Проект создания цифровой платформы Счетной палаты РФ

Цели проекта

Стратегия развития Счетной палаты РФ на 2018-2024 годы (далее «Стратегия») определяет приоритетные направления развития для реализации новых задач с учетом цифровизации информационного обеспечения деятельности Счетной палаты по осуществлению внешнего государственного аудита (контроля) на новом качественном уровне, где особо выделена задача создания цифровой инфраструктуры для поддержки аудита и аналитической деятельности,

На основании Стратегии Департаментом цифровой трансформации Счетной палаты разработана и утверждена Концепция цифровизации Счетной палаты РФ, которая определяет формат построения Цифровой платформы[1] и рабочего места «Цифрового

[1] **Цифровая платформа** – в общем смысле система алгоритмизированных взаимовыгодных взаимоотношений значимого количества независимых участников какой-либо сферы деятельности, которая приводит к принципиальному снижению транзакционных издержек за счет применения цифровых технологий работы с данными, устранения посредников и изменения системы разделения труда.

инспектора». С их помощью расширяется список используемых для аудита источников данных и повышается их качество, применяются современные методы управления информацией, снижается трудоемкость традиционных видов аудита и обеспечивается развитие стратегического аудита за счет создания инструментов риск-ориентированных и аналитических моделей и применения современных методов предиктивной аналитики.

В рамках реализации Концепции цифровизации в Счетной палате создается Цифровая платформа – программно-аппаратный комплекс, предоставляющий возможности дата-аналитикам иметь в своем распоряжении необходимые данные и на их основании с помощью инструментария Цифровой платформы обрабатывать их, получая риск-ориентированные модели, аналитические модели, визуальные инструменты и т.д.

Уникальность проекта

Уникальность проекта состоит в том, что:

1. Данные в «Озеро данных[2]» собираются из неопределенного заранее перечня источников всех возможных типов и этот перечень может со временем расширяться – в связи с этим необходимо было решить вопрос оптимального способа хранения разнородной информации, ее очистки, обработки, связывания данных из различных источников и далее построения витрин данных.
2. Источники данных находятся создаются различными системами, по большей части находящимися вне информационной системы Счетной палаты. Данные во внешних информационных системах построены на справочниках своей структуры и в результате связи между данными из различных источников не очевидны. Так же в качестве источников данных много файлов со машиночитаемой и машинно нечитаемой информацией, которые так же требуется привести к единому виду, распознать и связать с остальными данными.
3. Не готовность большого количества конечных пользователей видеть данные в «новом свете» через системы аналитические модели и визуальные инструменты и не умение пользоваться этим инструментарием, а привычка работать «по старинке».

Цифровая платформа Счетной палаты

Цифровая платформа Счетной палаты (Рисунок 1) включает себя:

- хранилище данных – так называемое «Озеро данных»;
- витрины данных;
- системы визуализации и аналитики;
- автоматизированные системы для обработки информации в целях облегчения деятельности инспекторского состава.

[2] Озера данных — это решения нового поколения для управления гибридными данными, позволяющие решать задачи в сфере больших данных и реализовать принципиально новые методы аналитики в реальном времени. Высокая масштабируемость решений обеспечивает поддержку очень больших объемов данных и возможность приема данных в исходных форматах из самых разнообразных источников. Озера данных помогают объединить разрозненные данные и дают организациям возможность получить полное представление о имеющейся у них информации.

Рисунок 1 Функциональная схема Цифровой платформы (целевая)

«Озеро данных» Счетной палаты РФ

«Озеро данных» Счетной палаты РФ построено на открытом ПО, в основе которого находится стек ПО от компании Arenadata, обогащенный открытым ПО от других производителей.

«Озеро данных» состоит из следующих основных компонентов:

4. Хранилище сырых данных (Arenadata Hadoop)
5. Системы загрузки данных (Apache AirFlow)
6. Хранилище метаданных (MongoDB)
7. Хранение витрин данных[3] и промежуточных таблиц (Arenadata DB)

Виртуализация данных

Технология хранения и обработки данных в озере строится по принципу виртуализации данных на основе семантического стека технологий, т.е. физически данные хранятся в хранилище сырых данных, с дополнительным слоем метаданных по модели RDF[4], описывающим данные, их структуру, связи и код на Python для преобразования при необходимости. Получение данных из хранилища «сырых» данных происходит с помощью специального языка запросов SPARQL[5], который на основе метаданных получают данные из хранилища сырых данных.

[3] **Витрины данных** – связанные, очищенные и обогащенные данные по одной тематике, представленные в виде плоской таблицы.

[4] **Resource Description Framework** (RDF, «среда описания ресурса») — это разработанная консорциумом Всемирной паутины модель для представления данных, в особенности — метаданных. RDF представляет *утверждения* о *ресурсах* в виде, пригодном для машинной обработки.

[5] SPARQL – язык запросов к данным, представленным по модели RDF, а также протокол для передачи этих запросов и ответов на них. SPARQL является рекомендацией консорциума W3C и одной из технологий семантической паутины. Предоставление SPARQL-точек доступа является рекомендованной практикой при публикации данных во всемирной паутине

Семантический стек технологий позволяет создать систему виртуализации данных полностью опираясь на открытые стандарты и продукты с открытым исходным кодом.

Этапы процесса работы с данными

Процесс работы с данными в озере происходит по следующим этапам (Рисунок 3):

1. При записи данных из источника дополнительно создаётся слой метаданных.
2. После поступления данных в хранилище сырых данных запускается процесс автоматического построения связей между загруженными данными из различных источников.
3. Результат автоматического построения контролируется через специально созданный интерфейс, при необходимости модифицируется, дополняется описательная часть метаданных и описываются в слое метаданных необходимые преобразования данных.
4. Следующим этапом через интерфейс заказывается создание витрины данных выбором из существующего набора с учетом связей.
5. Далее по заказу строятся витрины данных для дальнейшего использования.

Рисунок 2 – Этапы процесса работы с данными

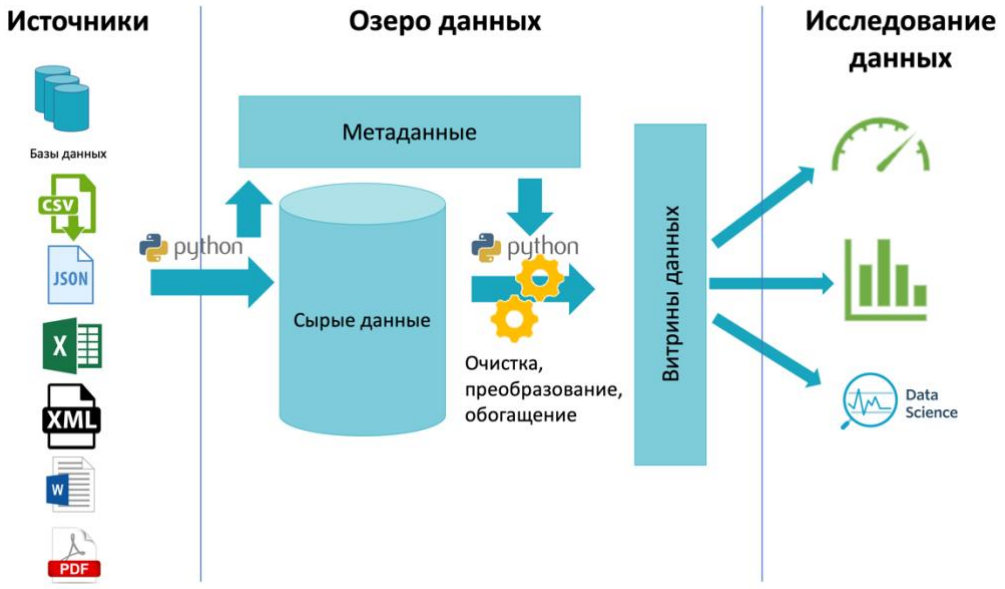

Рисунок 3 – Технологический процесс работы с данными

Хранение сырых данных

Для хранения сырых данных используется в Arenadata Hadoop.

Загруженные данные хранятся в файлах формата JSON с дополнительной служебной информацией, такой как дата и время загрузки, версия и прочее. Файлы формата JSON выбраны по причине их независимости от структуры источника.

Обработка данных из источника и сохранение их в хранилище сырых данных осуществляется одним из 3-х вариантов, в зависимости от типа источника данных:

Вариант 1: если данные загружаются из реляционной БД, то результат хранится в JSON файлах (Рисунок 4).

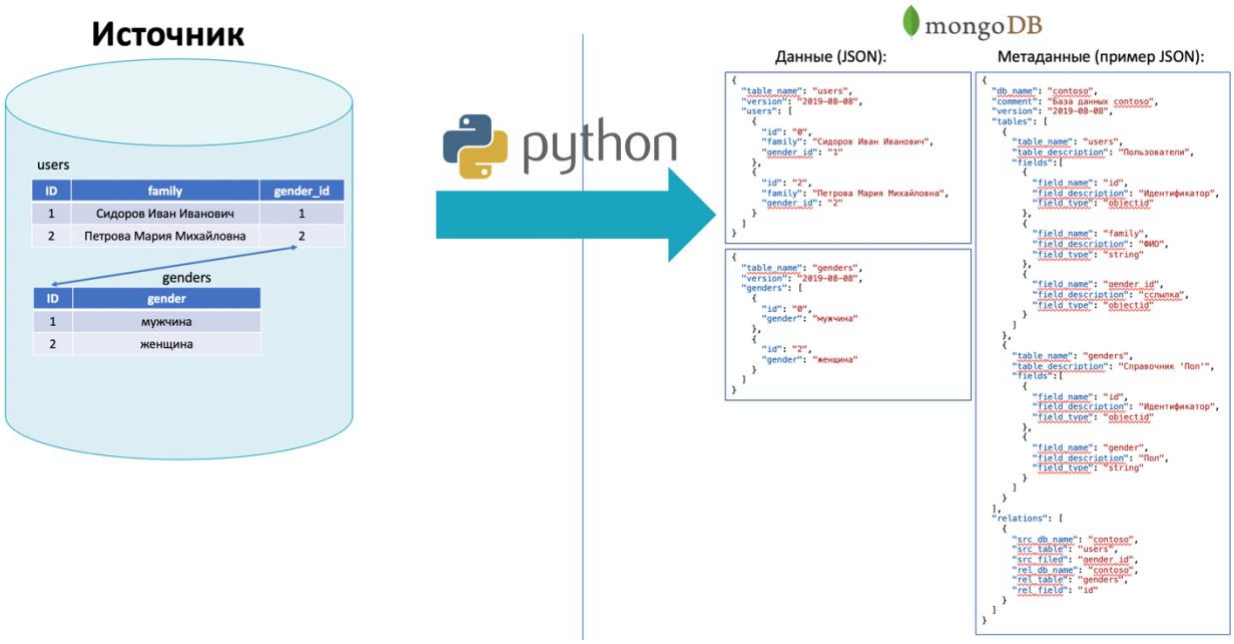

Рисунок 4 - пример записи таблиц из БД в хранилище сырых данных

Вариант 2: если источник является файлом со структурированными данными, то:

1. Загружается как есть в Hadoop и добавляется его описание в виде метаданных
2. Его содержимое сохраняется в виде набора JSON файлов
3. Дополняется слоем метаданных

Вариант 3: если источник является файлом со неструктурированными данными, то:

1. Файл хранится как есть в хранилище сырых данных
2. Дополняется слоем метаданных, описывающих этот файл.

Виртуализация данных на основе семантического стека технологий.

Виртуализация данных подразумевает под собой наличие слоя метаданных

Метаданные описываются с помощью OWL (Web Ontology Language — язык описания онтологий[6] для семантической паутины) и его расширения RDF **(Resource Description Framework** – «среда описания ресурса»).

[6] **Онтоло́гия** в информатике — это попытка всеобъемлющей и подробной формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области.

Фактически метаданные являются слоем для виртуализации данных, к которым можно обратиться программно или через интерфейс и далее соответствующей обработкой языком запросов SPARQL получить физические данные и сформировать витрину данных.

Метаданные позволяют:

1. Однозначно автоматически идентифицировать данные начиная с источника и заканчивая данными, из которых будут формироваться витрины данных.
2. Иметь человеческое описание данных с возможностью поиска
3. Описывать связи между данными по всем источникам.
4. Описывать необходимые преобразования данных

Связи между данными из различных источников создаются как минимум тремя способами:

1. Программно в коде загрузки данных в хранилище сырых данных – в результате этого создаётся первичное описание данных
2. С помощью специальной обработки, которая после загрузки данных в хранилище проходит по всем данным и с помощью различных методов, в том числе с помощью искусственного интеллекта, ищет и описывает связи между данными из различных источников.
3. Корректируются и дополняются метаданные так же вручную в специальном ПО.

Потоковая загрузка данных

Загрузка данных осуществляется под контролем Apache AirFlow.

Apache AirFlow – это библиотека для разработки, планирования и мониторинга рабочих процессов. Основная особенность AirFlow: для описания (разработки) процессов используется код на языке Python.

Apache AirFlow позволяет по заданию будильника или по событию (например, появлению файла в каталоге) запускать процесс обработки.

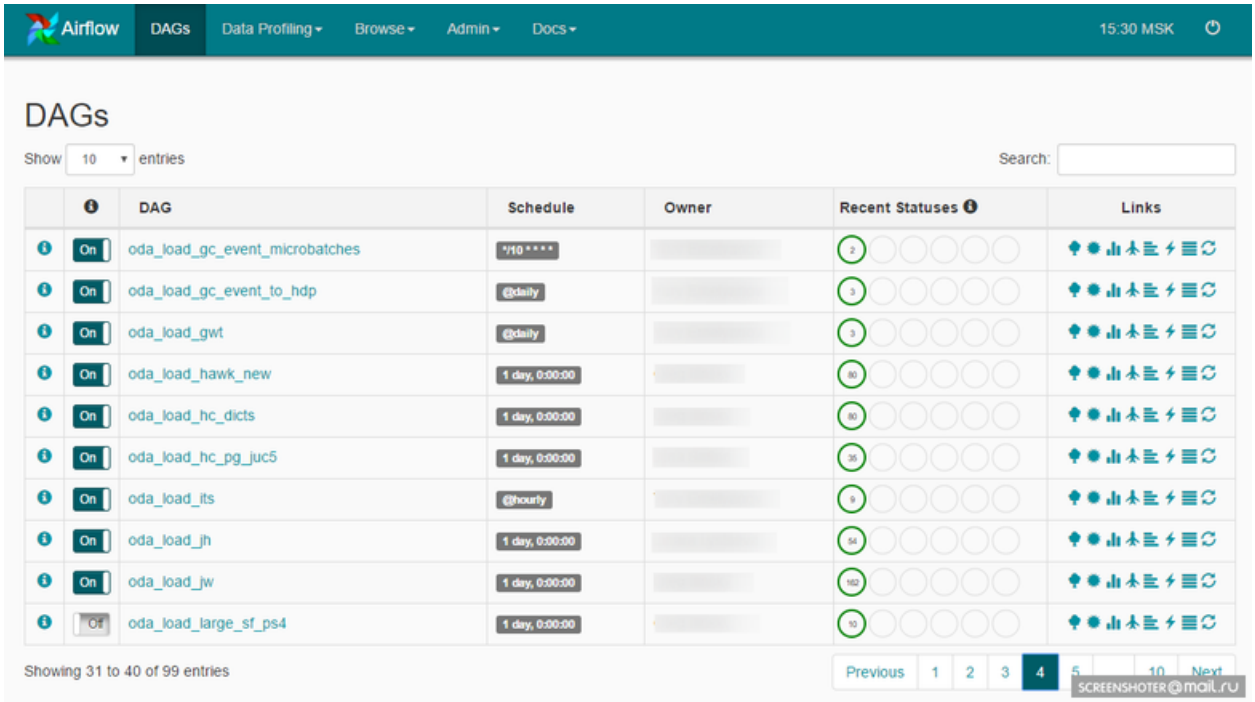

Рисунок 5

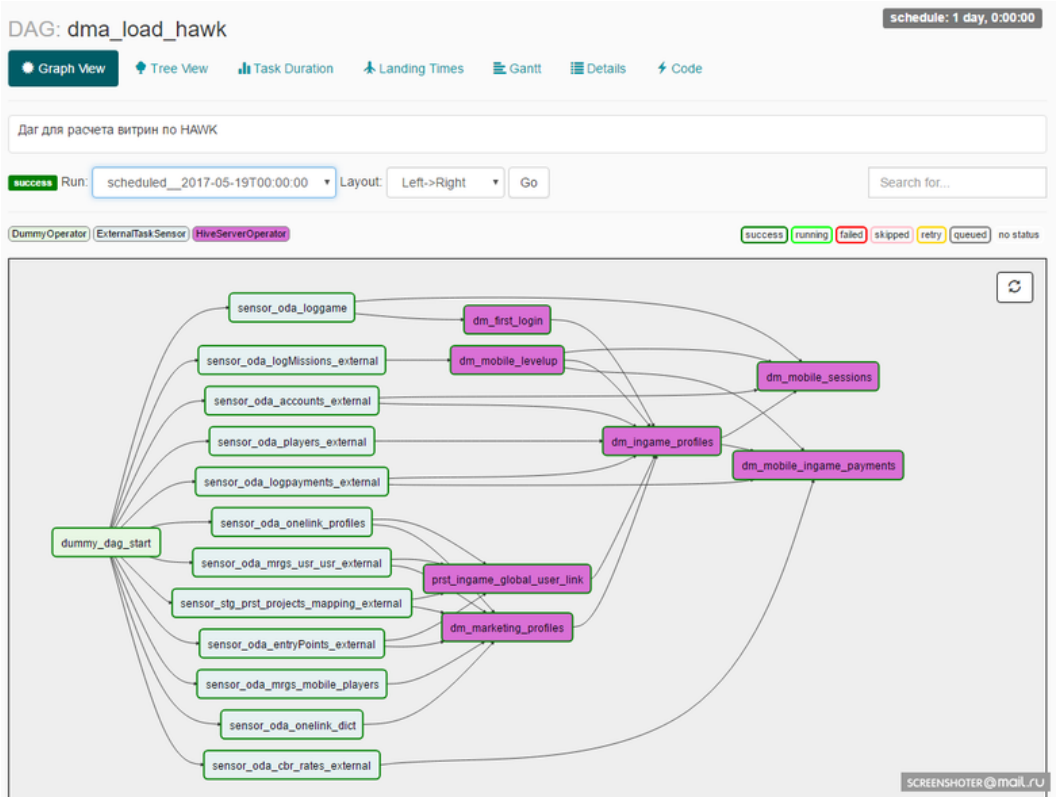

Рисунок 6

Витрины данных

Витрины данных (Рисунок 7) – это связанные наборы данных, которые в дальнейшем используются для построения визуализаций в BI системах и анализа аналитиками данных.

Рисунок 7 – пример витрины данных

Формирование витрины данных

Витрины данных формируются в автоматическом режиме. Для этого в разработанном интерфейсе выбираются из существующих данных те позиции, которые необходимы. Далее:

1. Автоматически генерируется скрипт на формирование и обновление витрины данных
2. Скрипт запускается, формируя витрину данных
3. Заказчику отправляется оповещение о готовности
4. В соответствии с указанной периодичностью обновления (по времени или по изменению исходных данных) настраивается процедура запуска скрипта на обновление данных.
5. После обновления данных заказчику отправляется оповещение.

Витрина имеет описательную часть, которая содержит следующую информацию:

1. Название витрины данных
2. Описание витрины
3. Используемые данные и процедуры их получения из сырых данных
4. Заказчик витрины
5. Дата формирования витрины
6. Срок действия витрины.

Удаление временной витрины

Удаление витрины сопровождается оповещением заказчика за указанное время до окончания срока действия. При этом заказчик может выйти с просьбой продлить срок действия витрины данных. Изменение срока действия витрины осуществляется службой поддержки.

Обновление витрины

Данные в витрине обновляются автоматически по мере обновления данных в хранилище сырых данных или через указанное время, в зависимости от выбранного варианта обновления.

По окончанию процесса приходит уведомление заказчику витрины.

Системы аналитических представлений

Системы аналитических представлений используются для всестороннего визуального анализа данных. В Счетной палате для этих целей используются Open Source инструменты Pentaho BI и Metabase. К сожалению, из-за ограничений по возможностям Open Source инструментов одним инструментом покрыть все необходимые требования не получается, поэтому используются оба инструмента и каждый из них закрывает свою нишу:

- Pentaho BI является более громоздкой и тяжелой системой, весьма требовательной к знаниям и умениям пользователя, но позволяющей строить более разнообразные аналитические представления, обогащая их возможностями языка Java Script. С использованием Pentaho BI созданы панели для оценки текущей деятельности Счетной палаты (Рисунок 8).

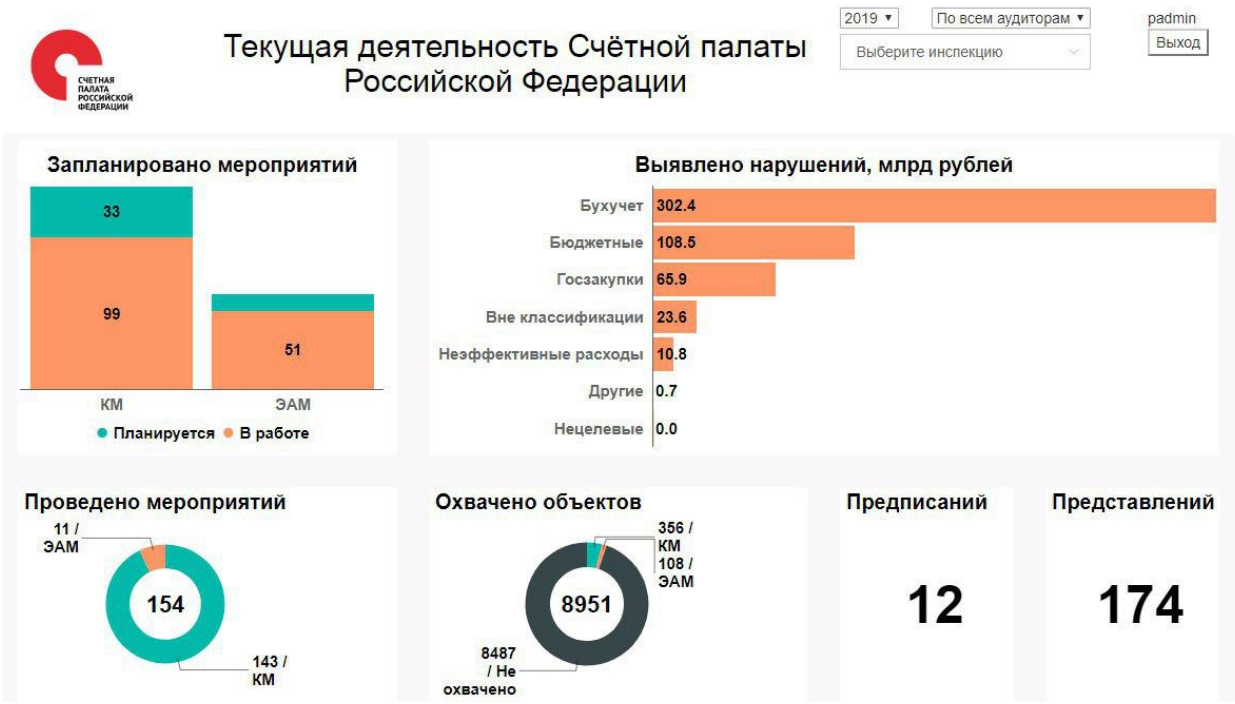

Рисунок 8 – пример визуализации в Pentaho BI

- Metabase (Рисунок 9) в свою очередь позволяет быстро строить аналитику собирая данные на ходу из различных источников и не очень требовательна к знаниям пользователя. Достаточно того, если пользователь умеет строить сводные таблицы и графики в Microsoft Excel, что несколько сложнее, чем делать подобное в Metabase. Но недостатком Metabase является отсутствие возможности настраиваемого перехода между наборами дашбордов, что есть в Pentaho BI, а также ограниченным набором аналитических представлений. С использованием данного инструмента созданы:
    - Аналитические представления для анализа деятельности сотрудников Счетной палаты по заполнению ими отчетов о проделанной работе
    - Аналитические представления по витрине данных, созданной для формирования аналитической записки по анализу бюджета ФОМС и дополнительных данных, получаемых от ФОМС.
    - Аналитические представления для анализа

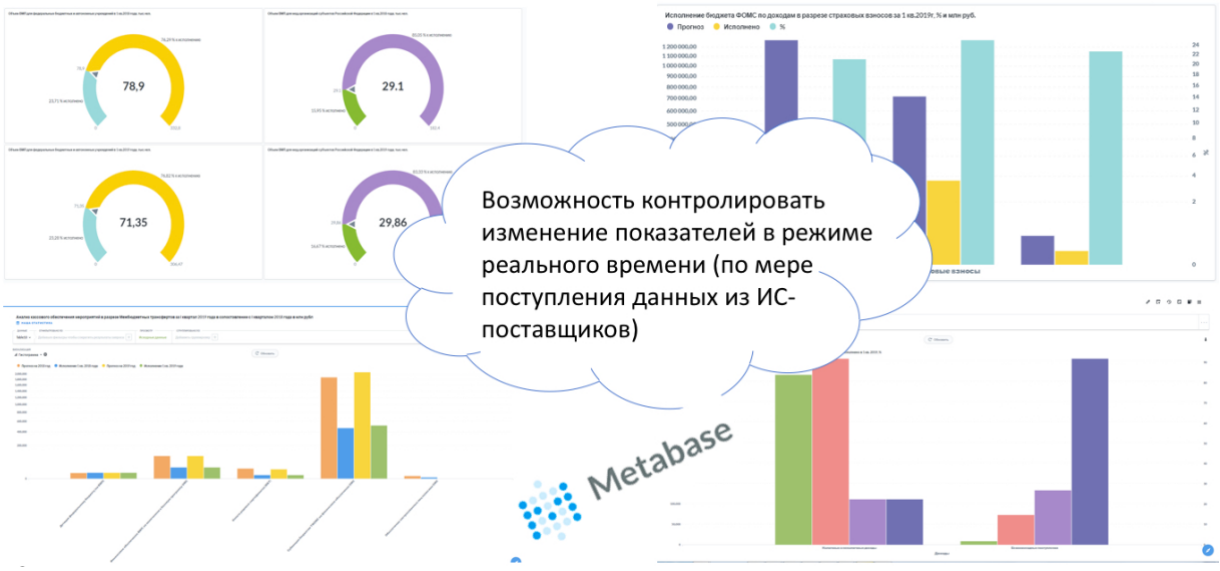

Рисунок 9 – пример представления в Metabase

Автоматизированные системы для сбора и обработки информации

Автоматизированные системы для сбора и обработки информации предназначены для получения качественных данных от объектов контроля и сокращения трудоемкости производства продуктов Счетной палаты и повышения их качества за счет внедрения новых технологий, аналитических возможностей и добавления новых источников данных и автоматической обработки этих данных.

Автоматизированная информационная система «Единая проектная среда» (АИС ЕПС)

Запущена в эксплуатацию автоматизированная информационная система «Единая проектная среда» (АИС ЕПС).

Основной целью взаимодействия Участников со Счетной палатой посредством АИС ЕПС является предоставление Участником в электронном виде сведений, запрашиваемых Счетной палатой в рамках проведения оперативного анализа исполнения и контроля за организацией исполнения федерального бюджета в текущем финансовом году, последующего аудита (контроля) и иных проверок путем поведения контрольных, экспертно-аналитических и иных мероприятий.

Пилотный проект «Цифровой департамент»

Пилотный проект «Цифровой департамент» Департамент цифровой трансформации Счетной палаты РФ проводит совместно с Департаментом аудита социальной сферы и науки.

Цель проекта: разработать технологии, позволяющие:

1. Кратно снизить трудовые и временные затраты ресурсы на рутинные операции традиционного аудита:
    a. Работы на объектах аудита
    b. Сбор и обработку информации
    c. Визуализацию данных
    d. Формирование типовых отчетов
    e. Подготовку первичных аналитических материалов
    f. Представление результатов
2. Повысить качество данных, обеспечить их надежность, полноту и единый формат

Один из результатов работы в рамках пилотного проекта – создание системы построения аналитической записки анализа бюджета ФОМС.

Эта автоматизированная система позволяет в автоматическом режиме получить данные из необходимых источников, сформировать витрину данных и с помощью специально разработанного шаблона, описания технологии получения данных для заполнения шаблона и программы-шаблонизатора формировать предзаполненную данными аналитическую записку с возможностью доводки полученного документа в Microsoft Word.

Для этого были созданы:

- Карта витрины данных (Рисунок 10) – описание источников данных, а также как и в каком виде эти данные должны собраться в витрину данных.
- Созданы требования к шаблону аналитической записки
- Разработана технология формирования шаблона аналитической записки (Рисунок 11)
- Разработана технология и форма описания переменных шаблона и кода для автоматического получения данных из витрины (Рисунок 12)
- Разработан программный код шаблонизатора на Python для заполнения шаблона

Рисунок 10 - карта витрины данных

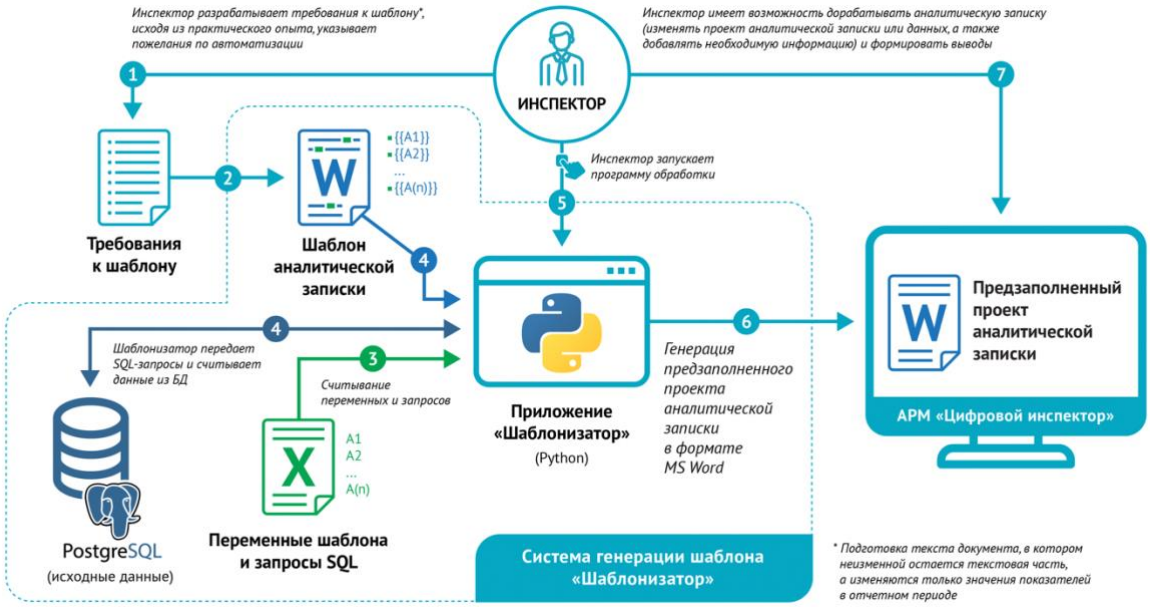

Рисунок 11 - Генерация предзаполненного проекта аналитической записки

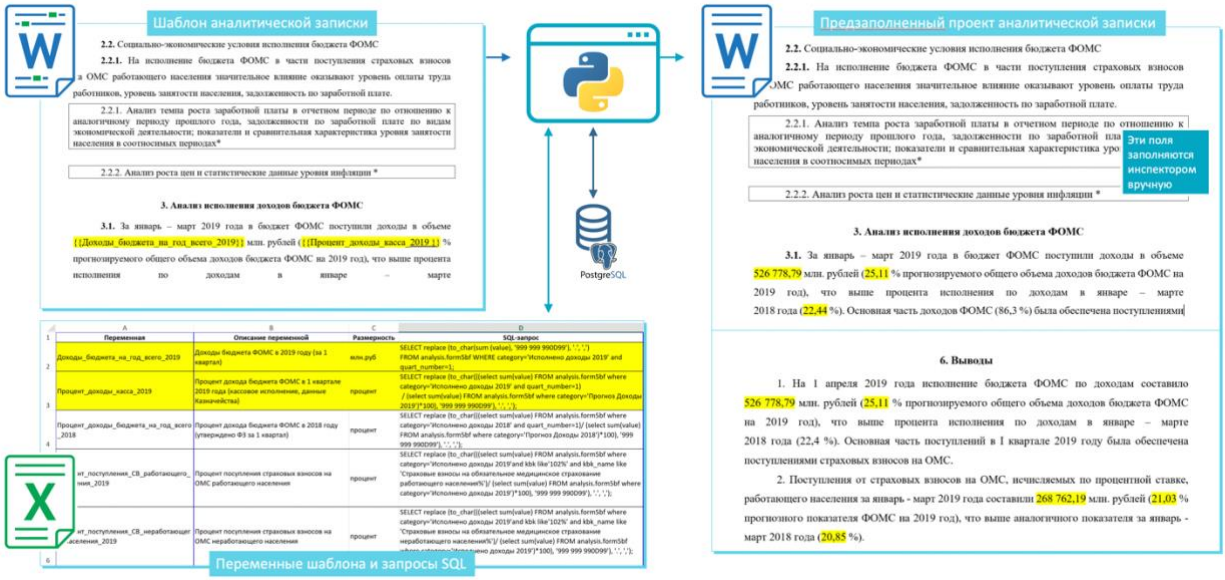

Рисунок 12 - процесс генерации проекта аналитической записки

Так же был создан набор аналитических представлений в Metabase (Рисунок 13) как для анализа используемых для построения аналитического отчета данных, так и более глубокого анализа используемой информации, обогащенной дополнительными данными из других источников.

Рисунок 13 - Аналитическая витрина данных для оперативного анализа исполнения бюджета ФОМС

Прототип автоматической классификации нарушений

Задача автоматической классификации нарушений по их описанию возникла из-за сложности и ошибок при ручной классификации нарушений. В прототипе решения задачи использовалась библиотека для Python - PyTorch[7].

Данные

- Получено – 15 000 примеров текстов с классом,
- в экспериментах участвуют 123 класса нарушений
- Не включены: “прочие нарушения” и 114 классов, у которых менее 10 примеров

Результаты

- Классификация по всем классам (123 шт.)
  - угадываем 1 класс по тексту - точность[8] 92%
  - даем 3 варианта на текст - точность угадывания правильного 94%
- Классификация по категориям верхнего уровня
  - точность угадывания категории по тексту - 96%

Аналитическая модель «Анализ профиля бедности в Ростовской области»

Цель: предоставить современный инструмент для контроля снижения уровня бедности в два раза до 2024 года (Указ Президента Российской Федерации от 07.05.2018 № 204)

Ожидаемый результат**:** создание информационного ресурса, содержащего полные данные о получателях мер социальной поддержки, необходимых для**:**

- проведения оценки реального уровня и структуры бедности,
- анализа причин бедности граждан и семей,

7 Библиотека машинного обучения для языка Python с открытым исходным кодом, созданная на базе Torch. Используется для решения различных задач: компьютерное зрение, обработка естественного языка.

8 Точность - количество правильных ответов ко всем ответам

- создание региональных реестров граждан с доходами ниже прожиточного минимума,
- развитие системы социальной помощи и её предоставления исходя из принципов адресности.

В результате создан прототип, позволяющий, при подключении к нему в качестве источника реальных данных ФНС, ПФР и других госорганов, построить набор аналитических представлений для анализа профиля бедного гражданина и бедной семьи.

Прототип представлен в виде преднастроенной виртуальной машины. Решение является масштабируемым и открытым, с дружелюбным современным интерфейсом (Рисунок 14, Рисунок 15) и возможностью гибко манипулировать данными.

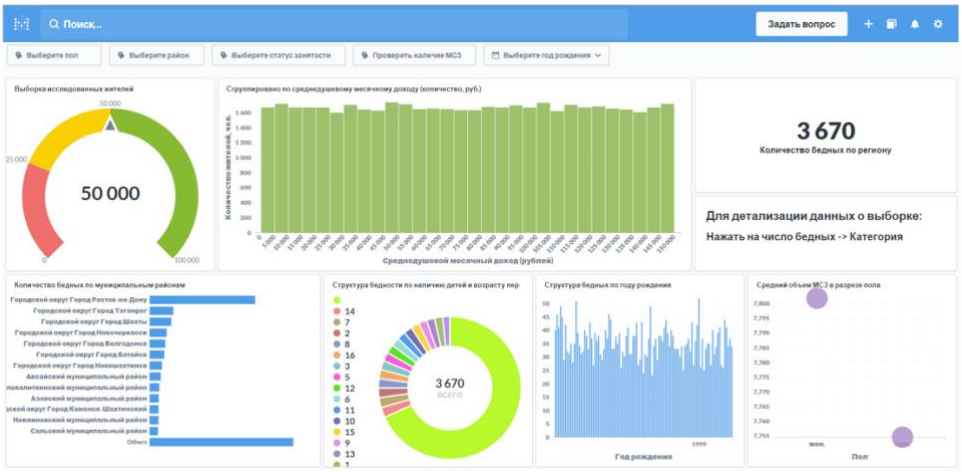

Рисунок 14

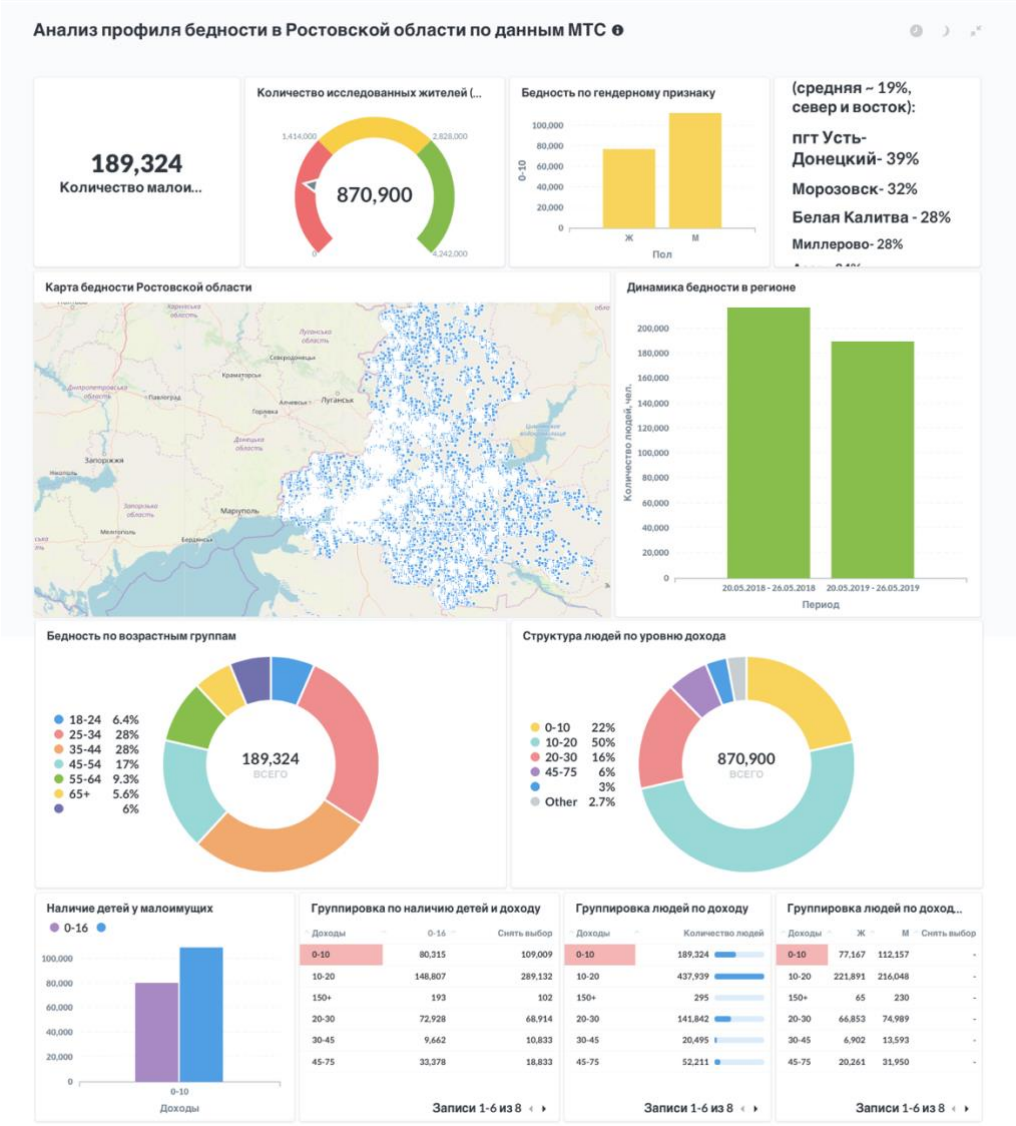

Рисунок 15

Мы рассматривали варианты с различными инструментами, в первую очередь open source. Связку PostgreSQL, Python, R и Metabase сочли оптимальным решением (Рисунок 16).

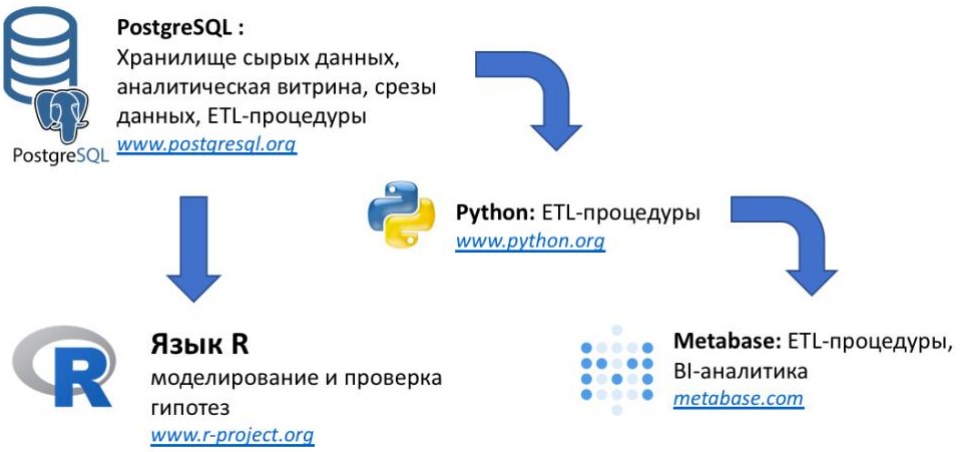

Рисунок 16 - Схема решения