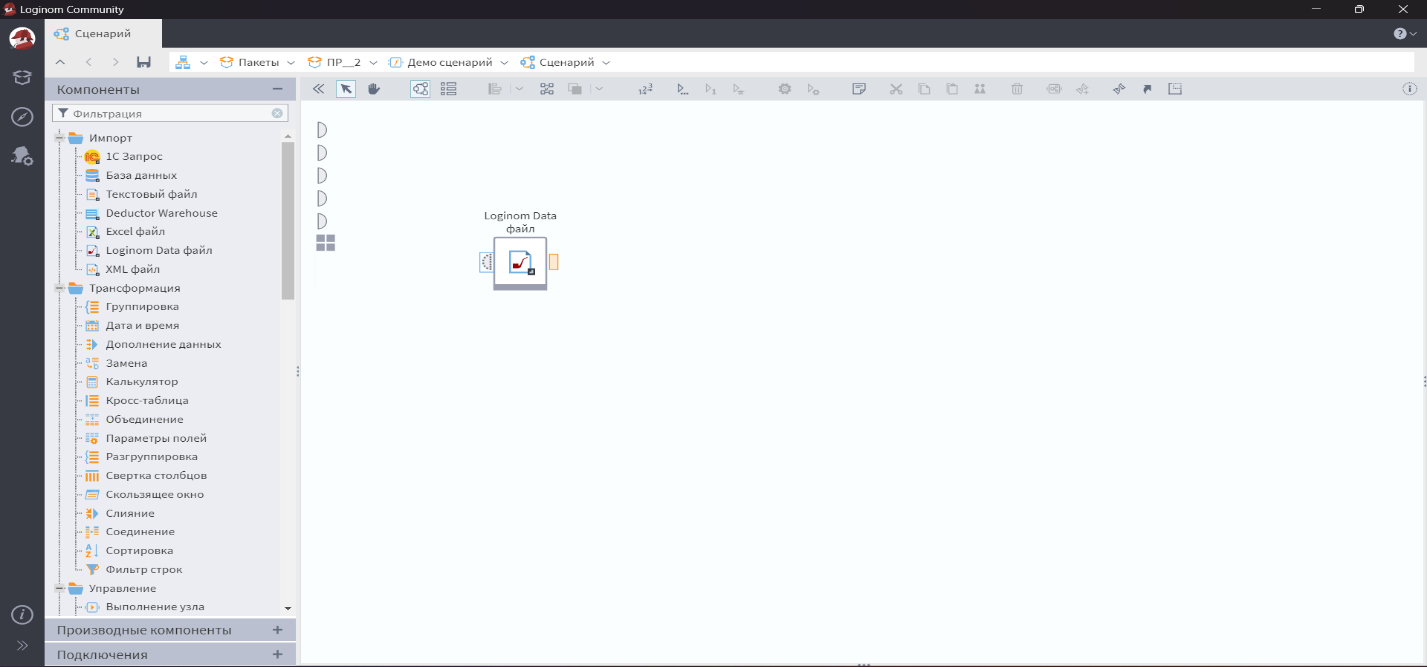
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | | МИНОБРНАУКИ РОССИИ | | | | Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА** – **Российский технологический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

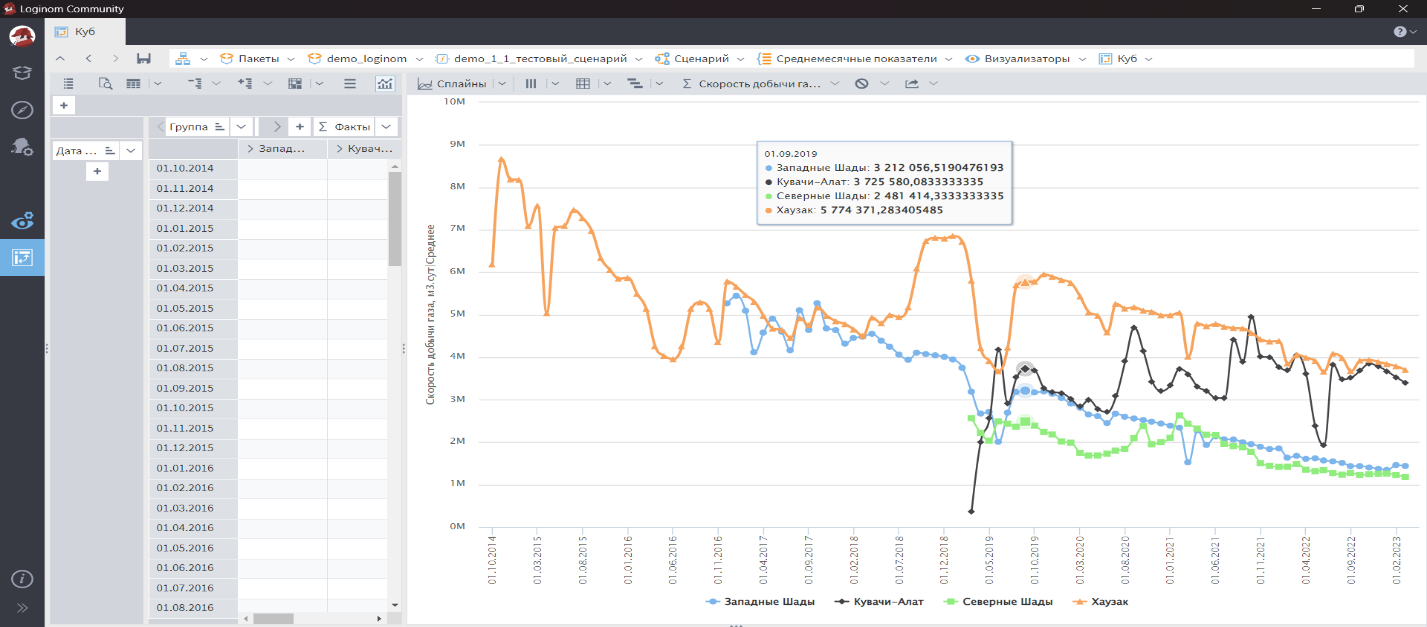
|  |  |
| --- | --- |
| **ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 2** | |
| **по дисциплине «Большие данные»** | |
|  | |
| Выполнил студент группы КМБО-06-22 | Масленников И. А. |
| Проверил ассистент кафедры ПМ ИИТ | ФИО Семинариста |

Москва 2024

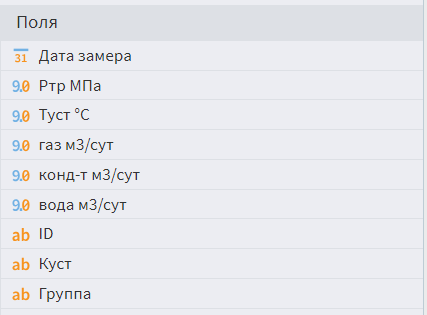
Задание №1

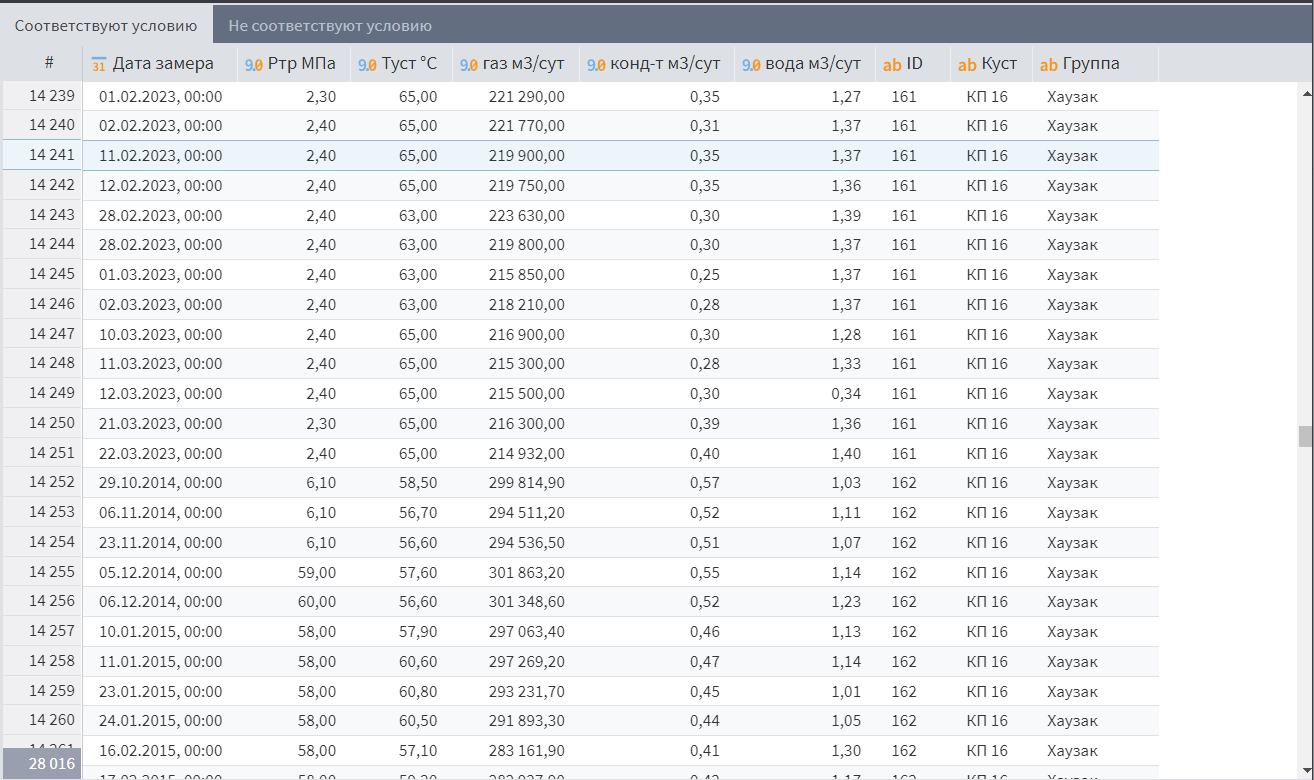
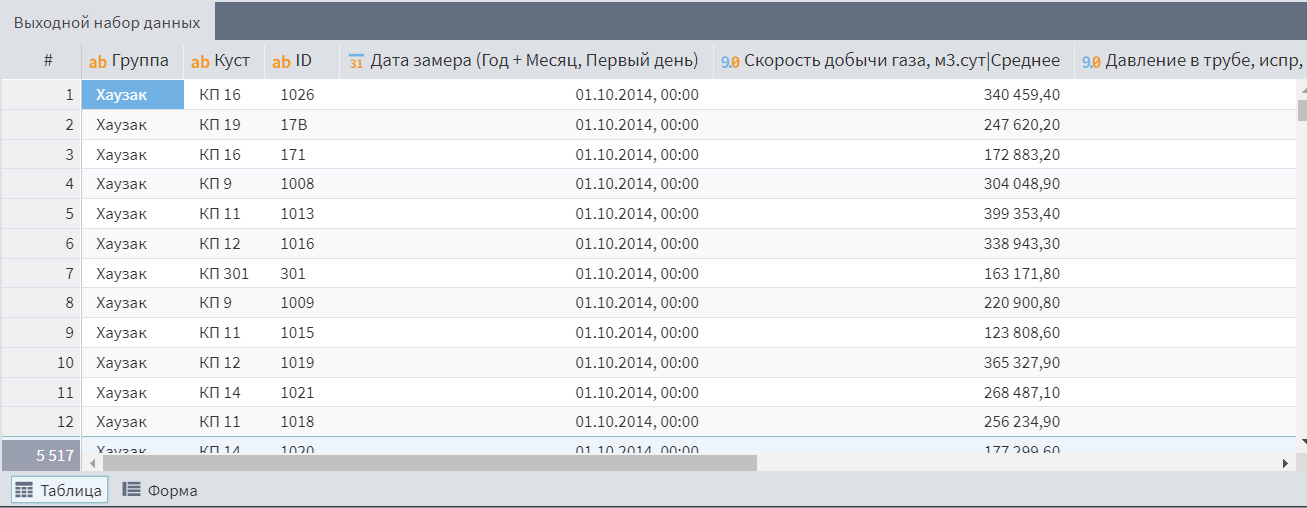
1. Создал пакет .lgp через Loginom
2. Сохранил под названием “ПР №2”
3. Создал модуль внутри пакета с названием «Демо сценарий» с помощь функции редактирования
4. Перенес на полотно компонент «Loginom Data файл»
5. Сохранил пакет.

Задание №2

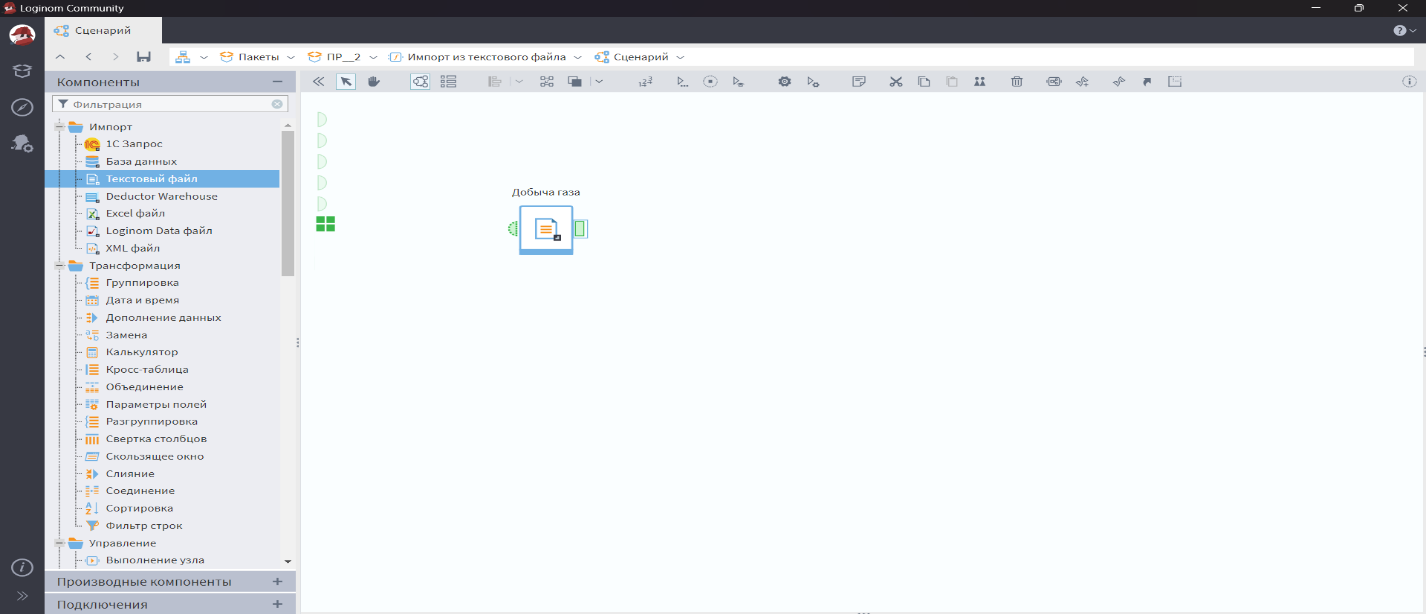
1. Открыл пакет demo\_loginom.lgd
2. Открыл модуль demo\_1\_1\_тестовый\_сценарий
3. Запустил все узлы кнопкой «выполнить все», сценарий выполняется без ошибок
4. Исследовал выходные и входные порты узлов на основе настройки входных портов и предварительного просмотра порта данных
5. Нажал на «глаз» под узлом среднемесячные показатели, просмотрел визуализацию

* На выходном порте таблицы данных «Импорт добычи газа» имеется 9 полей



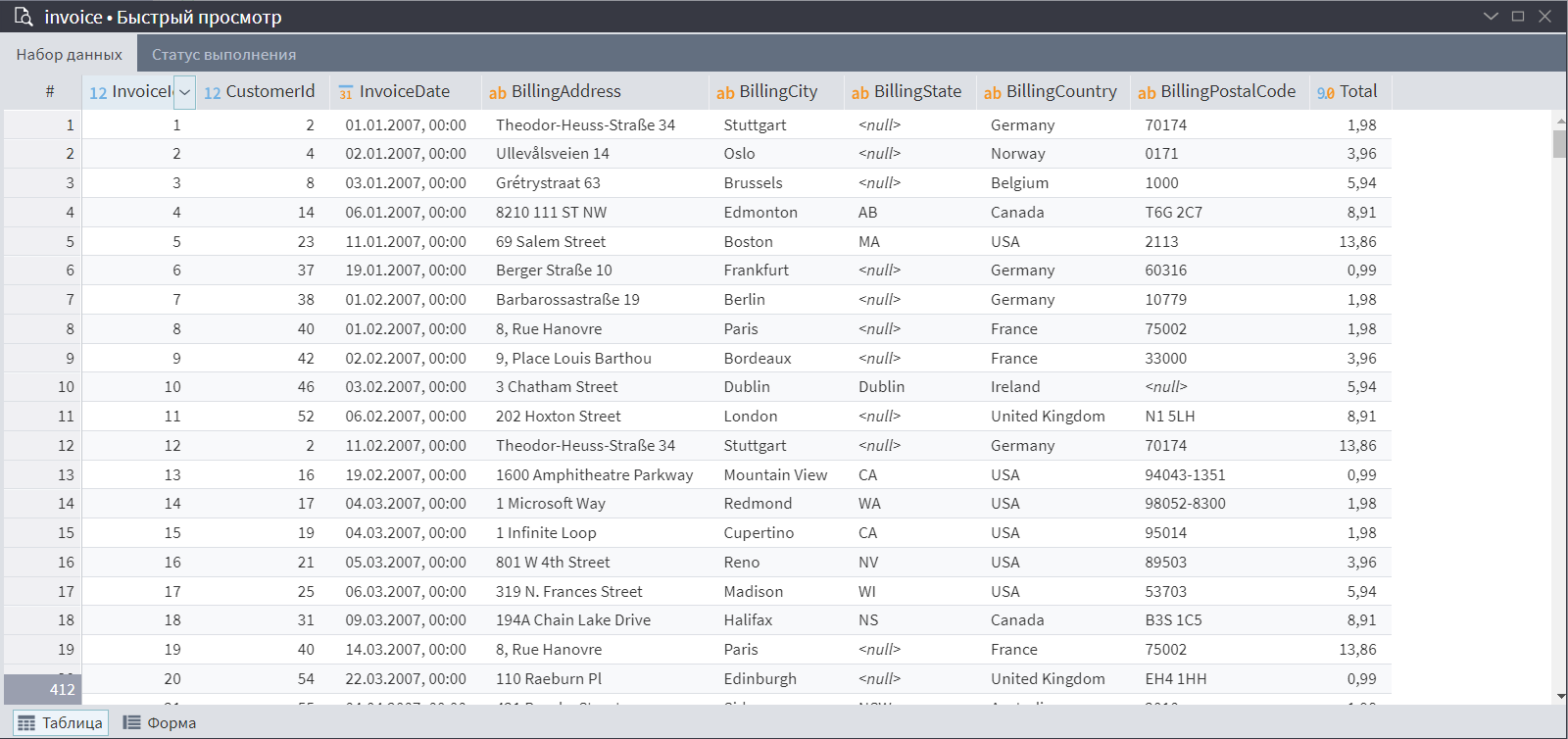
* В записях импортируемой таблицы данных имеются пропуски
* На верхнем порте узла «Полные записи» 28016 полных записей
* После выполнения узла «Начало месяца» появилось поле “Дата замера (Год + Месяц, Первый день)”
* В калькуляторе «Преобразование полей» добавляются поля:
  + Давление в трубе, испр, МПа
  + Температура устьевая, Цельсий
  + Скорость добычи газа, м3.сут
  + Температура устьевая, Кельвин
* После агрегации на выходе из узла «Среднемесячные показатели» остается 5517 записей

Задание №3

1. Открыл пакет с расширением «.lgp», созданный ране
2. Создал новый модуль и изменил его название на «Импорт из текстового файла»
3. Зашел в сценарий и перенес из панели компонентов на полотно сценария компонент «текстовый файл»
4. Настроил импорт данных из файла «data\\GAZ\\GAZ\_n.csv»:
5. Кодировка UTF-8;
6. Разделитель табуляция;
7. Пусто – null.
8. Убедился в корректности импорта данных воспользовавшись предпросмотром выходного порта

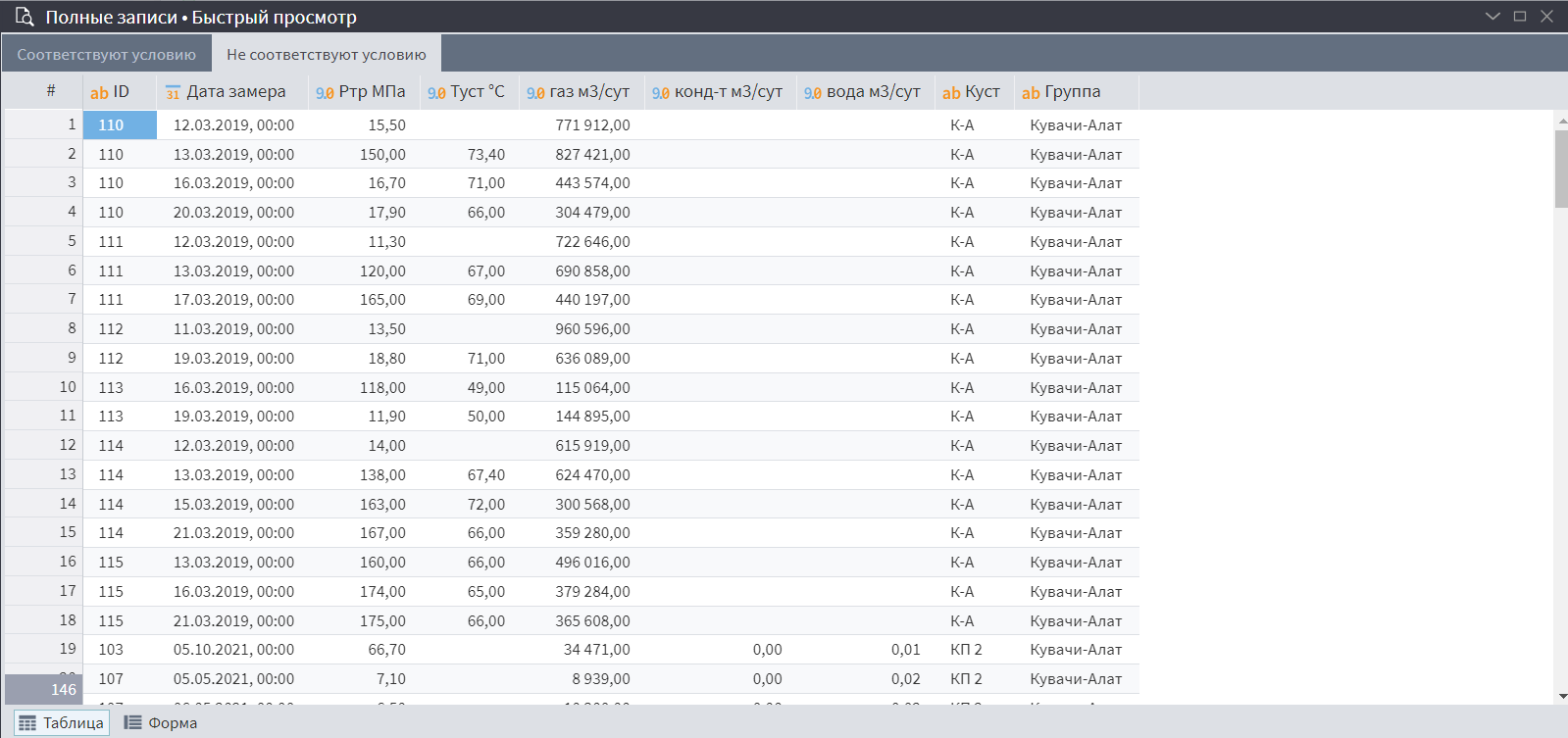
* В выходной таблице в результате выполнения узла получилось 9 полей
* Какие типы данных имеет каждое поле?
* Перешел в настройку форматов импорта, нашел поле ID и изменил его тип данных на строковый. Сохранил изменения. Выполнил узел. Вид данных в поле ID изменился на двоичный код
* В результате импорта данных из внешнего файла GAZ\_n.csv получилось 28162 строки
* Мы настроили чтение данных из файла .csv с разделением табуляцией с помощью раздела настроек форматов импорта в функции “Разделитель столбцов”

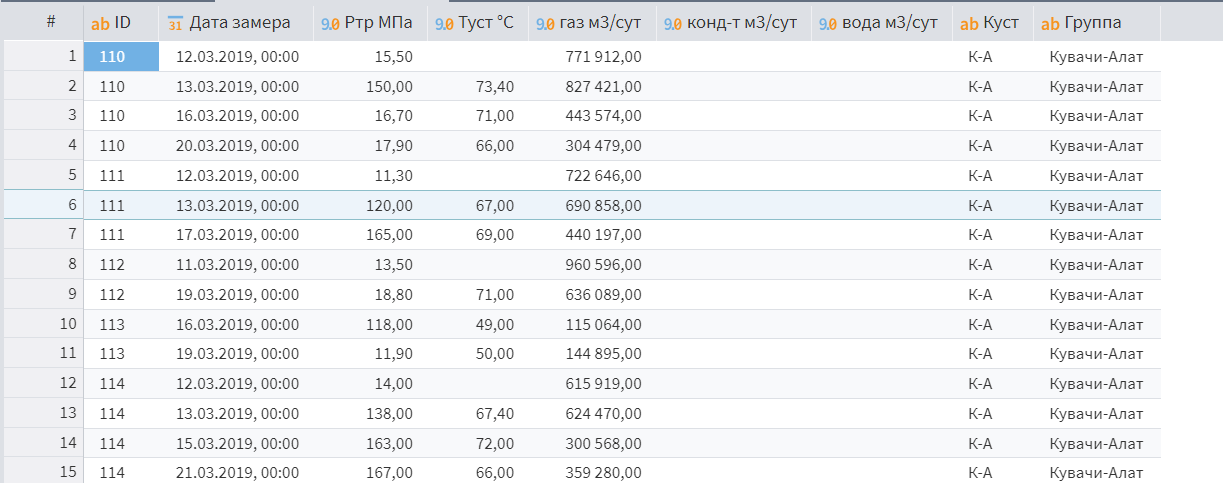
Задание №4

1. Открыл модуль demo\_2\_3\_импорт\_бд
2. Создал новый модуль для данного задания перешел в него
3. Внутри модуля перешел во вкладу «Подключения»
4. Из папки «Базы данных» перенес SQLite на полотно справа
5. Два раза нажал на новое созданное подключение для его настройки
6. В настройках указал в поле «строка подключения» путь к экземпляру базы данных Data/SQLite/Chinook.db
7. Проверил корректность соединения с SQLite и нажал на кнопку «Подключить»
8. Перешел во вкладку сценарий и снизу раскрыл «Подключения», где нашел свой пакет и данный модуль, где расположено мое новое соединение с базой данных
9. Нажал на него правой кнопкой мыши и выбрал опцию «Добавить узел подключения в сценарий»
10. Перенес на сценарий компонент «База данных»
11. Перетащил выходной порт подключения ко входному порту базы данных схожего вида и активировал соединение
12. Настроил базу данных на импорт таблицы Invoice. Проверил данные на выходном порте.

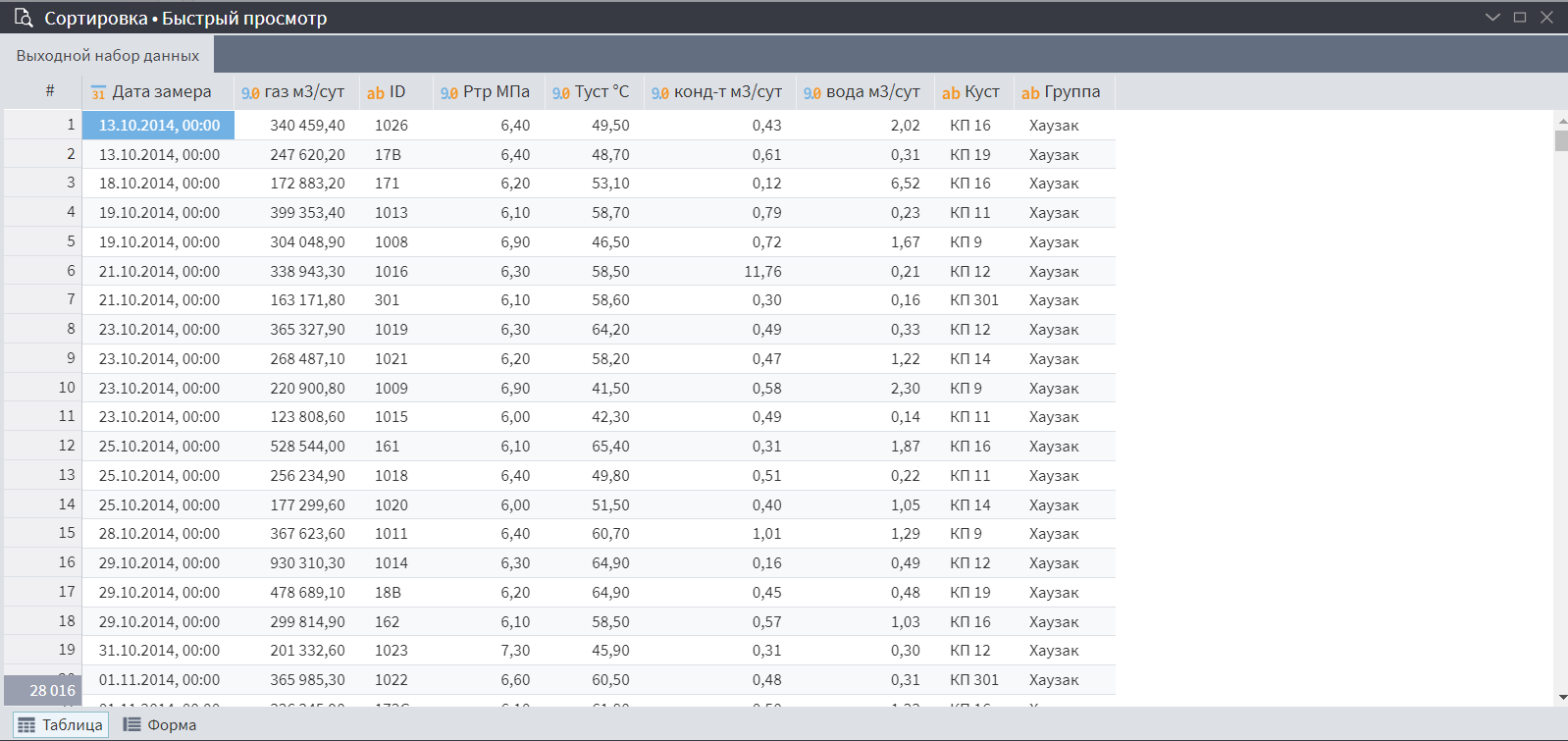
* Возможность быстрого просмотра информации для выходного порта узла подключения к базе данных существует. Открыть Быстрый просмотр можно:
  + кликнуть дважды левой кнопкой мыши по нужному активному порту
  + кликнуть правой кнопкой мыши по нужному порту и выбрать в контекстном меню Быстрый просмотр
  + При первом запуске Быстрый просмотр открывается в виде модального окна
* Количество записей находящихся в таблице Invoice равно 412
* Метка имеющая поле первичного ключа для таблицы Invoice это InvoiceID

Задание №5

1. Открыл модуль demo\_3\_1\_Фильтрация\_данных
2. Для импортированных данных из задания №3 провел фильтрацию импортированных данных на основе фильтрации записей с пустыми полями
3. Соединил выход порта узла импорта данных из текстового файла со входом в узел фильтр строк
4. Настроил узел фильтр строк для фильтрации данных
5. При настройке фильтрации данных нажал на «+» и выберал первое поле после «Номер строки» и задал для него условие «не пустой»
6. Нажал применить и повторил действие для всех остальных столбцов
7. Нажал далее и настроил название метки узла как «полные записи»
8. Для нижнего выходного порта с помощью быстрого просмотра выяснил, сколько в нем находится записей, а именно 146

* 28016 строк данных удовлетворяют условию полных записей
* 146 строк данных не удовлетворяют условию, т.е. содержат пропуски в данных
* 0.5% от общего количества строк занимают неполные записи
* В выборке присутствуют неполные записи, которые имеют пропуски в нескольких полях?

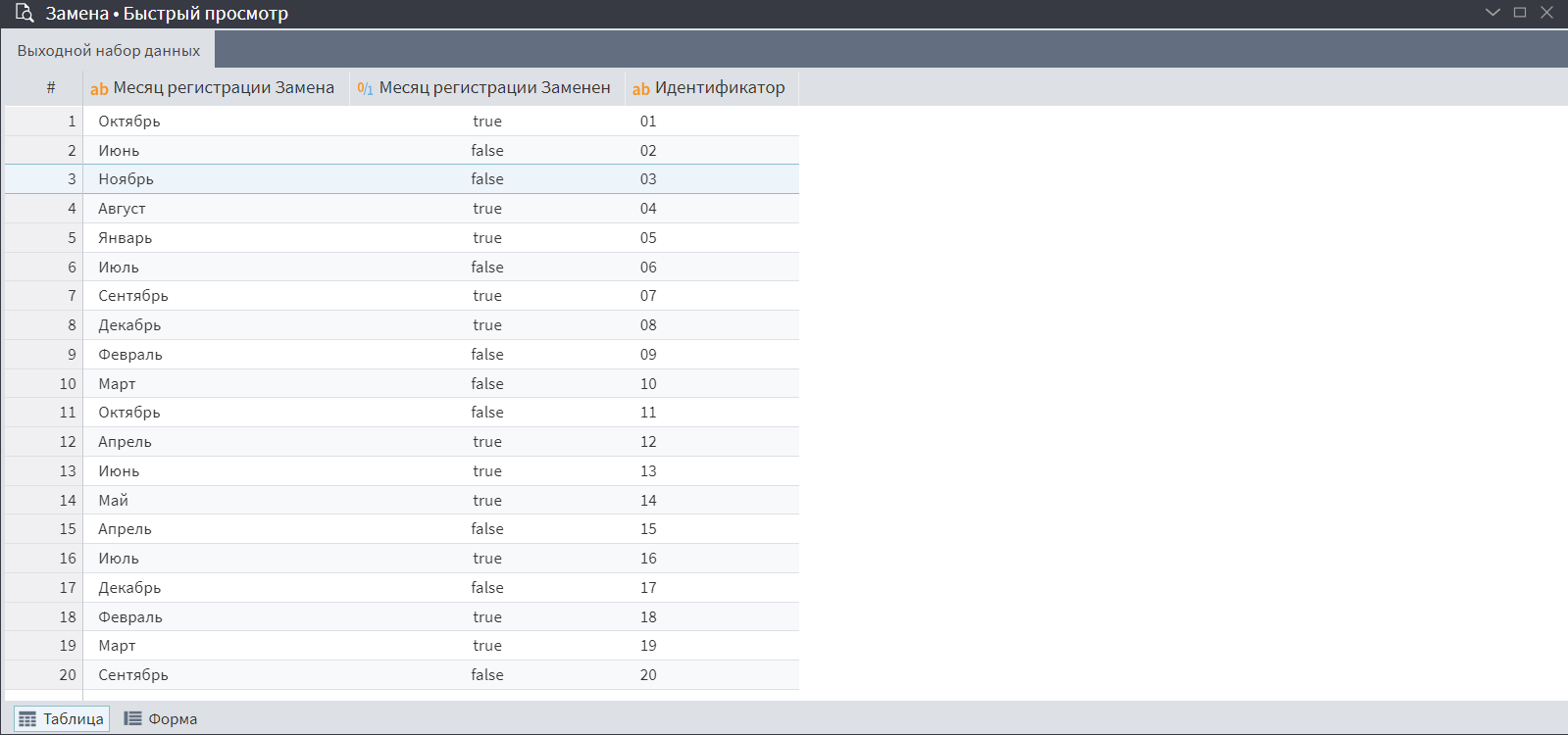
Задание №6

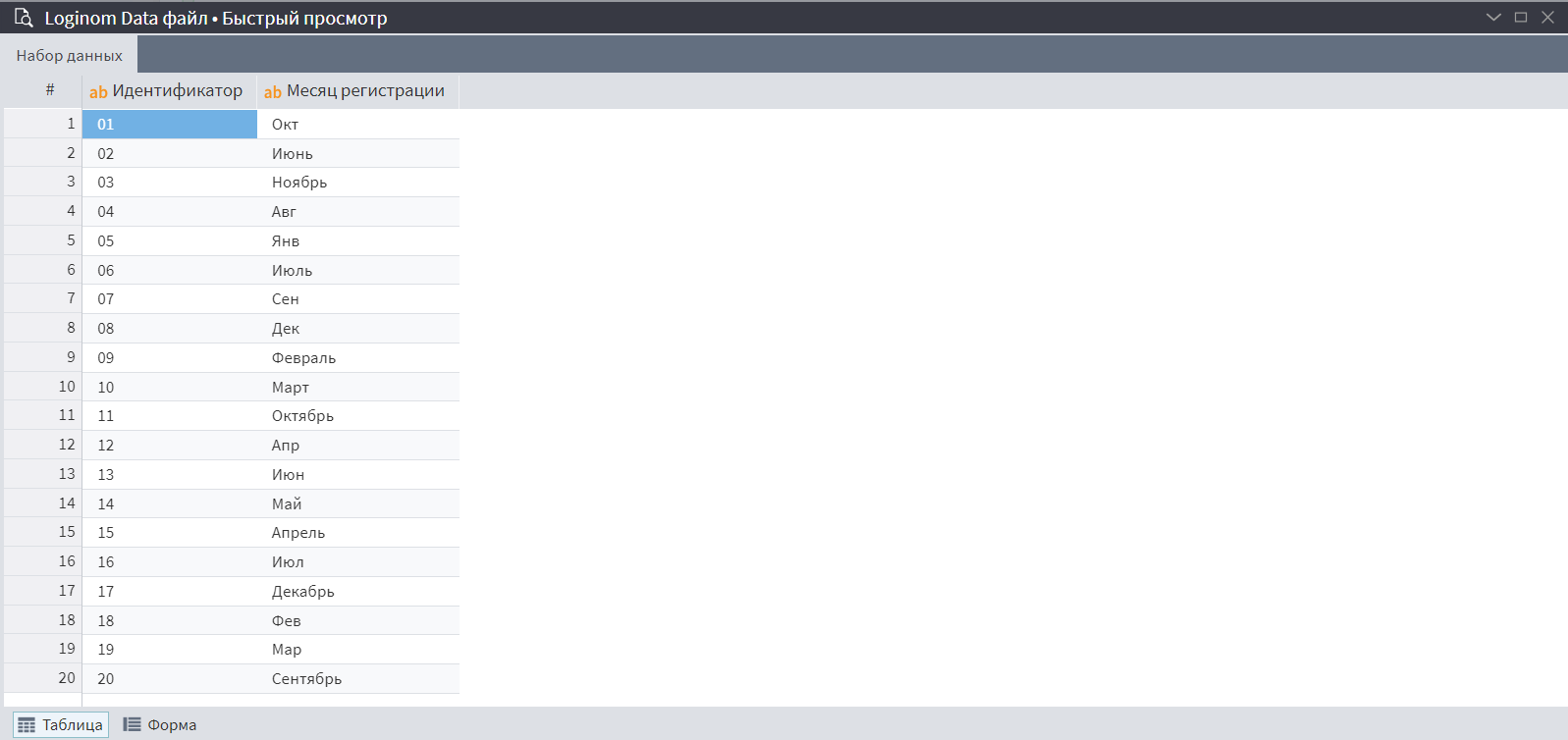
1. Открыл модуль demo\_3\_2\_Сортировка\_данных
2. Для данных, удовлетворяющих условию фильтрации из предыдущего задания выполнил сортировку записей по дате замера (по возрастанию) и скорости добычи газа (по убыванию). Перенеся в настройках сортировки поля «Дата замера» и «газ м3/сут» в поля сортировки
3. Сохранил настройки узла и выполнил его в сценарии
4. Убедился в корректности сортировки данных на основе быстрого просмотра выходного порта сортировки

* Дата замера “13.10.2014 00:00” оказывается самой ранней в таблице добычи газа
* ID 1026 соответствует самой ранней записи замеров скорости добычи
* Количество записей в выходной таблице по итогам сортировки таблицы по сравнению с количеством записей на входе не изменилось
* Дата замера “24.03.2023 00:00” является самой последней

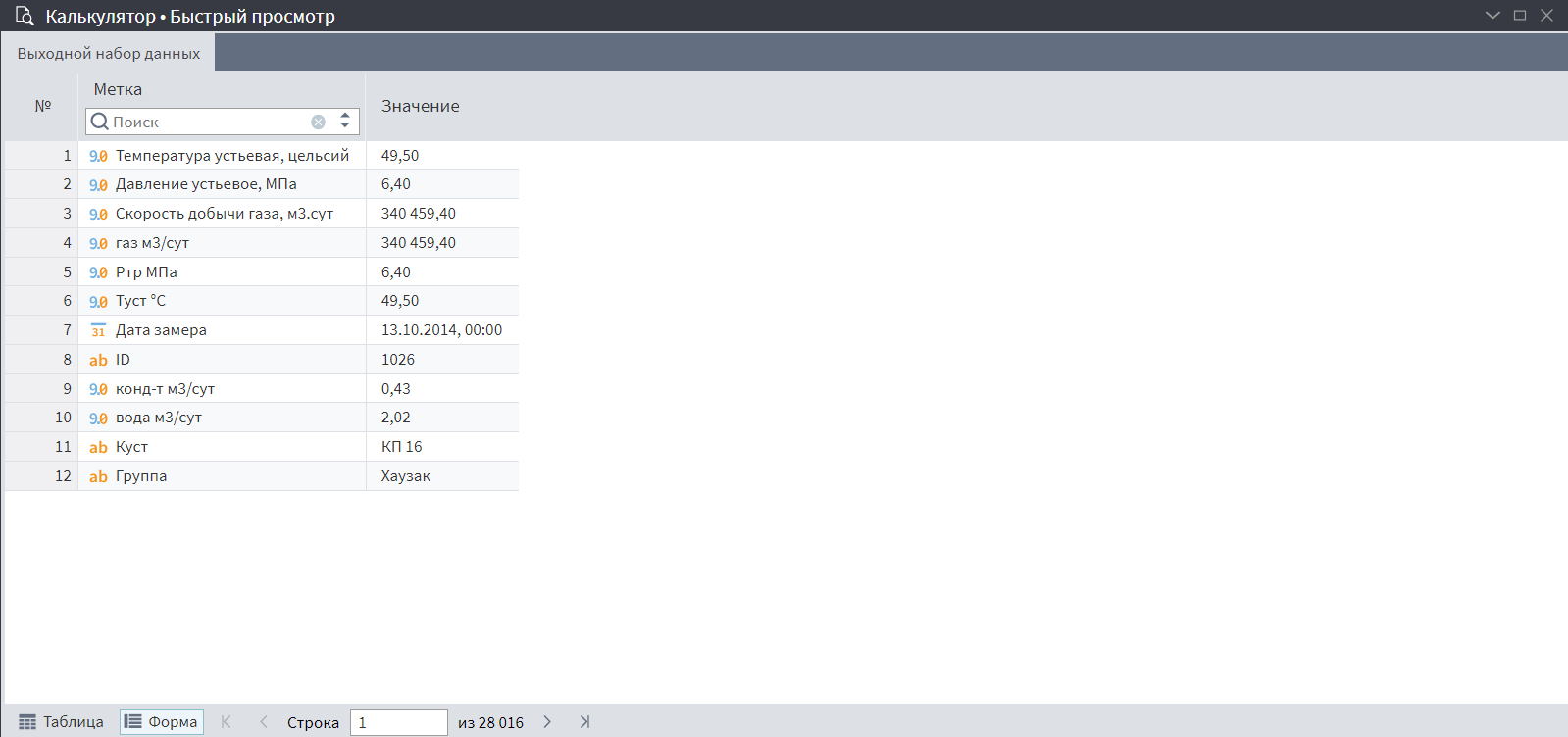
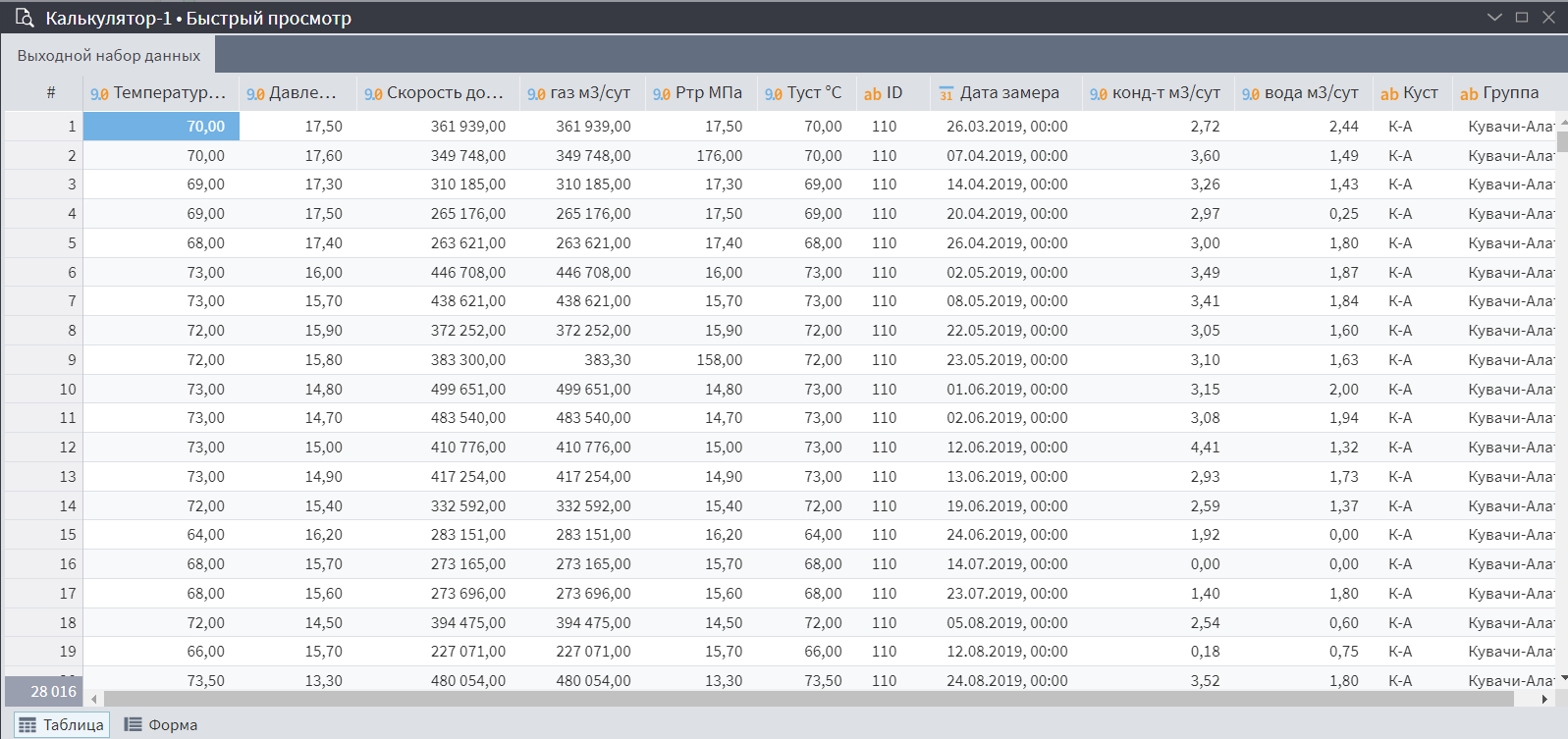
Задание №7

1. Открыл модуль demo\_3\_3\_замена
2. В поле построения сценария добавил 2 компонента импорта: Loginom data файл (Data/Demo/data\_months.lgd) и Текстовый файл (Data/Demo/months.txt).
3. Добавил в область построения сценария компонент Замена и подключил наборы данных к соответствующим портам узла и открыл настройку узла Замена и выбрал для поля Месяц регистрации способ Таблица замен
4. Назначением поля Краткое название установим Значение, поля Полное название – Замена.
5. Сохранил настройки. Выполнил узел. В результате получилось:

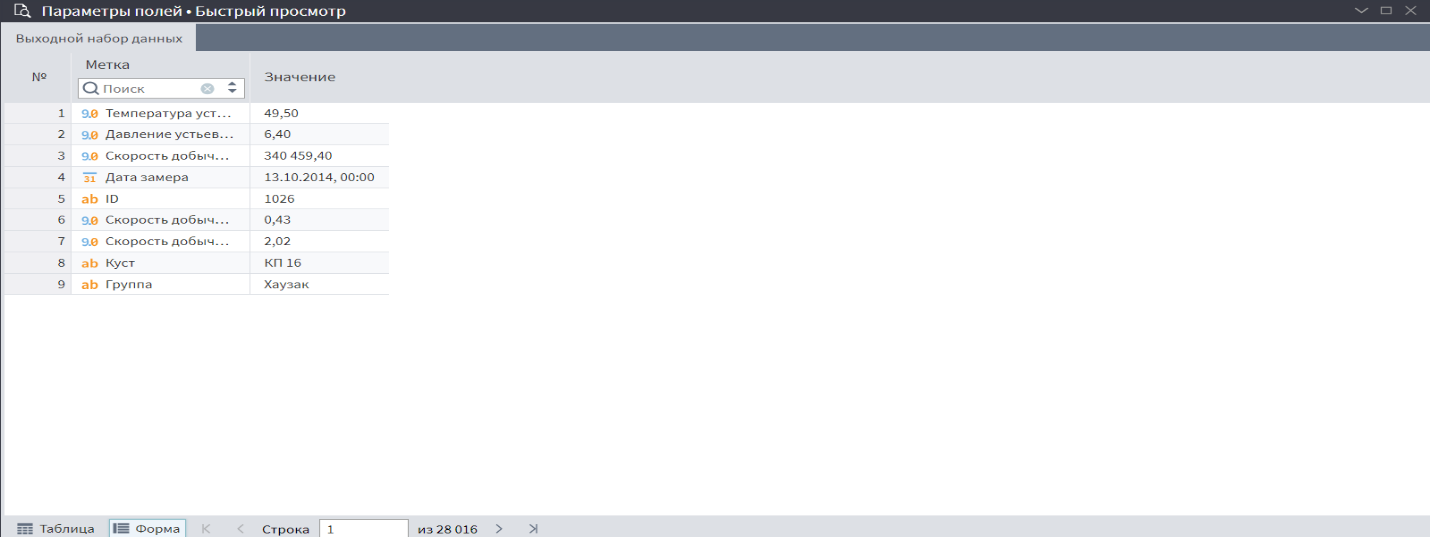
*  Не всем записям в выходной таблице данных соответствует значение true в поле Месяц регистрации заменен
* Во входной таблице данных были такие записи, заменять которые было не нужно

Задание №8

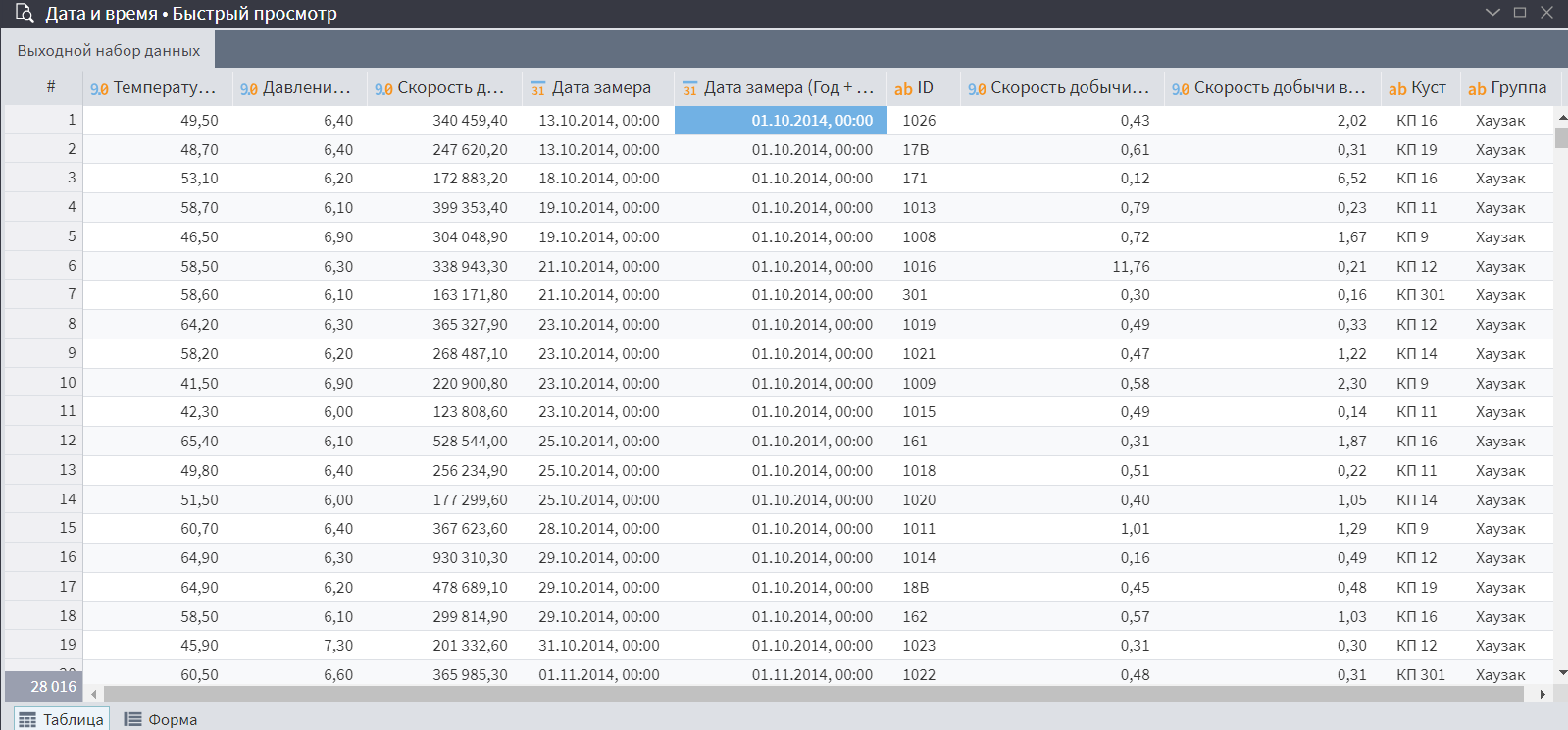
1. Открыл модуль demo\_3\_4\_калькулятор
2. Добавил в область построения сценария компонент калькулятор и соединил входной порт калькулятора с выходным портом сортировки из задания №6.
3. В калькуляторе в поле выражения с помощью кнопки добавить выражение добавил новое вычисляемое поле с именем Temp\_tube, и меткой Температура устьевая, цельсий и записал 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
4. Также создадал поля Pressure\_tube и Volume\_tube с соответствующими выражениями для их вычисления
5. С помощью быстрого просмотра выходного порта просмотрел какие поля получились в итоге

* 3 новых поля появилось в выходной таблице данных после вычисления выражений в узле калькулятор
* Скопировал настроенный узел калькулятора, вставил его в сценарий и подал на его вход данные из верхнего выходного порта фильтра строк. Выполнил данную копию узла.
* Просмотрел выходную таблицу данных у копии калькулятора.
* Сравнив поля: Скорость добычи газа, м3.сут и газ м3/сут. В 9 строчке произошли изменения, т.к газ м3/сут не соответствовал условию >1000, то произошло умножение на 1000.
* Также произошли другие изменения в вычисленных полях в сравнении с их базовыми полями согласно условию калькулятора.

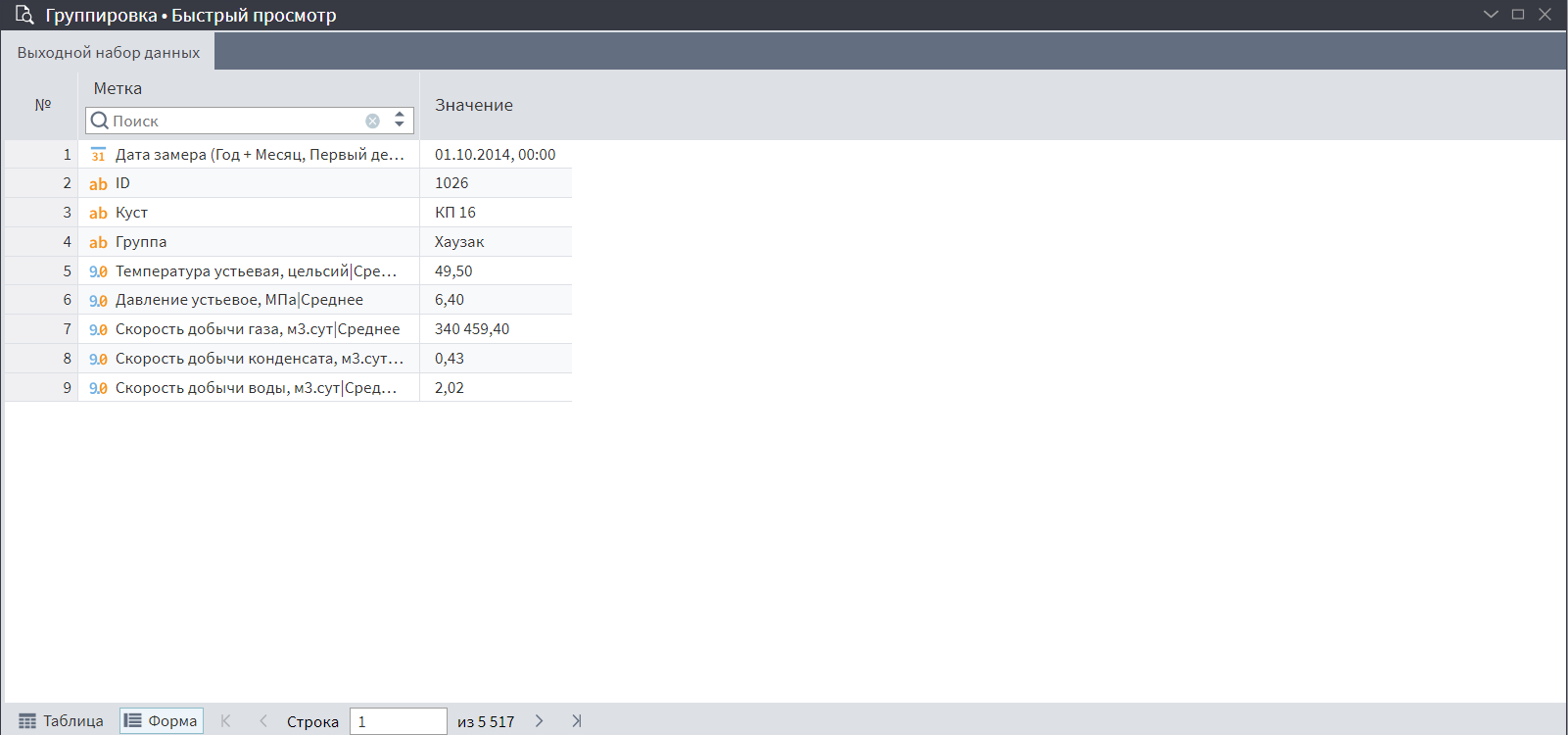
Задание №9

1. Открыл модуль demo\_3\_5\_параметры\_полей
2. Для предыдущего задания добавил в сценарий обработки данных компонент параметры полей
3. Подал выходной порт калькулятора на входной порт «Параметры полей»
4. Настроил 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
5. Выполнил узел и быстрый просмотр выходного порта из узла «Параметры полей»

Задание №10

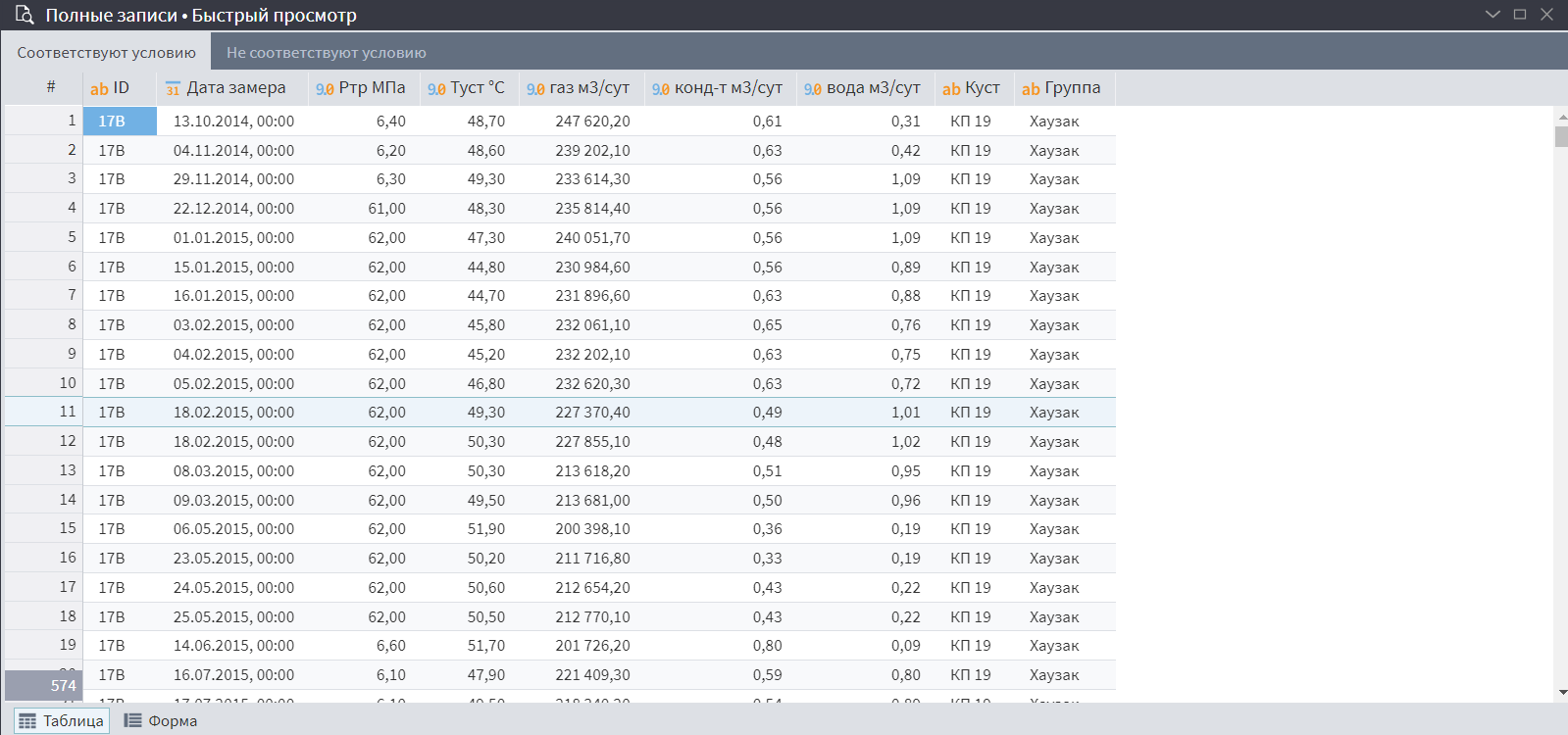
1. Открыл модуль demo\_3\_6\_дата\_время
2. Для созданного сценария в задании №9 добавил компонент даты и времени
3. Добавим в сценарий модуля обработки данных компонент «Дата и время» и соединил выходной порт «Параметров полей» с входным портом «Дата и время»
4. Для поля «Дата замера» выбрал разбиение по дате начала Год + месяц
5. На выходном порте узла активировал быстрый просмотр, чтобы убедиться в результате работы компонента

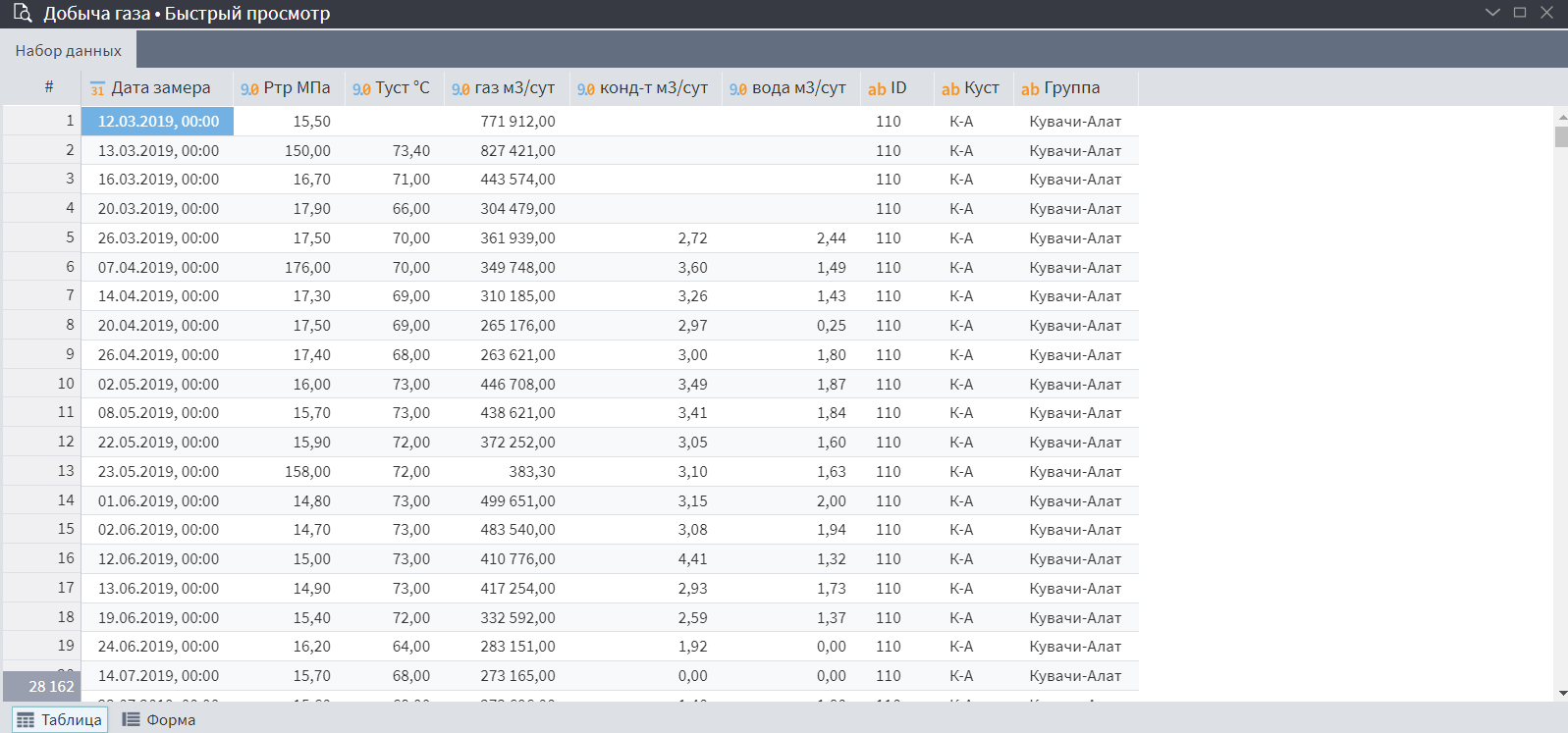
Задание №11

1. Открыл модуль demo\_3\_7\_группировка\_данных
2. Добавил в сценарий обработки данных расчёт среднемесячных показателей
3. Добавил в сценарий из задания №10 компонент «Группировка» и соединил выходной порт «Дата и время» с входным портом «Группировка»
4. В настройке группировки выбрал следующие поля в качестве групп:
   1. Дата замера (Год + месяц)
   2. ID
   3. Куст
   4. Группа
5. В настройке группировки выбрал все числовые поля в качестве показателей и рассчитал для них среднее

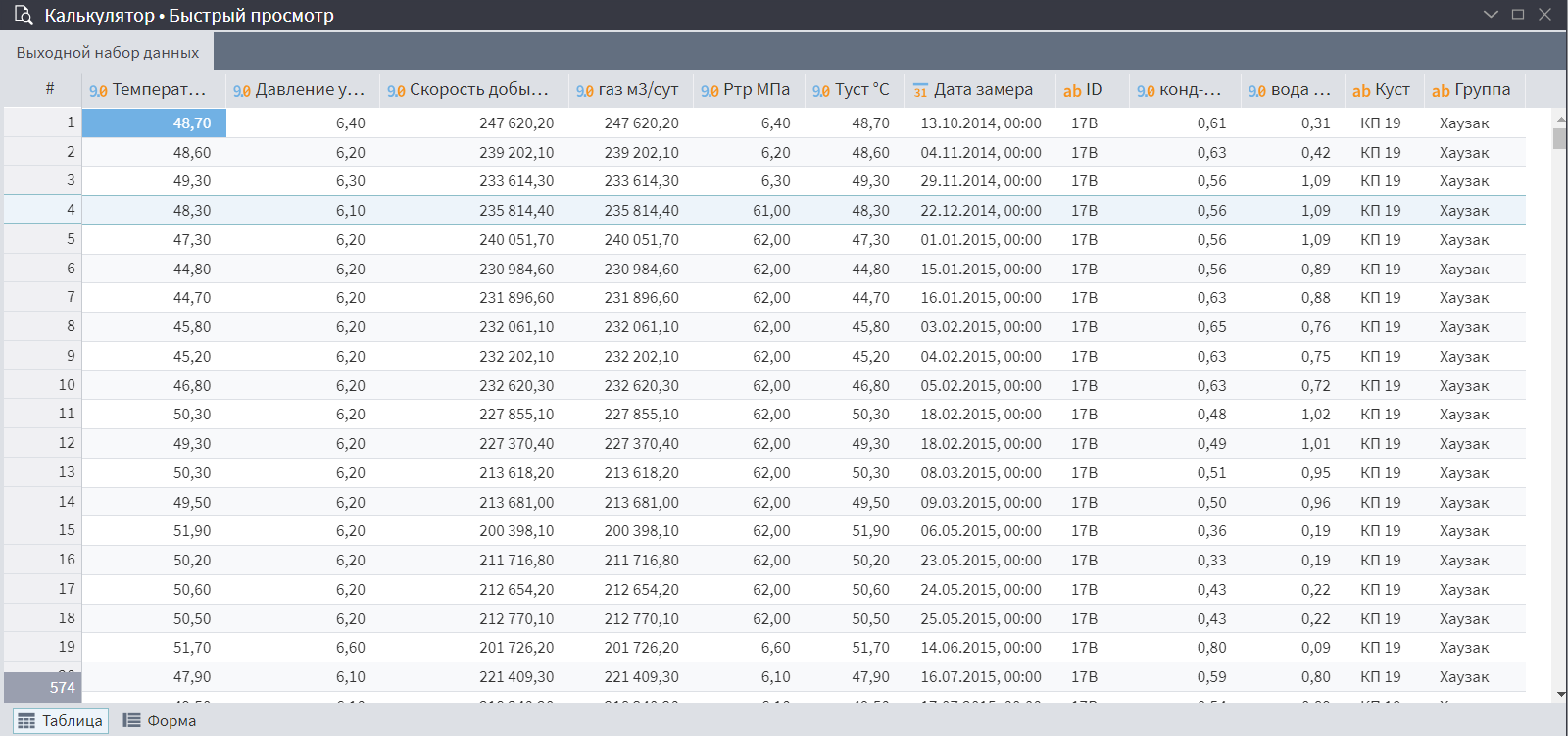
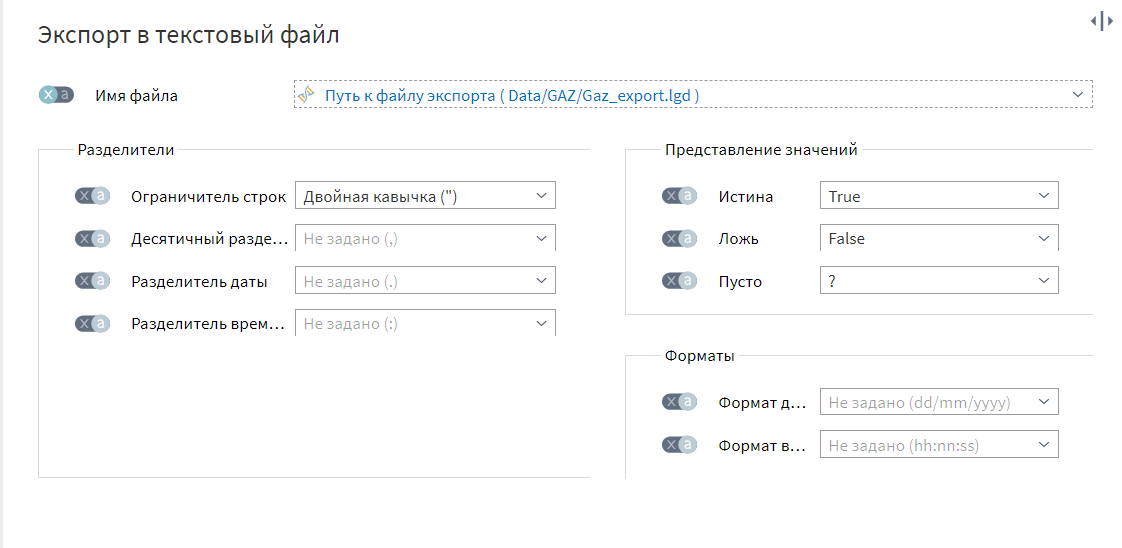
* 5517 записей мы получили в результате операции группировки данных
* 10 месяц и 2014 год были первыми в списке после группировки?
* Скважины по ID: 1026, 17B, 171, 1013 ,1008, 1016, 301, 1019, 1021, 1009, 1015, 161, 1018, 1020, 1011, 1014, 18B, 162, 1023 наблюдались(19 скважин) в первый(10) месяц в выходной выборке данных
* Какова максимальная скорость добычи конденсата в первый месяц наблюдения 11,76 м3.сут

Задание №12

1. Создал модуль “Задание №12”
2. С помощью ПКМ по узлу и выбора опции меню «Показать порт управляющих переменных» для фильтра строк создал у него порт переменных, который сможет принимать входные параметры
3. Создал новую переменную как показано на слайде и отключил автосинхронизацию
4. Настроил узел «Фильтр строк» добавив в перечень условий ещё одно условие по совпадению ID скважины со значением из внешней переменной(порт был настроен корректно)
5. На выходном порте данных, соответствующих условию наблюдаем ожидаемую функциональность
6. Далее добавил в сценарий пользовательские переменные, показанные на слайде, которые можно будет использовать для эффективной параметризации сценария
7. Соединил порт переменных пользователя с входной переменной добычи газа и настройках порта удалил все переменные, полученные от порта переменных пользователя оставив только переменную строкового типа path\_import
8. В настройках узла импорта данных включил передачу значения для Имя файла из внешней переменной, остальные настройки импорта оставляем без изменений
9. Убедился в том, что данные успешно импортированы в сценарий обработки данных



1. Передал переменные пользователя на вход фильтра строк, указал в качестве входного значения для настроенной входной переменной поле Скважина(id)

* Добавил 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
* Добавил узел экспорта данных, имя файла для записи в котором будет определяться внешней переменой пользователя

Самостоятельная работа

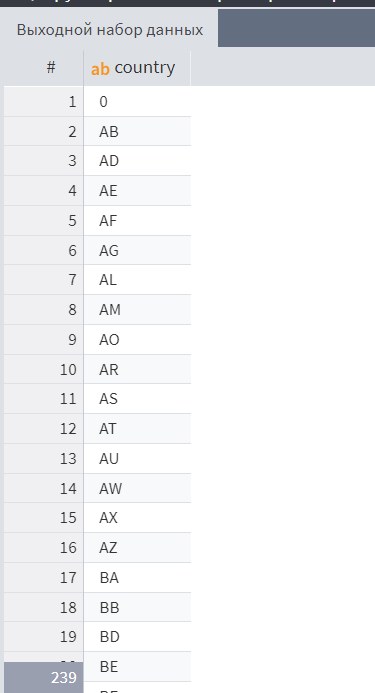
Работа на выполнение №1

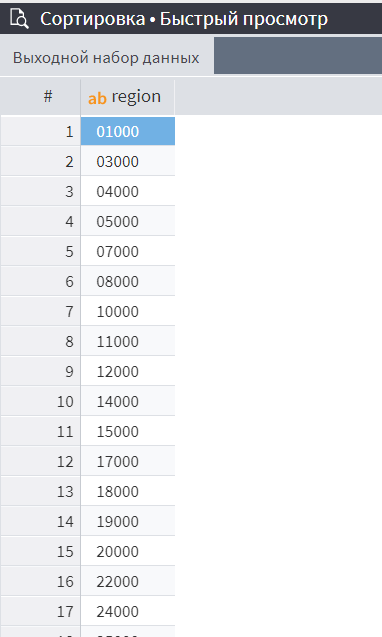
* 3 710 474 строк имеет выборка данных за 2016 год
* 10 столбцов имеет данная выборка
* Информация о каждом поле в отдельности:
  + Direction - Направление поставок: экспорт (ЭК) или импорт (ИМ)
  + Month - Месяц и год осуществления поставки
  + Country - Кодовое обозначение страны участника торговли в формате Alpha2р
  + Code - Кодовое обозначение товара в соответствии с 10-значным классификатором ТН ВЭД
  + Measure - Краткое обозначение дополнительных единиц измерения товара
  + Value - Суммарный объём товаров в поставках по таможенным декларациям в долларах США
  + Netto - Суммарный вес нетто товаров в поставках по таможенным декларациям в килограммах
  + Quantity - Суммарное количество товаров по таможенным декларациям в соответствии с дополнительными единицами измерения
  + Region – Регион
  + District - Федеральный округ
* Данные необходимые, чтобы расшифровать значения в полях country, code, region, district это фактические наименования для всех уникальных значений в полях соответственно
* В данных отсутствуют пропуски
* 239 уникальных стран участников торговли перечислено в поле country
* 82 уникальных регионов России участвовало в торговле (поле region)

Работа на выполнение №2

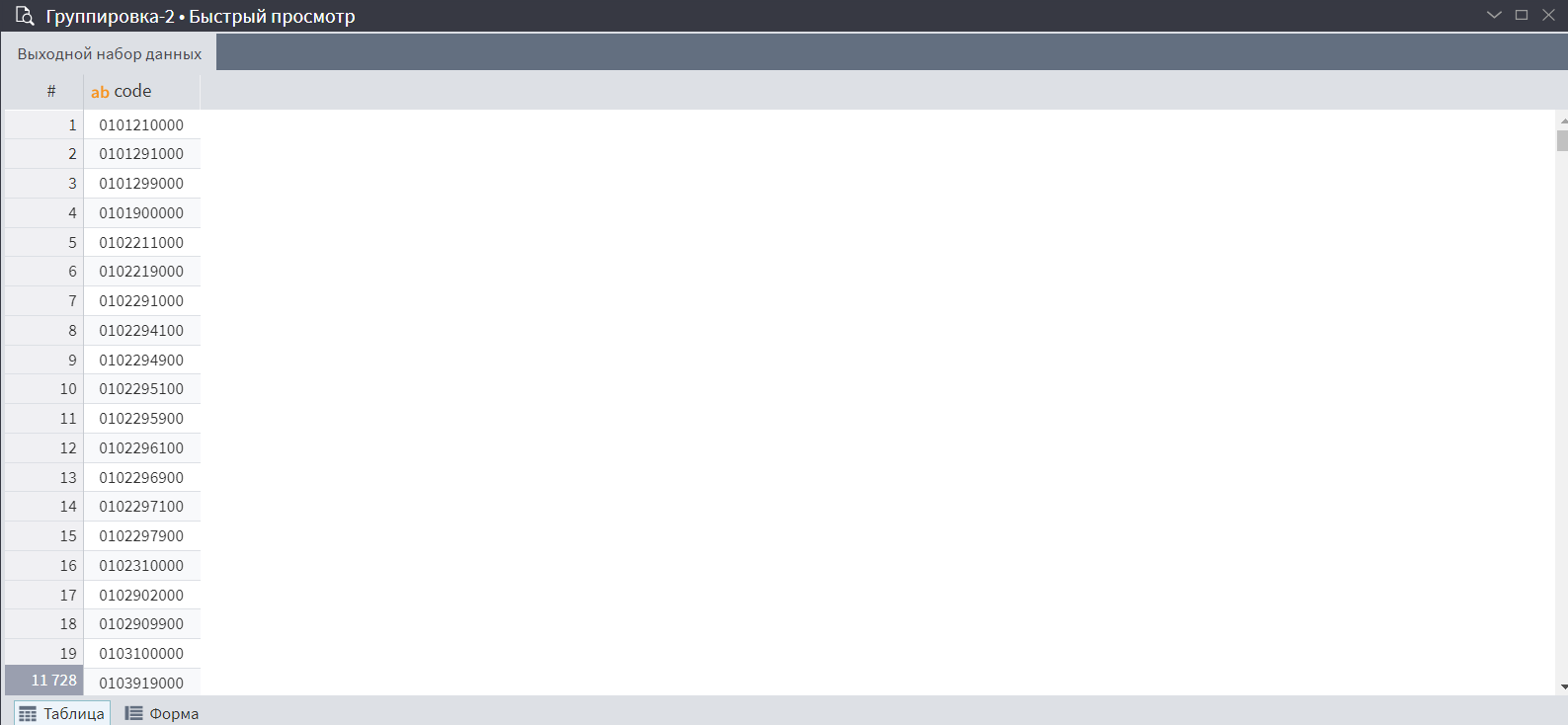
* 4 466 172 строки имеет выборка данных за 2018 год
* 10 столбцов имеет данная выборка
* В данных пропуски? На основе компонентов таблица в переменные можно утверждать о том, что в поле value - 4 256 531 пропусков, netto – 3 096 973 пропусков, quantity – 138 653 пропуска
* 236 уникальных стран участников торговли перечислено в поле country
* 85 уникальных регионов России участвовало в торговле (поле region)
* Динамике качества данных собираемой статистики во всех возможных периодах имеет пустоты данных и не особо зависит от конкретных дат. Поля, дальнейший анализ по которым будет репрезентативен и валиден это поля в которых наименьшее количество пропусков или они отсутствуют, а именно:
  + country
  + month
  + direction
  + code
  + measure
  + region
  + district

Задание на выполнение №3  
Для импортированной выборки за 2016 год на основе компонентов трансформации данных отобразите:

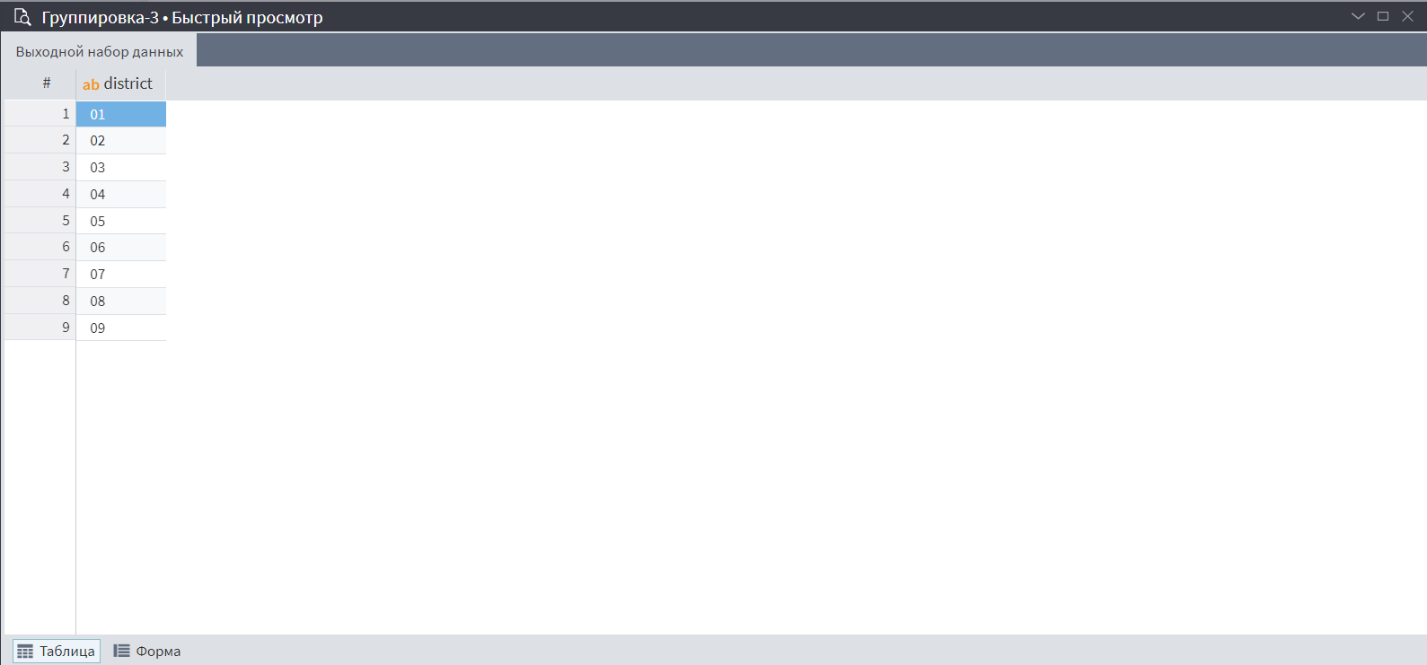
* Перечень уникальных стран-участниц торговли в поле country с помощью компонента группировка (отсортируйте значения по возрастанию)
* Перечень уникальных регионов России участников торговли в поле region с помощью компонента группировка (отсортируйте значения по возрастанию)



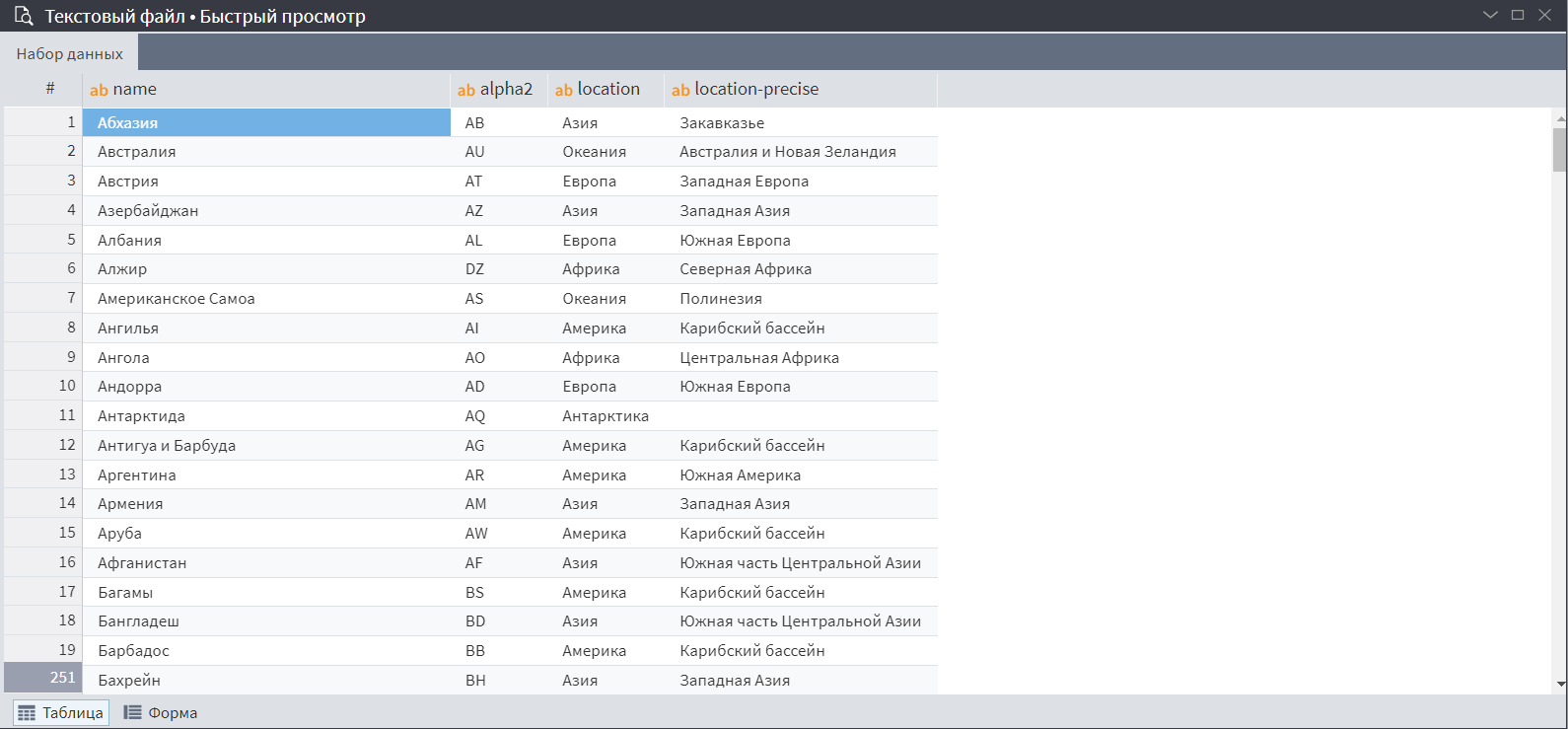
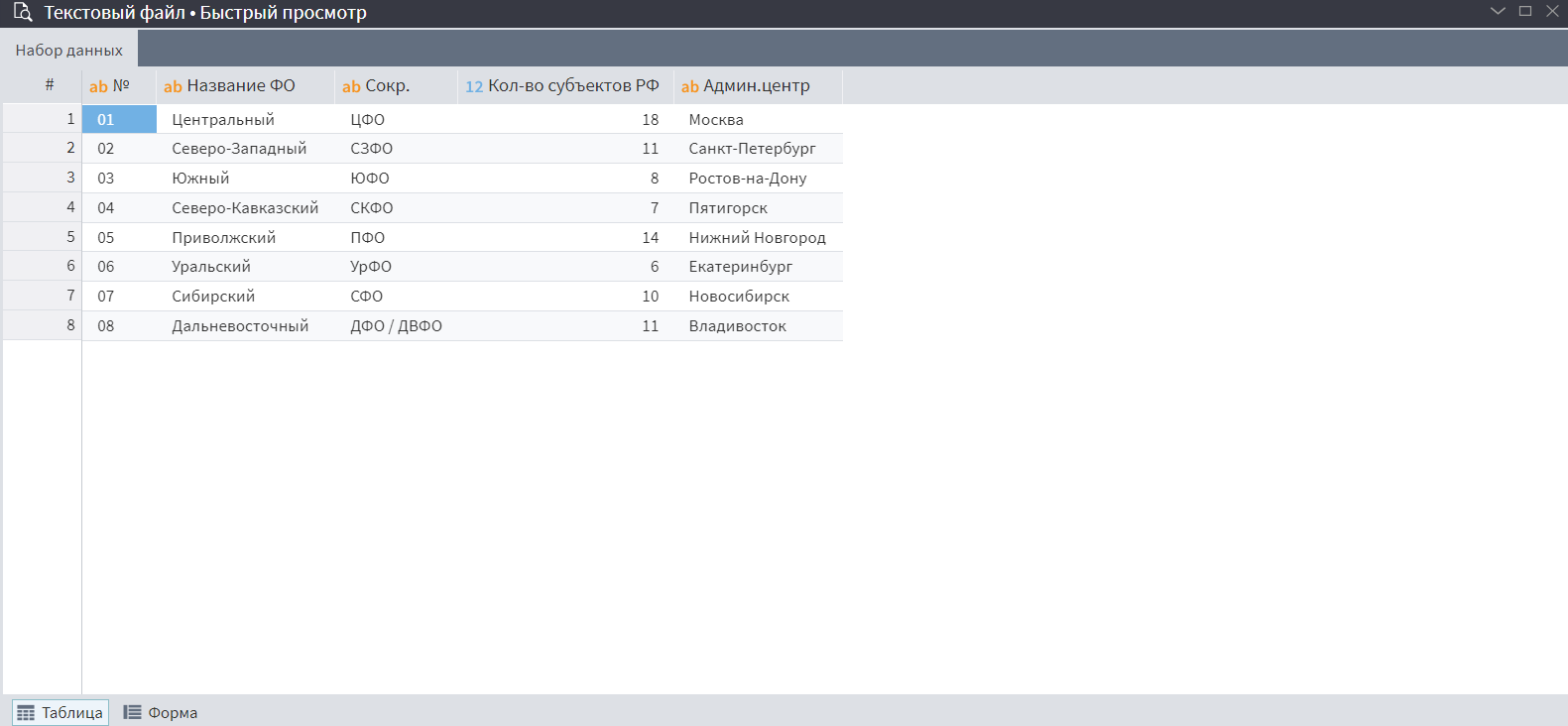
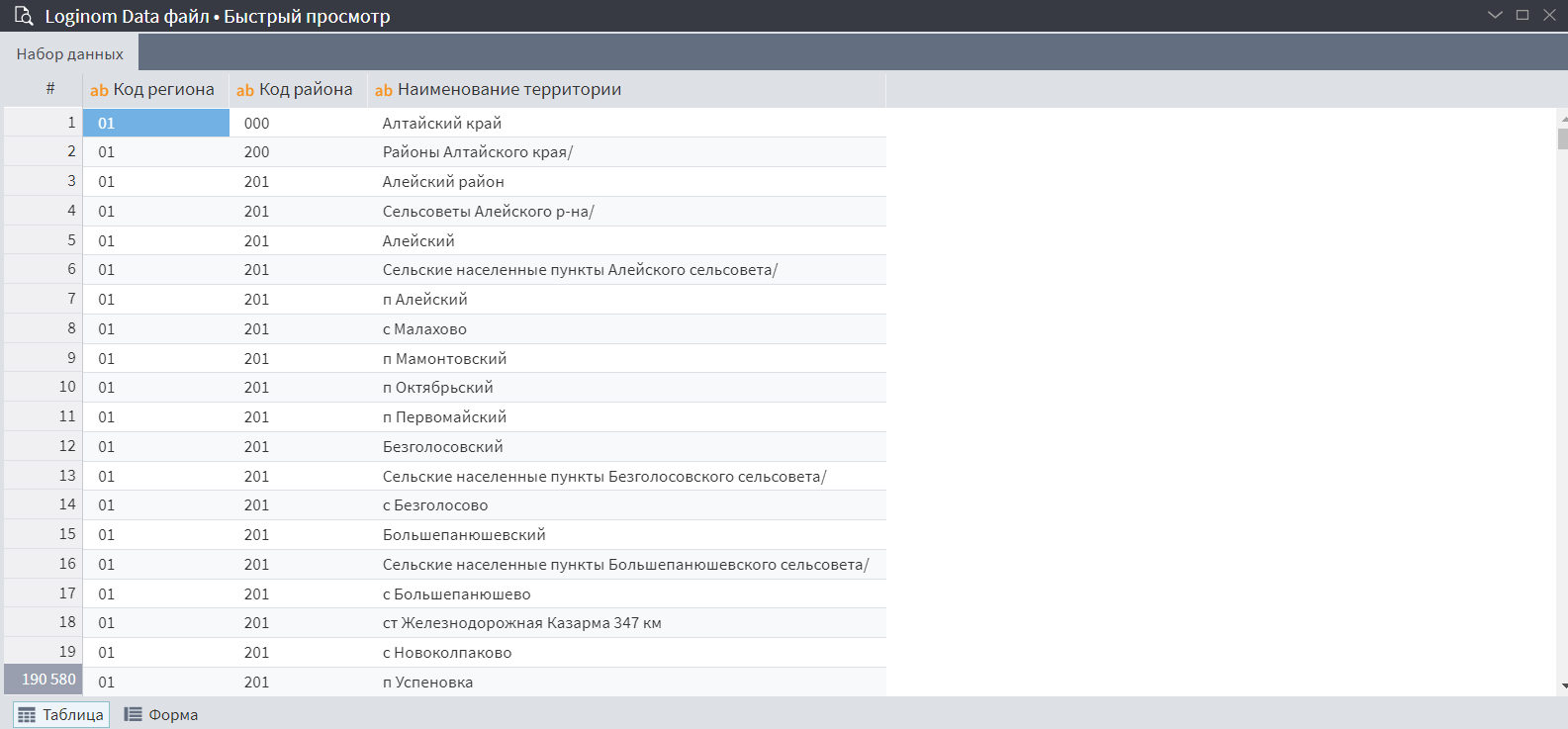
* Перечень уникальных категорий товаров по ТН ВЭД, перевозимых через таможню в поле code с помощью компонента группировка (отсортируйте значения по возрастанию

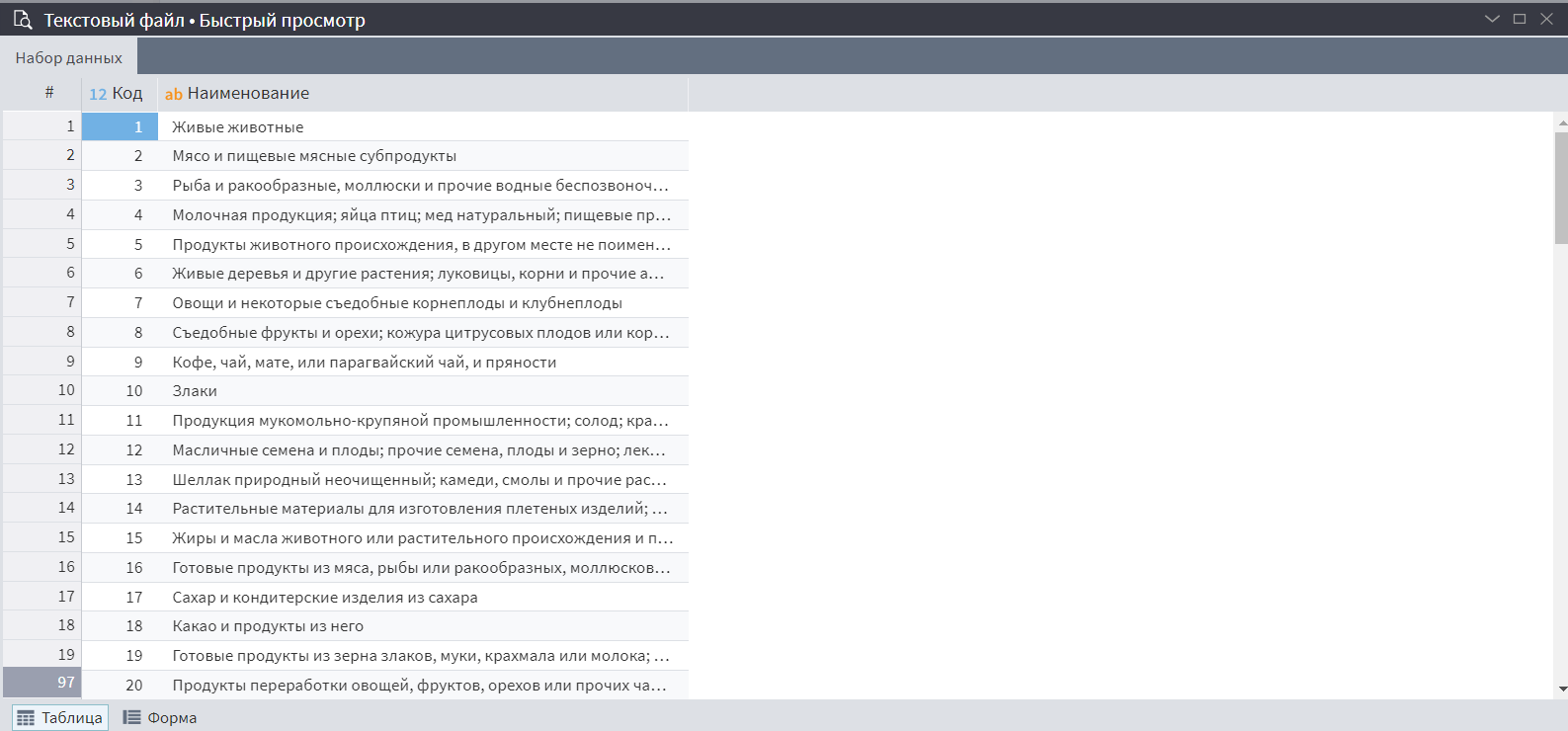


* Перечень уникальных кодов федеральных округов по полю district с помощью компонента группировка (отсортируйте значения по возрастанию)

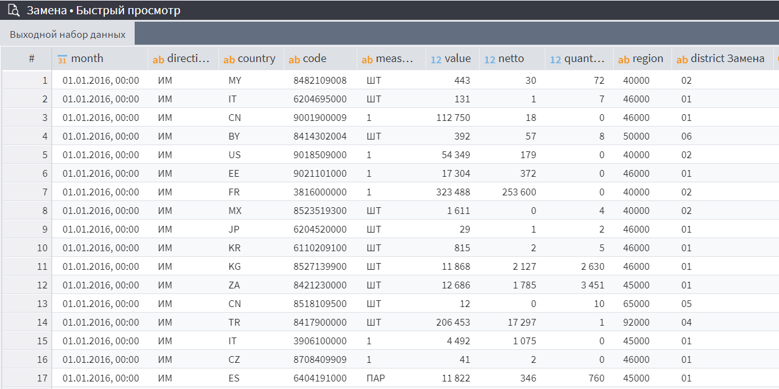


Задание на выполнение №4

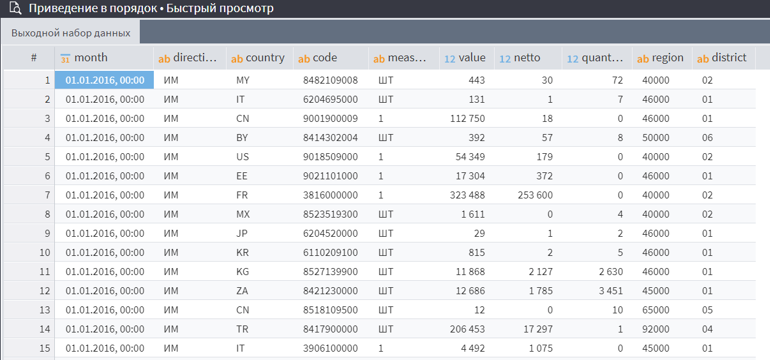
1. Импортируйте в сценарий следующие выборки данных:
   * Файл данных с кодами стран мира ОКСМ в формате alpha2 по пути: Data/Alpha2/alpha2\_codes.csv
   * в кодировке UTF-8,
   * с разделителем столбцов табуляцией.
   * Исключите из полей импортируемой таблицы: fullname, english, alpha3, iso.
2. Импортируйте в сценарий следующие выборки данных:
   * Файл данных с информацией о федеральных округах Data/ОКАТО/fed\_codes.tsv.txt,
   * в кодировке ANSI,
   * с разделителем столбцов табуляцией.
   * Измените тип поля № на строковый тип данных.
   * Исключите из рассмотрения все поля кроме полей с метками: №, Название ФО, Сокр., Кол-во субъектов, Админ. Центр.
3. Импортируйте в сценарий следующие выборки данных:
   * Файл данных с информацией о регионах РФ в соответствии с ОКАТО Data/ОКАТО/okato\_regions.lgd,
   * Исключите из рассмотрения все поля кроме полей с метками: Код региона, Код района, Наименование территории.
4. Импортируйте в сценарий следующие выборки данных:
   * Файл данных с информацией о разделах товарной номенклатуры ТН ВЭД Data/ТН ВЭД/tnved\_codes12.txt
   * в кодировке ANSI,
   * с разделителем столбцов табуляцией.
   * Оставим остальные настройки импорта данных по умолчанию.

Задание на выполнение №5

* В датасете в поле district присутствует федеральный округ с кодом «09», что есть девятый федеральный округ «Крымский ФО», упраздненный в 2016 году
* 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)