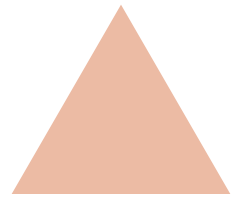




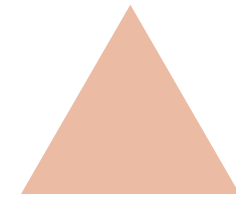
# ИС БД Росстат

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

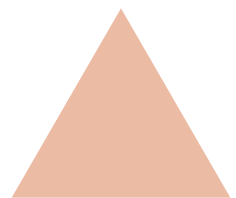
# Проблематика рынка



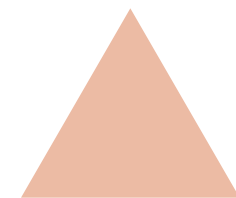
Сложность сбора и  
качество обработки  
больших данных



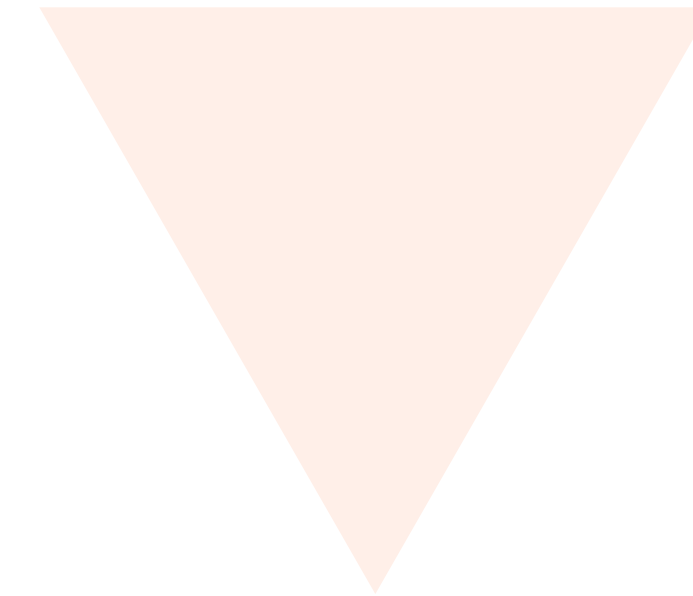
Отсутствие универсальных  
инструментов по сбору и  
обработке данных



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

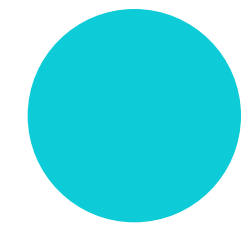


Высокая стоимость разработки  
систем по работе с большими  
данными

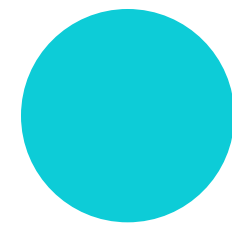


# Решение

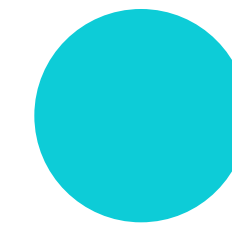
Разработан универсальный алгоритм, который позволяет:



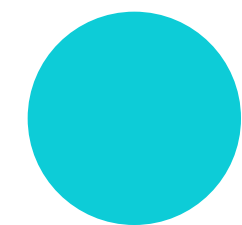
Обрабатывать большие данные и находить паттерны в информации



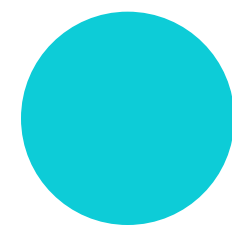
Собирать данные из различных источников: новостные сайты, чеки, блоги и т. д.



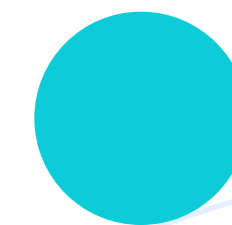
Автоматизировать процесс сбора при сокращении затрат на разработку не менее 40%.



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



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



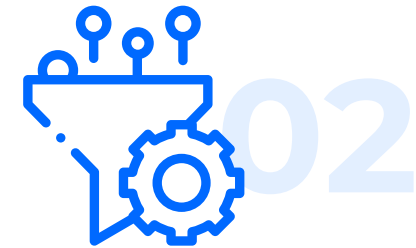
Универсальность технологии. Легкая адаптируемость под различные тематические области (Например, медицинская тематика, трейдинг, продуктовая и др.)

\*Подтверждено классификацией чеков Росстата

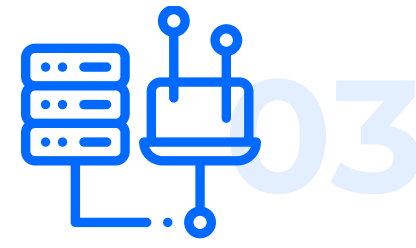
# Возможности продукта



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



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



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

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



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



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

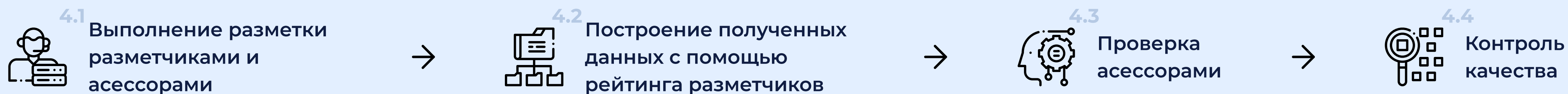
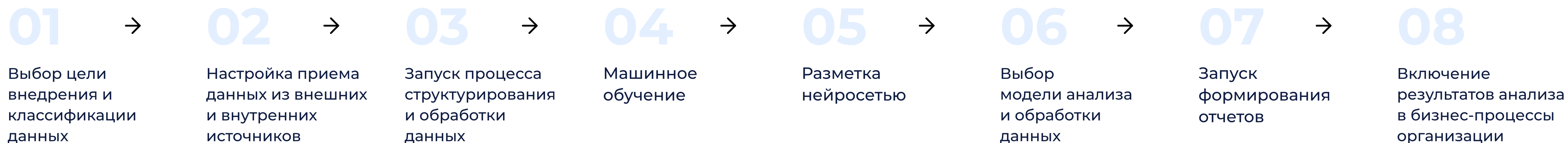
Хранение данных: Ceph, Spark SQL, Elastic Enterprise Search, Redis, Prometheus

Загрузка данных: Spark, Python

Существующий корпоративных ландшафт: Zabbix, Grafana

# Установка

Сохраняем конфиденциальность данных



## 4.1.2



Проверка полученных данных

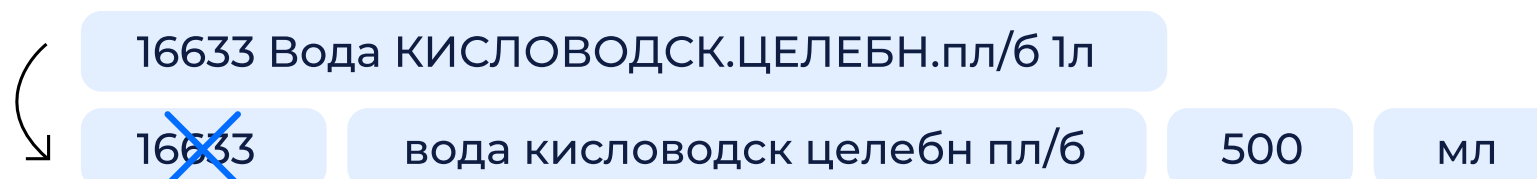


## 4.1.3



### Обработка данных делится на 2 этапа

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



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



### 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.

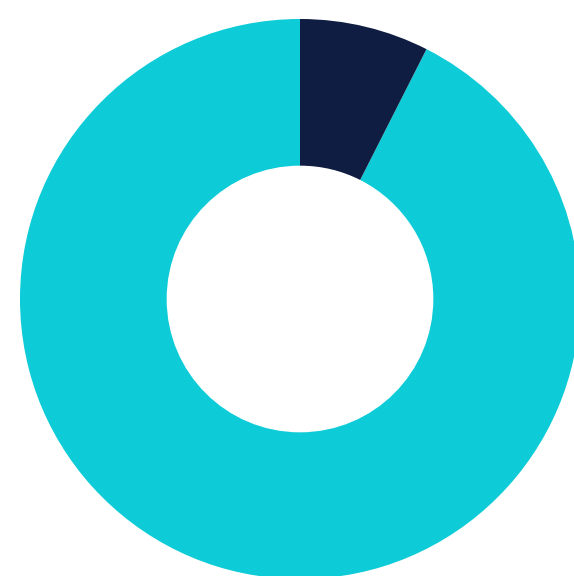
Разработан web-интерфейс для работы с ИС БД.



# 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

ОБРАБОТКА ДАННЫХ | ИДЕНТИФИКАЦИЯ ДАННЫХ | КОНТРОЛЬ КАЧЕСТВА | РАЗМЕТКА ДАННЫХ | АДМИНИСТРИРОВАНИЕ

Главная / Идентификация данных

### Категории

Поиск по наименованию и

Наименование
Колбаса сырокопченая
Бромгексин
Консервы мясные для д
Говядина
Таурин
Ксилометазолин (Галазо
Соки фруктовые
Колбаса вареная
Консервы мясные
Пельмени, манты, равио

К < 1 2

Фразы ИПЦ | Фразы ОКПД2 | Инструкция | Машинное обучение

Категория: Говядина | Код ИПЦ: 1504

Экспорт всех фраз | Импорт всех фраз

Поиск единиц измерения: Поиск | Присвоенные единицы измерения: Грамм x, Тонна x

Основная фраза: ГОВЯДИНА | Экспорт | Импорт

Исключения из фразы: нарезка x, felix x, колбаски x, тушёная x, балык x, perfectfit x, perfect x, gourmet x, доширак x, д/собак x, соб x, собак x, тефтели x, шницель x, холодец x, барбекю x, по-восточному x, по-строгановски x, по-восточному x, по-еврейски x, по-мусульмански x, по-тайски x, картофелем x, wok-гуляш x, пастроми x, лангет x, кубики x, тушен x, +

Основная фраза: ГОВЯД | Экспорт | Импорт

Исключения из фразы: нарезка x, felix x, колбаски x, тушёная x, балык x, perfectfit x, perfect x, по-еврейски x, по-мусульмански x, по-тайски x, картофелем x, wok-гуляш x, пастроми x, лангет x, кубики x, тушен x, +

Основная фраза: ГОВ | Исключения из фразы: чаппи x, корм x, шаурма x, вискас x, косточки x, бек x, гуляш x

Сохранить | Добавить фразу | Удалить | Отмена

© 2021 Информационная система по сбору больших данных. версия: 2.1 | Система | Документы | Форум | Те

# Дизайн системы

ОБРАБОТКА ДАННЫХ ИДЕНТИФИКАЦИЯ ДАННЫХ КОНТРОЛЬ КАЧЕСТВА РАЗМЕТКА ДАННЫХ АДМИНИСТРИРОВАНИЕ

Главная / Список задач / Картофель

### Картофель

3 из 10  Только не размеченные Черновики

**Наименование товара:** Картофель красный 1 кг код 14  
**Цена:** 28.0

Не верно  Верно  Затрудняюсь ответить

**Наименование товара:** Жев/резинка Картошка 1шт  
**Цена:** 29.1

Не верно  Верно  Затрудняюсь ответить

**Наименование товара:** САМСА (говядина, картошка)  
**Цена:** 75.0

Не верно  Верно  Затрудняюсь ответить

**Наименование товара:** Пирожок с печен и картофель  
**Цена:** 25.0

Не верно  Верно  Затрудняюсь ответить

**Наименование товара:** Пирожок жар.карт.печень  
**Цена:** 30.0

Не верно  Верно  Затрудняюсь ответить

**Наименование товара:** Картофель Российский вес  
**Цена:** 30.39

Не верно  Верно  Затрудняюсь ответить

302 Консервы мясные для детского питания, кг 415 Кальмары мороженые, кг

401 Рыба живая и охлажденная, кг 501 Сельдь соленая, кг

© 2021 Copyright Федеральная служба государственной статистики О системе Документы Форум Техподдержка

ОБРАБОТКА ДАННЫХ ИДЕНТИФИКАЦИЯ ДАННЫХ

Главная / Разметка данных / Список задач

### Список задач

556

ИПЦ КИПЦ ОКЦ

Поиск по коду или названию

- 111 Говядина (кроме бескостного мяса), кг
- 112 Говядина бескостная, кг
- 113 Свинина (кроме бескостного мяса), кг
- 117 Свинина бескостная, кг
- 116 Баранина (кроме бескостного мяса), кг
- 114 Куры охлажденные и мороженые, кг
- 115 Окорочка куриные, кг
- 119 Мясо индейки, кг
- 107 Печень говяжья, кг
- 105 Фарш мясной, кг

© 2021 Copyright Федеральная служба государственной статистики

ОБРАБОТКА ДАННЫХ ИДЕНТИФИКАЦИЯ ДАННЫХ

Главная / Обработка данных

### Обработка данных

2021 Октябрь

Понедельник	Вторник	Среда
30 ● 7. Отчет опубликован	31 ● 7. Отчет опубликован	1 ● 7. Отчет опубликован
6 ● 1. Прием данных ФНС	7 2. Токенизация 55%	8 ● 1. Прием данных ФНС
13 ● 1. Прием данных ФНС	14 ● 1. Прием данных ФНС	15 ● 1. Прием данных ФНС
20 ● 1. Прием данных ФНС	21 ● 1. Прием данных ФНС	22 ● 1. Прием данных ФНС
27 ● 1. Прием данных ФНС	28 ● 1. Прием данных ФНС	29 ● 1. Прием данных ФНС
30 ● 1. Прием данных ФНС	1	2 ● 1. Прием данных ФНС
		3

### Инструкция по разметке

**Описание категории**

Бараночные изделия среднего ценового класса: «Каравай», «Кроха», «Октябрь» и т. п.

**Общие правила разметки**

Категория – группа товаров, которые покупатель воспринимает как взаимосвязанные, взаимозаменяемые.

В описании категорий расписано, что к какой категории относится.

**Пример правильной категоризации**

Категория «Яблоки», товар «Яблоки красные».

К категории может быть ошибочно отнесен товар, который к нему на самом деле не относится.

**Примеры плохой категоризации**

- Товары «Корм д/кош с рыбой» и «Сельдь под шубой» ошибочно отнесены к категории «Рыба замороженная».
- Товар «Рыба замороженная» ошибочно отнесен к категории «Корм д/кош с рыбой».
- Товар «Яблочный сок», «Запеченное яблоко», «Дет пит Фрутоняня яблоко» ошибочно отнесены к категории «Яблоки».



# 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

