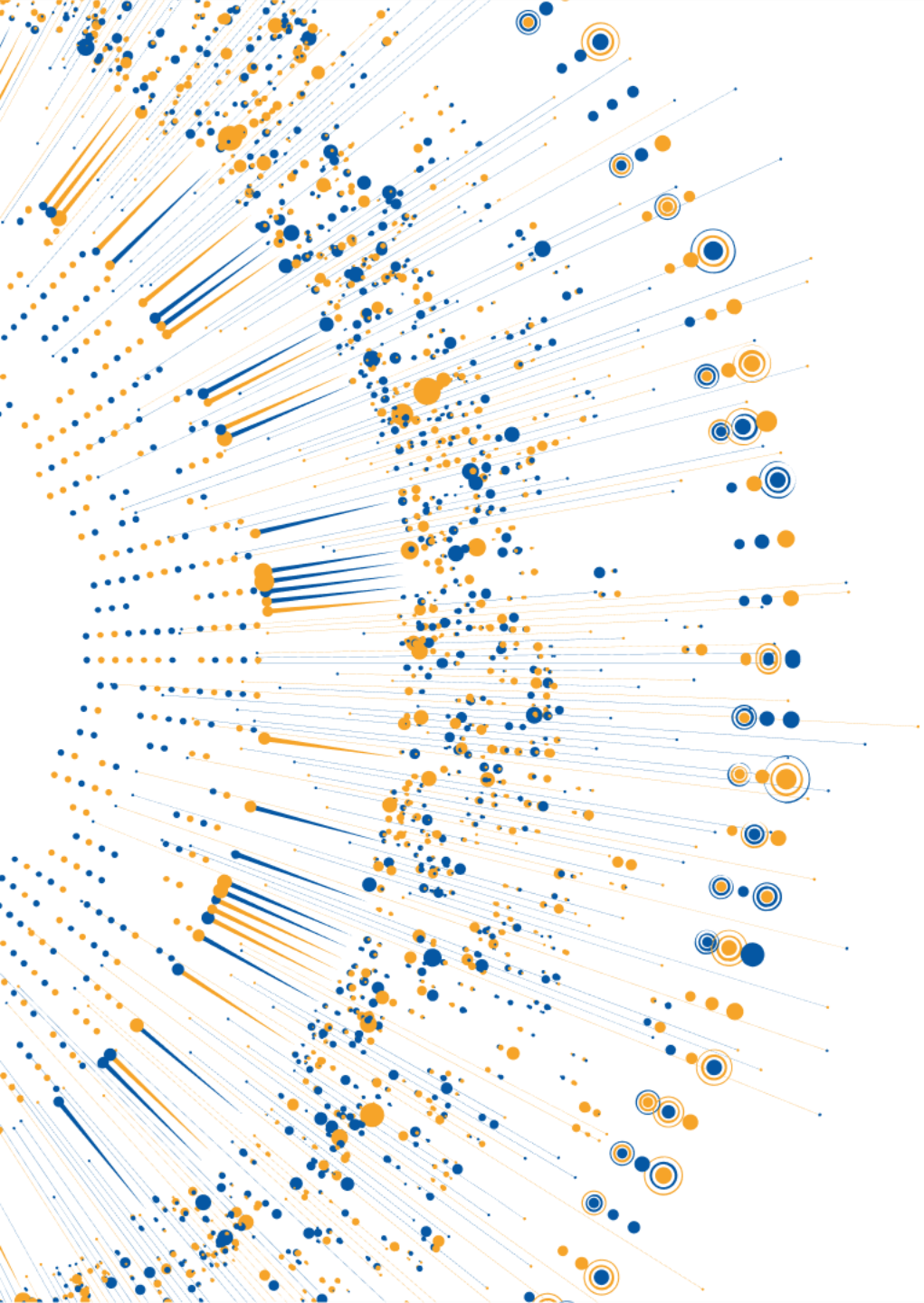


QAZSTAT

БЮРО НАЦИОНАЛЬНОЙ СТАТИСТИКИ
АГЕНТСТВА ПО СТРАТЕГИЧЕСКОМУ
ПЛАНИРОВАНИЮ И РЕФОРМАМ РЕСПУБЛИКИ
КАЗАХСТАН

Использование больших данных при формировании ИПЦ в Казахстане

ноябрь 2025 г.

Активная стадия реформирования Бюро национальной статистики

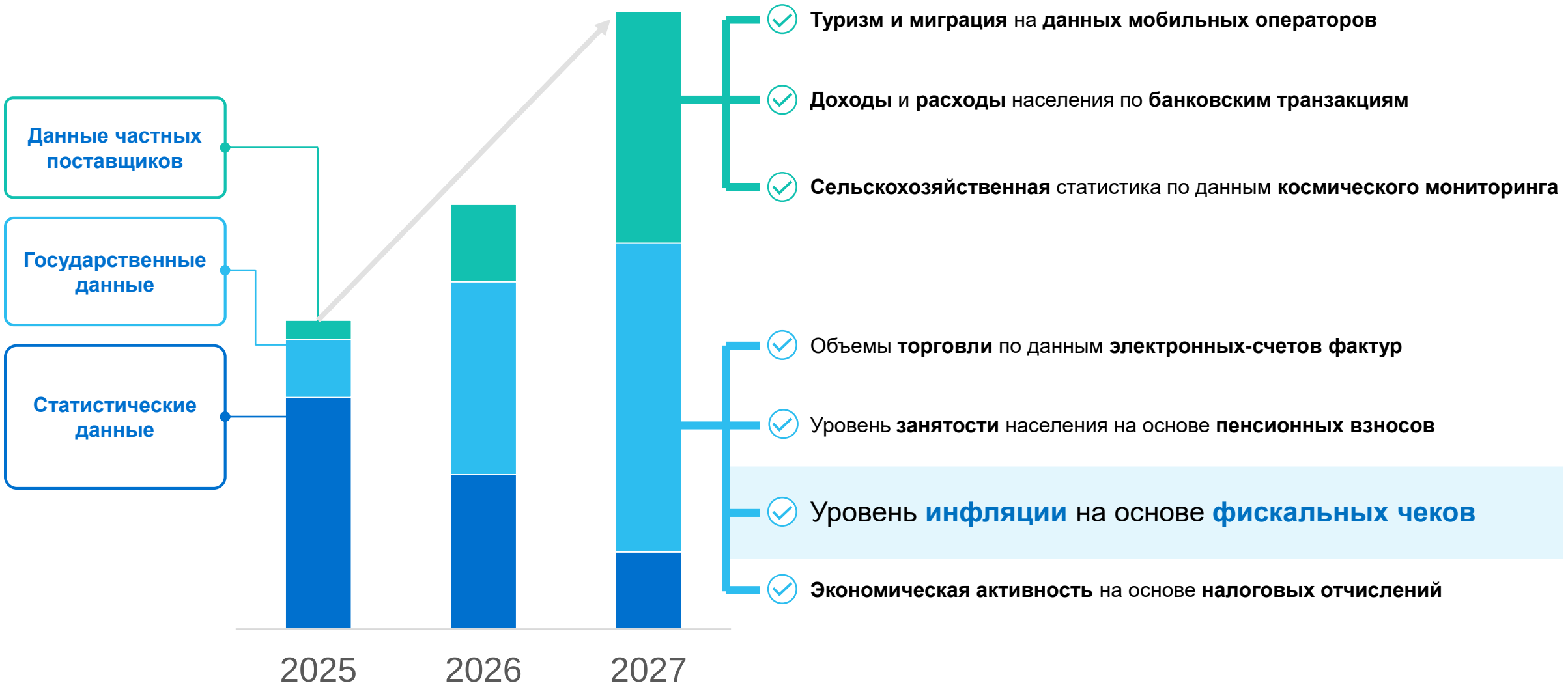
- **На активной стадии реализация Концепции развития государственной статистики и национальной экосистемы данных**

В реализацию Концепции утверждена Дорожная карта реформирования БНС:

1. Институционально-организационные меры
2. Активное использование **административных** и **альтернативных** источников
3. Укрепление аналитического потенциала и расширение доступа к данным

Реформы реализуются поэтапно **с учетом мнения экспертного сообщества**, **пользователей и респондентов** для соблюдения интересов всех заинтересованных сторон

Переход на альтернативные источники данных



Касательно административных данных ОФД

Операторы фискальных данных (ОФД) — связующее звено для передачи сведений в комитет государственных доходов от контрольно кассовых машин

- В стране операторы фискальных данных (ОФД) появились в 2015 году, когда АО «Казахтелеком» было назначено постановлением правительства оператором, отвечающим за передачу сведений о денежных расчетах налогоплательщиков в органы налоговой службы в режиме онлайн.

Обязательные пользователи онлайн-касс:

- Субъекты осуществляющие розничную торговлю и услуги населению
- Субъекты обязанные фиксировать наличные и безналичные расчеты через кассы
- Торговые автоматы и терминалы оплаты услуг, если они принимают наличные средства

Освобожденные от обязательного использования онлайн-касс:

- Частные судебные исполнители, адвокаты и медиаторы
- Организации и граждане, осуществляющие перевозки общественным транспортом
- Национальный Банк РК, банки второго уровня, религиозные объединения
- Налогоплательщики в местах без связи, допускается использование только «старых» касс, без передачи данных

Ограничения данных ОФД

- Не охватывают неформальный сектор экономики
- В случае технических сбоев (кассовое ПО, серверы ОФД, перебои связи) возможна задержка поступления чеков
- Не всегда в чеке указывается подробное наименование товара (например, только «товар», «продукт»)
- В случае задержки передачи данных от КГД возможны сбои в полноте и своевременности анализа, что может повлиять на оперативность расчетов ИПЦ

Отличительная особенность данных ОФД это возможность оперативного мониторинга цен на товары

Внедрение инструментов ИИ для автоматизации обработки и анализа данных

Цель проекта – разработка и внедрение автоматизированной системы классификации наименований товаров с использованием современных языковых моделей и методов машинного обучения для ускорения обработки фискальных данных и упрощения анализа товарного ассортимента

Задачи

- Категоризация 362 млн. товарных наименований для дальнейшего использования при расчете инфляции на основе данных ОФД

Ожидаемые результаты

- Полностью автоматизированная система классификации наименований товаров (перед началом реализации проекта было 20 товаров)
- Расширение перечня товаров для формирования статистической информации

Ключевые эффекты

- Точный мониторинг инфляции по большему количеству товаров
- Детальный расчет показателей розничной торговли

Источники данных

- Данные Операторов фискальных данных (ОФД)
- Данные Статистического бизнес регистра
- Справочник по адресам контрольно-кассовых машин

Этапы построения выборки и очистки для категоризации данных

Этапы построения выборки

1. **Фильтрация по видам деятельности** исключительно розничной торговли, исключение онлайн торговли, аптек и прочего
2. **Фильтрация по обороту**, компании с объемом продаж **свыше 10 млн тенге**
3. **Фильтрация по длине символов**, длина от 2 до 150 символов, исключаются технические записи *(TO_EMPTY, BAD_BIN)*
4. **GTIN-склейка** – в случае, если в чеке имеется GTIN, он используется для унификации в рамках одного субъекта

Этапы очистки

1. Приведение наименований товаров **к нижнему регистру**
2. **Очистка от неинформативных данных** *(спецсимволы, служебные наименования – «чек», «касса», «скидка», прочее)*
3. **Нормализация символов**, латинские буквы, похожие на кириллицу (a→а, o→о, p→р и т.д.), заменяются.
4. **Специфическая обработка топлива** для строк с наименованием «АИ-92», «АИ-95», «ДТ», «газ» отдельные правила.

Обучена модель bge_m3 (600 млн параметров), которая по качеству сопоставима с GPT-5 на задачах категоризации товаров по данным ОФД

Картофель

Выбросы

Модель машинного обучения на базе обработки естественного языка

- **bge_m3** ***(~600 млн параметров)*** компактная модель, дообученная на данных ОФД. Работает быстро и дешевле в эксплуатации.
- **GPT-5** ***(сотни миллиардов параметров)*** универсальная модель для множества задач, при этом гораздо тяжелее и дороже.
- **LLaMA-7B** *(7 млрд параметров)* пример «среднего размера» модели, которая тоже требует серьезных ресурсов.

От выборочного мониторинга к масштабному применению данных фискальных чеков

Традиционный метод

120 тыс. котировок

12 тыс.
базовых объектов
(магазины, рынки)

Цены фиксируется регистраторами цен *(свыше 300 сотрудников)*

Использование данных фискальных чеков

36,9 млн. котировок

124 тыс.
объектов розничной торговли охваченных категоризацией

Используются данные фискальных чеков с применением ИИ

Внедрение инструментов ИИ для автоматизации обработки и анализа данных

Текущие результаты

КГД

КОЛИЧЕСТВО ПОСТУПИВШИХ ЗАПИСЕЙ

20,6 млрд

КОЛИЧЕСТВО УНИКАЛЬНЫХ НАИМЕНОВАНИЙ ТОВАРОВ

362 млн

После фильтрации, очистки и разметки

БНС

КОЛИЧЕСТВО УНИКАЛЬНЫХ ТОВАРОВ ПОСЛЕ ФИЛЬТРАЦИИ ПО КРИТЕРИЯМ

36,9 млн

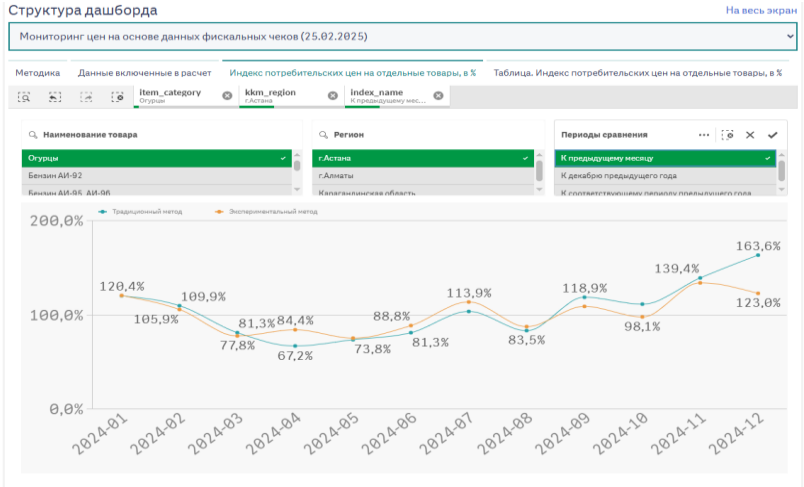
КОЛИЧЕСТВО ОХВАЧЕННЫХ КАТЕГОРИЙ ТОВАРОВ

290 или **57%** *(от всего перечня товаров ИПЦ)*

КОЛИЧЕСТВО ОХВАЧЕННЫХ ЗАПИСЕЙ ПО 290 КАТЕГОРИЯМ

8,3 млрд или **40,3%**
(от всего кол-ва записей)

Индекс потребительских цен на основе анализа диапазона цен



Данные обновляются на дашборде для внутреннего пользования

Примечание: в расчет взяты данные с января 2023 года по август 2025 года

Комбинированный метод сбора

Вместо бумажного процесса регистрации цен переход на цифровой сбор посредством сканирования **штрих кодов, фиксирования пути и места нахождения регистратора цен, а также централизованный сбор посредство парсинга**

Внедрен модуль ввода данных мобильного приложения (МП)
С марта 2024г. сбор цен на СЗПТ производится с помощью МП

Преимущества:

- повышение качества первичных данных за счет минимизации ошибок при вводе (возможность передачи скринов товара, выявление ошибок)
- передача данных в онлайн-режиме и более оперативная обработка данных

Разработан и используется инструмент парсинга
С января 2024г. осуществляется сбор по 43 наименованиям ИПЦ

Преимущества:

- web scraping позволяет автоматический собирать данные о ценах с множества сайтов, что значительно экономит время по сравнению с традиционным методом
- передача данных в онлайн-режиме и более оперативная обработка данных

Осуществлен экспериментальный расчет индексов цен **на основе данных контрольно-кассовых машин**. Результаты расчетов опубликованы на официальном сайте Бюро в рубрике «Экспериментальная статистика и исследования» в виде дашборда и аналитической записки к дашборду.