ИС БД Росстат

Технология обработки «больших данных» (big data) из различных источников с использованием искусственного интеллекта

Проблематика рынка

Сложность сбора и качество обработки больших данных

Отсутствие универсальных инструментов по сбору и обработке данных

Отсутствие коробочных решений с ИИ, которые можно быстро адаптировать под конкретную задачу или бизнес

Высокая стоимость разработки систем по работе с большими данными

Решение

Разработан универсальный алгоритм, который позволяет:

Обрабатывать большие данные и находить паттерны в информации

Собирать данные из различных источников: новостные сайты, чеки, блоги и т. д.

Автоматизировать процесс сбора при сокращении затрат на разработку не менее 40%.

Подключаться к различным системам заказчика быстро с помощью готовых скриптов и шаблонов

Обрабатывать данные с использованием ИИ с качественным показателем от 90%*

Универсальность технологии. Легкая адаптируемость под различные тематические области (Например, медицинская тематика, трейдинг, продуктовая и др.)

*Подтверждено классификацией чеков Росстата

Возможности продукта

01

Система собирает данные из различных источников – базы данных, сайты, блоги, новостные ресурсы, интеграции (легко настраивается в административной панели)

02

Система классифицирует и распределяет данные по категориям/ параметрам/бизнес-процессам (визуальная настройка)

03

Анализирует и визуализирует полученные данные (конструктор отчетов более 150 шаблонов).

Например, прогноз закупок для крупных ритейлеров или для успешной торговли на бирже.

04

Наш продукт – это приложение с возможностью настройки под конкретные задачи/ процессы/функции

05

Точность распределения, обработки, классификации не менее 90%.

Подтверждено Росстатом.

Компонентная схема

Бизнес приложения (front-end): Vue.js

Сервисы (back-end): PyTorch, NumPy, Spring, Elastic Enterprise Search, Pandas

Хранение данных: Ceph, Spark SQL, Elastic Enterprise Search, Redis, Prometheus

Загрузка данных: Spark, Python

Существующий корпоративных ландшафт: Zabbix, Grafana

Установка

Сохраняем конфиденциальность данных

01 Выбор цели внедрения и классификации данных →

02 Настройка приема данных из внешних и внутренних источников →

03 Запуск процесса структурирования и обработки данных →

04 Машинное обучение →

05 Разметка нейросетью →

06 Выбор модели анализа и обработки данных →

07 Запуск формирования отчетов →

08 Включение результатов анализа в бизнес-процессы организации

4.1 Выполнение разметки разметчиками и асессорами →

4.2 Построение полученных данных с помощью рейтинга разметчиков →

4.3 Проверка асессорами →

4.4 Контроль качества

↓↑

4.1.2 Проверка полученных данных

↓↑

4.1.3 Toloka

Обработка данных делится на 2 этапа

Токенизация – это разделение наименований чека на отдельные позиции, выделение единиц измерения и объема, приведение к нижнему регистру.

16633 Вода КИСЛОВОДСК.ЦЕЛЕБН.пл/б 1л

~~16633~~	вода кисловодск целебн пл/б	500	мл

Первичная классификация – это унификация единиц измерения и разметка наименований из чеков по категориям.

вода кисловодск целебн пл/б	0,5	л.

Как работает нейросеть?

Обучение нейросети производится с помощью датасетов, подготовленных разметчиками. Пример разметки категории «Бараночные изделия»:

Наименование товара: Сушка мини от Каролины с маком 1 кг / **Цена:** 206

- (•) Верно
- ( ) Не верно
- ( ) Затрудняюсь ответить

ИС БД Росстат

Описание задачи

Создать систему обработки и классификации больших данных для анализа статистики потребительских цен и статистики торговли.

Создать удобное web-приложение, для управления функциями математической обработки и контроля качества полученных данных.

Что было сделано

Внедрена система автоматизированного мониторинга данных фискальных чеков с применением алгоритмов big data.

Снижена нагрузка на отчитывающихся респондентов и сокращен ручной труд по сбору информации о ценовых котировках.

Применена технология больших данных с перспективой их использования в других направлениях статистики.

Разработан web-интерфейс для работы с ИС БД.

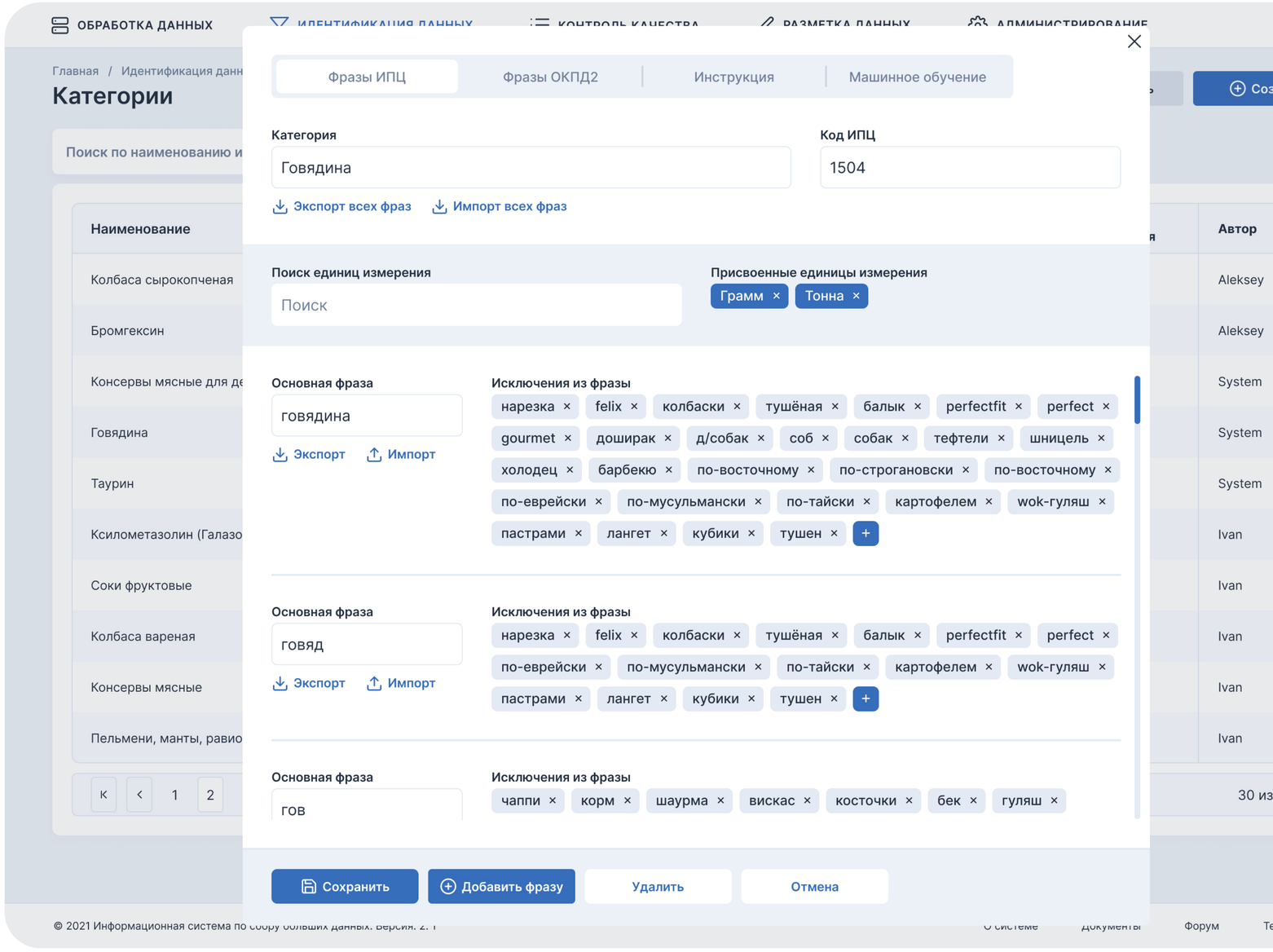
ИС БД Росстат

Результат

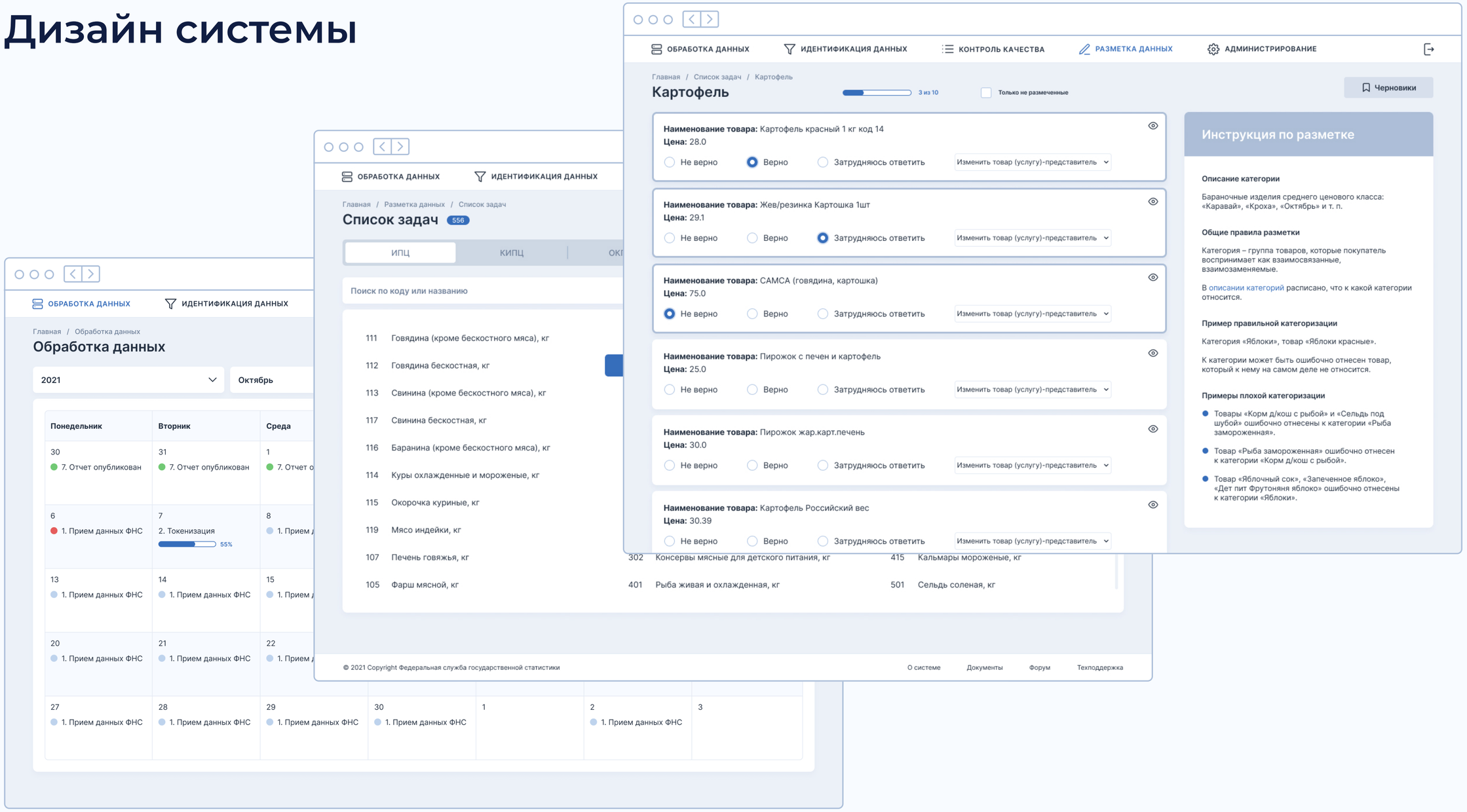
Внедрена в промышленную эксплуатацию система, позволяющая **проводить статистические исследования на основе данных ККТ**, получаемых на регулярной основе от ФНС России.

90%

ТОЧНОСТЬ КЛАССИФИКАЦИИ ДАННЫХ ИЗ ЧЕКОВ НЕЙРОСТЕТЬЮ



Дизайн системы



Компонентная схема

Бизнес приложения (front-end)		Front-end подсистем бизнес приложений	
Сервисы (back-end)	    	Back-end бизнес приложений Сервис идентификации товаров Сервис контроля качества	Поисковая программа Сервис машинного обучения
Хранение данных	    	Подсистемы хранения первичных и обработанных данных Подсистемы хранения проиндексированных рассчитанных данных	Подсистема хранения кеш данных для оперативного доступа Витрины данных для смежных систем Подсистема хранения логов
Загрузка данных	 	Подсистема распределенной обработки данных Подсистема загрузки данных	
Существующий корпоративных ландшафт	 	Мониторинг инфраструктуры Визуализация собранных метрик	

Функциональная схема

Обеспечивающие подсистемы

- Интеграция
- Администрирование

Большие данные

Прикладные подсистемы

- Прием данных
- Хранение данных
- Идентификация товаров
- Контроль качества
- Машинное обучение
- Математическая обработка и визуализация