|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА** – **Российский технологический университет»**  **РТУ МИРЭА** | | | | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики | |

|  |  |
| --- | --- |
| **ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 2** | |
| **по дисциплине «Большие данные»** | |
|  | |
| Выполнил студент группы КМБО-07-22 | Невский В.Е. |
| Проверил ассистент кафедры ПМ ИИТ | Горячев А.А. |

Москва 2024

Оглавление

[Задание 1 3](#_Toc162571035)

[Задание 2 4](#_Toc162571036)

[Задание 3 6](#_Toc162571037)

[Задание 4 8](#_Toc162571038)

[Задание 5 10](#_Toc162571039)

[Задание 6 11](#_Toc162571040)

[Задание 7 12](#_Toc162571041)

[Задание 8 13](#_Toc162571042)

[Задание 9 15](#_Toc162571043)

[Задание 10 16](#_Toc162571044)

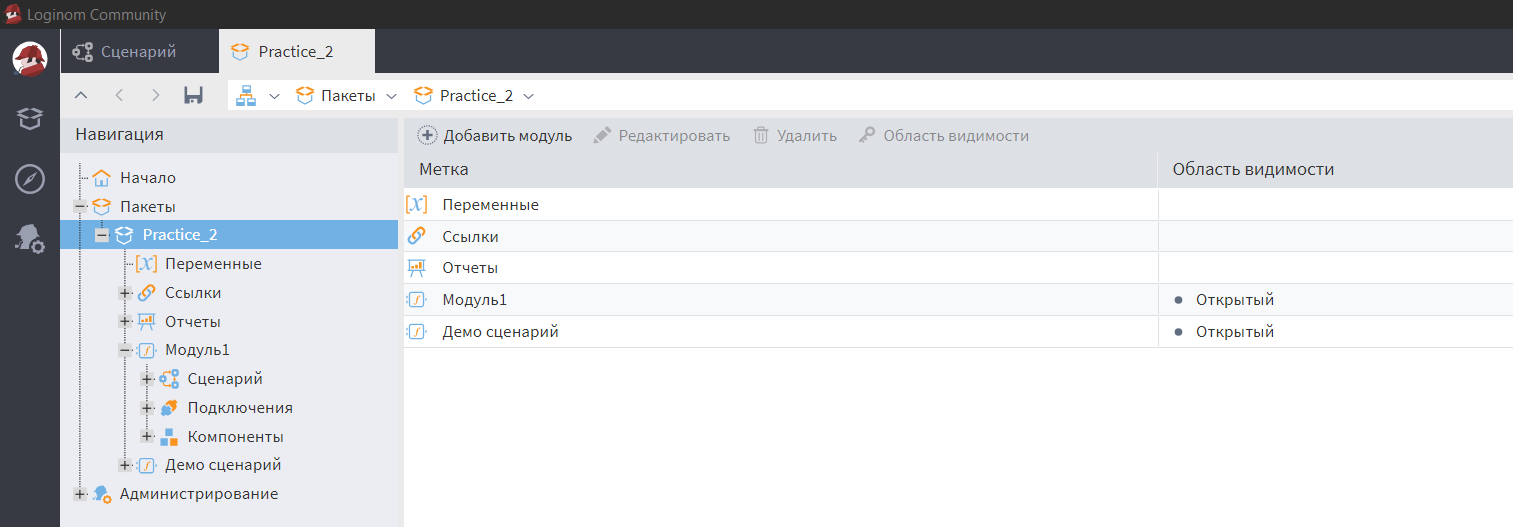
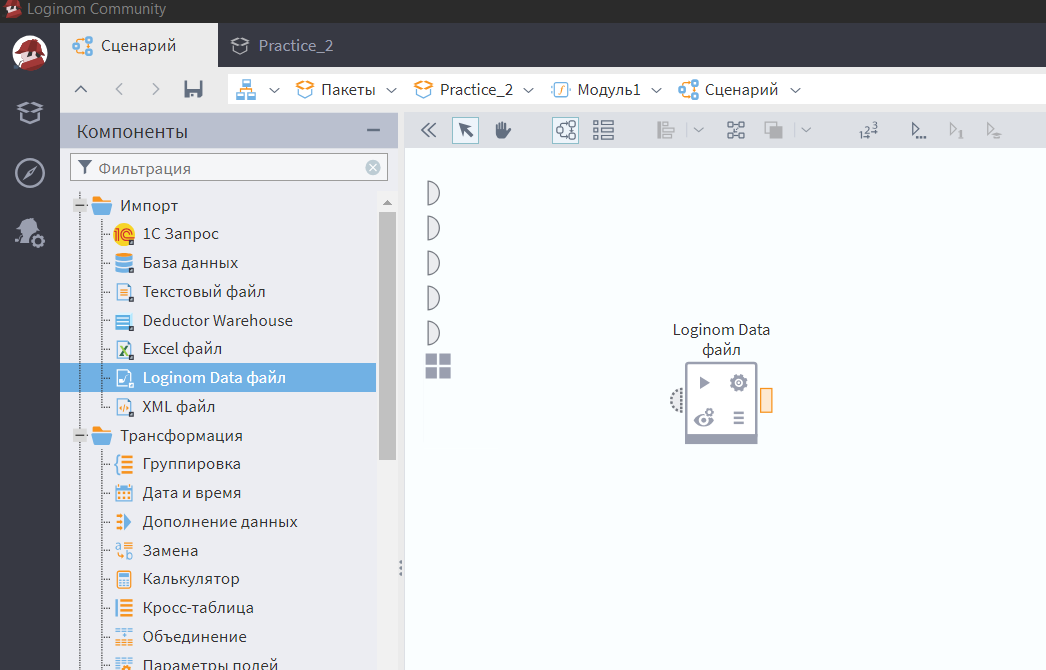
[Задание 11 17](#_Toc162571045)

[Задание 12 18](#_Toc162571046)

[Самостоятельная работа 19](#_Toc162571047)

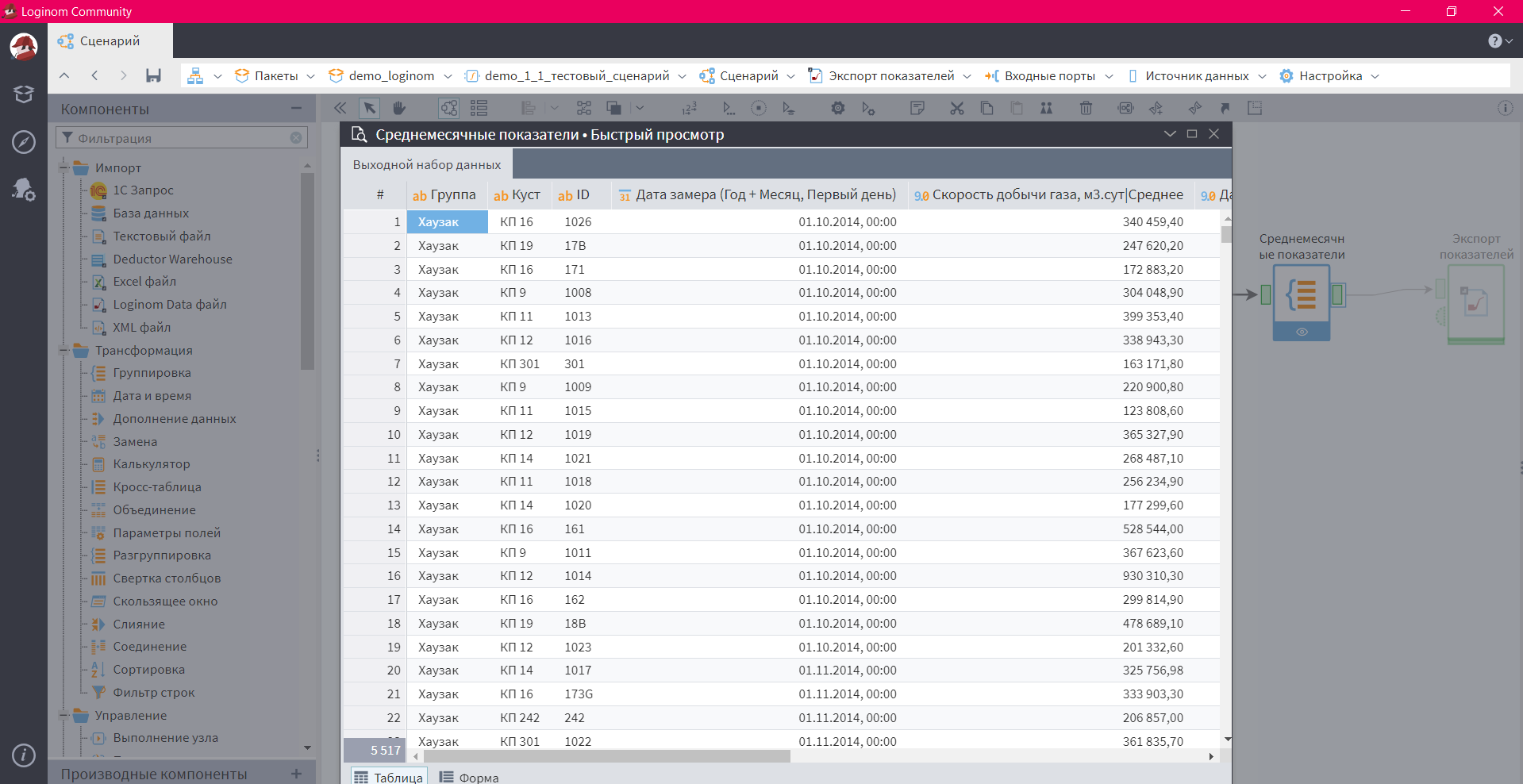
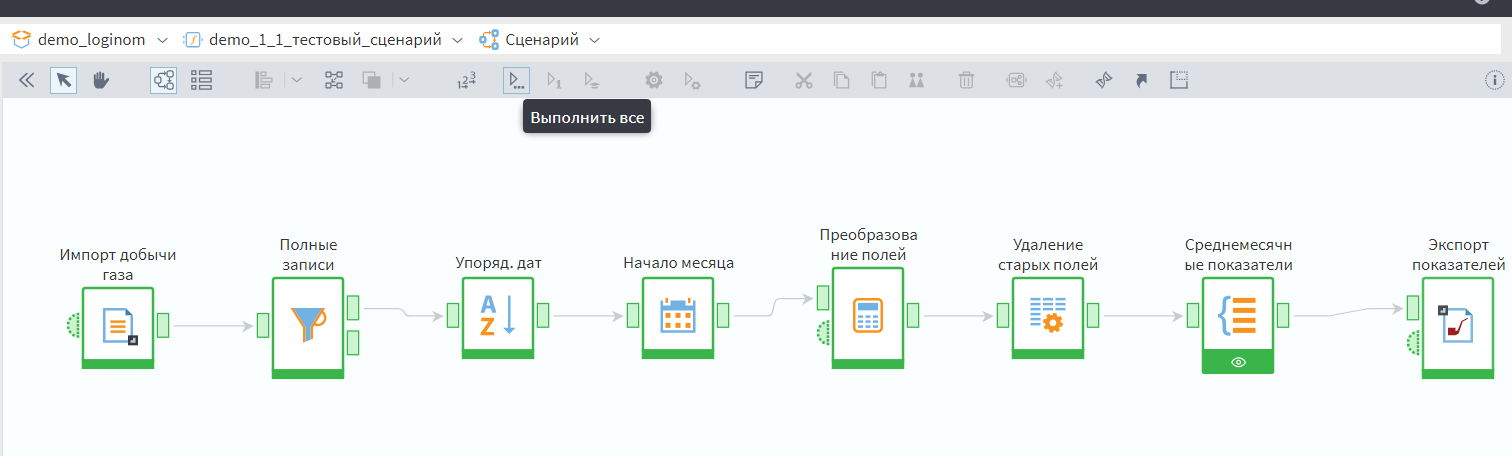
# Задание 1

* Создать пакет Loginom, сохранить его. Создать модуль внутри пакета с названием “Демо сценарий”. Перенести на полотно компонент “Loginom Data файл”. Сохранить пакет.



# Задание 2

* Открыть пакет demo\_loginom.lgd. Открыть модуль demo\_1\_1\_тестовый\_сценарий. Запустить все узлы. Исследовать входные и выходные данные.



* Сколько полей имеет таблица данных на выходном порте «Импорт добычи газа»?

Ответ: 9.

* Есть ли пропуски в записях импортируемой таблицы данных?

Ответ: Да.

* Сколько полных записей на верхнем порте узла «Полные записи»?

Ответ: 28016.

* Какое поле появилось после выполнения узла «Начало месяца»?

Ответ: Дата замера (Год + Месяц, Первый день).

* Какие поля добавляются в калькуляторе «Преобразование полей»?

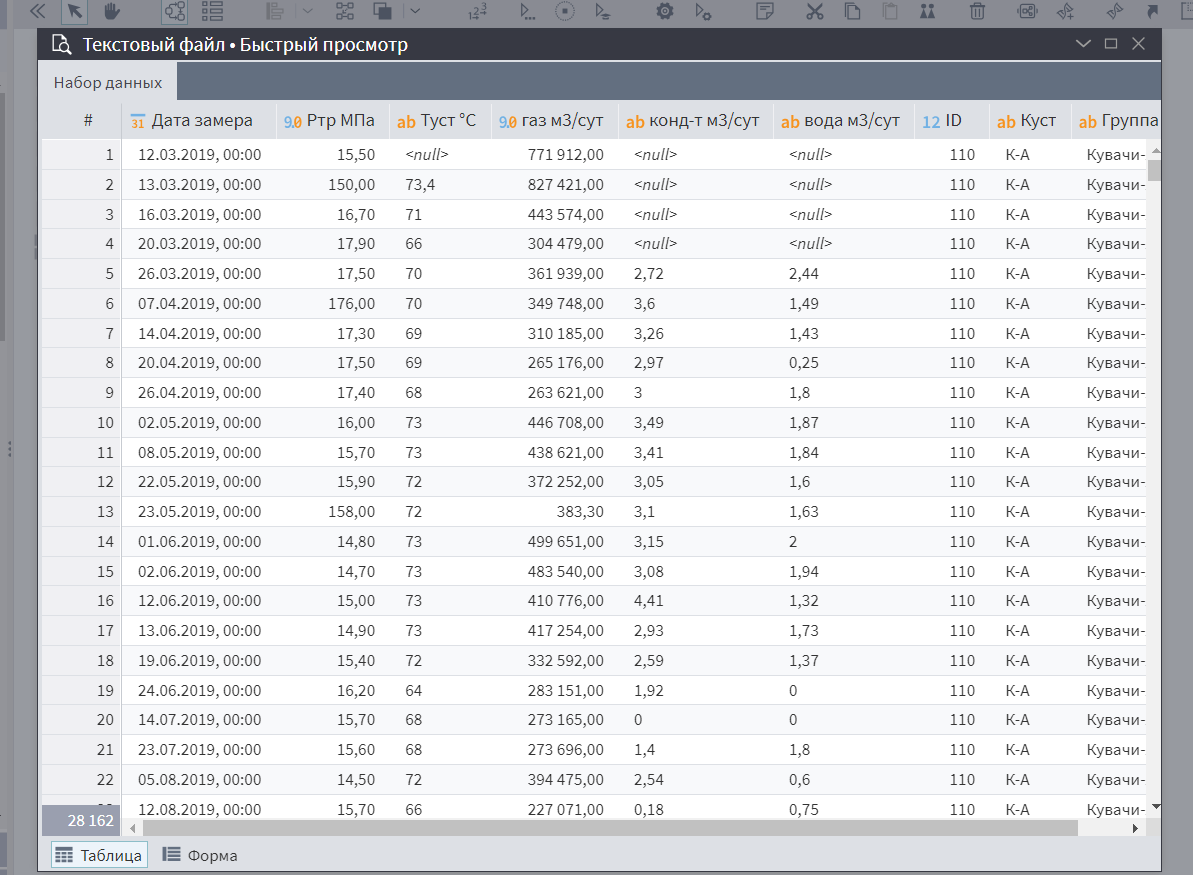
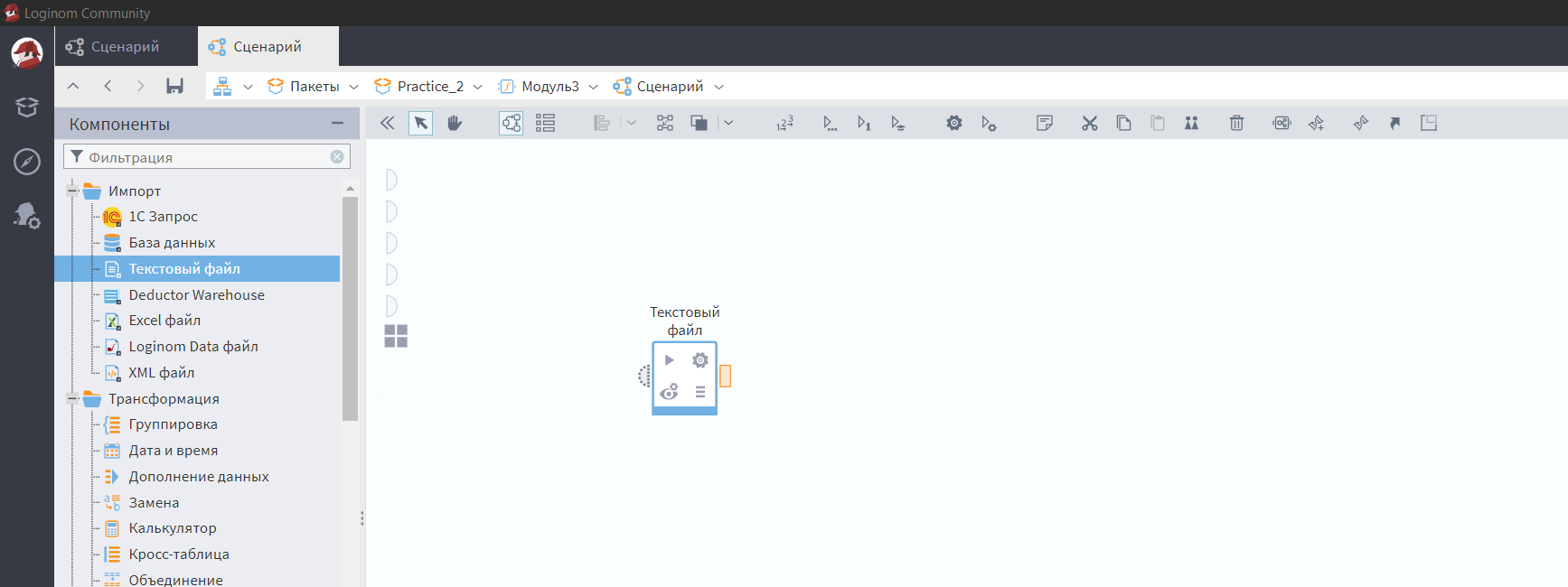
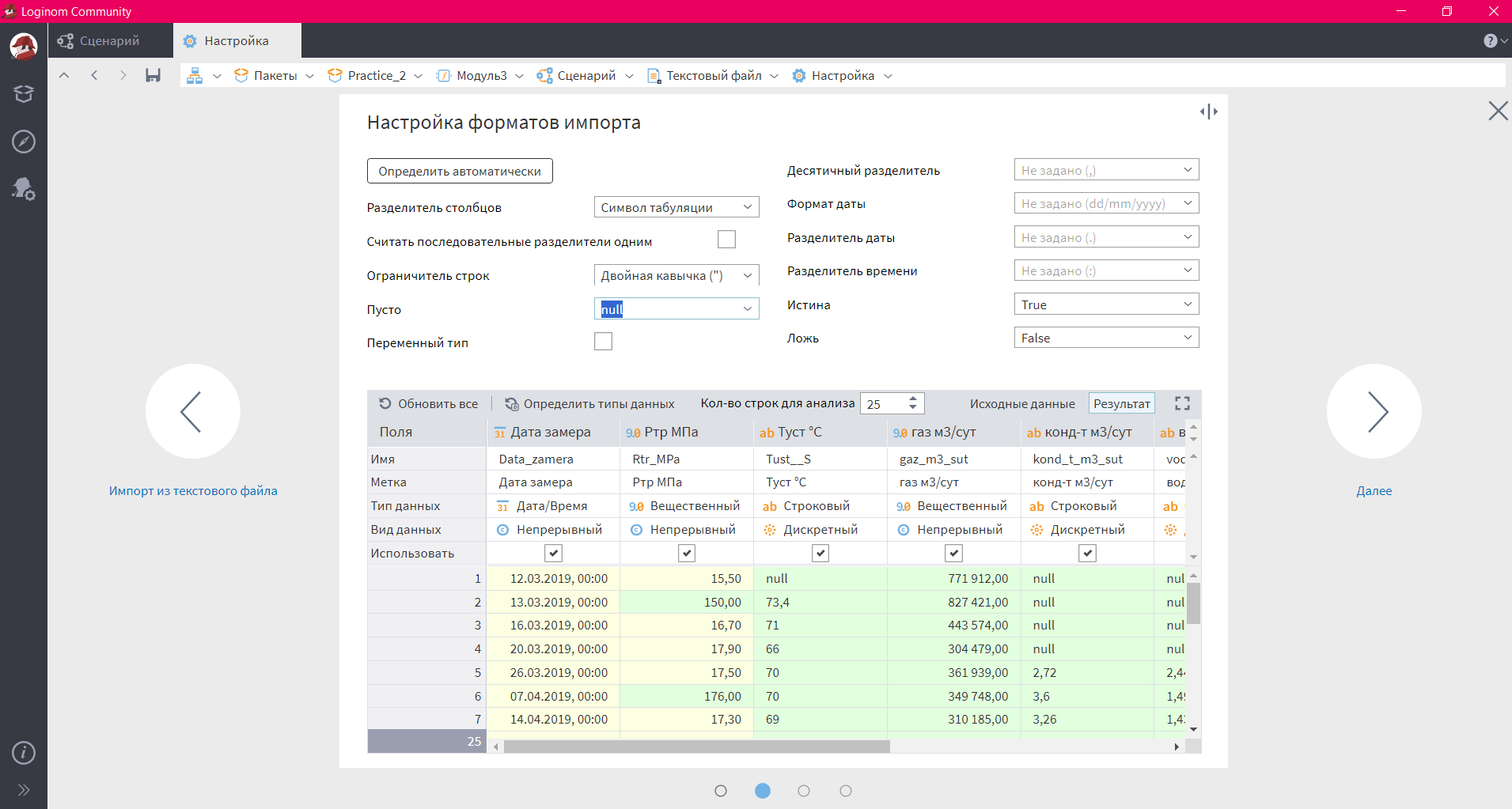
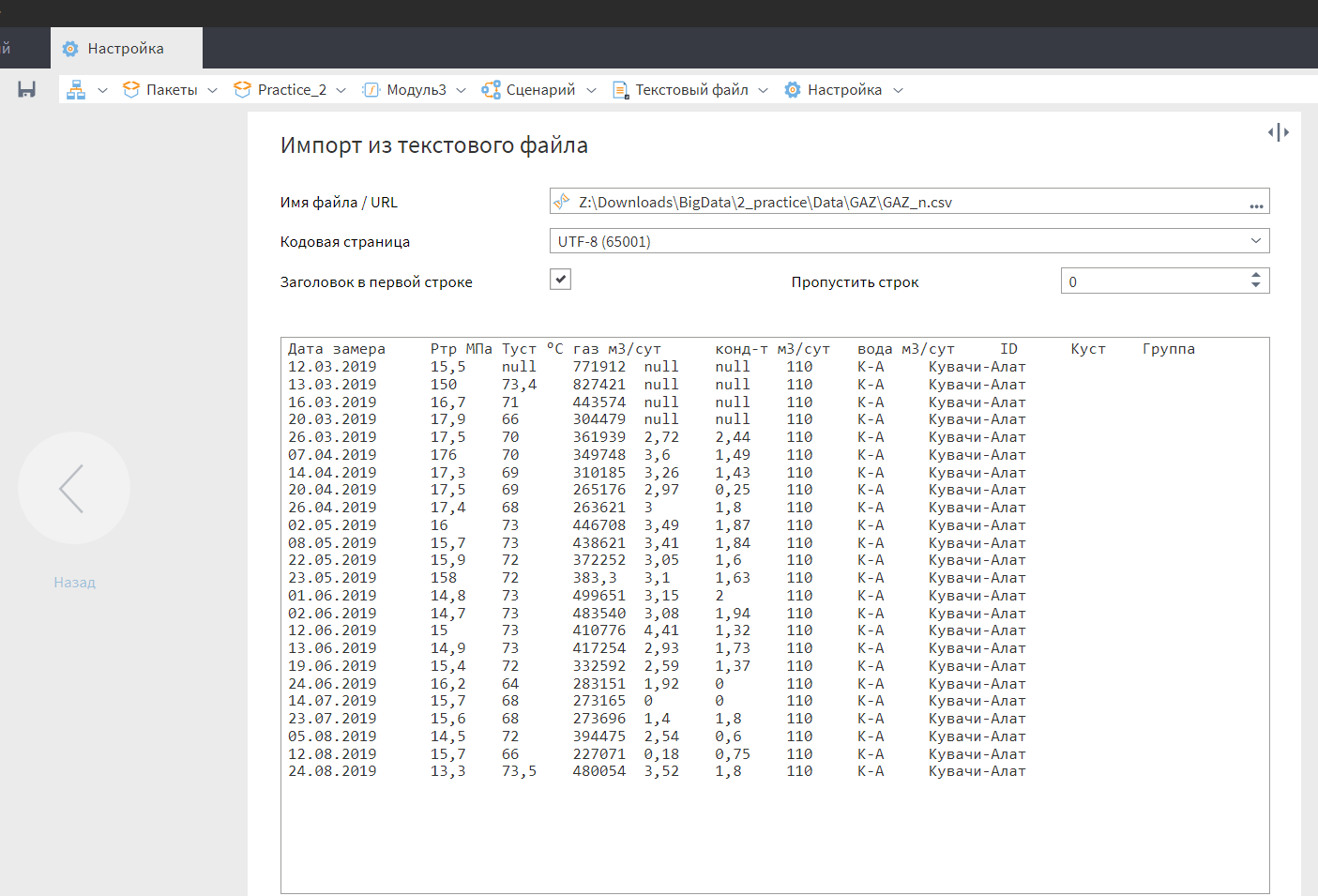
Ответ: Давление в трубе, Мпа, Температура устьевая, Цельсий, Скорость добычи газа, м3., Температура устьевая, Кельвин.

* Сколько записей остается после агрегации на выходе из узла «Среднемесячные показатели»?

Ответ: 5517.

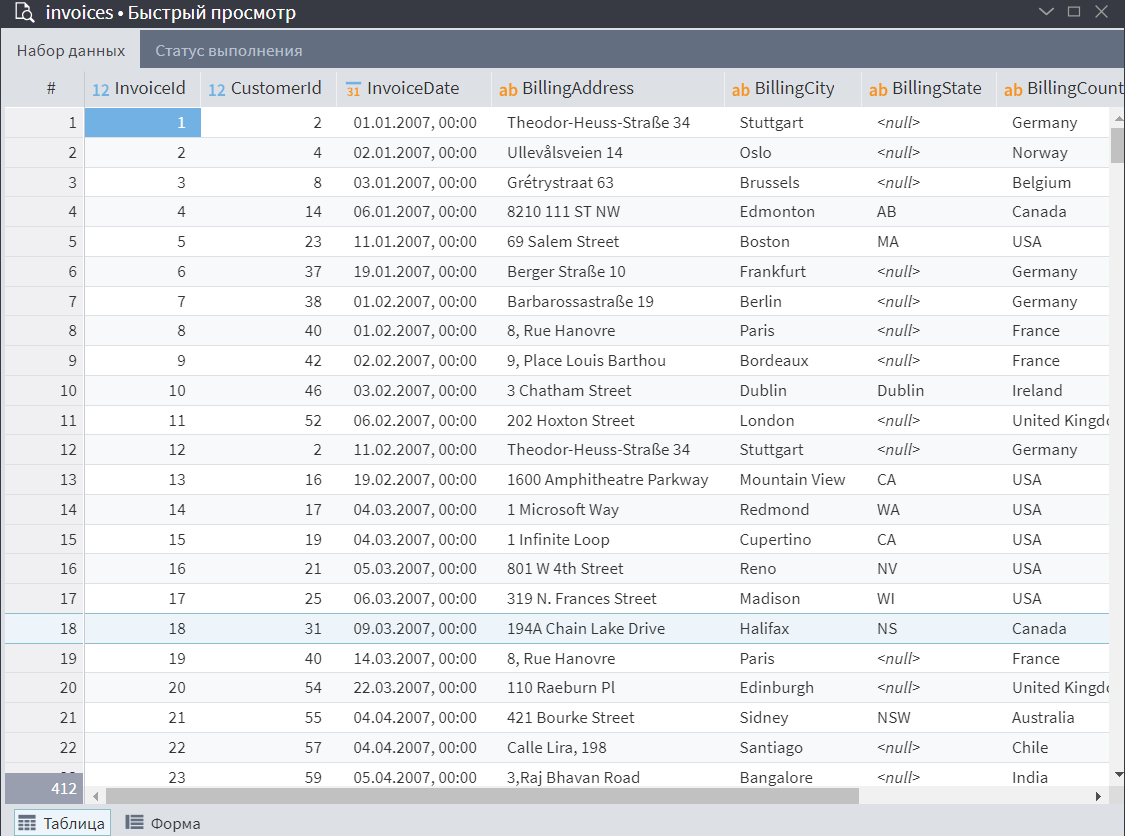
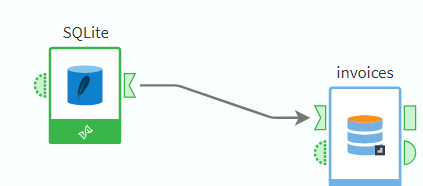
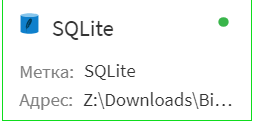
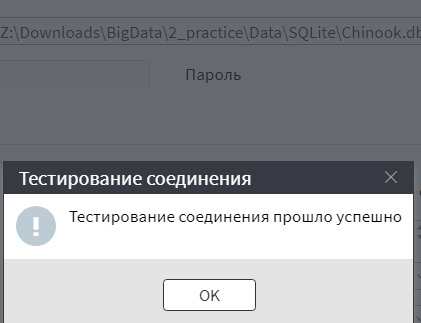
# Задание 3

* Открыть ранее созданный пакет lgp. Создать новый модуль под названием “Импорт из текстового файла”. Зайти в сценарий и перенести компонент “текстовый файл”. Настроить импорт данных.



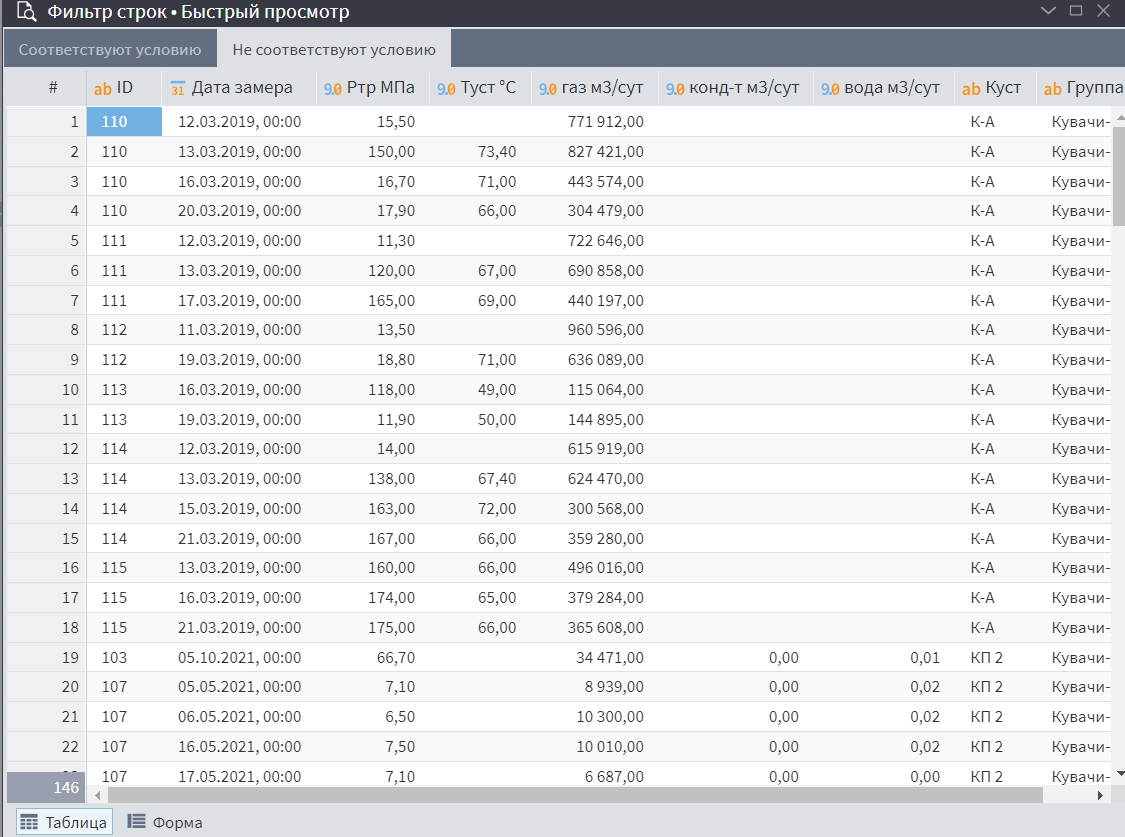
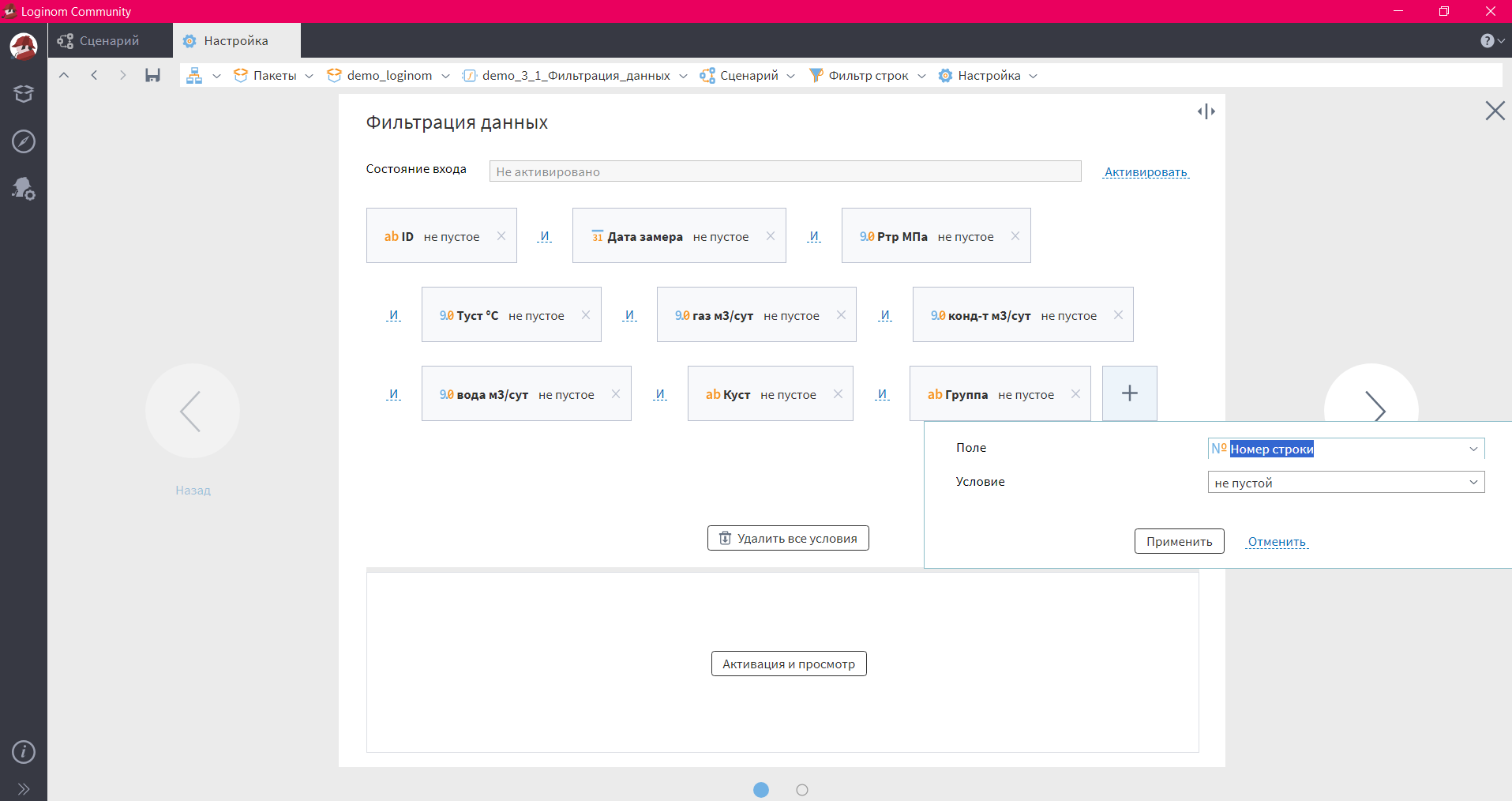
# Задание 4

* Модуль demo\_2\_3\_импорт\_бд, создать новый модуль и во вкладке "подключения” добавить компонент SQLite.



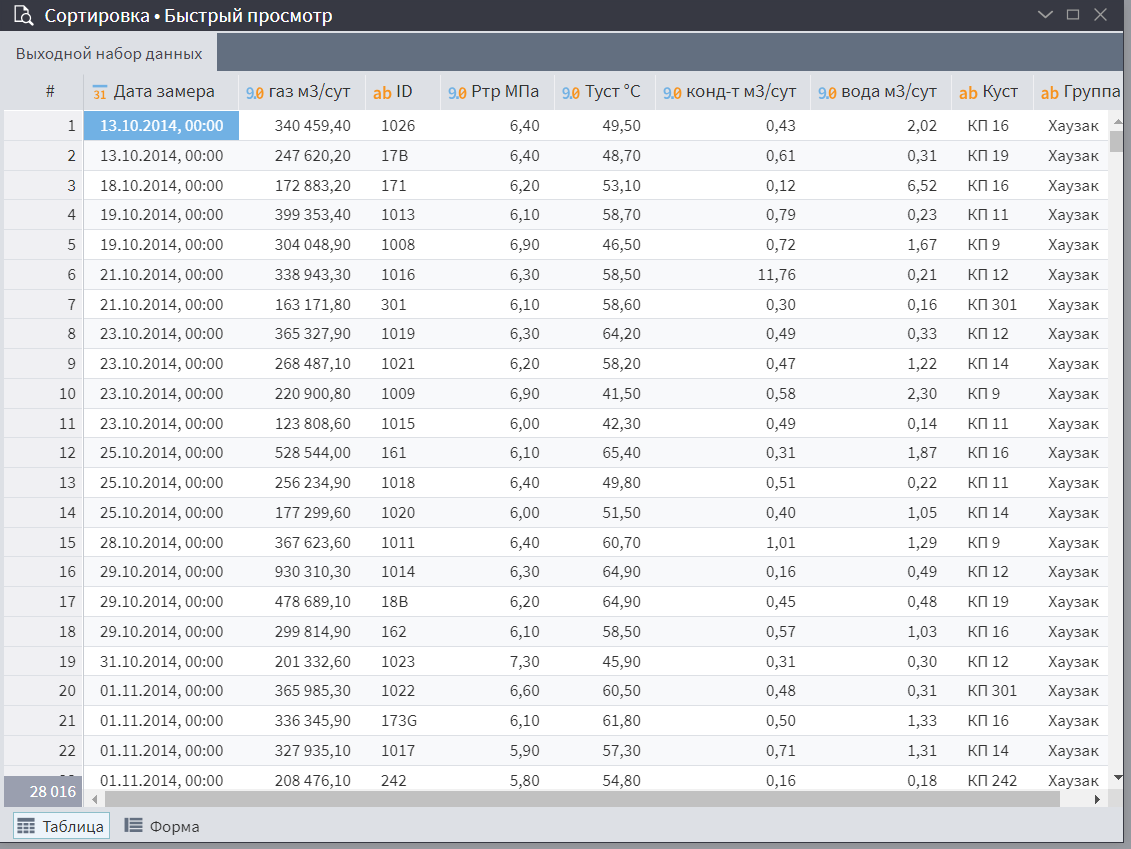
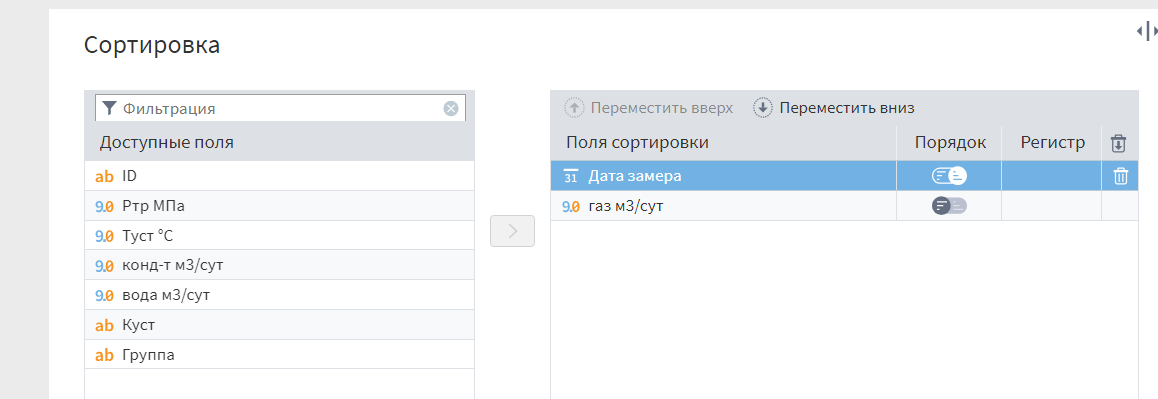
# Задание 5

* Модуль\_3\_1\_Фильтрация\_данных. Настроить фильтр для указанных данных.



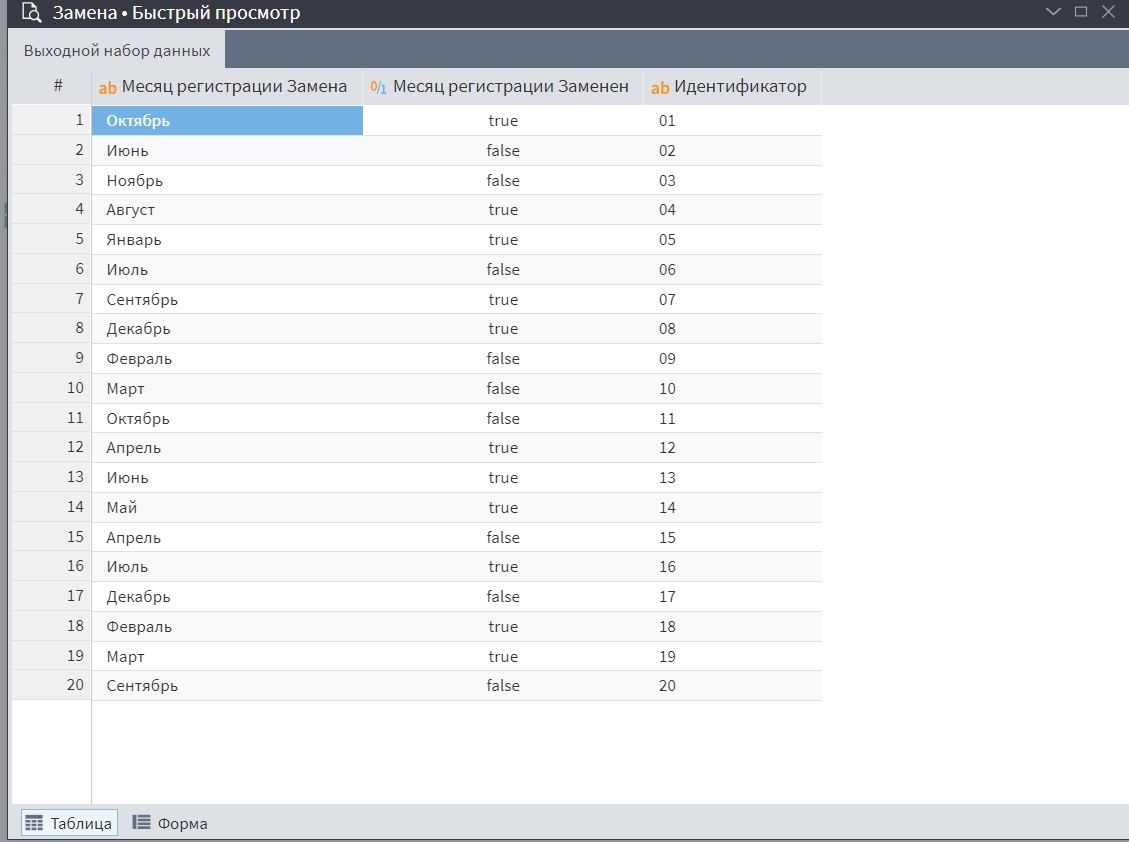
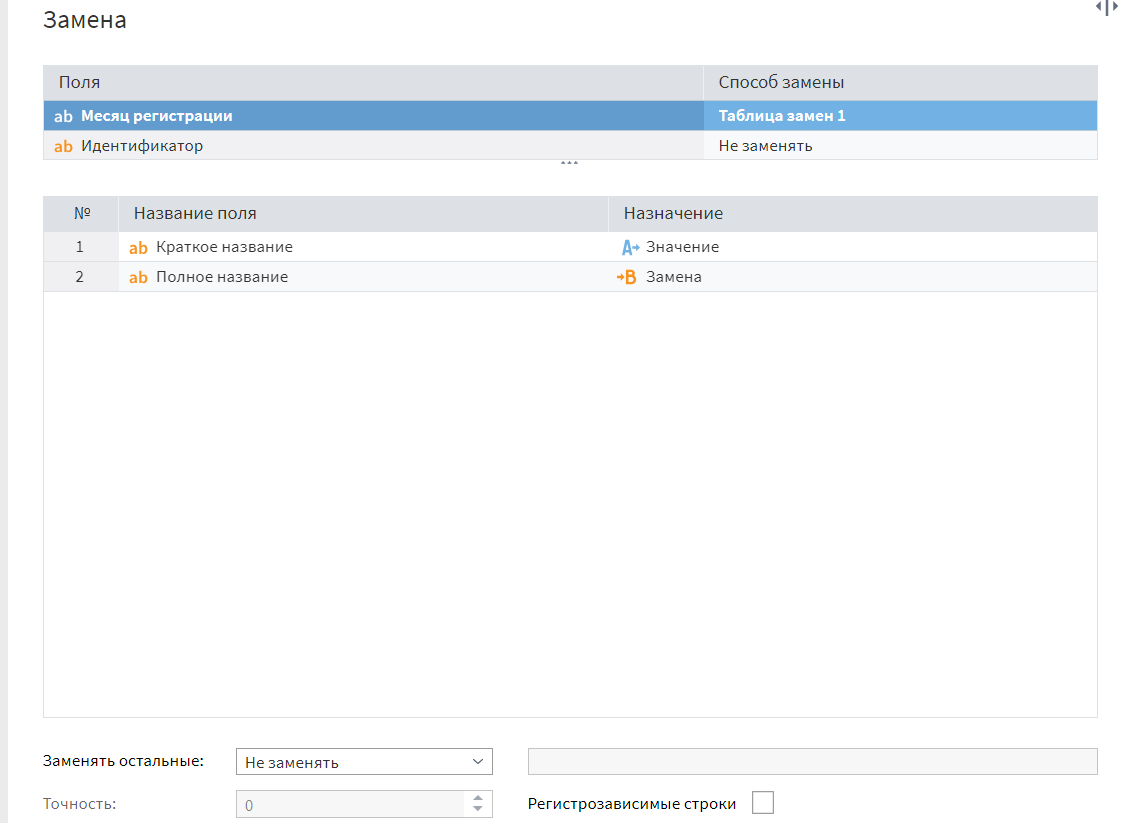
# Задание 6

* Модуль demo\_3\_2\_Сортировка\_данных. Для данных из предыдущего пункта применить сортировку.



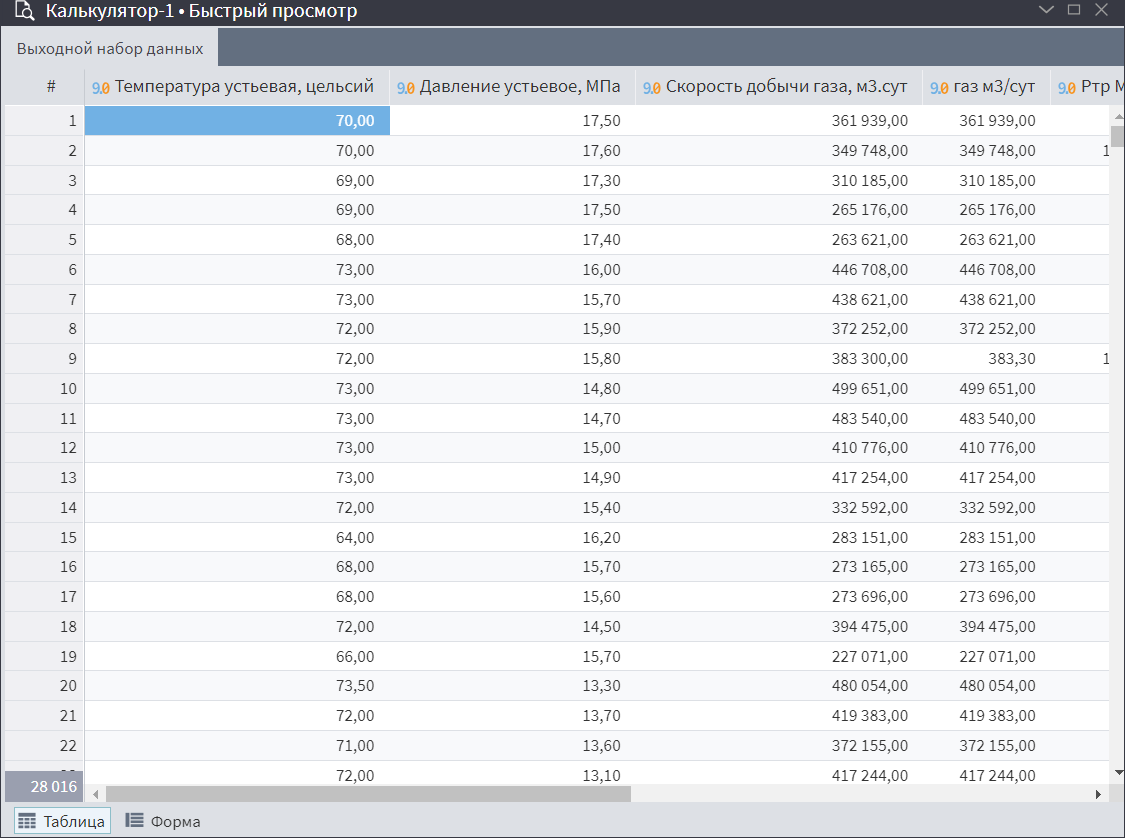
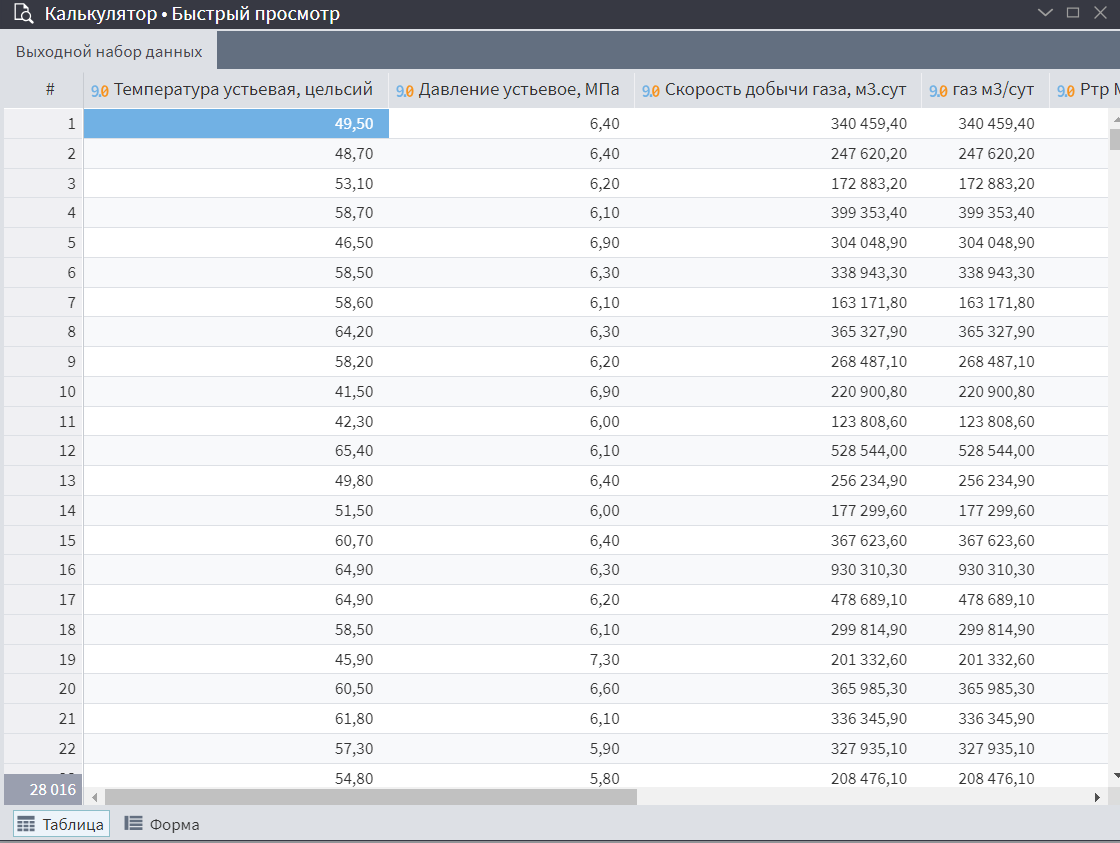
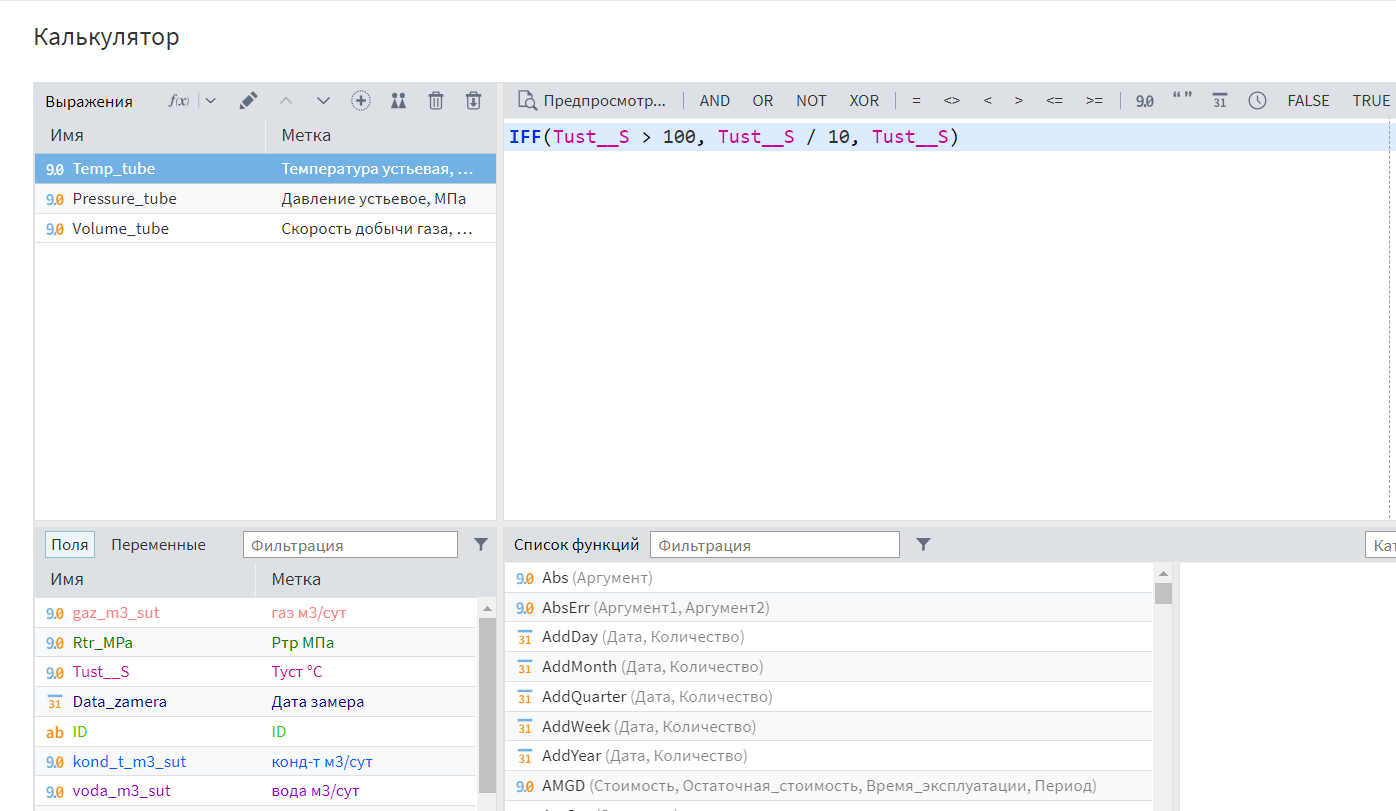
# Задание 7

* Модуль demo\_3\_3\_замена, необходимо заменить краткие названия месяцев на полные.



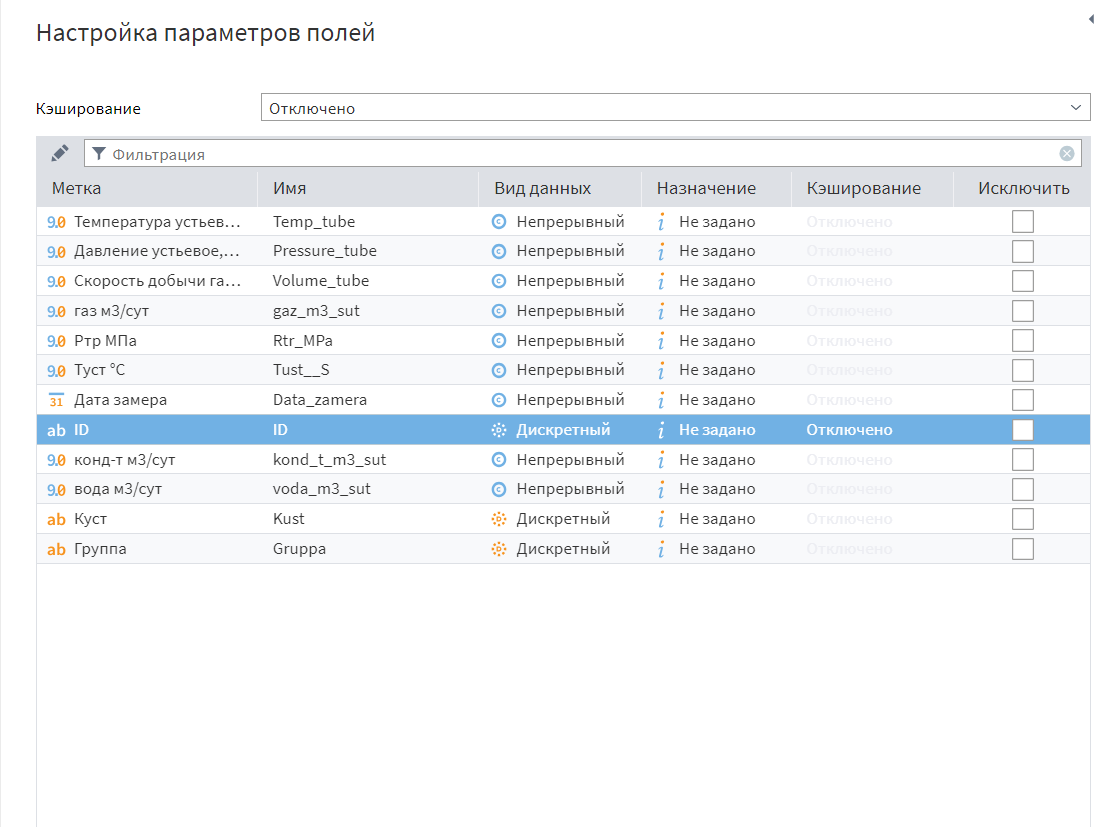
# Задание 8

* Модуль demo\_3\_4\_калькулятор, подкорректировать числовые значения некоторых полей.



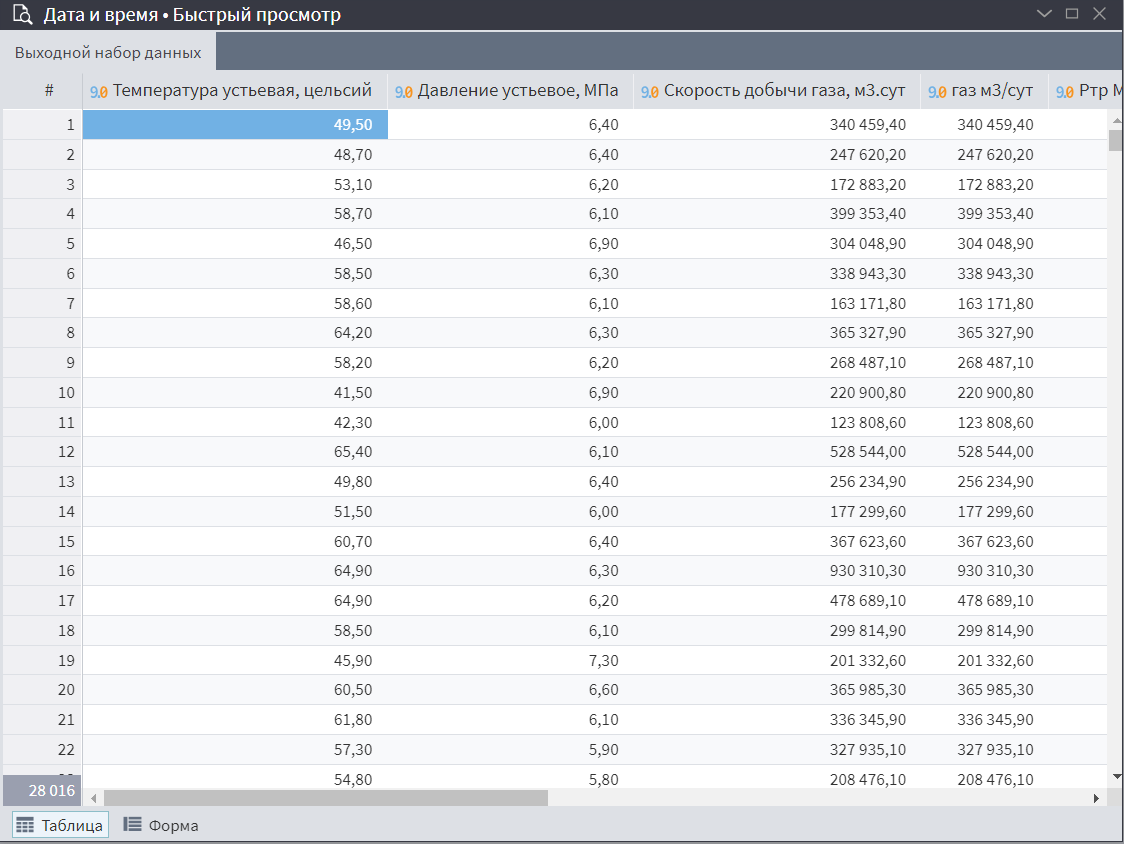
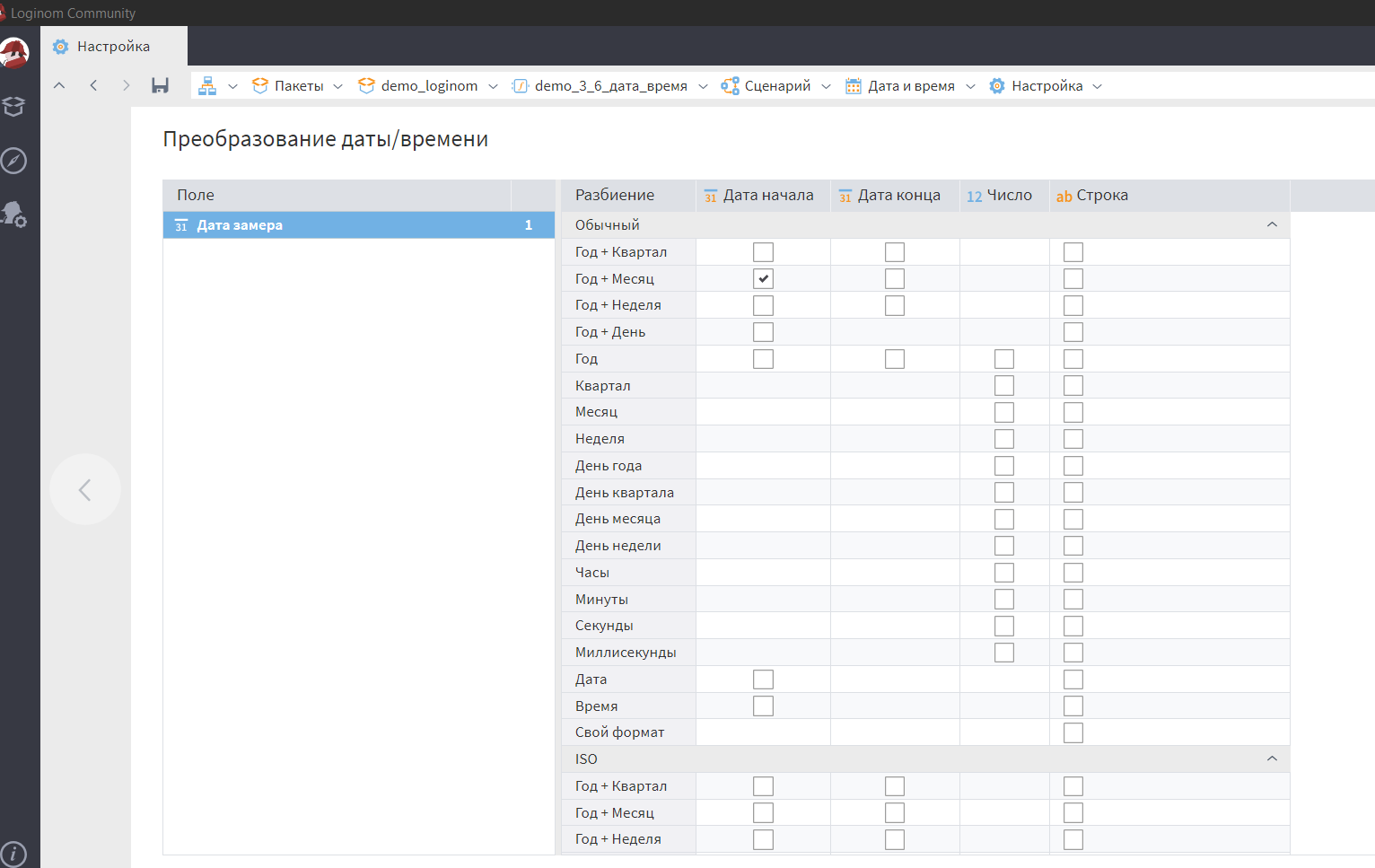
# Задание 9

* Модуль demo\_3\_4\_калькулятор, добавить компонент “параметр полей” и сделать некоторые преобразования.



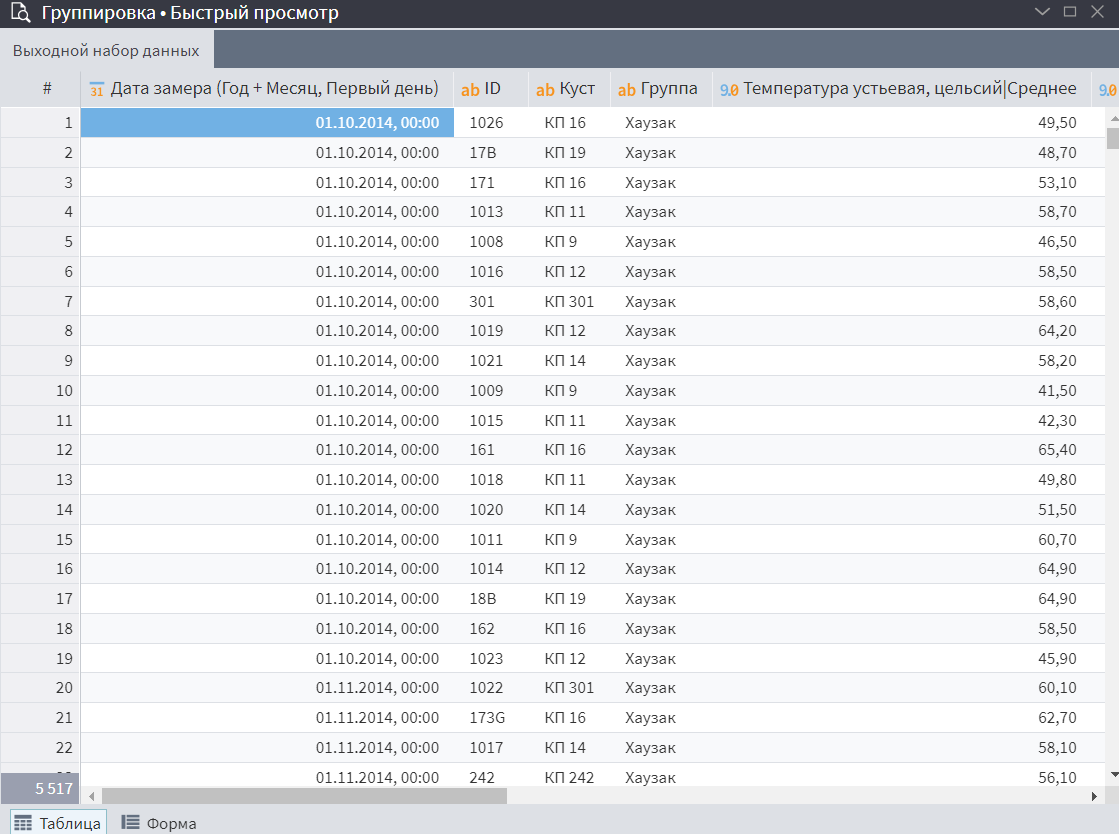
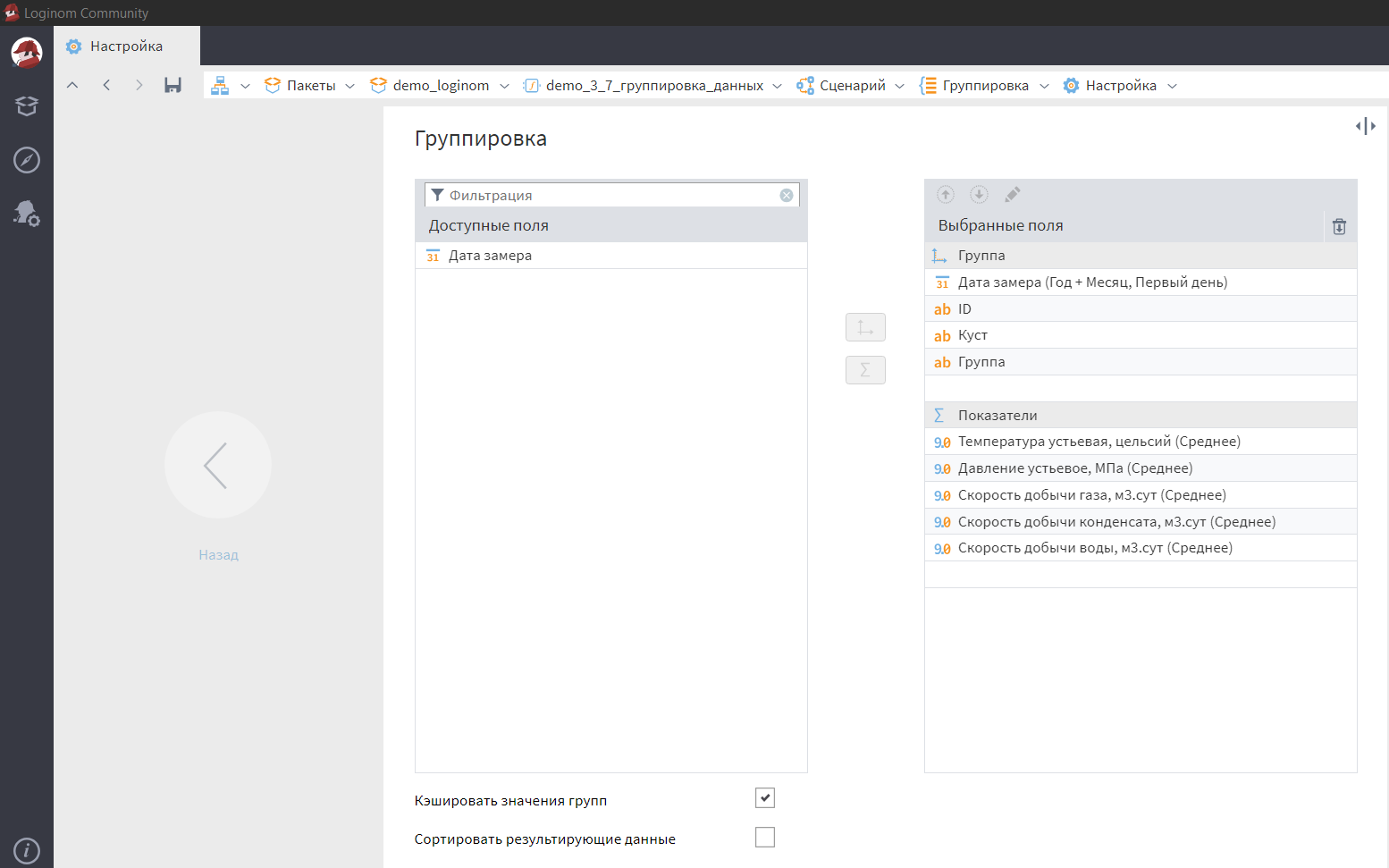
# Задание 10

* Модуль demo\_3\_6\_дата\_время, добавить компонент даты и времени и настроить его.



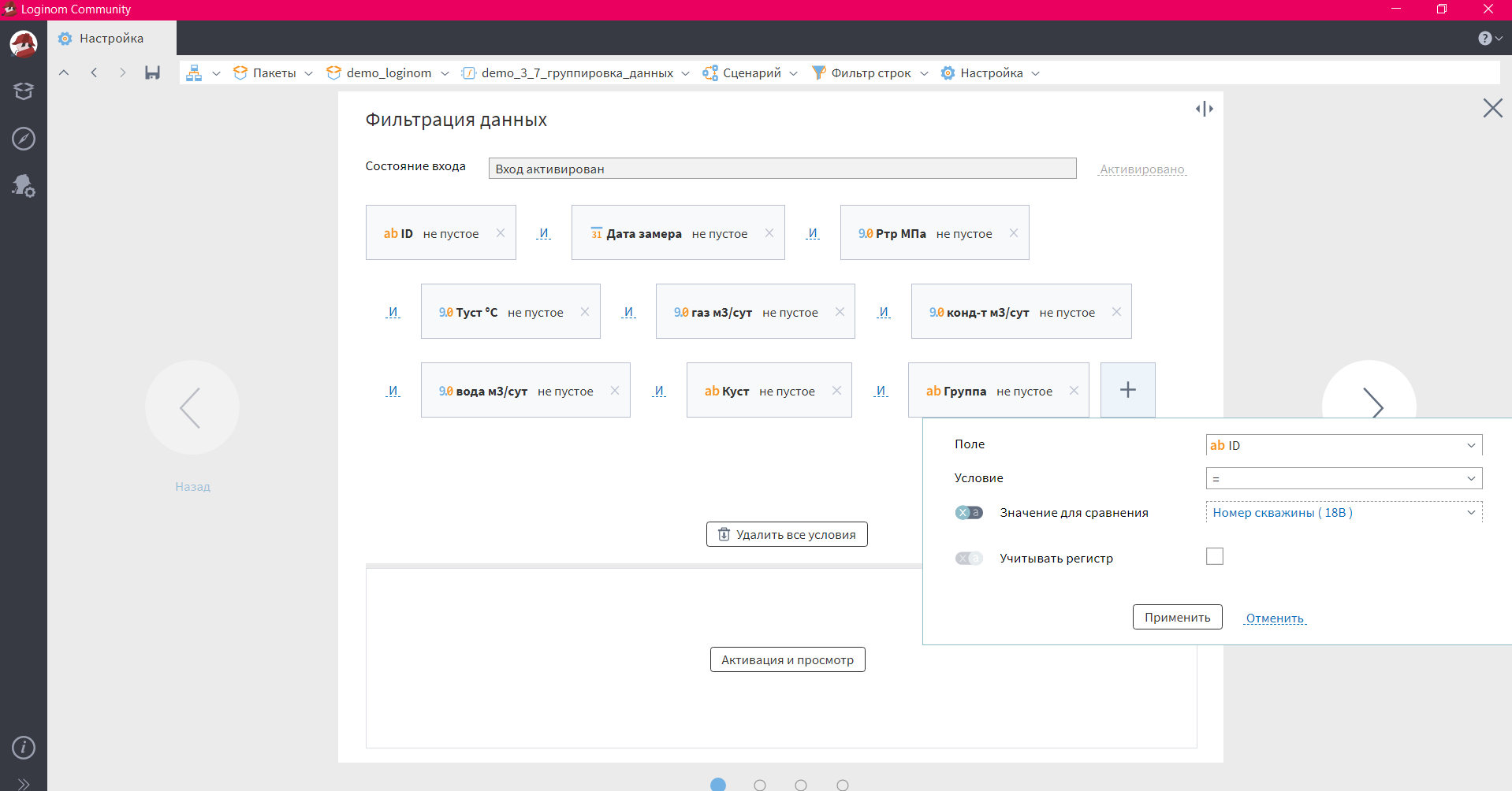
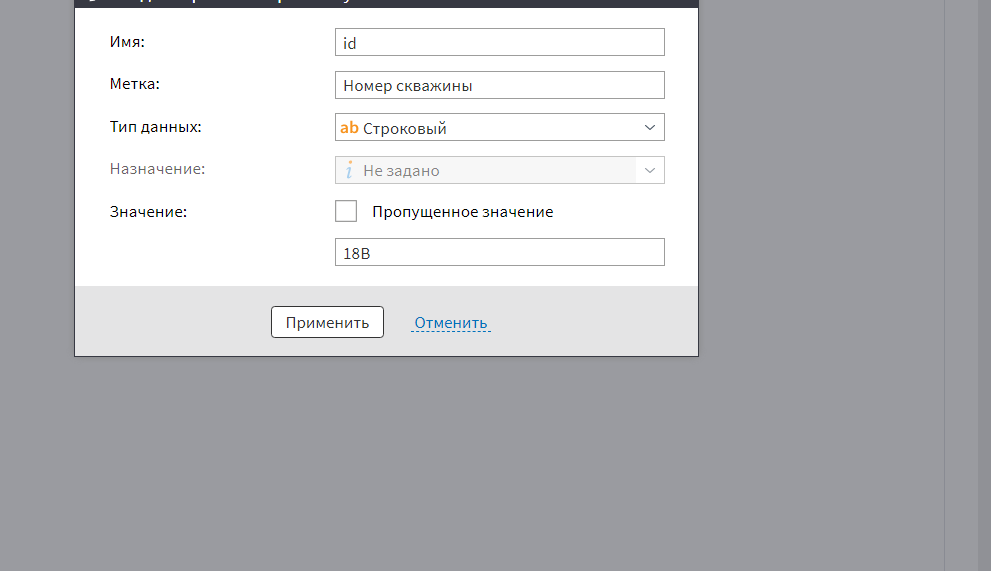
# Задание 11

* Модуль demo\_3\_7\_группировка\_данных, добавить компонент “группировка” и сделать группировку по параметрам.



# Задание 12

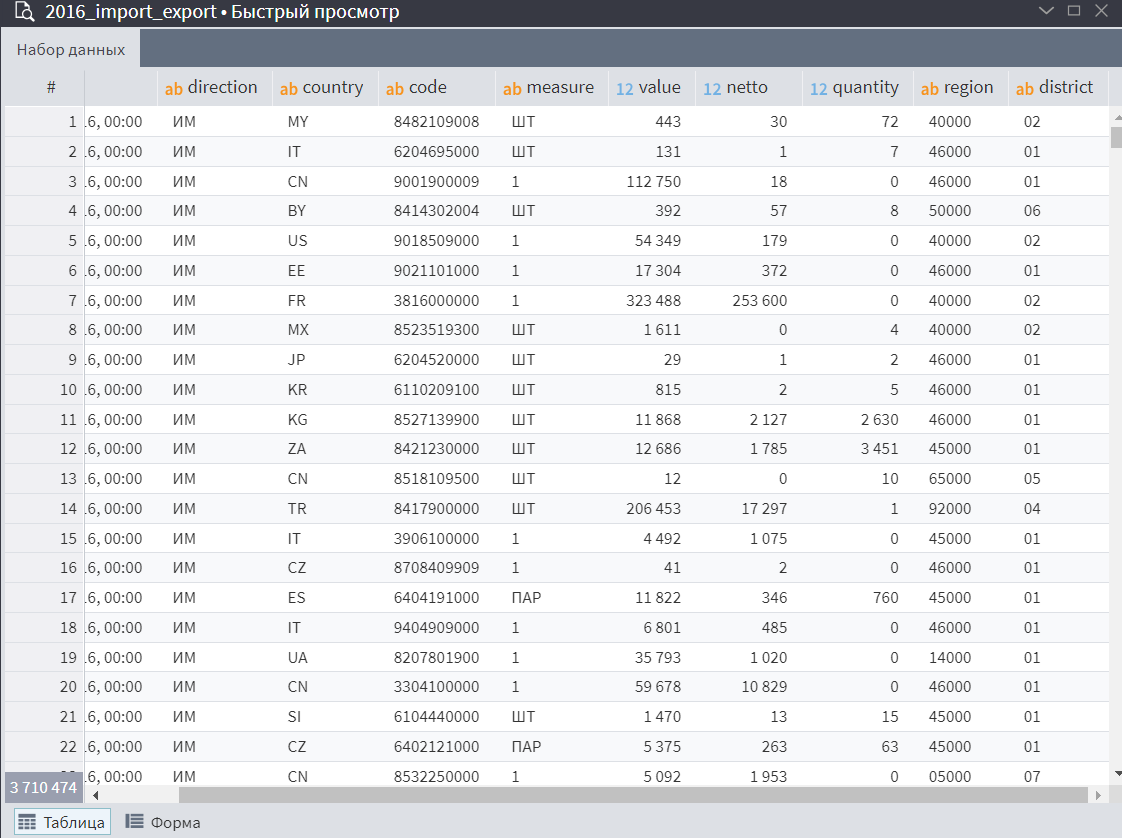
* Добавить переменную и проверить работу с ней на примере сравнения в фильтре.



# Самостоятельная работа

## Часть 1

Импортировать данные таможенной статистики по экспорту и импорту в регионах РФ за 2016 год.



* Сколько строк имеет выборка данных за 2016 год?

Выборка имеет 3 710 474 строки с данными.

* Какое количество столбцов имеет данная выборка?

Выборка имеет 10 столбцов.

* Какую информацию хранит каждое поле в отдельности? Какие данные необходимо еще добыть, чтобы расшифровать значения в полях country, code, region, district?

Direction хранит тип сделки (импорт или экспорт),

Country содержит информацию про страну, с которой совершается сделка,

Code является уникальным кодом приобретённого товара,

Measure хранит единицы измерения приобретённого товара,

Value содержит сумму сделки,

Netto содержит массу товара,

Quantity содержит количество купленного товара,

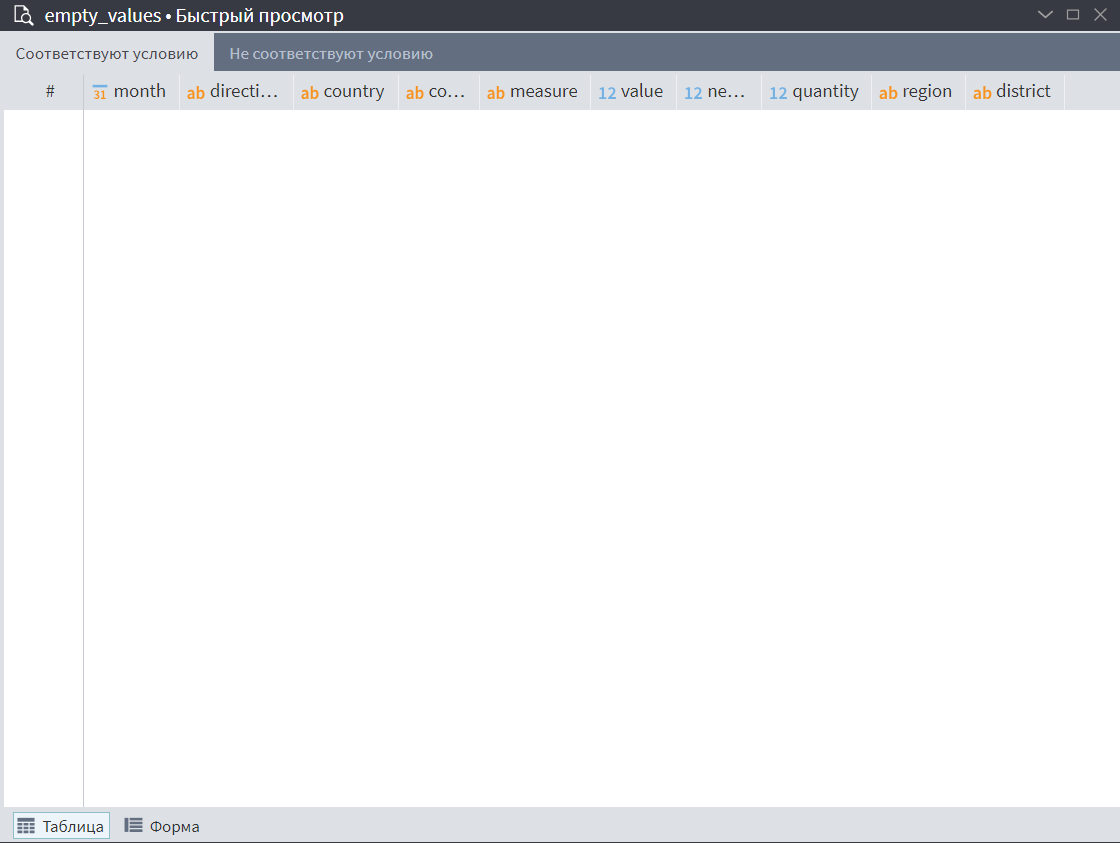
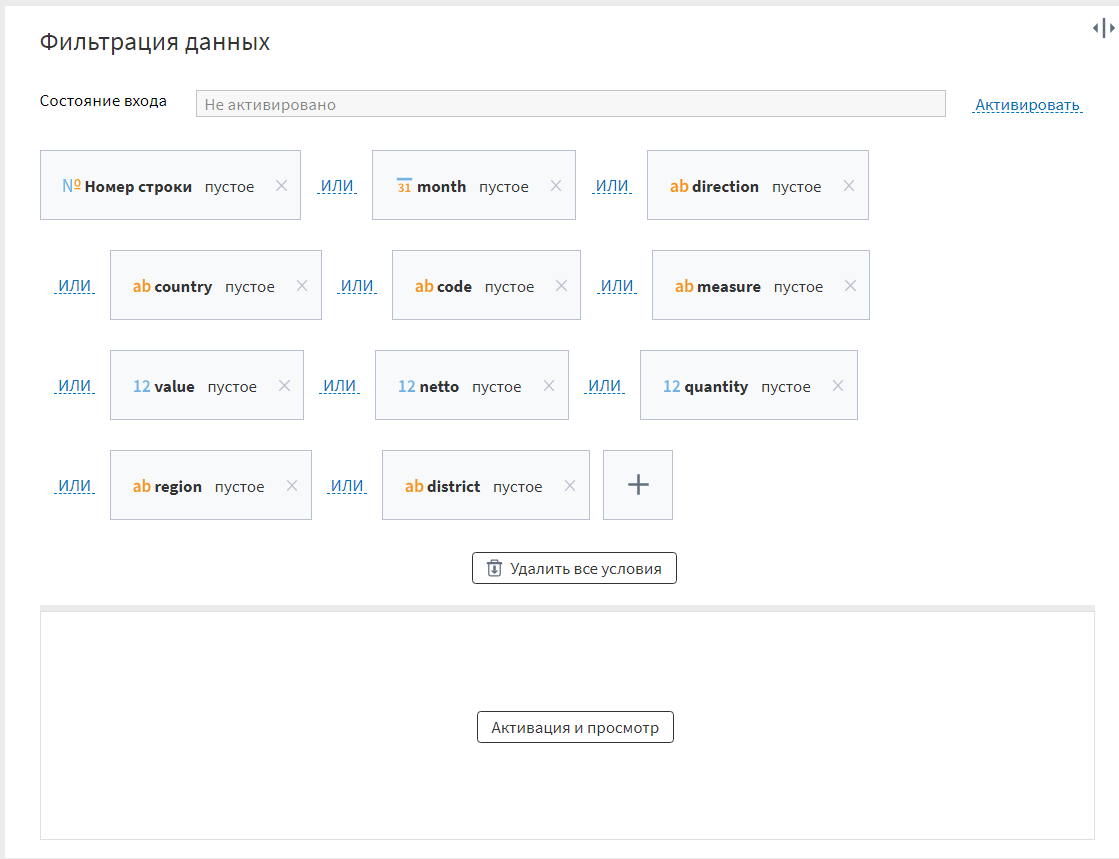
Region содержит уникальный номер региона страны (РФ),

District содержит номер федерального округа.

Чтобы больше разобраться в данных, необходимо найти информацию, в которой указанные ID для разных атрибутов имеют свои расшифровки.

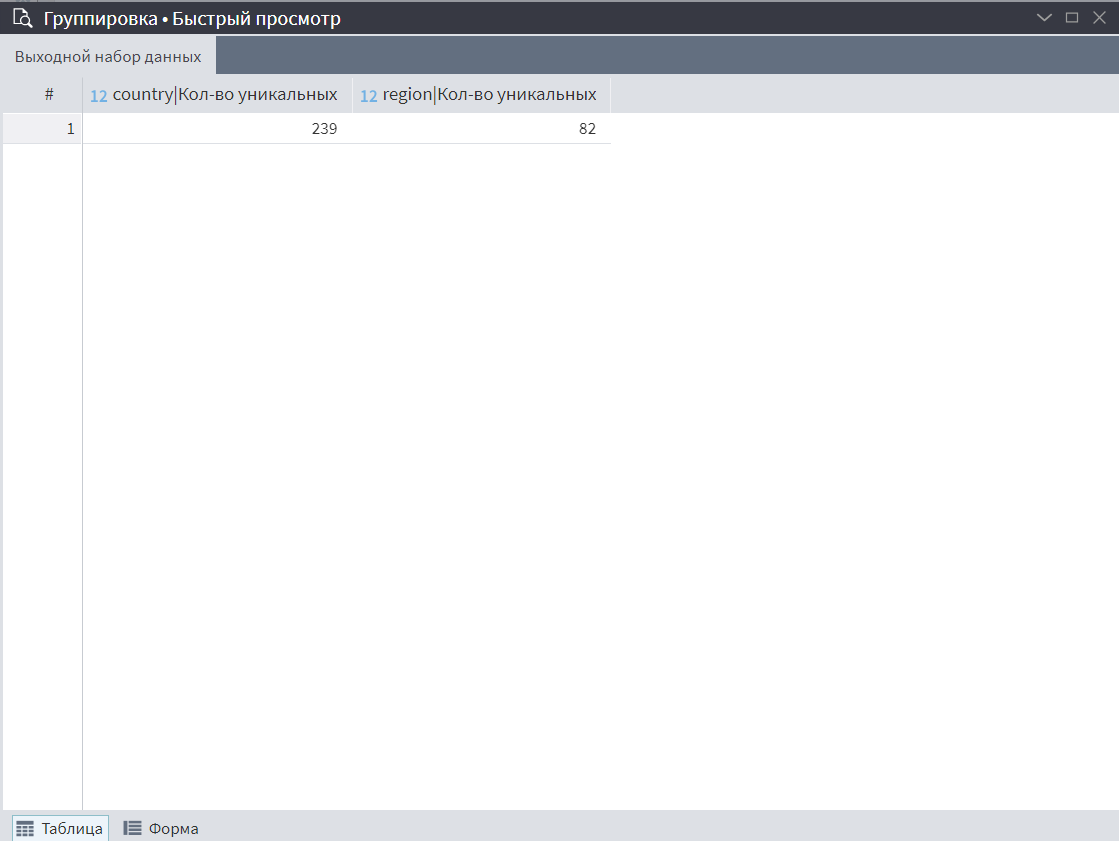
* Имеются ли в данных пропуски? На основе компонентов таблица в переменные, группировка, визуализатор статистика, ответьте на вопрос о количестве пропущенных значений в каждом из столбцов выборки данных (можно использовать что-то одно).

Сделаем фильтр и найдём все пустые значения.



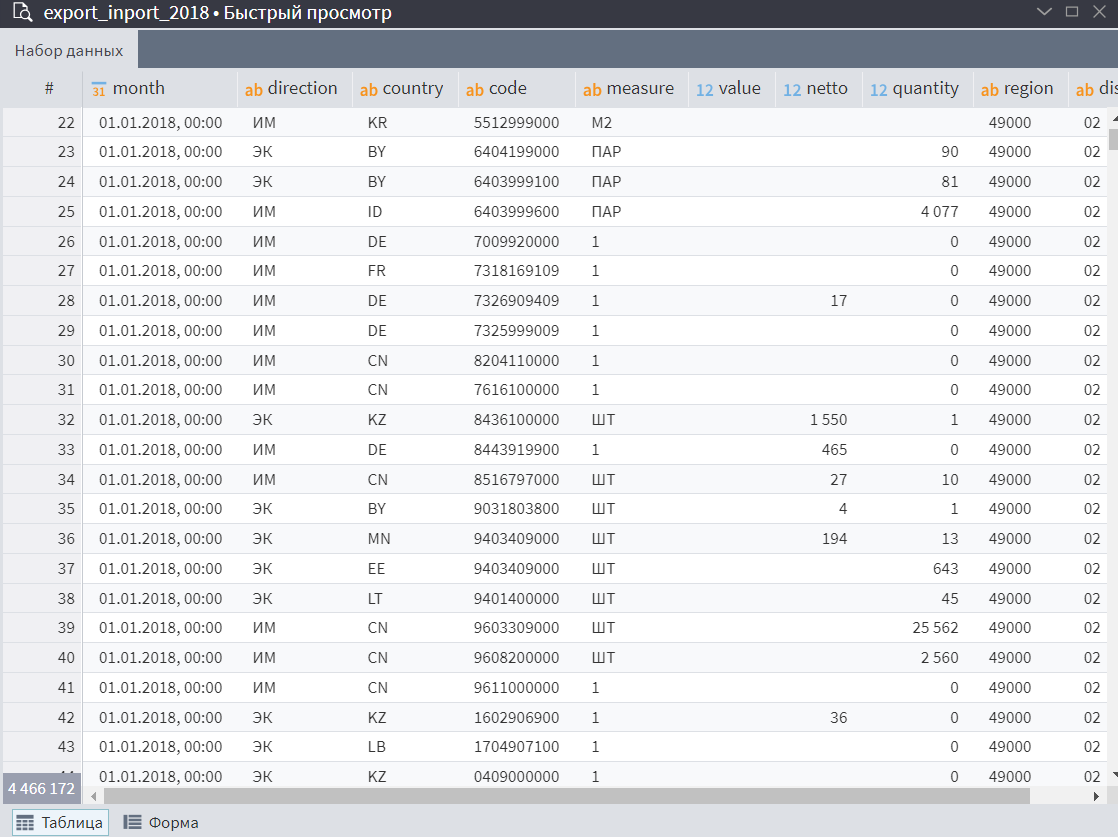
В данных отсутствуют какие-либо следы пропусков.

* Сколько уникальных стран участников торговли перечислено в поле country?
* Сколько уникальных регионов России участвовало в торговле (поле region)?



## Часть 2

Импортировать данные таможенной статистики по экспорту и импорту в регионах РФ за 2018 год.



* Сколько строк имеет выборка данных за 2018 год?

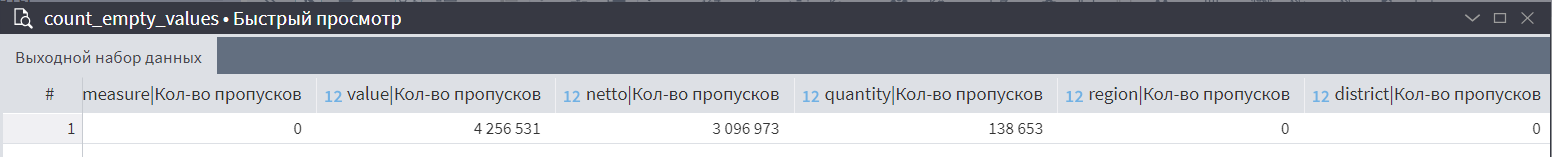
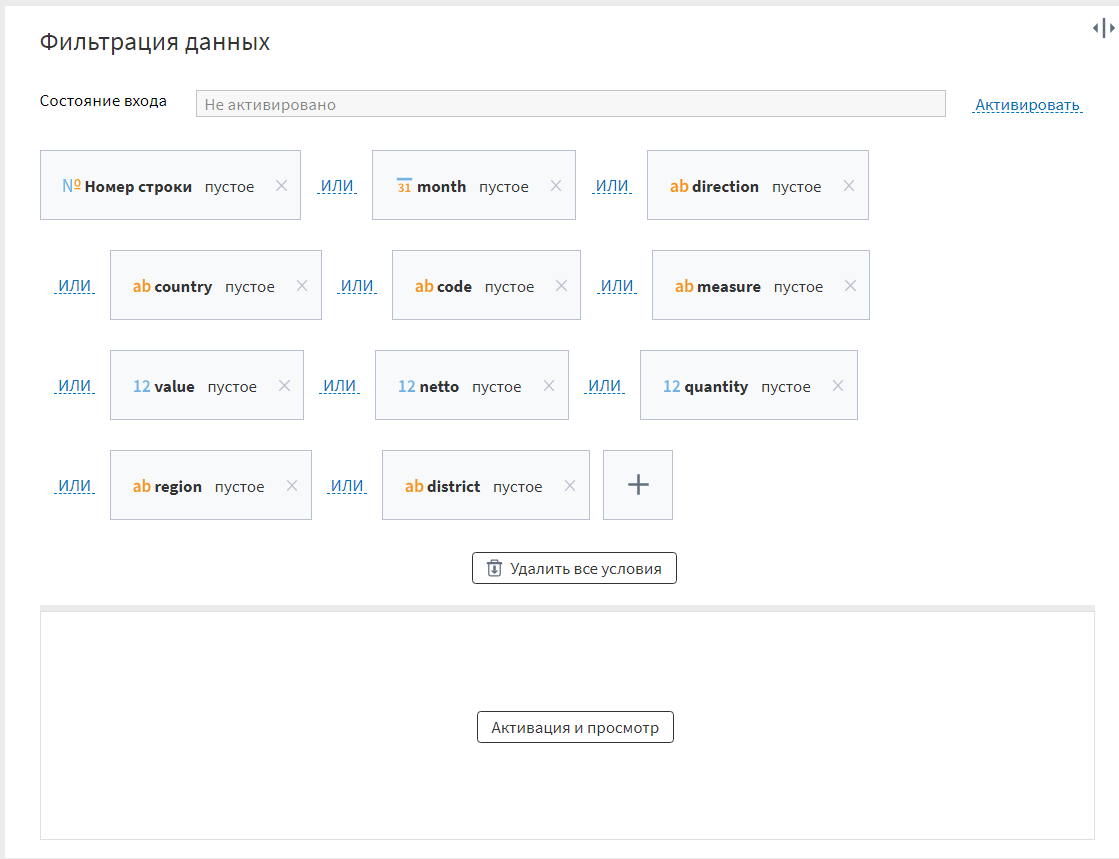
Данная выборка имеет 4 466 172 строк.

* Какое количество столбцов имеет данная выборка?

В данной выборке так же 10 столбцов.

* Имеются ли в данных пропуски? На основе компонентов **таблица в переменные, группировка,** визуализатор **статистика**, ответьте на вопрос о количестве пропущенных значений в каждом из столбцов выборки данных (можно использовать что-то одно).

Проверим на наличие пропусков с помощью фильтра.



Видим, что в данной выборке данных имеются пропуски, а именно:

Value имеет 4 256 531 пропуск, в netto 3 096 973 пропусков, а quantity содержит 138 653 пропуска.

* Сколько уникальных стран участников торговли перечислено в поле **country**?
* Сколько уникальных регионов России участвовало в торговле (поле **region**)?



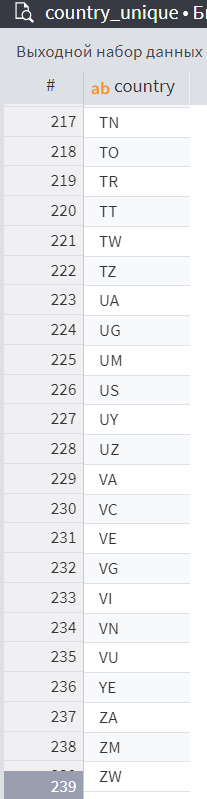
Сформируйте вывод о динамике качества данных собираемой статистики во времени и перечислите поля, дальнейший анализ по которым будет репрезентативен и валиден.

Вывод: Качество данных собираемой статистики с 2016 по 2018 значительно ухудшилось в виду большого количества пустых значений.

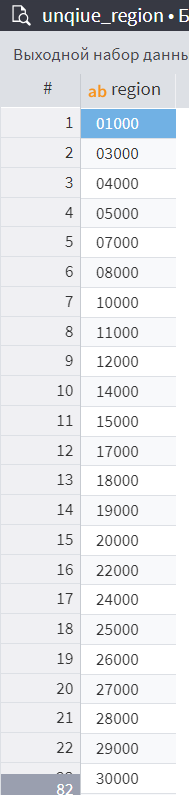
## Часть 3

Для импортированной выборки за 2016 год на основе компонентов трансформации данных отобразите:

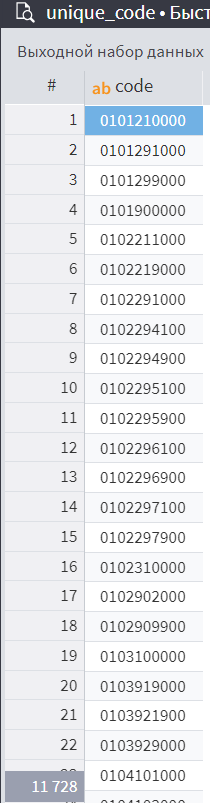
* Перечень уникальных стран-участниц торговли в поле **country** с помощью компонента группировка (отсортируйте значения по возрастанию).



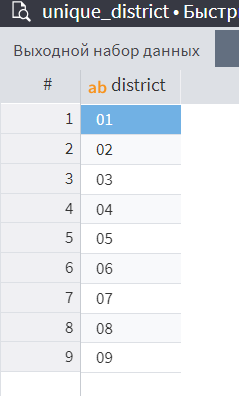
* Перечень уникальных регионов России участников торговли в поле **region** с помощью компонента группировка (отсортируйте значения по возрастанию).



* Перечень уникальных категорий товаров по ТН ВЭД, перевозимых через таможню в поле **code** с помощью компонента группировка (отсортируйте значения по возрастанию.



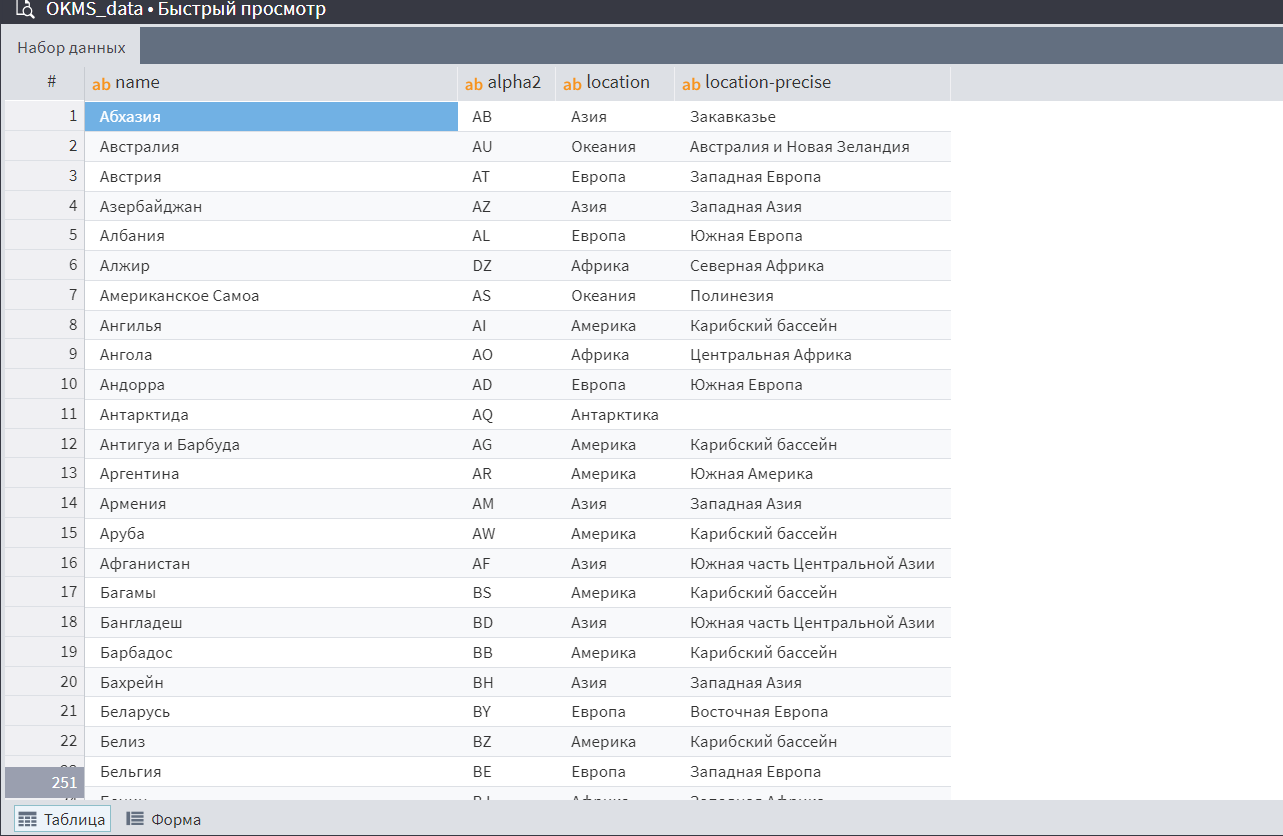
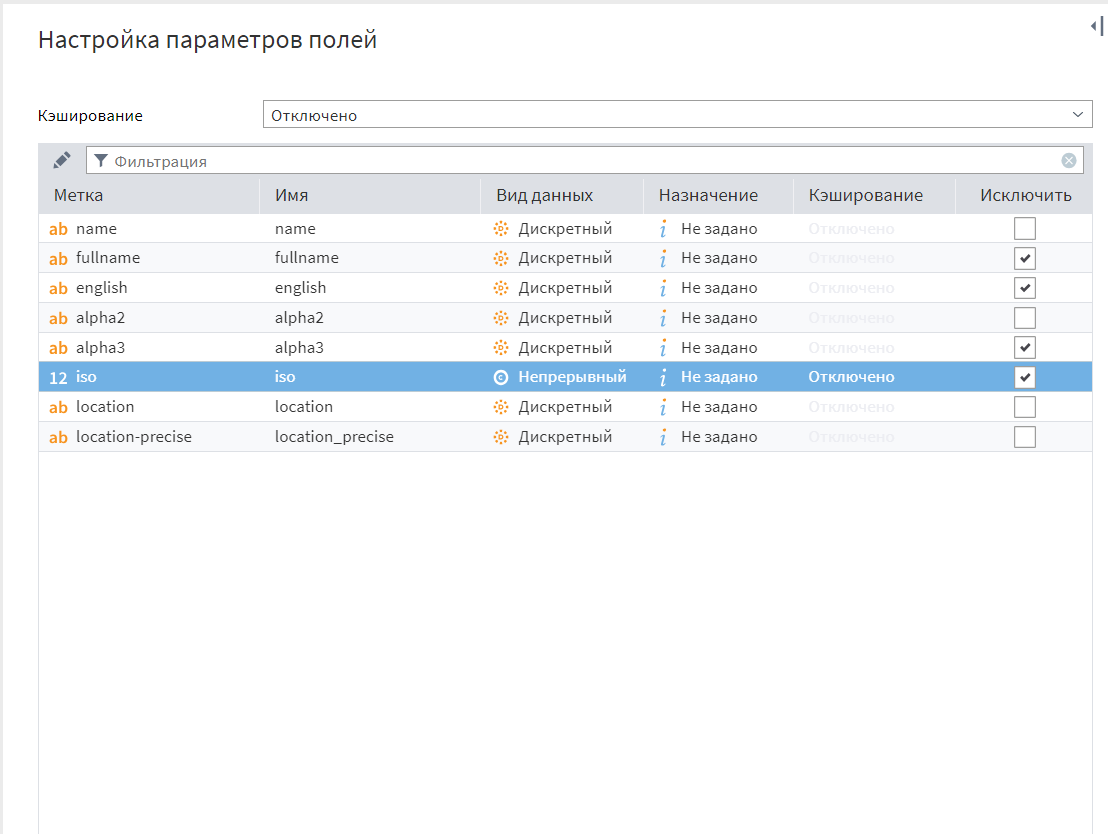
* Перечень уникальных кодов федеральных округов по полю **district** с помощью компонента группировка (отсортируйте значения по возрастанию).



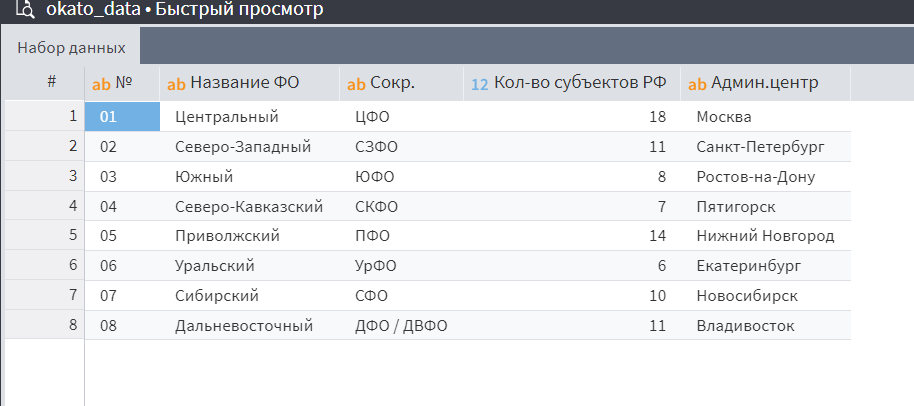
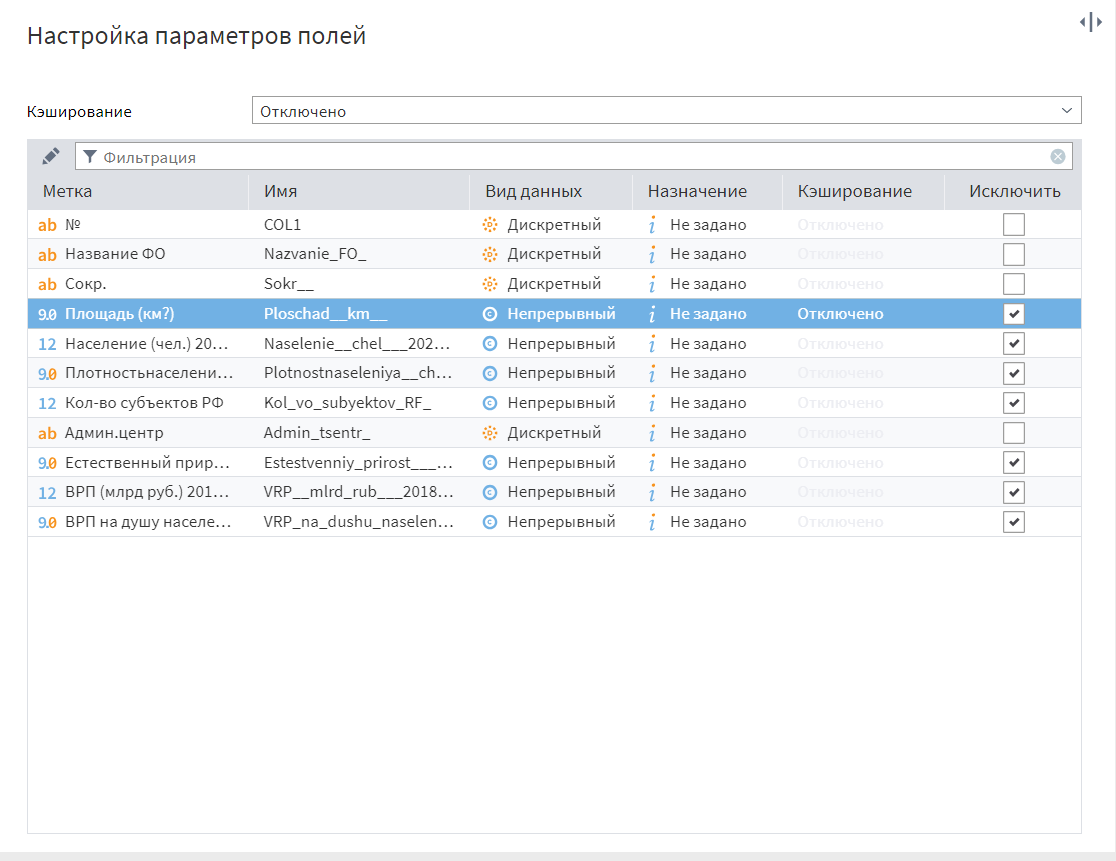
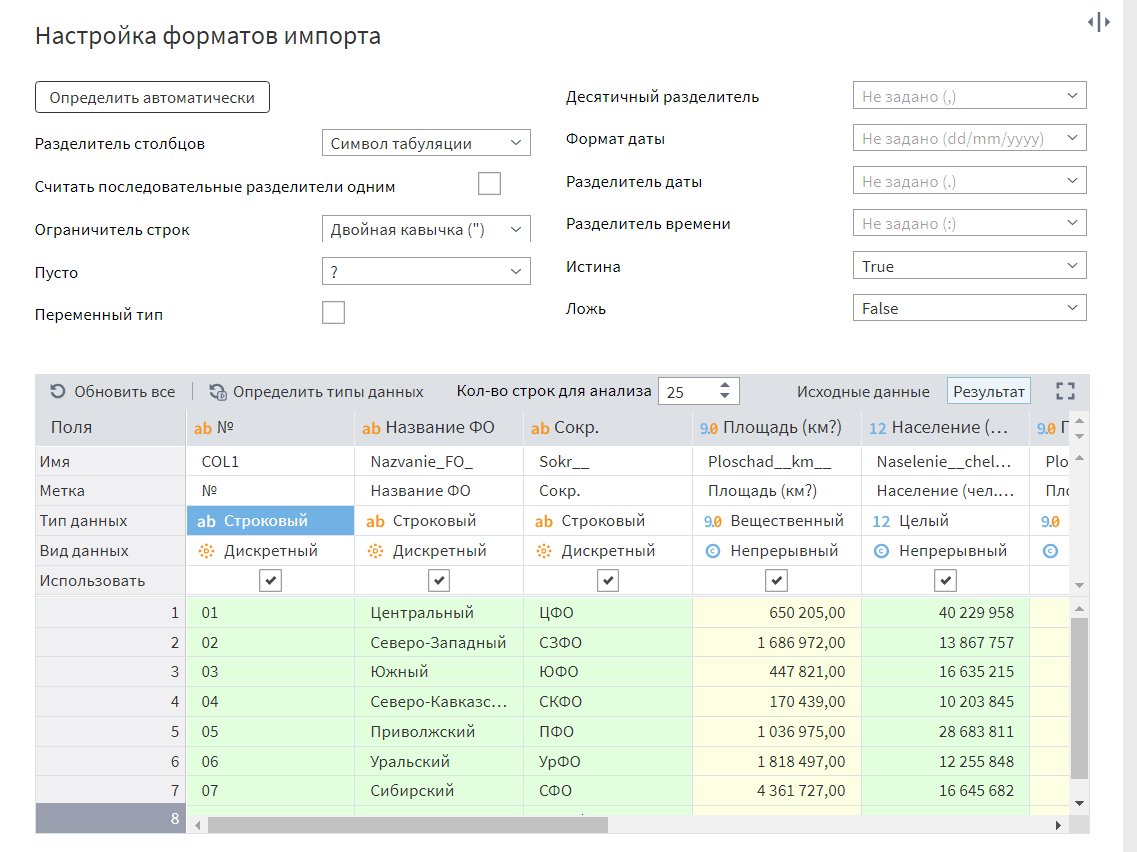
## Часть 4

Импортировать в сценарий выборки данных:

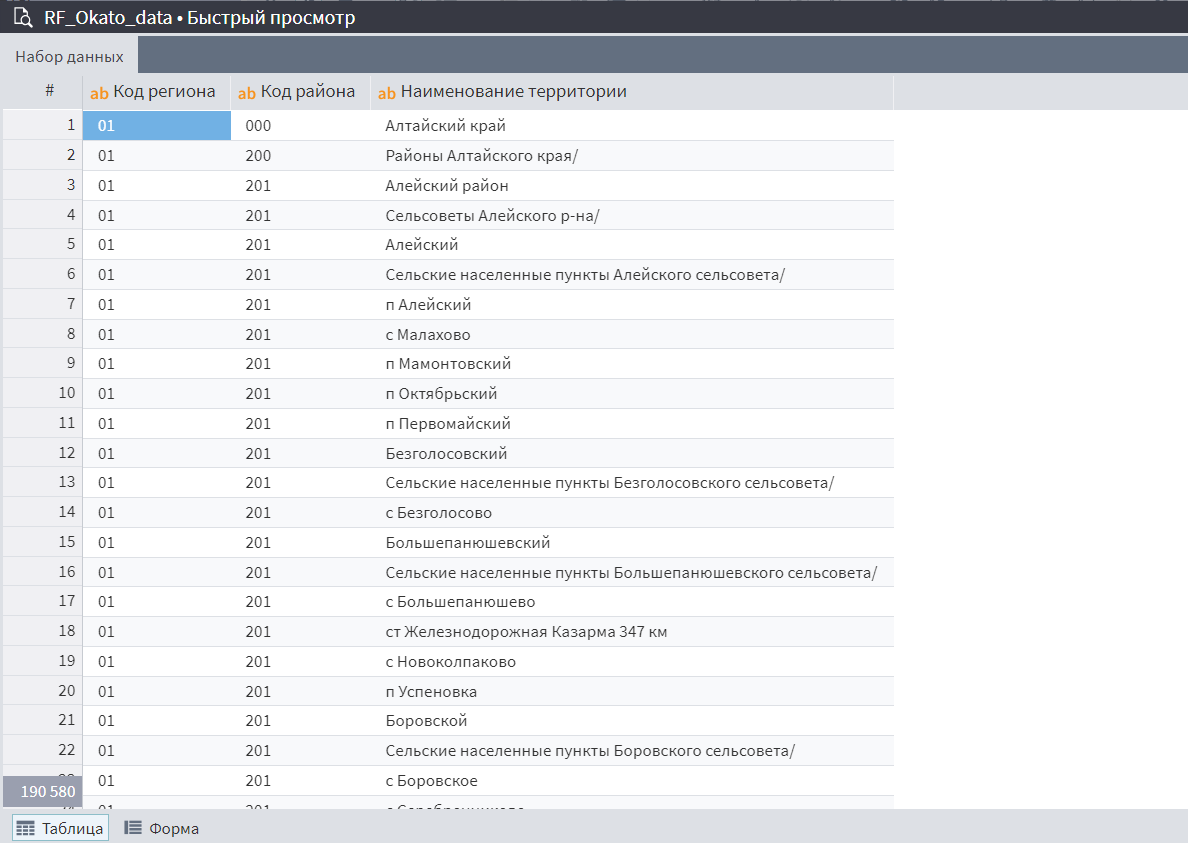
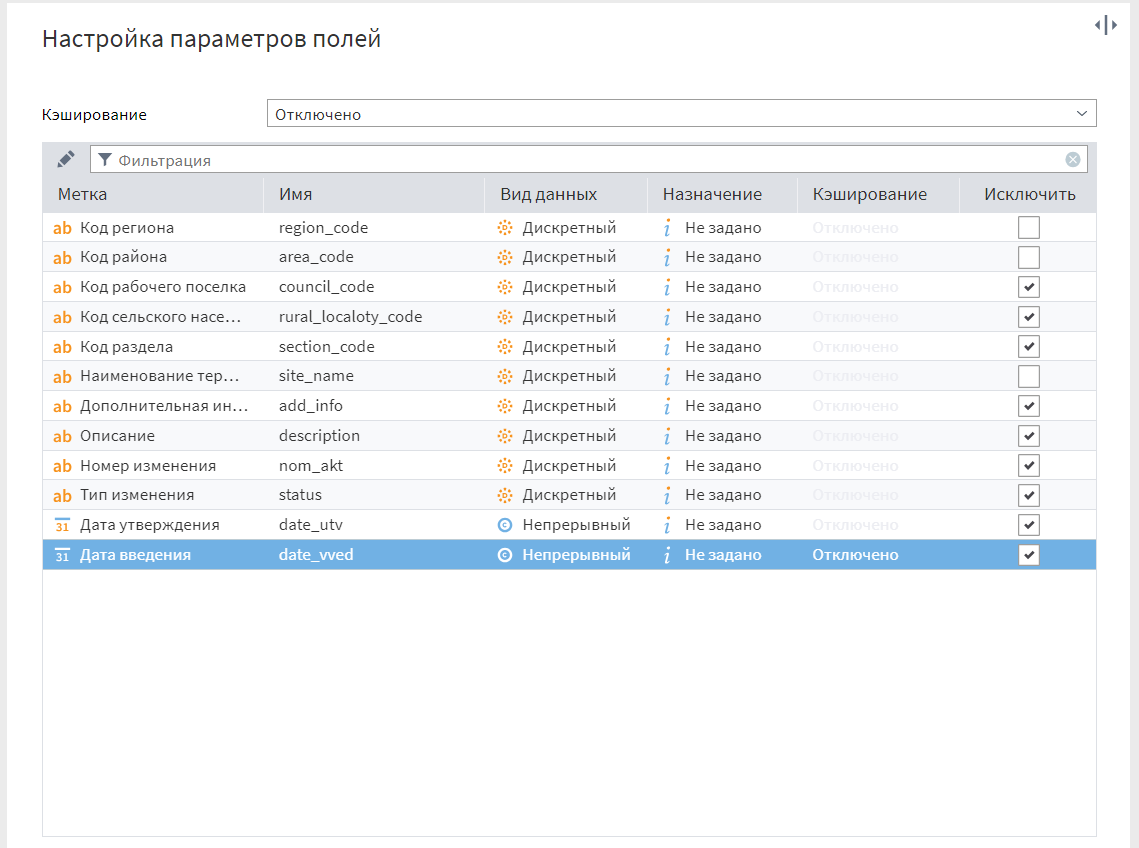
* Файл данных с кодами стран мира ОКСМ в формате alpha2, в кодировке UTF-8 с разделителем столбцов табуляцией. Исключить из полей импортируемой таблицы: fullname, English, alpha3, iso.



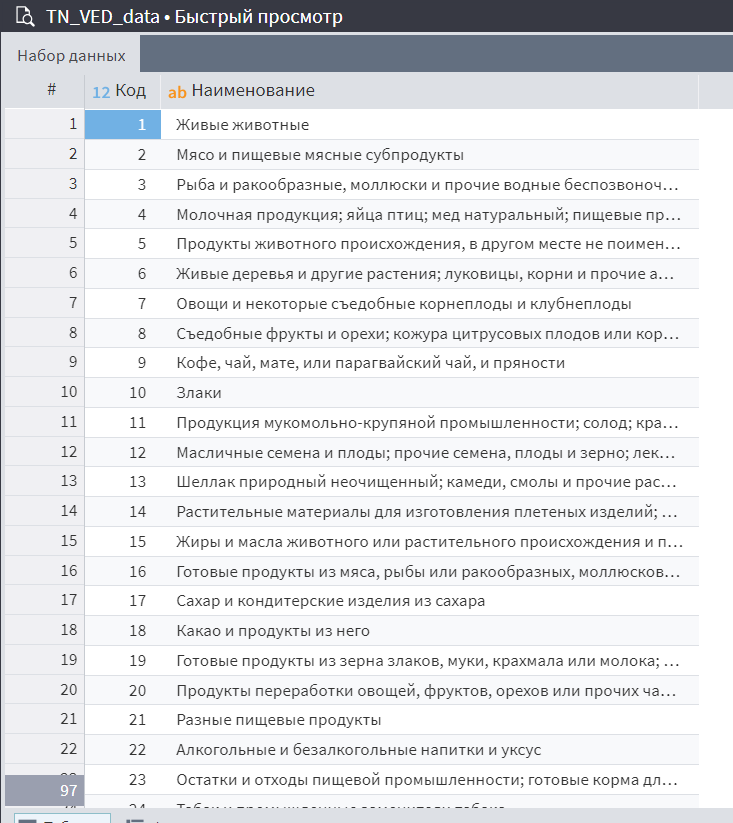
* Файл данных с информацией о федеральных округах, в кодировке ANSI с разделителем столбцов табуляцией. Изменить тип поля № на строковый тип данных. Исключить из рассмотрения все поля, кроме: №, Название ФО, Сокр., Кол-во субъектов, Админ. Центр.



* Файл данных с информацией о регионах РФ в соответствии с ОКАТО. Исключить из рассмотрения все поля, кроме: Код региона, Код района, Наименование территории.

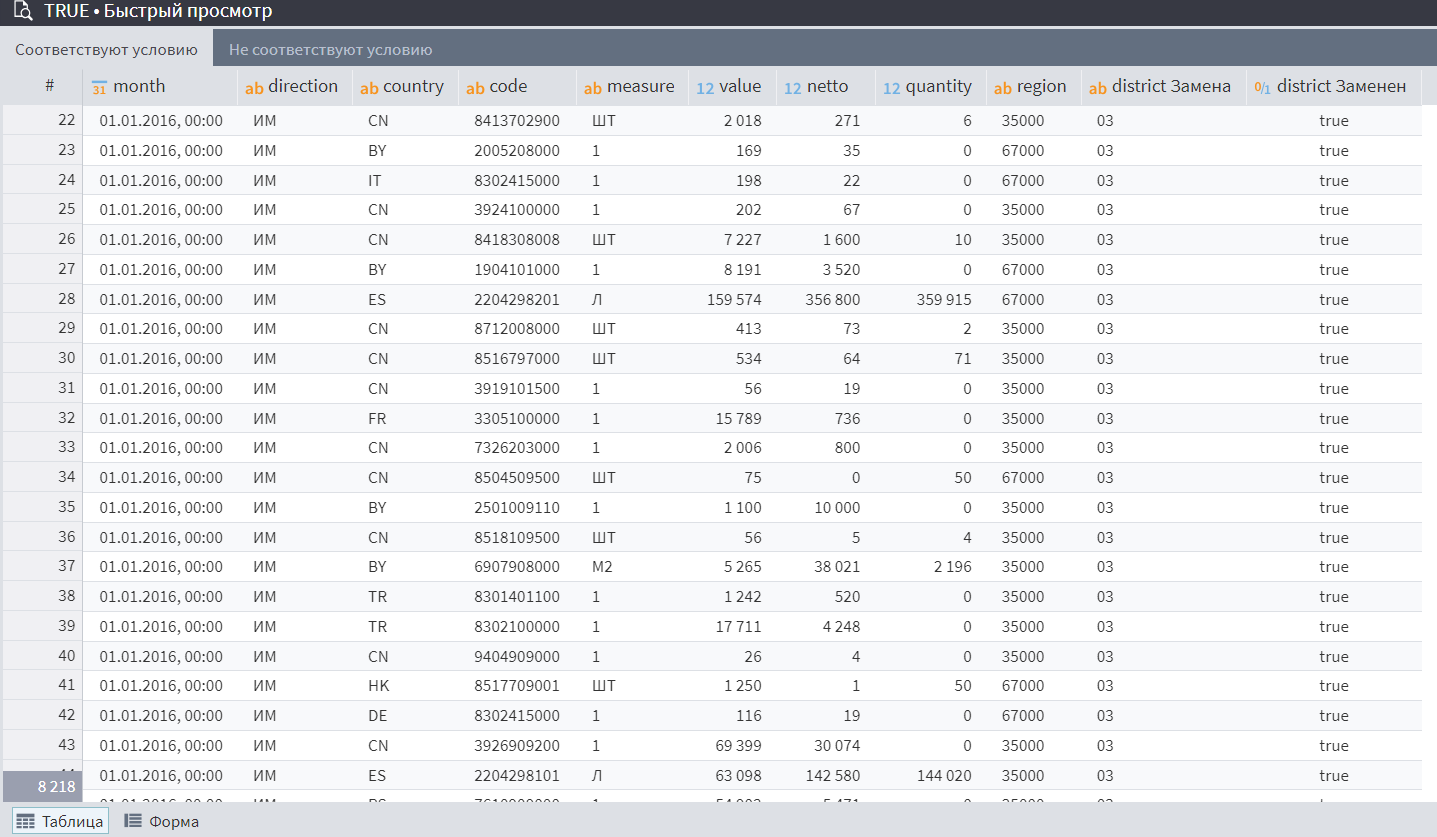


* Файл данных с информацией о разделах товарной номенклатуры ТН ВЭД, в кодировке ANSI с разделителем столбцов табуляцией.

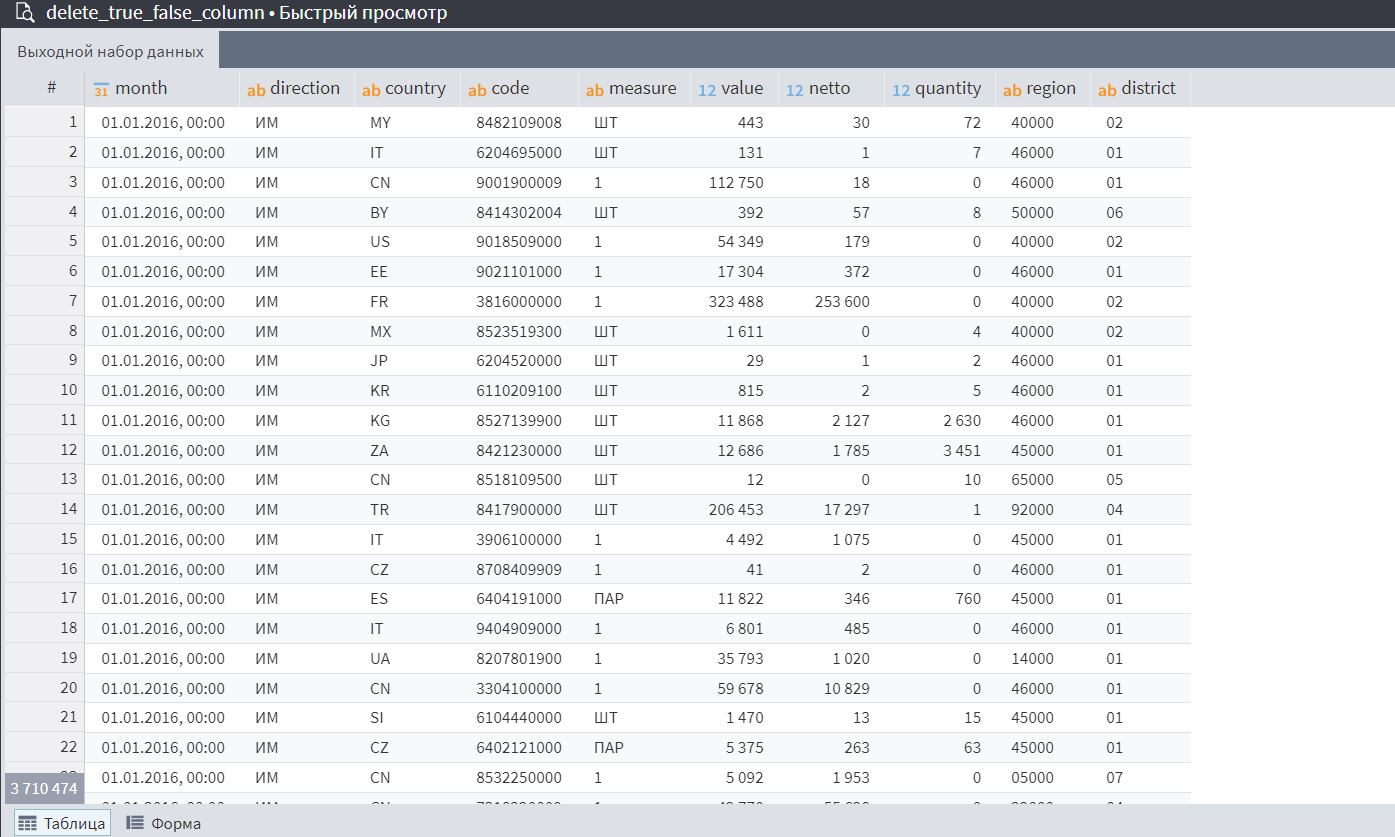


## Часть 5

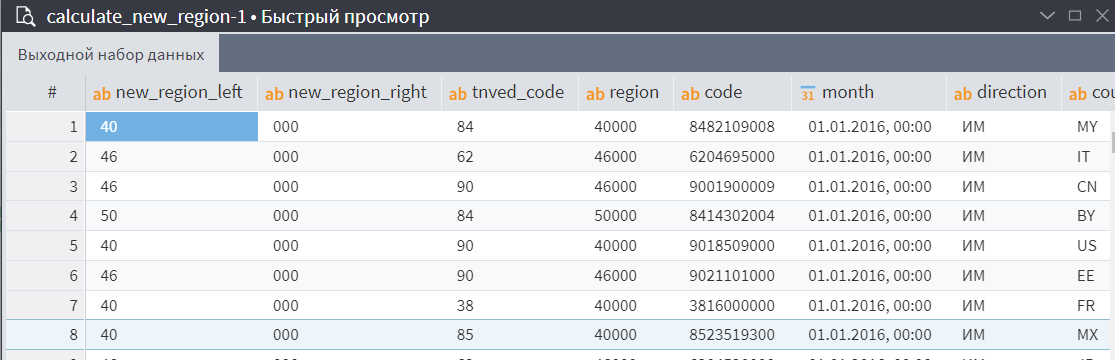
* Замените в исходном наборе данных в поле с названием **district** значения «09» на «03» (Южный федеральный округ) с помощью компонента «Замена».



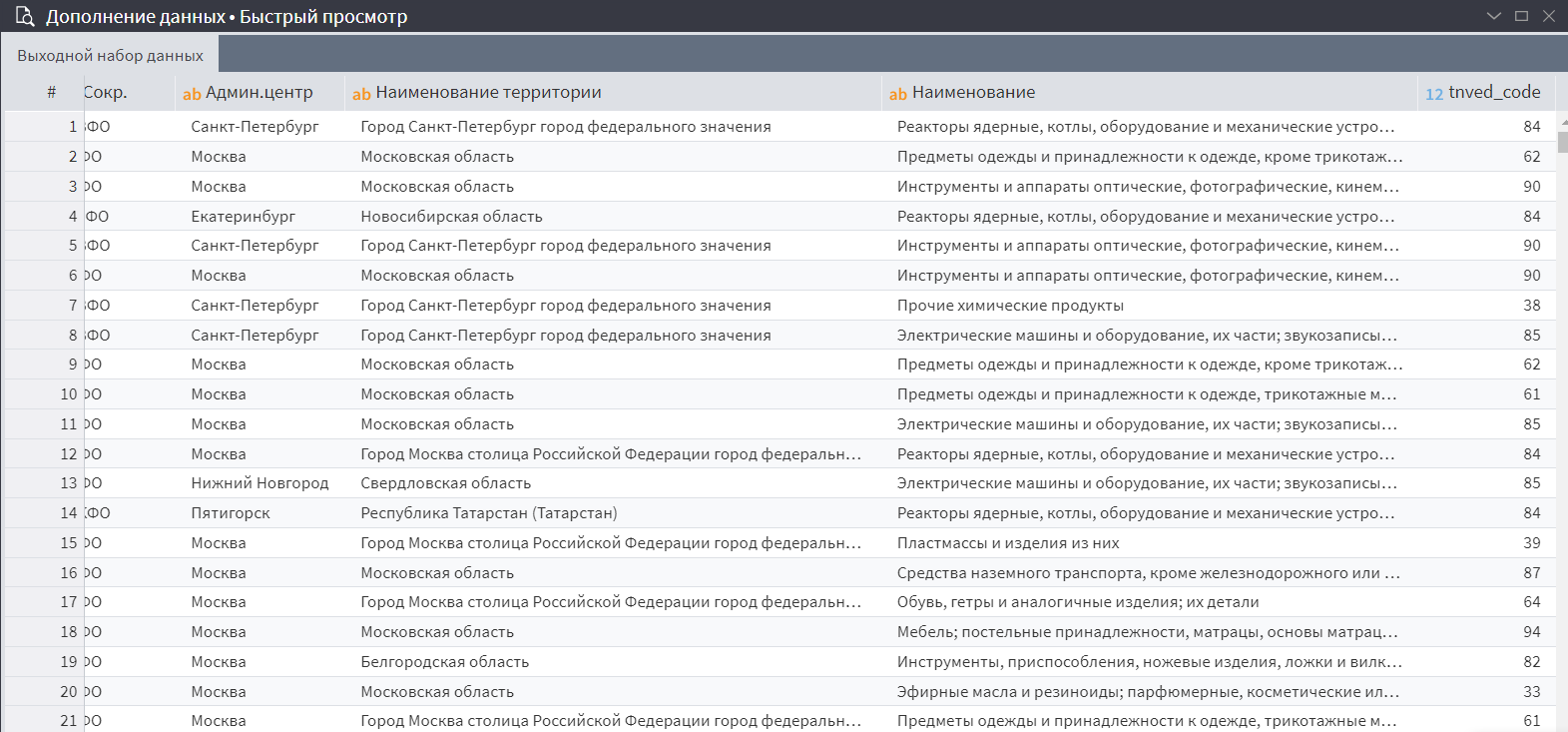
* С помощью компонента «Параметры полей» приведите выборку после замены к исходной структуре (количество полей и их названия).



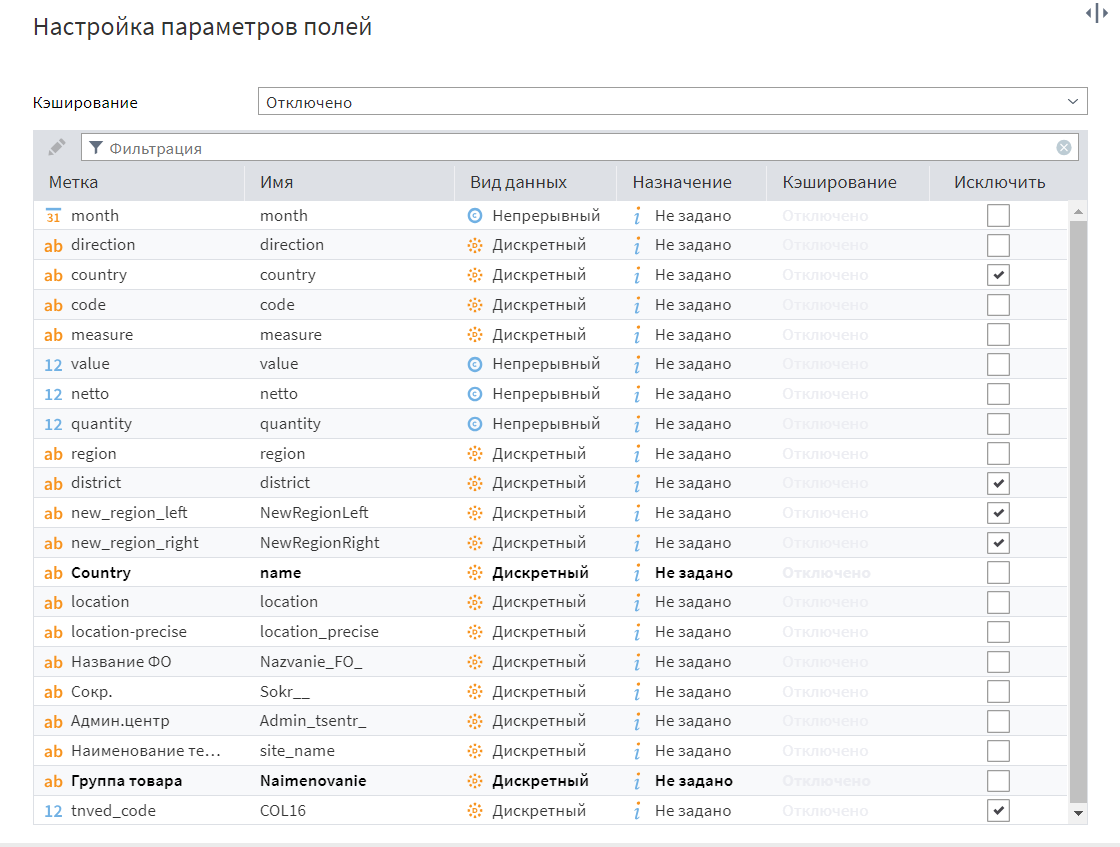
* Получите из поля **region** *новые поля* с первыми двумя цифрами из **region** и последующими тремя цифрами из **region**, применив функции **Left()** и **Right()** из компонента калькулятор.



* Дополните преобразованную выборку данных новыми полями из импортированных таблиц с помощью компонента **«Дополнение данных».**



* Исключите с помощью **«Параметры полей»** все промежуточные ключи(district, region, country, новые поля из калькулятора).

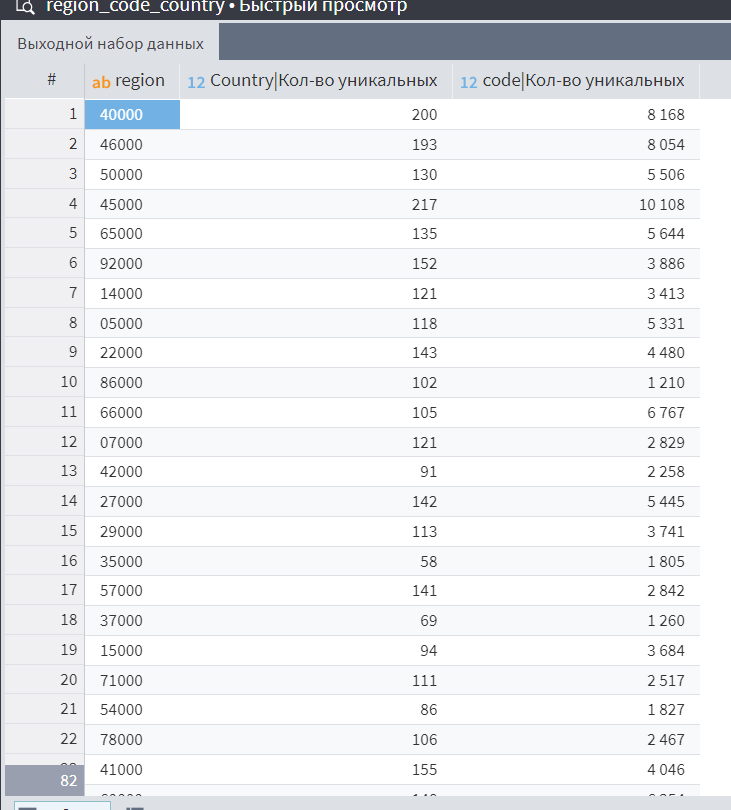


## Часть 6

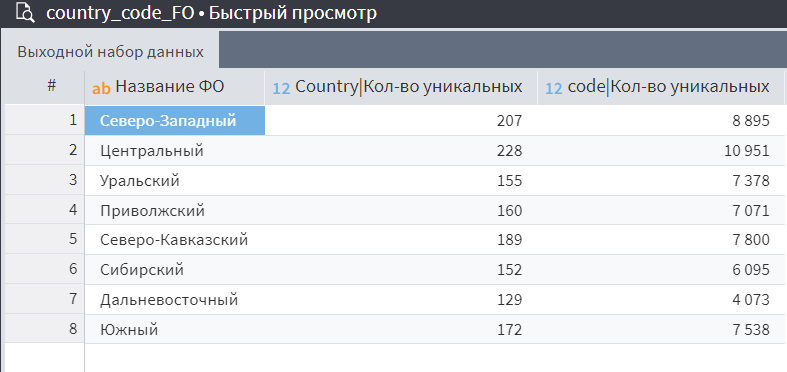
Сформируйте серию группировок данных на основе компонента «Группировка»:

* Какое количество уникальных кодов ТН ВЭД (поле **code**) и колчиество уникальных стран участниц торговли было задействовано в отчете об импорте и экспорте за 2016 код по: 1) каждому региону России, 2) Каждому Федеральному округу России.

1)



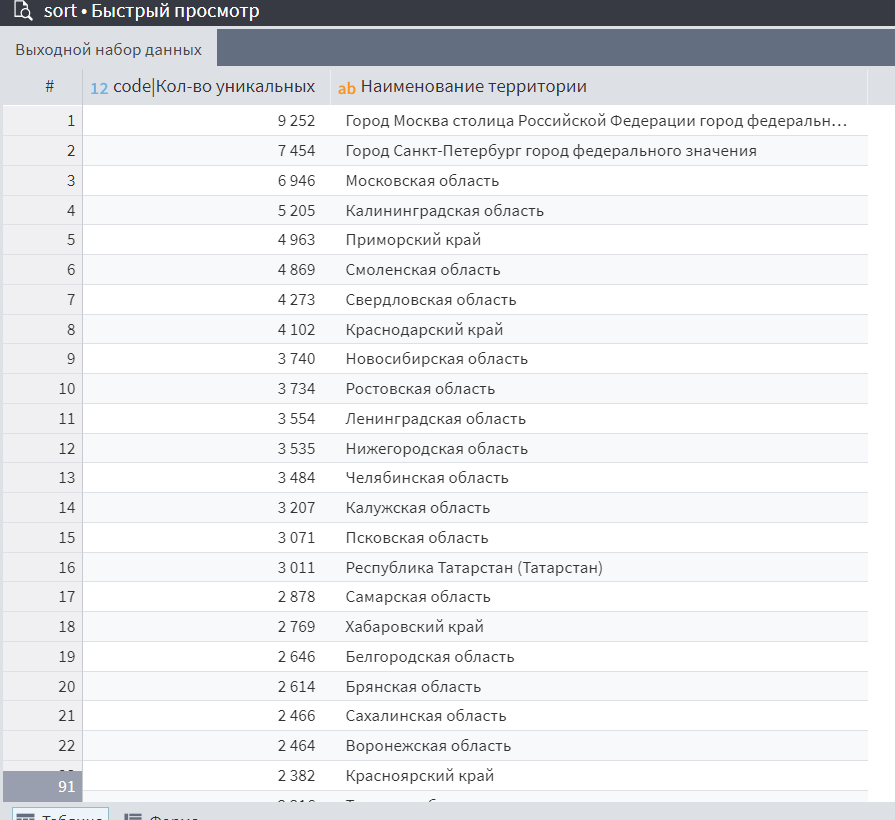
2)



* Найдите самое популярное направление экспорта за рубеж по количеству уникальных позиций экспортируемых товаров за 2016 год. Считаем общее число уникальных кодов по всем регионам России.

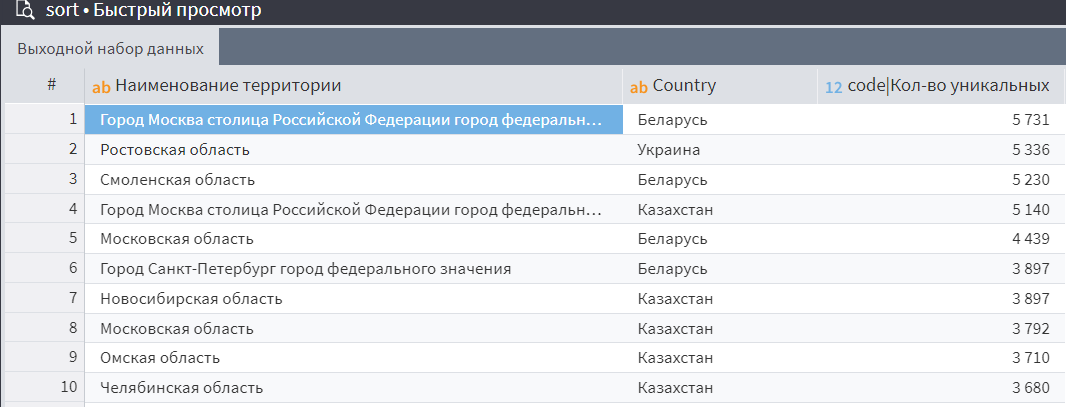


* Найдите самое популярное направление импорта в регион РФ из зарубежа. В какой регион поступает самый разнообразный спектр товаров.



* Найдите десятку самых популярных пар **регион**-**страна** по количеству уникальных позиций в: 1) экспорте товаров, 2) импорте товаров.

1)



2)

