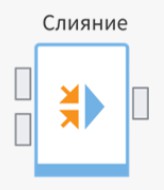
**Рабочая тетрадь №2. Часть 3**

**Обогащение данных Компонент Слияние**

Компонент **Слияние** позволяет дополнить основной набор данных полями из присоединяемого набора. Слияние осуществляется с помощью связей, которые создаются между ключевыми полями входных наборов.

В компоненте доступны 5 **операций слияния**:

* **Внутреннее соединение**
* **Левое соединение**
* **Правое соединение**
* **Полное соединение**
* **Разность**

**Практическое задание №4**

Рассмотрим указанные операции. Для этого нам понадобятся наборы данных

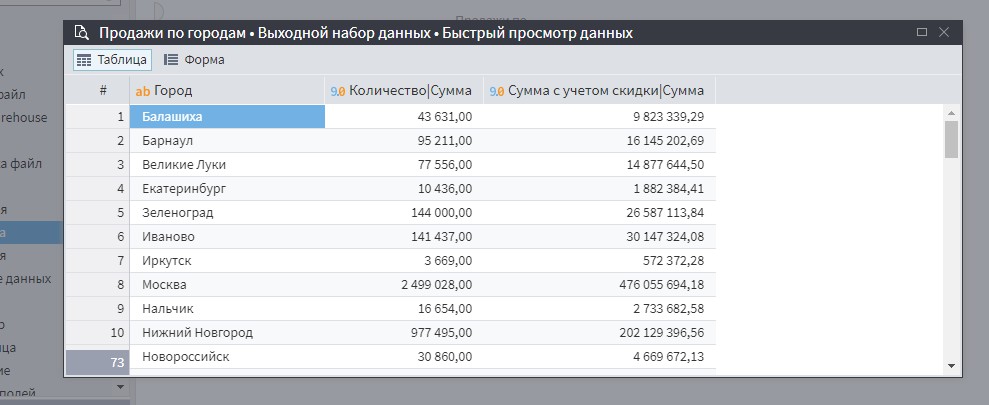
**sales.lgd** и **plan.lgd**.

Создадим в **Loginom** новый пакет. Импортируем указанные наборы данных.

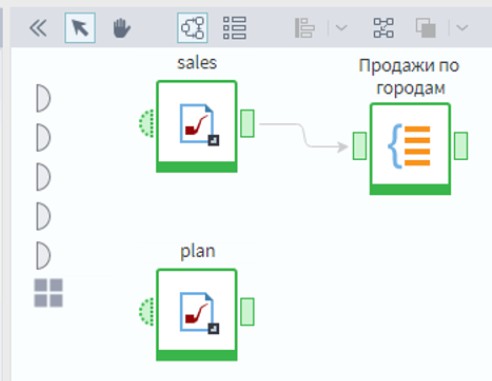
Файл **sales.lgd** содержит информацию о продажах строительных товаров за некоторый период времени в разных городах.

Сгруппируем продажи (**Сумма с учетом скидки**) по городам.

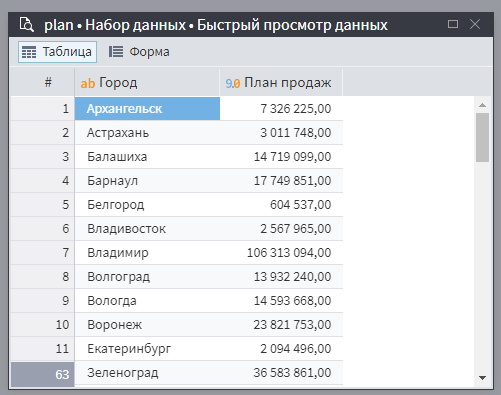
Мы получили набор из 73 городов, в которых произошли продажи.



Файл **plan.lgd** содержит информацию о плане продаж (суммы) по городам.

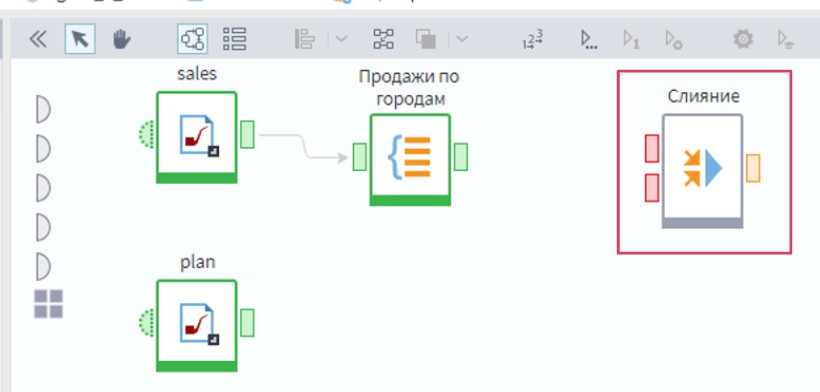


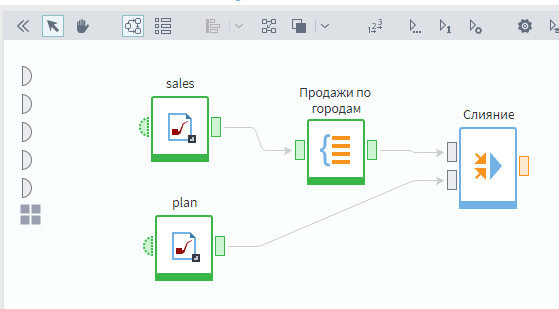
Пусть нам необходимо для дальнейшего анализа иметь в одном наборе данных суммы прошлых продаж и план продаж. Отметим, что план содержит только 63 города, то есть их меньше, чем городов, по которым были продажи.



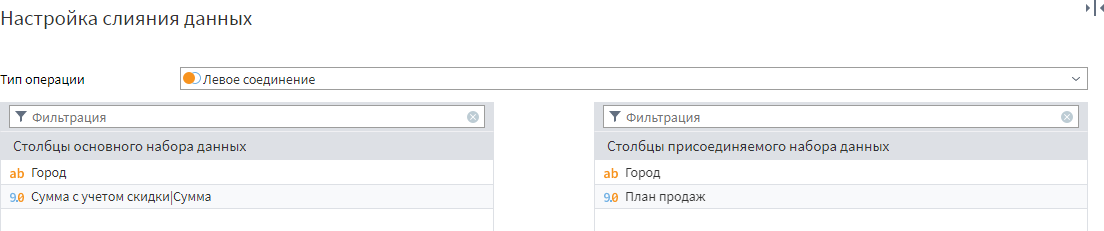
Перенесем в область построения компонент **Слияние**.

Он имеет два входных порта; **Главная таблица** и **Присоединяемая таблица** и только один выходной порт. Результат обработки будет зависеть от выбранного типа операции слияния.

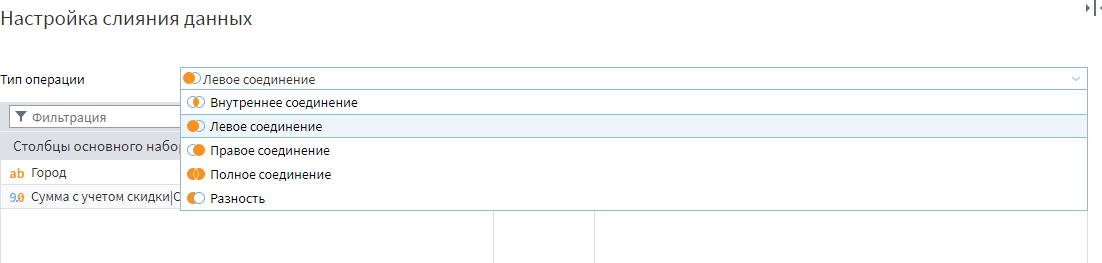


Подадим на порт **Главная таблица** данные с узла группировки **Продажи по городам**, а на порт **Присоединяемая таблица** данные набора **plan.lgd** и перейдем в настройку узла.

Для начала необходимо определиться с типом операции слияния. По умолчанию указан тип **Левое соединение**.



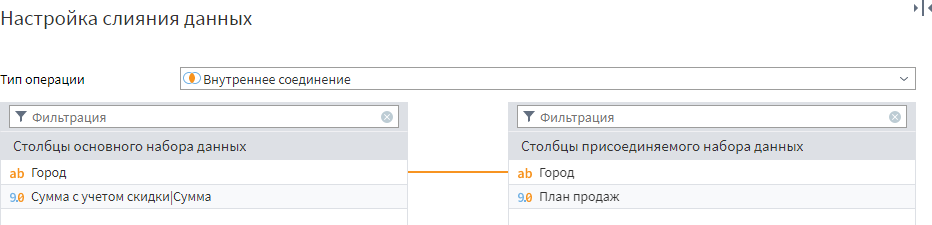
Раскроем список типов операций.



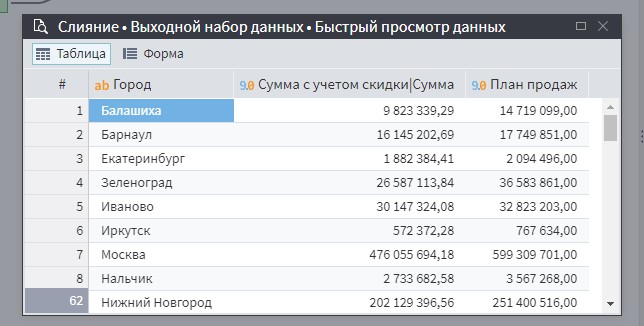
Вариантов слияния пять.

1. Выберем **Внутреннее соединение**.

Основную часть окна настройки занимает область сопоставления полей. Она содержит два списка: **Столбцы основного набора данных** и **Столбцы присоединяемого набора данных**. Здесь создается связь между наборами: каким полям из главной таблицы соответствуют поля присоединяемой.

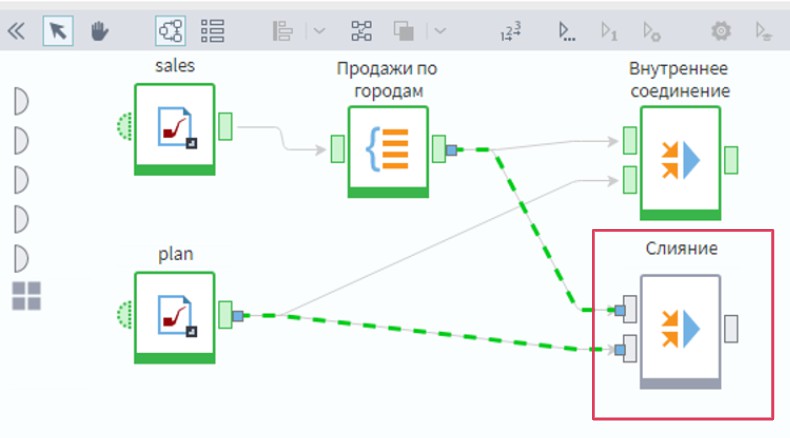
Для связи необходимо с помощью механизма **Drag&Drop** перетащить поле из одного списка на поле, с которым его нужно связать, из другого списка. Связывание допускается только для полей с одинаковыми типами данных.

Создадим связь, как показано на рисунке, и перейдем к просмотру результата.



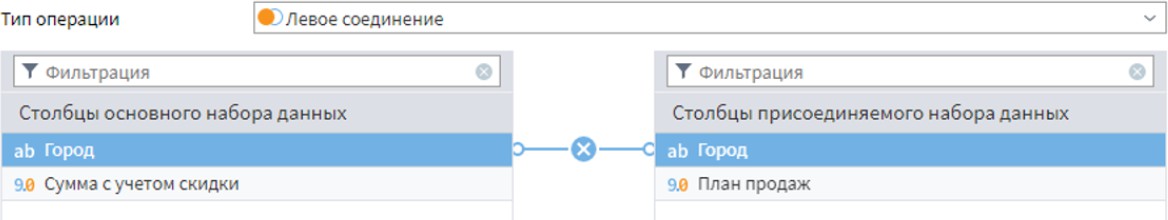
На выходе мы получили набор данных с новым полем **План продаж**. В наборе содержатся только те записи, которые присутствовали в обоих входных наборах, это 62 города.

Добавим еще один узел слияния и создадим такие же связи с его портами.



1. В настройках оставим значение по умолчанию – **Левое соединение**, – связи между полями будут такие же, как и в предыдущем узле.

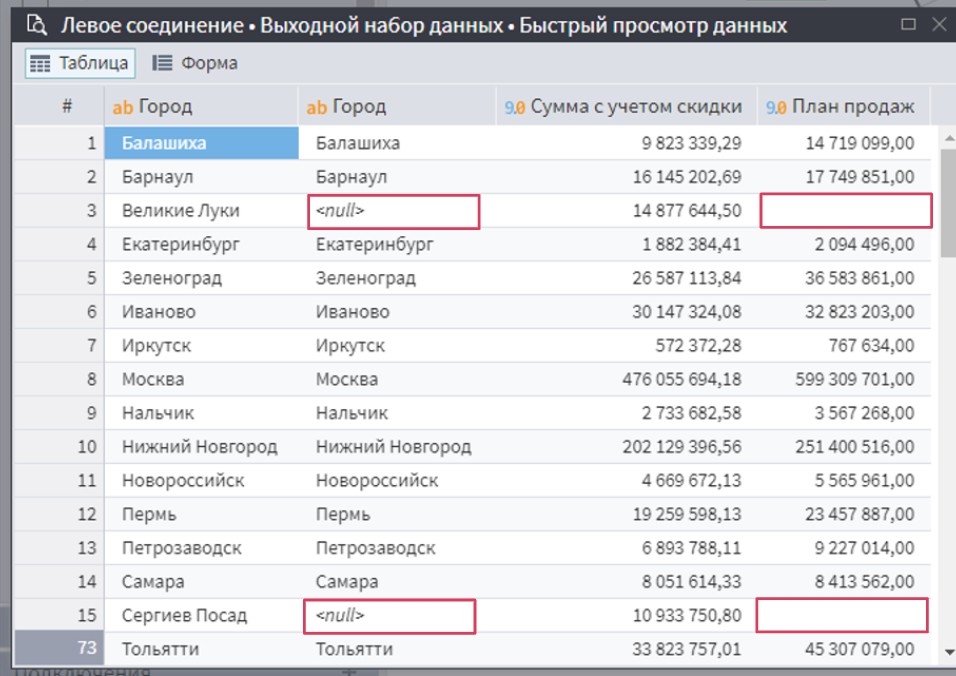
Дополнительно установим флаг **Добавлять присоединяемые ключевые поля**. Это позволит на выходе узла видеть оба ключевых поля **Город**: и из главной таблицы, и из присоединяемой.



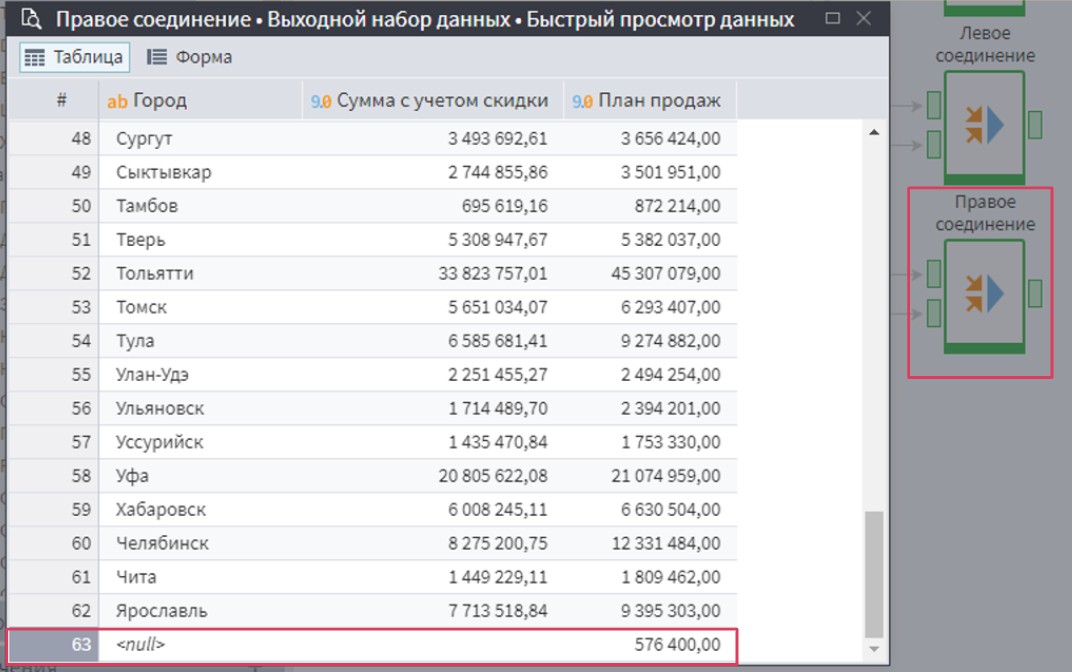
Если флаг не устанавливать, на выходе будет только поле главной таблицы, как в предыдущем узле.

В результате мы получили 73 записи по всем городам, которые содержались в наборе с продажами.

Если для города не нашлось записи в файле **plan.lgd**, для него указаны пустые значения (**null**) в полях **Город** и **План продаж** из присоединяемого набора.



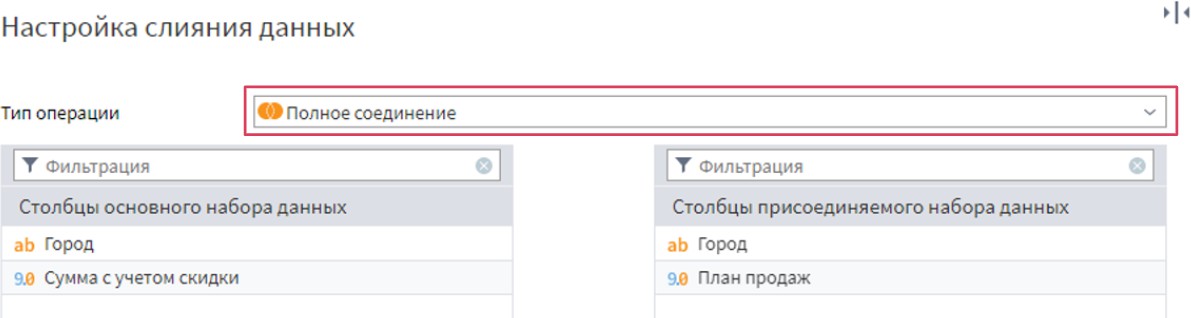
1. Если выбрать **Правое соединение**, мы получим 63 записи, как в наборе **plan.lgd** (здесь и далее флаг **Добавлять присоединяемые ключевые поля** устанавливать не будем).



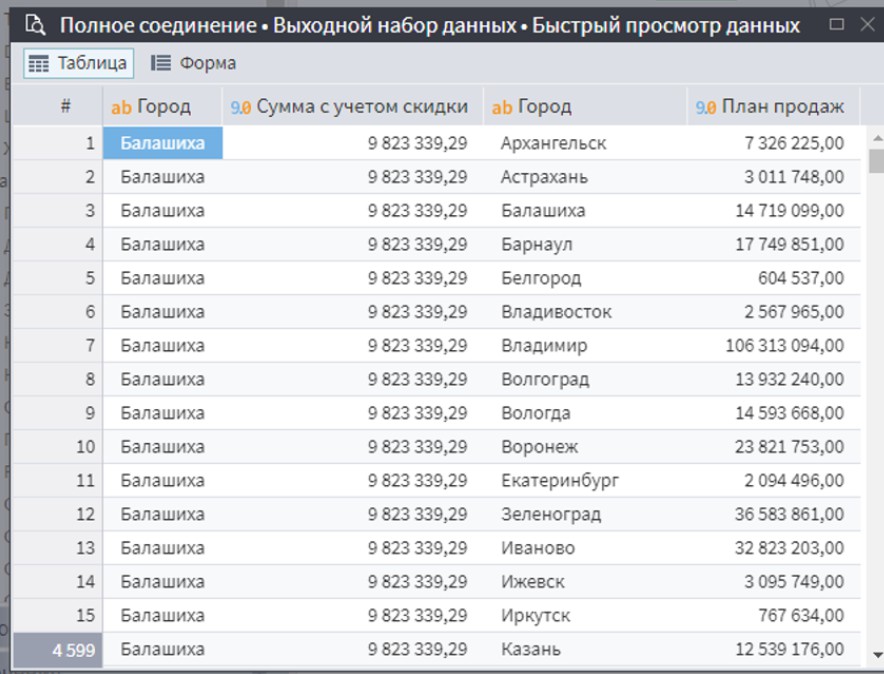
Однако в записях из этого набора, для которых город отсутствует в таблице

продаж, поля **Город** и **Сумма с учетом скидки** будут пусты, это связано с тем, что главной в слиянии является таблица по продажам, и значения этих полей берутся из нее. Поэтому при проведении слияния важно правильно определить главную таблицу.

1. Теперь рассмотрим **Полное соединение**. Оно позволяет соединить каждую строку одной таблицы с каждой строкой второй таблицы, давая в результате все возможные сочетания строк. Это может понадобиться, например, в случае необходимости сравнить каждую строку одного набора с каждой строкой другого с помощью расстояния Левенштейна между строками.



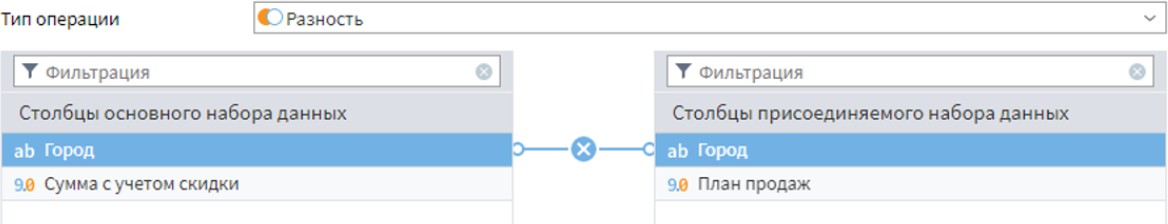
В полном соединении допускается не создавать ни одной связи. Тогда на выходе мы получим все поля из обеих таблиц независимо от того, установлен ли флаг, – так как ключевые поля не выбраны. Перейдем к просмотру результата.



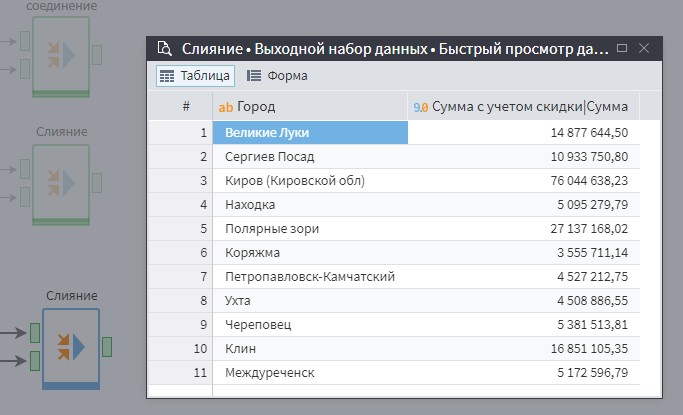
Мы получили 4 599 записей. Каждая запись таблицы с продажами повторяется 63

раза, так как к ней добавилась каждая запись из набора с планом продаж.

1. Последняя операция, которая доступна в узле, **Разность**. При ее использовании мы получаем на выходе только записи главной таблицы, для которых не нашлось соответствий по ключевым полям в присоединяемой таблице.



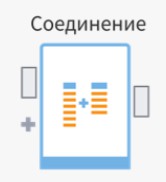
В нашем случае таких записей 11. Для этих городов план продаж не был сформирован.



**Компонент соединение**

Перейдем к рассмотрению следующего компонента – **Соединение**.

С его помощью один набор данных можно дополнить полями из других наборов, либо переменными.

В отличие от слияния, здесь нет ключевых полей. К каждой записи главного набора присоединяются записи из дополнительных наборов с тем же порядковым номером, то есть первая запись соединяется с первой, вторая со второй и так далее.

Кейсы использования: *добавление позиций в рейтинге*.

С помощью компонента **Соединение** возможно, к примеру, добавить в набор данных поле с позициями объектов в рейтинге. Например, в ежемесячном рейтинге по продажам сети магазинов достаточно будет отсортировать магазины по убыванию

суммы продаж, после чего можно добавить поле с позициями.

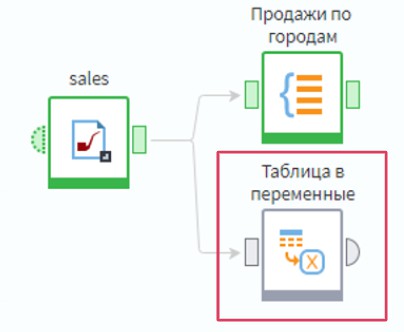


Кейсы использования: *добавление одного значения*.

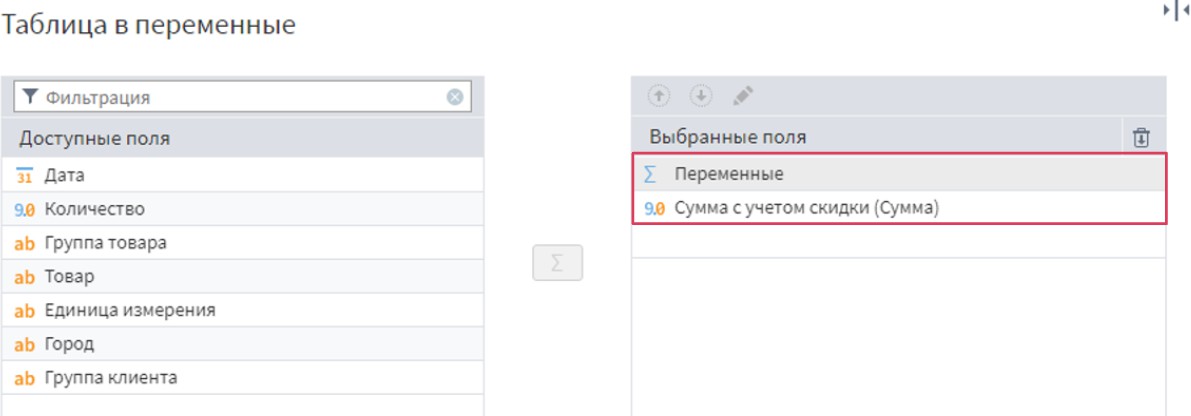
Еще один, наиболее распространенный случай использования **Соединения**, – необходимость добавить в набор данных поле с каким-либо определенным значением. Это значение может содержаться как в таблице, так и в переменных.

**Практическое задание №5**

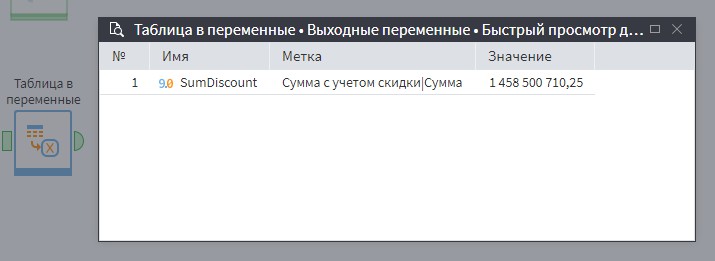
Рассмотрим данную ситуацию на примере. Для этого снова используем набор данных **sales.lgd**.

В новом модуле импортируем указанный набор и сгруппируем продажи но городам. Пусть мы хотим добавить к получившемуся набору столбец, содержащий общую сумму продаж по всем городам. Добавим в сценарий узел **Таблица в переменные**, чтобы получить эту сумму.

В узле выполним настройку, как показано рисунке.

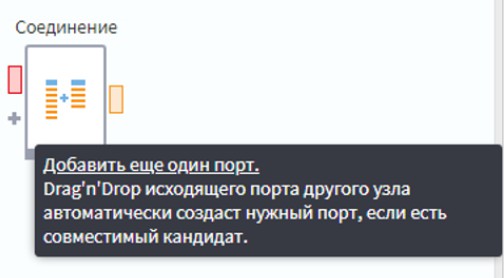


На выходе получили переменную, которая содержит искомое значение.

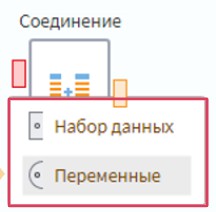


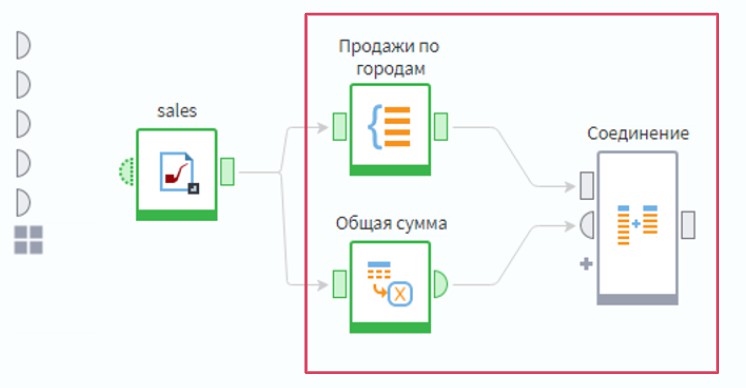
Теперь добавим в область построения узел **Соединение**. По умолчанию он имеет всего один порт **Главная таблица**. Дополнительные порты нам необходимо добавить самостоятельно. Это связано с тем, что они могут быть двух типов: **Набор данных** и **Переменные.**

Нажмем кнопку **Добавить еще один порт**.

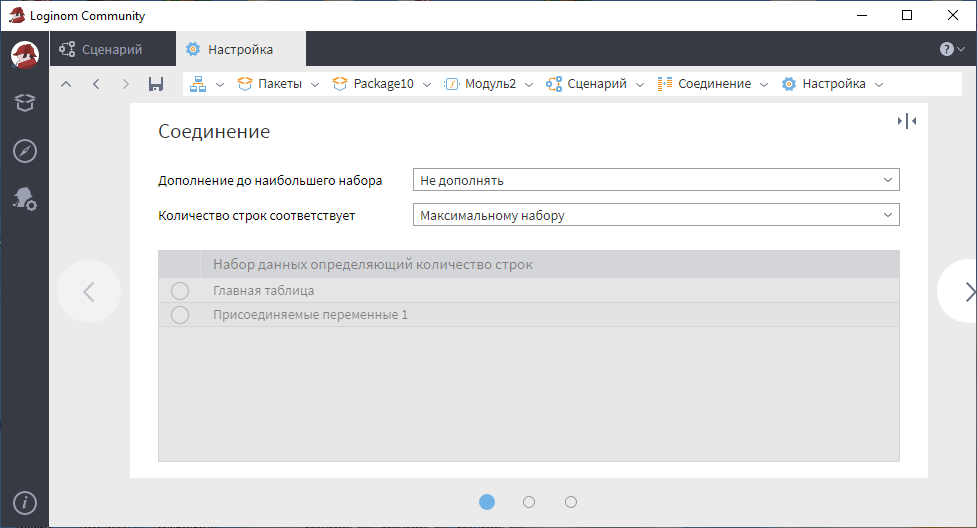


Выберем вариант **Переменные**.



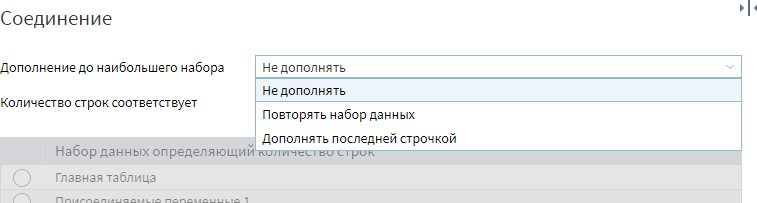
Теперь мы можем подать на порты соответствующие данные. Можно добавить любое количество дополнительных портов. Кроме указанного способа это можно сделать, протянув к кнопке **Добавить еще один порт** связь от выходного порта другого узла

Рассмотрим параметры окна настройки узла **Соединение**.



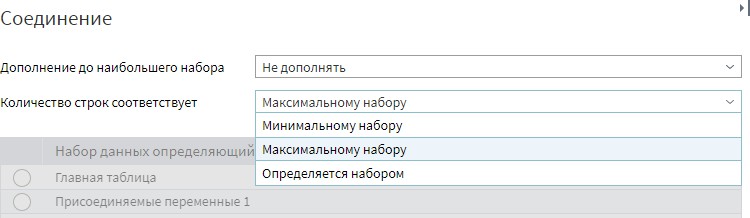
В поданных на вход узла наборах данных может быть разное количество строк. Данный параметр определяет, каким образом меньший набор будет дополняться до большего:

* **Не дополнять** – меньший набор будет дополнен до большего пустыми значениями.
* **Повторять набор данных** – значения в полях меньшего набора будут повторяться: после последней строки поля снова будет идти первая, затем вторая и так далее.
* **Дополнять последней строчкой** – значение из последней записи добавляемого поля меньшего набора заполнит все недостающие строки.

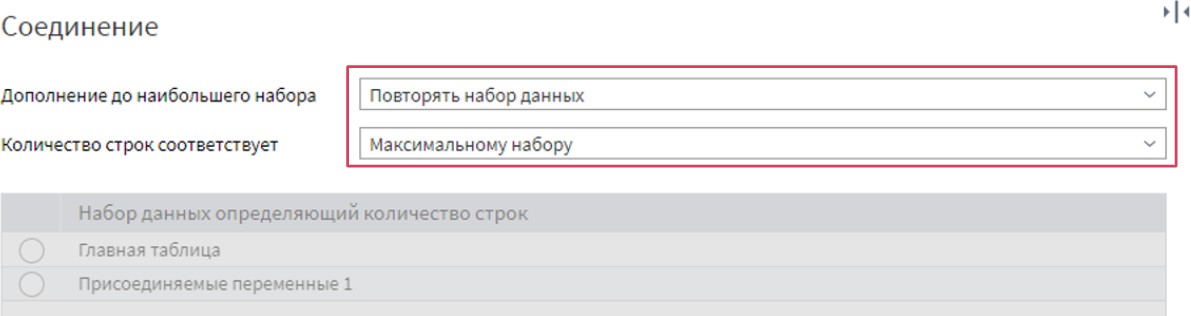


Следующий параметр позволяет выбрать, какому из входных наборов будет соответствовать количество строк:

* + **Максимальному набору** – значение по умолчанию.
  + **Минимальному набору**.
  + **Определяется набором** – при выборе данного варианта станет активна серая область ниже, где можно будет выбрать любой из поданных на вход узла наборов данных.

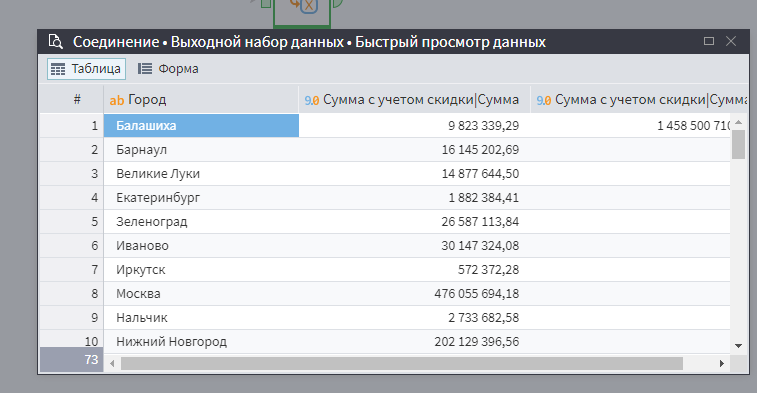


Добавим в главную таблицу столбец со значением переменной. Выберем параметры, как показано на рисунке и перейдем к просмотру результата:



К нашему набору по продажам добавилось еще одно поле — Общая сумма, —

которое содержит значение одноименной переменной



**Компонент Дополнение данных**

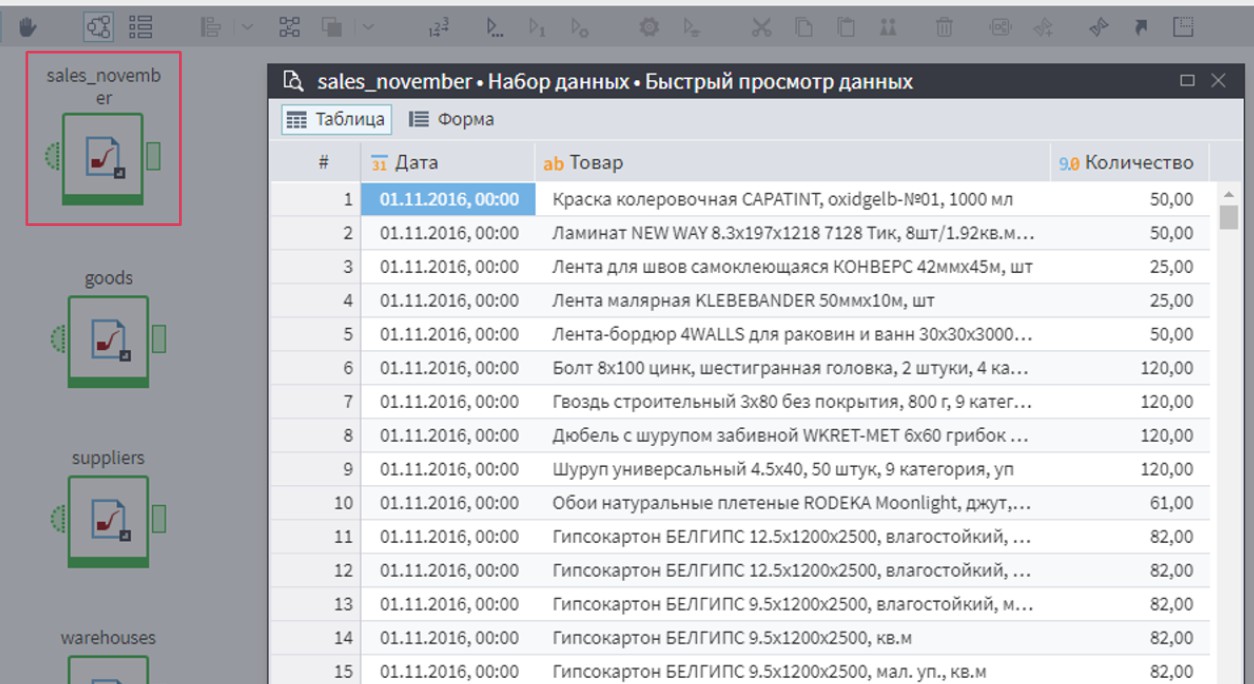
Компонент позволяет обогатить данные путем присоединения к главной таблице полей из других таблиц по принципу левого соединения.

Как и в слиянии, здесь используются связи по ключевым полям, но число присоединяемых таблиц не ограничено.

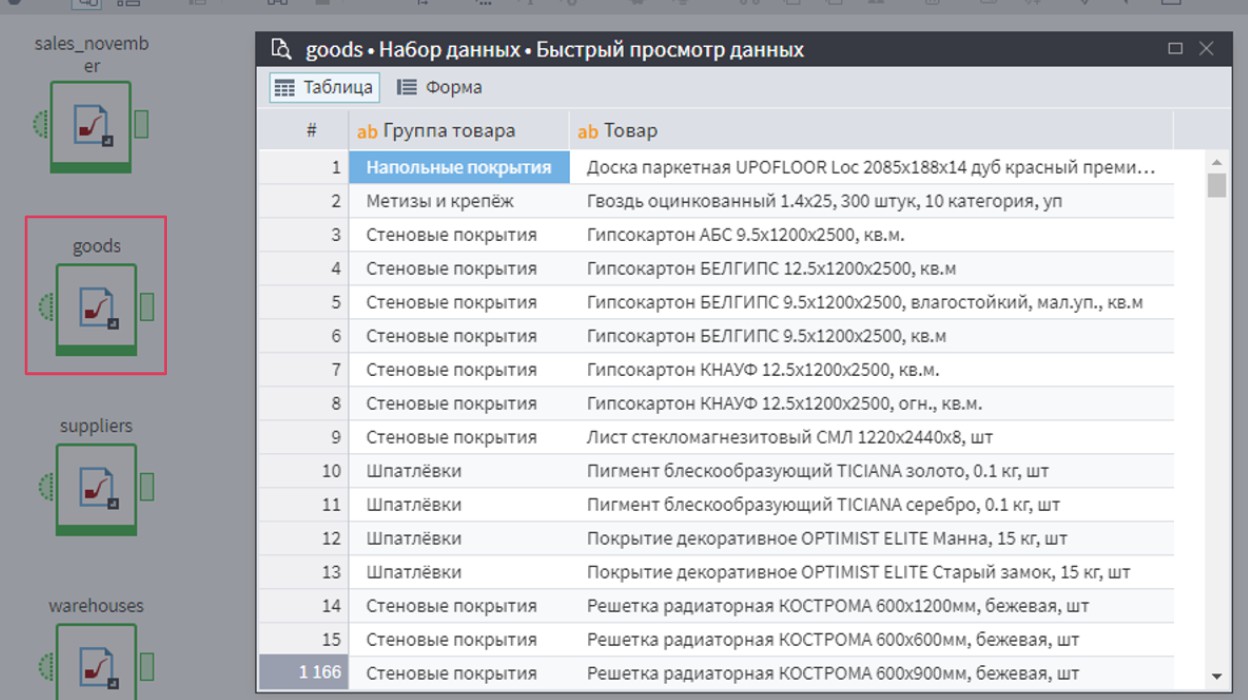
**Практическое задание №6**

Рассмотрим пример обогащения набора данных из нескольких справочников. Нам понадобятся наборы **sales\_november.lgd**, **goods.lgd**, **suppliers.lgd**, **warehouses.lgd** (прилагаются к практикуму).

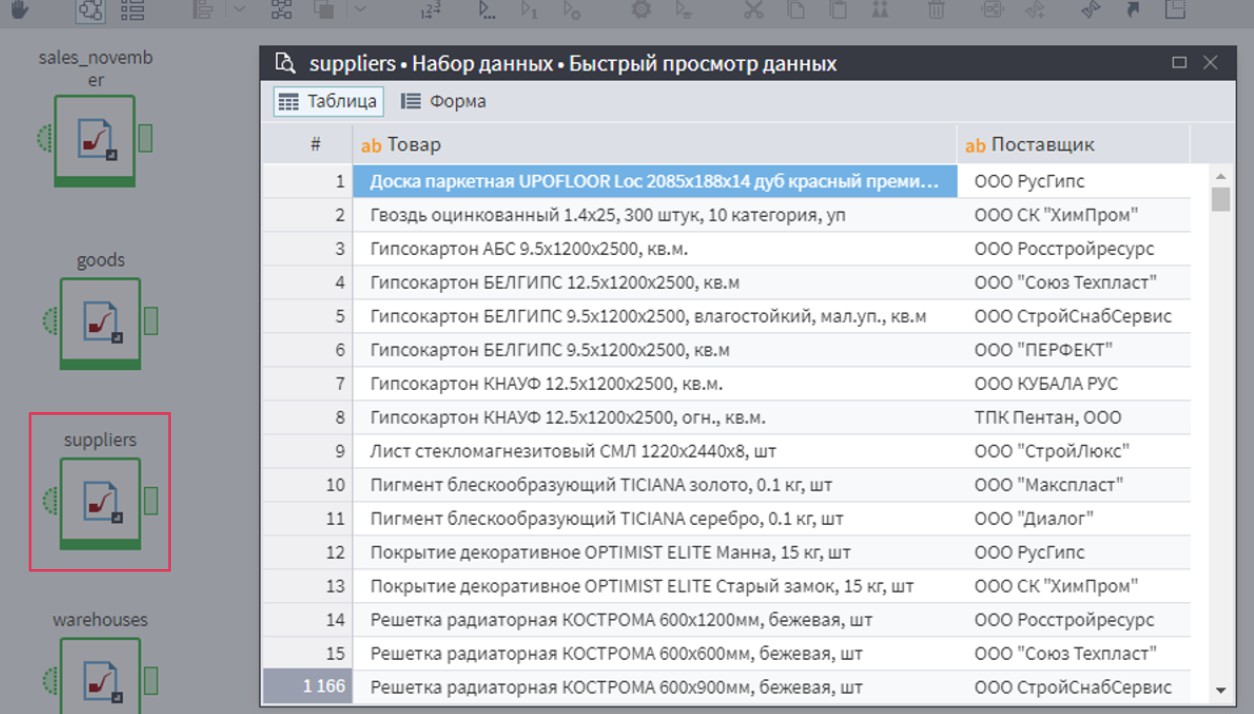
Импортируем перечисленные наборы.

Набор **sales\_november.lgd** содержит данные о продажах строительных товаров за ноябрь с указанием даты продажи и количества проданного товара. Пусть мы хотим, добавить к этим данным информацию о группах товаров, их остатках на складе и поставщиках.

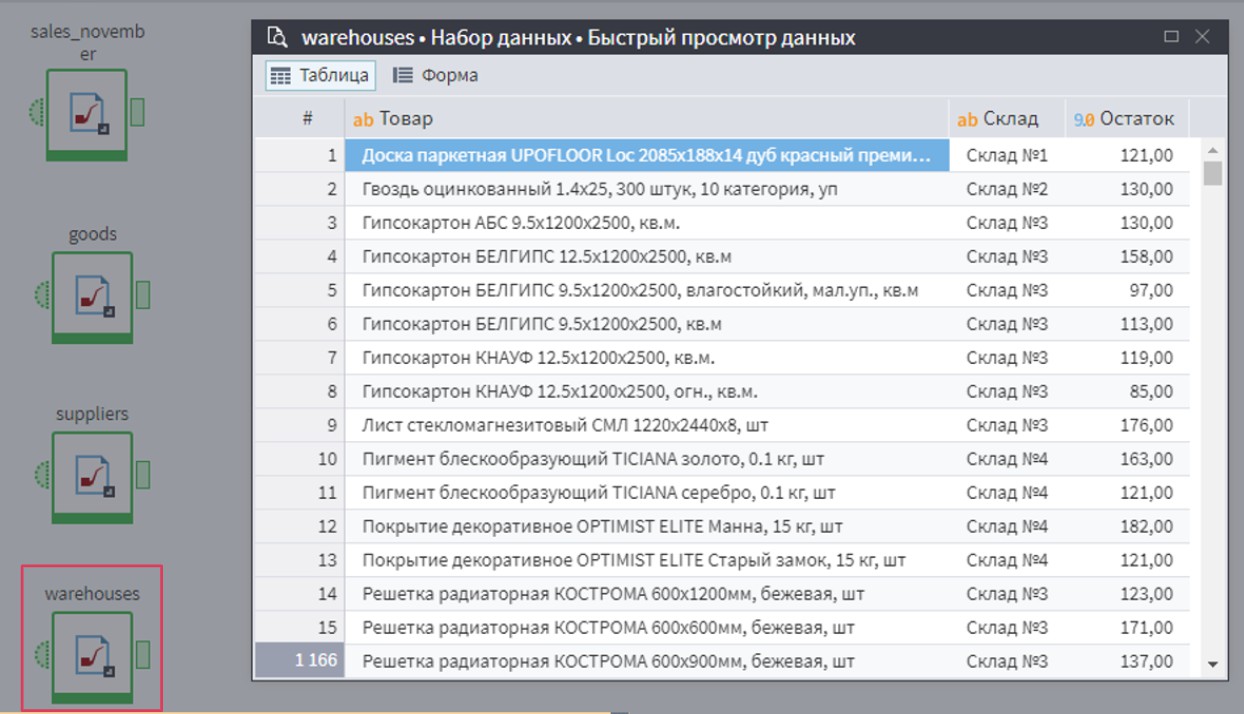
В наборе **goods.lgd** хранится полный список товаров с указанием групп, к которым они относятся.



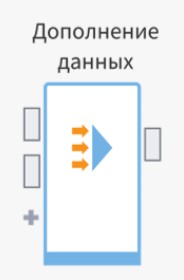
Набор **suppliers.lgd** содержит список товаров с указанием поставщика.



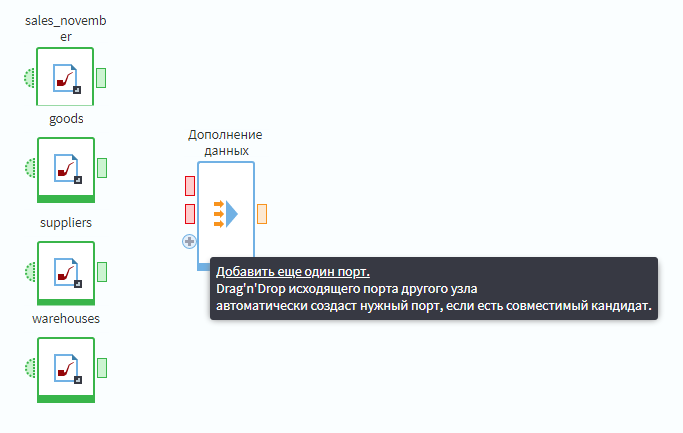
Последний набор – **warehouses.lgd** – содержит информацию о складе, на котором находятся товары, и остатках на 1 ноября 2016 года.



Добавим в область построения узел **Дополнение данных**.

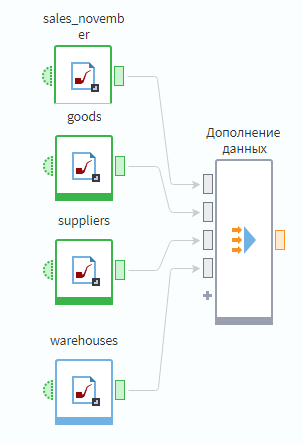


По умолчанию он имеет два входных порта.

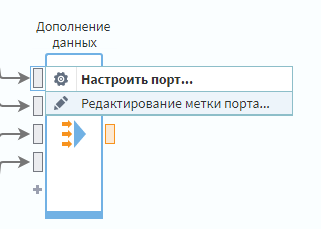


Для того чтобы добавить дополнительные порты, можно нажать кнопку **Добавить еще один порт** или протянуть к этой кнопке связь от выходного порта другого узла.

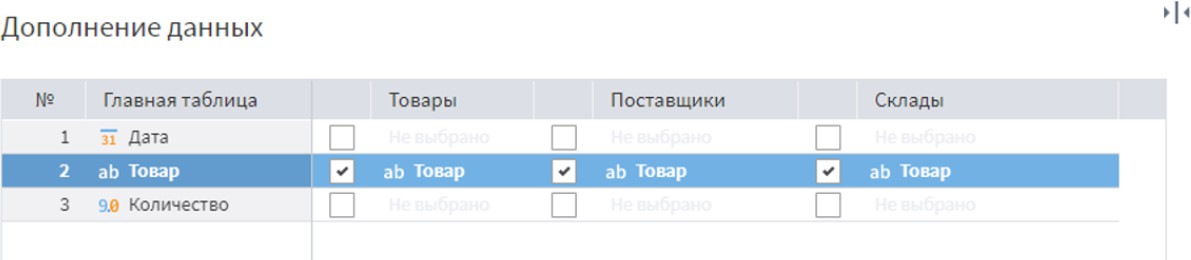
Подадим на входы все наши наборы данных.



Для удобства переименуем порты и перейдем в настройку узла.



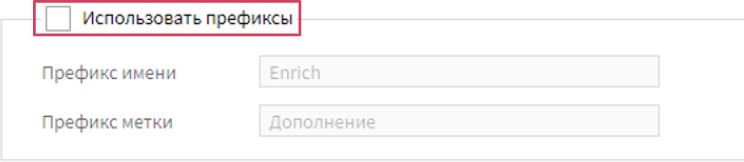
Основную часть окна занимает **Область настройки ключевых полей**. Здесь необходимо напротив ключевого поля главной таблицы выставить флаг в столбце каждой присоединяемой таблицы. Поставим флаг для таблицы **Товары** напротив поля **Товар главной таблицы**.



Стал активен список доступных полей набора данных **Товары** Доступны поля с таким же типом данных. Автоматически установилась связь по полю **Товар**.

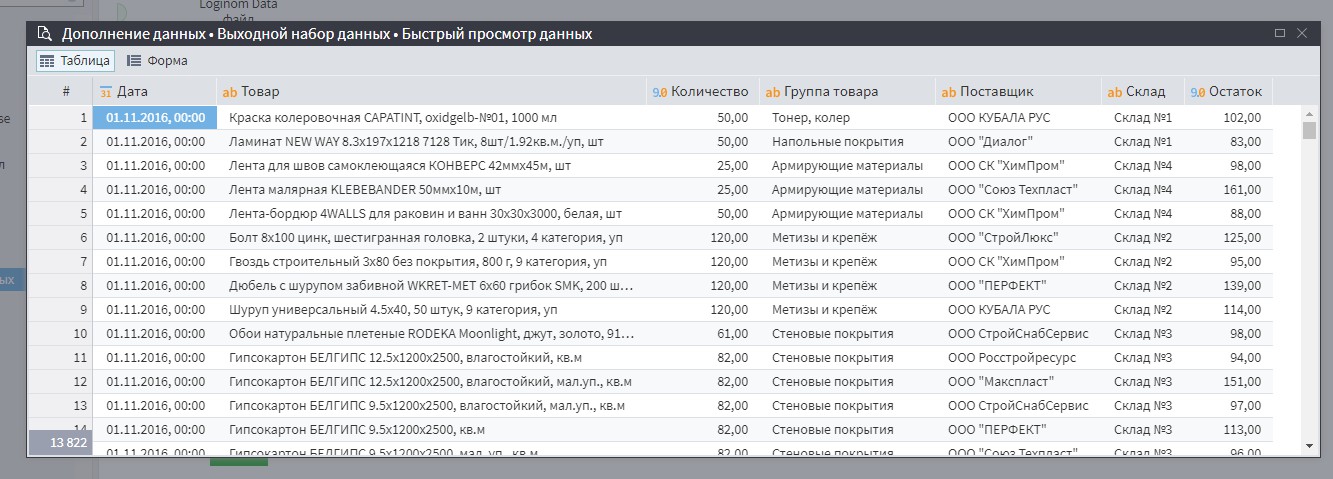
Значок фильтрации включен по умолчанию и означает, что в списке доступны только те поля, которые еще не связаны с ключевыми полями главной таблицы. Если выключить фильтрацию, в списке будут отображаться все поля, совместимые по типу данных

В нижней части окна доступен флаг **Использовать префиксы**. Если его включить, именам и меткам полей, взятых из присоединяемых таблиц, на выходе будут добавлены префиксы, указанные, соответственно, в параметрах **Префикс имени** и **Префикс метки**.



Проставим соответствия для остальных наборов и посмотрим на результат.

На выходе получили все тот же список продаж за ноябрь 2016, к которому добавилась информация из выбранных полей. Таким образом можно с помощью одного узла легко собрать в один набор информацию из разных таблиц.



**Компонент Объединение**

С помощью трех рассмотренных выше компонентов можно дополнять набор данных *полями* из других наборов. Компонент **Объединение** позволяет дополнить набор, поданный на порт **Главная таблица**, *записями* из присоединяемых наборов.

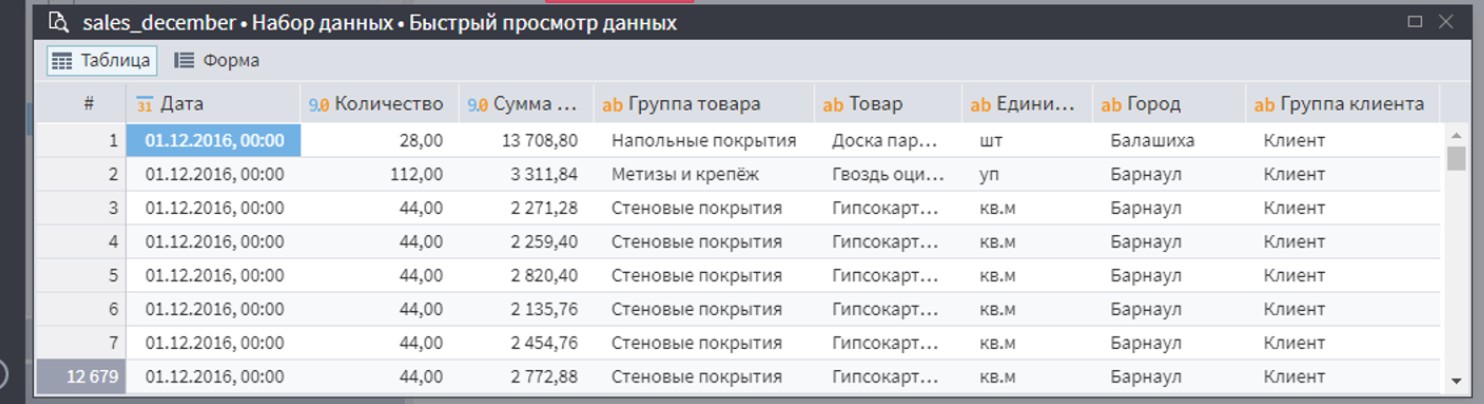
**Практическое задание №7**

Для рассмотрения работы компонента нам понадобятся наборы данных **sales.lgd**

и **sales\_december.lgd**.

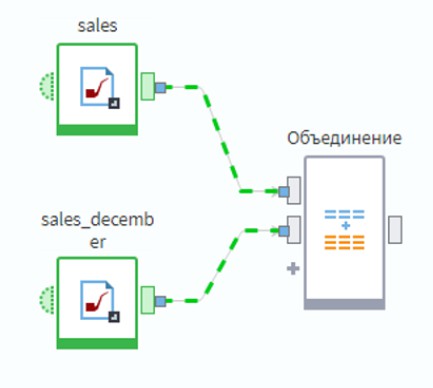
Импортируем указанные наборы.

Набор **sales\_december.lgd** содержит такие же данные по продажам, что и **sales.lgd**, но за декабрь 2016-го года. Задача состоит в том, чтобы добавить к данным по продажам за несколько месяцев декабрьские продажи.



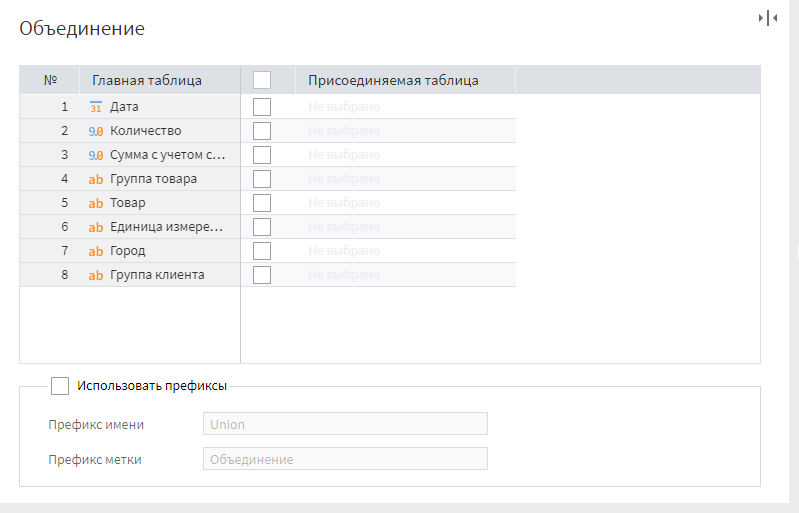
Создадим узел **Объединение** и подадим на его порты наши наборы, как показано

на рисунке.



При необходимости у узла можно добавить любое количество портов.

Настройка узла похожа на настройку **Дополнения данных**. Здесь также необходимо проставить флаги напротив полей главной таблицы, которым соответствуют поля присоединяемых. Однако в данном случае мы дополняем основной набор записями из присоединяемого, поэтому соответствия необходимо проставить для всех возможных полей.

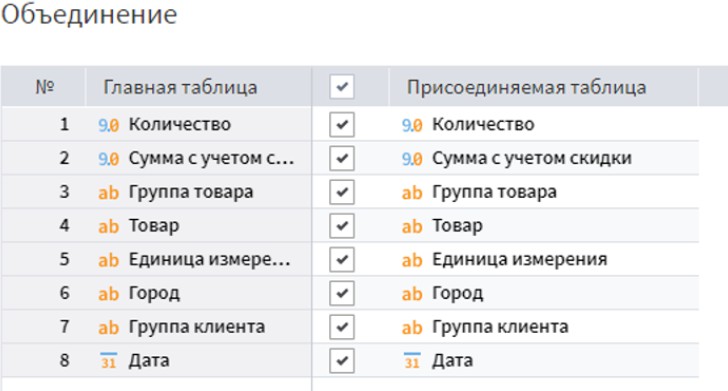


Если в присоединяемой таблице есть поля, для которых в главной отсутствует

(или не проставлено) соответствие, они добавятся в выходной набор в качестве отдельных полей. Для них можно использовать префиксы.

В нашем случае все поля присоединяемой таблицы аналогичны полям главной.

При проставлении соответствий необходимо помнить, что сопоставить можно только поля с одинаковым типом данных



На выходе мы получаем набор данных с такими же полями, как и в главной таблице, но с большим количеством строк: 111150.

