структурированных ответов транспортных компаний предложенный подход обеспечивает практически идеальную полноту данных. У популярных профильных морских линий (MSC, MAERSK, HAPAG-LLOYD, которые суммарно составляют 36,33% от всего трафика портала по внутренним данным за апрель 2025 года) и железнодорожных операторов (ТРАНСКОНТЕЙНЕР, который составляет 31,32%), где в ответах транспортной компании содержатся преимущественно только релевантные логистические объекты (порты и станции), показатели покрытия стабильно превышают 95%, а в отдельных случаях достигают значений, близких к 100%. Эти результаты подтверждают, что при адекватной структуре исходных данных унификация через внутренний справочник позволяет получать практически полное покрытие маршрута без необходимости ручных доработок или дополнительных источников.

Кроме того, новый подход предусматривает построение внутреннего справочника, изначально заполненного только верифицированными логистическими объектами — морскими портами и железнодорожными станциями. Таким образом, если в процессе обработки данных перевозчика происходит сопоставление локации с записью в справочнике, можно быть уверенным, что это действительно релевантный для маршрута объект, чего нельзя было гарантировать в предыдущем подходе. Это решение устраняет риск случайных сопоставлений и создаёт надёжную основу для унификации локаций в ответе сервиса и дальнейшей аналитической обработки маршрутов.

## 5. Инструменты унификации статусов перевозки

Глава описывает программные компоненты, разработанные для поддержки процесса сопоставления статусов перевозки с унифицированной системой кодов: инструмент визуализации переходов между статусами и систему управления соответствиями.

### 5.1. Визуализация переходов между статусами

Для реализации аналитических инструментов, работающих на основе логов, был создан отдельный микросервис ELASTIC UTILS. Он обеспечивает доступ к данным, хранящимся в ELASTICSEARCH, и предоставляет API-интерфейсы для агрегации, фильтрации и визуализации информации, извлекаемой из логов парсеров. Такой подход позволил изолировать вспомогательные задачи анализа от основной бизнес-логики системы и упростить расширение функциональности. Часть архитектуры сервиса трекинга после внедрения сервиса ELASTIC UTILS выглядит, как представлено на рисунке ниже.

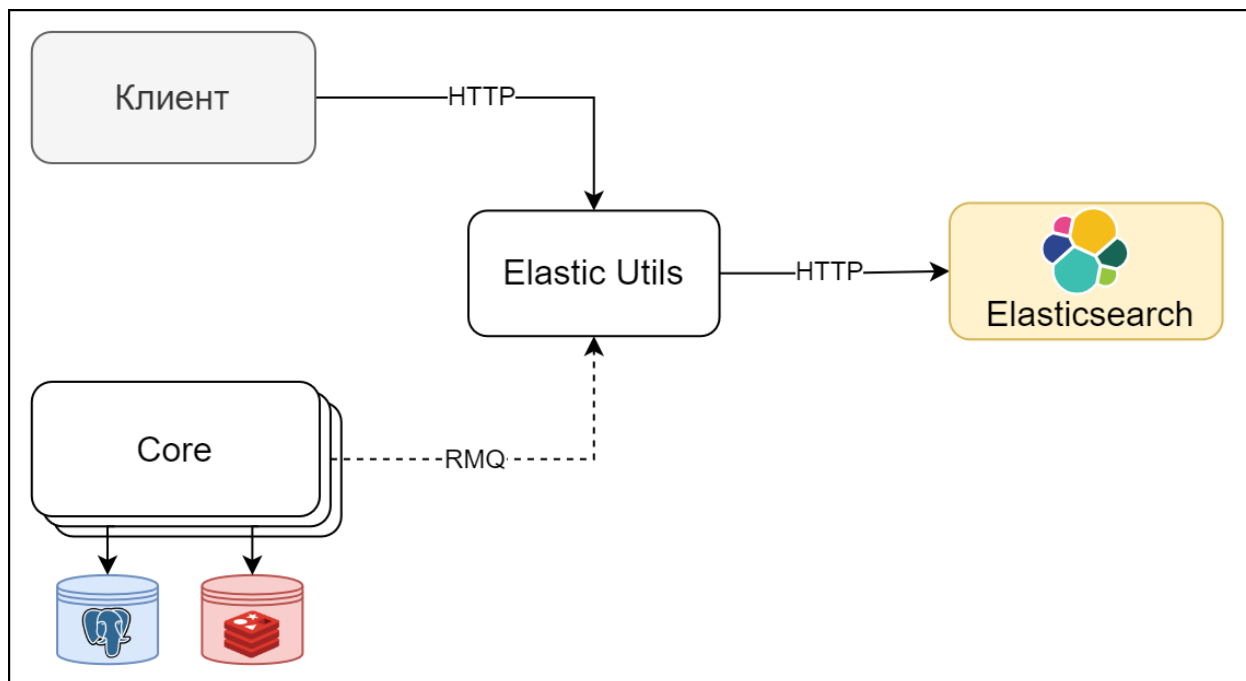


Рис. 6: Внедрение Elastic Utils

В рамках микросервиса ELASTIC UTILS был реализован инструмент визуализации переходов между статусами перевозки. Этот инструмент позволяет на основе логов парсеров построить ориентированный граф, отображающий фактические последовательности событий, встречающиеся в маршрутах контейнеров конкретной транспортной компании.

Через API-интерфейс сервиса можно задать название компании и объём выборки — количество последних логов, которые следует проанализировать. На основании извлекаемых данных формируется агрегированная структура переходов между текстовыми статусами. Для каждого маршрута определяется упорядоченная последовательность событий, из которой выделяются пары переходов. Далее по всей выборке подсчитываются частоты этих переходов и формируется ориентированный граф.

В графе:

- узлы представляют собой уникальные текстовые статусы;
- рёбра отражают возможные переходы между ними;
- вес рёбер соответствует частоте появления соответствующего перехода.

Построение и визуализация графа реализованы с использованием формата DOT [8] — текстового языка описания графов, применяемого в системе визуализации GRAPHVIZ [14]. Для генерации изображения в формате SVG используется библиотека VIZ.JS [36], которая представляет собой WEBASSEMBLY-порт GRAPHVIZ и позволяет выполнять отрисовку на стороне сервера без необходимости установки внешних утилит. Сформированный граф может быть получен через API-эндпоинт `/tracking/raw-state-graph/:company` (в формате DOT) или `/tracking/raw-state-graph/:company/svg` (в формате SVG). Оба эндпоинта поддерживают параметр `size`, определяющий объём выборки логов. Такой подход обеспечивает простую интеграцию инструмента с внешними системами и позволяет использовать полученные диаграммы

как для визуального анализа, так и для последующей автоматической обработки.

Чтобы упростить восприятие диаграммы:

- начальный и конечный узлы (`begin` и `end`) выделяются цветом и формой;
- толщина рёбер масштабируется пропорционально их весу;
- цветовая кодировка отражает позицию перехода в маршруте: зелёный — для начальных, красный — для завершающих, и оттенки серого (в зависимости от веса ребра) — для остальных.

Пример такой диаграммы, построенной для компании FESCO на основе 1000 последних запросов, представлен на рисунке 7.

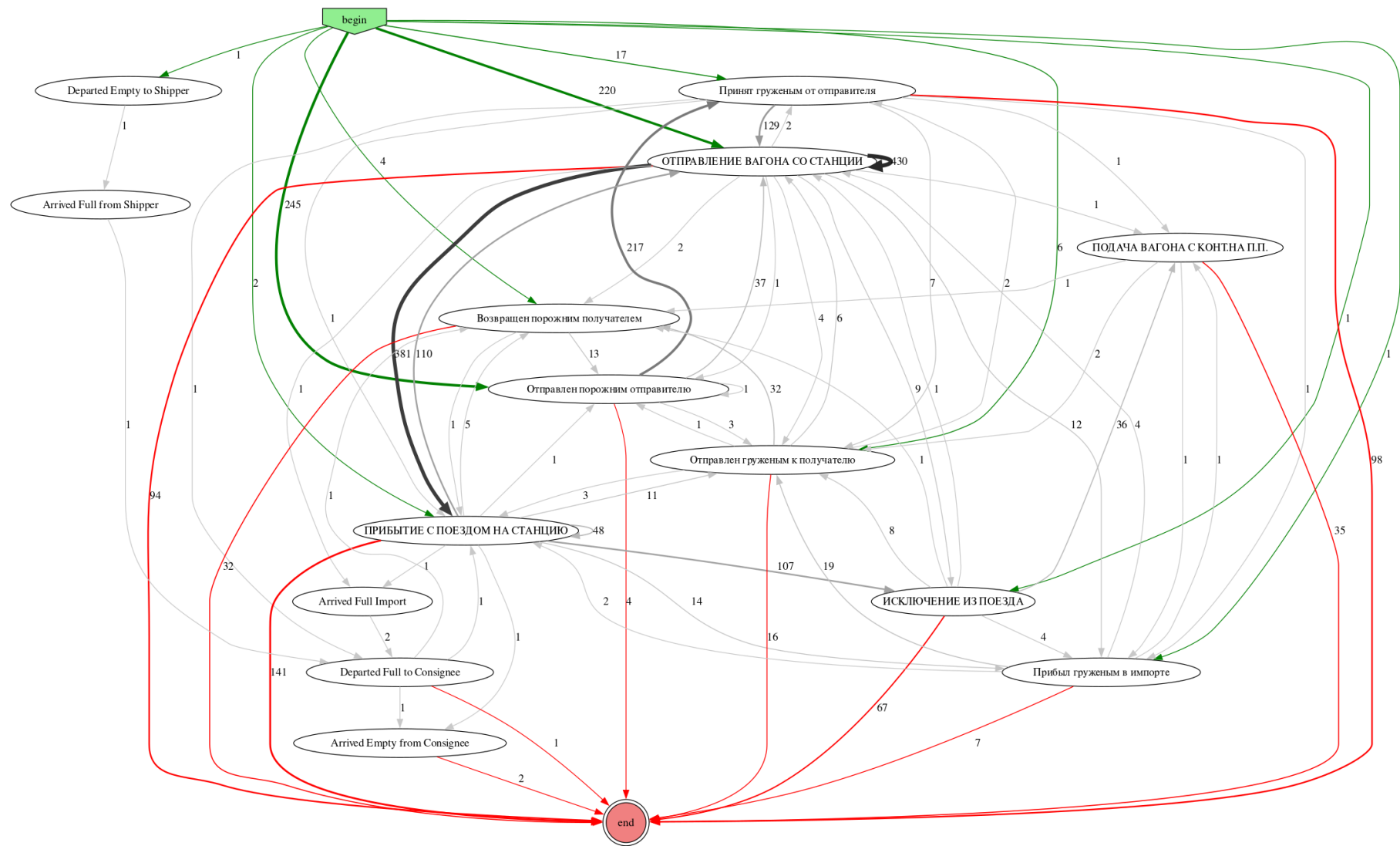


Рис. 7: Диаграмма переходов статусов FESCO



Таким образом, реализованный инструмент визуализации позволяет наглядно представить структуру маршрутов, характерных для конкретного перевозчика, и служит вспомогательным средством при интерпретации статусной информации. Использование графа переходов упрощает принятие экспертных решений при сопоставлении текстовых статусов с унифицированной системой кодов.

## 5.2. Система управления унификацией статусов

Для обеспечения возможности сопоставления текстовых статусов, поступающих от различных транспортных компаний, с внутренней системой кодов был реализован специализированный компонент в составе микросервиса CORE. Размещение хранилища сопоставлений именно в этом сервисе обусловлено тем, что парсеры, обрабатывающие ответы от компаний и выполняющие приведение статусов к унифицированной форме, также находятся внутри CORE и напрямую используют эти данные в процессе работы.

Хранилище реализовано в базе данных микросервиса в виде отдельной таблицы `status_mappings`. Таблица содержит следующие поля:






















- `id` — уникальный идентификатор записи;
- `company_id` — внешний ключ на таблицу компаний;
- `status` — текстовое значение статуса, полученное от компании;
- `status_code` — соответствующий унифицированный код из внутренней классификации.

Для работы с таблицей `status_mappings` реализован CRUD-интерфейс, позволяющий просматривать, добавлять, редактировать и удалять записи. Это необходимо для поддержки актуальности сопоставлений, так как статусная информация у компаний может со временем меняться или дополняться.

Для удобства работы эксперта с системой унификации статусов был также реализован пользовательский интерфейс на NEXT.JS. Он представляет собой интерактивную таблицу с поддержкой фильтрации, сортировки, пагинации, добавления записей, а также их построчного редактирования и удаления. Интерфейс представлен на рисунке ниже.

Статусы компаний

+ Добавить

ID	Компания	Статус	Код	Действия
	Компания Все	Статус	Код Все	
1	FESCO	Отправлен порожним отправителю	CGO	 
2	FESCO	ОТПРАВЛЕНИЕ ВАГОНА СО СТАНЦИИ	CDC	 
3	FESCO	ПРИБЫТИЕ С ПОЕЗДОМ НА СТАНЦИЮ	VDL	 
4	FESCO	ИСКЛЮЧЕНИЕ ИЗ ПОЕЗДА	UNK	 
5	FESCO	Прибыл груженым в импорте	CGO	 
6	FESCO	ПОДАЧА ВАГОНА С КОНТ.НА П.П.	CDC	 
7	MAERSK	GATE-OUT	VDL	 
8	MAERSK	GATE-IN	UNK	 
9	MAERSK	CONTAINER ARRIVAL	CGO	 
10	MAERSK	RAIL_DEPARTURE	CDC	 

Показывать по: 10 1-10 of 10 < >

Рис. 8: Интерфейс системы управления унификацией статусов

Таким образом, разработанная система управления унификацией статусов обеспечивает централизованное хранение, удобное редактирование и высокую гибкость в поддержке актуальности статусной информации, что является необходимым условием для корректной интерпретации событий перевозки.

## 6. Оптимизация парсинга данных о судах

Данная глава описывает подходы, реализованные с целью повышения скорости получения информации о судах из внешних источников. В рамках оптимизации сервиса MARINE были применены два ключевых метода: асинхронная обработка запросов с управлением степенью нагрузки и фоновая актуализация наиболее востребованных данных с локальным кешированием.

### 6.1. Фоновая актуализация и кеширование

Для ускорения доступа к информации о судах и снижения нагрузки на внешние источники в сервисе MARINE была реализована система фоновой актуализации и кеширования данных. Вместо классического подхода, при котором данные о судне запрашиваются в момент обращения пользователя, сервис сохраняет информацию о наиболее часто используемых судах в собственной базе данных и периодически обновляет её в фоновом режиме.

Пользователи портала запрашивают информацию о контейнерах. В процессе обработки таких запросов микросервис CORE обращается к транспортным компаниям, и если в ответе содержится наименование судна, оно фиксируется в логах. Сервис MARINE с заданным интервалом анализирует эти логи, подсчитывает количество упоминаний каждого судна и определяет список наиболее часто встречающихся.

Актуальная информация о судах сохраняется в таблице `container_ships`. Ниже приведена её упрощённая структура, отражающая ключевые поля, используемые в логике фонового обновления:

- `id` — уникальный идентификатор записи;
- `name` — наименование судна;
- `rating` — рейтинг судна (число упоминаний в логах);
- `lat, lon` — координаты текущего местоположения;

- `last_checked` — дата и время последней проверки логов на предмет упоминания судна;
- `last_updated` — дата и время последнего успешного обновления информации о судне;
- `info_source` — наименование источника, откуда была получена информация.

Алгоритм работы включает перечисленные ниже шаги.

1. **Анализ логов.** Сервис периодически анализирует логи и обновляет поле `rating` в таблице `container_ships`. Также фиксируется время обновления статистики в поле `last_checked`.
2. **Выбор судов для обновления.** Из таблицы отбираются записи с рейтингом выше заданного порога. Исключаются записи, которые были недавно обновлены (`last_updated`), либо не были найдены в логах при последних проверках (`last_checked`).
3. **Фоновое обновление.** Для выбранных судов выполняется парсинг внешнего источника. Полученные данные сохраняются в поля `lat`, `lon`, `info_source` и `last_updated`.

Такой подход обеспечивает быстрый доступ к данным о судах при повторных обращениях, снижает количество внешних запросов и повышает устойчивость системы к временной недоступности источников.

## 6.2. Многопоточная обработка запросов

Вторая ключевая оптимизация сервиса MARINE связана с ускорением обработки пакетных запросов — случаев, когда необходимо получить информацию сразу о нескольких судах. Такая ситуация возникает, например, при запросе большого числа контейнеров, оказавшихся на разных судах, либо в рамках фоновой актуализации, когда требуется массово обновить наиболее популярные суда. В этих сценариях последовательный парсинг приводит к значительным задержкам.

Сервис MARINE уже поддерживает репликацию, что позволяет распределять нагрузку при большом количестве одиночных запросов. Однако при пакетной обработке, когда вся нагрузка сосредотачивается в одном экземпляре, этого оказывается недостаточно. Для таких случаев была реализована внутренняя многопоточная обработка, позволяющая эффективно использовать ресурсы машины за счёт параллелизма в браузерном движке CHROMIUM.

В основе этой реализации лежит библиотека PLAYWRIGHT, позволяющая программно управлять экземплярами браузера и создавать отдельные вкладки для парсинга сайтов. Несмотря на то, что NODE.JS является однопоточной средой исполнения, каждая вкладка CHROMIUM функционирует в отдельном потоке на уровне операционной системы. Таким образом, при корректной организации запросов удаётся задействовать параллелизм браузерного движка, обеспечивая прирост производительности при I/O-bound операциях: NODE.JS асинхронно запускает задачи, а их фактическое выполнение распределяется между потоками браузера. На рисунке ниже проиллюстрирована схема такого взаимодействия:

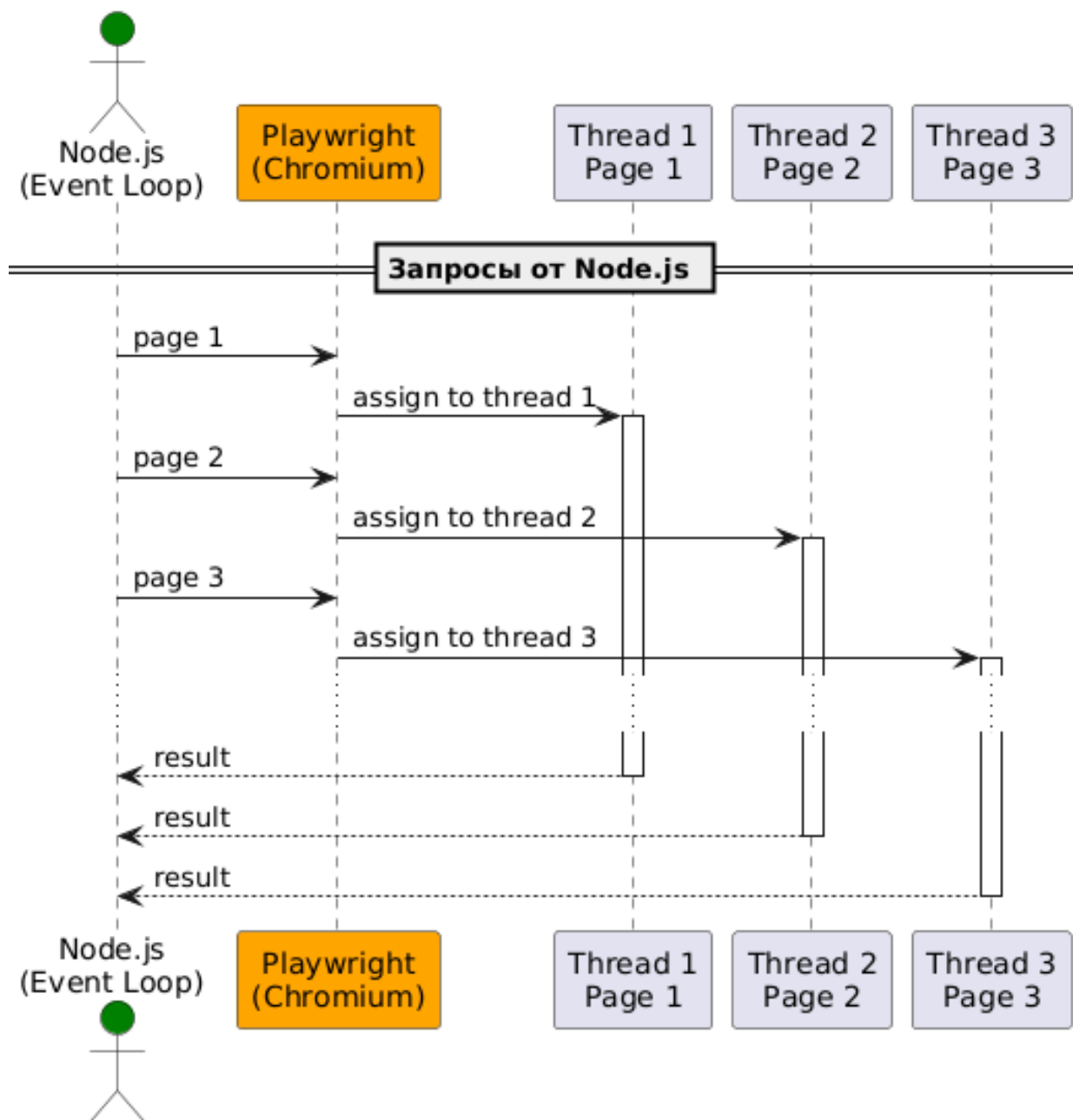


Рис. 9: Асинхронное взаимодействие с PLAYWRIGHT

При использовании многопоточности необходимо учитывать риски, связанные с высокой частотой запросов, в частности вероятность появления САРТСНА [2] на стороне целевых сайтов. Чтобы избежать блокировок и контролировать степень нагрузки, был реализован механизм ограничения количества одновременных задач и обработки ошибок.

Реализация представлена в виде метода `mapConcurrently()`, который асинхронно обрабатывает массив элементов и принимает в качестве параметров:

- `items` — массив элементов, подлежащих асинхронной обработке;
- `mapper` — вызываемая функция-обработчик;
- `concurrent` — ”степень асинхронности” (максимальное число асинхронных вызовов);
- `timeout` — лимит по времени на работу функции;
- `retryCount` — количество повторных попыток вызова функции.

Такой подход позволяет контролировать интенсивность нагрузки на внешние источники, обеспечивая стабильную и предсказуемую работу сервиса при массовой обработке.

### 6.3. Результаты

На момент написания работы экспериментальная оценка была проведена только для компонента многопоточной обработки. Измерение эффективности фоновой актуализации и кеширования потребует накопления статистики использования и анализа долгосрочного поведения системы, что запланировано на следующем этапе разработки.

Для оценки влияния многопоточности на производительность был проведён эксперимент, в котором сравнивалось время выполнения парсинга при синхронной и асинхронной обработках запросов. Тестирование проводилось на страницах с данными портов: каждый запрос содержал случайный набор портов из заранее подготовленного списка. Измерения выполнялись для различных объёмов входных данных — от 6 до 192 портов в одном запросе. Для каждого объёма проводилось по 20 повторов, после чего рассчитывались средние значения времени отклика и стандартное отклонение.

Программно-аппаратная конфигурация приведена в таблице ниже.

<b>Процессор</b>	Intel Core i7-8750H CPU 2.20GHz (Coffee Lake), 1 CPU, 12 logical and 6 physical cores
<b>Объём ОЗУ</b>	16 GB
<b>Операционная система</b>	Windows 11 23H2
<b>Версия Node.js</b>	22.9.0
<b>Версия Playwright</b>	1.43.1

Таблица 3: Конфигурация

Эксперимент проводился с использованием инструмента нагрузочного тестирования k6 [13], имитирующего последовательную отправку запросов к API сервиса (1 виртуальный пользователь). Влияние качества интернет-соединения в эксперименте не учитывалось, поскольку время отклика страниц и объём передаваемого трафика незначительны относительно затрат на саму обработку страниц. В ходе предварительных измерений было установлено, что оптимальными значениями параметров для заданной конфигурации являются следующие величины:

- “степень асинхронности” – 6;
- кол-во повторных попыток – 5;
- таймаут – 30 с.

Именно с данными значениями проводились основные замеры.

Результаты представлены на графиках ниже (рис. 10–11). Первый график отображает зависимости в линейной шкале, второй — с логарифмической шкалой по оси X. Закрашенные области вокруг кривых соответствуют стандартному отклонению и отражают разброс значений между повторами.



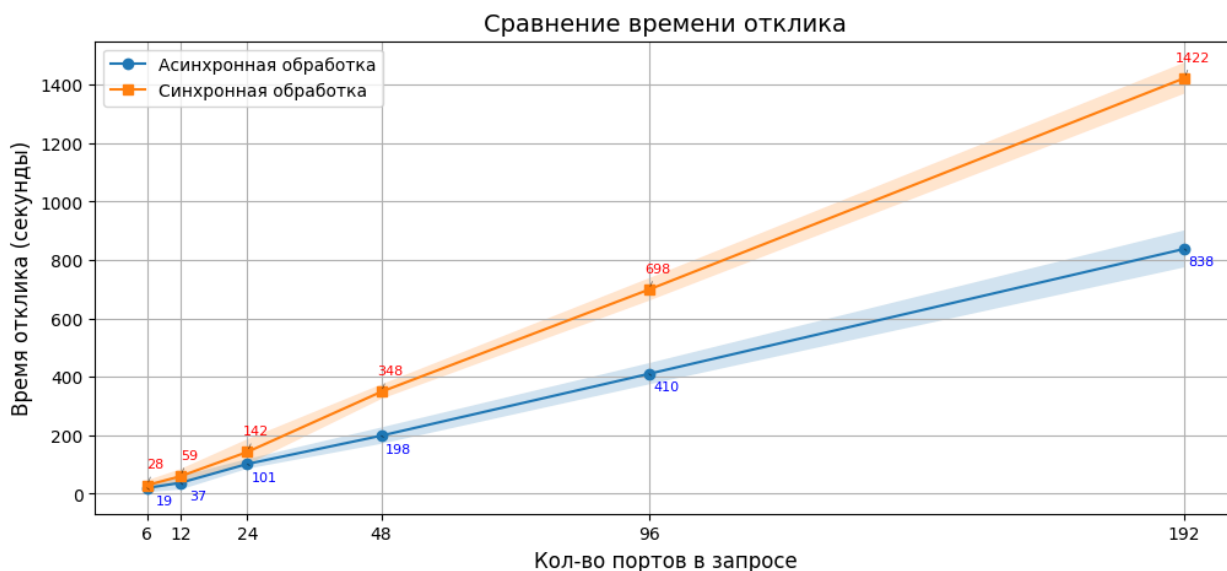


Рис. 10: Сравнение времени отклика сервиса MARINE

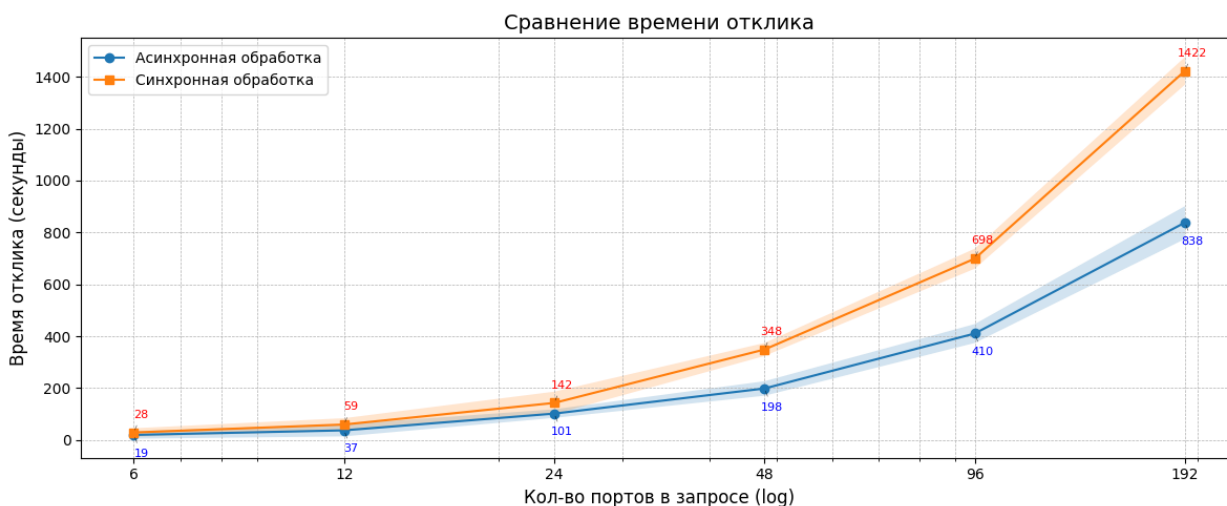


Рис. 11: Сравнение времени отклика сервиса MARINE (логарифмическая шкала)

На графиках видно, что асинхронная реализация демонстрирует стабильное преимущество по времени отклика по сравнению с синхронной. Среднее ускорение составило около 1,7 раза. При многопоточной обработке могло бы ожидатьсся более заметное сокращение времени, однако достигнутый результат остаётся обоснованным с учётом внешних ограничений. В частности, при параллельной загрузке страниц САРТСНА-защита на стороне целевого сайта срабатывала существенно чаще: практически каждая страница требовала 2–3 повторных попыток. В син-

хронном режиме такие ситуации возникали единично, что обеспечивало более равномерную обработку. Это увеличивало совокупное время выполнения в асинхронном варианте и объясняет, почему ускорение оказалось ниже потенциально достижимого при идеальных условиях.

Несмотря на указанные ограничения, полученные результаты подтверждают эффективность выбранного подхода. Даже в условиях повышенной нагрузки асинхронная обработка даёт стабильный прирост производительности. Дополнительные меры — такие как интервалы между задачами, ротация IP-адресов или внедрение случайных задержек — могут позволить улучшить устойчивость и сократить число повторов.

Также стоит отметить универсальность данного подхода в том смысле, что его можно применять не только для получения информации от парсеров портов и судов, но также и для массовых запросов к любым другим парсерам, реализованным с помощью библиотеки `PLAYWRIGHT`. Это делает его подходящим для задач фоновой актуализации, обработки пакетных запросов и других операций с большим числом однотипных обращений.

## Заключение

Целью работы являлось повышение качества предоставляемой информации о контейнерных перевозках и скорости доступа к ней путём разработки и внедрения пайплайна автоматизированной обработки, стандартизации и обогащения логистических данных из различных источников.

В ходе работы были выполнены все поставленные для достижения цели задачи, перечисленные ниже.

- Разработан механизм стандартизации и обогащения сырых данных о локациях контейнерной перевозки, использующий локальное хранилище данных. Покрытие summary- и port-локаций увеличено на 15,55% и 10,7% соответственно.
- Разработан интерактивный инструмент для экспертного сопоставления разнородных статусов перевозки с унифицированной системой кодов.
- Обеспечен высокоскоростной доступ к данным о судах путем внедрения фоновой актуализации и многопоточной обработки запросов. Время отклика сервиса сократилось в среднем в 1,7 раза.
- Разработанный пайплайн обработки и стандартизации сырых данных внедрён в портал CARGOTIME.

Качество предоставляемой информации о контейнерных перевозках было повышено за счет увеличения покрытия возможных локаций и минимизации риска экспертной ошибки при сопоставлении статусов перевозки. Скорость доступа к этой информации была повышена благодаря сокращению времени отклика сервиса парсинга данных о судах.

Полученные результаты могут стать основой для дальнейших исследований, направленных на разработку модели прогнозирования сроков доставки контейнеров на основе унифицированных логистических данных. Это позволит повысить практическую ценность системы и расширить её функциональность.

## Список литературы

- [1] Bing Maps Documentation. — URL: <https://learn.microsoft.com/en-us/bingmaps/> (дата обращения: 17 марта 2025 г.).
- [2] CAPTCHA. — URL: <http://captcha.net/> (дата обращения: 25 декабря 2024 г.).
- [3] China Railway Map. — URL: <http://cnrail.geogv.org/enus/about> (дата обращения: 6 апреля 2025 г.).
- [4] Chromium. — URL: <https://www.chromium.org/Home/> (дата обращения: 25 декабря 2024 г.).
- [5] Container Tracking. — URL: <https://www.visiwise.co/tracking/container/> (дата обращения: 1 октября 2024 г.).
- [6] DCSA Standards. — URL: <https://dcsa.org/standards> (дата обращения: 14 марта 2025 г.).
- [7] DCSA Track Trace Standard documentation. — URL: <https://dcsa.org/standards/track-and-trace/standard-documentation-track-and-trace> (дата обращения: 14 марта 2025 г.).
- [8] DOT Language. — URL: <https://graphviz.org/doc/info/lang.html> (дата обращения: 15 марта 2025 г.).
- [9] Docker. — URL: <https://www.docker.com/> (дата обращения: 25 декабря 2024 г.).
- [10] ELK Stack. — URL: <https://www.elastic.co/elastic-stack> (дата обращения: 25 декабря 2024 г.).
- [11] English GB/T 10302-2010 PDF. — URL: <https://www.chinesestandard.net/PDF/English.aspx/GBT10302-2010> (дата обращения: 6 апреля 2025 г.).

- [12] Find the best Freight Quote. — URL: <https://www.searates.com/> (дата обращения: 1 октября 2024 г.).
- [13] Grafana k6. — URL: <https://k6.io/> (дата обращения: 21 апреля 2025 г.).
- [14] Graphviz. — URL: <https://graphviz.org/> (дата обращения: 15 марта 2025 г.).
- [15] Hapag most reliable top-13 carrier in June 2024: Sea-Intelligence. — URL: <https://www.itln.in/shipping/hapag-most-reliable-top-13-carrier-in-june-2024-sea-intelligence> (дата обращения: 1 октября 2024 г.).
- [16] Impacts of a Tracking and Tracing System for Containers in a Port-Based Supply Chain / Jesús Muñuzuri, Luis Onieva, Alejandro Escudero, Pablo Cortés // [Brazilian Journal of Operations and Production Management](#). — 2016. — . — Vol. 13, no. 3. — P. 352–359. — URL: <http://dx.doi.org/10.14488/BJOPM.2016.v13.n3.a12>.
- [17] MarineTraffic: Global Ship Tracking Tracking Intelligence. — URL: <https://www.marinetraffic.com/> (дата обращения: 25 декабря 2024 г.).
- [18] Meijer Ricardo. Predicting the ETA of a Container Vessel Based on Route Identification Using AIS Data : Master's thesis / Ricardo Meijer ; Delft University of Technology. — Delft, The Netherlands, 2017. — . — URL: <https://resolver.tudelft.nl/uuid:cba0ef59-dd23-49aa-91d5-bed239e27395>.
- [19] NestJS. — URL: <https://nestjs.com/> (дата обращения: 25 декабря 2024 г.).
- [20] Next.js by Vercel. — URL: <https://nextjs.org/> (дата обращения: 17 марта 2025 г.).
- [21] Node.js. — URL: <https://nodejs.org/en> (дата обращения: 25 декабря 2024 г.).

- [22] Online Container Tracking - Your GoComet Portal. — URL: <https://www.gocomet.com/online-container-tracking> (дата обращения: 1 октября 2024 г.).
- [23] Park Kikun, Sim Sunghyun, Bae Hyerim. Vessel estimated time of arrival prediction system based on a path-finding algorithm // *Maritime Transport Research*. — 2021. — Vol. 2. — P. 100012. — URL: <https://www.sciencedirect.com/science/article/pii/S2666822X21000046>.
- [24] Parolas Ioannis. ETA prediction for containerships at the Port of Rotterdam using Machine Learning Techniques : Master's thesis / Ioannis Parolas ; Delft University of Technology. — Delft, The Netherlands, 2016. — . — URL: <https://resolver.tudelft.nl/uuid:9e95d11f-35ba-4a12-8b34-d137c0a4261d>.
- [25] Playwright. — URL: <https://playwright.dev/> (дата обращения: 25 декабря 2024 г.).
- [26] PostgreSQL. — URL: <https://www.postgresql.org/> (дата обращения: 14 марта 2025 г.).
- [27] Puppeteer. — URL: <https://pptr.dev/> (дата обращения: 25 декабря 2024 г.).
- [28] RabbitMQ. — URL: <https://www.rabbitmq.com/> (дата обращения: 25 декабря 2024 г.).
- [29] React. — URL: <https://react.dev/> (дата обращения: 17 марта 2025 г.).
- [30] Redis. — URL: <https://redis.io/> (дата обращения: 25 декабря 2024 г.).
- [31] Review of Maritime Transport. — URL: <https://unctad.org/publication/review-maritime-transport-2024> (дата обращения: 17 марта 2025 г.).

- [32] Ship Container Tracking - VesselFinder. — URL: <https://www.vesselfinder.com/> (дата обращения: 25 декабря 2024 г.).
- [33] Tracking and Tracing of Global Supply Chain Network: Case Study from a Finnish Company / Ahm Shamsuzzoha, Michael Ehlers, Richard Addo-Tengkorang, Petri Helo // Proceedings of the 23rd International Conference on Enterprise Information Systems. — SCITEPRESS - Science and Technology Publications, 2021. — P. 118–125. — URL: <http://dx.doi.org/10.5220/0010515401180125>.
- [34] TypeScript. — URL: <https://www.typescriptlang.org/> (дата обращения: 25 декабря 2024 г.).
- [35] UN/LOCODE Code List by Country and Territory. — URL: <https://unece.org/trade/cefact/unlocode-code-list-country-and-territory> (дата обращения: 6 апреля 2025 г.).
- [36] Viz.js. — URL: <https://viz-js.com/> (дата обращения: 15 марта 2025 г.).
- [37] Yue Zhongyun, Mangan John. A framework for understanding reliability in container shipping networks // *Maritime Economics and Logistics*. — 2023. — . — Vol. 26, no. 3. — P. 523–544. — URL: <http://dx.doi.org/10.1057/s41278-023-00269-7>.
- [38] Единая сетевая разметка. — URL: <https://osm.sbin.ru/esr/> (дата обращения: 8 апреля 2025 г.).
- [39] Коды ЖД станций Китая – Бесттранслинк. — URL: <http://www.besttranslink.com/about//2019/0901/26.html> (дата обращения: 21 марта 2025 г.).
- [40] Логистический портал Cargotime. — URL: <https://cargotime.ru/> (дата обращения: 25 декабря 2024 г.).
- [41] Организация сотрудничества железных дорог. — URL: <https://osjd.org/> (дата обращения: 21 марта 2025 г.).

- [42] Перечень грузовых станций Китайских железных дорог. — URL: <https://osjd.org/api/media/resources/10839> (дата обращения: 21 марта 2025 г.).
- [43] Логистический портал Cargotime: реализация модуля по построению маршрута и отслеживанию судов : Отчет по учебной практике / Санкт-Петербургский государственный университет ; исполн.: Захар Сергеевич Фадеев. — Санкт-Петербург, Россия : 2023.