

# **Практическое занятие 5**

## **ETL-операции с набором данных**

Модуль 1 «Введение в аналитику данных»

# Доступ к платформе

1. Loginom Community Edition на сайте


[www.loginom.ru/download](http://www.loginom.ru/download)

2. Облачный доступ на демо-сервере

[demo.loginom.ru](http://demo.loginom.ru).

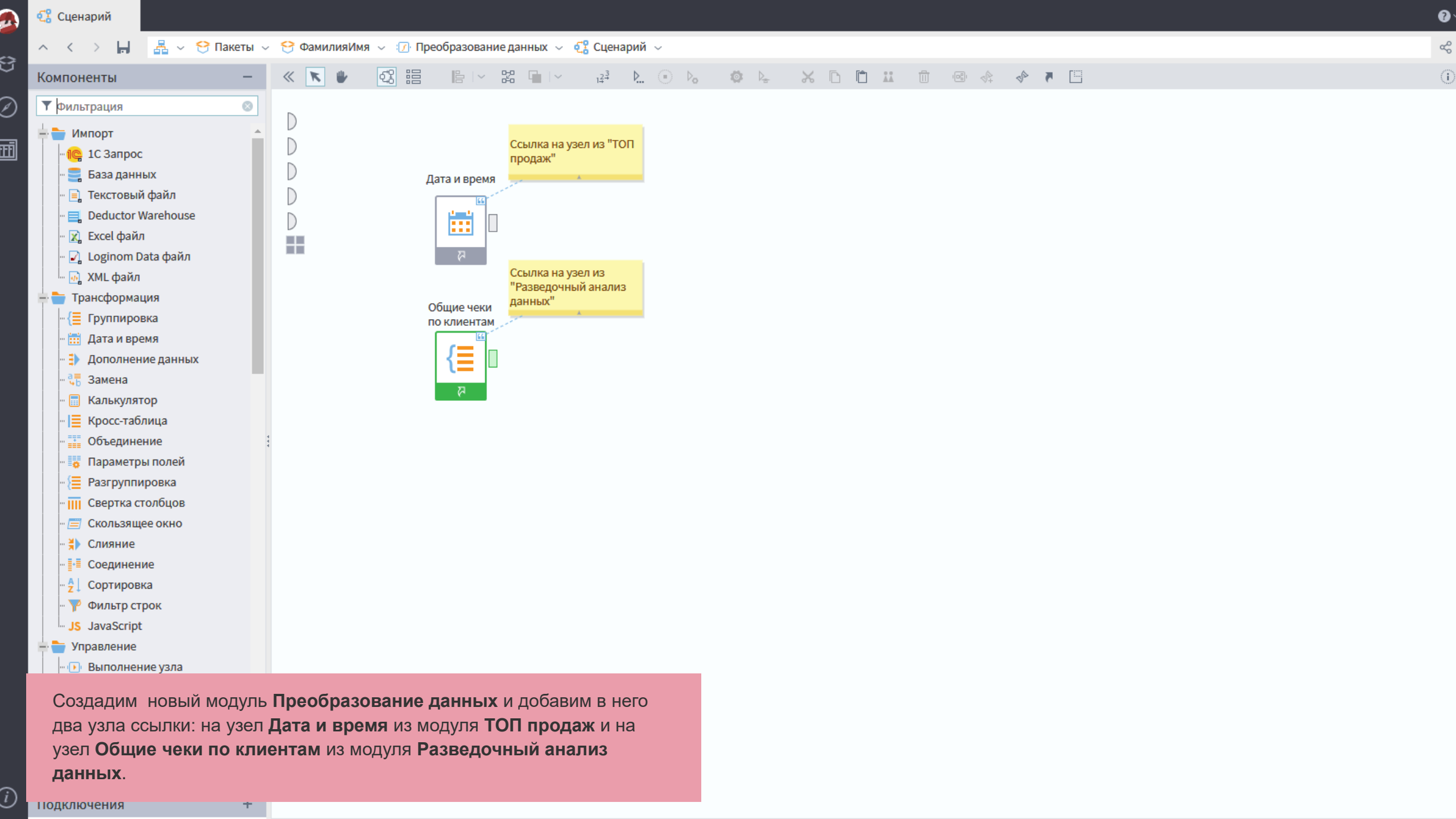
- Логин и пароль – fin\_academy. Сервер публичный, скорость работы ограничена.

# Порядок выполнения

1. Убедитесь, что вы скачали файл с набором данных **transactions\_diy.lgd**
  2. Повторяйте все действия по пошаговым скриншотам и выполняйте задания. Ответы на вопросы вам понадобятся при заполнении контрольного теста по практическому занятию. Вы должны ответить **верно** на все вопросы.
- 

# Скользящее окно



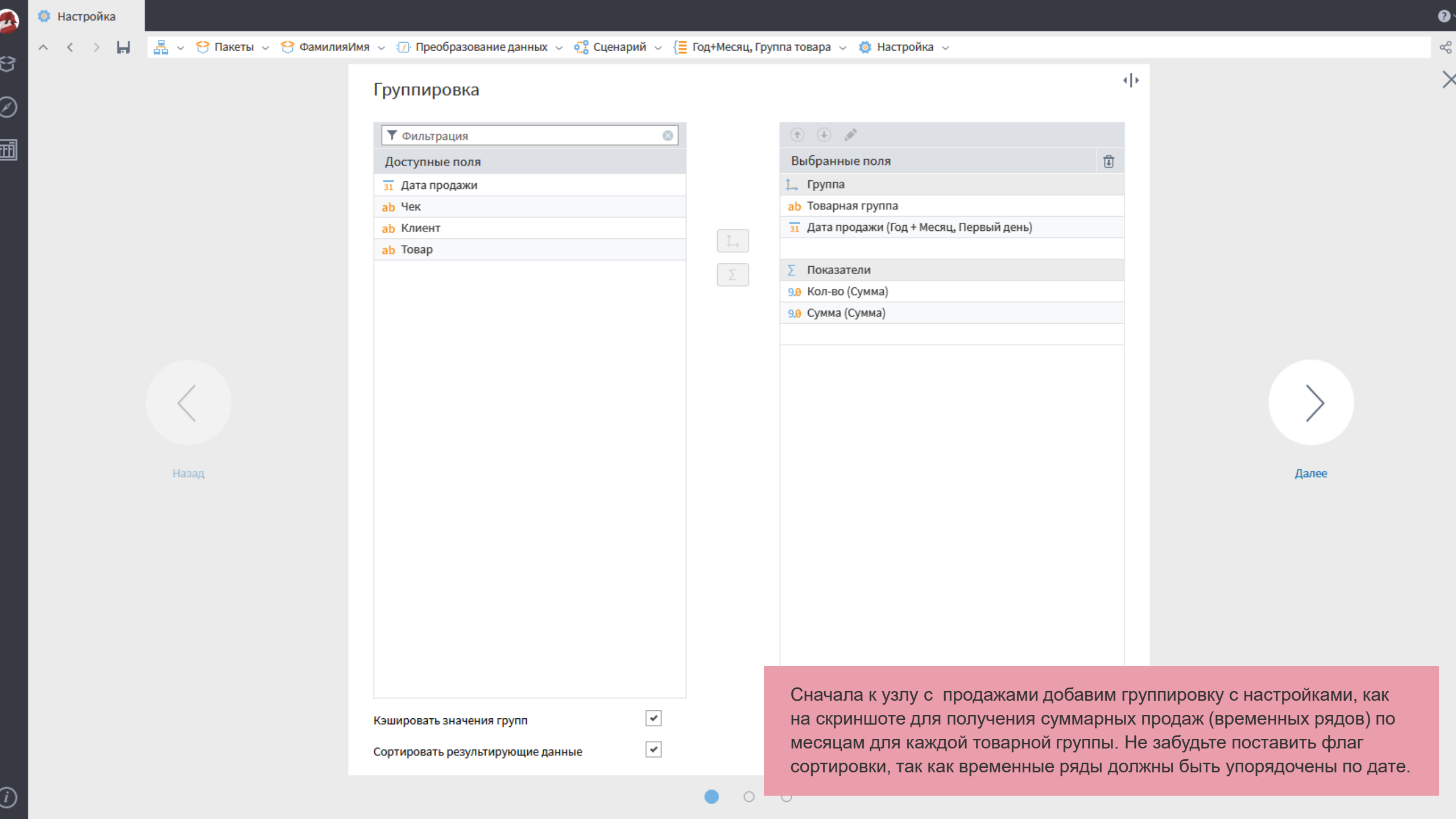


Создадим новый модуль **Преобразование данных** и добавим в него два узла ссылки: на узел **Дата и время** из модуля **ТОП продаж** и на узел **Общие чеки по клиентам** из модуля **Разведочный анализ данных**.

# Скользящее окно

Сначала покажем применение скользящего окна на примере решения задачи построения прогноза спроса на товарную группу по «наивной» модели скользящего среднего: величина спроса равна средним продажам в штуках за последние три месяца.





Настройка

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Год+Месяц, Группа товара

Выходные порты

Выходной набор данных

Настройка

Настройка выходных столбцов

Таблица

Связи

Фильтрация

Входные	Выходные	Имя	Вид данных	Назначение	
ab Товарная группа	ab Товарная группа	ItemGroup	Дискретный	Не задано	
31 Дата продажи (Год + ...	31 Дата (Год + Месяц )	TrDte_YM	Непрерывн...	Не задано	
9.0 Кол-во Сумма	9.0 Кол-во	Quantity	Непрерывн...	Не задано	
9.0 Сумма Сумма	9.0 Сумма	Amount	Непрерывн...	Не задано	

Назад

Просмотр

Сохранить

Выполнить

Для удобства в выходном порту группировки переименуем метки полей на более читабельные. Не забудьте отключить автосинхронизацию перед переименованием.



Сценарий

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Компоненты

Фильтрация

Импорт

- 1С Запрос
- База данных
- Текстовый файл
- Deductor Warehouse
- Excel файл
- Loginom Data файл
- XML файл

Трансформация

- Группировка
- Дата и время
- Дополнение данных
- Замена
- Калькулятор
- Кросс-таблица
- Объединение
- Параметры полей
- Разгруппировка
- Свертка столбцов
- Скользящее окно
- Слияние
- Соединение
- Сортировка
- Фильтр строк
- JavaScript

Управление

- Выполнение узла
- Подмодель
- Узел-ссылка

Дата и время

Ссылка на узел из "ТОП продаж"

Год+Месяц, Группа товара

Скользящее окно

Общие чеки по клиентам

Ссылка на узел из "Разведочный анализ данных"

Поскольку для расчета прогноза нам нужно три последних месяца, нам потребуется компонент **Скользящее окно**.

Подключения

Настройка

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Скользящее окно

Настройка

Скользящее окно

Столбец	Глубина истории	Горизонт прогноза
ab Товарная группа	Не выбрана	Не выбран
9.0 Кол-во	3	Не выбран
9.0 Сумма	Не выбрана	Не выбран
31 Дата (Год + Месяц )	Не выбрана	Не выбран

Назад

Далее

Удалять все неполные записи

Для поля **Кол-во** зададим глубину истории 3.

Сценарий

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Компоненты

Фильтрация

Импорт

- 1С Запрос
- База данных
- Текстовый файл
- Deductor Warehouse
- Excel файл
- Loginom Data файл
- XML файл

Трансформация

- Группировка
- Дата и время
- Дополнение данных
- Замена
- Калькулятор
- Кросс-таблица
- Объединение
- Параметры полей
- Разгруппировка
- Свертка столбцов
- Скользящее окно
- Слияние
- Соединение
- Сортировка
- Фильтр строк
- JavaScript

Управление

- Выполнение узла
- Подмодель
- Узел-ссылка

Скользящее окно • Выходной набор данных • Быстрый просмотр данных

#	ab Товарная группа	9.0 Кол-во[-3]	9.0 Кол-во[-2]	9.0 Кол-во[-1]	9.0 Кол-во	9.0 Сумма	31 Дата (Год + Месяц)
1	Ванная комната	326,00	413,00	510,00	399,00	986 907,00	01.12.2017, 0
2	Ванная комната	413,00	510,00	399,00	536,00	1 412 932,00	01.01.2018, 0
3	Ванная комната	510,00	399,00	536,00	624,00	1 474 027,00	01.02.2018, 0
4	Ванная комната	399,00	536,00	624,00	526,00	1 094 837,00	01.03.2018, 0
5	Ванная комната	536,00	624,00	526,00	494,00	1 219 927,00	01.04.2018, 0
6	Ванная комната	624,00	526,00	494,00	517,00	1 123 003,00	01.05.2018, 0
7	Ванная комната	526,00	494,00	517,00	363,00	889 116,00	01.06.2018, 0
8	Ванная комната	494,00	517,00	363,00	624,00	1 513 103,00	01.07.2018, 0
9	Ванная комната	517,00	363,00	624,00	1 300,00	2 585 017,00	01.08.2018, 0
10	Ванная комната	363,00	624,00	1 300,00	727,00	1 694 683,00	01.09.2018, 0
11	Ванная комната	624,00	1 300,00	727,00	757,00	1 874 162,00	01.10.2018, 0
12	Ванная комната	1 300,00	727,00	757,00	814,00	2 068 176,00	01.11.2018, 0
13	Ванная комната	727,00	757,00	814,00	954,00	2 329 715,00	01.12.2018, 0
14	Ванная комната	757,00	814,00	954,00	733,00	1 784 480,00	01.01.2019, 0
15	Ванная комната	814,00	954,00	733,00	591,00	1 644 751,00	01.02.2019, 0
16	Ванная комната	954,00	733,00	591,00	765,00	1 698 453,00	01.03.2019, 0
17	Ванная комната	733,00	591,00	765,00	810,00	2 058 760,00	01.04.2019, 0
18	Ванная комната	591,00	765,00	810,00	832,00	2 155 023,00	01.05.2019, 0
19	Ванная комната	765,00	810,00	832,00	845,00	2 043 836,00	01.06.2019, 0
20	Ванная комната	810,00	832,00	845,00	1 094,00	3 146 688,00	01.07.2019, 0
21	Ванная комната	832,00	845,00	1 094,00	1 347,00	3 157 052,00	01.08.2019, 0
22	Ванная комната	845,00	1 094,00	1 347,00	1 012,00	2 862 295,00	01.09.2019, 0
439	Ванная комната	1 094,00	1 347,00	1 012,00	964,00	2 546 617,00	01.10.2019, 0

Посмотрим, что получилось. В нашем наборе данных после скользящего окна появилось три новых столбца с предыдущими значениями количества продаж.

Сценарий

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Компоненты

Фильтрация

Импорт

- 1С Запрос
- База данных
- Текстовый файл
- Deductor Warehouse
- Excel файл
- Loginom Data файл
- XML файл

Трансформация

- Группировка
- Дата и время
- Дополнение данных
- Замена
- Калькулятор
- Кросс-таблица
- Объединение
- Параметры полей
- Разгруппировка
- Свертка столбцов
- Скользящее окно
- Слияние
- Соединение
- Сортировка
- Фильтр строк
- JavaScript

Управление

- Выполнение узла
- Подмодель
- Узел-ссылка

Дата и время

Ссылка на узел из "ТОП продаж"

Год+Месяц, Группа товара

Скользящее окно

Прогноз по среднему

Общие чеки по клиентам

Ссылка на узел из "Разведочный анализ данных"

Калькулятор

Выражения	Имя	Метка
12 Forecast	Прогноз	

Поля/Переменные

Фильтрация

Имя

Метка

Поля

Quantity\_H\_3

Кол-во[-3]

Quantity\_H\_2

Кол-во[-2]

Quantity\_H\_1

Кол-во[-1]

ItemGroup

Товарная группа

Quantity

Кол-во

Amount

Сумма

TrDte\_YM

Дата (Год + Месяц)

Список функций

Фильтрация

Категории

Abs

(Аргумент)

AbsErr

(Аргумент1, Аргумент2)

AddDay

(Дата, Количество)

AddMonth

(Дата, Количество)

AddQuarter

(Дата, Количество)

AddWeek

(Дата, Количество)

AddYear

(Дата, Количество)

AMGD

(Стоимость, Остаточная\_с...

ArcCos

(Значение)

Теперь компонентом **Калькулятор** напомним формулу прогноза по среднему по трем последним месяцам. Новое поле назовем **Прогноз (Forecast)** и оно будет целого типа.

Сценарий

Пакеты

ФамилияИмя

Преобразование данных

Сценарий

Компоненты

Прогноз по среднему • Выходной набор данных • Быстрый просмотр данных

#	12 Прогноз	9.0 Кол-во[-3]	9.0 Кол-во[-2]	9.0 Кол-во[-1]	ab Товарная группа	9.0 Кол-во	9.0 Сумма	31 Дат
1	416	326,00	413,00	510,00	Ванная комната	399,00	986 907,00	
2	441	413,00	510,00	399,00	Ванная комната	536,00	1 412 932,00	
3	482	510,00	399,00	536,00	Ванная комната	624,00	1 474 027,00	
4	520	399,00	536,00	624,00	Ванная комната	526,00	1 094 837,00	
5	562	536,00	624,00	526,00	Ванная комната	494,00	1 219 927,00	
6	548	624,00	526,00	494,00	Ванная комната	517,00	1 123 003,00	
7	512	526,00	494,00	517,00	Ванная комната	363,00	889 116,00	
8	458	494,00	517,00	363,00	Ванная комната	624,00	1 513 103,00	
9	501	517,00	363,00	624,00	Ванная комната	1 300,00	2 585 017,00	
10	762	363,00	624,00	1 300,00	Ванная комната	727,00	1 694 683,00	
11	884	624,00	1 300,00	727,00	Ванная комната	757,00	1 874 162,00	
12	928	1 300,00	727,00	757,00	Ванная комната	814,00	2 068 176,00	
13	766	727,00	757,00	814,00	Ванная комната	954,00	2 329 715,00	
14	842	757,00	814,00	954,00	Ванная комната	733,00	1 784 480,00	
15	834	814,00	954,00	733,00	Ванная комната	591,00	1 644 751,00	
16	759	954,00	733,00	591,00	Ванная комната	765,00	1 698 453,00	
17	696	733,00	591,00	765,00	Ванная комната	810,00	2 058 760,00	
18	722	591,00	765,00	810,00	Ванная комната	832,00	2 155 023,00	
19	802	765,00	810,00	832,00	Ванная комната	845,00	2 043 836,00	
20	829	810,00	832,00	845,00	Ванная комната	1 094,00	3 146 688,00	
21	924	832,00	845,00	1 094,00	Ванная комната	1 347,00	3 157 052,00	
22	1 095	845,00	1 094,00	1 347,00	Ванная комната	1 012,00	2 862 295,00	
439	1 151	1 094,00	1 347,00	1 012,00	Ванная комната	964,00	2 546 617,00	

Закреть

Скользящее окно

Прогноз по среднему

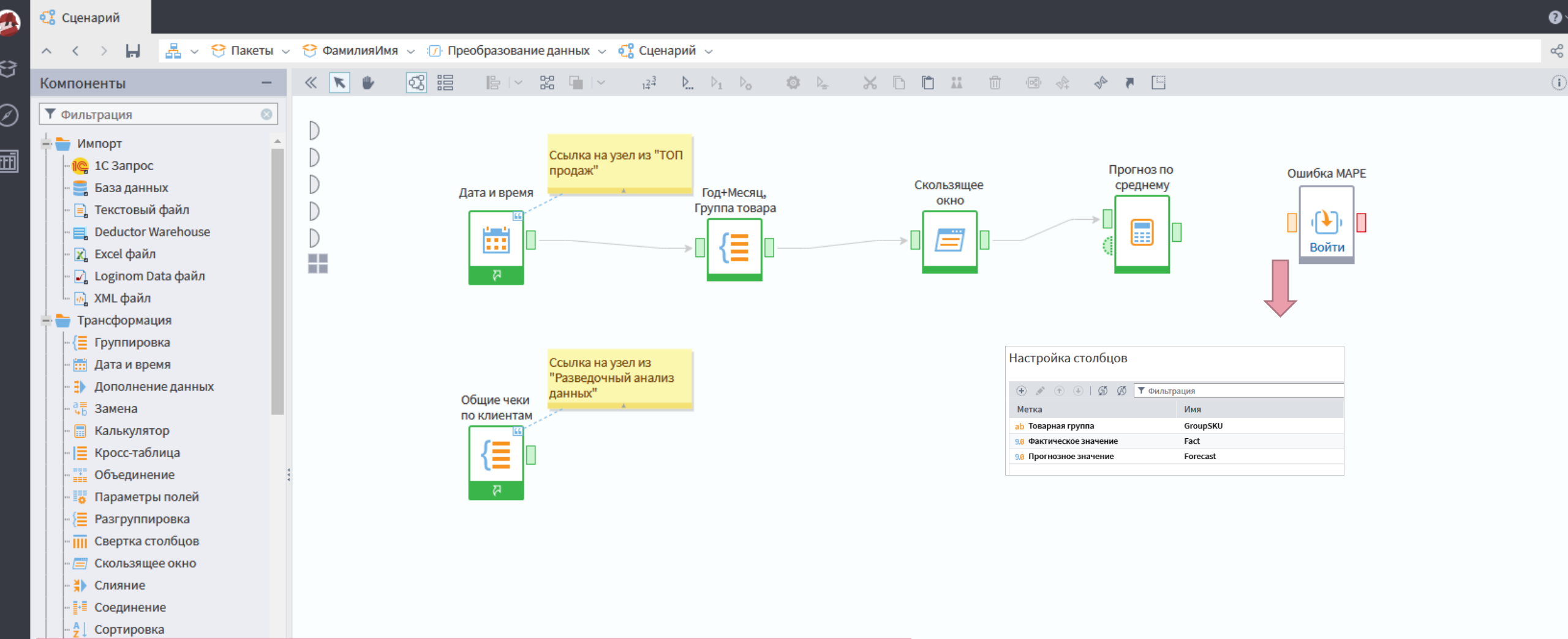
Мы получили прогноз спроса, но это на самом деле ретропрогноз, так как фактическое значение, какое было продано, нам известно и оно находится в поле **Кол-во**. Поэтому мы можем рассчитать ошибку прогнозирования, которая дает модель, сравнив прогноз с фактом.

# Ошибка прогнозирования

Усредненная ошибка MAPE считается по формуле ( $y^*$  – прогнозное значение,  $y$  – фактическое значение,  $n$  – число наблюдений):

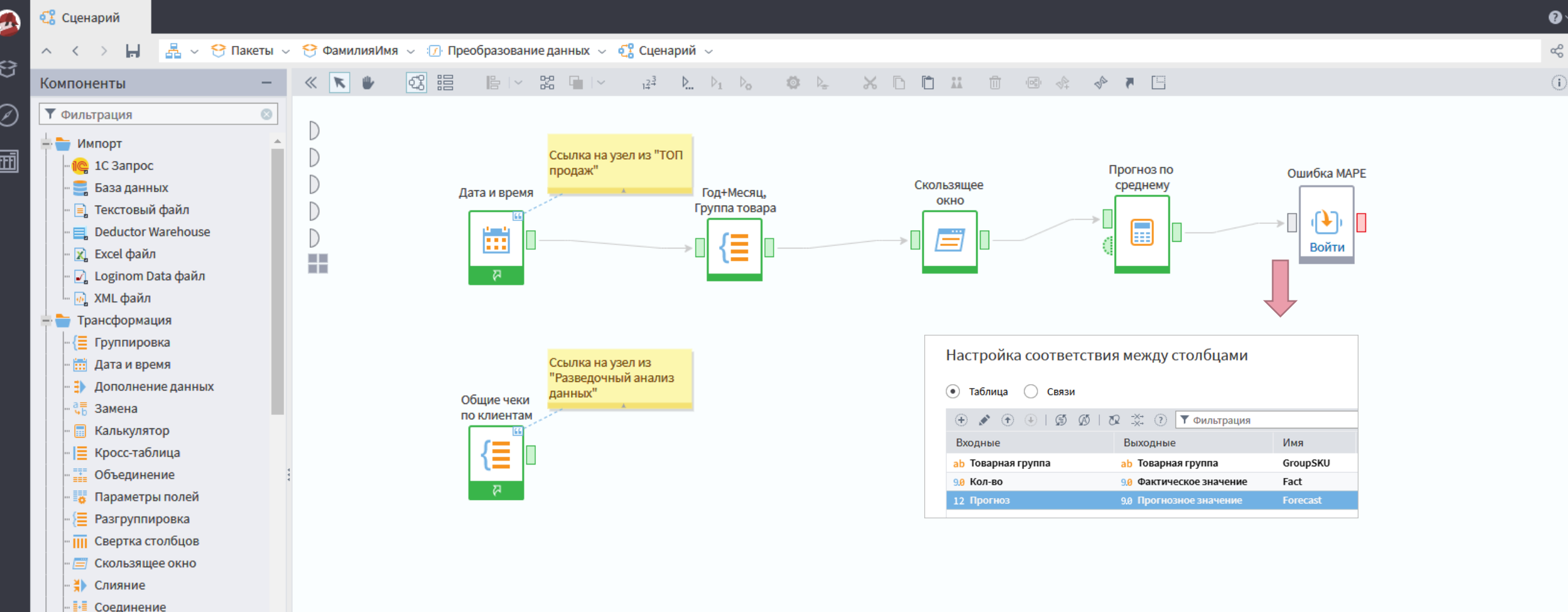
$$\frac{\sum \frac{|y^* - y|}{y}}{n}$$

В нашем случае ошибку MAPE лучше считать для каждой товарной группы в отдельности.



Рассчитывать ошибку прогноза будем в отдельной подмодели. На вход она будет требовать набор данных с тремя полями **Товарная группа**, **Фактическое значение** и **Прогнозное значение**.

Не соединяя подмодель **Ошибка MAPE** с калькулятором, зададим в настройках портов требуемые входы и выходы. То есть мы сначала формируем требования к входным данным подмодели (так называемый подход к проектированию логики обработки без данных).



Соединим узел **Прогноз по среднему** и **Ошибка MAPE** и в настройках входного порта поставим соответствия между входами и выходами.

Теперь осталось реализовать саму логику расчета ошибки MAPE внутри подмодели. Это необходимо сделать самостоятельно. В итоге вы получите список ошибок MAPE для каждой из товарных групп.

Далее постройте настоящий прогноз с учетом самого последнего имеющегося в данных месяца и получите прогнозные цифры спроса для каждой из товарных групп.

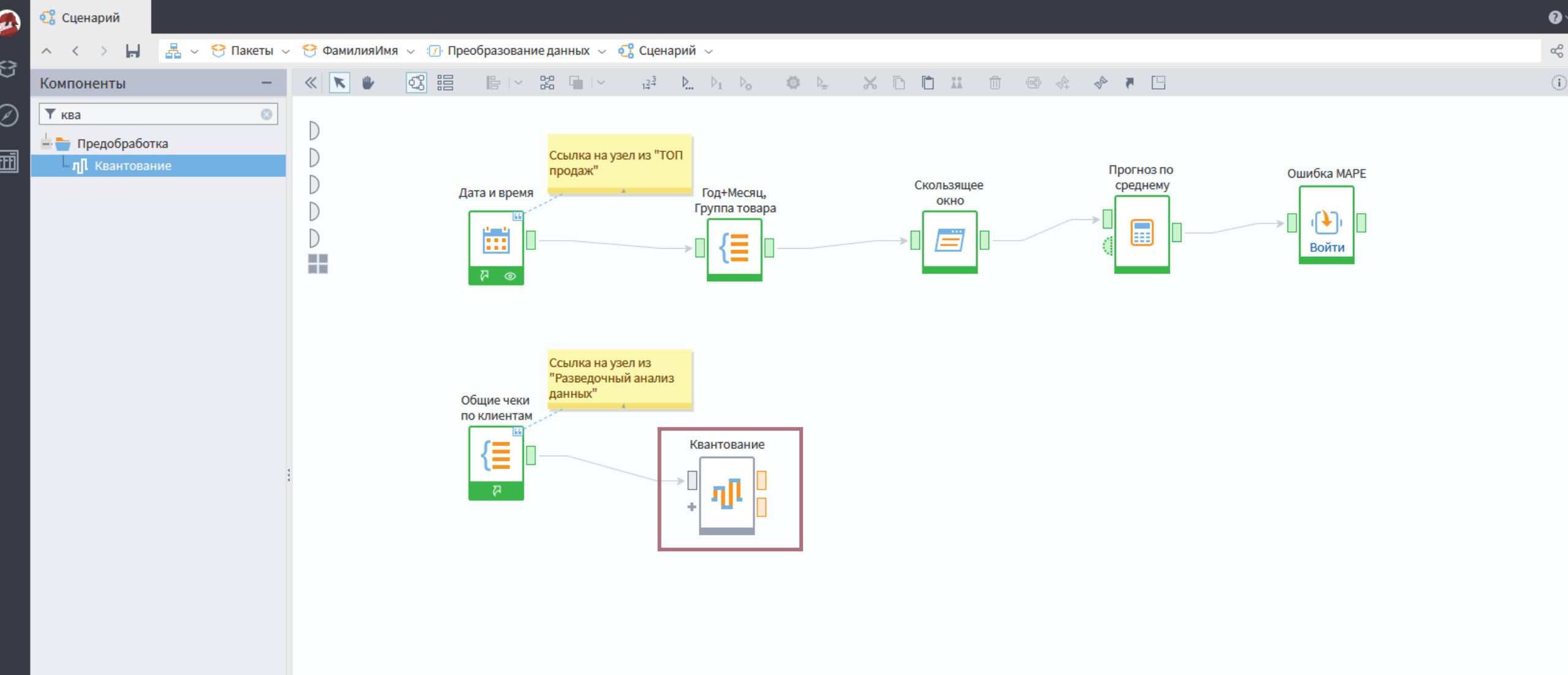


# Задание 1

- Какое значение MAPE имеет товарная группа **Освещение**?  
Ответ округлите до двух знаков после запятой.
- Сколько, согласно модели прогноза по среднему, будет продано товаров группы с самой минимальной ошибкой MAPE в ноябре 2019 года?

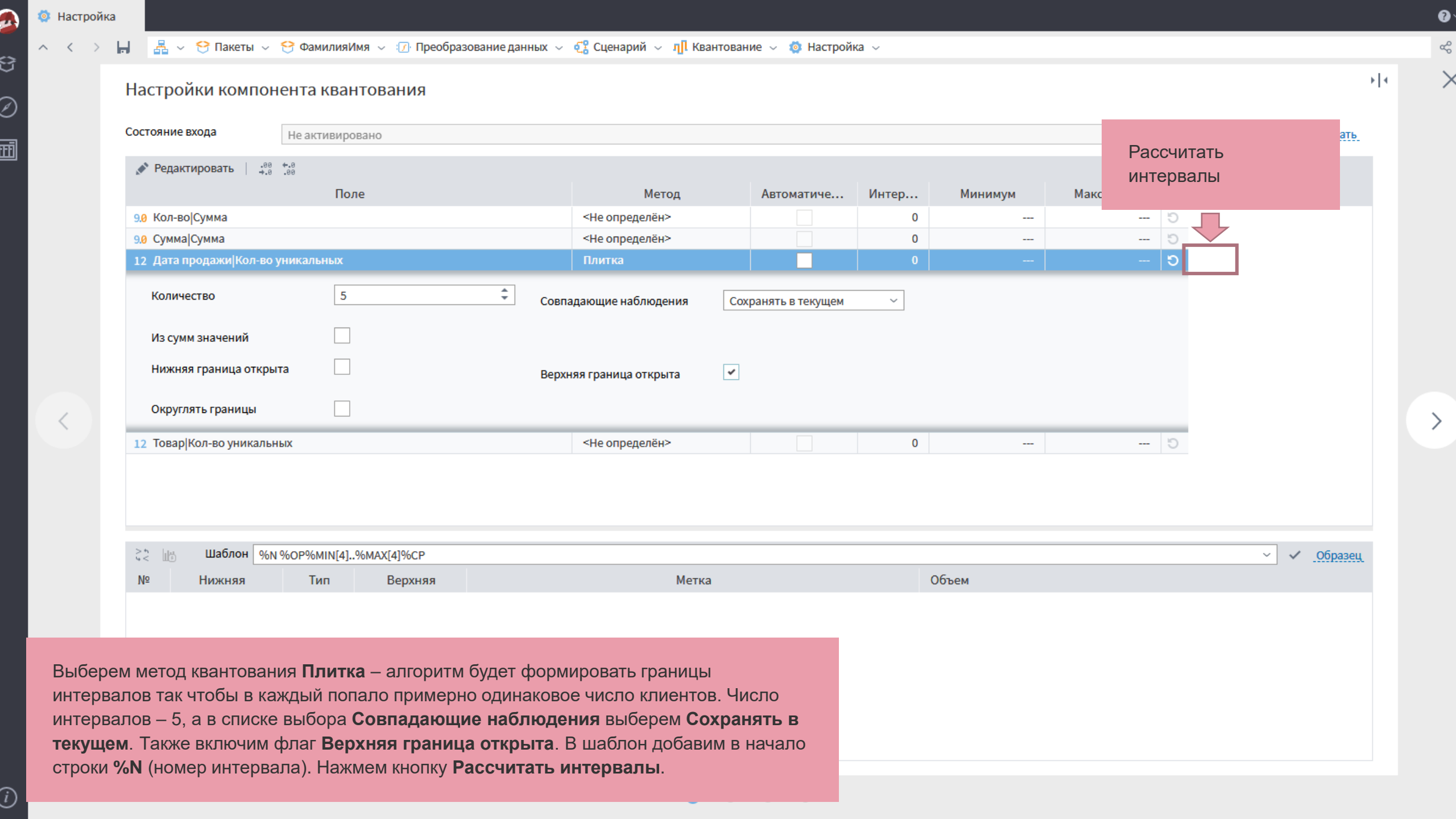
# Квантование





Ранее мы считали количество визитов каждого клиента в магазин. Это непрерывное поле (**Дата продажи|Кол-во уникальных**). Разобьём его на несколько интервалов каким-либо автоматическим методом.

Для этого к узлу **Общие чеки по клиентам** присоединим компонент **Квантование** и войдем в мастер его настройки.



## Настройки компонента квантования

Состояние входа

Не активировано

Редактировать

Поле	Метод	Автоматиче...	Интер...	Минимум	Макс
9.0 Кол-во Сумма	<Не определён>	<input type="checkbox"/>	0	---	---
9.0 Сумма Сумма	<Не определён>	<input type="checkbox"/>	0	---	---
12 Дата продажи Кол-во уникальных	Плитка	<input checked="" type="checkbox"/>	0	---	---

Количество

5

Совпадающие наблюдения

Сохранять в текущем

Из сумм значений

☐

Нижняя граница открыта

☐

Верхняя граница открыта

☒

Округлять границы

☐

12 Товар Кол-во уникальных	<Не определён>	<input type="checkbox"/>	0	---	---
----------------------------	----------------	--------------------------	---	-----	-----

Шаблон %N %OP%MIN[4]..%MAX[4]%CP

№	Нижняя	Тип	Верхняя	Метка	Объем
---	--------	-----	---------	-------	-------

Рассчитать  
интервалы

Выберем метод квантования **Плитка** – алгоритм будет формировать границы интервалов так чтобы в каждый попало примерно одинаковое число клиентов. Число интервалов – 5, а в списке выбора **Совпадающие наблюдения** выберем **Сохранять в текущем**. Также включим флаг **Верхняя граница открыта**. В шаблон добавим в начало строки **%N** (номер интервала). Нажмем кнопку **Рассчитать интервалы**.

## Настройки компонента квантования

Состояние входа

Вход активирован

Активировано

Редактировать

Поле

Метод

Автоматиче...

Интер...

Минимум

Максимум

Кол-во|Сумма

Сумма|Сумма

12 Дата продажи|Кол-во уникальных

Кол-во

5

Совпадающие наблюдения

Сохранять в текущем

Из сумм значений

Нижняя граница открыта

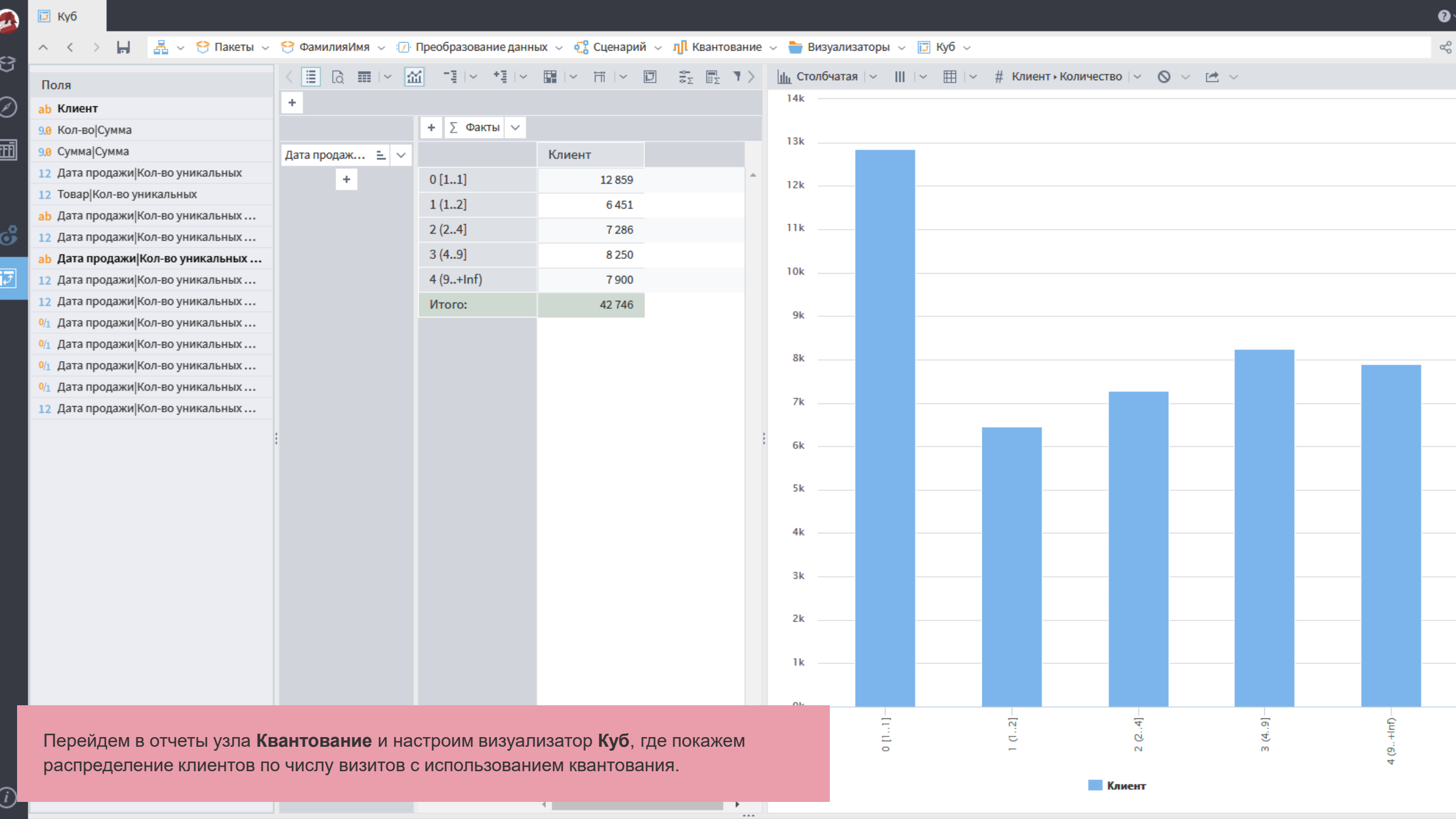
Верхняя граница открыта

Округлять границы

12 Товар|Кол-во уникальных

Получим результат квантования с границами интервалов 1, 2, 4, 9. Из-за большого числа одноразовых алгоритм не смог сделать полностью равные группы – клиентов, которые сделали одну покупку 30%.

Шаблон %N %OP%MIN[4]..%MAX[4]%CP							Образец
№	Нижняя	Тип	Верхняя	Метка	Объем		
0	1	$\leq x \leq$	1	0 [1..1]		30%	
1	1	$< x \leq$	2	1 (1..2]		15%	
2	2	$< x \leq$	4	2 (2..4]		17%	
3	4	$< x \leq$	9	3 (4..9]		19%	
4	9	$< x <$	---	4 (9..+Inf)		18%	



## Задание 2

Создайте копию узла **Квантование** и измените автоматический метод на **Количество**. Ограничьте нижнюю границу 0, верхнюю – 15, количество интервалов тоже установите равным 15. Запустите расчет.

Ответьте на вопрос: Сколько клиентов (кол-во человек) попало в последний интервал при разбиении методом **Количество** при настройках, указанных выше?

# Задание 3

Рассчитайте для каждого клиента его две «любимых» товарных группы. «Любимая» группа – та, на которую сделаны покупки на максимальную сумму (за все визиты).

Расчет любимых групп реализуйте в отдельной подмодели. На выходе должен быть набор данных с тремя полями **Клиент, Любимая группа 1, Любимая группа 2**.

Для реализации этой задачи вам понадобится пронумеровать строки в пределах каждого клиента. Это можно сделать функцией калькулятора **CumulativeSum()**.

Если у клиента только одна любимая группа, то в поле **Любимая группа 2** должно стоять значение NULL (пусто).

Ответьте на вопросы:

- У какого числа клиентов **Любимая группа 1** это "Освещение"? Имейте ввиду, что может возникнуть ситуация, когда у клиента совпадают суммы по нескольким группам. Предусмотрите это, отсортировав группы по алфавиту.
- У какого числа клиентов нет любимой группы 2?
- *Если вы правильно спроектировали подмодель для расчета любимых групп, вы быстро сможете получить ответ на следующий вопрос: У какого числа клиентов **Любимая группа 1** это "Освещение", если расчет производить не по сумме, а количеству товара?*