Предложения по организации автоматизированного сбора данных из национальных статистических служб государств-участников СНГ

Текущая ситуация

Статистические показатели собираются в виде вопросников, в которых содержатся все показатели. Показатели вводятся в Excel таблицу в ненормализованном и машинно-нечитаемом виде. Справочники не определены, факты не определены, формального описания структуры нет. Разрезности (измерения) и справочники являются заголовками колонок, клетками таблицы.

Из этих таблиц вручную создаются показатели в ПАК ИАП, они называются «базовые показатели». Для этого используется некоторый конструктор показателя, в котором пользователь может выбрать таблицу и описать какие клетки относятся к каким разрезам и справочникам, а какие к фактам. Это крайне трудоемкая и нетворческая работа, требующая, однако, высокой квалификации и опыта исполнителя.

Для оптимизации работ по предоставлению в Статкомитет СНГ и сбору статистических данных, мы подготовили следующие Предложения по автоматизации.

Общие принципы:

1. Сбор наборов данных показателя вместо сбора ненормализованных таблиц.
2. Один набор данных должен содержать сведения за один или нескольких периодов по одной стране и одному показателю.
3. Наборы данных будут полностью нормализованы. Они будут содержать измерения – колонки с кодами справочников, атрибуты – колонки с дополнительной информацией, например с примечаниями, и факты – колонки с цифрами.
4. Для каждого показателя в ЕИАС автоматически будет генерироваться описание структуры и форматов для сбора данных.
5. Поддерживается одновременно несколько форматов:
    a. CSV – простейший текстовый формат
    b. JSON – простой технологичный формат обмена данными, не содержащий прикладных метаданных
    c. JSON-LD – простой, но мощный формат обмена данными, содержащий определение структуры показателя, состав используемых справочников, используемые в данных показателя записи справочников, набор данных, ссылающийся на справочники.
    d. SDMX – сложный, и очень мощный формат, уже применяющийся НСС в отдельных случаях.
6. Загрузка данных будет выполняться полностью автоматически, стандартными механизмами хранилища данных, с отражением результатов в журнале загрузки данных, дашборде «Мониторинг загрузки» и специальном журнале ошибок, предназначенном для поставщиков данных и разработчиков коннекторов.
7. Полное описание показателей, включая формальное описание структуры, будет содержаться в едином реестре показателей. В нем для каждого показателя можно будет выгрузить динамически генерируемую документацию по всем форматам обмена данными и метаданными.

8. Все справочники будут содержаться в реестре справочников. НСС, в плане создания (при необходимости) коннекторов, смогут выгрузить описание справочников и элементы справочников для разработки коннекторов.
9. В системе должны будут настраиваться «переходные ключи» для автоматической перекодировки данных из системы классификации НСС в централизованную систему классификации Статкомитета СНГ.
10. Для разработки коннекторов и перекодировки будет применен искусственный интеллект.
11. Загрузка данных будет выполняться через API ЕИАС.
12. В ЕИАС создается календарь сбора данных, позволяющий видеть сроки предоставления данных и статус предоставления данных в разрезах страна, показатель, период.

Методы сбора данных

Предлагается одновременно использовать два режима сбора данных:

1. Толкающий (push) – когда НСС загружает данные в ЕИАС своим коннектором через API ЕИАС. Этот режим наиболее предпочтительный для нас, более надежный технически и методически, но более трудозатратный для НСС.
2. Тянущий режим (pull) – когда коннектор ЕИАС Статкомитета соединяется с API статистической системы НСС, выполняет запросы и загружает данные в ЕИАС. Этот режим снижает трудоемкость для НСС, но затратен для Статкомитета СНГ и влечет технологические и методические риски. Вместе с этим статистические системы НСС тоже должны быть готовы к забору данных из них.

Толкающий режим

Преимущества толкающего режима:

1. Коннектор с ЕИАС запускается в статистической системе по мере готовности данных за новый период в информационной системе НСС;
2. В разработке коннекторов участвует большое количество команд и экспертов;
3. Данные загружаются сразу в системе показателей и справочников Статкомитета, а переклассификацией занимаются люди, которые уже знают, как это делать, поскольку делали это для заполнения вопросников.

Статкомитет СНГ публикует API ЕИАС для загрузки данных, все необходимые метаданные – реестр показателей с описанием структуры каждого показателя, справочники с их кодами и именами, описание поддерживаемых форматов и примеры.

Сотрудники или подрядчики НСС разрабатывают коннекторы с ЕИАС, которые забирают данные из системы НСС, в идеале перекодируют их в систему классификации Статкомитета СНГ (но не обязательно) и загружают через API ЕИАС по расписанию.

В случае если данные имеют кодировку ЕИАС, их инкрементальные обновления автоматически загружаются в общие для всех стран временные ряды показателей.

Если данные имеют кодировку в классификации страны поставщика информации, то они автоматически перекодируются средствами ХД, исходные коды элементов справочников сохраняются во временных рядах для выверки и контроля.

Ручная загрузка

Ручная загрузка является разновидностью толкающего режима. В случае, если коннектор разработать не удается, а также для нужд отладки, набор данных показателя можно загрузить в имеющейся в ЕИАС странице загрузки данных как файл в одном из поддерживаемых форматов.

Тянущий режим

НСС предоставляет Статкомитету СНГ адрес, описание API и описание формата, в котором публикуются данные и обеспечивает их полный состав.

Подрядчики Статкомитета разрабатывают коннекторы с статистическими системами НСС по технологии и правилам, которые применяются для разработки коннекторов ЕИАС с международными источниками.

В лучшем случае используется API информационной системы НСС, в худшем – статистические данные скачиваются как файлы, опубликованные на сайте.

Скачивание файлов, как правило, приводит к трудностям выявления изменений, наличия данных за следующий период, если файлы выкладываются вручную, то их именование может быть произвольным, и т.д., и т.п.

Подрядчики разрабатывают коннекторы на языке JavaScript по определенным правилам, так, что коннектор устанавливается в ХД ЕИАС как плагин, в пользовательском интерфейсе, регистрируется в реестре коннекторов, его настройки выносятся в специальные справочники ХД, а не хранятся в коде, он запускается планировщиком заданий ХД по расписанию.

Разработка коннекторов с использованием ИИ

Коннекторы для обоих режимов можно разрабатывать с помощью ИИ. Для этого используются следующие подходы:

1. Алгоритм чтения данных источника индивидуален.
2. Алгоритм загрузки данных в ХД универсален, используется встроенный механизм.
3. Справочник типов ошибок содержит все виды ошибок сбора данных для контроля, а также промпты для ИИ, по которым он должен разрабатывать процедуры контроля входных данных, такие как неверная кодировка, неверная структура (много типов ошибок), пропущенные данные, нарушение ссылочной целостности, доменной целостности и т.д. Для каждого типа ошибок указывается их серьезность, вес и возможность автоматического исправления.
4. Программный код коннектора генерируется, для толкающих коннекторов на языке, удобном для системы НСС, для тянущих на JavaScript для Node.js. Во время работы коннектора ИИ не используется, в коде содержатся все правила, варианты ошибок, методы их исправления. Код обучается, то есть многократно перегенерируется ИИ по промптам, описывающим обнаруженные проблемы.
5. Для удобства разработчика предлагается подробная документация по API, форматам обмена данными, возможность выгружать как примеры в выбранном формате все метаданные, в том числе справочники, и все данные, уже имеющиеся в ХД.

Ожидаемые результаты

1. Полная автоматизация регулярной подготовки данных и передачи данных из НСС. Снижение трудоемкости в НСС.
2. Полная автоматизация сбора данных в Статкомитете. Снижение трудоемкости в Статкомитете.
3. Повышение качества и сопоставимости данных за счет внедрения мощной системы автоматической переклассификации и детального контроля, невозможности загрузить данные с грубыми ошибками.
4. Резкое повышение оперативности. Сразу после завершения загрузки будут автоматически обновляться кубы и данные, после утверждения (или без) будут опубликованы на портале.
5. Возможность расширения номенклатуры собираемых данных за счет упрощения и удешевления процесса.
6. Создание модельной технологии, которую можно использовать в НСС, а также в университетах для обучения студентов.

Проблемы, которые придется решать

1. Не все показатели, собираемые Статкомитетом СНГ, производятся НСС, часть из них производится другими ведомствами, министерствами здравоохранения, внутренних дел, центральными банками и т.д. Вероятно, что НСС нужно будет получать эти показатели и отправлять их, или публиковать для Статкомитета СНГ.
2. Проблема первоначальной трудоемкости и требования к квалификации с обеих сторон.
3. Сопоставление показателей, справочников и классификаторов, записей справочников и классификаторов, гармонизация единиц измерения и уровней агрегации, гармонизация периодичностей.
4. При толкающем режиме потребуются трудовые и финансовые ресурсы со стороны НСС.
5. При тянущем режиме выявятся несоответствие потребностей Статкомитета и публикуемых НСС открытых данных, трудовые и финансовые ресурсы также потребуются с обеих сторон.

В этой связи предлагаем:

1. *Для снижения трудоемкости сбора данных от национальных статистических служб государств – участников Содружества Независимых Государств поручить Статкомитету СНГ в 2026-2027 годах провести апробацию процесса перехода от сбора данных путем ручного заполнения вопросников и ручного преобразования их в базовые показатели на автоматизированную загрузку данных непосредственно в хранилище данных Единой информационно-аналитической системы (ЕИАС).*

2. Апробация перехода на автоматизированный сбор должна выполняться поэтапно с учетом готовности национальных статистических служб государств-участников СНГ, с выбором механизмов сбора, наиболее подходящих для конкретной статистической службы.

3. Пилотными участниками данного проекта определить национальные статистические службы Казахстана, Кыргызстана и Узбекистана.