|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | 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 № 1** | |
| **по дисциплине «Большие данные»** | |
|  | |
| Выполнил студент группы БИСО-02-22 | Зубарев В.С. |
| Проверил ассистент кафедры ПМ ИИТ | Тетерин Н.Н. |

Оглавление

[Создание пакета Loginom 3](#_Toc162529831)

[Демо сценарий Loginom 4](#_Toc162529832)

[Импорт из текстовых файлов 5](#_Toc162529833)

[Импорт данных из РСУБД 7](#_Toc162529834)

[Фильтрация пустых полей 10](#_Toc162529835)

[Сортировка записей 12](#_Toc162529836)

[Использование таблицы замен 14](#_Toc162529837)

[Вычисление новых полей 16](#_Toc162529838)

[Параметры полей 18](#_Toc162529839)

[Выделение начала месяца 20](#_Toc162529840)

[Группировка данных в добыче 22](#_Toc162529841)

[Работа с переменными Loginom 24](#_Toc162529842)

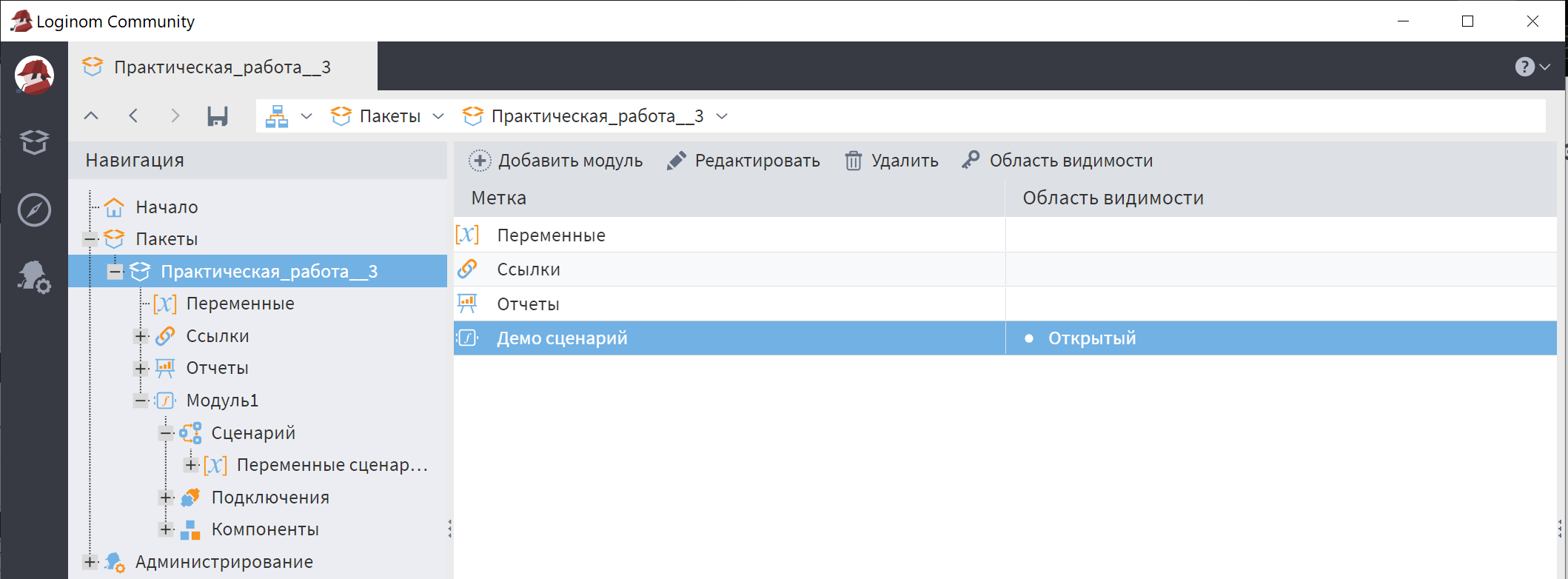
[Самостоятельная работа 30](#_Toc162529843)

# Создание пакета Loginom

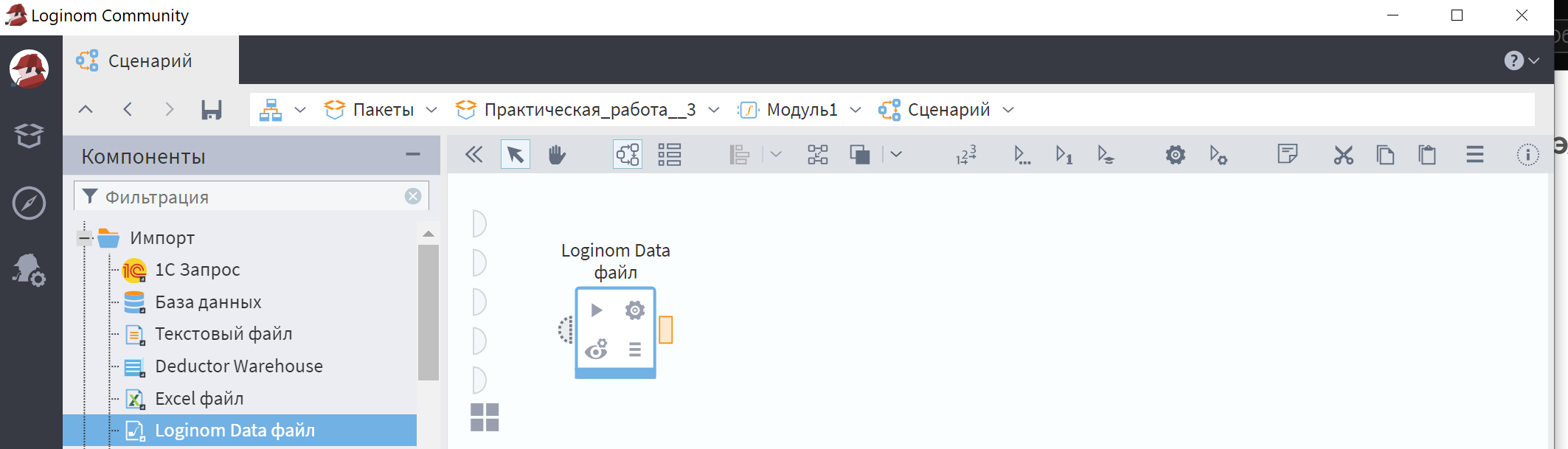
1. Создайте пакет Loginom одним из удобных для вас способов

2. Сохраните пакет под названием «Практическая работа №3»

3. Создайте модуль внутри пакета с названием «Демо сценарий» (функция **редактировать**)



4. Перенесите на полотно компонент «Loginom Data файл»



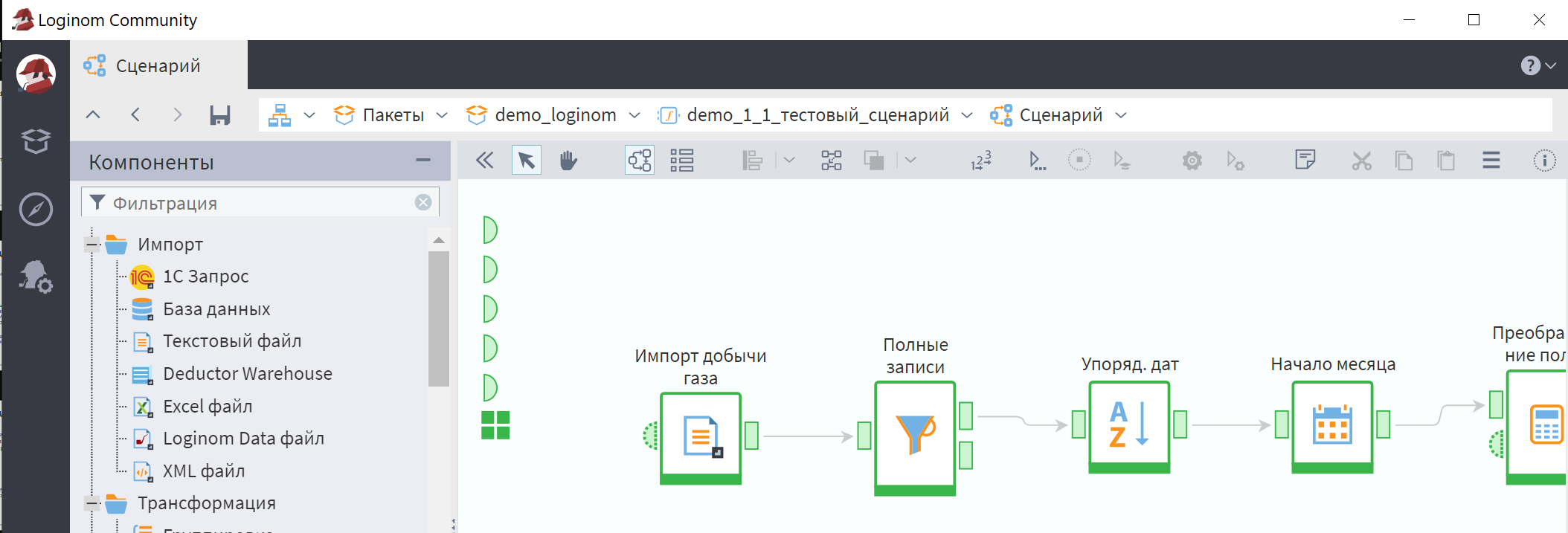
5. Сохраните пакет.

# Демо сценарий Loginom

1. Откройте пакет **demo\_loginom.lgd**

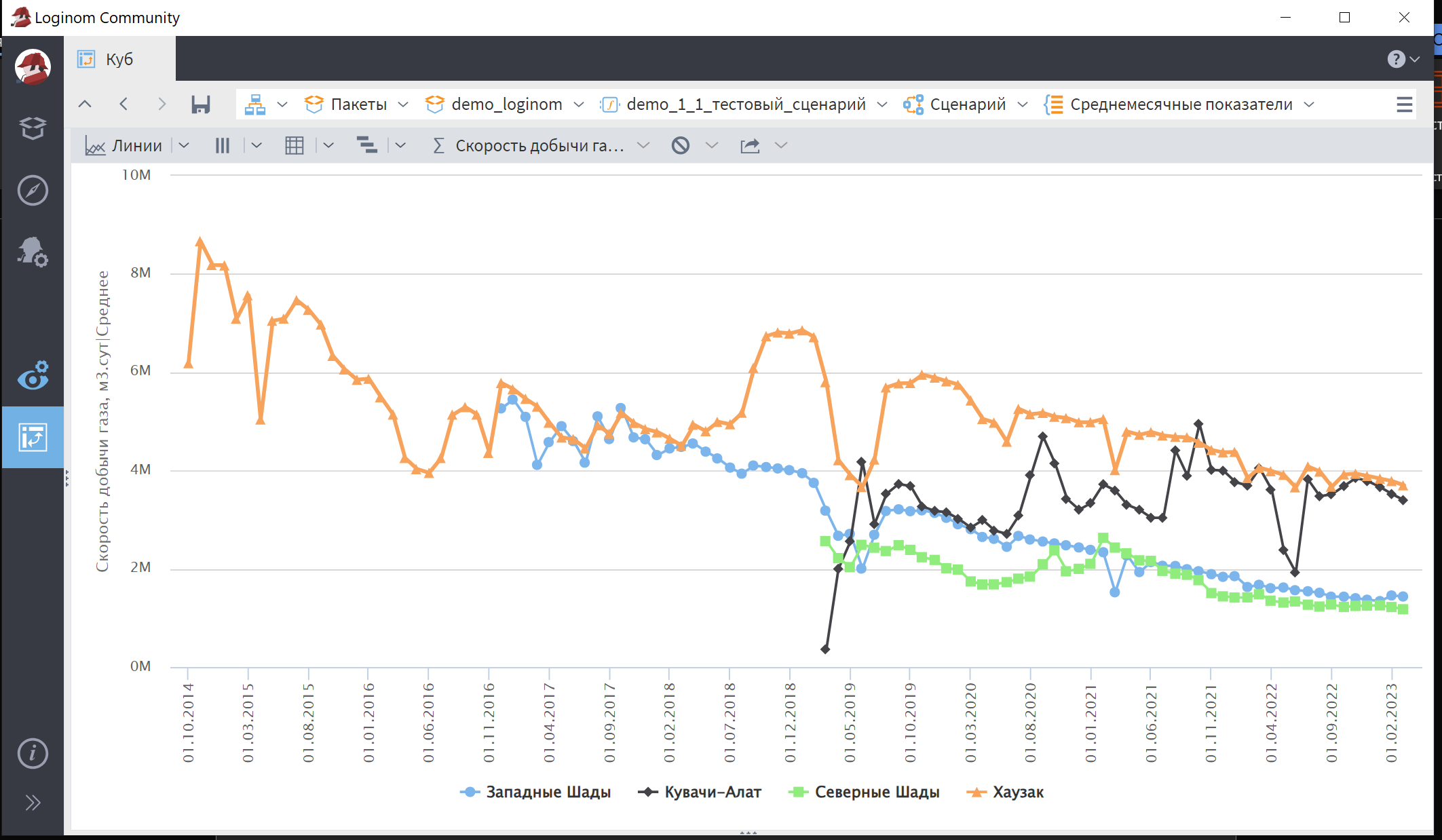
2. Откройте модуль **demo\_1\_1\_тестовый\_сценарий**

3. Запустите все узлы кнопкой «**выполнить все**», убедитесь в том, что сценарий выполняется без ошибок



4. Исследуйте выходные и входные порты узлов на основе **настройки входных портов** и **предварительного просмотра** порта данных

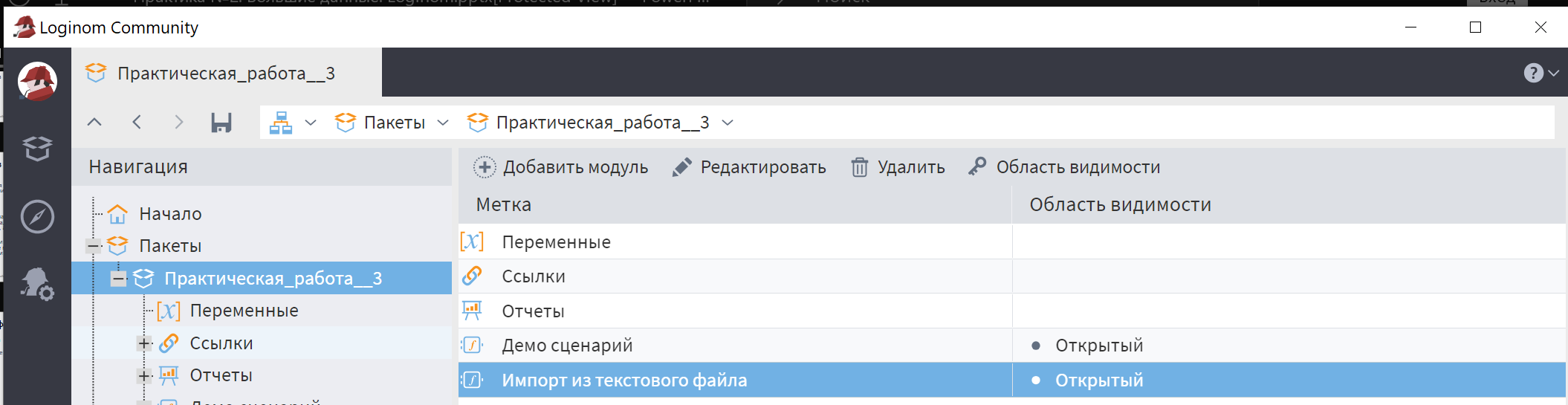
5. Нажмите на «глаз» под узлом **среднемесячные показатели**, просмотрите визуализацию



# Импорт из текстовых файлов

1. Откройте пакет с расширением «.lgp», созданный вами ранее;

2. Создайте новый модуль и измените его название на «Импорт из текстового файла»

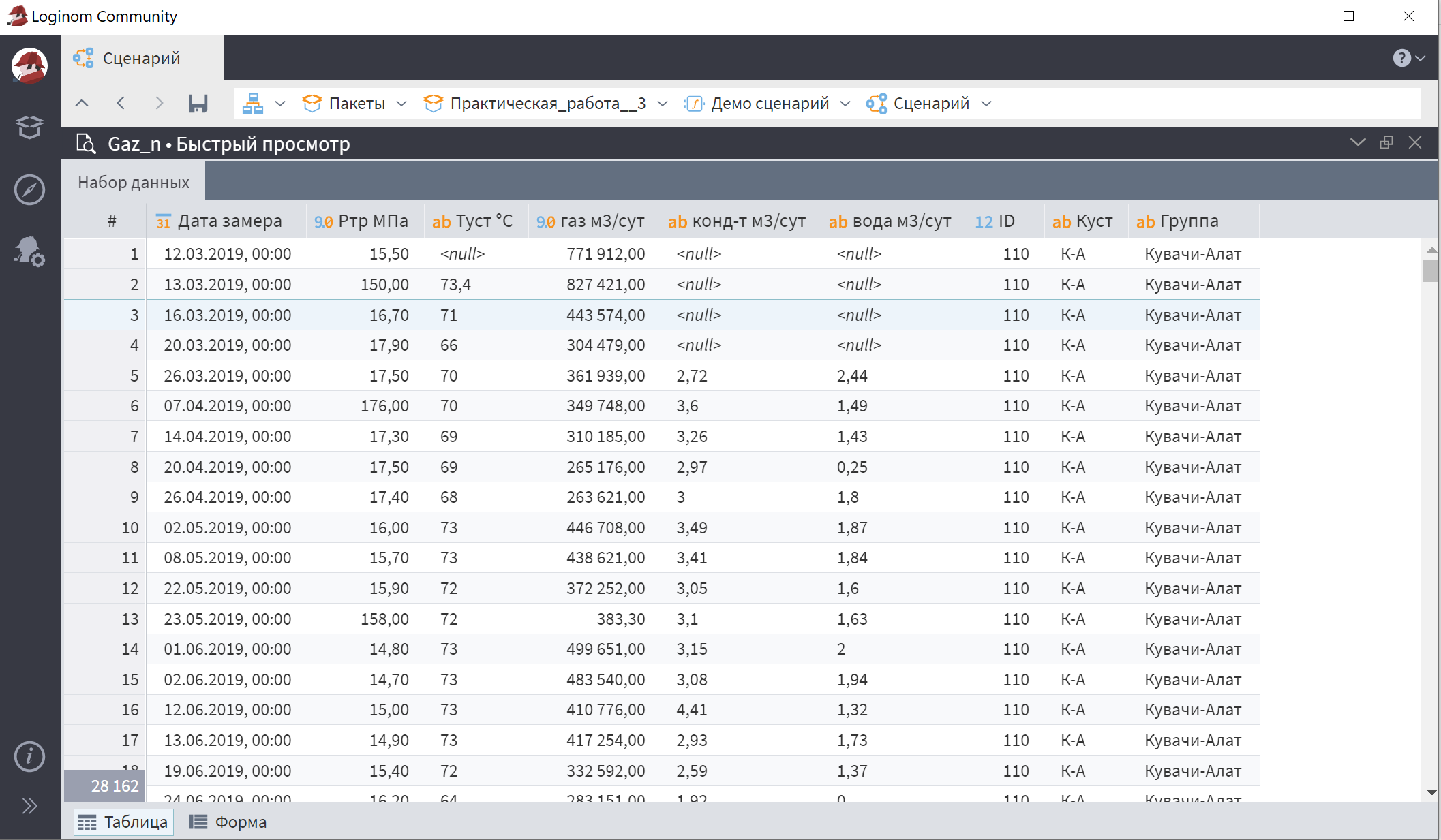


3. Зайдите в сценарий и перенесите из панели компонентов на полотно сценария компонент «текстовый файл»

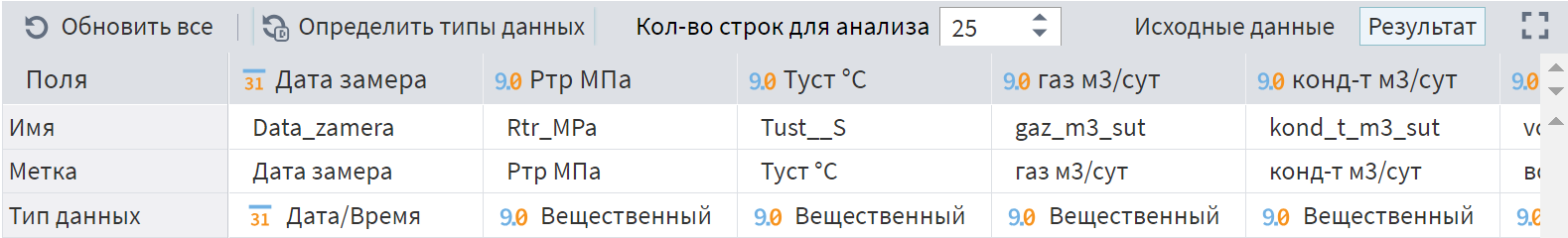
4. Настройте импорт данных из файла «data\\GAZ\\GAZ\_n.csv»:

* Кодировка **UTF-8**;
* Разделитель **табуляция**;
* Пусто – **null.**

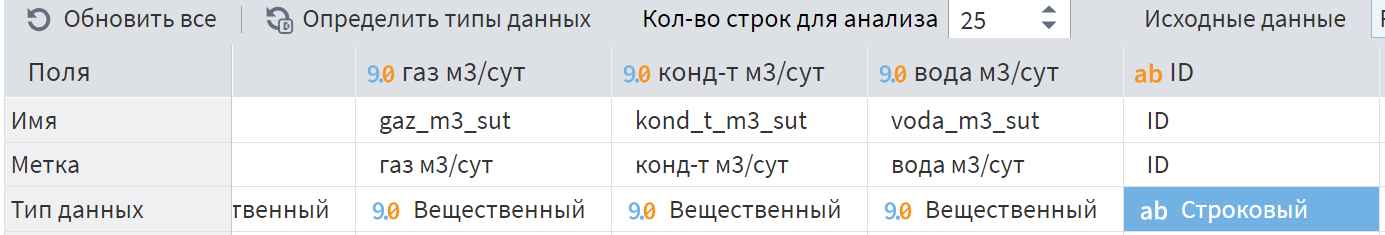
**5. Убедитесь в корректности импорта данных воспользовавшись предпросмотром выходного порта**



* Какие типы данных имеет каждое поле?



* Перейдите в настройку форматов импорта, найдите поле **ID** и измените его тип данных на строковый. Сохраните изменения . Выполните узел.   
  Как изменился вид данных в поле ID?



Переместился в конец вывода

* Сколько строк данных получилось в результате импорта данных из внешнего файла **GAZ\_n.csv**?

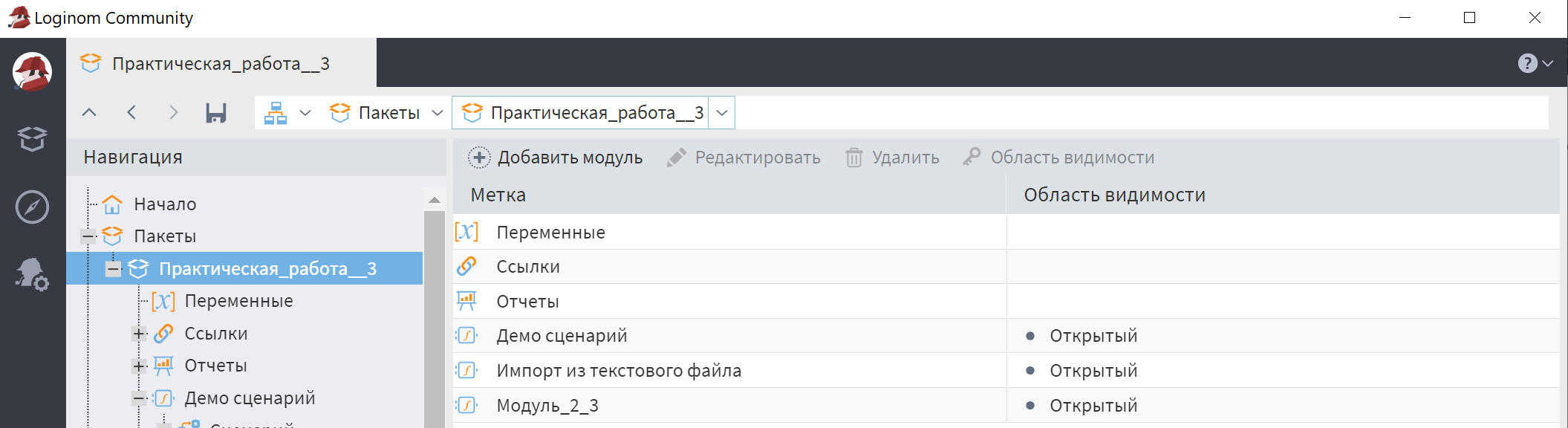
28162

* Каким образом мы настроили чтение данных из файла .csv с разделением табуляцией?

Через настройки импорта

# Импорт данных из РСУБД

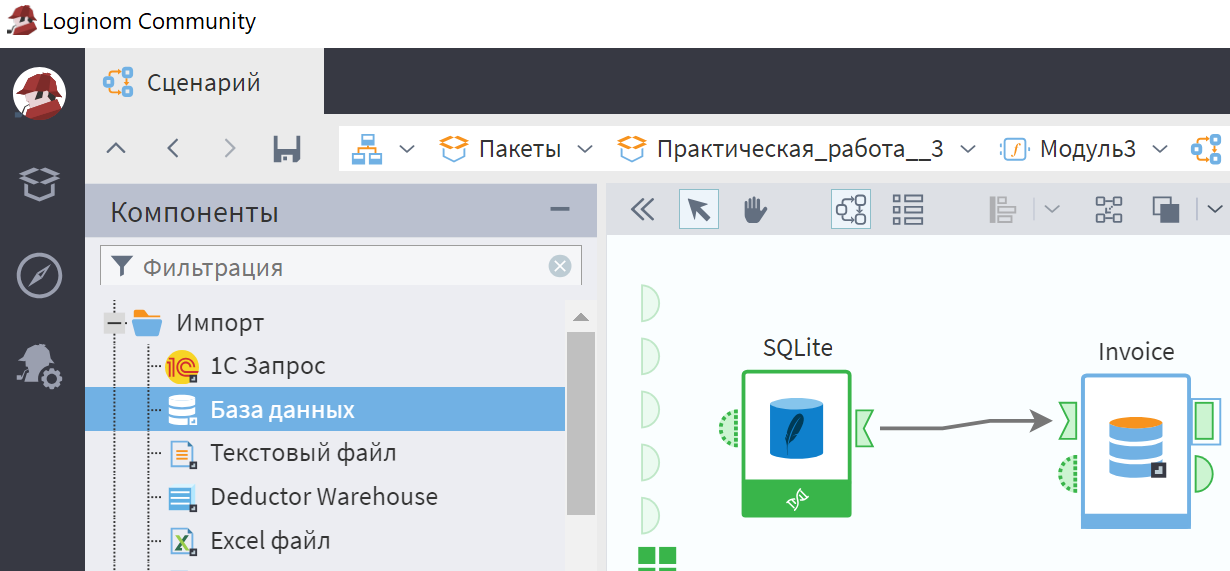
1. Модуль demo\_2\_3\_импорт\_бд
2. Создайте новый модуль для данного задания перейдите в него

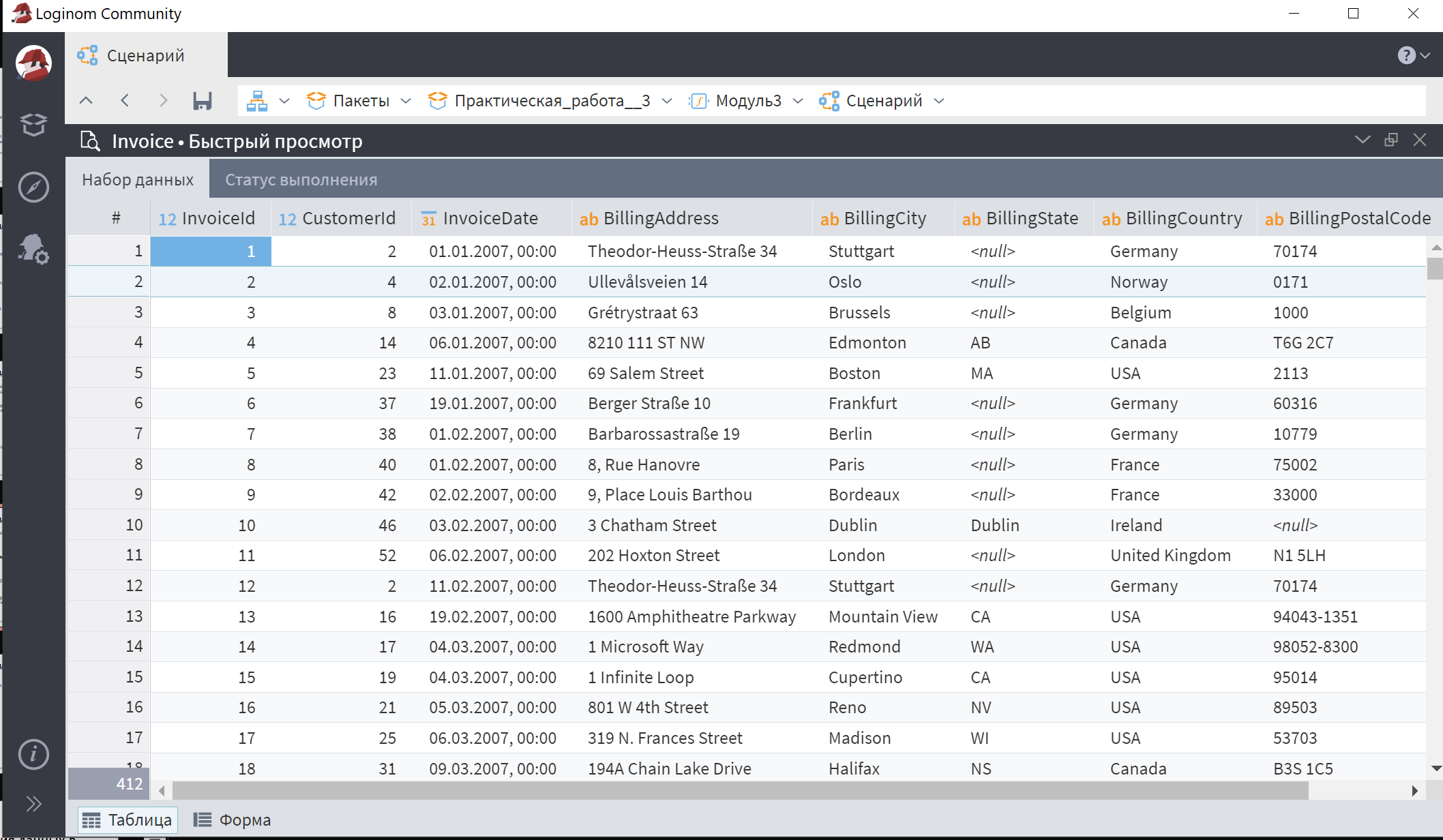


1. Внутри модуля перейдите во вкладу «Подключения»
2. Из папки «Базы данных» перенесите на полотно справа или   
   ПКМ->Добавить подключение компонент SQLite
3. Два раза нажмите на новое созданное подключение для его настройки
4. В настройках укажите в поле «строка подключения» путь к экземпляру базы данных Data/SQLite/Chinook.db
5. Нажмите на кнопку «Тестировать» для проверки корректности соединения с SQLite
6. Нажмите на кнопку «Подключить»
7. Перейдите во вкладку сценарий
8. Снизу раскройте «**Подключения**»
9. В раскрытых подключениях найдите свой пакет и данный модуль, где и будет расположено ваше новое соединение с базой данных



1. Нажмите на него правой кнопкой мыши и выберите опцию «Добавить узел подключения в сценарий»
2. Раскройте вкладку «**Компоненты**» и перенесите на сценарий компонент «**База данных**»
3. Перетащите выходной порт подключения ко входному порту базы данных схожего вида. Активируйте соединение
4. Настройте базу данных на импорт таблицы Invoice. Проверьте данные на выходном порте.





1. Существует ли возможность быстрого просмотра информации для выходного порта узла подключения к базе данных?

Нет

1. Какое количество записей находится в таблице Invoice?

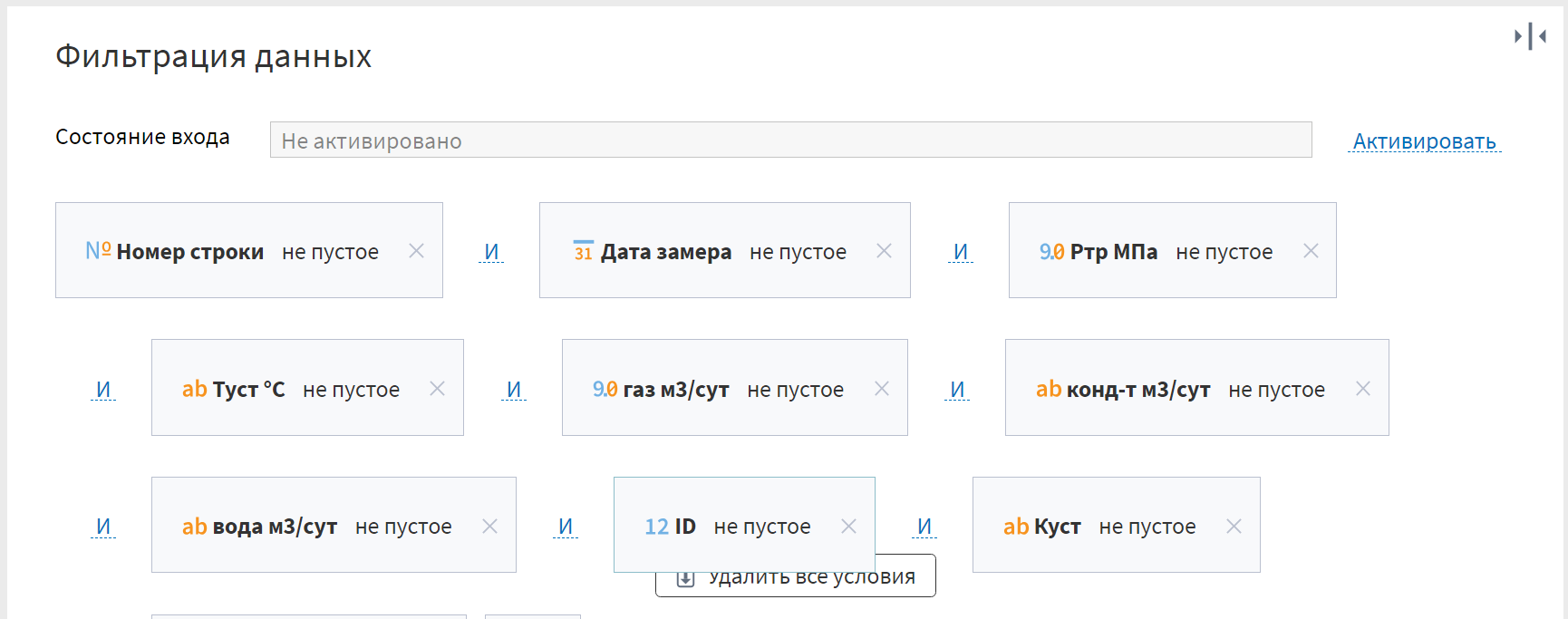
412

1. Какую метку имеет поле первичного ключа для таблицы Invoice?

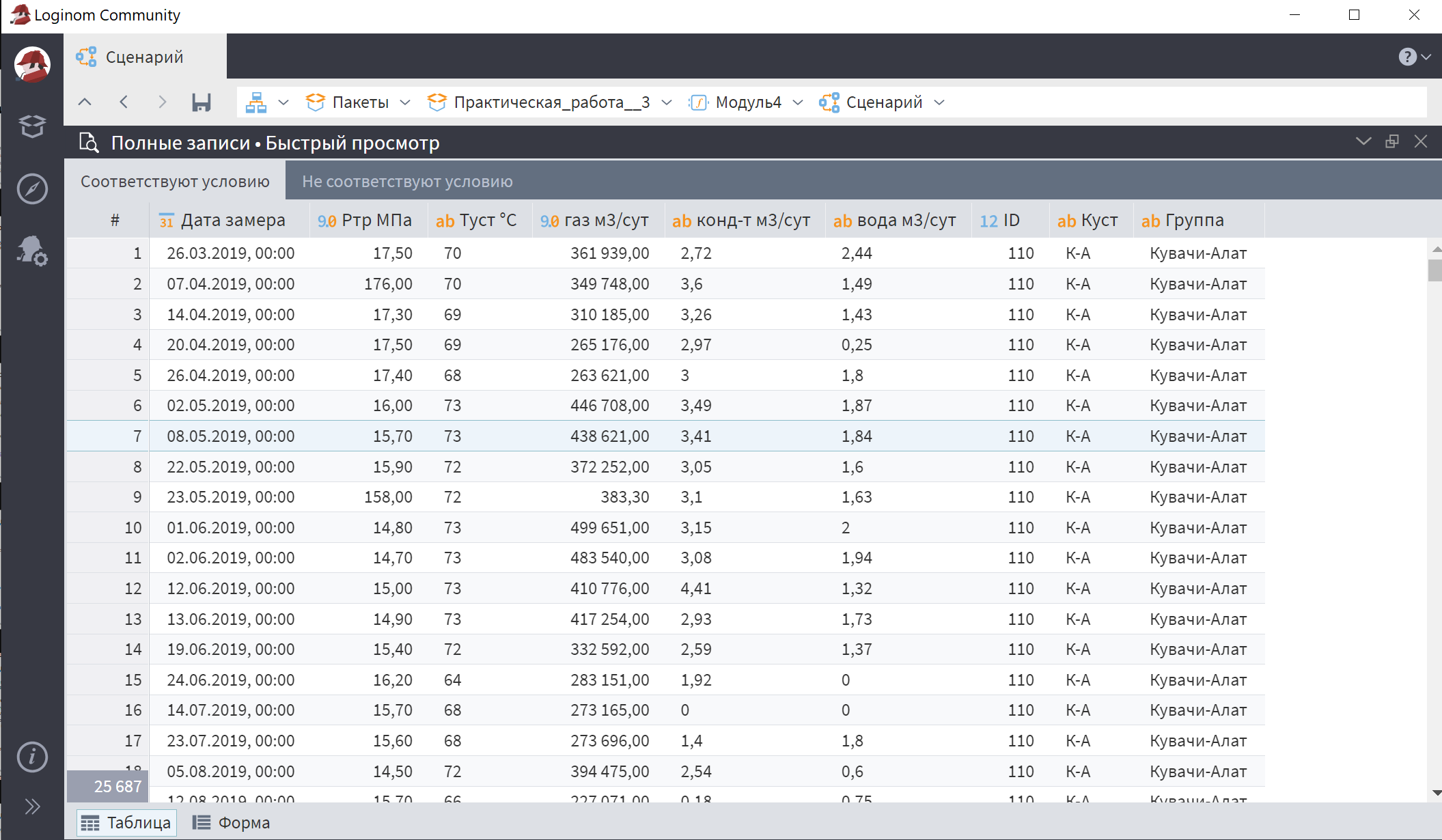
Целочисленный

# Фильтрация пустых полей

1. Модуль demo\_3\_1\_Фильтрация\_данных
2. Для импортированных данных из задания №3 проведите фильтрацию импортированных данных на основе фильтрации записей с пустыми полями
3. Соедините выход порта узла **импорта данных из текстового файла** со входом в узел **фильтр строк**
4. Настройте узел фильтр строк для фильтрации данных
5. При настройке фильтрации данных нажмите на «**+**» и выберите первое поле после «**Номер строки**»
6. Задайте для него условие «**не пустой**»
7. Нажмите **применить** и повторите действие для всех остальных столбцов
8. Нажмите **далее** и настройте название метки узла как «полные записи» и нажмите **выполнить**



1. Для нижнего выходного порта с помощью быстрого просмотра скажите, сколько в нем находится записей



28162

1. Сколько строк данных удовлетворяют условию полных записей?

25687

1. Сколько строк данных не удовлетворяют условию, т.е. содержат пропуски в данных?

2475

1. Какой процент от общего количества строк занимают неполные записи?

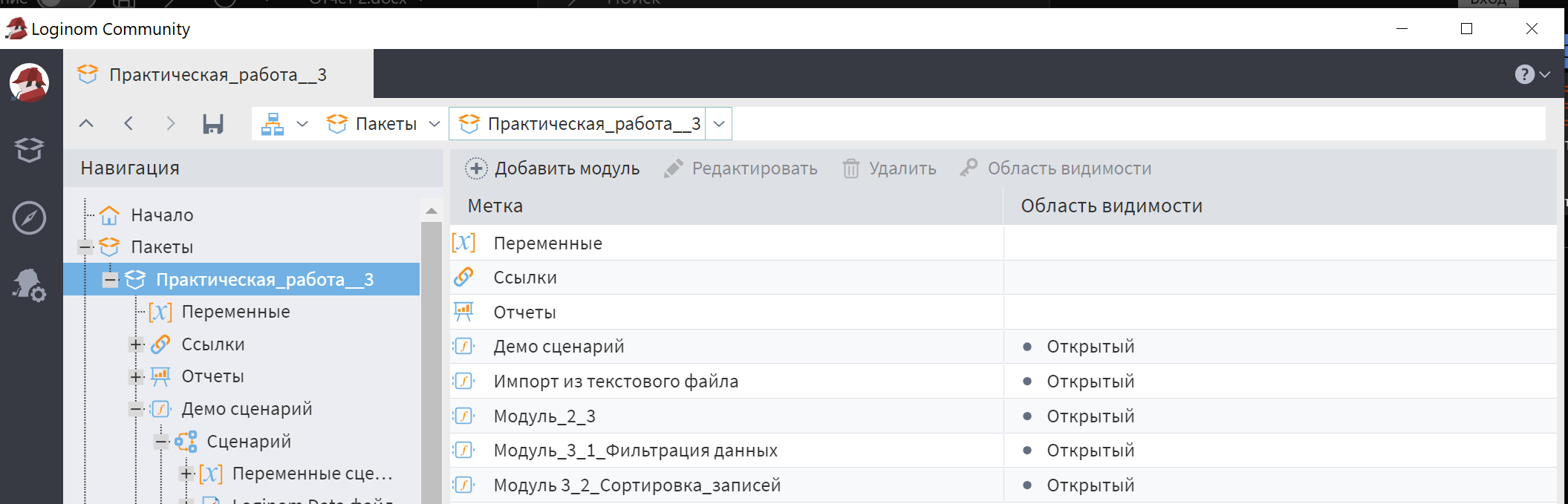
8.78%

1. На основе предпросмотра убедитесь, присутствуют ли в выборке неполных записей те, которые имеют пропуски в нескольких полях?

Да

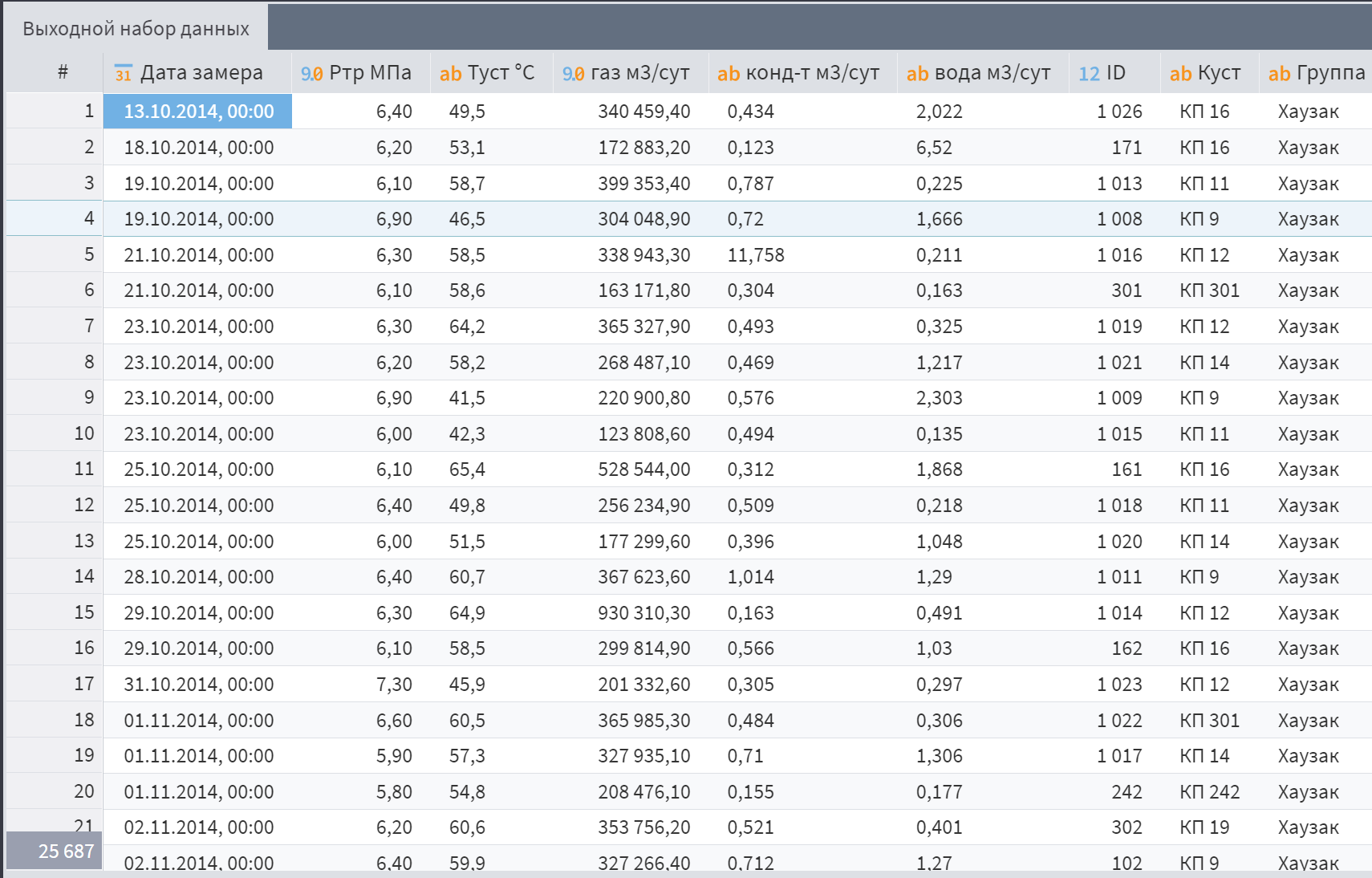
# Сортировка записей

1. Модуль demo\_3\_2\_Сортировка\_данных



1. Для данных, удовлетворяющих условию фильтрации из предыдущего задания произведите сортировку записей по дате замера (по возрастанию) и скорости добычи газа (по убыванию)
2. Перенесите с помощью мыши верхний выходной порт фильтра строк и входной порт сортировки
3. Перейдите в настройку узла сортировки
4. Перенесите в настройках сортировки поля «Дата замера» и «газ м3/сут» в **поля сортировки**
5. Переключателем настройте порядок сортировки как указано ранее
6. Сохраните настройки узла и выполните его в сценарии
7. Убедитесь в корректности сортировки данных на основе быстрого просмотра выходного порта сортировки





1. Какая дата замера оказывается самой ранней в таблице добычи газа?

13.10.2014

1. Какой ID скважины соответствует самой ранней записи замеров скорости добычи?

1026

1. Изменилось ли количество записей в выходной таблице по итогам сортировки таблицы по сравнению с количеством записей на входе?

нет

1. Какая дата замера является самой последней?

24.03.2023

# Использование таблицы замен

1. Модуль demo\_3\_3\_замена
2. Необходимо заменить краткие названия месяцев на полные.
3. В поле построения сценария добавим 2 компонента импорта: Loginom data файл (Исходники) и Текстовый файл (Замена).
4. Подать на вход набор данных из файлов **Data/Demo/data\_months.lgd** и **Data/Demo/ months.txt** соответственно.
5. Таблица исходных данных **Data/Demo/ data\_months.lgd** содержит Идентификатор – Краткое название месяца.
6. Таблица замен **Data/Demo/months.txt** содержит пары значений: **Краткое название месяца** – **Полное название месяца**.
7. Добавим в область построения сценария компонент **Замена**.
8. Подключим наборы данных к соответствующим портам узла и откроем настройку узла **Замена**.
9. Для полей исходных данных можно настроить **Способ замены**:

**Не заменять**.

**Ввод вручную** – формируем внутреннюю таблицу замен.

**Таблица замен** – используем внешнюю таблицу замен.

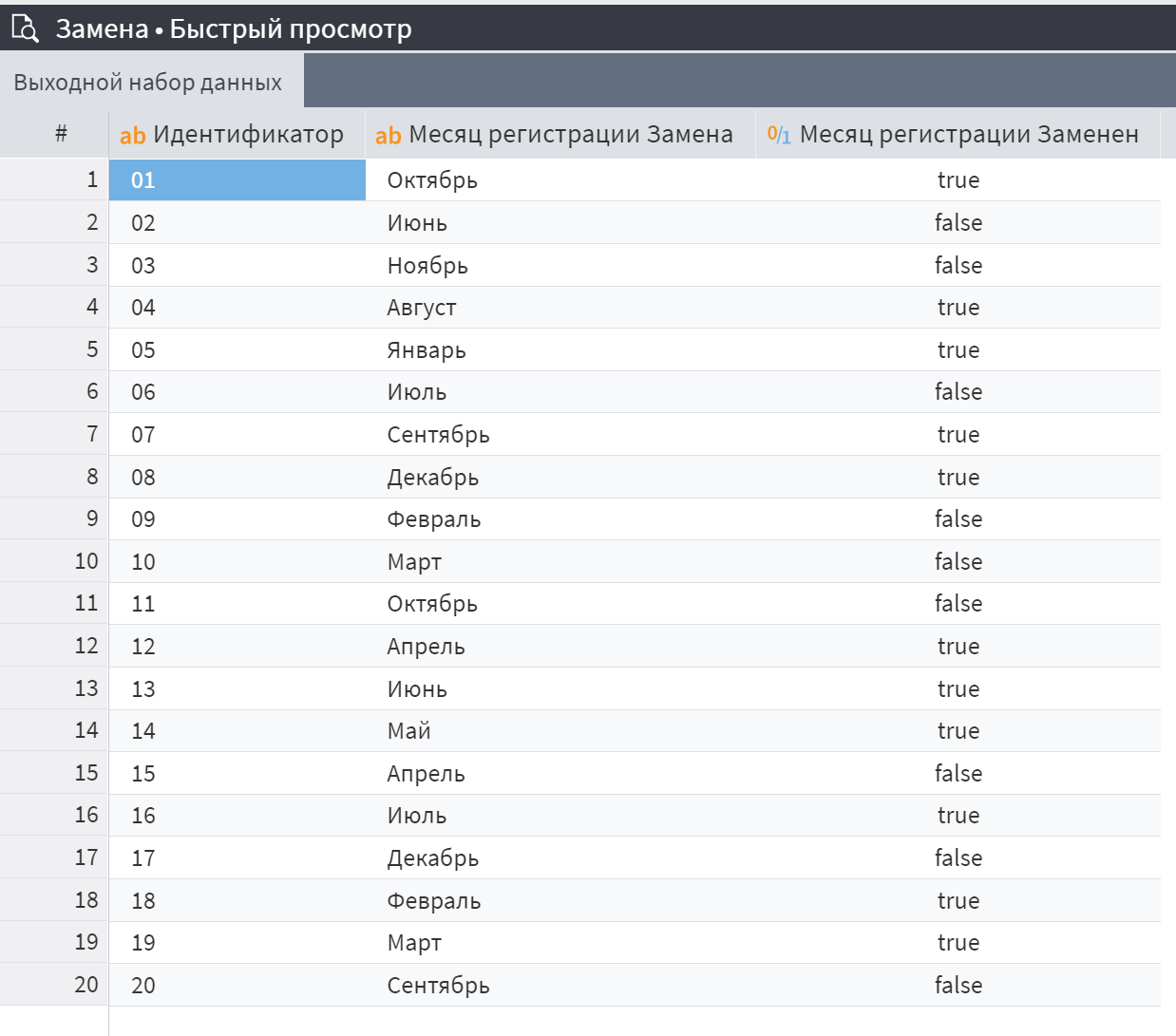
1. Выберем для поля **Месяц** **регистрации** способ **Таблица замен**

Назначение поля *Краткое название* установим **Значение**, поля *Полное название* – **Замена**.

Если в исходных данных встретятся значения, для которых не будет совпадений в таблице замен, можно выбрать вариант их замены во вкладке **Заменять остальные**.

На следующем шаге настройки, в окне *Настройка соответствия между столбцами* по умолчанию используется соответствие столбцов Замена: столбец с исходными значениями заменится на столбец с новыми значениями.

1. Всем ли записям в выходной таблице данных соответствует значение **true** в поле **Месяц регистрации заменен**?



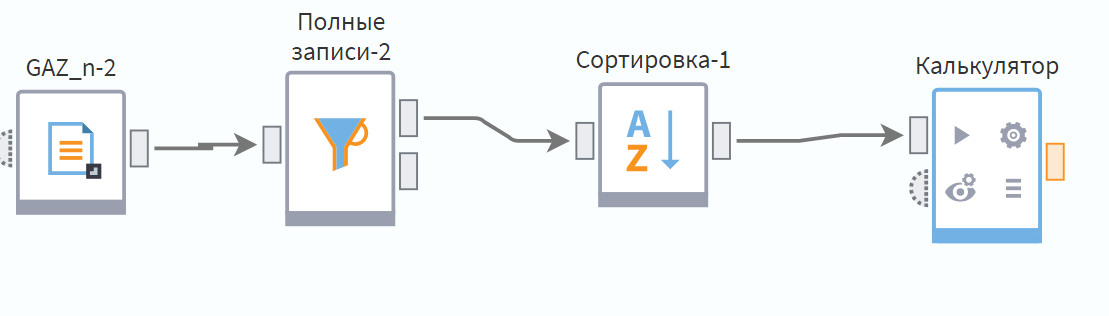
Нет

1. Есть ли во входной таблице данных такие записи, заменять которые было не нужно

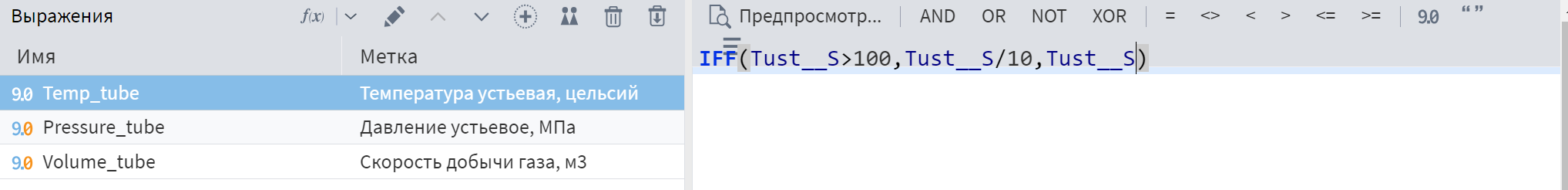
Да

# Вычисление новых полей

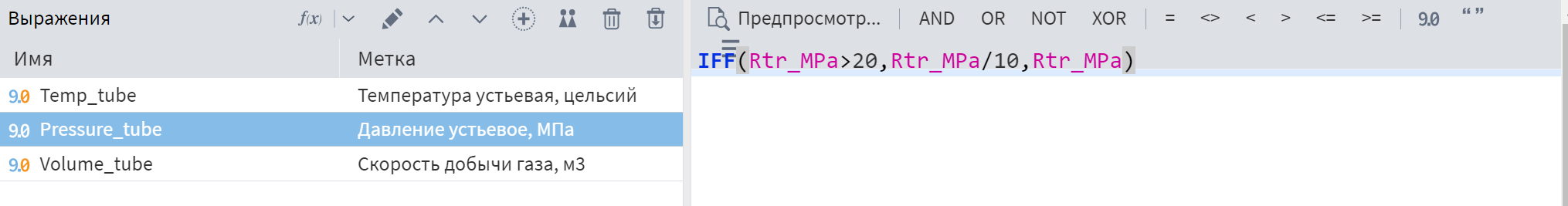
1. Модуль demo\_3\_4\_калькулятор
2. В данном задании для упорядоченной выборки необходимо подкорректировать числовые значения некоторых полей
3. Для задания с сортировкой данных продолжим построение тестового сценария обработки данных добычи газа
4. Добавим в область построения сценария компонент калькулятор

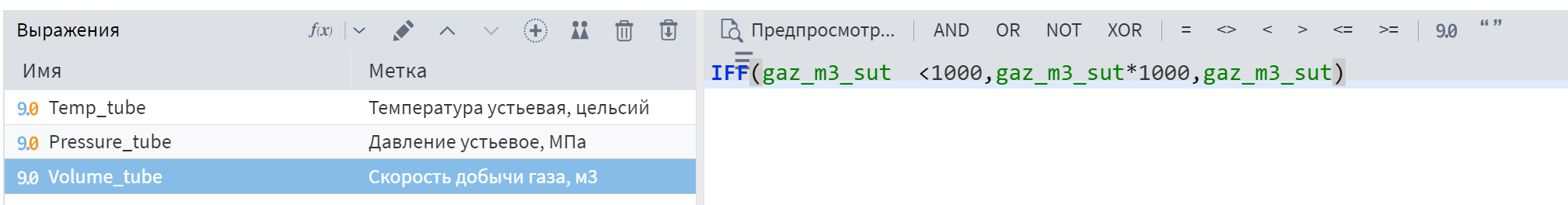


1. Соединим входной порт калькулятора с выходным портом сортировки из задания №6. Перейдем к настройке узла «Калькулятор»
2. В поле выражения с помощью кнопки *добавить выражение* добавим новое вычисляемое поле с именем **Temp\_tube**, и меткой *Температура устьевая, цельсий*
3. Запишем для его вычисления следующее условное выражение, которое в поле **Tust\_\_S** проверяет условие и в случае успешного прохождения производит деление значения в записи на 10



1. При написании выражений можно просто два раза кликнуть ЛКМ на нужное входное поле, чтобы оно переместилось в выражение
2. Также создадим поля **Pressure\_tube** и **Volume\_tube** с соответствующими выражениями для их вычисления





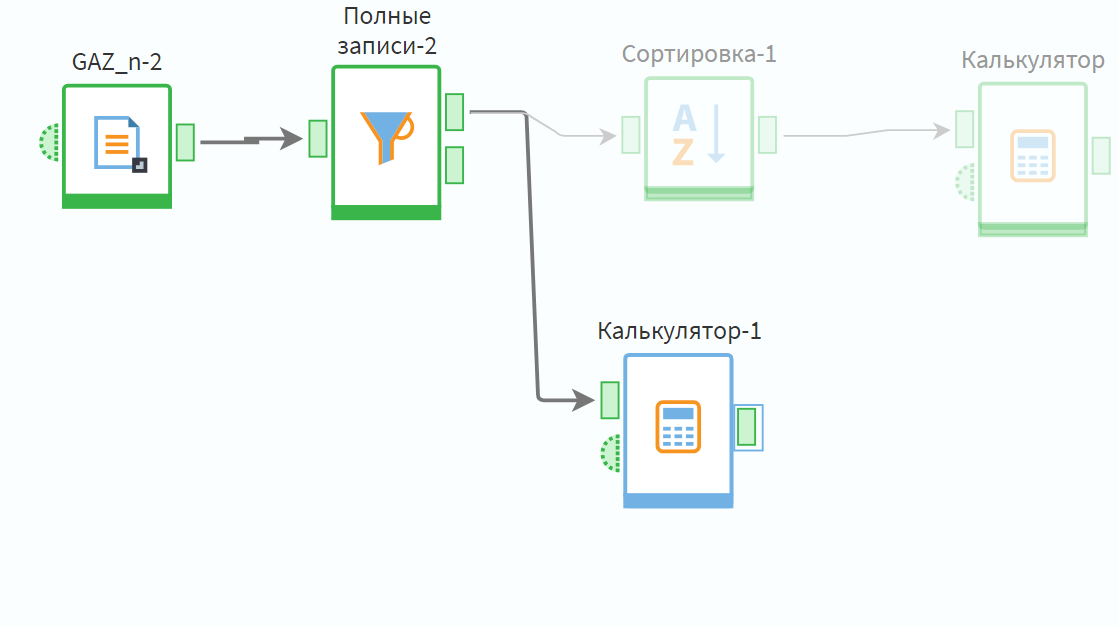
1. Данные выражения основаны на анализе предметной области и обзора данных с точки зрения анализа качества выборки
2. Сохраните настройки узла и выполните узел
3. С помощью быстрого просмотра выходного порта просмотрите какие поля получились в итоге



1. Сколько новых полей появилось в выходной таблице данных после вычисления выражений в узле калькулятор?

0

1. Скопируйте настроенный узел калькулятора, вставьте его в сценарий и подайте на его вход данные из верхнего выходного порта фильтра строк.
2. Выполните данную копию узла.



1. Просмотрите выходную таблицу данных у копии калькулятора.
2. Сравните поля: **Скорость добычи газа, м3.сут** и **газ м3/сут**. В какой строчке произошли изменения?

8

1. **Также сравните другие вычисленные поля с их базовыми**

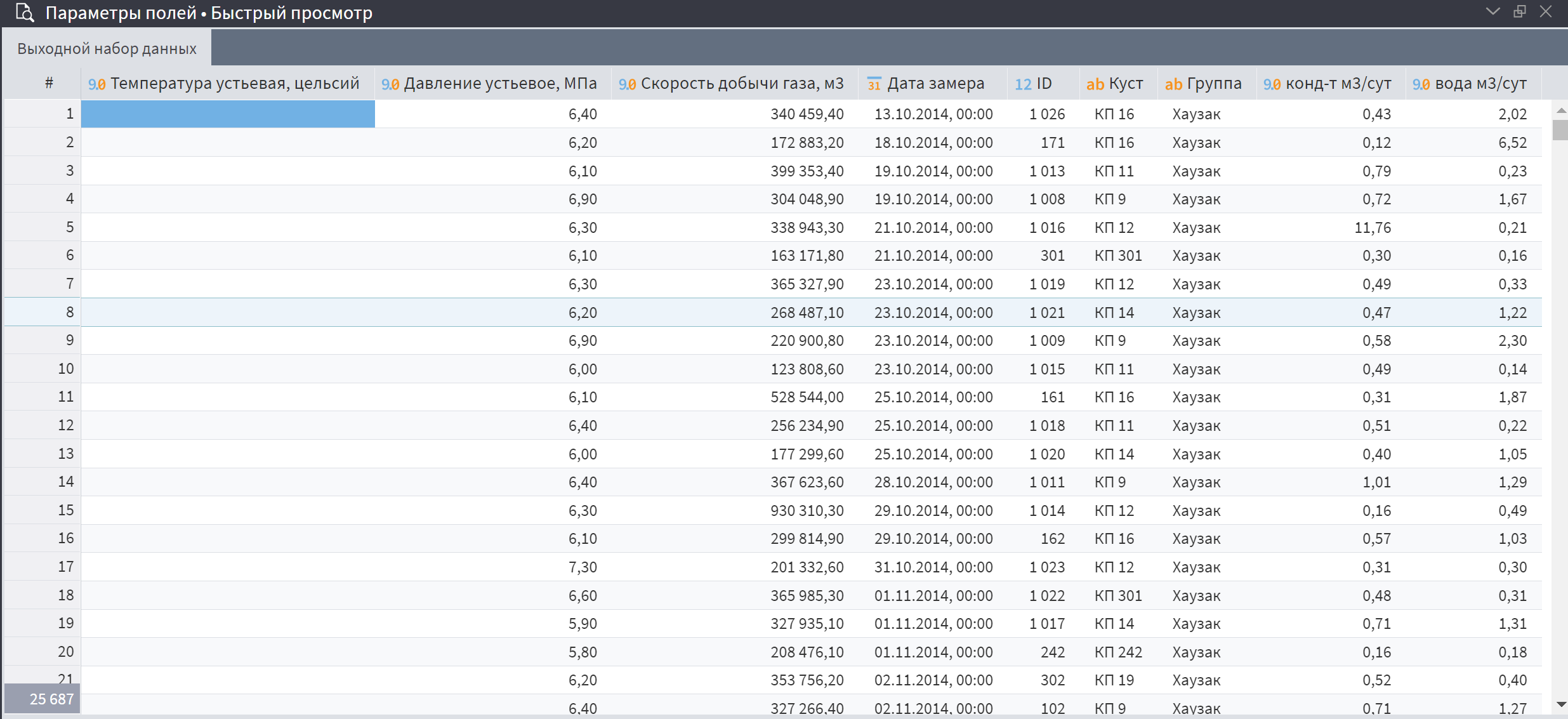
Других изменений нет

# Параметры полей

1. Модуль demo\_3\_5\_параметры\_полей
2. Для предыдущего задания добавьте в сценарий обработки данных компонент параметры полей, чтобы явно преобразовать структуру данных после проведения расчетов в калькуляторе
3. Подайте выходной порт калькулятора на входной порт «**Параметры полей**»
4. Настройте параметры полей как показано на слайде, исключив поля, на основе которых проводилось вычисление новых, а также изменив метку для полей **конд-т** и **вода** на новые метки



1. Выполните узел и выполните быстрый просмотр выходного порта из узла «Параметры полей»

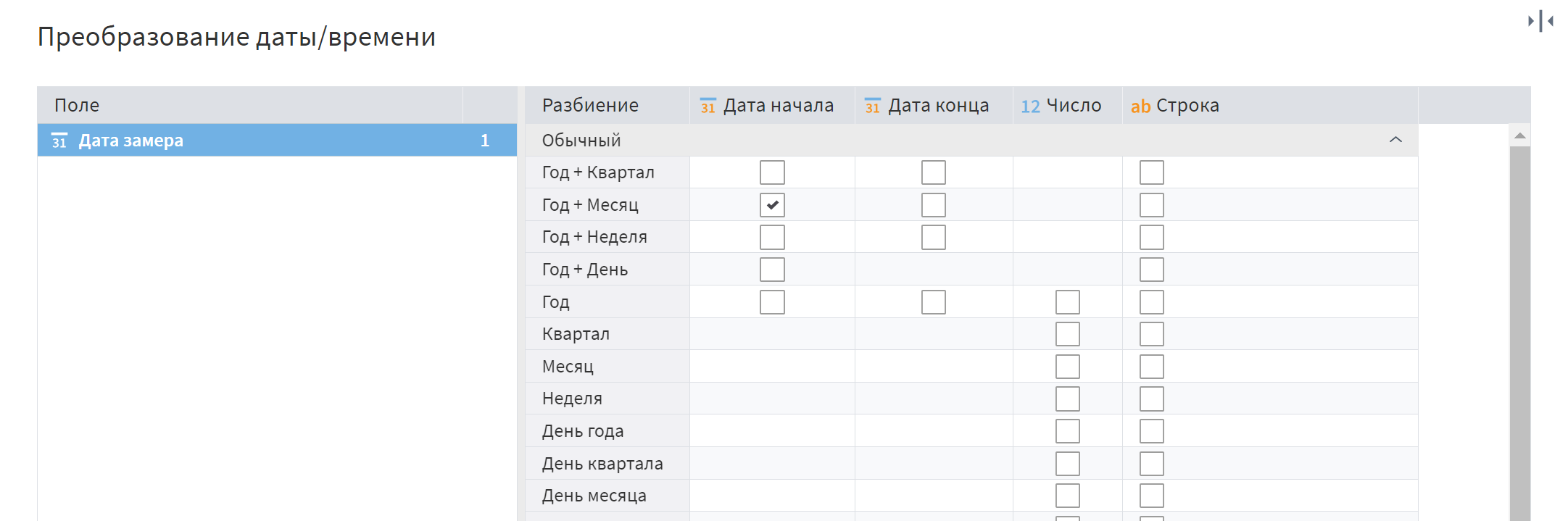


# Выделение начала месяца

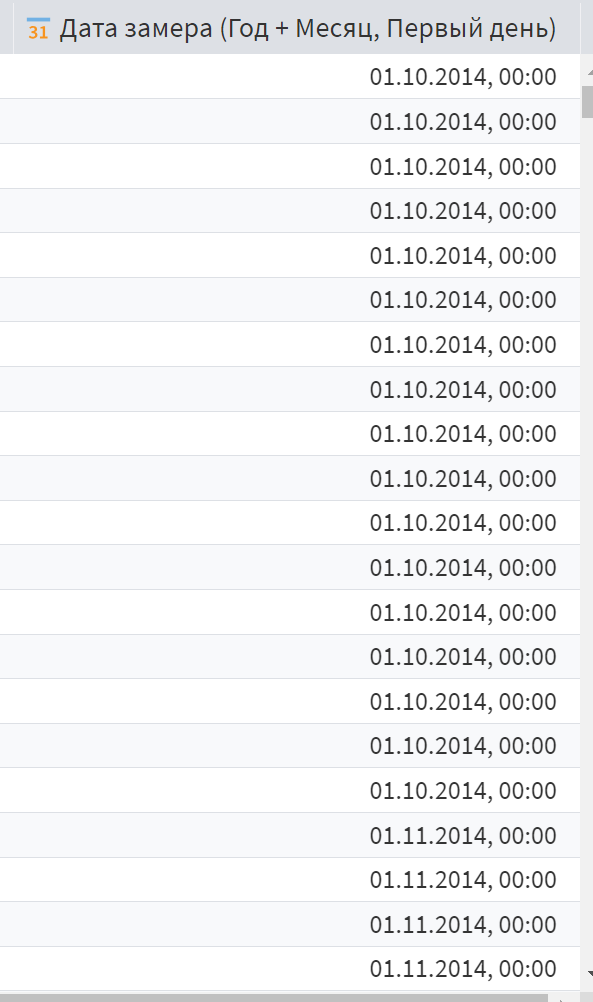
1. Модуль demo\_3\_6\_дата\_время
2. Для созданного сценария в задании №9 добавим компонент даты и времени для выделения даты начала месяца добычи газа.



1. Даты замеров в нашей выборке данных размещены неравномерно, что мешает обрабатывать данные временных рядов.
2. Добавим в сценарий модуля обработки данных компонент «**Дата и время**»
3. Соединим выходной порт «Параметров полей» с входным портом «Дата и время»
4. Для поля «**Дата замера**» выберем разбиение по дате начала *Год+месяц*

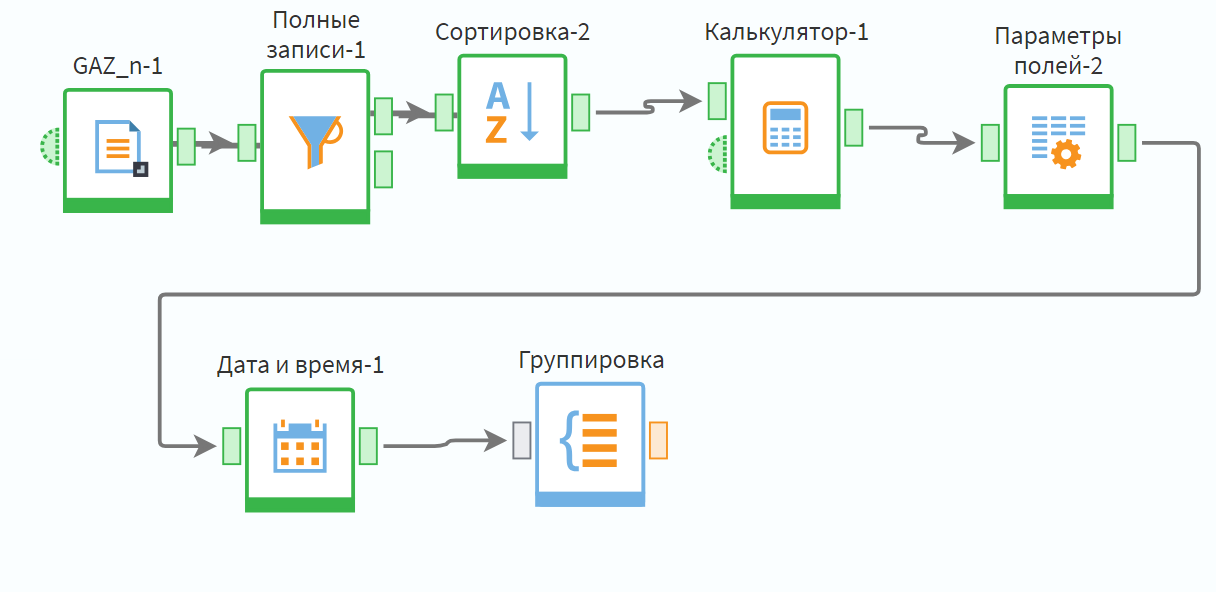


1. Сохраним настройки узла и выполним его
2. На выходном порте узла активируем быстрый просмотр, чтобы убедиться в результате работы компонента

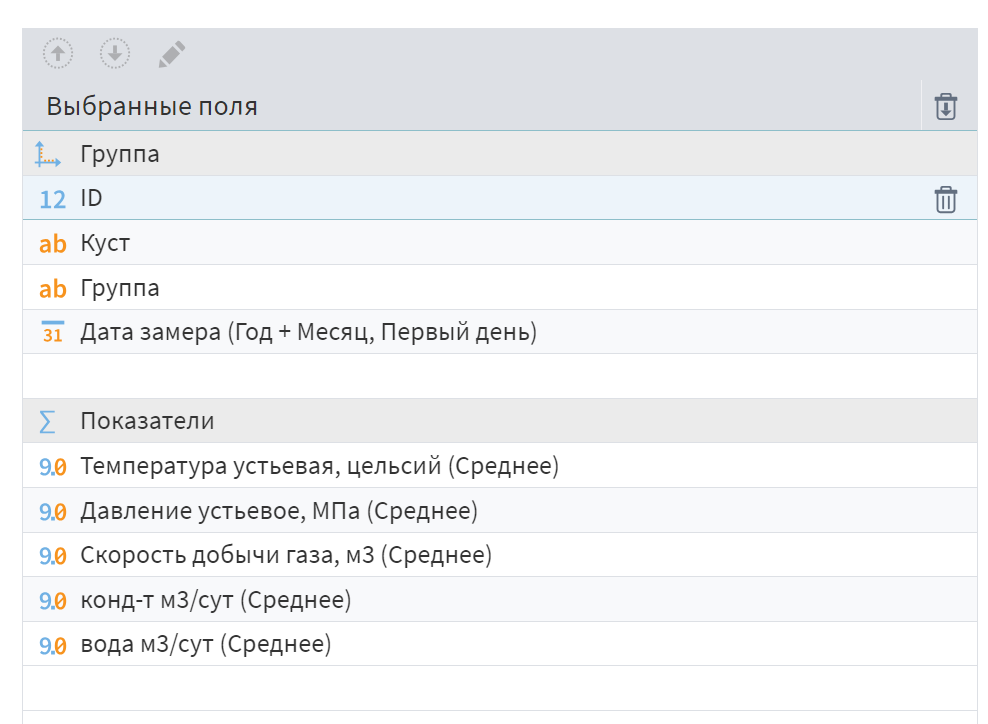


# Группировка данных в добыче

1. Модуль demo\_3\_7\_группировка\_данных
2. Добавим в сценарий обработки данных расчёт среднемесячных показателей для того, чтобы нормировать временные показатели измерений в одну интервальную шкалу
3. Добавим в сценарий из задания №10 компонент «Группировка»
4. Соединим выходной порт «Дата и время» с входным портом «Группировка»



1. В настройке группировки выберем следующие поля в качестве групп:
2. Дата замера (Год+месяц)
   * ID
   * Куст
   * Группа
3. В настройке группировки выберем все числовые поля в качестве показателей и рассчитаем для них **среднее**

****

1. Сколько записей мы получили в результате операции группировки данных?

79

1. Какие месяц и год были первыми в списке после группировки?

01.10.2014

1. Какие скважины по ID наблюдались в первый месяц в выходной выборке данных?

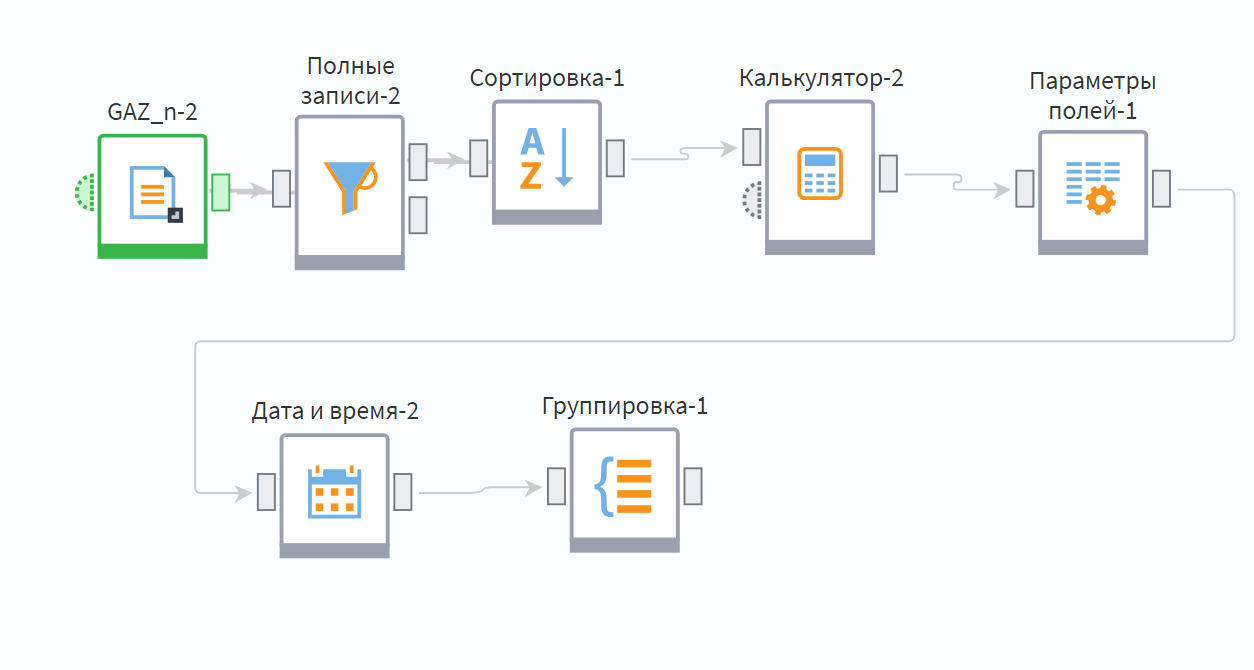
|  |
| --- |
| 171 |
| 1 013 |
| 1 008 |
| 1 016 |
| 301 |
| 1 019 |
| 1 021 |
| 1 009 |
| 1 015 |
| 161 |
| 1 018 |
| 1 020 |
| 1 011 |
| 1 014 |
| 162 |
| 1 023 |

1. Какова максимальная скорость добычи конденсата в первый месяц наблюдения? (до 2 знака после запятой)

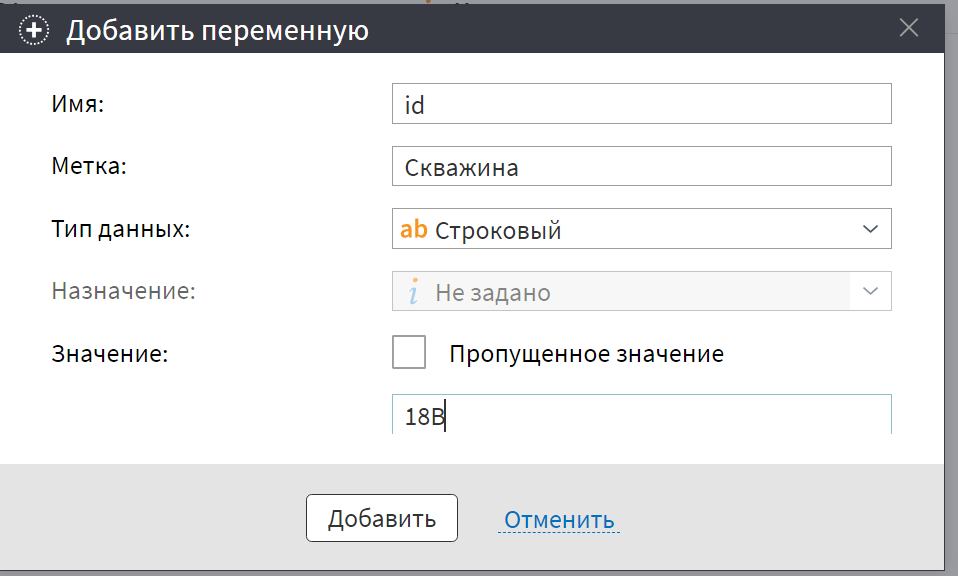
930 310,3

# Работа с переменными Loginom

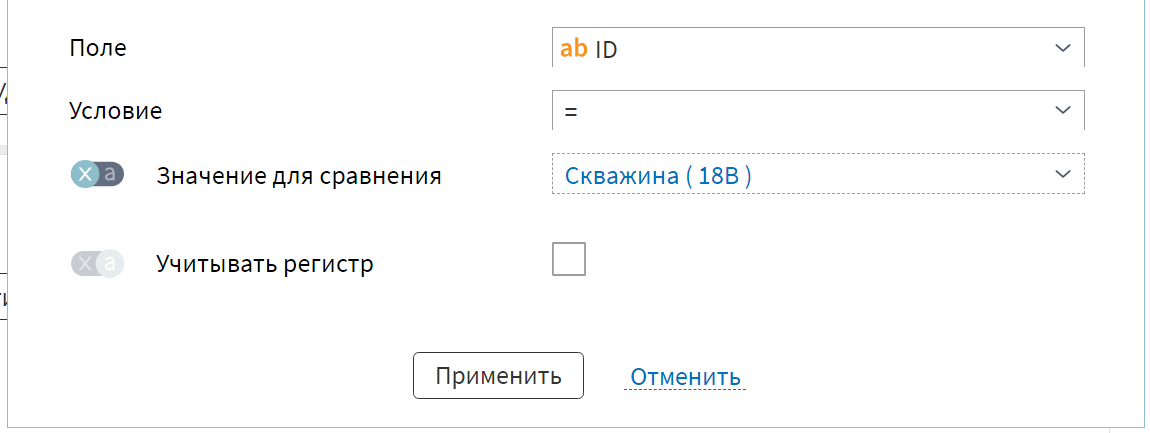
1. Зададим параметризацию нашего сценария обработки данных добычи газа на месторождениях
2. Добавим в сценарий возможность считывать имя файла из входной переменной
3. Осуществлять фильтрацию данных по параметру ID скважины
4. Задавать ограничения в калькуляторе с помощью переменных



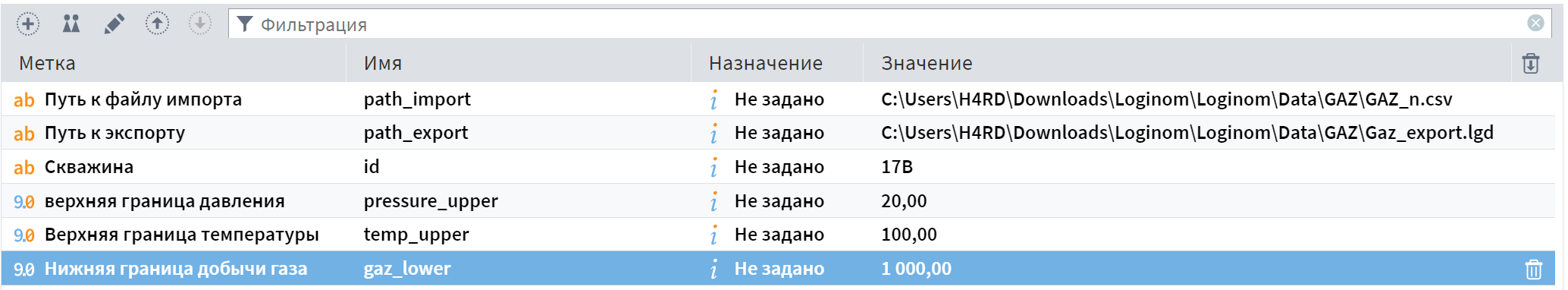
1. С помощью ПКМ по узлу и выбора опции меню «Показать порт управляющих переменных» для фильтра строк создадим у него порт переменных, который сможет принимать входные параметры
2. С помощью ПКМ по появившемуся порту перейдем в его настройки
3. С помощью кнопки **добавить** мы сможем задать новую подаваемую переменную заданного типа с значением по умолчанию
4. Создадим новую переменную как показано на слайде
5. Отключим **автосинхронизацию** с помощью кнопки рядом с кнопкой **добавить**



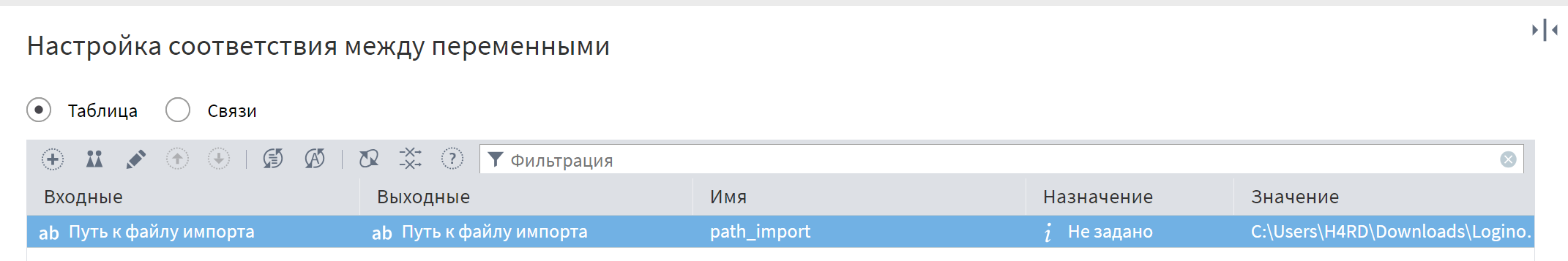
1. После создания новой переменной на входе порта управляющей переменной сохраним все настройки порта и вернемся в область построения сценария
2. Настроим узел «Фильтр строк»
3. В перечень условий добавим ещё одно условие по совпадению ID скважины со значением из внешней переменной
4. Если порт управляющих переменных был настроен корректно, то в значении для сравнения будет успешно добавлен переключатель, позволяющий передавать в данное поле значение из внешней переменной



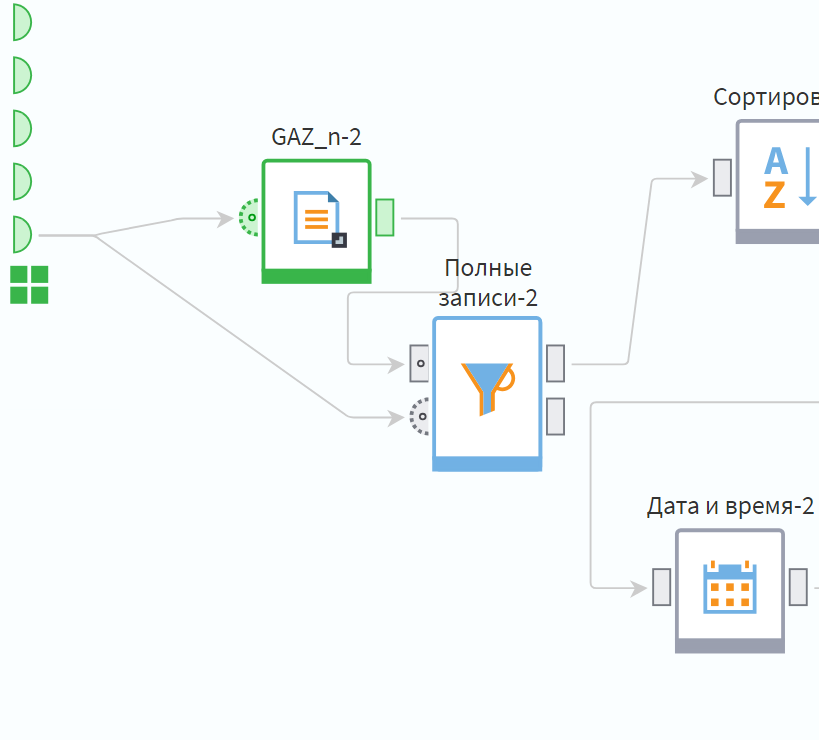
1. На выходном порте данных, соответствующих условию наблюдаем ожидаемую функциональность, что поданное значение по умолчанию для фильтрации данных отрабатывает в штатном режиме
2. Далее добавим в сценарий пользовательские переменные, которые можно будет использовать для эффективной параметризации сценария
3. На порте переменных пользователя щелкнем ПКМ и перейдем к настройке порта
4. Создадим на порте набор переменных, показанных на слайде



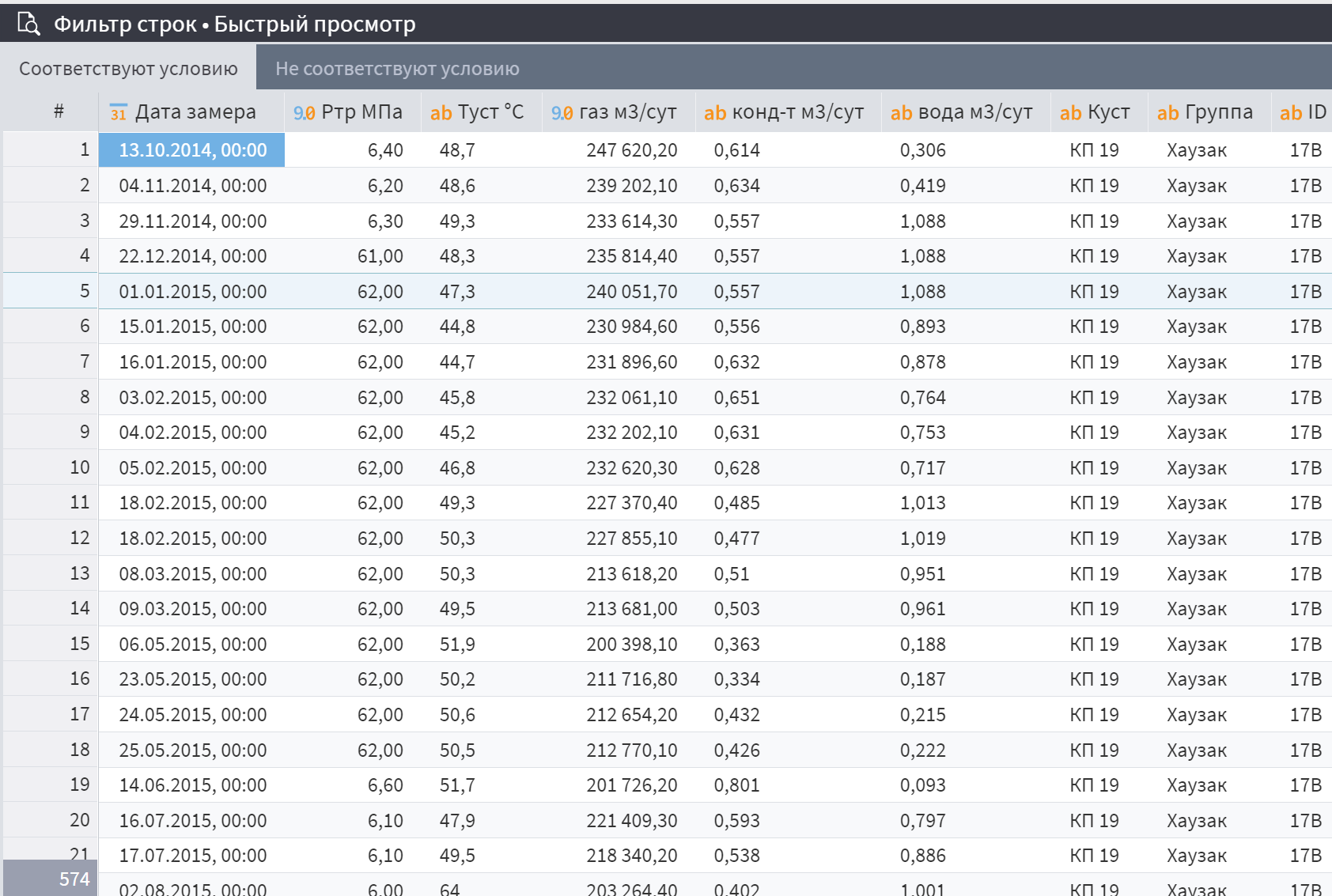
1. Сохраним настройки
2. Соединим порт переменных пользователя с входной переменной добычи газа



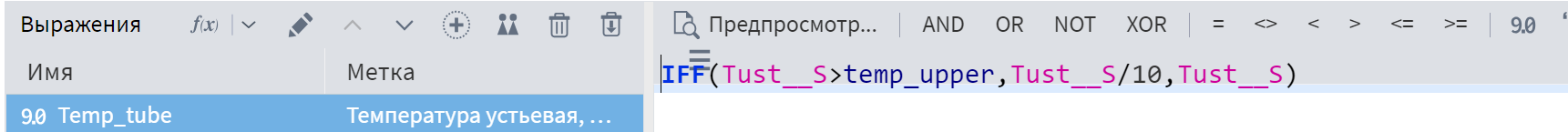
1. Перейдем к настройке порта входных переменных узла импорта данных
2. В настройках порта удалим все переменные, полученные от порта переменных пользователя и оставим только переменную строкового типа **path\_import**
3. Автосинхронизация отключается автоматически при удалении полей на порте узла
4. Сохраним настройки порта и перейдем к настройкам узла импорта данных
5. В настройках узла импорта данных включим передачу значения для **Имя файла** из внешней переменной, нажав на переключатель рядом с полем
6. Заметим, что значение из переменной успешно передано
7. Остальные настройки импорта оставляем без изменений
8. Включим импорт данных, на основе быстрого просмотра убедимся в том, что данные успешно импортированы в сценарий обработки данных
9. Передадим переменные пользователя на вход фильтра строк и перейдем к его настройкам

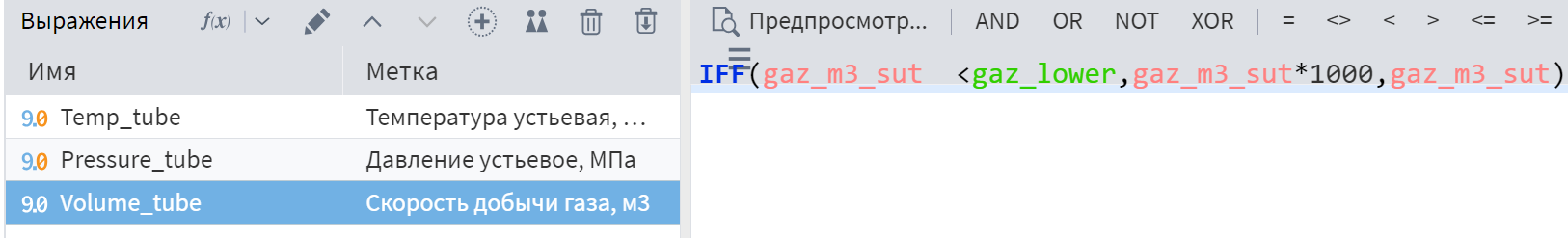
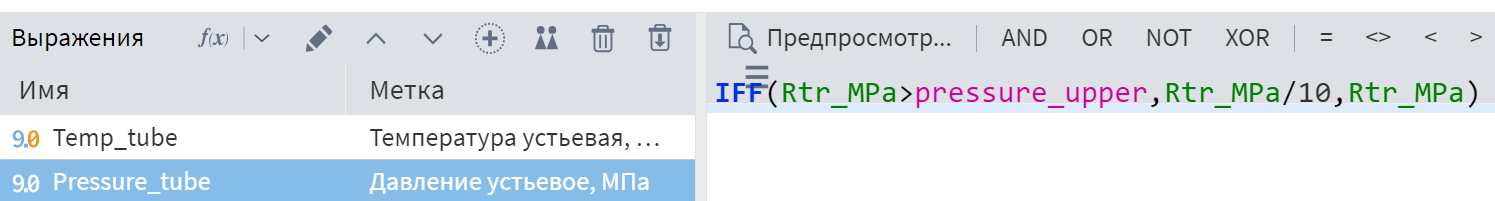


1. В порте фильтра строк укажем в качестве входного значения для настроенной входной переменной поле   
   **Скважина (id)**.
2. Сохраним настройки и запустим выполнение узла
3. На основе быстрого просмотра поля данных, соответствующих условию убедимся в том, что фильтр строк работает в соответствии с поданным на него значением для сравнения из порта входных переменных

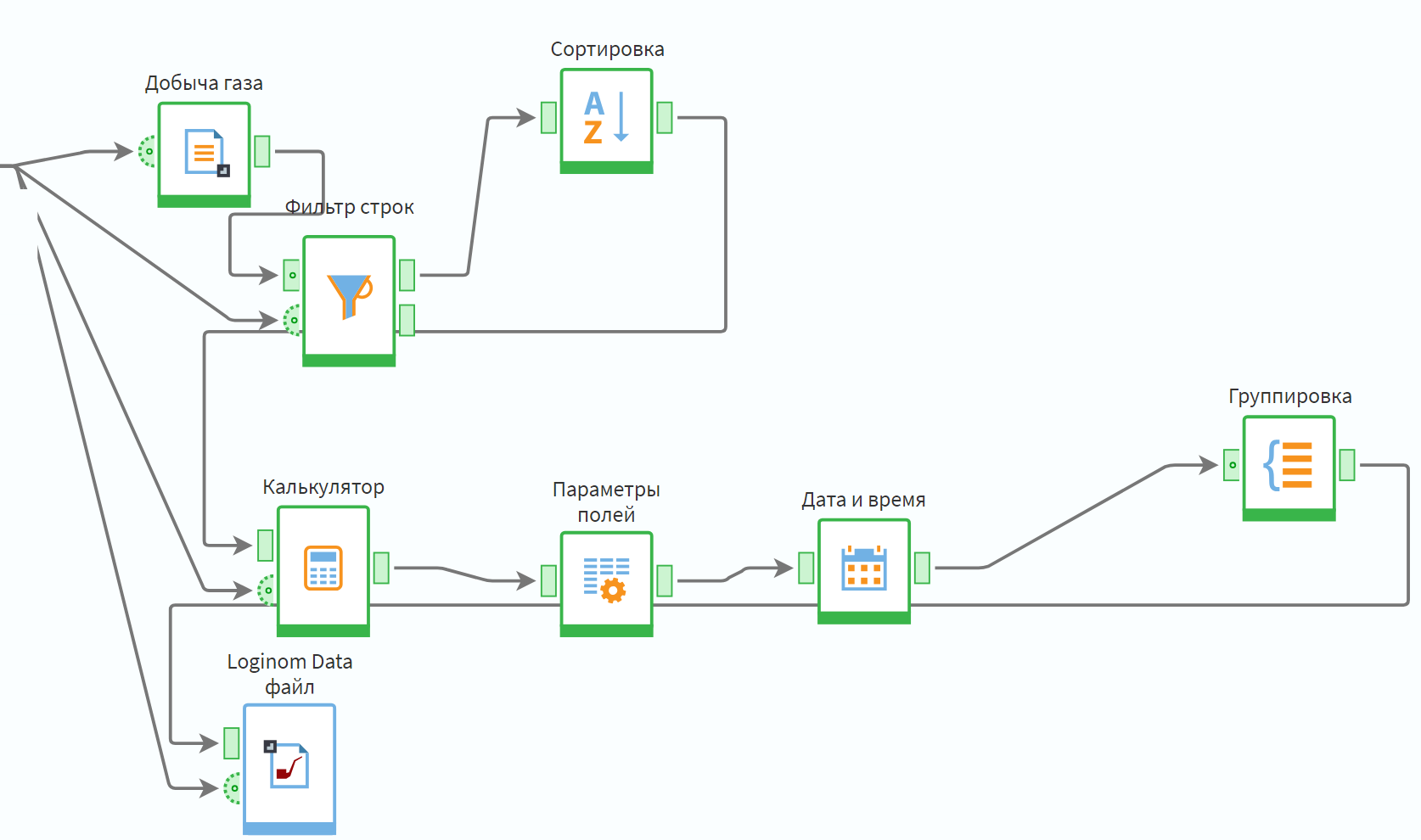


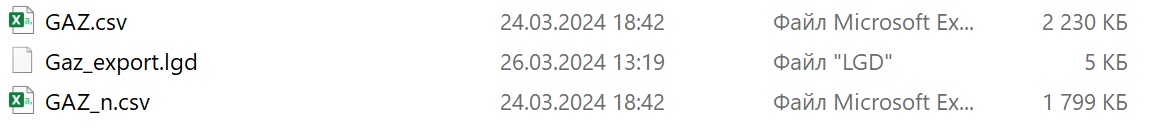
1. Добавьте на входном порте переменных калькулятора переменные, отвечающие за прохождение условия для вычисляемых в нем значений в соответствии с названием и меткой созданных переменных пользователя и полей калькулятора из задания №8





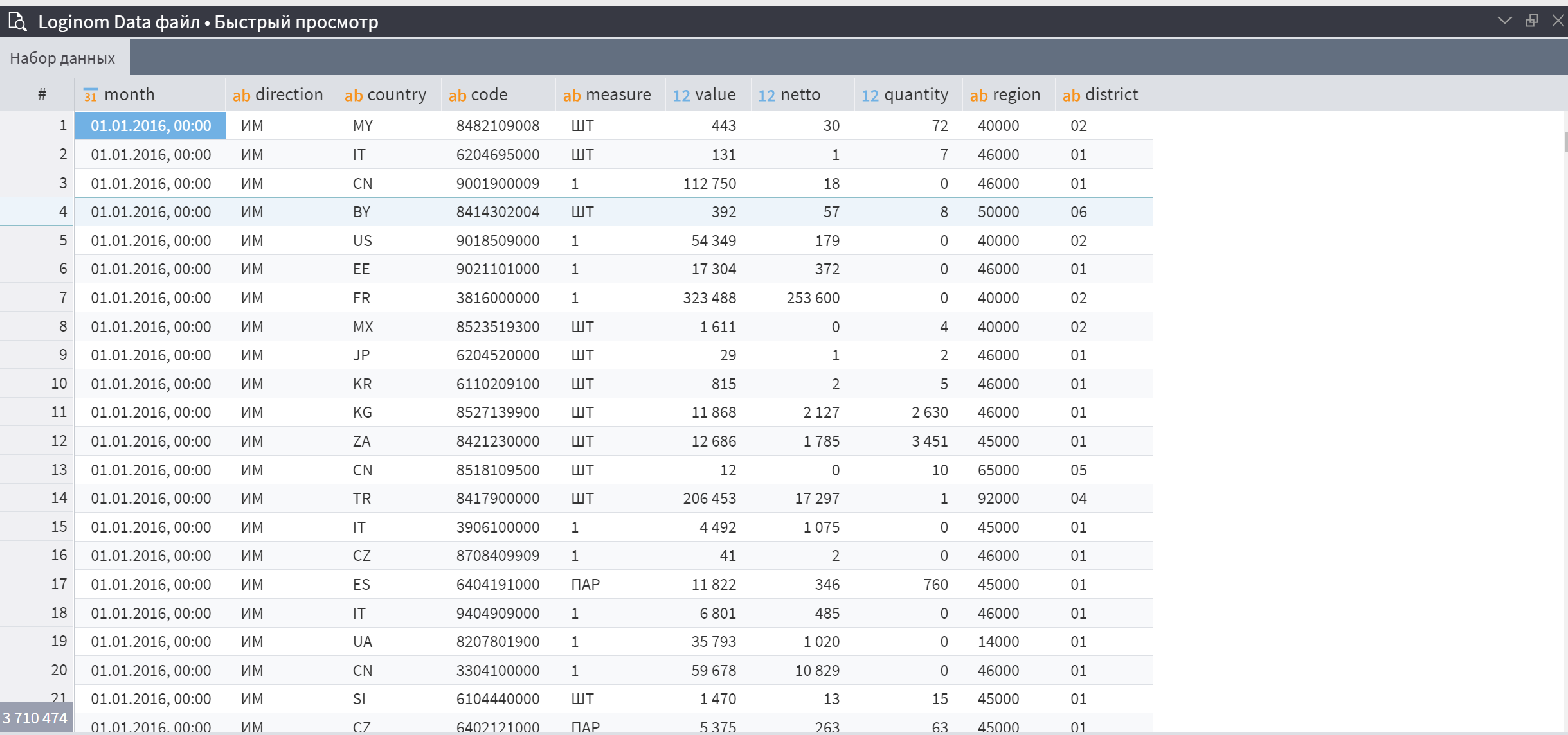
1. Добавьте узел экспорта данных, имя файла для записи в котором будет определяться внешней переменой пользователя





# Самостоятельная работа

1. Импортировать данные таможенной статистики по экспорту и импорту в регионах РФ за 2016 год
2. Сколько строк имеет выборка данных за 2016 год?



3 710 474

1. Какое количество столбцов имеет данная выборка?

10

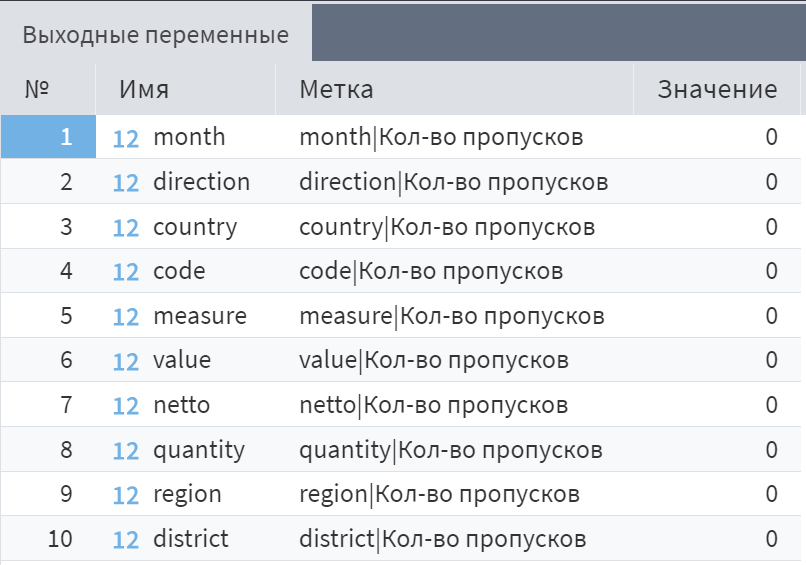
1. Какую информацию хранит каждое поле в отдельности? Какие данные необходимо еще добыть, чтобы расшифровать значения в полях **country, code, region, district**?

Списки сокращений полных названий для country

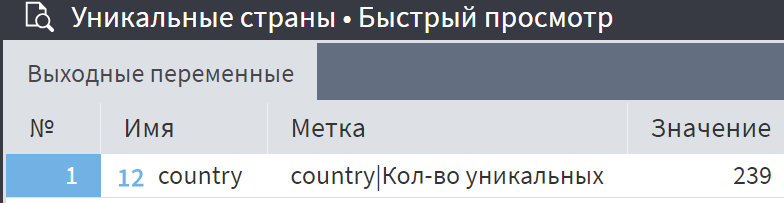
Список преобразования кода в значение для code

Список преобразования сокращений регионов и районов для region и district

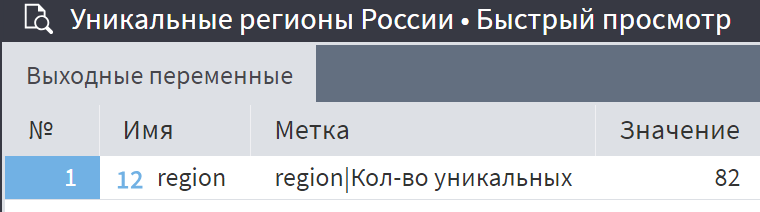
1. Имеются ли в данных пропуски? На основе компонентов **таблица в переменные, группировка,** визуализатор **статистика**, ответьте на вопрос о количестве пропущенных значений в каждом из столбцов выборки данных (можно использовать что-то одно).



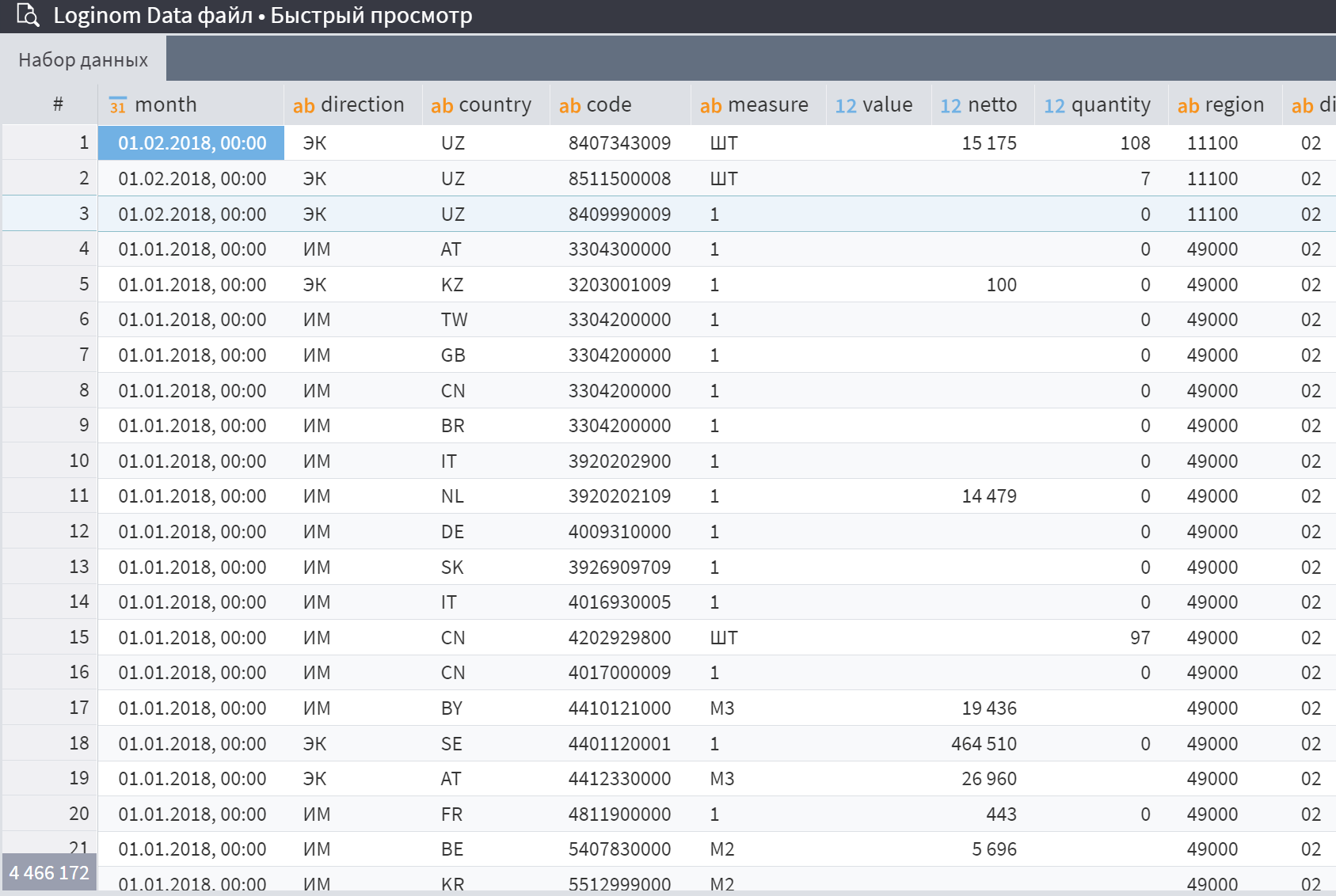
1. Сколько уникальных стран участников торговли перечислено в поле **country**?



1. Сколько уникальных регионов России участвовало в торговле (поле **region**)?



1. Импортировать данные таможенной статистики по экспорту и импорту в регионах РФ за 2018 год
2. Сколько строк имеет выборка данных за 2018 год?

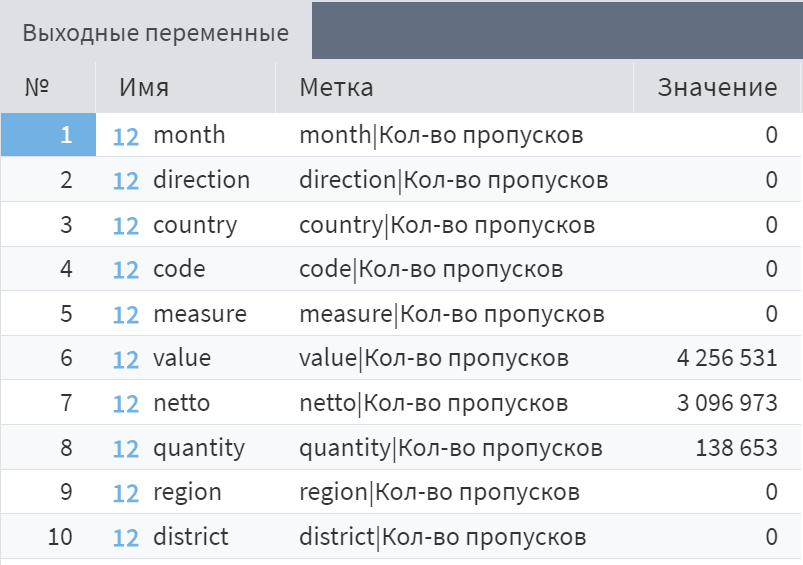


4 466 172

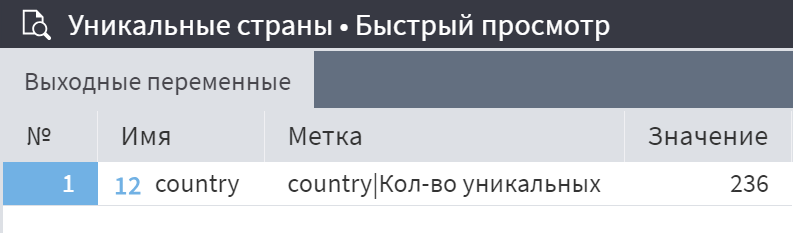
1. Какое количество столбцов имеет данная выборка?

10

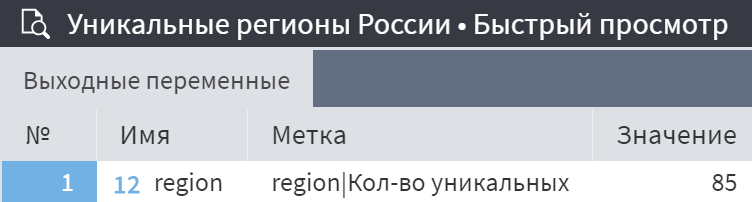
1. Имеются ли в данных пропуски? На основе компонентов **таблица в переменные, группировка,** визуализатор **статистика**, ответьте на вопрос о количестве пропущенных значений в каждом из столбцов выборки данных (можно использовать что-то одно).



1. Сколько уникальных стран участников торговли перечислено в поле **country**?



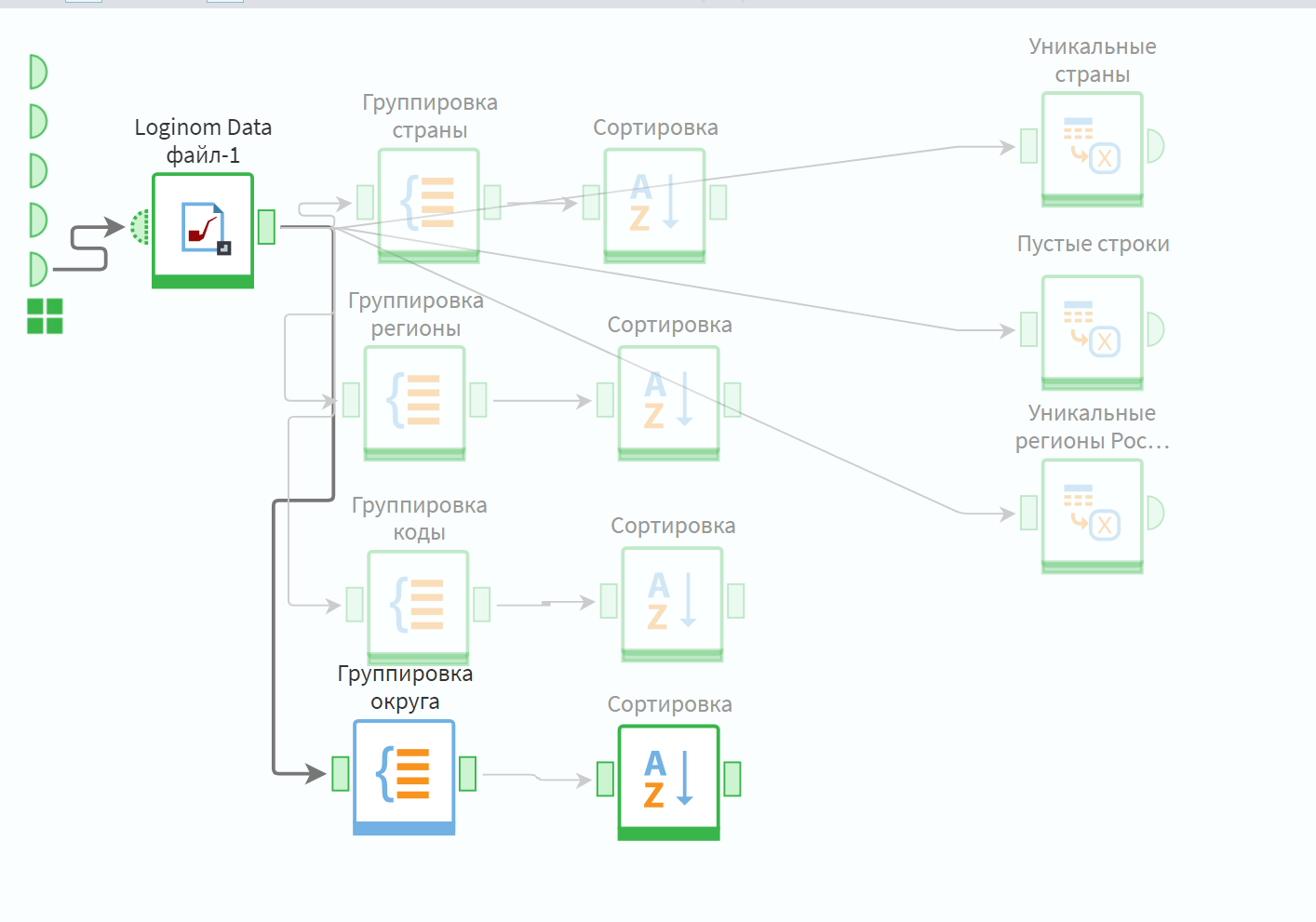
1. Сколько уникальных регионов России участвовало в торговле (поле **region**)?



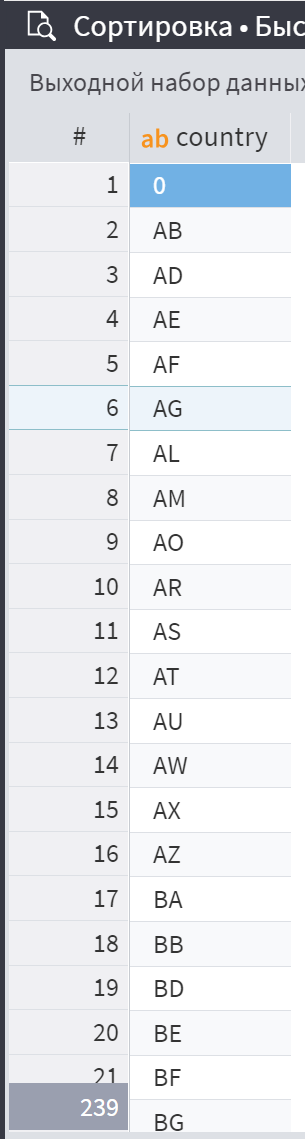
1. Сформируйте вывод о динамике качества данных собираемой статистики во времени и перечислите поля, дальнейший анализ по которым будет репрезентативен и валиден.

Не валидным будет анализ данных по полям value, nettо и quantity

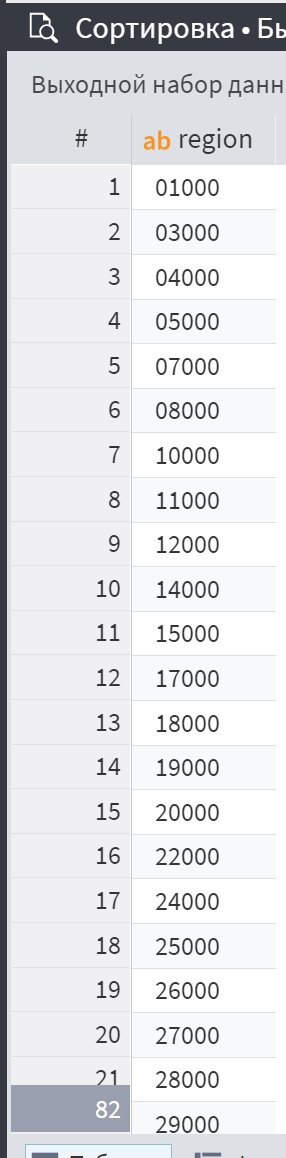
Для импортированной выборки за 2016 год на основе компонентов трансформации данных отобразите:



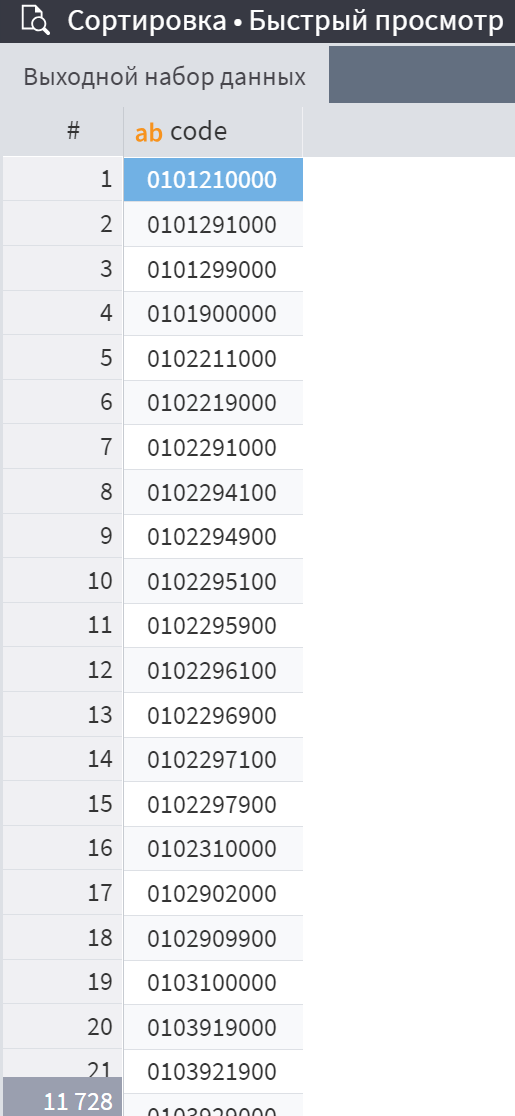
1. Перечень уникальных стран-участниц торговли в поле **country** с помощью компонента группировка (отсортируйте значения по возрастанию)



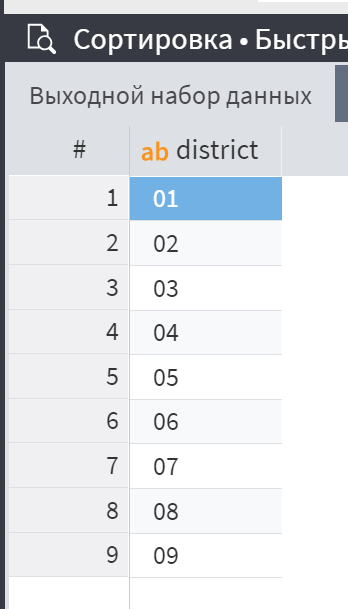
1. Перечень уникальных регионов России участников торговли в поле **region** с помощью компонента группировка (отсортируйте значения по возрастанию)



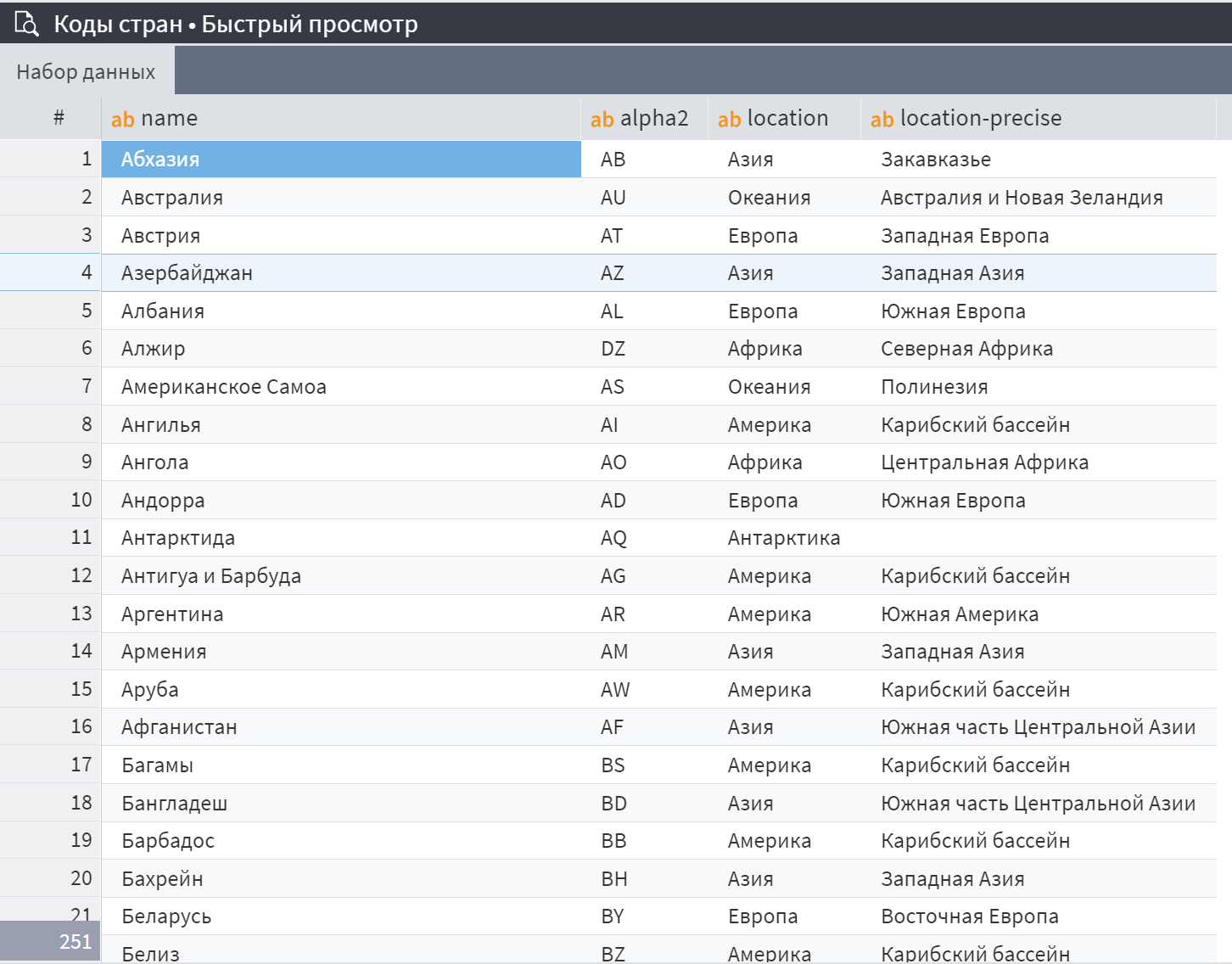
1. Перечень уникальных категорий товаров по ТН ВЭД, перевозимых через таможню в поле **code** с помощью компонента группировка (отсортируйте значения по возрастанию)



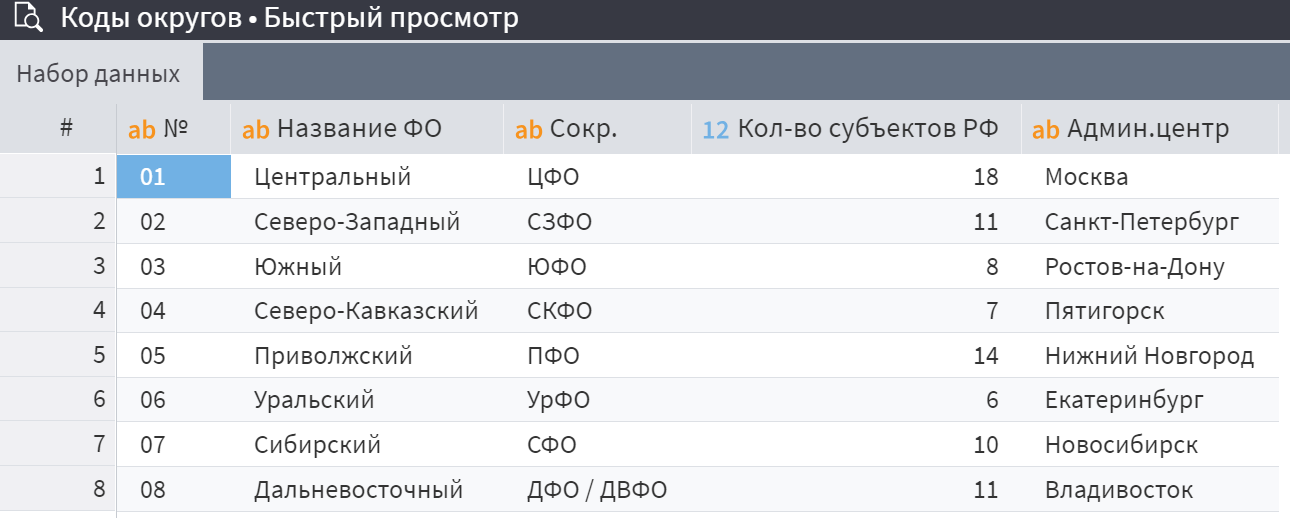
1. Перечень уникальных кодов федеральных округов по полю **district** с помощью компонента группировка (отсортируйте значения по возрастанию)



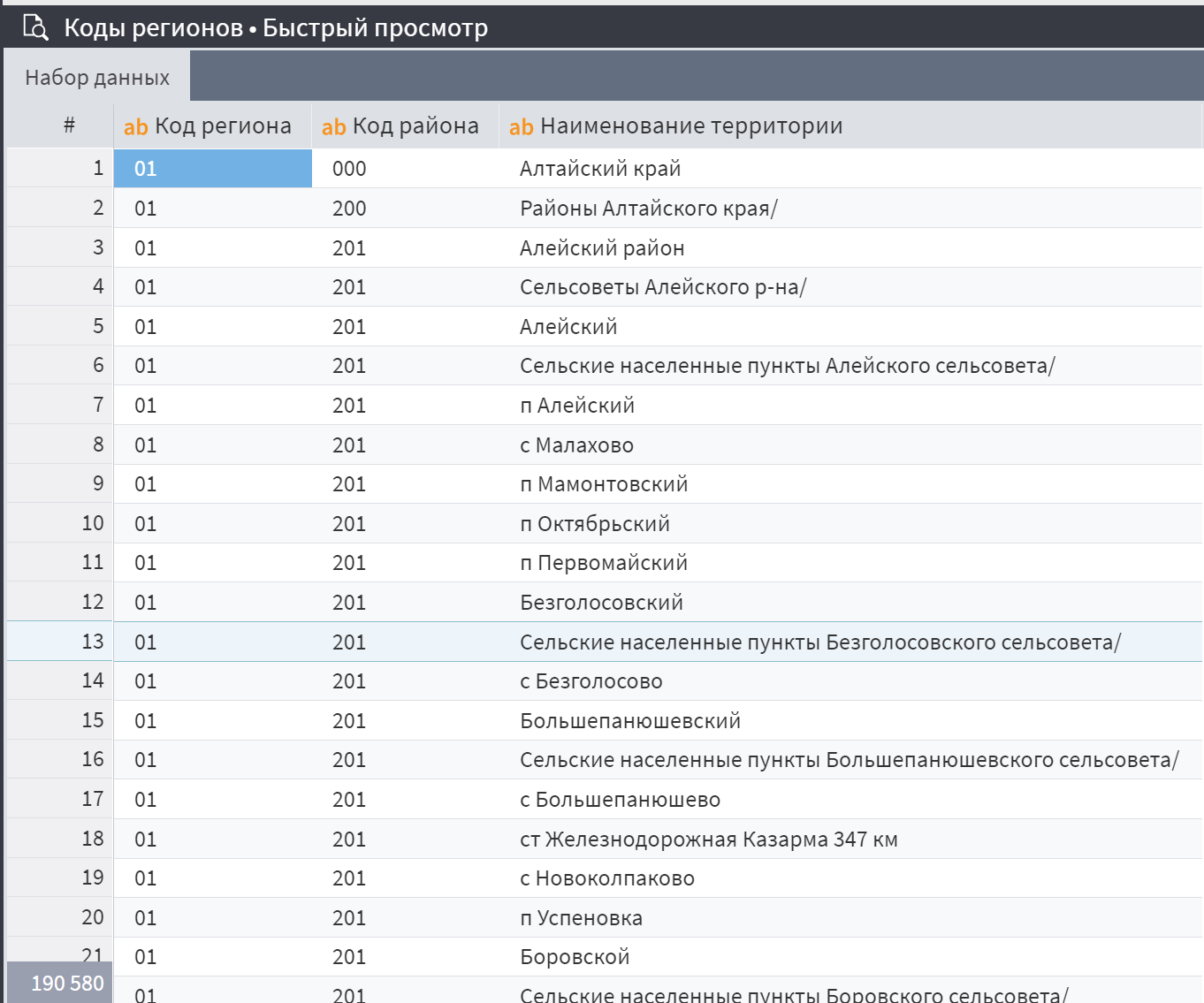
1. Импортируйте в сценарий следующие выборки данных:
2. Файл данных с кодами стран мира ОКСМ в формате alpha2 по пути: Data/Alpha2/alpha2\_codes.csv,
3. в кодировке **UTF-8**,
4. сразделителем столбцов **табуляцией**.
5. Исключите из полей импортируемой таблицы: **fullname, english, alpha3, iso**.



1. Файл данных с информацией о федеральных округах Data/ОКАТО/fed\_codes.tsv.txt,
2. в кодировке **ANSI**,
3. сразделителем столбцов **табуляцией**.
4. Измените тип поля **№** на строковый тип данных.
5. Исключите из рассмотрения все поля кроме полей с метками: **№, Название ФО, Сокр., Кол-во субъектов, Админ. Центр**.

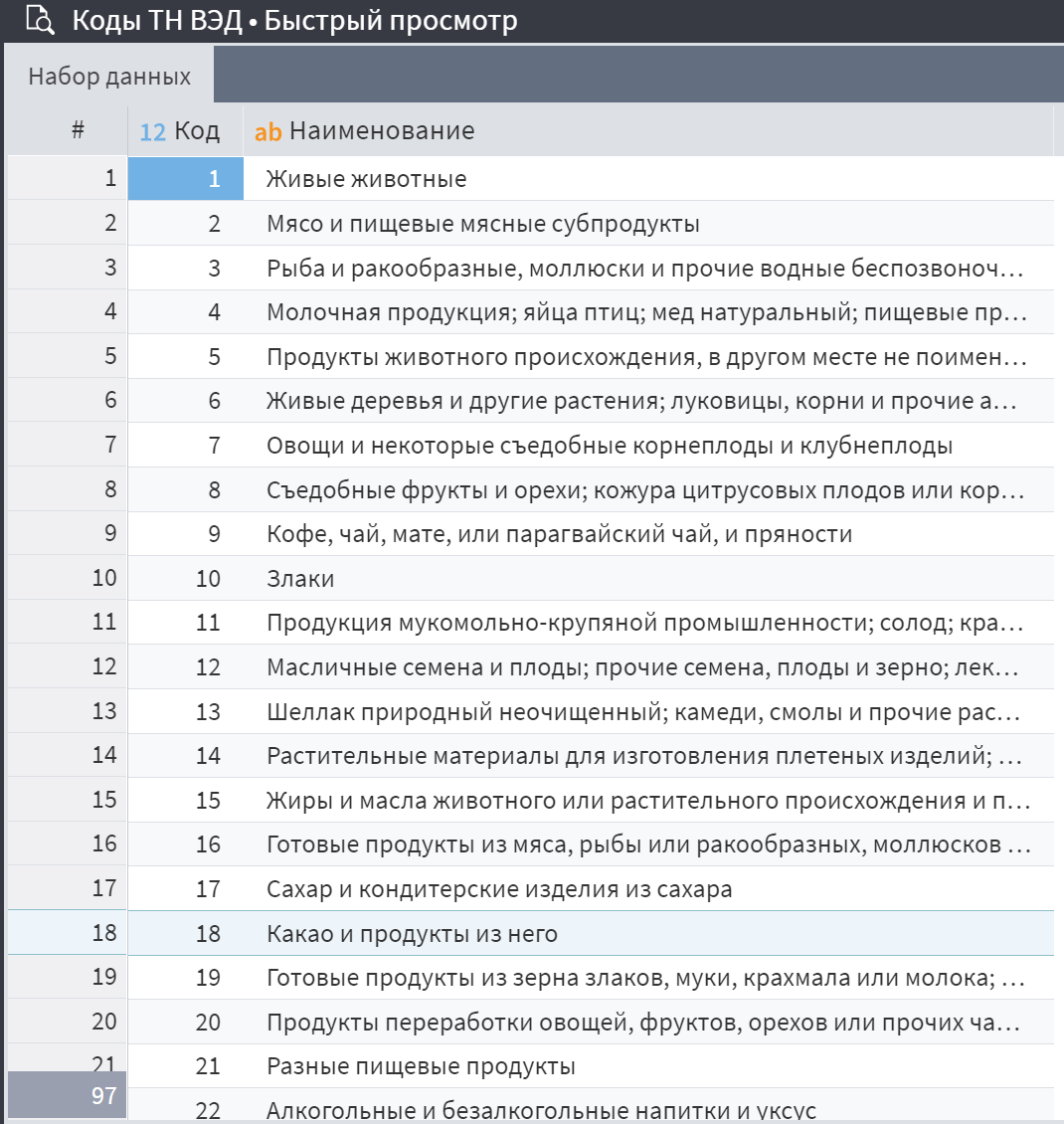


1. Импортируйте в сценарий следующие выборки данных:
2. Файл данных с информацией о регионах РФ в соответствии с ОКАТО Data/ОКАТО/okato\_regions.lgd,
3. Исключите из рассмотрения все поля кроме полей с метками: **Код региона, Код района, Наименование территории**.

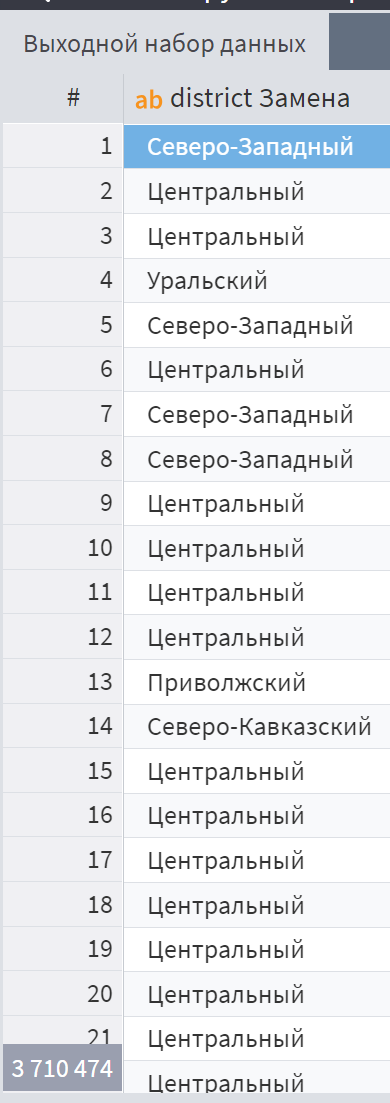


1. Импортируйте в сценарий следующие выборки данных:
2. Файл данных с информацией о разделах товарной номенклатуры ТН ВЭД Data/ТН ВЭД/tnved\_codes12.txt
3. в кодировке **ANSI**,
4. сразделителем столбцов **табуляцией**.
5. Оставим остальные настройки импорта данных по умолчанию.

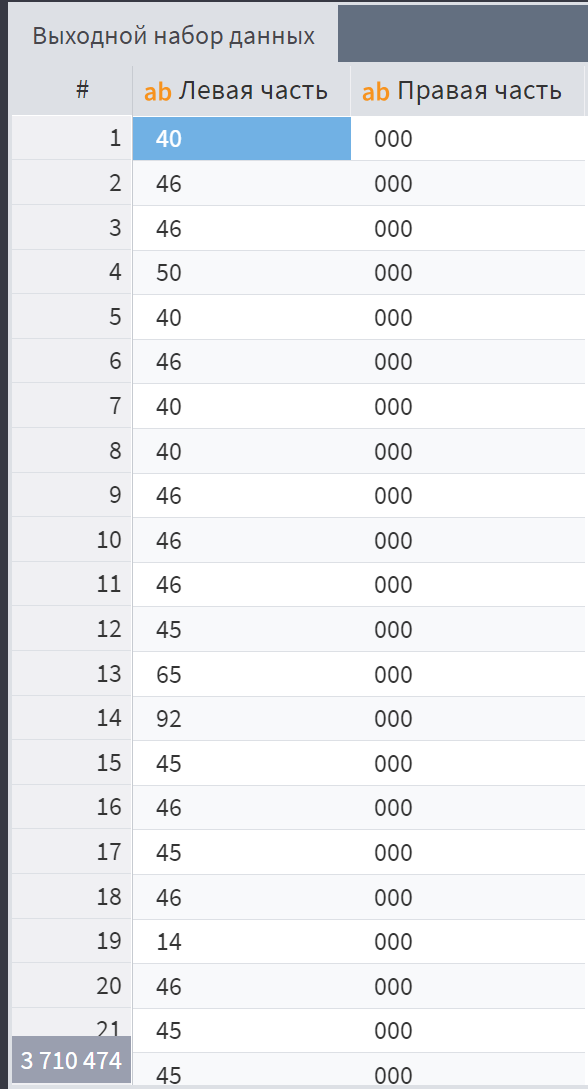
ТН ВЕД

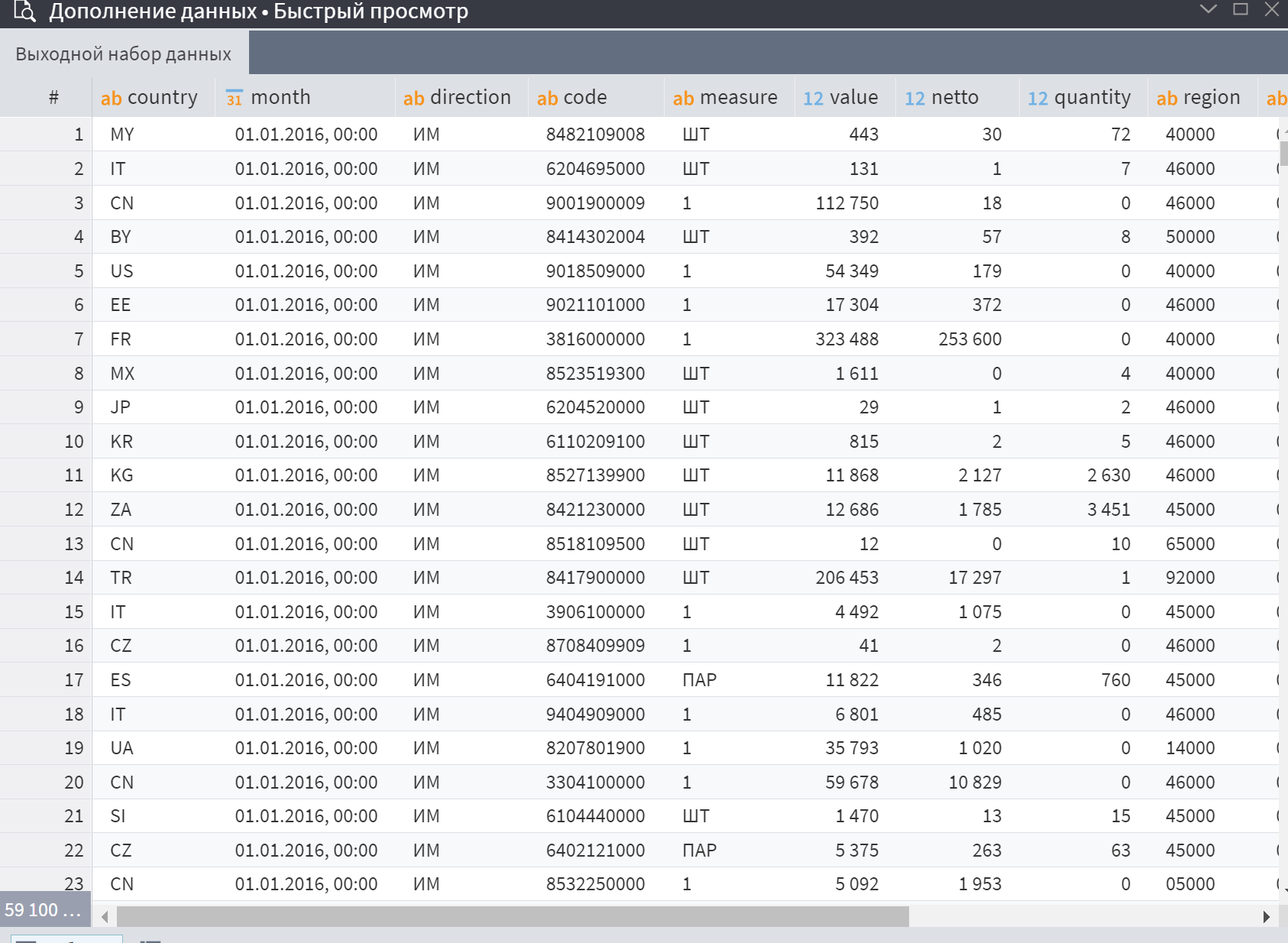


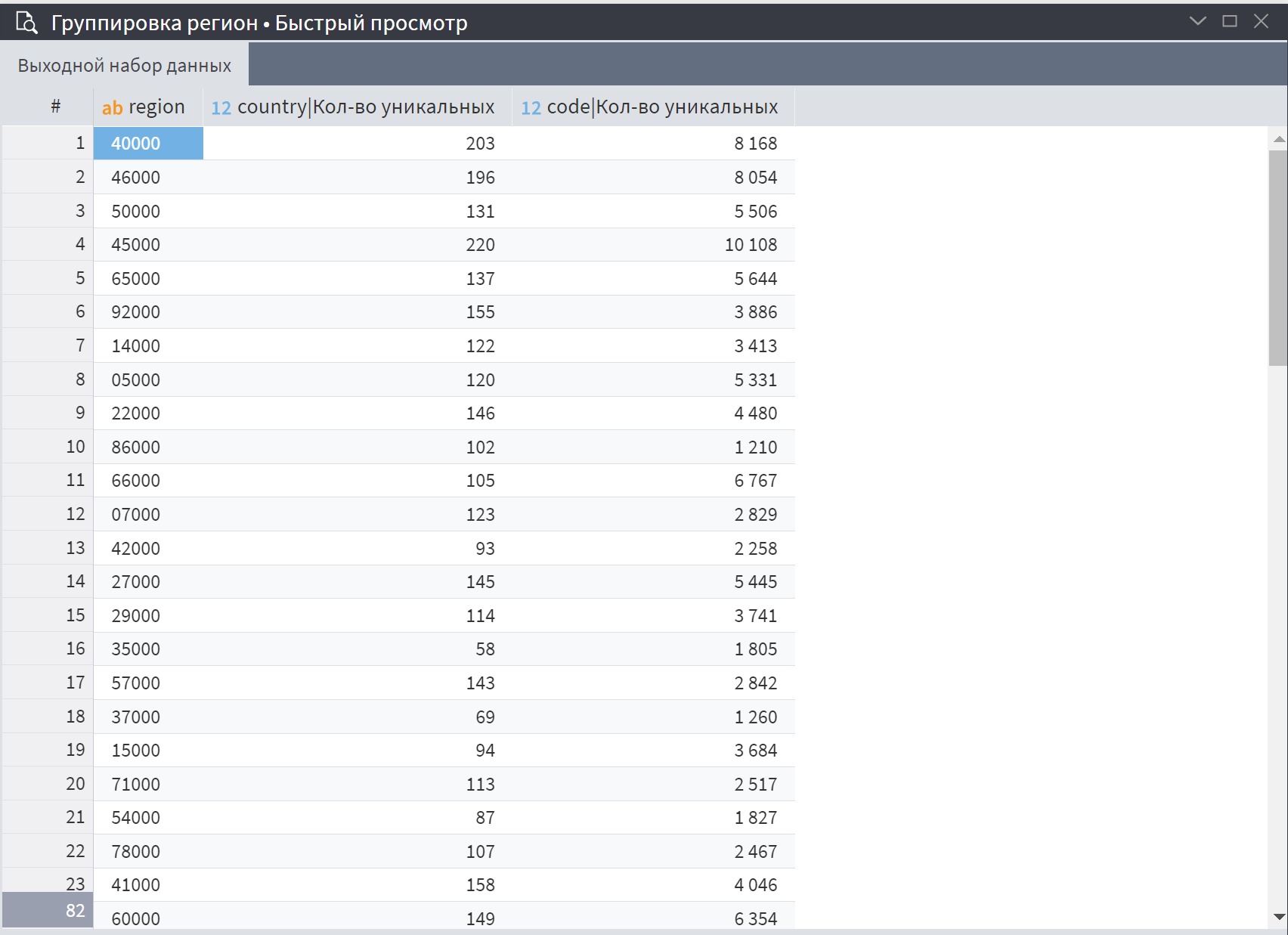
1. В датасете в поле **district** присутствует федеральный округ с кодом «09», что есть девятый федеральный округ «Крымский ФО», упраздненный в 2016 году
2. Замените в исходном наборе данных в поле с названием **district** значения «09» на «03» (Южный федеральный округ) с помощью компонента «Замена».

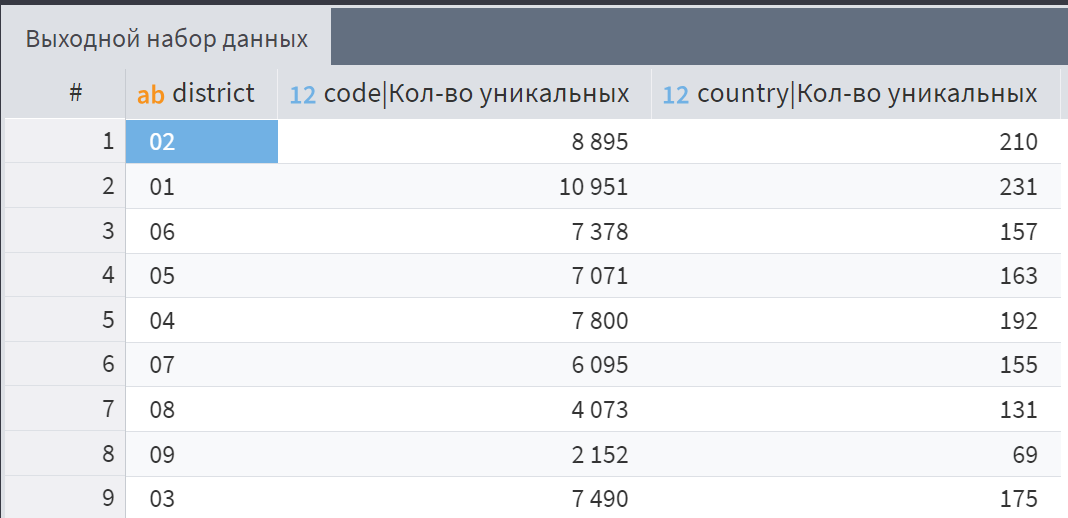


1. С помощью компонента «Параметры полей» приведите выборку после замены к исходной структуре (количество полей и их названия).
2. Получите из поля **region** *новые поля* с первыми двумя цифрами из **region** и последующими тремя цифрами из **region**, применив функции **Left()** и **Right()** из компонента калькулятор.



1. Дополните преобразованную выборку данных новыми полями из импортированных таблиц с помощью компонента **«Дополнение данных»**
2. **Исключите с помощью «Параметры полей» все промежуточные ключи** (district, region, country, новые поля из калькулятора)
3. 
4. Сформируйте серию группировок данных на основе компонента «Группировка»:
5. 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

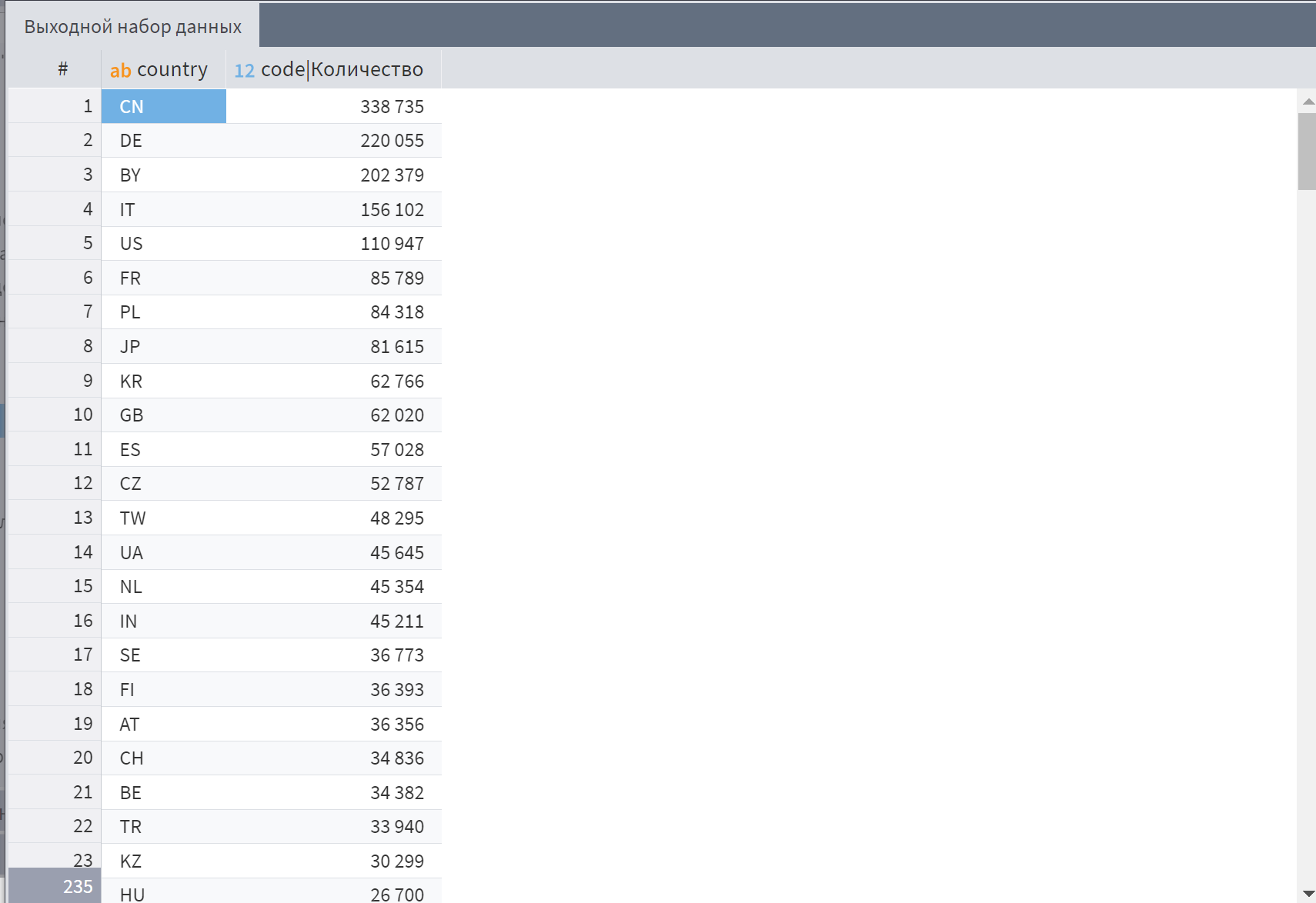


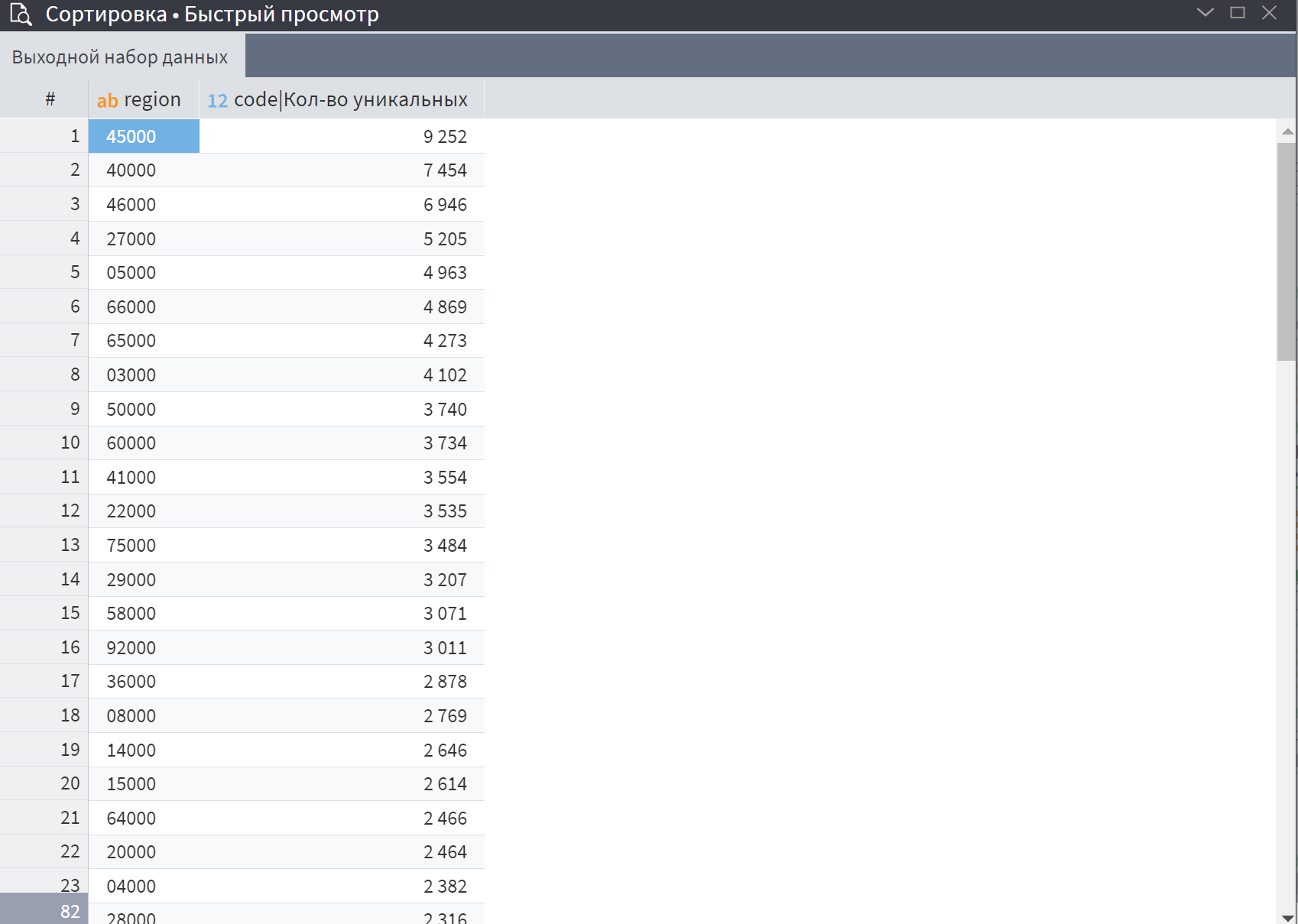


1. Найдите самое популярное направление экспорта за рубеж по количеству уникальных позиций экспортируемых товаров за 2016 год. Считаем общее число уникальных кодов по всем регионам России



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





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

