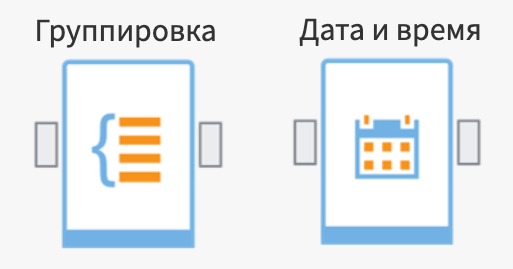
# Рабочая тетрадь №2. Часть 2

**Компоненты «Группировка» и «Дата и время»**

В **Loginom** существуют специальные компоненты для всех методов преобразования данных.

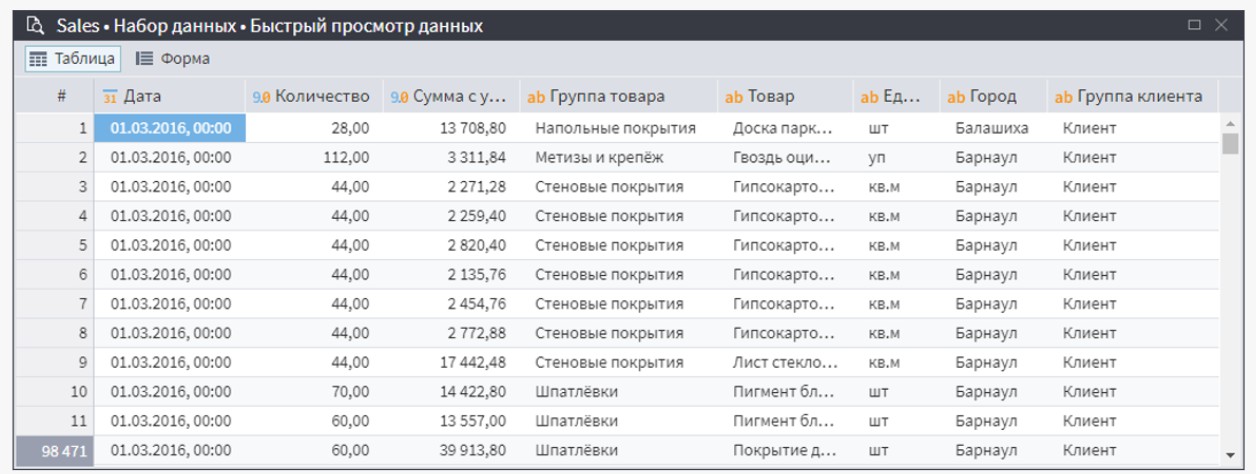
За операции группировки и преобразования даты отвечают компоненты

# Группировка и Дата и время.

Для начала остановимся на первом компоненте, преобразования будем проводить над набором данных **sales.lgd**.

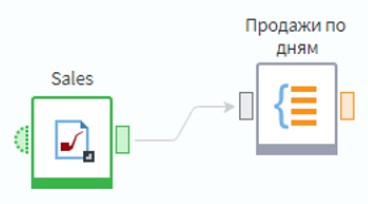
# Постановка задачи №3

В файле **sales.lgd** хранятся данные о продажах строительных товаров: их названия и товарные группы, дата продажи, количество проданного и единица измерения, сумма с учетом скидки, а также информация о городе продажи и типе клиента- покупателя. Всего в наборе 98 471 запись.

Нам необходимо получить набор данных с продажами каждого товара по дням. Поскольку один и тот же товар мог быть продан в один день несколько раз в разных городах, нам потребуется провести группировку. Для этого создадим новый пакет и импортируем данный набор в **Loginom**.

Компонент **Группировка** позволяет объединять записи по выбранным полям в

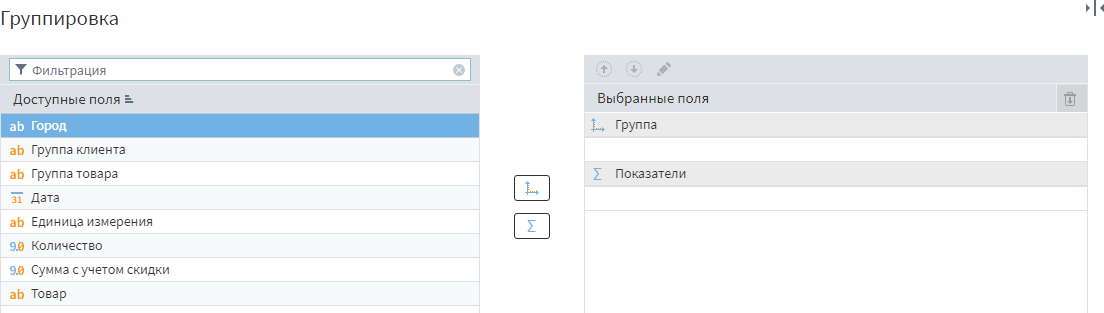
группы, агрегируя данные в полях, которые выбраны в качестве показателей, с помощью различных статистических функций. Для каждой группы возвращается одна строка.

Создадим в области построения узел **Группировка**. Свяжем его с узлом импорта и перейдем в настройку.

Окно настройки узла группировки состоит из двух основных областей.

По умолчанию входные поля попадают в область **Доступные поля** в левой части окна.

После соответствующих настроек поля попадают в область **Выбранные поля**, где распределяются по спискам **Группа** и **Показатели**.



Мы хотим получить на выходе суммарные продажи по дням и товарам, значит, в списке групп нам нужны поля **Дата** и **Товар**. Выделим их и нажмем кнопку

# Переместить в Группу .

Добавить поля в соответствующий список области выбранных полей можно также перетаскиванием, с помощью контекстного меню или горячих клавиш:

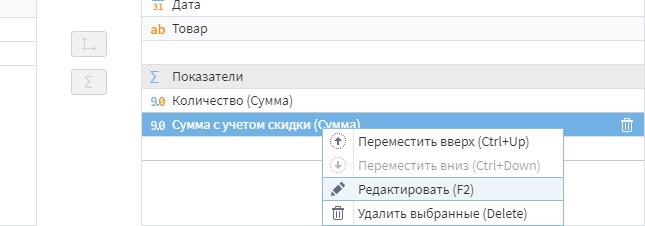
* **ALt+G** – в список **Группа**;
* **Alt+S** – в список **Показатели.**

Выбранные поля отобразились в списке **Группа.**

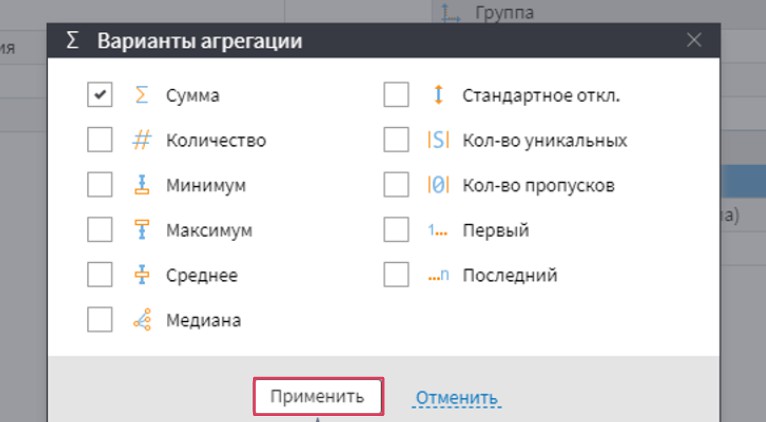
Аналогичным образом добавили поля **Количество** и **Сумма** с учетом скидки в

список показателей. Для них автоматически установился вариант агрегации **Сумма**,

используемый по умолчанию для числовых и вещественных полей.

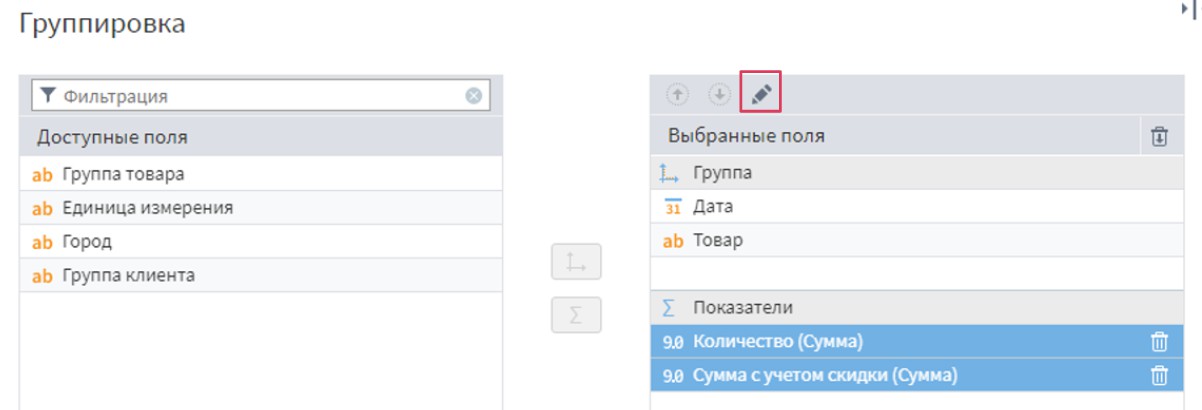
При необходимости изменить вариант агрегации для поля необходимо дважды щелкнуть по нему левой кнопкой мыши.

Появится окно вариантов агрегации.



В узле доступны те же варианты агрегации, что и в компоненте **Таблица в переменные**. Не будем ничего менять и нажмем кнопку **Применить.**

Если для нескольких полей нужно установить одинаковые варианты агрегации, нужно выделить их в списке показателей и нажать кнопку **Редактировать** на панели инструментов.

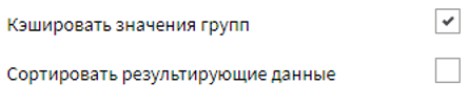


Тогда выбранные варианты установятся для каждого выделенного поля.

Если нам необходимо убрать какие-то поля из области выбранных, необходимо щелкнуть на каждом левой кнопкой мыши при зажатой клавише **Ctrl** на клавиатуре, после чего использовать один из трех вариантов:

* нажать клавишу **Delete**;
* перетащить поля в область доступных полей;
* использовать команду **Удалить выбранные** контекстного меню.

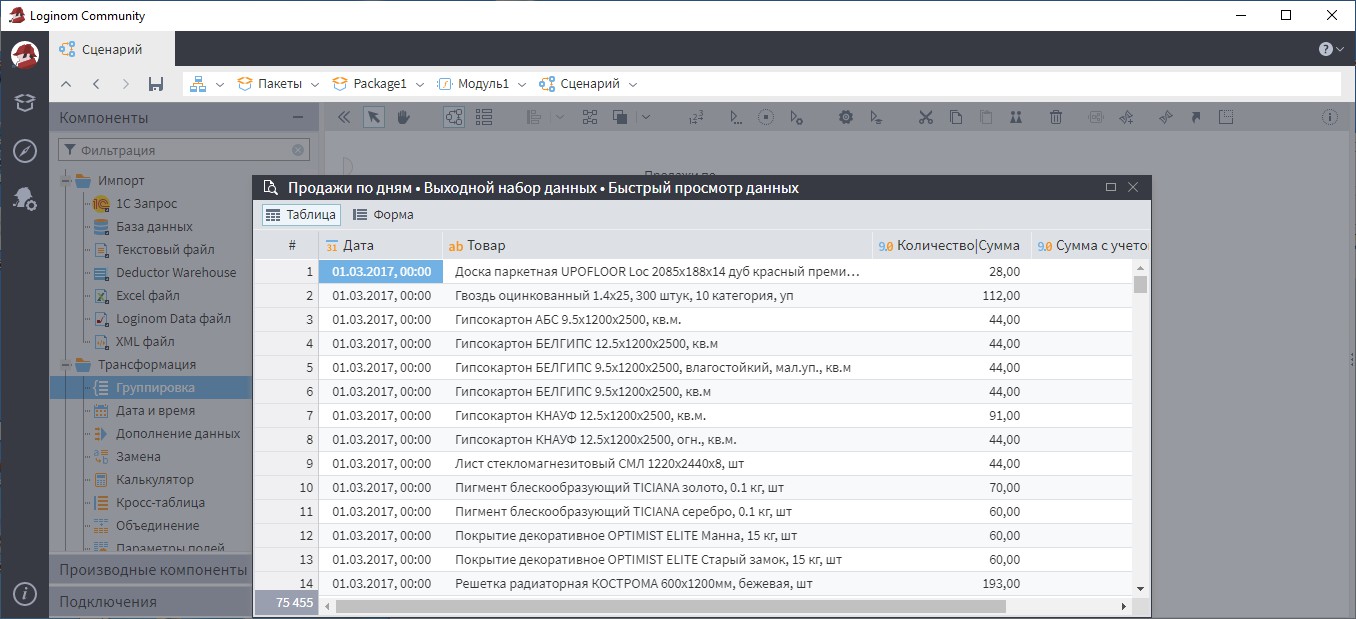
Рассмотрим еще два параметра настроек узла.

Флаг **Кэшировать значения групп** включен по умолчанию и позволяет сохранить полученные после группировки значения групп. Это может существенно ускорить последующую обработку, особенно при подаче большого набора данных, однако необходимо учитывать, что чем больше набор, тем больше памяти займут сохраненные значения.

При установке флага **Сортировать результирующие данные** записи на выходе будут отсортированы по полям в списке **Группы** с учетом их расположения в списке.

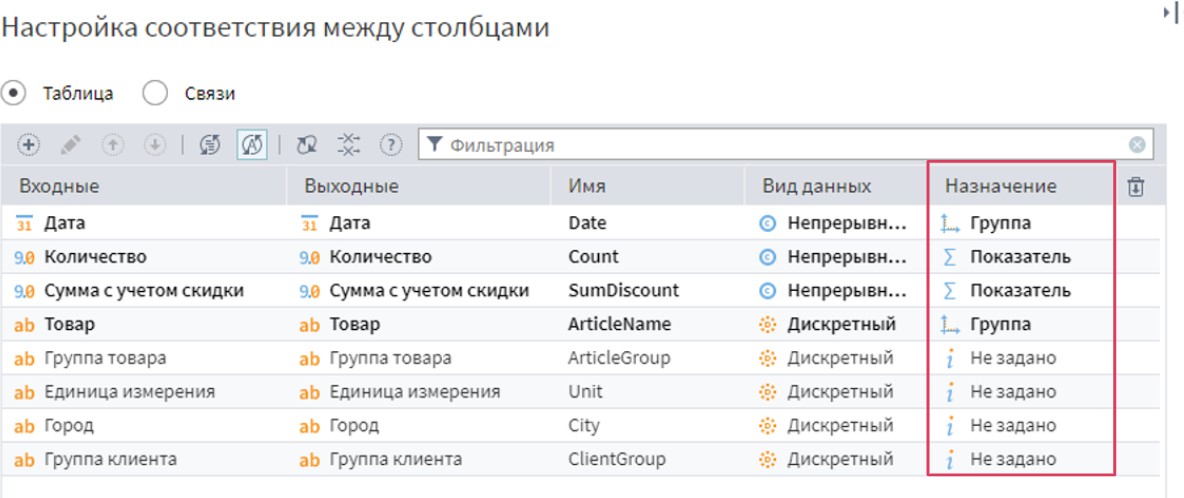
Оставим настройки без изменений и перейдем к просмотру результатов обработки.

Получили набор данных по продажам товаров по дням. Обратите внимание, что при группировке мы потеряли информацию о значениях показателей в разрезе исключенных групп.



На вход узла можно подать поля с назначением, настроенным заранее (например, с помощью компонента **Параметры полей**). Кроме того назначение можно изменить непосредственно на входном порте.

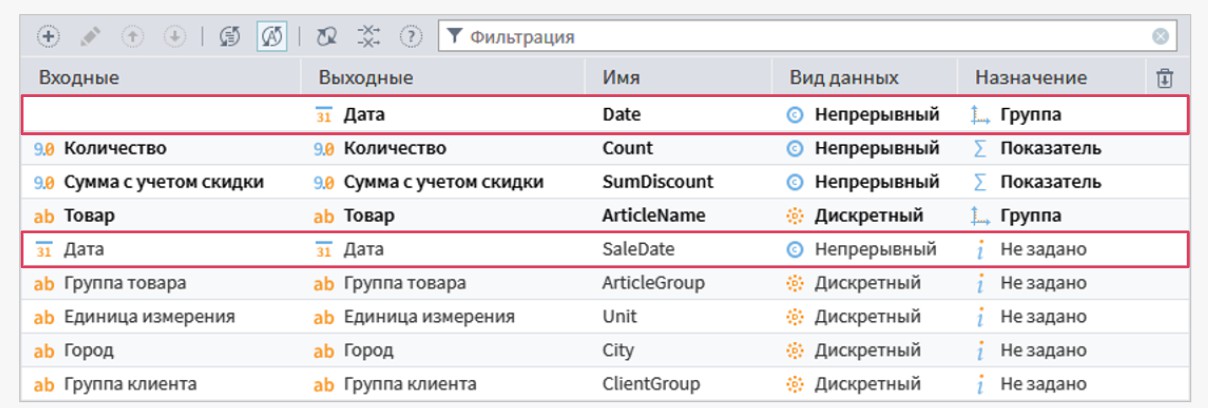
Перейдем в настройку входного порта узла **Продажи по дням**. После того, как узел настроен, назначение полей во входном порте автоматически меняется в соответствии с заданными настройками. Чтобы изменить этот параметр необходимо дважды щелкнуть на нужном поле левой кнопкой мыши.



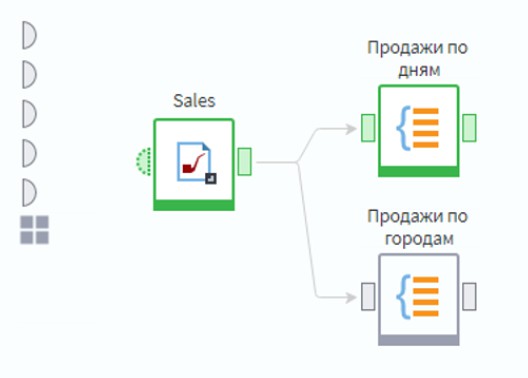
# Назначение полей и автосинхронизация

Обратите внимание, что при включенной автосинхронизации работать с узлом группировки нужно с осторожностью – если имена полей в новых входных данных не совпадут с заданными на входе узла, новые поля, даже если у них настроено

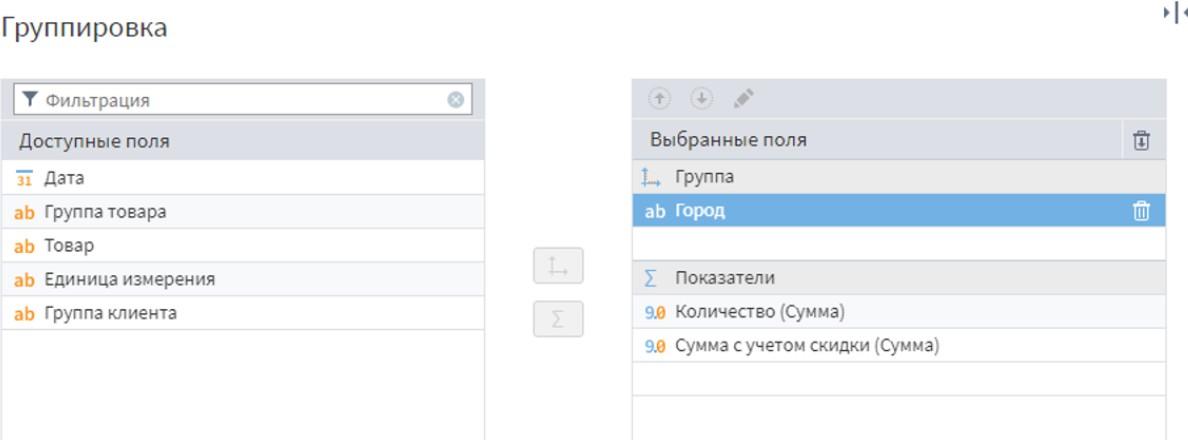
назначение, не свяжутся с имеющимися, а добавятся к ним.

Рассмотрим пример. Если в новом наборе поле **Дата** будет иметь имя **SaleDate** вместо **Date**, в таблице на входном порте будет два поля: входное поле с именем **SaleDate** будет связано с выходным полем с именем **SaleDate**, а для поля **Date** входного поля в таблице не будет, что приведет к ошибке.

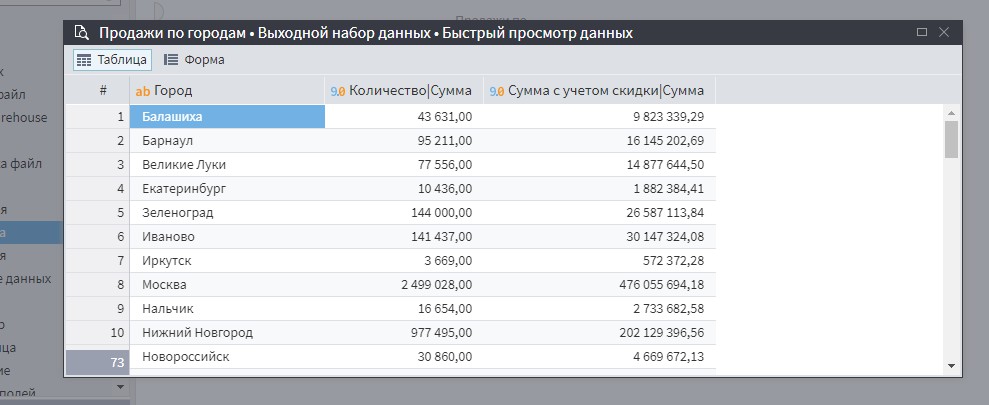
Добавим в сценарий еще один узел группировки и переименуем его, как показано на рисунке:



В новом узле настроим группировку продаж по сумме и количеству по городам.

Для этого нам понадобится указать в списке групп только одно поле – **Город**, а показатели остаются такими же.

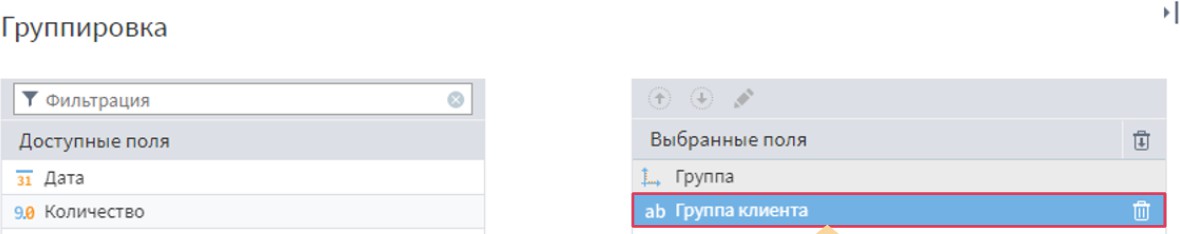
На выходе мы получили общие продажи по каждому из 73 городов.



Иногда требуется получить список всех значений какого-либо поля. В этом случае можно провести группировку без показателей.

Обратите внимание: если выбрать несколько полей, то сохранится иерархия, которая присутствовала в исходных данных.

Проведем группировку по полю **Группа клиента**.

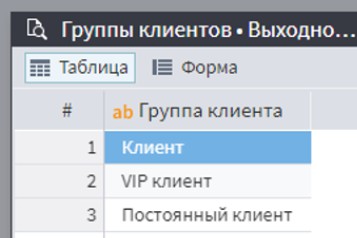


Получили список всех существующих в нашем наборе групп клиентов.

Можно также использовать группировку без измерений, тогда мы получим агрегированное значение по выбранному полю.

Отметим, что для этой операции рекомендуется использовать компонент

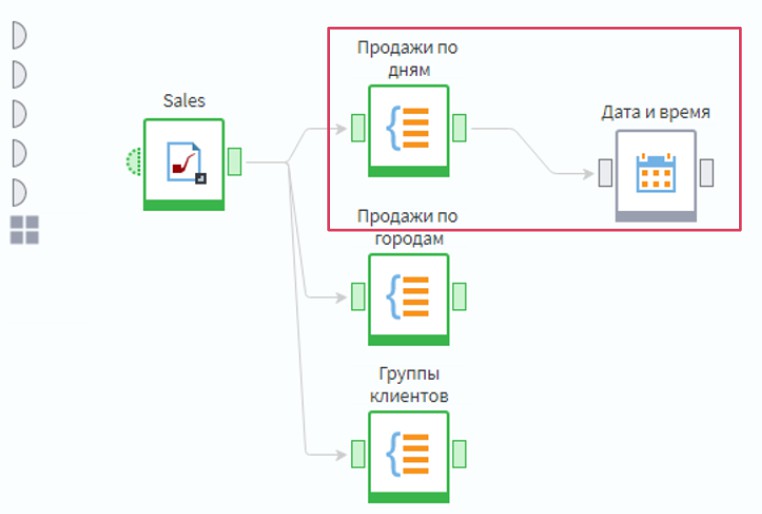
# Таблица в переменные.

****

**Компонент Дата и Время**

**Практическое задание №4**

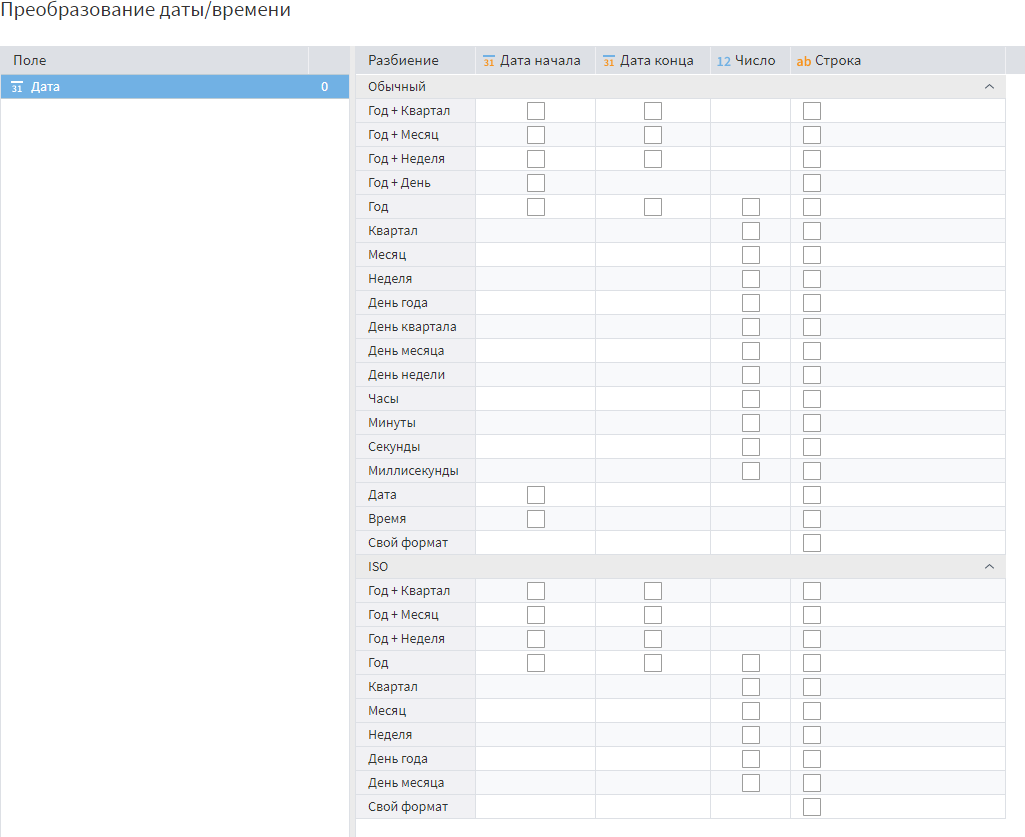
Теперь добавим в наш сценарий узел **Дата и время** и подадим на его вход данные с узла группировки **Продажи по дням**. По сути эти данные – временные ряды продаж товаров.

****Пусть нам необходимо получить продажи по месяцам года. Очевидно, что для этого нужно сделать группировку, но поля с месяцем продажи в нашем наборе нот. Однако его легко получить из поля **Дата** с помощью узла **Дата и Время.**

Данный узел позволяет провести следующее преобразование: на основе поля типа **Дата/Время** формируется одно или несколько дополнительных полей, в которых указывается, к какому заданному интервалу времени (год, квартал и так далее) принадлежит запись данных.

Перейдем в настройку. Перед нами окно задания параметров преобразования. Слева расположена область **Поле**, в нее попадают только те поля входного набора данных, которые имеют тип **Дата/Время**.

Справа находится область **Разбиение**, где можно выбрать формат и тип данных, которые мы хотим получить на выходе. По умолчанию ничего не выбрано.

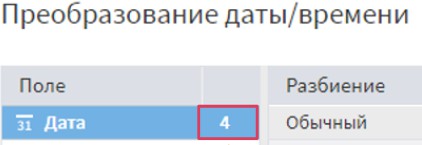


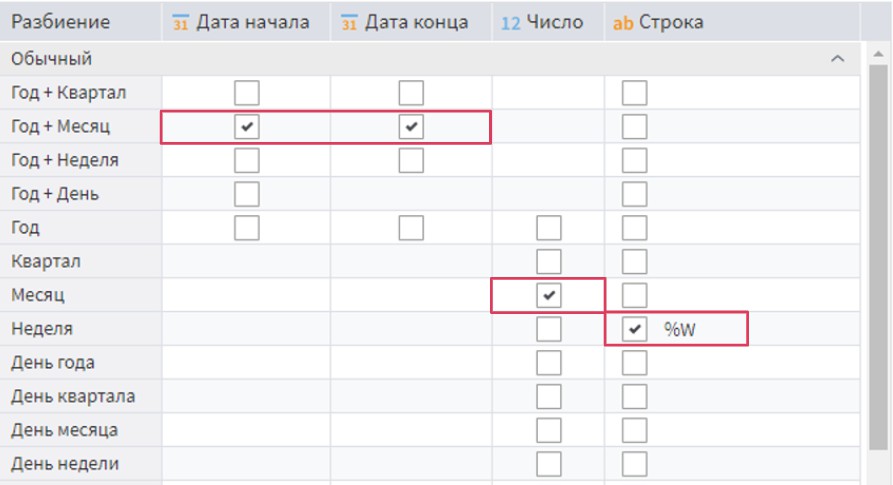
Нам необходимо поле с месяцем года, поэтому выберем разбиение **Год+Месяц** с типом **Дата/Время** (выберем оба варианта: **Дата начала** и **Дата конца**).

Дополнительно выделим из поля **Дата** номер месяца и неделю продажи с типами

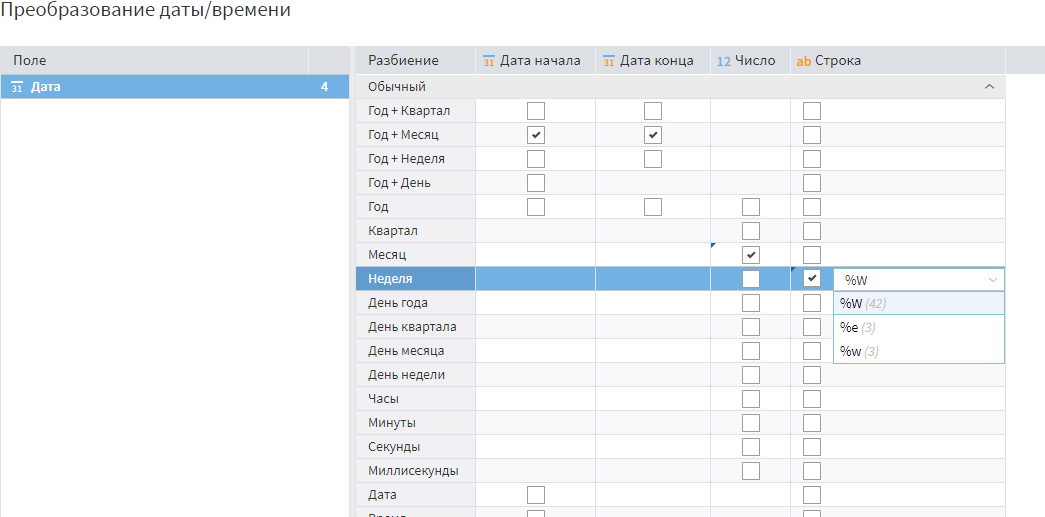
**Число** и **Строка** соответственно.

Каждый выбранный вариант преобразования будет являться отдельным полем выходного набора данных. Возле каждого поля в области списка полей считается количество выбранных для него вариантов.

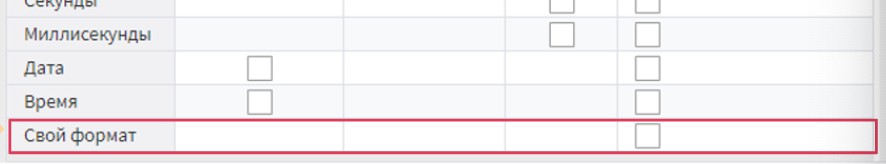




При выборе типа данных **Строка** в выпадающем списке справа от установленного флага можно выбрать формат представления даты в строковом виде.



Кроме того для строкового типа можно задать произвольный формат при помощи строки форматирования.



# Строка форматирования даты и времени

Формат представления даты в строковом виде задается с помощью специальных маркеров.

Рассмотрим маркеры, доступные для представления недели:

* **%W** – номер недели в году;
* **%е** – номер недели в квартале;
* **%w** – номер недели в месяце.

При задании пользовательского формата можно использовать сочетание произвольного текста и этих маркеров. Например, для даты **01.01.2018** можно задать значение:

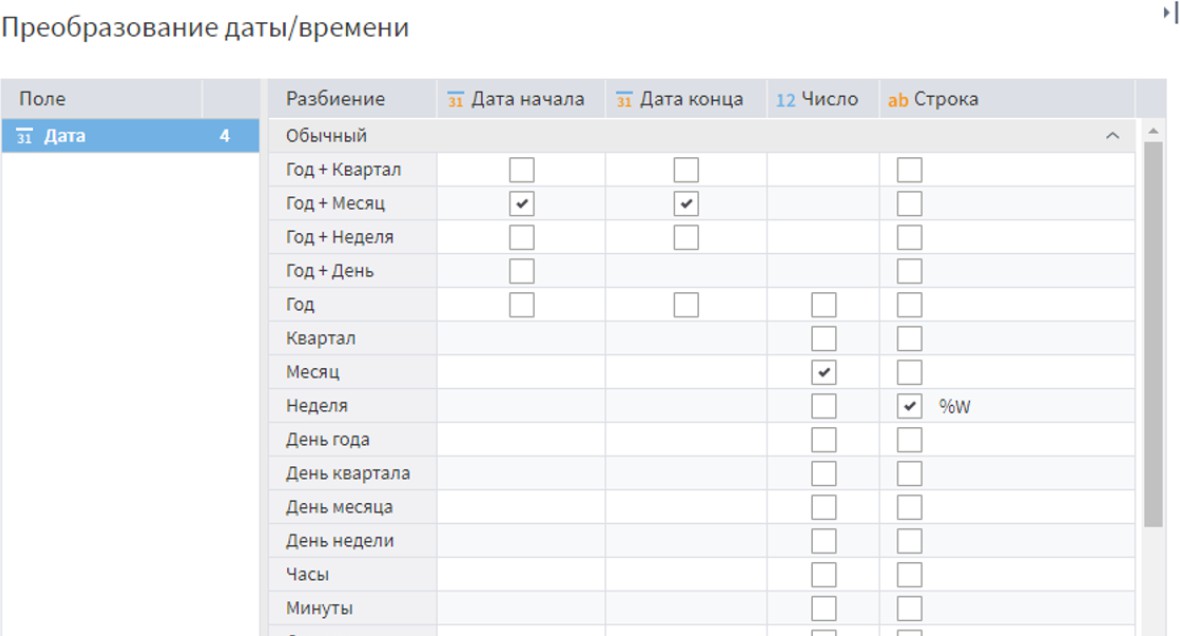
# произвольный текст % W, произвольный текст

В результате которого на выходе мы получим:

# произвольный текст 1, произвольный текст

Количество маркеров в тексте не ограничено. С полным списком маркеров можно ознакомиться в справке по **Loginom**.

Итак, мы выбрали четыре варианта преобразования. Посмотрим, что мы получим на выходе.

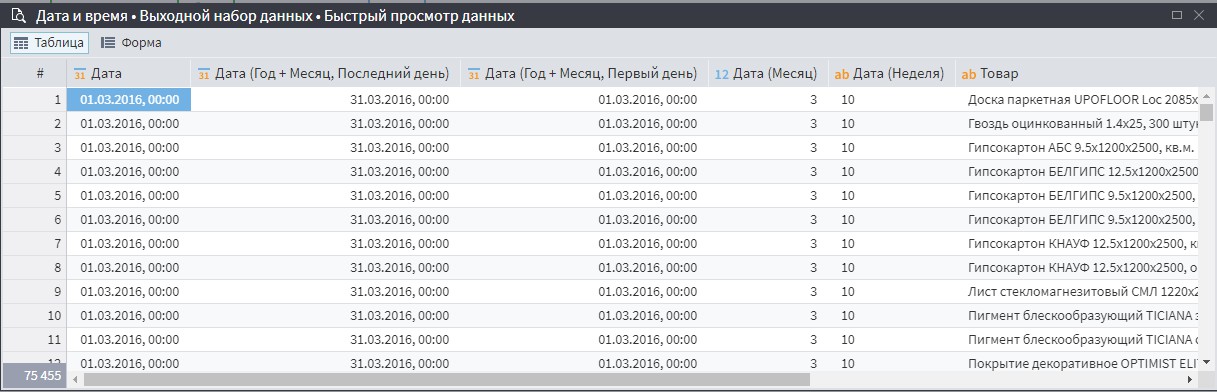


В наш набор данных добавились четыре новых поля, которые характеризуют поле

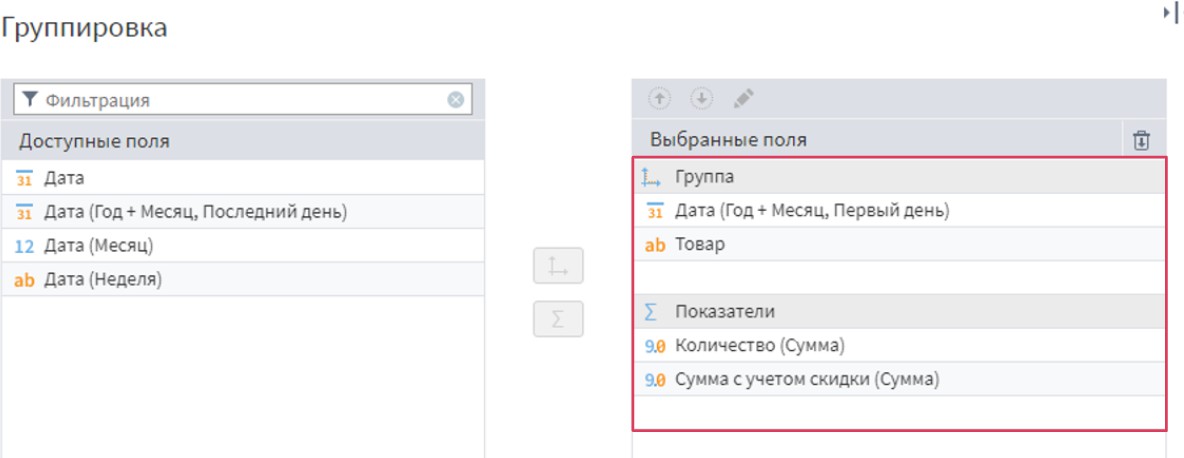
# Дата.

Обратите внимание, что поля **Дата** (**Год+Месяц**, **Последний день**) и **Дат**а

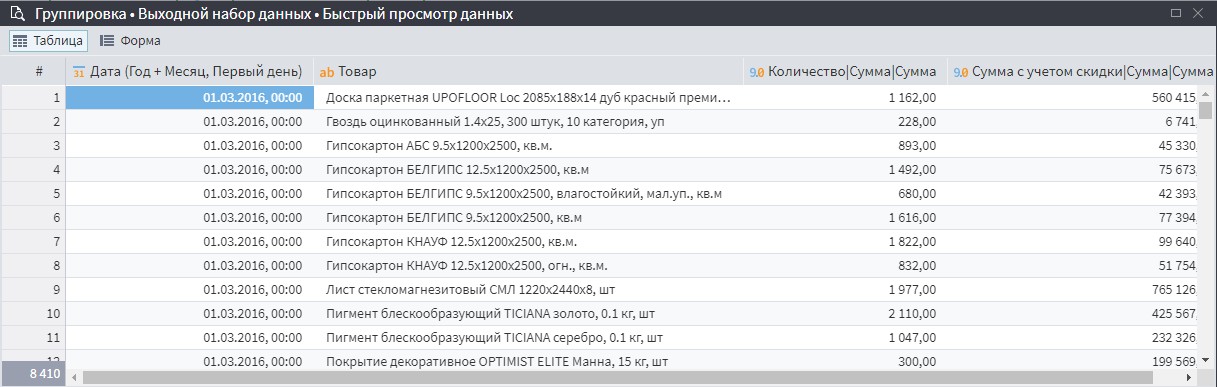
(**Год+Месяц**, **Первый день**) обозначают один и тот же период, только в одном случае указывается дата конца этого периода, а в другом – дата начала.



Теперь мы можем сгруппировать продажи по месяцам года.



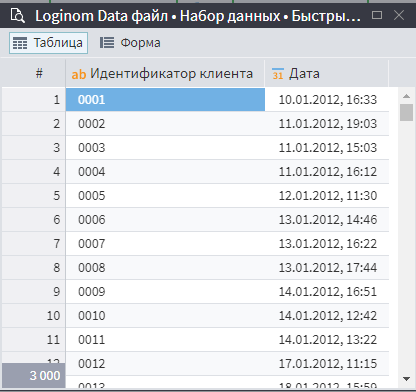
В результате количество записей сильно сократилось (до 8 410), так как мы получили продажи по месяцам.



# Форматирование даты по ISO

Компонент **Дата и время** позволяет также получить интервал времени, к которому принадлежит запись данных, в формате, определенном международным

стандартом **ISO 8601**.

Новый период (год, квартал, месяц) в данном формате всегда начинается с понедельника и не всегда совпадает с календарным. Первой неделей года считается та, которая содержит первый четверг в году по Григорианскому календарю. Аналогичным образом определяется первая неделя квартала и месяца.

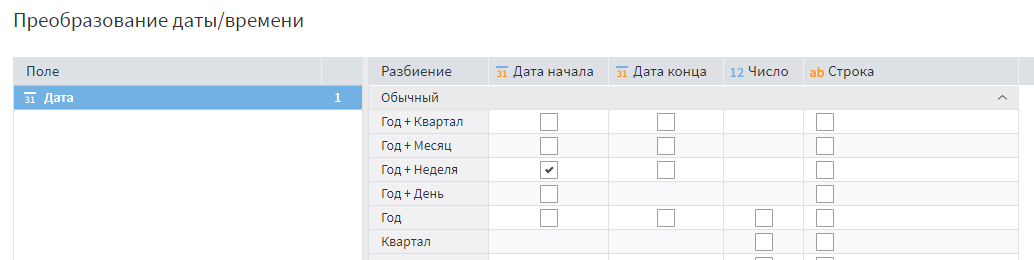
Использование данного формата при группировке позволяет не делить данные на стыке периодов, например, лет.

Рассмотрим, как это происходит с помощью набора данных **transactions.lgd.**

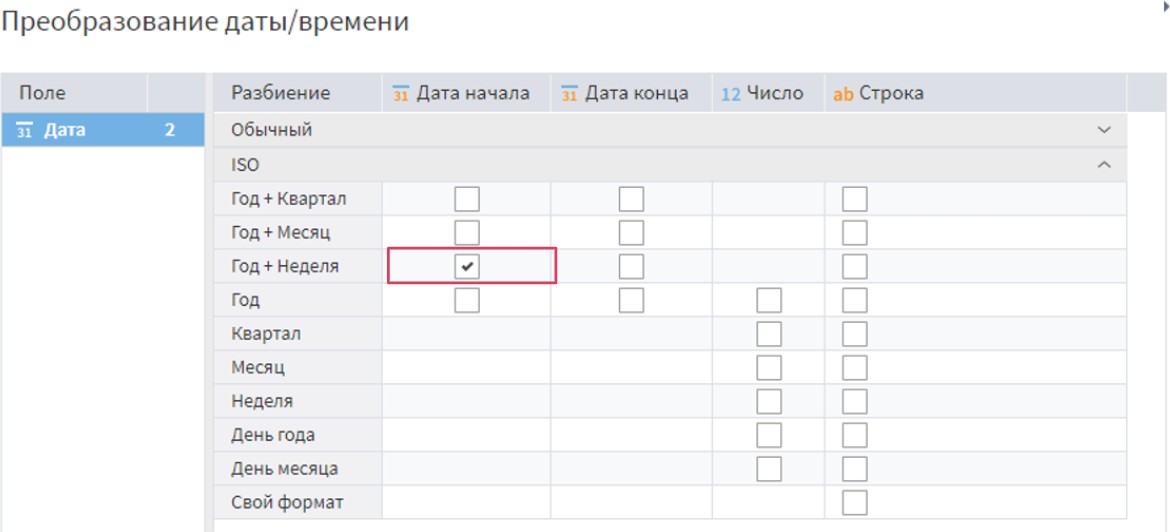
Импортируем указанный набор. Он содержит 3 000 записей о клиентах магазина и датах совершенных ими транзакций за несколько лет.

Пусть нам необходимо получить информацию о том, какое количество клиентов в неделю совершало транзакции за эти годы.

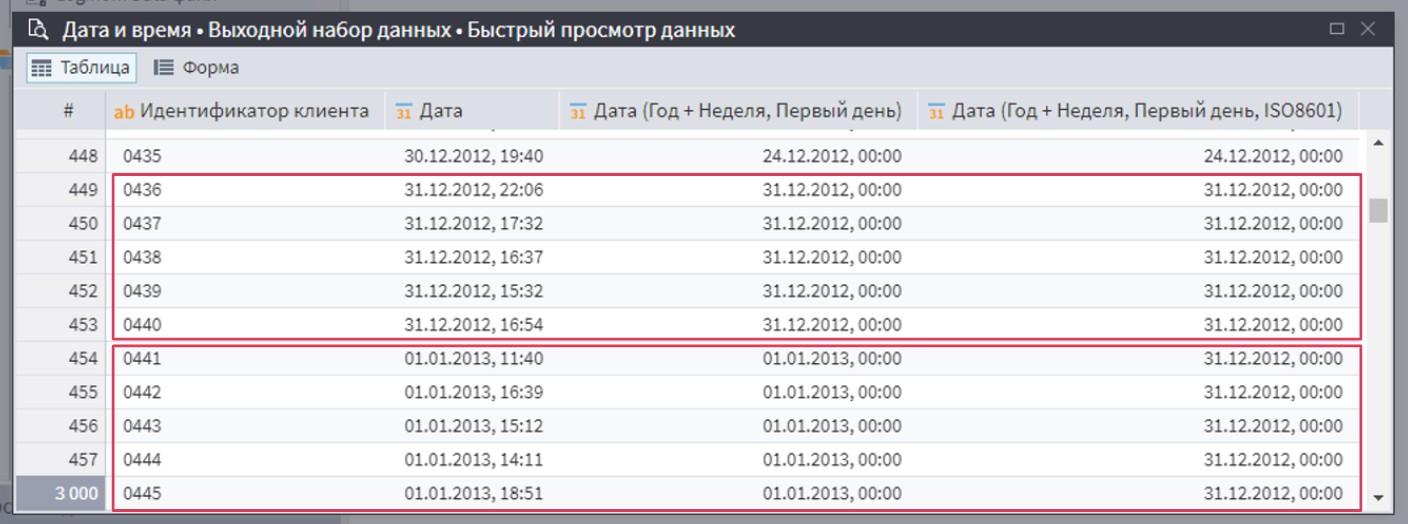
Для начала нам понадобится выделить из дат информацию о неделе транзакции.

Для этого добавим в сценарий узел **Дата и время** и перейдем в его настройку. Выберем вариант **Год+ Неделя** на вкладке **Обычный** и свернем ее.

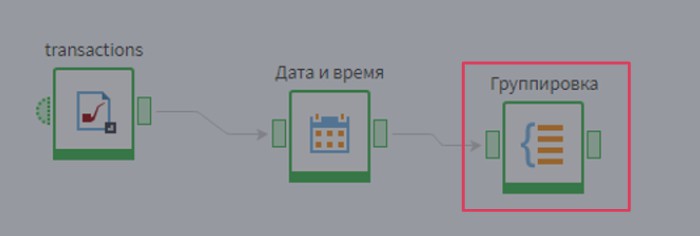
Перед нами варианты интервалов времени, которые можно представить в формате **ISO**. Выберем такое же значение – **Год+Неделя** – и перейдем к просмотру результатов.



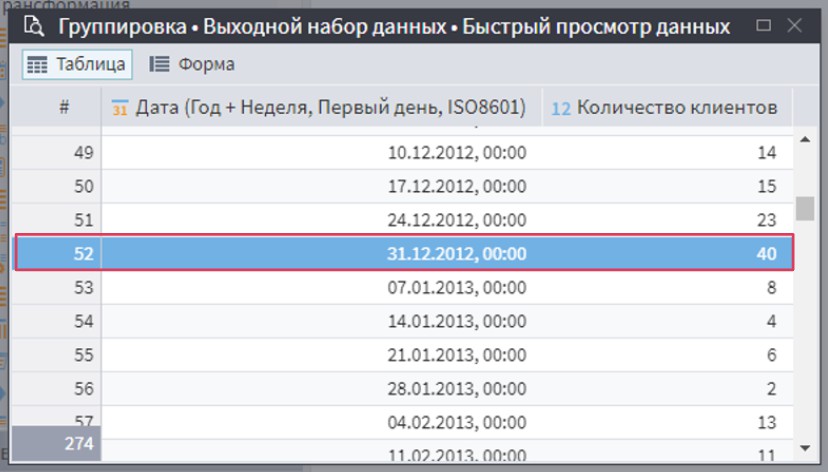
Для того чтобы увидеть разницу, перейдем к 449-ой записи выходного набора. 31 декабря 2012 года понедельник, для клиентов из записей 449 453, совершивших транзакции в этот день, в обоих форматах: обычном и ISO, – указана одна и та же дата начала недели – 31.12.2012.

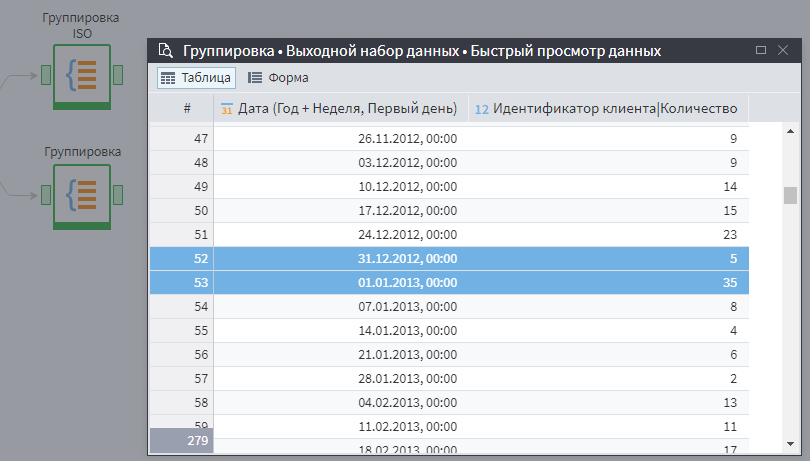
Однако, начиная с 454-ой строки идут записи о клиентах, совершивших транзакции 1 января 2013 года. В обычном формате эти клиенты относятся уже к другой неделе – первой неделе нового года. начинающейся со вторника. Однако в формате **ISO** они по-прежнему относятся к неделе, которая началась 31 декабря 2012 года, и именно она является первой неделей 2013 года по **ISO**.

Проведем группировку, где в список групп добавим поле **Дата** (**Год+Неделя**,

**Первый день**, **ISO8601**), а в список показателей – поле **Идентификатор клиента** с вариантом агрегации **Количество**, после чего посмотрим выходной набор данных.

Перейдя к 52-ой строке, мы видим, что в неделю, начавшуюся 31.12.2012 года, транзакции совершили 40 клиентов.



Если же мы проводим группировку по неделям в обычном формате, за тот же период мы получаем две записи: 5 клиентов попадают в последнюю недолю 2012-го года с 31.12.2012 по 31.12.2012, а еще 35 – в первую неделю 2013 года С 01.01.2013 по 06.01.2013.