Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Северо-Кавказский федеральный университет»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по выполнению практических работ
по дисциплине «Интеллектуальный анализ данных в профессиональной сфере»
для студентов специальности
38.05.02 Таможенное дело

Направленность (профиль) «Таможенные платежи»

Ставрополь, 2026

ВВЕДЕНИЕ

Самостоятельная работа - планируемая учебная, учебно-исследовательская, научно-исследовательская работа студентов, выполняемая во внеаудиторное (аудиторное) время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия (при частичном непосредственном участии преподавателя, оставляющем ведущую роль за работой студентов).

Самостоятельная работа студентов в ВУЗе является важным видом учебной и научной деятельности студента. Самостоятельная работа студентов играет значительную роль в рейтинговой технологии обучения.

Целью освоения дисциплины является формирование общепрофессиональной (ОПК-8) компетенции будущего специалиста по направлению подготовки 43.03.02 Туризм.

Основными задачами изучения дисциплины «Интеллектуальный анализ данных в профессиональной сфере» являются:

– изучение основных методов интеллектуального анализа данных;

– изучение основных терминов в области интеллектуальных информационных технологий и анализа данных;

– изучение методик выбора алгоритмов и методов интеллектуального анализа данных при решении профессиональных задач;

– умение осуществлять сбор и систематизацию экспериментальных данных в электронной форме;

– умение проводить предварительную подготовку данных для анализа;

– умение подобрать подходящие алгоритмы и методы интеллектуального анализа данных исходя из поставленной задачи и характеристик выборки данных;

– приобретение опыта решения практических задач в профессиональной области с использованием конкретных программных средств.

Общая характеристика самостоятельной работы студента при изучении дисциплины «Интеллектуальный анализ данных в профессиональной сфере»

Целью самостоятельной работы студентов является овладение фундаментальными знаниями, профессиональными умениями и навыками деятельности по профилю, опытом творческой, исследовательской деятельности. Самостоятельная работа студентов способствует развитию самостоятельности, ответственности и организованности, творческого подхода к решению проблем учебного и профессионального уровня.

Задачами СР являются:

- систематизация и закрепление полученных теоретических знаний и практических умений студентов;
- углубление и расширение теоретических знаний;
- формирование умений использовать нормативную, правовую, справочную документацию и специальную литературу;
- развитие познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности;
- формирование самостоятельности мышления, способностей к саморазвитию, самосовершенствованию и самореализации;
- развитие исследовательских умений;
- использование материала, собранного и полученного в ходе самостоятельных занятий на практических и лабораторных занятиях, при написании курсовых и выпускной квалификационной работ.

План-график выполнения самостоятельной работы

№	Наименование разделов и тем дисциплины, их краткое содержание; вид самостоятельной работы	Форма контроля	Зачетные единицы (часы)
	3 семестр		
1	Подготовка к лабораторной работе	Выполнение лабораторных работ	4,0
2	Подготовка к лекции	Опрос	4,0
3	Самостоятельное изучение литературы	Опрос	2,0
	Итого за 3 семестр		10,0
	Итого		10,0

Методические рекомендации по изучению теоретического материала

Работа с книгой

При работе с книгой необходимо подобрать литературу, научиться правильно ее читать, вести записи. Для подбора литературы в библиотеке используются алфавитный и систематический каталоги.

Важно помнить, что рациональные навыки работы с книгой - это всегда большая экономия времени и сил.

Правильный подбор учебников рекомендуется преподавателем, читающим лекционный курс. Необходимая литература может быть также указана в методических разработках по данному курсу.

Изучая материал по учебнику, следует переходить к следующему вопросу только после правильного уяснения предыдущего, описывая на бумаге все выкладки и вычисления (в том числе те, которые в учебнике опущены или на лекции даны для самостоятельного вывода).

При изучении любой дисциплины большую и важную роль играет самостоятельная индивидуальная работа.

Особое внимание следует обратить на определение основных понятий курса. Студент должен подробно разбирать примеры, которые поясняют такие определения, и уметь строить аналогичные примеры самостоятельно. Нужно добиваться точного представления о том, что изучаешь. Полезно составлять опорные конспекты. При изучении материала по учебнику полезно в тетради (на специально отведенных полях) дополнять конспект лекций. Там же следует отмечать вопросы, выделенные студентом для консультации с преподавателем.

Выводы, полученные в результате изучения, рекомендуется в конспекте выделять, чтобы они при перечитывании записей лучше запоминались.

Опыт показывает, что многим студентам помогает составление листа опорных сигналов, содержащего важнейшие и наиболее часто употребляемые формулы и понятия. Такой лист помогает запомнить формулы, основные положения лекции, а также может служить постоянным справочником для студента.

Различают два вида чтения; первичное и вторичное. *Первичное* - эти внимательное, неторопливое чтение, при котором можно остановиться на трудных местах. После него не должно остаться ни одного непонятного слова. Содержание не всегда может быть понятно после первичного чтения.

Задача *вторичного* чтения полное усвоение смысла целого (по счету это чтение может быть и не вторым, а третьим или четвертым).

Правила самостоятельной работы с литературой

Как уже отмечалось, самостоятельная работа с учебниками и книгами (а также самостоятельное теоретическое исследование проблем, обозначенных преподавателем на лекциях) - это важнейшее условие формирования у себя научного способа познания. Основные советы здесь можно свести к следующим:

• Составить перечень книг, с которыми Вам следует познакомиться;

• Сам такой перечень должен быть систематизированным.

• Обязательно выписывать все выходные данные по каждой книге (при написании курсовых и дипломных работ это позволит очень сэкономить время).

• Разобраться для себя, какие книги (или какие главы книг) следует прочитать более внимательно, а какие - просто просмотреть.

• При составлении перечней литературы следует посоветоваться с преподавателями и научными руководителями (или даже с более подготовленными и эрудированными сокурсниками), которые помогут Вам лучше сориентироваться, на что стоит обратить большее внимание, а на что вообще не стоит тратить время...

•Естественно, все прочитанные книги, учебники и статьи следует конспектировать, но это не означает, что надо конспектировать «все подряд»: можно выписывать кратко основные идеи автора и иногда приводить наиболее яркие и показательные цитаты (с указанием страниц).

Чтение научного текста является частью познавательной деятельности. Ее цель - извлечение из текста необходимой информации. От того на сколько осознанна читающим собственная внутренняя установка при обращении к печатному слову (найти нужные сведения, усвоить информацию полностью или частично, критически проанализировать материал и т.п.) во многом зависит эффективность осуществляемого действия.

Выделяют ***четыре основные установки в чтении научного текста***:

1. информационно-поисковый (задача - найти, выделить искомую информацию);
2. усваивающая (усилия читателя направлены на то, чтобы как можно полнее осознать и запомнить как сами сведения, излагаемые автором, так и всю логику его рассуждений);
3. аналитико-критическая (читатель стремится критически осмыслить материал, проанализировав его, определив свое отношение к нему);
4. творческая (создает у читателя готовность в том или ином виде - как отправной пункт для своих рассуждений, как образ для действия по аналогии и т.п. - использовать суждения автора, ход его мыслей, результат наблюдения, разработанную методику, дополнить их, подвергнуть новой проверке).

Основные виды систематизированной записи прочитанного:

1. Аннотирование - предельно краткое связное описание просмотренной или прочитанной книги (статьи), ее содержания, источников, характера и назначения;
2. Планирование - краткая логическая организация текста, раскрывающая содержание и структуру изучаемого материала;
3. Тезирование- лаконичное воспроизведение основных утверждений автора без привлечения фактического материала;
4. Цитирование - дословное выписывание из текста выдержек, извлечений, наиболее существенно отражающих ту или иную мысль автора;
5. Конспектирование - краткое и последовательное изложение содержания прочитанного.

Конспект - сложный способ изложения содержания книги или статьи в логической последовательности. Конспект аккумулирует в себе предыдущие виды записи, позволяет всесторонне охватить содержание книги, статьи. Поэтому умение составлять план, тезисы, делать выписки и другие записи определяет и технологию составления конспекта.

Методические указания по составлению конспекта

1. Внимательно прочитайте текст. Уточните в справочной литературе непонятные слова. При записи не забудьте вынести справочные данные на поля конспекта.
2. Выделите главное, составьте план.
3. Кратко сформулируйте основные положения текста, отметьте аргументацию автора.
4. Законспектируйте материал, четко следуя пунктам плана. При конспектировании старайтесь выразить мысль своими словами. Записи следует вести четко, ясно.
5. Грамотно записывайте цитаты. Цитируя, учитывайте лаконичность, значимость мысли.

В тексте конспекта желательно приводить не только тезисные положения, но и их доказательства. При оформлении конспекта необходимо стремиться к емкости каждого предложения. Мысли автора книги следует излагать кратко, заботясь о стиле и выразительности написанного. Число дополнительных элементов конспекта должно быть логически обоснованным, записи должны распределяться в определенной последовательности, отвечающей логической структуре произведения. Для уточнения и дополнения необходимо оставлять поля.

Овладение навыками конспектирования требует от студента целеустремленности, повседневной самостоятельной работы.

Задания для подготовки к лабораторным работам

Задание 1.

Определить, сколько выручки приносят товарные группы по периодам.

1. Заполнить пропуски
2. Выделить периоды
3. Визуально оценить результаты
4. Получить расчетных данных

Инструментарий: Стандартные компоненты в Loginom.

Задание 2.

Анализ ассортимента

Цель: узнать, какие продукты и услуги компании приносят наибольшую прибыль, а от каких лучше отказаться

Результат: оптимизация ассортимента

Базовые методы:

1. Анализ рентабельности ассортиментной группы товаров
2. ABC-анализ
3. Анализ по адаптивной матрице BCG

4. Анализ по методу ДиббаСимкина

5. Анализ по матрице совместных покупок

Задание 3.

Задание на настройку ETL-процесса:

Вы работаете в крупной компании, которая использует множество различных систем для управления своими бизнес-процессами. Вам нужно настроить ETL-процесс с использованием платформы Loginom для извлечения данных из разных источников и их загрузки в целевую базу данных. Конкретно, вам нужно извлечь данные из Salesforce, Google Analytics и Excel-файла, объединить их и загрузить в базу данных Microsoft SQL Server. В качестве задания вам необходимо настроить соединения с каждым из источников данных, определить правила преобразования данных, настроить согласование данных и загрузить их в базу данных.

Задание 4.

Задание на настройку бизнес-процесса:

Вы работаете в крупной международной компании и вам нужно настроить бизнес-процесс с использованием платформы Loginom для управления и контроля производственным процессом на заводе в другой стране. Конкретно, вам нужно создать цепочку действий, которые будут выполняться автоматически, когда производственный процесс завершится, начиная с отправки уведомлений и заканчивая запуском автоматических заказов на закупку необходимых материалов. В качестве задания вам нужно определить этапы производственного процесса, создать соответствующие шаги для Loginom, задать условия выполнения действий и определить порядок выполнения шагов.

Список рекомендуемой литературы

Основная литература:

1. Нестеров, С. А. Интеллектуальный анализ данных средствами MS SQL Server 2008 / С.А. Нестеров. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 338 с. : ил. - http://biblioclub.ru/. - Библиогр. в кн

2. Пальмов, С.В. Интеллектуальный анализ данных Электронный ресурс : учебное пособие / С.В. Пальмов. - Самара : Поволжский государственный университет телекоммуникаций и информатики, 2017. - 127 с. - Книга находится в базовой версии ЭБС IPRbooks.

3. Управление данными : учебник / Ю.Ю. Громов, О.Г. Иванова, А.В. Яковлев, В.Г. Однолько ; Министерство образования и науки Российской Федерации ; Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тамбовский государственный технический университет». - Тамбов : Издательство ФГБОУ ВПО «ТГТУ», 2015. - 192 с. : ил., табл., схем. - http://biblioclub.ru/. - Библиогр. в кн. - ISBN 978-5-8265-1385-9

Дополнительная литература:

1 Васюков, О. Г. Управление данными : учебно-методическое пособие / О.Г. Васюков ; Министерство образования и науки РФ ; Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Самарский государственный архитектурно-строительный университет». - Самара : Самарский государственный архитектурно-строительный университет, 2014. - 161 с. : табл., ил. - http://biblioclub.ru/. - Библиогр. в кн. - ISBN 978-5-9585-0608-8

2 Козлов, А. Ю. Статистический анализ данных в MS EXCEL : учеб. поосбие / А. Ю. Козлов, В. С. Мхитарян, В. Ф. Шишов. - М. : ИНФРА-М, 2012. - 320 с. - (Высшее образование). - Гриф: Рек. УМО. - ISBN 978-5-16-004579-5

3 Мельниченко, А. С. Математическая статистика и анализ данных Электронный ресурс : Учебное пособие / А. С. Мельниченко. - Математическая статистика и анализ данных,2019-09-01. - Москва : Издательский Дом МИСиС, 2018. - 45 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-906953-62-9

Интернет-ресурсы:

1. Официальный сайт библиотеки ФГАОУ ВО СКФУ Режим доступа: http://cataloq.ncstu.ru|cataloq –.
2. Информационная справочная система ГАРАНТ.РУ // Режим доступа: http://www.garant.ru/
3. Информационная справочная система КонсультантПлюс. // Режим доступа: http://www.consultant.ru
4. Профессиональная база данных «Всероссийская система данных о компаниях и бизнесе «За честный бизнес» // Режим доступа: https://zachestnyibiznes.ru
5. Профессиональная база данных Росстата // Режим доступа: Росстат — Базы данных (rosstat.gov.ru)

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Северо-Кавказский федеральный университет»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по организации самостоятельной работы
по дисциплине «Интеллектуальный анализ данных в профессиональной сфере»
для студентов специальности
38.05.02 Таможенное дело

Направленность (профиль) «Таможенные платежи»

Ставрополь, 2026

Введение

Целью освоения дисциплины является формирование общепрофессиональной (ОПК-8) компетенции будущего специалиста по направлению подготовки 43.03.02 Туризм.

Основными задачами изучения дисциплины «Интеллектуальный анализ данных в профессиональной сфере» являются:

– изучение основных методов интеллектуального анализа данных;

– изучение основных терминов в области интеллектуальных информационных технологий и анализа данных;

– изучение методик выбора алгоритмов и методов интеллектуального анализа данных при решении профессиональных задач;

– умение осуществлять сбор и систематизацию экспериментальных данных в электронной форме;

– умение проводить предварительную подготовку данных для анализа;

– умение подобрать подходящие алгоритмы и методы интеллектуального анализа данных исходя из поставленной задачи и характеристик выборки данных;

– приобретение опыта решения практических задач в профессиональной области с использованием конкретных программных средств.

Лабораторная работа № 1.

Тема: Технологии анализа данных.

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

Loginom — low-code платформа для продвинутой аналитики данных. Визуальный конструктор позволяет настроить все процессы анализа от интеграции и подготовки данных до моделирования и визуализацию.

Редакции платформы:

- Community — бесплатная настольная версия для некоммерческого использования
- Personal — коммерческая настольная версия для автономной аналитической обработки.
- Team — редакция, предназначенная для небольших команд до 5 человек, ориентированная на решение базовых аналитических задач.
- Standard — редакция для реализации аналитических проектов и обработки существенных объемов данных в рамках средних организаций или департаментов крупных компаний.
- Enterprise — редакция, ориентированная на создание отказоустойчивых систем принятия решений и обработки больших объемов данных.

Полезные ресурсы:

- Обучающие ролики
- Демопримеры
- Документация

При работе с данными в «обывательском» режиме кажется, что пирамида задач выглядит так:

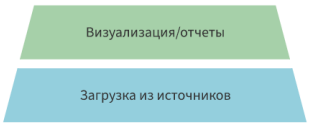


Рис.1 Задачи дата-аналитики (ожидание)

Нужно подключиться к источникам, построить отчеты, сделать выводы или отправить отчет выше по цепочке. Однако в реальности при работе с данными следует держать в голове больше промежуточных этапов от загрузки до визуализации.

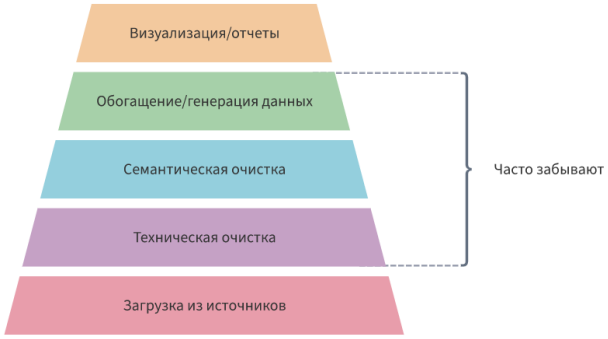


Рис.2 Задачи дата-аналитики (реальность)

Этапы подготовки данных

Перед тем, как приступить к визуализации необходимо выполнить определенные действия над данными:

Загрузка из источников. Нужно подключиться к месту хранения данных. Чаще всего это база данных или файлы. Может потребоваться загрузка множества одинаковых файлов или разворачивание сводных таблиц в плоские. Частая задача — объединение записей из нескольких источников.

Техническая очистка. Данные нужно проверить на наличие типовых проблем: дубли, пропуски, противоречия. Даже если информация получена из надежного хранилища, не стоит игнорировать проверку. В конце концов, многие ошибки технического плана возникают «бесшумно». Например, разработчик не учел реальную структуру данных в написании запроса объединения таблиц. Как следствие появились дубли. Ошибка может пройти незамеченной и проявить себя при получении отчета руководством.

Семантическая очистка. Технически идеальные данные могут содержать смысловые ошибки. Одна из причин — объединение несовместимых понятий, процессов или показателей, например, построение портрета среднего клиента на основе смешанных оптовых и розничных продаж.

Обогащение/генерация данных. Визуализируемые данные желательно обогащать дополнительными аналитическими признаками. В этом случае пользователь отчета сможет быстро отфильтровать клиентов по статусам или товары по стабильности продаж, и, следовательно, проанализировать показатели в этих разрезах. Это делает отчеты проще и понятнее.

Ценность правильной подготовки данных

Известно множество примеров, когда неправильные настройки визуализации приводят к некорректному восприятию данных. Классика жанра — при отсчете оси Y не от нуля разница между столбцами выглядит более серьезной, чем на самом деле.

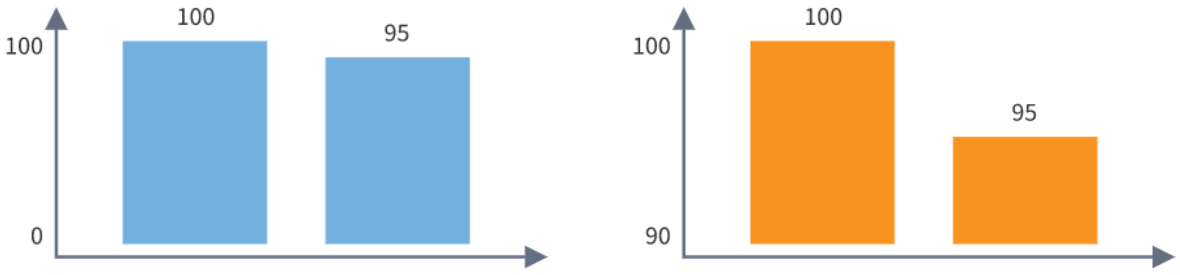


Рис. 3 – Искажение в восприятии информации

Однако если проблемы кроются на уровне самих данных, а не визуализации, высока вероятность запутать всех, включая самого себя, не подозревая об этом. Даже если у аналитика черный пояс по созданию дашбордов.

Проблемы могут быть технического характера: на дашбордах числа больше/меньше, чем в учетных системах или других отчетах. Потребуется время, чтобы разобраться почему так и кто не прав. Такие ошибки, как правило, быстро выявляются, хоть и сопровождаются нападками в сторону аналитика :)

Не менее опасны смысловые ошибки, которые могут создать предпосылки для неправильного принятия решений, например, завышенных прогнозов. Если в отчете средний чек клиента 15 000, а на самом деле 7 500, то будет построена слишком оптимистичная финансовая модель.

Решения, принятые на основе такой информации, приведут к потерям. И хотя технически в данных нет ошибок, последствия могут быть печальными. Задача очистки данных сделать так, чтобы проблемы были исправлены или хотя бы выявлены, а не маскировались за красивыми картинками.

Интерфейс Loginom. С чего начать

Прежде, чем начинать практиковаться на реальных данных, нужно разобраться, как работает Loginom. **Занятие будет посвящено не построению сценария обработки с нуля, а изучению готового проекта, на котором можно понять основные механики платформы.**

Loginom — система визуального проектирования процессов преобразования данных в режиме low-code. Это значит, что для решения большинства задач не требуется программировать. При этом, если у аналитика есть навыки кодирования на JavaScript или Python, то он может включать в сценарии скрипты на этих языках, а при работе с базами данных использовать SQL.

В итоге, основная рабочая область в Loginom выглядит примерно так:

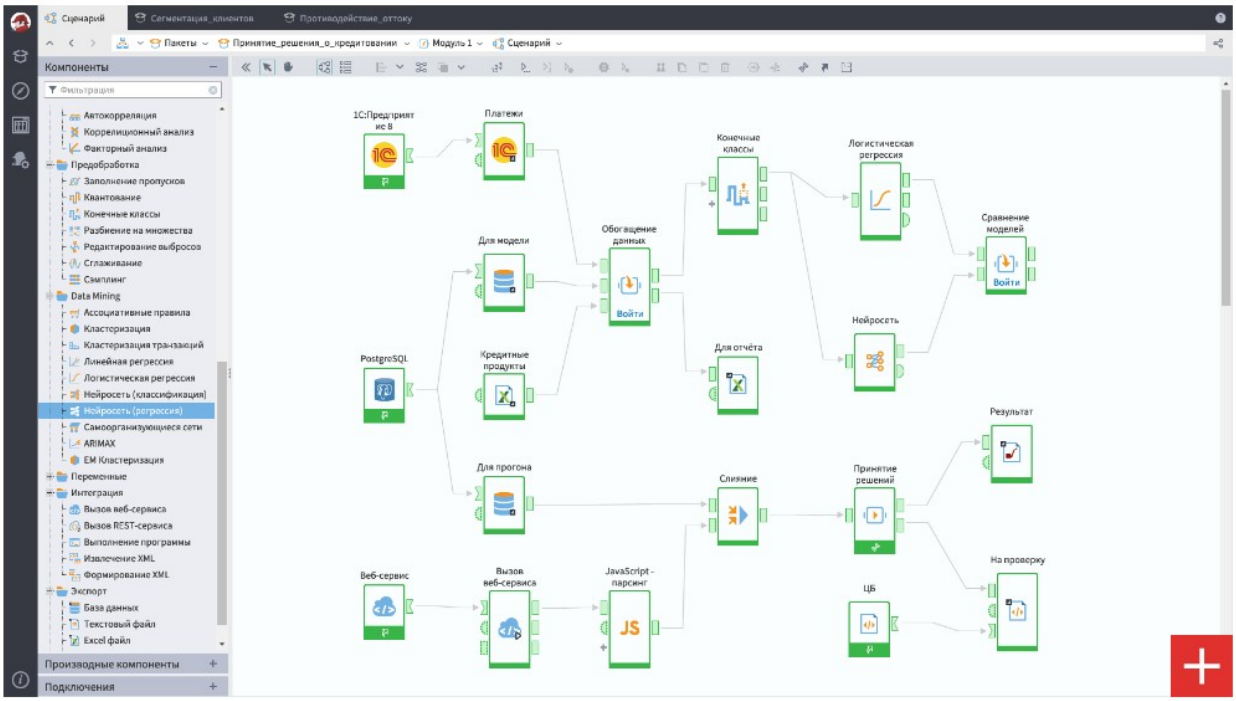


Рис.4 пример сценария в Loginom

Low-code подход в бизнес-аналитике

Аналитика данных — сложная задача, в которой можно обеспечить высокую эффективность только за счет комбинирования технических навыков с бизнес-экспертизой.

Чаще всего глубокая бизнес-экспертиза и хорошие технические знания по работе с данными идут порознь. В конце концов, человек становится профессионалом в той сфере, в которую инвестирует большую часть времени.

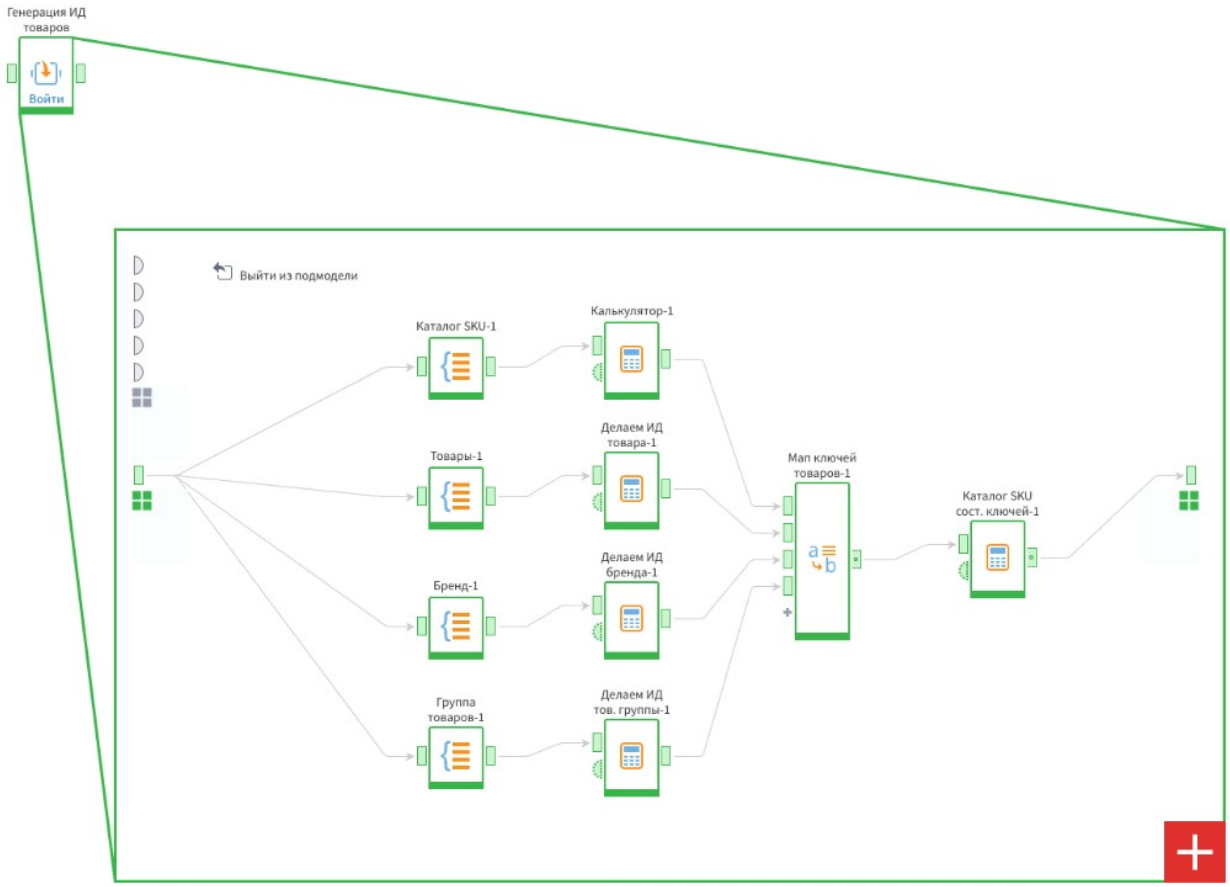
Поэтому, если у аналитика хорошая бизнес-экспертиза, но технические навыки хромают, то с помощью low-code он сможет легко решить свои проблемы. Это займет гораздо меньше времени, чем попытка объяснить программисту постановку задачи и добиться от него корректной реализации. Не говоря уж о том, что чаще всего техническую работу делать некому, т.к. компании испытывают большой дефицит IT-специалистов.

Если же аналитик сфокусирован на выполнении входящих задач от бизнес-пользователей, то работа в low-code формате позволит ему предоставлять результат быстрее и легче вносить в него корректировки. Потому что никто не знает, что заказчик попросит добавить или изменить завтра:)

Накопление экспертизы

Работа в Loginom интересна тем, что позволяет не только выполнить некие действия над данными, но и создать собственные переиспользуемые компоненты, которыми могут воспользоваться другие сотрудники организации.

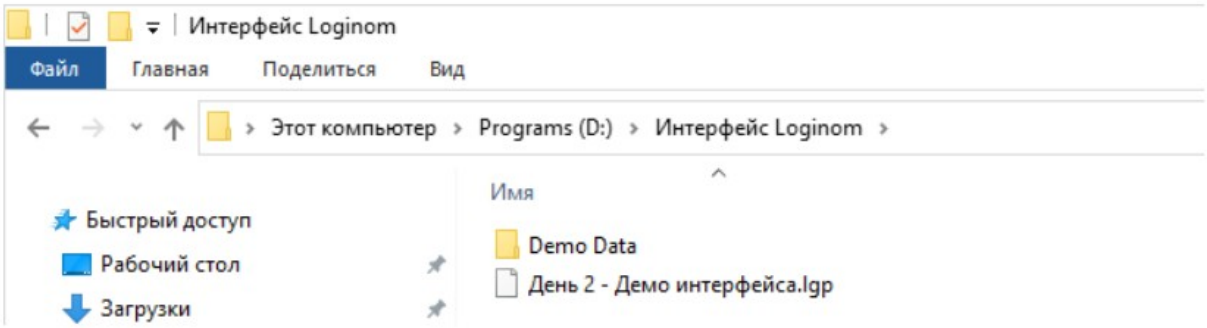
Эти компоненты могут быть как преднастроенными подключениями к данным в разных источниках, так и подмоделями — сложными параметризованными процессами из десятков узлов. Подмодель выглядит для пользователя как узел сценария, в который можно войти.



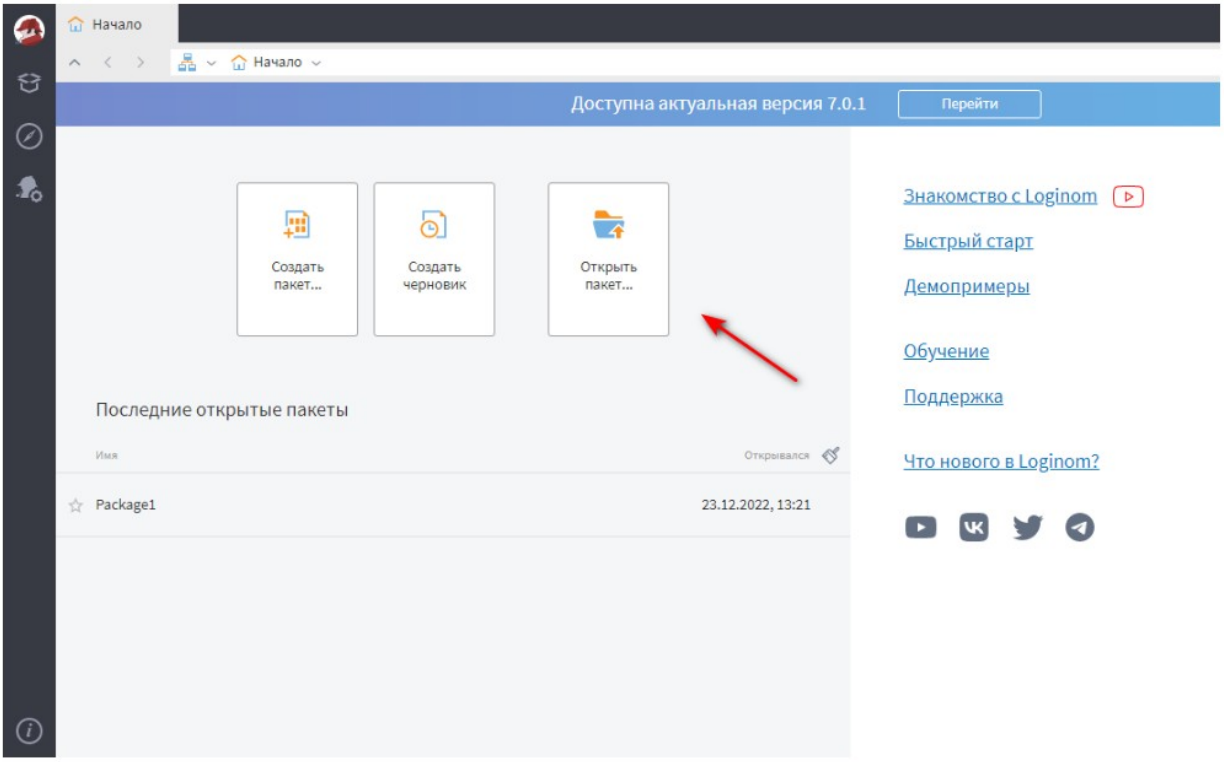
Таким образом, на базе Loginom можно организовать единое пространство работы с любыми структурированными данными с добавлением пользовательских моделей.

Аналитик может создать свою библиотеку компонентов как с написанием скриптов на языках программирования, так и без единой строчки кода, и предоставить коллегам без серьезного технического бэкграунда свои наработки. А опытный специалист может обернуть свою экспертизу в компоненты и создать на базе этого функционала свой собственный аналитический продукт.

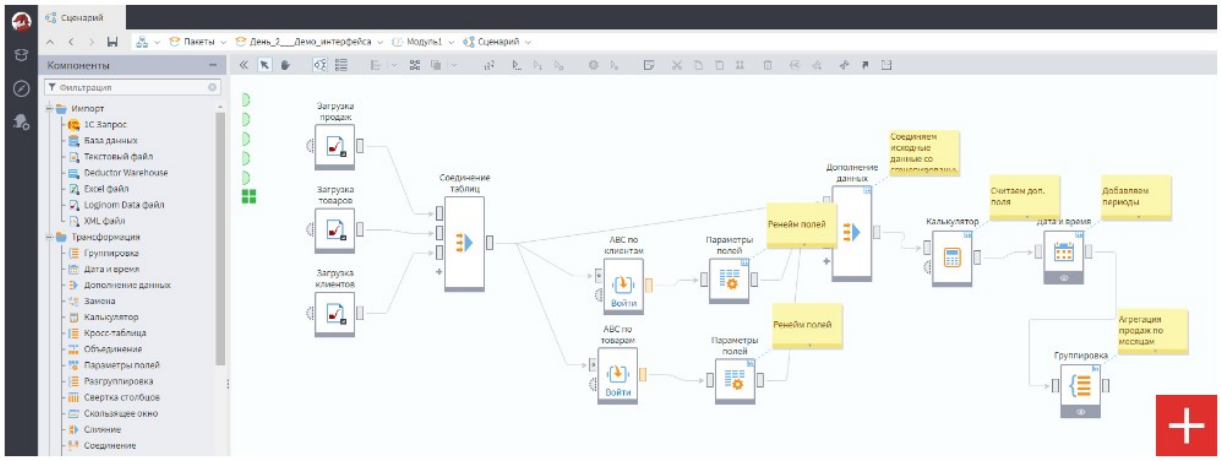
Состав пакета Loginom



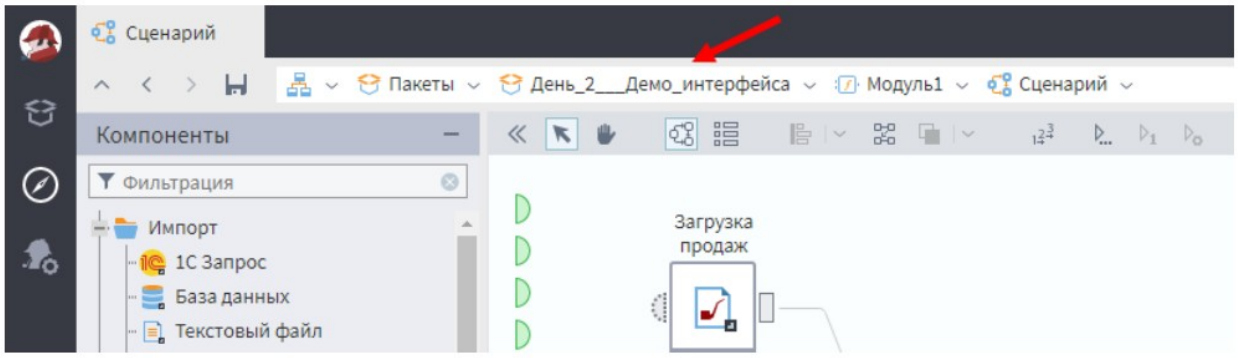
Откройте Loginom. Отобразится начальный экран, где доступны действия открыть или создать пакет. Пакет — это базовая проектная единица в Loginom, представленная в виде файла с расширением **lgp**. Щелкните по кнопке «Открыть пакет» и откройте скачанный файл.



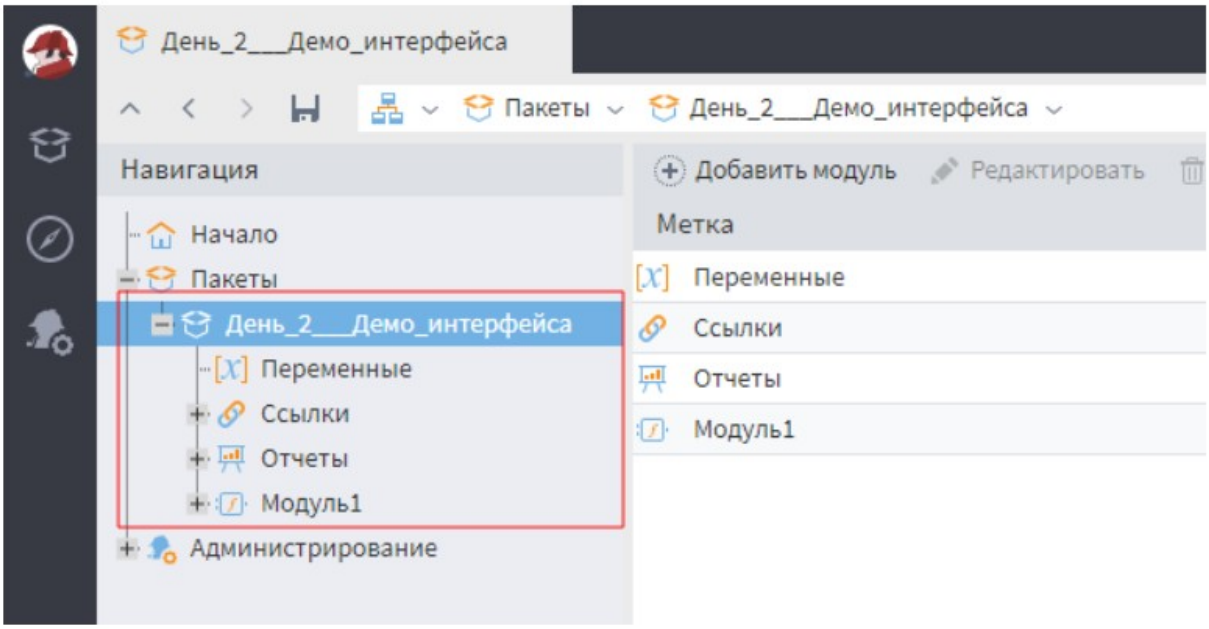
Перед вами откроется схема обработки данных. Она называется Сценарий.



Но не будем погружаться в нее прямо сейчас. Для начала изучим структуру пакета. Для этого кликнем левой кнопкой мыши (ЛКМ) по названию пакета в строке навигации.



В левой части экрана вы увидите дерево со структурой открытых в данный момент пакетов. Возможно, придется выделить ее мышкой, чтобы она отображалась как на картинке. В правой части экрана видно содержимое текущего выбранного узла в дереве.



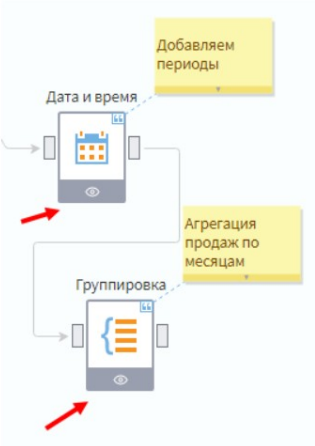
Перемещаться между уровнями можно либо через выделение нужного узла в дереве слева, либо через двойной клик ЛКМ по нужному элементу в центральной области.

Что тут есть?

Переменные (а точнее, переменные пакета) — список переменных (изменяемых параметров), которые будут доступны на любом уровне в сценариях (схемах обработки данных) пакета. Вообще, переменные в Loginom можно создавать практически в любой момент выполнения сценария. Поэтому сюда обычно выносятся **Самые Главные Переменные** — те, которые должны быть легко доступны в одном месте.

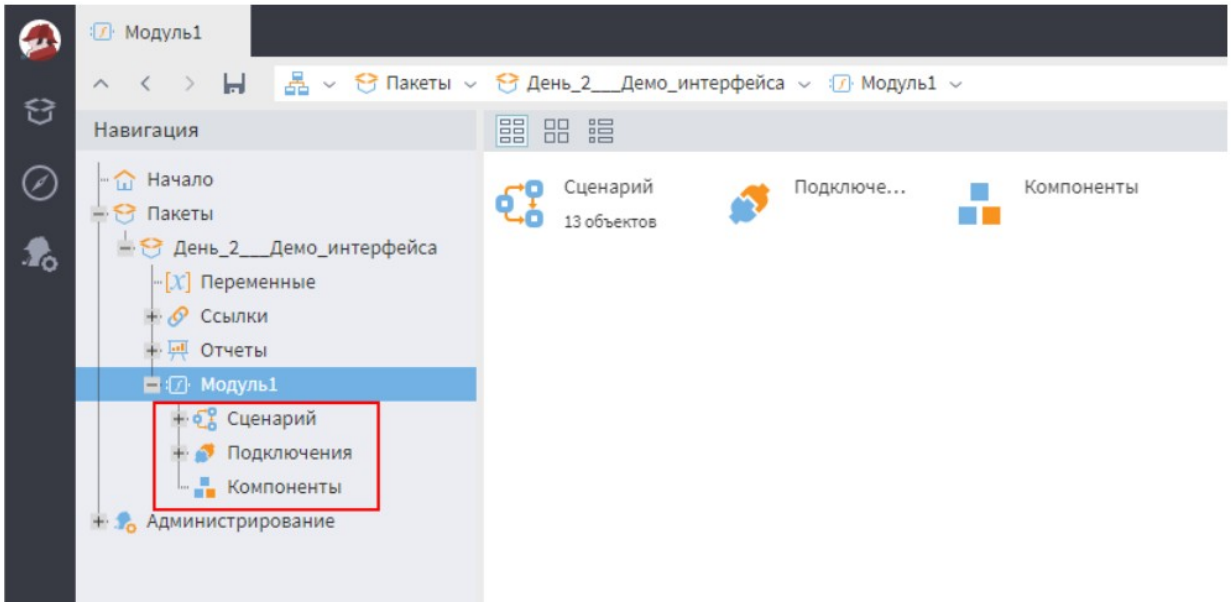
Ссылки — ссылки на другие пакеты, чьи компоненты могут быть переиспользованы. Это тот самый функционал, который позволяет формировать подключаемые библиотеки ваших и сторонних компонентов. Мы подробнее рассмотрим эту механику в будущем.

Отчеты — единое место для отображения отчетов текущего пакета. Вообще, все отчеты делаются в привязке к определенным узлам сценария. Узлы с отчетами можно определить по наличию пиктограммы глаза.



Как вы понимаете, реальные сценарии могут быть очень большими и многоуровневыми, и выискивать, какой именно узел содержит нужный отчет, — тяжкий труд. Поэтому в каждом пакете существует сводная область Отчеты, куда могут быть добавлены отчеты из разных узлов.

Модуль — пожалуй, главный раздел любого пакета. В одном пакете может быть несколько модулей, и один модуль может содержать в себе другие модули.



Каждый модуль всегда содержит в себе 3 подраздела:

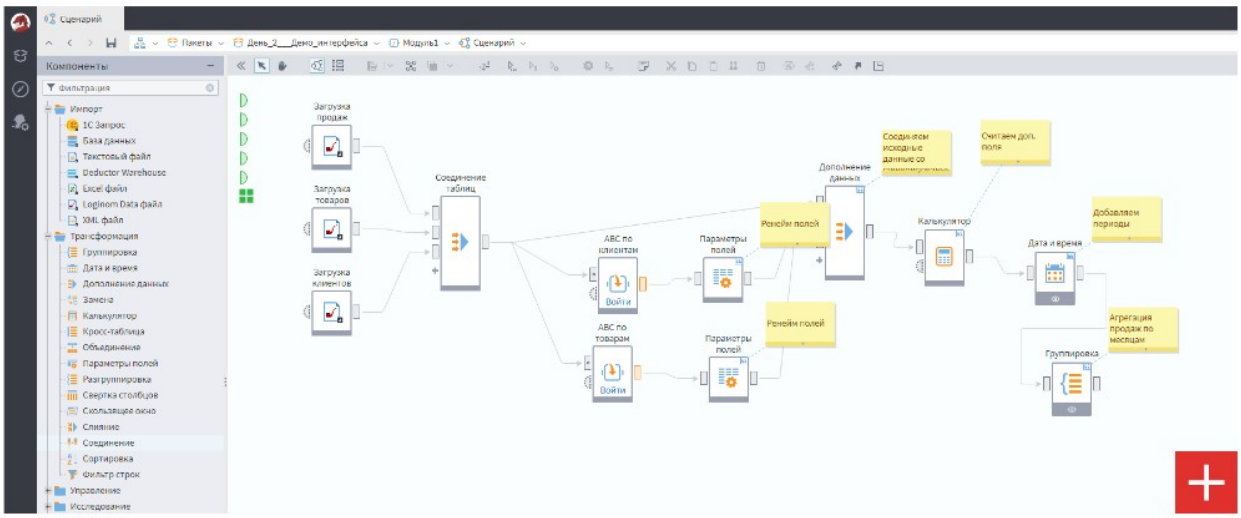
Сценарий — схема обработки данных, выстроенная как последовательность соединенных между собой узлов.

Подключения — преднастроенные подключения к источниками типа баз данных (детально рассмотрим в будущем) и REST-подключений.

Компоненты — перечень узлов сценария, разрешенных к переиспользованию. Именно они будут появляться в других пакетах при добавлении ссылок.

Интерфейс сценария Loginom

Сценарий представляет собой набор соединенных между собой узлов. Каждый узел реализует определенный функционал работы с данными. Доступные компоненты перечислены в левой части экрана.



Глобально узлы можно разделить на следующие группы:

Узлы импорта. Отвечают за загрузку данных из разных источников.

Узлы трансформации — преобразование данных, сходное с тем, что обычно делается базовыми SQL-запросами. Т.е. разнообразное соединение таблиц (JOIN/UNION), создание дополнительных вычисляемых полей, группировка, сортировка и т.д.

Узлы управления. Используются, когда нужно реализовать сложную логику в сценарии. В частности, ветвление сценария по условию (IF) и выполнение действий в цикле.

Узлы программирования. Позволяют вставлять в сценарий код на JavaScript или Python. Дают возможность применять библиотеки этих языков и при необходимости дописывать недостающий функционал.

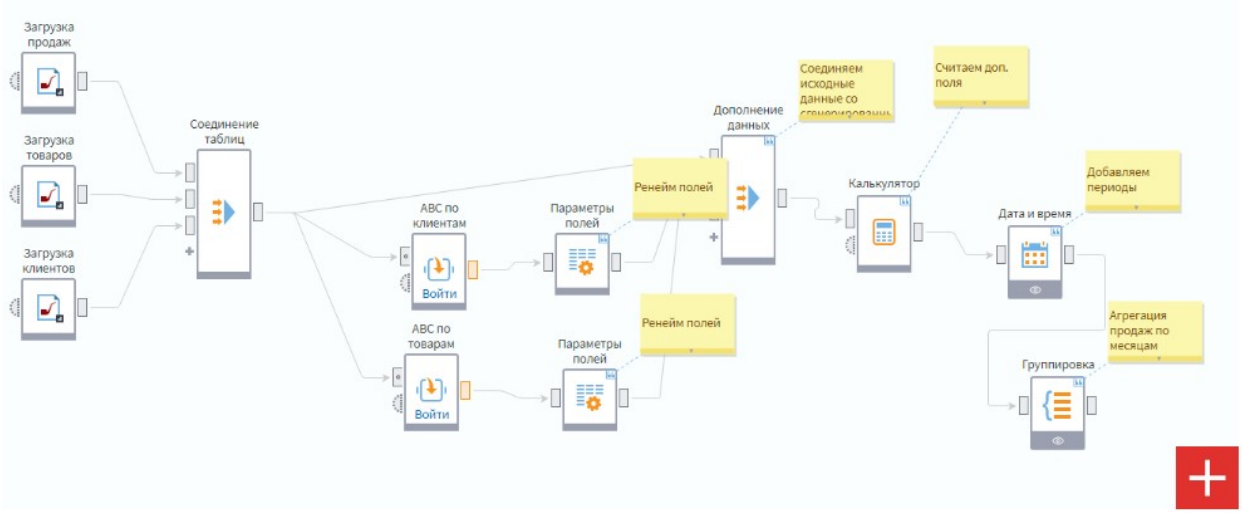
Узлы интеграции. Позволяют инициировать запуск внешних программ или сервисов из сценария Loginom. Используются в сложных интеграционных кейсах.

Узлы статистической обработки. Большой набор узлов, посвященный исследованию и очистке данных, а также применению алгоритмов машинного обучения.

Узлы переменных. Отдельный набор узлов по формированию переменных, включая генерацию переменных из таблицы (очень удобно для выноса параметров управления во внешний контур), вычисление переменных через выражения, группировка потоков переменных.

Узлы экспорта. Все таблицы, которые формируются внутри Loginom, могут быть выгружены в виде файлов или экспортированы в базы данных для переиспользования в других аналитических системах и визуализаторах.

Рассмотрим подробнее сам сценарий.



То, что каждый компонент этой схемы называется узлом, мы уже выяснили. Узлы соединяются между собой через порты. Порты — это выступающие блоки слева и справа на узлах. Порты, которые находятся слева, называются входными. В них поступают данные, которые будут обработаны в узле. Порты, которые находятся справа, — выходные. На них выводится результат обработки данных.

Порты бывают разных типов, которые определяются их формой. В 99% случаев вы будете использовать 2 типа портов:

Порт данных — прямоугольный порт. Он передает строго одну таблицу.

Порт переменных — полукруглый порт. Может передавать несколько переменных, которые будут использованы внутри узла для подстановки в параметры узла или выражения.

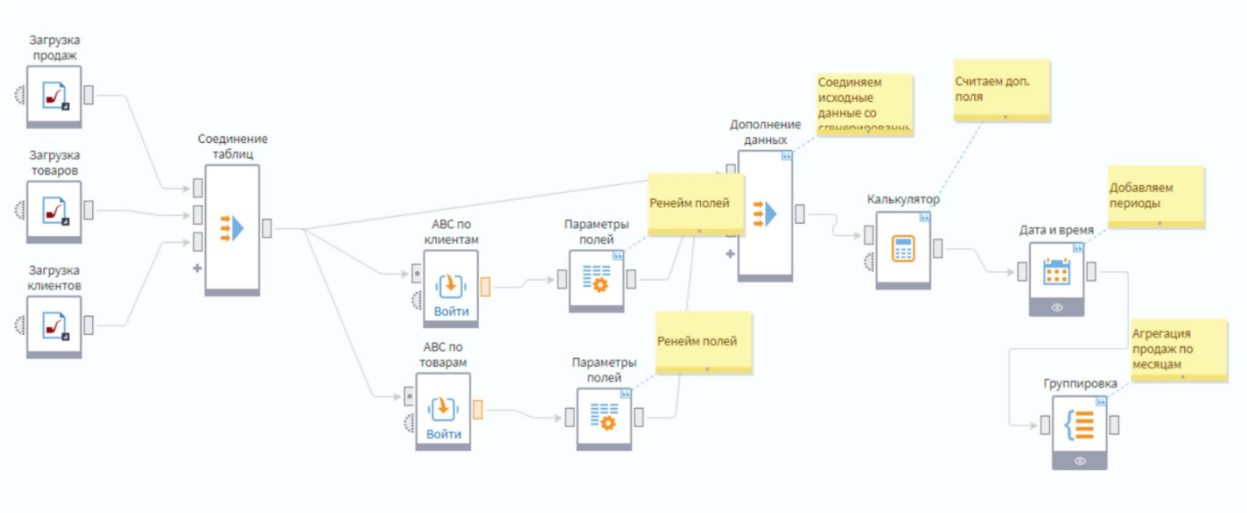
Порты могут быть обязательными (имеют сплошную границу, как порт данных на картинке вверху) и необязательными (имеют пунктирную обводку, как порт переменных на картинке вверху). Если порт обязательный, то это значит, что в него должно быть что-то введено. Иначе при выполнении сценария возникнет ошибка.

А еще, некоторые порты имеют внутри себя точку. Это означает, что для данного порта строго определен набор полей и типы данных, которые он должен получить извне или которые он отдает вовне, если это выходной порт. Такое часто применяется в переиспользуемых компонентах и служит не только ограничителем нежелательных действий, но и подсказкой для пользователя при использовании такого узла.

Детали работы с этими режимами мы разберем на следующих занятиях.

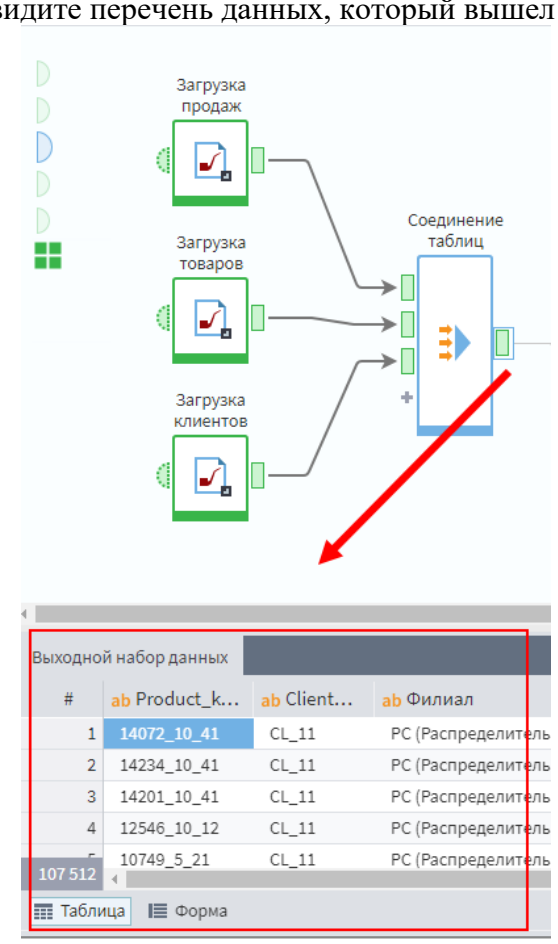
Запуск сценариев

Давайте попробуем Loginom в деле. Сейчас все узлы перед нами серого цвета. Это значит, что они не активированы. Т.е. данные в них не загружены, действия не выполнены. Самый простой способ активировать узлы — нажать кнопку Play в панели инструментов.



Как результат этого действия, все узлы станут зелеными. Это значит, что сценарий отработал успешно.

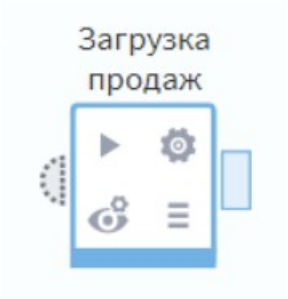
Что дает активация узлов? Во-первых, при двойном клике ЛКМ по любому выходному порту вы увидите перечень данных, который вышел из узла в этом порту.



Так можно контролировать процесс преобразования данных.

Предпросмотр — это способ быстрого отображения результата, который выводит до 1 млн строк. Анализ и более глубокий аудит данных выполняется другими инструментами.

Также можно активировать не все узлы сразу. Для этого нужно кликнуть ЛКМ в центр пиктограммы узла. В результате появится мини-панель управления с 4-мя кнопками.



Кнопка Play/Stop — активация/деактивация узла в зависимости от текущего состояния. Зачем может потребоваться деактивировать узел? Например, если вначале данные были импортированы, а затем в них внесли изменения.

Узел хранит в себе информацию, загруженную в момент его активации. Для обновления данных узел требуется деактивировать и активировать повторно.

Деактивация узла приведет к деактивации всех зависимых от него узлов, их также придется активировать повторно. Запуск единичного узла активирует только те узлы, которые необходимы для его выполнения. Это полезно для оптимизации производительности при разработке и отладке.

Настройки узла (Шестеренка). Здесь можно задать специфичные для узла настройки. Вход в настройки и их изменение приводит к деактивации узла и всей последующей цепочки узлов.

Визуализаторы (Глаз с шестеренкой) — инструменты для отображения данных. Подробнее ниже по тексту.

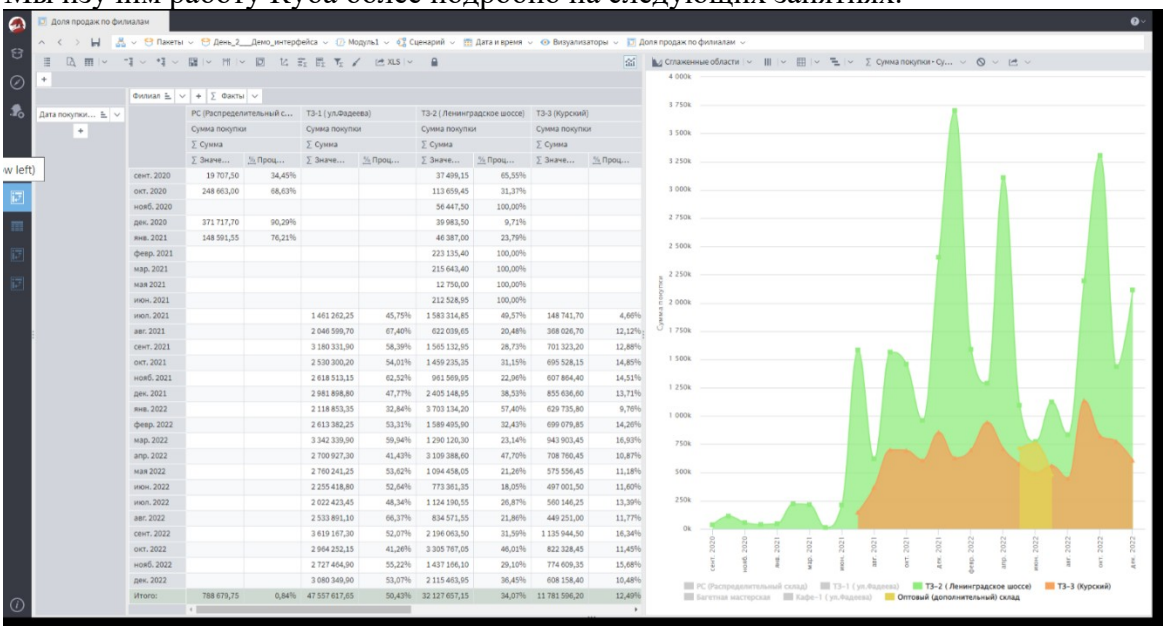
Контекстное меню (Бутерброд) открывает меню со списком возможных действий над узлом. Также может вызываться через клик правой кнопкой мыши (ПКМ) по узлу.

Визуализаторы в Loginom

На каждый выходной порт данных может быть назначено несколько визуализаторов с разными настройками. Сразу обозначим, что Loginom — не система построения дашбордов и не имеет функционала для работы с моделью данных. Для построения дашбордов данные из Loginom экспортируются во внешние приемники и отображаются в сторонних BI-системах.

Каждый визуализатор работает с данными строго одной таблицы. Самый популярный визуализатор для исследования данных — Куб. Он позволяет построить сводную таблицу с динамической фильтрацией и простыми диаграммами, а также детализировать ячейки до базовых данных.

Мы изучим работу Куба более подробно на следующих занятиях.

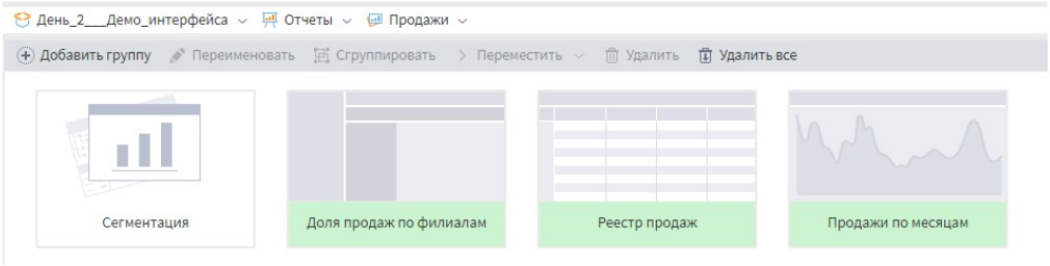


Если в узле есть визуализатор, то это будет видно по пиктограмме глаза в его нижней части. При клике на ней, откроется последний используемый визуализатор.

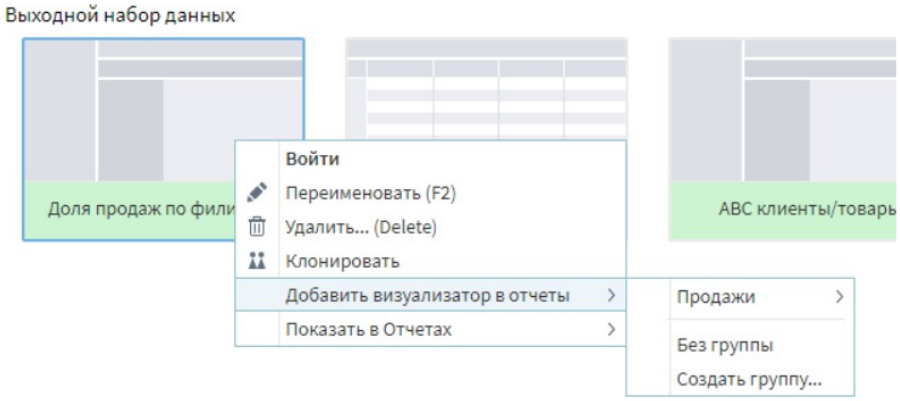
Если в узле используется несколько визуализаторов, то на боковой панели появятся пиктограммы, при помощи которых можно переключаться между визуализаторами.

Когда сценарий разрастается, и в нем используется множество визуализаторов в разных узлах, искать нужный отчет по лабиринтам узлов становится неудобно. Для облегчения этой задачи визуализаторы разных узлов могут быть добавлены в Отчеты — единое пространство визуализаторов в пакете. Самый быстрый способ попасть в этот раздел — кликнуть по стрелочке рядом с названием пакета в строке навигации и нажать на кнопку Отчеты.

Отчеты могут быть скомпонованы в группы для более удобной навигации.

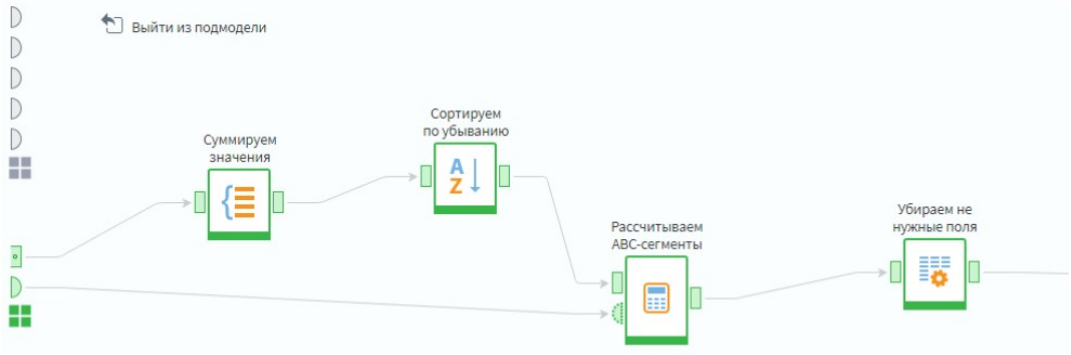


Чтобы добавить визуализатор в каталог отчетов, нужно сначала зайти в настройки визуализаторов соответствующего узла через глаз с шестеренкой. А потом через клик ПКМ по нужному отчету назначить ему существующую группу или создать новую.



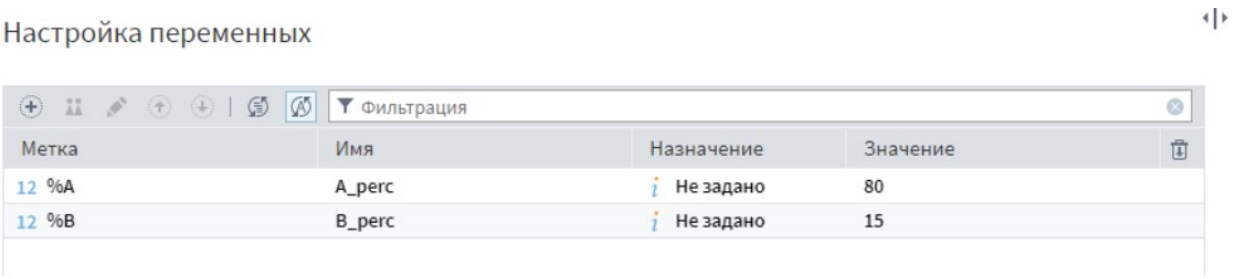
Использование переменных

Последний штрих в базовых механиках — работа с переменными. Для подмоделей можно определить порты переменных. Узел Подмодель отличается от других узлов наличием кнопки «Войти». Подмодель помимо сценария содержит входные и выходные порты: прямоугольные — таблица, полукруглые — переменные.



По точке на входном порту с таблицей можно понять, что подать на этот порт можно таблицу со строго определенным набором полей.

Вернувшись на уровень выше (кликнув «Выйти из подмодели») и щелкнув 2 раза ЛКМ по входному порту переменных на узле АВС, можно увидеть список переменных, которые создаются на этом порту.



Порты передачи данных (прямоугольные) не могут создавать данные внутри себя. Они должны быть доставлены в порт от узлов импорта или других обработчиков. В отличие от них переменные могут быть созданы в любом узле, у которого есть соответствующий порт. Там же им могут быть присвоены значения по умолчанию.

Это значит, что порты с переменными могут быть использованы как пункт быстрых настроек узла. Динамические значения переменных могут быть получены через узлы Калькулятор переменных и Таблица в переменные.

Задание.

Изучить системы интеллектуального анализа данных.

Установите Loginom Community Edition.

Изучить интерфейс программы Loginom

Скачайте архив с учебными файлами.

ZIPДанные для практики. День 2.zip

Если у вас не установлен архиватор для его открытия, то его можно скачать тут. Распакуйте содержимое архива в любую удобную папку. Главное, чтобы файл .lgp был в той же папке, что и каталог Demo Data.

Изучите в разделе отчеты имеющиеся визуализаторы. Попробуйте добавить в любой визуализатор Куб дополнительные разрезы строк и столбцов. Узел ABC по товарам делит всю номенклатуру на 3 категории: A — товары, давшие 80% продаж, B — товары, давшие 15% продаж, C — товары, давшие 5% продаж.

Измените настройки переменных подмодели, чтобы получилась следующая сегментация: A — 50% продаж, B — 30% продаж, C — 20% продаж (доля C не задается в явном виде, она считается автоматически как 100-A-B). Зайдите в любую подмодель ABC и проследите, как переменные с ее входа используются внутри подмодели. Какое поле рассчитывается с применением этих переменных?

Добавьте в узле «Дата + время» новый визуализатор Куб, попробуйте собрать свой отчет. Попробуйте добавить после узла Загрузка продаж узел Фильтр строк, и подать в сценарий только те продажи, которые произошли с 01.01.2022 по 31.12.2022. Определите через предпросмотр данных (двойной клик ЛКМ по выходному порту), сколько записей получилось в таблице, выходящей из узла ABC по клиентам.

Походите по настройкам разных узлов, попробуйте их поменять. Обратите особое внимание на узел, рассчитываем ABC-сегменты в подмоделях.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1. Что такое методы ИАД и каково их назначение?
2. Охарактеризуйте области применения методов ИАД.
3. Каковы этапы исследований методами ИАД?

Лабораторная работа № 2.

Тема: Импорт данных из Excel и CSV

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

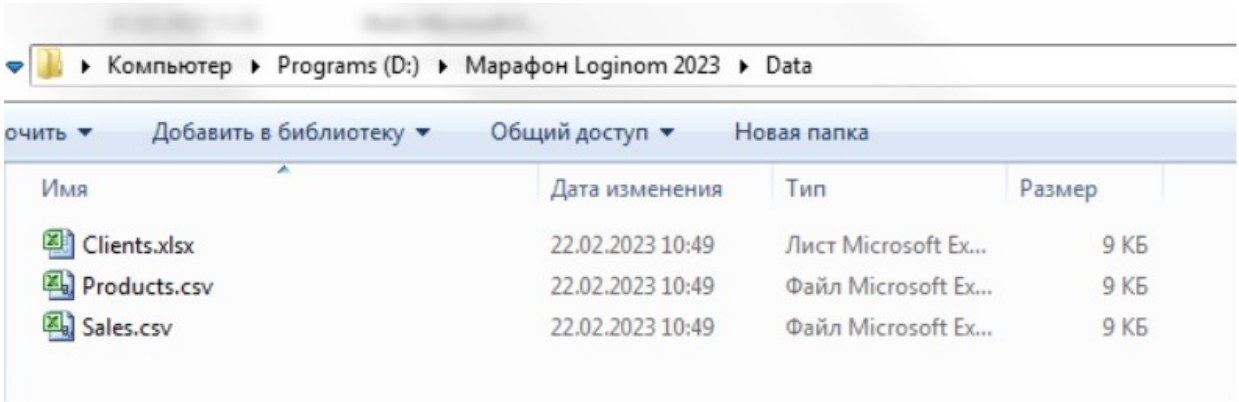
Исходный датасет состоит из 3-х таблиц:

Таблица продаж	Справочник клиентов	Справочник товаров
Дата покупки	Client_ID	Product Key
Product_key	Покупатель	SKU name (наименование товара)
Филиал	Contractor Name (юрлицо)	Brand_name (бренд товара)
Номер чека		SKU group_name (товарная группа)
Количество		
Сумма покупки		
Сумма скидки		
Client_ID		
Себестоимость		
Машина доставки		
Адрес		

Несмотря на существование множества СУБД и всевозможных API, загрузка данных из файлов остается одним из самых распространенных сценариев. В конце концов, даже если в компании есть настроенное хранилище данных, нет гарантий, что завтра не придется анализировать эту информацию вместе с какой-нибудь выгрузкой из базы таможни. Поэтому начнем с импорта данных из файлов.

Подготовка рабочего проекта

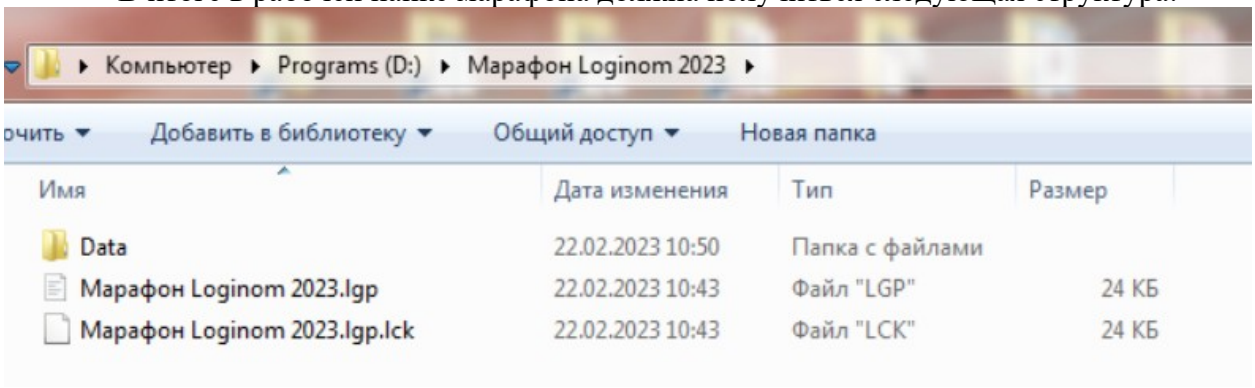
Создайте на диске в удобном месте папку с любым названием (например, «Марафон Loginom 2023»). Это будет рабочей папкой марафона. Скачайте учебные данные.**ZIPДанные для практики. День 3.zip**Разместите учебные файлы так, чтобы они находились в **рабочей папке марафона в папке Data**.



Такое расположение папок важно для легкого использования демонстрационных пакетов, т.к. по умолчанию в узлах импорта файлов прописываются относительные пути.

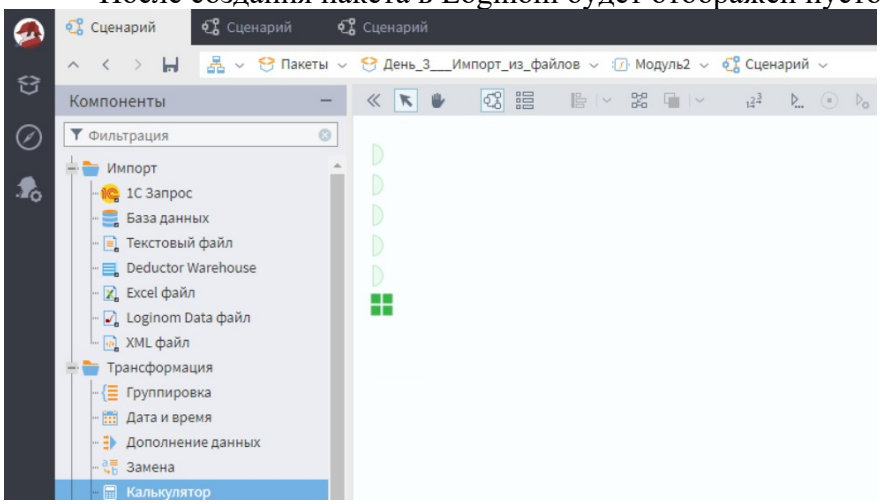
Откройте Loginom и создайте новый пакет. Вы можете назвать его как угодно. Самое главное — создать пакет внутри рабочей папки марафона, чтобы он лежал рядом с папкой **Data**.

В итоге в рабочей папке марафона должна получиться следующая структура:



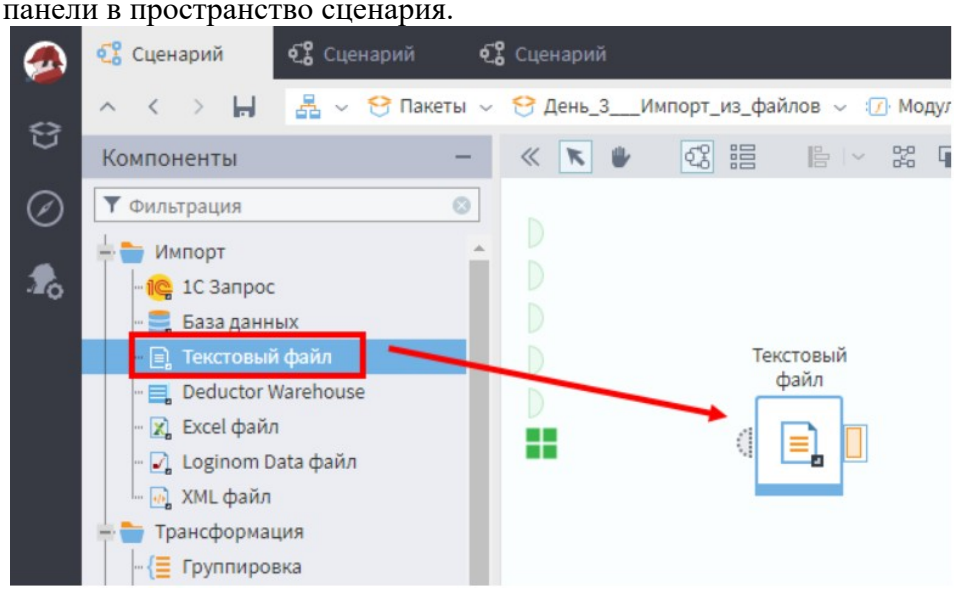
Узлы импорта из файлов

После создания пакета в Loginom будет отображен пустой сценарий.

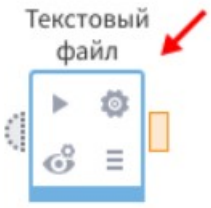


Сначала данные должны быть загружены в Loginom. За это отвечают узлы из группы Импорт, которая находится в самом верху панели компонентов. Как можно заметить, анализируемые данные представлены двумя CSV-файлами и одним файлом Excel.

Начнем с импорта CSV. За него отвечает компонент Текстовый файл. Его надо перетащить из левой панели в пространство сценария.

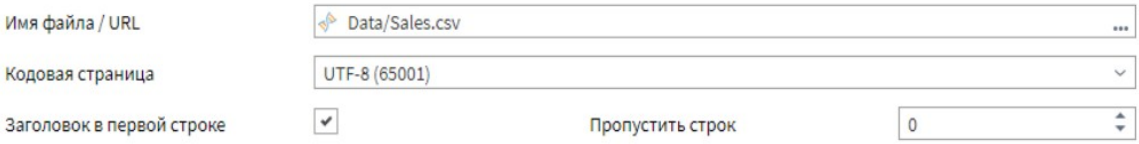


Затем узел необходимо настроить. Щелкните ЛКМ в его центре, чтобы появились пиктограммы управления узлом. Щелкните по шестеренке, чтобы перейти в настройки узла.



Настройки узла импорта осуществляются в несколько шагов. На первом нужно указать путь к файлу и определить его кодировку. Укажите путь к файлу Sales.csv в папке Data. Правильная кодировка — UTF-8.

На этом этапе можно включить определение заголовков полей по первой строке и задать количество пропускаемых строк сверху. Это требуется, когда данные начинаются не с первой строки или в таблице есть шапка с ненужной информацией.

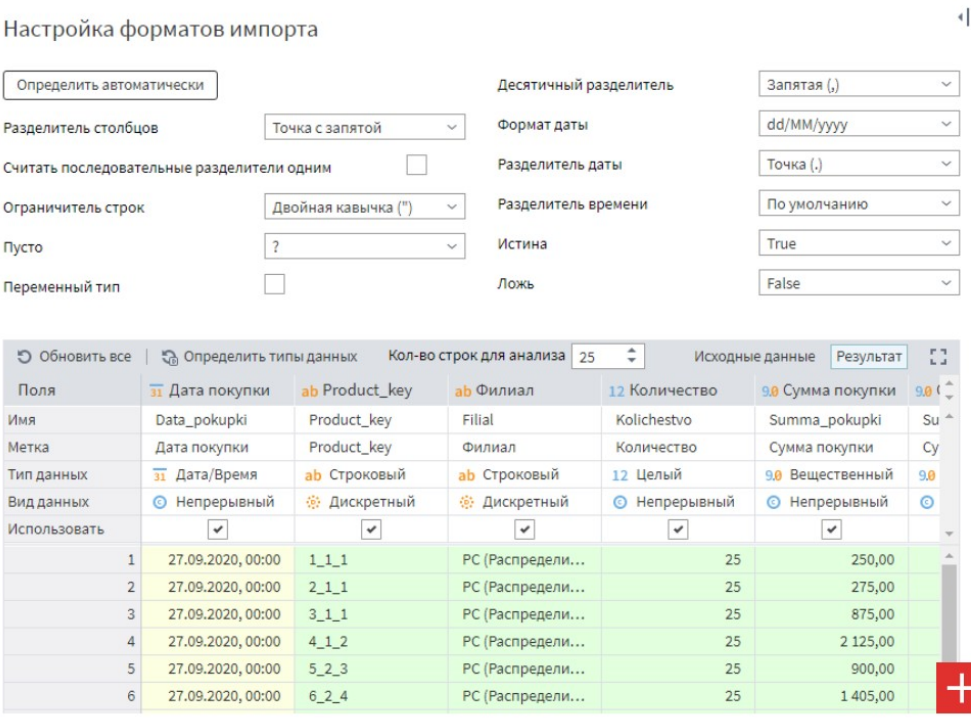



```
Дата покупки;Product_key;Филиал;Количество;Сумма покупки;Сумма скидки;Client_ID;Себестоимость;Машина доставки;
27.09.2020;1_1_1; РС (Распределительный склад);25;250;0;CL_01;-112,25;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;2_1_1; РС (Распределительный склад);25;275;0;CL_01;-183,65;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;3_1_1; РС (Распределительный склад);25;875;0;CL_01;-356,74;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;4_1_2; РС (Распределительный склад);25;2125;0;CL_01;?;H290ЛО;410779, г. Саратов, ул. Спортивная, 9,
27.09.2020;5_2_3; РС (Распределительный склад);25;900;0;CL_01;-557,94;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;6_2_4; РС (Распределительный склад);25;1405;0;CL_01;-584,88;H290ЛО;410779, г. Саратов, ул. Спортивн
27.09.2020;7_2_4; РС (Распределительный склад);25;1080;0;CL_01;-628,08;H290ЛО;410779, г. Саратов, ул. Спортивн
27.09.2020;8_3_5; РС (Распределительный склад);25;475;0;CL_01;-144,17;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;9_1_4; РС (Распределительный склад);25;325;0;CL_01;-122,14;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;10_2_3; РС (Распределительный склад);25;1692,5;0;CL_01;-1086,41;H290ЛО;410779, г. Саратов, ул. Спор
27.09.2020;11_2_3; РС (Распределительный склад);25;935;0;CL_01;-495,8;H290ЛО;410779, г. Саратов, ул. Спортивна
27.09.2020;12_2_3; РС (Распределительный склад);25;1125;0,0225000000000364;CL_01;-501,63;H290ЛО;410779, г. Сар
27.09.2020;13_2_3; РС (Распределительный склад);25;1260;1,4210854715202E-13;CL_01;-471,93;H290ЛО;410779, г. Са
27.09.2020;14_2_3; РС (Распределительный склад);25;1485;0;CL_01;-851,33;H290ЛО;410779, г. Саратов, ул. Спортив
28.09.2020;15_4_6; РС (Распределительный склад);20;5500;0;CL_11;-1663,41;H290ЛО;400586, г. Волгоград, ул. Моло
30.09.2020;16_5_7; ТЗ-2 ( Ленинградское шоссе);10;251;0;CL_21;-83,95;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;17_5_7; ТЗ-2 ( Ленинградское шоссе);10;251;0;CL_21;-151,97;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;18_6_8; ТЗ-2 ( Ленинградское шоссе);10;290;-0,0069999999999837;CL_21;-112,38;A645ДМ;400140, г. Волг
30.09.2020;19_6_8; ТЗ-2 ( Ленинградское шоссе);10;304;0;CL_21;-168,22;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;20_6_8; ТЗ-2 ( Ленинградское шоссе);10;304;0;CL_21;-142,55;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;21_6_8; ТЗ-2 ( Ленинградское шоссе);10;304;0;CL_21;-196,35;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;22_6_8; ТЗ-2 ( Ленинградское шоссе);10;304;0;CL_21;-211,01;A645ДМ;400140, г. Волгоград, ул. Ленина,
30.09.2020;23_6_8; ТЗ-2 ( Ленинградское шоссе);10;290;-0,0069999999999837;CL_21;-161;A645ДМ;400140, г. Волгогр
30.09.2020;24_6_8; ТЗ-2 ( Ленинградское шоссе);10;304;0;CL_21;-139,84;A645ДМ;400140, г. Волгоград, ул. Ленина,
```

Предпросмотр покажет, как Loginom считал содержимое файла. На этом этапе не заданы разграничители, и важно проверить, что в тексте нет «кракозябр». Это значит, что кодировка настроена правильно.

На втором шаге импорта размечается структура загружаемой таблицы. При импорте из текстовых файлов нужно задать настройки разделителей и ограничителей строк, чтобы данные не «поплыли» в итоговой таблице.

Если нет уверенности в правильности настроек, можно нажать на кнопку «Определить автоматически». Loginom достаточно точно определяет структуру файла самостоятельно. Кстати, нажмите ее на всякий случай.



Внизу отображается таблица, которую Loginom распознал. Здесь можно провести дополнительную настройку полей. Что тут интересного?

Во-первых, наименование полей в Loginom определяется двумя параметрами:

Имя поля — это техническое наименование на латинице без пробелов. Используется как технический идентификатор поля. В рамках одной таблицы должен быть уникальным.

Метка поля — произвольное наименование поля, допустима кириллица и пробелы. Используется для задания человекопонятных наименований, отображается в отчетах и всевозможных подписях.

Автоматически определенные имена и метки нас полностью устраивают — нет необходимости их менять.

Во-вторых, требуется настройка типа и вида данных. В Loginom используется строгая типизация, следовательно, нужно для каждого поля определить, данные какого типа в нем находятся.

От этого зависит, какие методы обработки и трансформации Loginom сможет применить к полям, и в какой роли он их сможет использовать.

Доступны следующие типы данных:

- логический — 1/0, Истина/Ложь или True/False;
- дата и время — данные, содержащие отметки времени и допускающие соответствующее форматирование;
- вещественный — числа с дробной частью;
- целый — числа без дробной части;
- строковый — произвольный текст;
- переменный — смешанный тип данных.

В зависимости от выбранного типа данных значения в полях подсвечиваются зеленым, желтым или красным цветом.

Поля	Дата покупки	Product_key	Филиал
Имя	Data_pokupki	Product_key	Filial
Метка	Дата покупки	Product_key	Филиал
Тип данных	Дата/Время	Строковый	Дата/Время
Вид данных	Непрерывный	Дискретный	Дискретный
Использовать	✔	✔	✔
1	27.09.2020, 00:00	1_1_1	
2	27.09.2020, 00:00	2_1_1	
3	27.09.2020, 00:00	3_1_1	
4	27.09.2020, 00:00	4_1_2	
5	27.09.2020, 00:00	5_2_3	
6	27.09.2020, 00:00	6_2_4	
7	27.09.2020, 00:00	7_2_4	
8	27.09.2020, 00:00	8_3_5	
9	27.09.2020, 00:00	9_1_4	
10	27.09.2020, 00:00	10_2_3	

Это экспресс-проверка на соответствие значения выставленному типу данных:

- Зеленый — тип точно соответствует;
- Желтый — скорее всего соответствует, но нет полной уверенности;
- Красный — точно не соответствует.

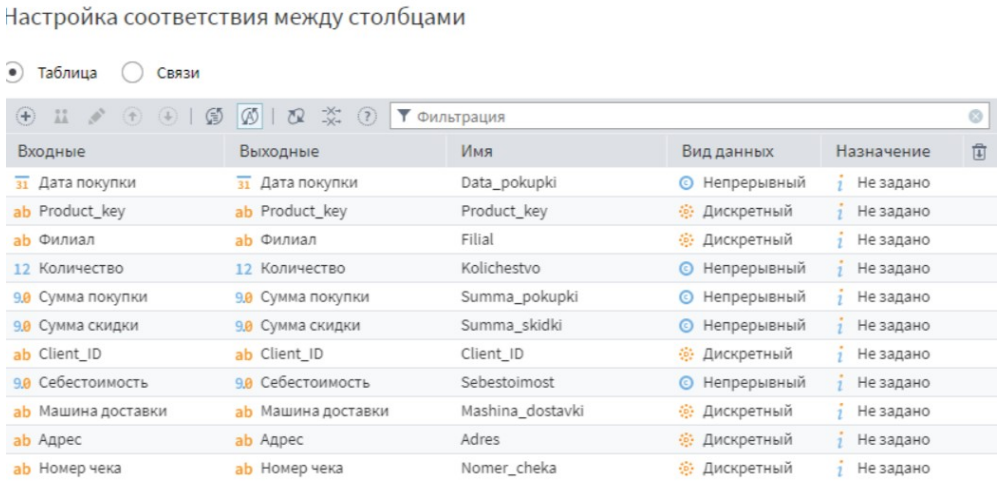
Попробуйте выставить полям разные типы и посмотреть, как реагирует проверка соответствия.

Когда наэксперементируетесь, верните все как было с помощью кнопки «Определить автоматически» слева вверху.

Параметр Вид данных имеет 2 значения: непрерывный или дискретный. Если задать «Непрерывный», то в некоторых функциях Loginom будет оценивать значения поля как последовательные значения одной оси (например, дата и время).

В частности, будет возможность искать пропуски в диапазонах значений в таких полях. Дискретный вид означает, что каждое значение — это просто отдельное значение. Например, номер чека, который может быть числовым, но при этом арифметические операции с ним лишены смысла.

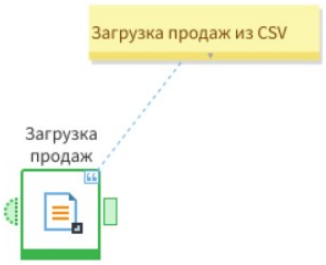
Для выполнения настроек на этом экране достаточно нажать кнопку «Определить автоматически». Сделайте это и двигайтесь на следующий шаг. Вы увидите экран «Настройка соответствия между столбцами». Если кратко, то здесь определяется, какие поля покинут узел, и как они будут называться.



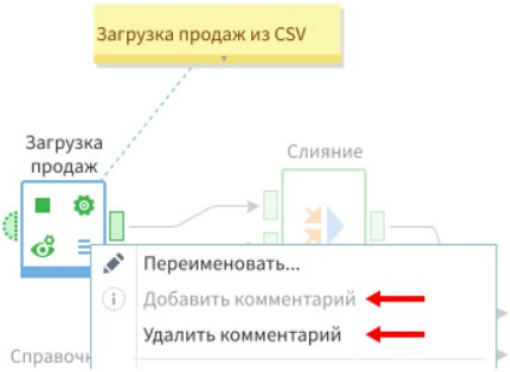
Для учебного примера этот экран не важен. Можно убедиться, что набор полей, их названия и типы данных соответствуют тому, что на картинке, и двигаться дальше.

Финальный этап настройки — присвоение метки узлу и опциональное добавление комментария.

Этот текст будет отображаться в схеме сценария. Поэтому рекомендуется добавлять подписи для улучшения читабельности.



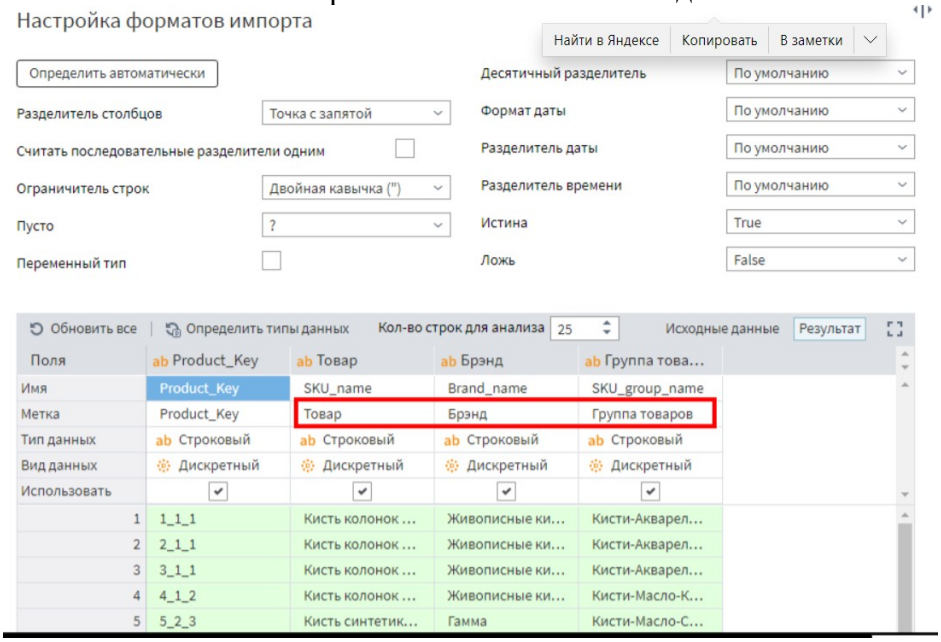
Кстати, менять подписи узлов, как и комментарии, можно через двойной клик ЛКМ по ним. Комментарии можно скрыть через клик по знаку кавычек в правом верхнем углу узла. Быстро добавить/удалить комментарий к узлу можно через контекстное меню, которое открывается по клику ПКМ на узле.



Описанный процесс надо повторить для файла Products.csv. Автоматического определения структуры файла будет достаточно. Однако для дальнейшего удобства задайте следующие метки полям:

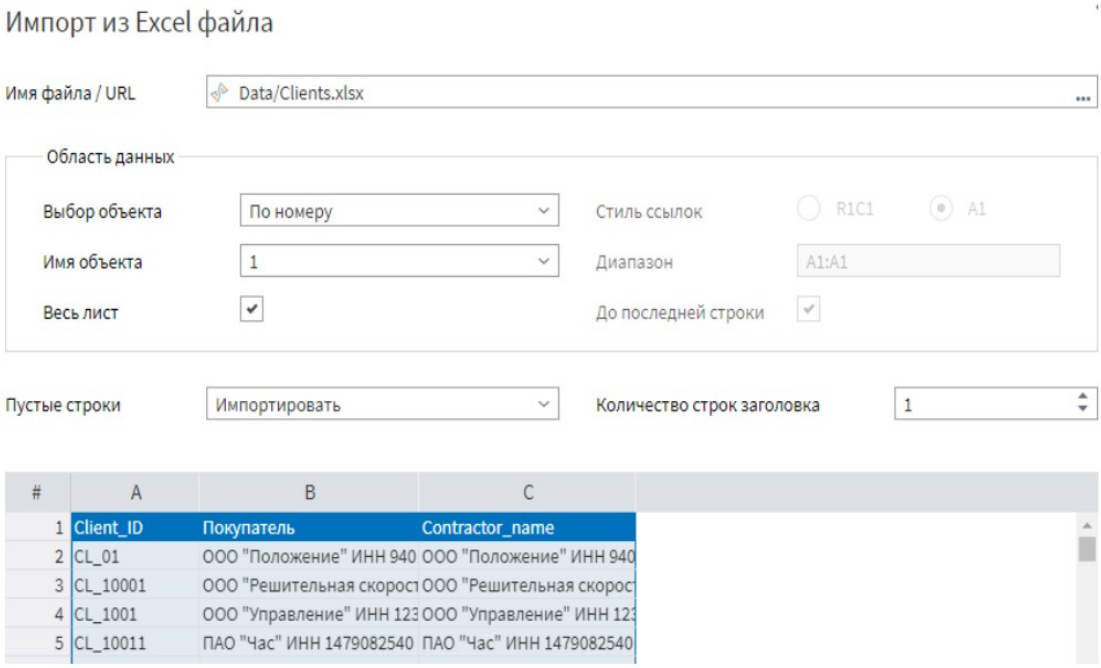
«Товар» для SKU_Name.

«Брэнд» для Brand_name.
«Группа» товаров для SKU_group_name.
Имена полей оставьте оригинальными. Все поля должны иметь текстовый тип.



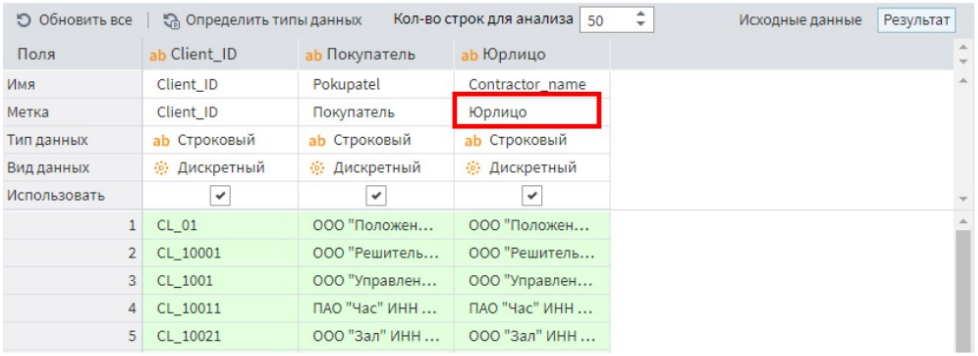
Следующий шаг — добавление узла импорта из Excel и загрузка файла **Clients.xlsx**. Можно заметить, что начальный экран импорта отличается от такового для csv-файлов.

При импорте из Excel нужно указать лист или именованный диапазон, из которого будут взяты данные. Лист задается через порядковый номер (изменение порядка листов в файле сломает импорт) или имя (изменение имени листа в файле сломает импорт).



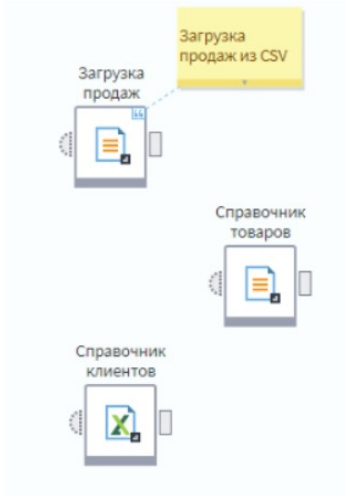
Настройку полей надо оставить как есть, только изменить метку поля «Contractor_name» на «Юрлицо». Как итог, должно загружаться 3 строковых поля.

Настройка полей

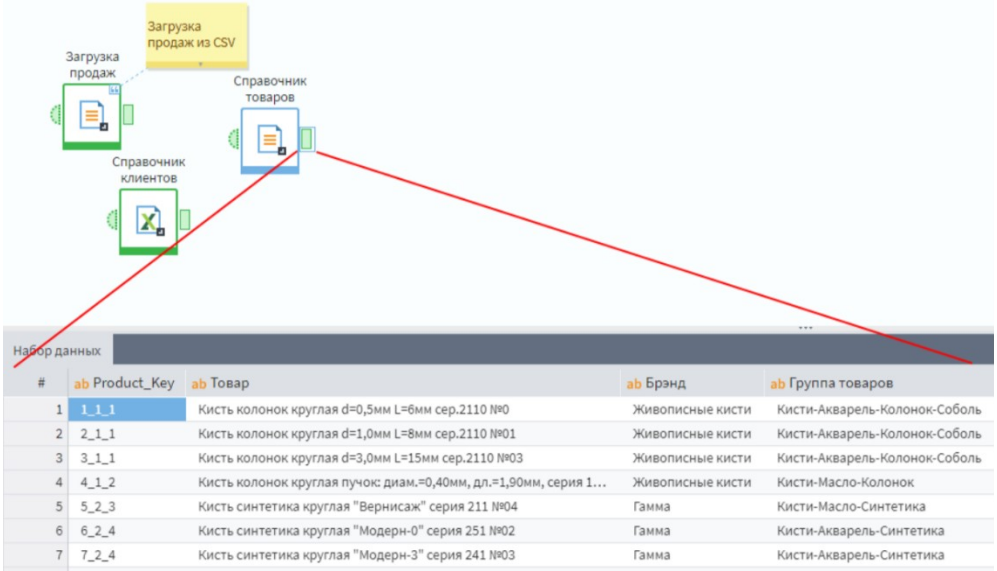


Поля	ab Client_ID	ab Покупатель	ab Юрлицо
Имя	Client_ID	Pokupatel	Contractor_name
Метка	Client_ID	Покупатель	Юрлицо
Тип данных	ab Строковый	ab Строковый	ab Строковый
Вид данных	Дискретный	Дискретный	Дискретный
Использовать	✔	✔	✔
1	CL_01	ООО "Положен...	ООО "Положен...
2	CL_10001	ООО "Решитель...	ООО "Решитель...
3	CL_1001	ООО "Управлен...	ООО "Управлен...
4	CL_10011	ПАО "Час" ИНН ...	ПАО "Час" ИНН ...
5	CL_10021	ООО "Зал" ИНН ...	ООО "Зал" ИНН ...

Сценарий будет выглядеть так:



Активируйте сценарий, чтобы все узлы стали зелеными. Кликните дважды ЛКМ по каждому выходному порту, чтобы посмотреть, какие данные были загружены.



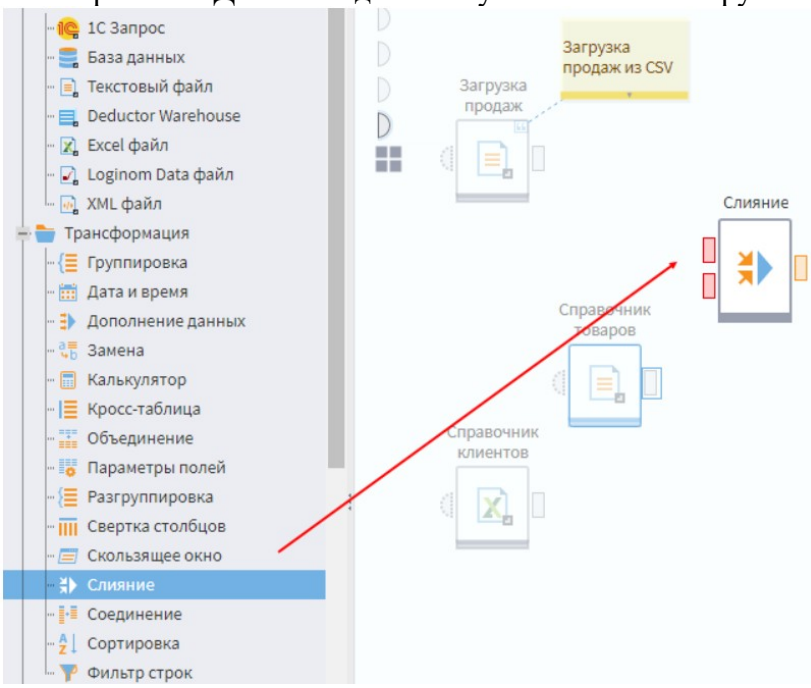
#	ab Product_Key	ab Товар	ab Брэнд	ab Группа товаров
1	1_1_1	Кисть колонок круглая d=0,5мм L=6мм сер.2110 №0	Живописные кисти	Кисти-Акварель-Колонок-Соболь
2	2_1_1	Кисть колонок круглая d=1,0мм L=8мм сер.2110 №01	Живописные кисти	Кисти-Акварель-Колонок-Соболь
3	3_1_1	Кисть колонок круглая d=3,0мм L=15мм сер.2110 №03	Живописные кисти	Кисти-Акварель-Колонок-Соболь
4	4_1_2	Кисть колонок круглая пучок: диам.=0,40мм, дл.=1,90мм, серия 1...	Живописные кисти	Кисти-Масло-Колонок
5	5_2_3	Кисть синтетика круглая "Вернисаж" серия 211 №04	Гамма	Кисти-Масло-Синтетика
6	6_2_4	Кисть синтетика круглая "Модерн-0" серия 251 №02	Гамма	Кисти-Акварель-Синтетика
7	7_2_4	Кисть синтетика круглая "Модерн-3" серия 241 №03	Гамма	Кисти-Акварель-Синтетика

Построение сводной отчетности в Loginom

Соединение таблиц через Join

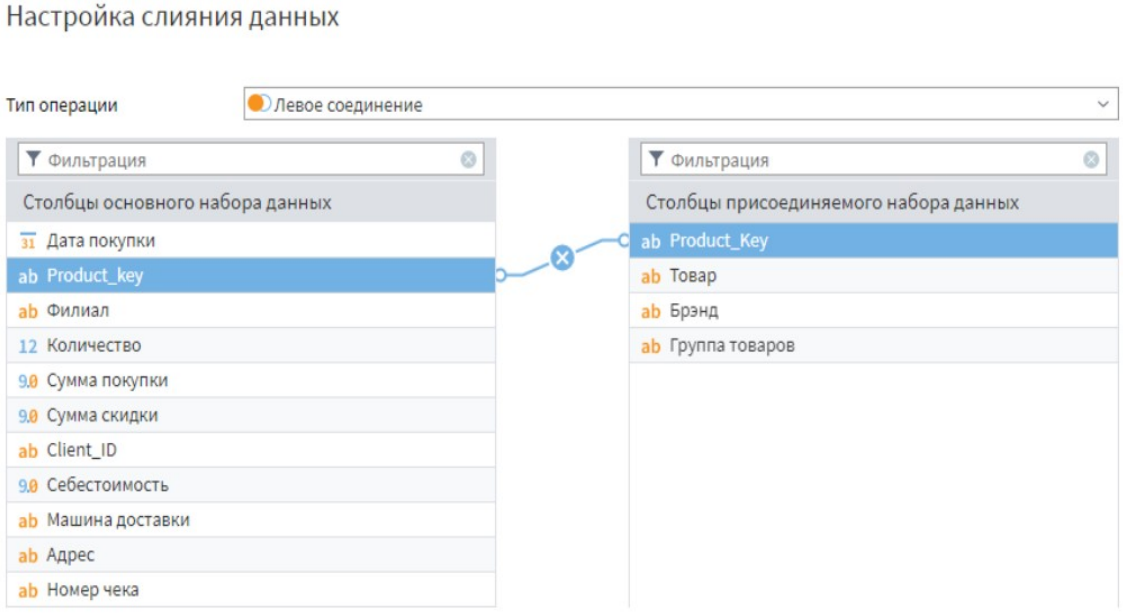
Для анализа данных из нескольких таблиц в Loginom предварительно их надо соединить в одну. По классике жанра в Loginom можно использовать соединение вида Join (добавление в таблицу столбцов другой таблицы на основе совпадающих ключевых значений) и соединение вида Union/Concatenate(добавление в таблицу строк другой таблицы).

Т.к. есть 1 таблица с данными и 2 справочника, будет правильным соединить данные со справочниками через Join. Для этого добавим узел Слияние из группы узлов Трансформация.

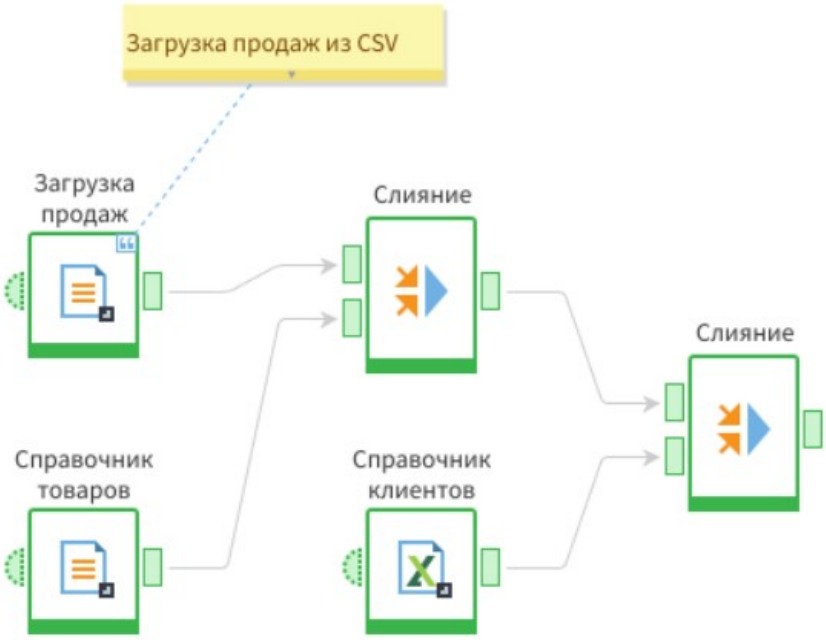


Узел Слияние всегда требует подать ему на вход 2 таблицы: основную и ту, которая будет присоединена. На верхний порт нужно подать данные по продажам, а на нижний — данные по товарам. Зайдите в настройки узла.

В мастере отображаются наборы полей 2-х таблиц. Чтобы Join сработал, нужно сопоставить ключевые поля. Свяжите эти 2 таблицы по полю Product_Key. В выпадающем списке сверху можно выбрать способ соединения таблиц. В данном случае требуется Левое соединение.



Этот же процесс надо повторить для объединенной таблицы продаж и справочника товаров, сделав левое соединение по полю Client_ID. В результате сценарий должен выглядеть вот так:

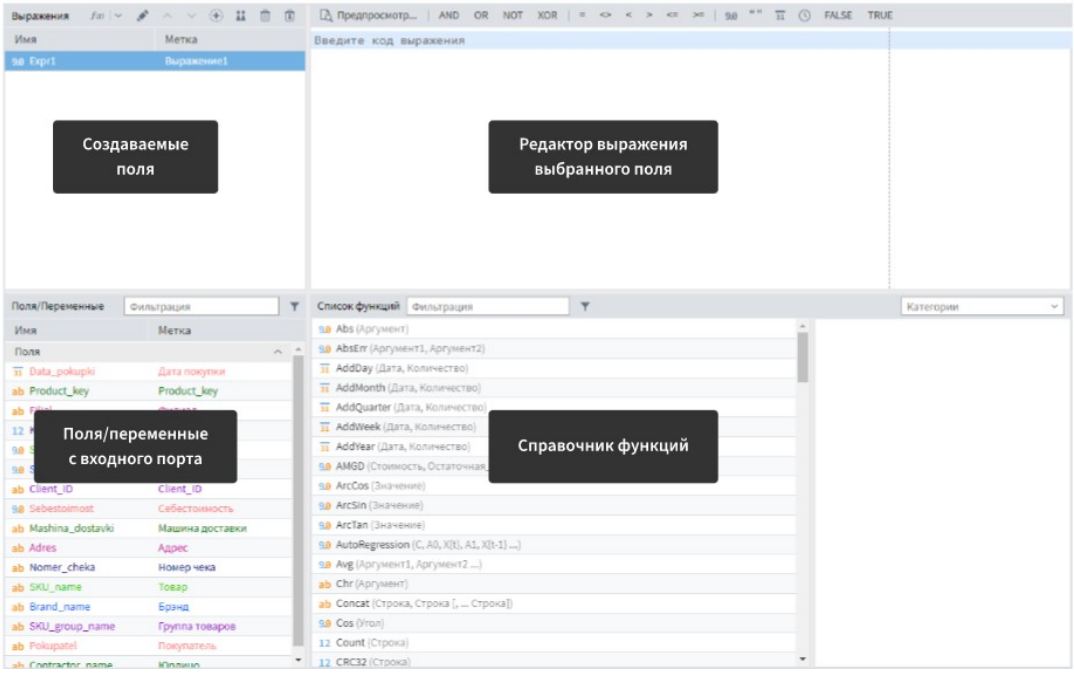


Создание вычисляемых полей

Далеко не всегда в источнике содержатся все необходимые для анализа атрибуты. Некоторые поля нужно создавать самостоятельно. Эту задачу решает узел Калькулятор из группы узлов Трансформация.

Надо добавить узел в сценарий и завести в него таблицу из второго Слияния. Далее войти в мастер настройки калькулятора. Мастер выглядит следующим образом.

Калькулятор



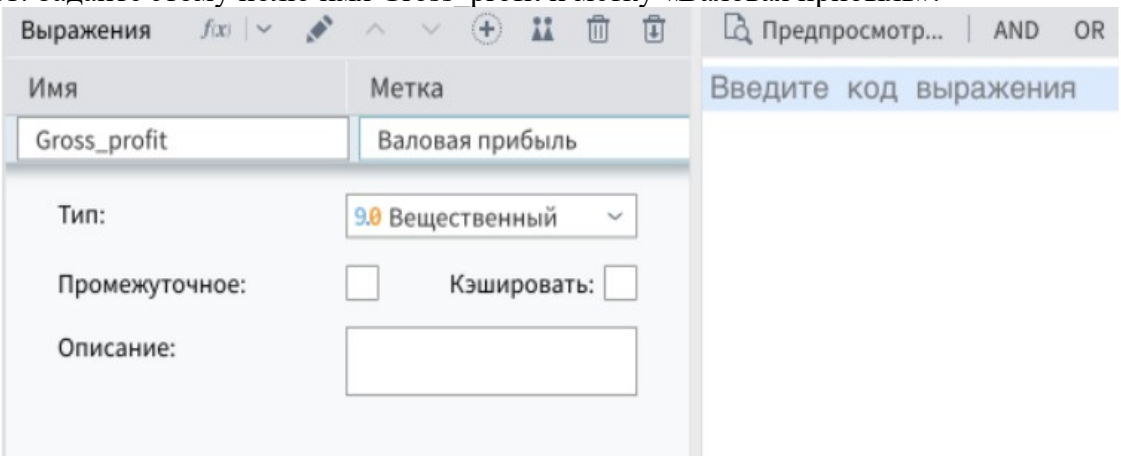
Калькулятор состоит из 4-х областей.

- Создаваемые поля. Сюда можно добавлять поля, которые будут вычисляться в узле. При этом в формулах можно ссылаться на названия других вычисляемых полей, если они стоят выше по списку.
- Редактор выражения — область для написания формул.
- Доступные поля/переменные — это то, что мы подали на вход узлу. Поля и переменные можно вставить в формулу через двойной клик по названию или через перетаскивание мышкой.

- Справочник функций — список доступных функций Loginom. Можно искать по категориям и сразу читать инструкцию для выбранной функции. Очень удобно!

В данных есть поля «Сумма покупки» и «Себестоимость». Но нет поля «Валовая прибыль».

Для его создания дважды кликните ЛКМ по заготовке нового поля, которое сейчас называется Expr1. Задайте этому полю имя Gross_profit и метку «Валовая прибыль».



Создаваемое поле имеет ряд настроек.

Тип данных определяет, какие результаты и в каком формате сможет хранить в себе поле, например, в числовое поле нельзя записать текст. Хотя в самом выражении поля можно использовать функции, относящиеся к обработке строковых значений. Главное, чтобы на выходе получалось число. Вещественный тип нас полностью устроит.

Опция «Промежуточное» исключает это поле с выходного порта калькулятора. Ее нужно использовать, когда составляются сложные выражения. Чтобы не писать формулу на 1000 символов, часть этого выражения можно разместить в промежуточное поле. Ссылаясь в итоговой формуле на промежуточное поле, проще написать и понять формулу. При этом само промежуточное поле дальше калькулятора не выйдет и не будет засорять выходную таблицу.

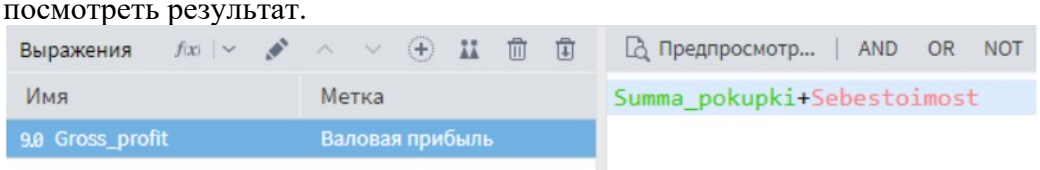
Опция «Кэширование» сохраняет рассчитанные значения поля в оперативной памяти, а не рассчитывает их на лету при обращении из других узлов. Зачем это нужно?

Для ускорения работы сценария, если используется сложная формула. Вместо затрат процессорного времени на вычисления будут браться готовые значения из оперативной памяти.

Для стабилизации значений, использующих функцию random(). Если не закэшировать значения, то в каждом последующем узле случайное число генерируется заново, и данные будут отличаться. Иногда это нужно, а иногда нет.

Чтобы активировать возможность рекурсивно ссылаться в формуле поля на само себя. Полезно для реализации накопительных вычислений и формул-счетчиков. Пример можно посмотреть в пакете второго дня, в калькуляторе подмодели ABC анализа (использование функции data).

Пропишите формулу как «Сумма покупки» + «Себестоимость», дважды кликнув ЛКМ по именам этих полей в списке внизу для быстрой подстановки. Используйте кнопку предпросмотра, чтобы посмотреть результат.



Предпросмотр показывает 100 первых строк таблицы, включая вычисляемые поля, поля которые используются в вычислениях, переменные, которые используются в вычислениях. Если в выражении есть ошибка, пользователь получит предупреждение.

Предпросмотр

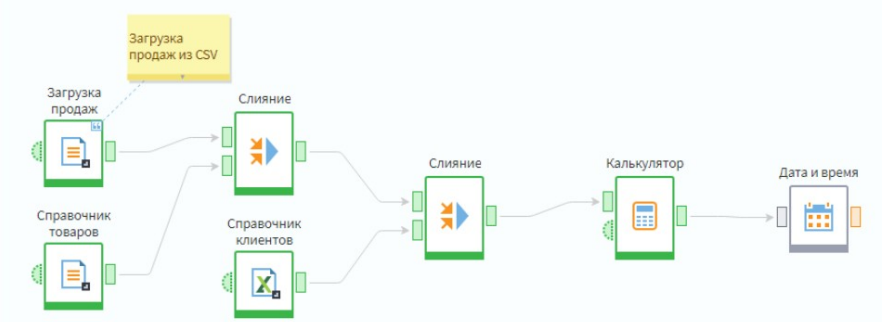
#	Валовая прибы…	Себестоимо…	Сумма покупки
42			170,00
43			170,00
44	59,72	-110,28	170,00
45	59,72	-110,28	170,00
46	512,76	-242,24	755,00
47	512,76	-242,24	755,00
48	1 106,43	-1 143,57	2 250,00
49	1 106,43	-1 143,57	2 250,00
50			951,00
51			951,00
52	165,23	-306,77	472,00
53	165,23	-306,77	472,00
54	8 114,66	-4 473,24	12 587,90

Выходите из калькулятора.

Добавление периодов

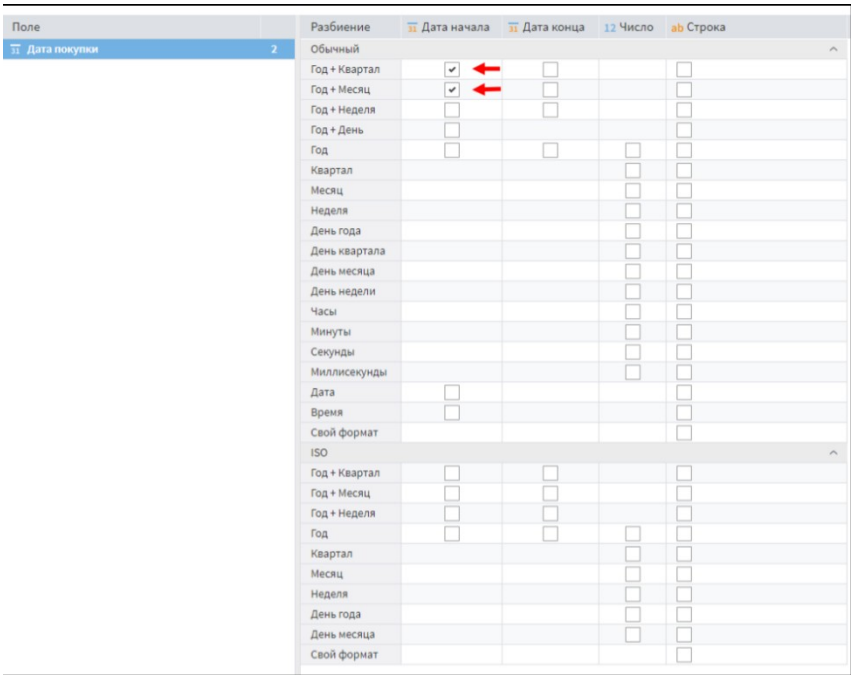
Последний штрих, прежде чем будет построен интерактивный отчет: в данных есть поле с датой продажи, но в отчетах наверняка потребуется группировка по неделям, месяцам и т.д.

Самый быстрый способ создать периоды от поля даты — использование узла Дата и время, из группы узлов Трансформация. Добавьте его в сценарий и заведите в него выход из Калькулятора.



В настройках узла для каждого поля даты можно задать множество вариантов производных периодов. Для группировки по более крупным временным периодам рекомендуется выбирать опцию «Дата начала» нужного вам периода.

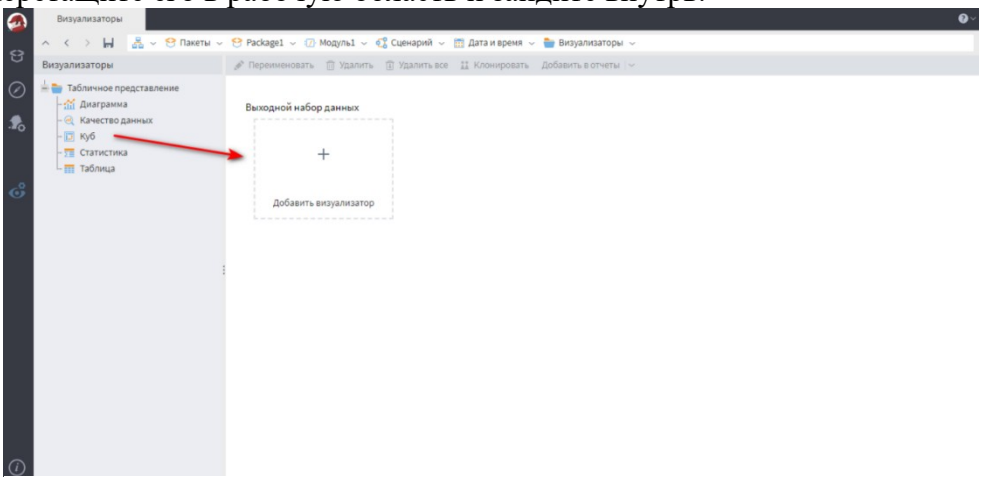
Это позволит группировать данные на нужном уровне и при этом сохранить числовую суть значений поля. Т.е. при сортировке апрель не будет идти перед декабрем, потому что первый начинается на «а».



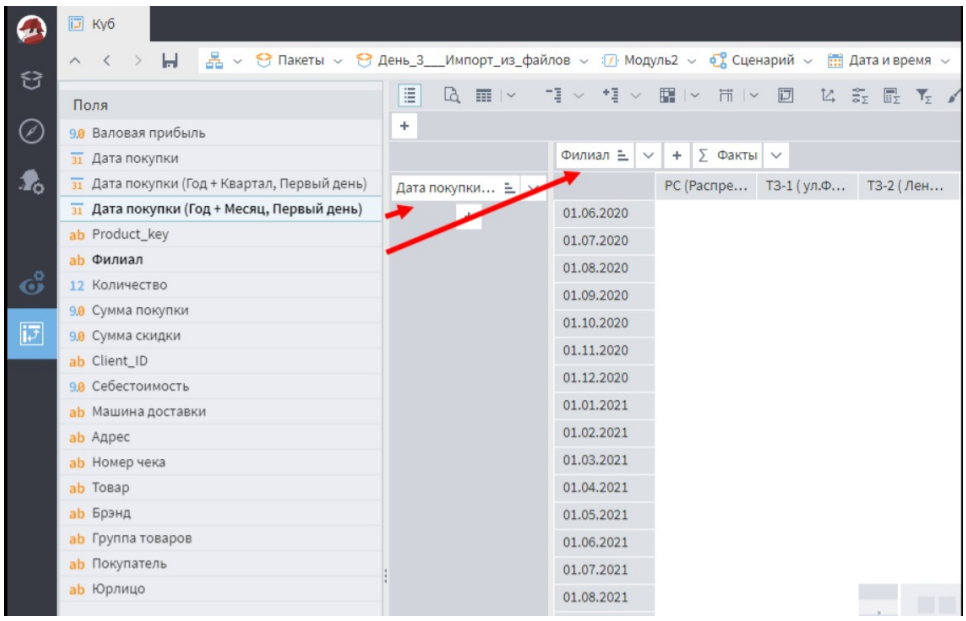
Построение отчета

Теперь можно построить отчет. Для этого требуется настроить визуализатор для порта с итоговыми данными: кликните ЛКМ в центр узла Дата и время, а потом на пиктограмму глаза с шестеренкой.

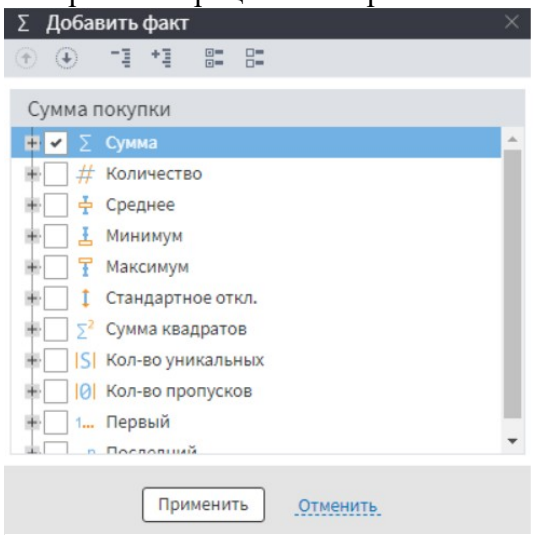
Главный инструмент для построения интерактивных отчетов в Loginom — визуализатор Куб. Перетащите его в рабочую область и зайдите внутрь.



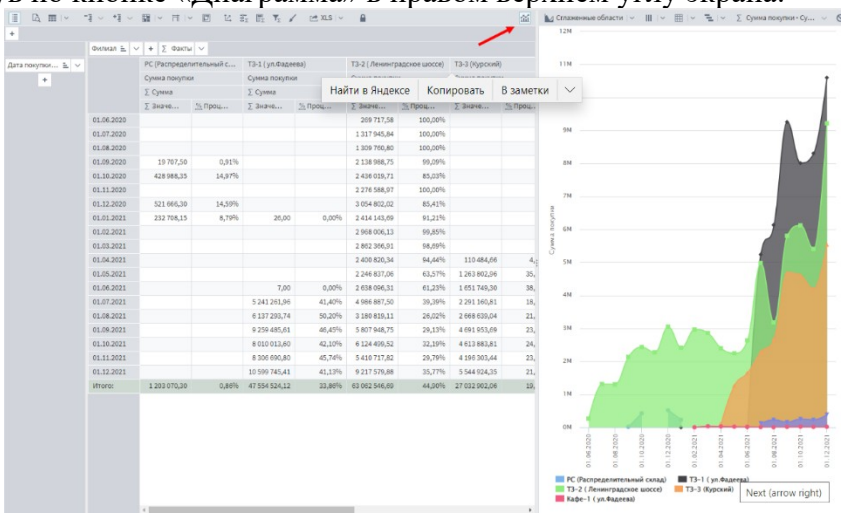
Для человека, хоть раз работавшего со сводными таблицами, все должно быть очень знакомо. Из списка полей слева перетащите месяц покупки в разрез строк, а поле «Филиал» — в столбцы.



Перетащите в центр таблицы поле «Сумма покупки». Так добавляются поля для расчета показателей. При добавлении показателя (факта) надо выбрать способ агрегации. Можно отметить несколько вариантов сразу. По умолчанию выбирается «Сумма». Раскройте его и дополнительно отметьте вариант «Процент по горизонтали».



Размещенные в таблице данные можно визуализировать на диаграмме, если открыть ее, кликнув по кнопке «Диаграмма» в правом верхнем углу экрана.



Несмотря на простой вид, у визуализатора Куб есть множество интересных функций, которые будут рассмотрены на будущих занятиях. Но если интересно узнать детали, можно посмотреть ролик.

Задание.

ZIPДанные для практики. День 3.zip

Поэкспериментируйте с добавлением показателей и разрезов в Куб. Напишите в чат, какие проблемы в данных вы уже подметили.

Задачка со звездочкой. Создайте визуализатор вида Диаграмма, в котором отражался бы график продаж по месяцам. Учитывайте, что для этого визуализатора необходимо использовать массив данных, в котором одному месяцу будет соответствовать одно значение суммы. Попробуйте сформировать такой массив с помощью узла Группировка.

Бонусная задача. Добавьте в одну группу на панели отчетов визуализаторы Куб и Диаграмму, чтобы их можно было найти в одном пространстве, а не переключаться между узлами сценария.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1.Что такое методы ИАД и каково их назначение?

2.Охарактеризуйте области применения методов ИАД.

3.Каковы этапы исследований методами ИАД?

Лабораторная работа № 3.
Тема: Импорт из баз данных на примере SQLite

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

Это типичная ситуация — консолидация данных чаще всего предполагает загрузку информации из разнородных источников: файлы, базы данных, учетные системы, веб-сервисы. И Loginom предоставляет инструменты для объединения этих данных в один массив. Давайте разбираться, как платформа позволяет работать с базой данных.

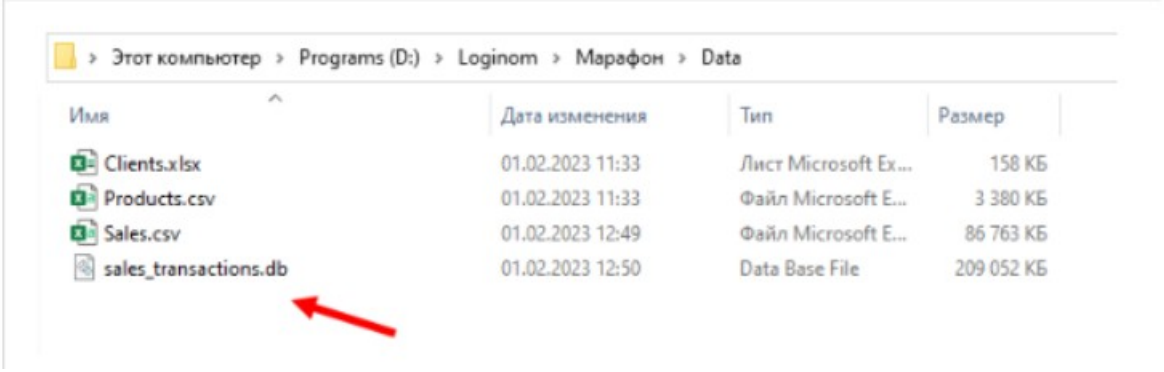
Подготовка к занятию

Скачайте учебную базу данных и разместите ее в папке Data, где уже лежат табличные файлы с предыдущего занятия. Если вы уже загружали данный файл в предыдущем занятии, то повторно скачивать не требуется.

ZIPДанные для практики. День 4.zip

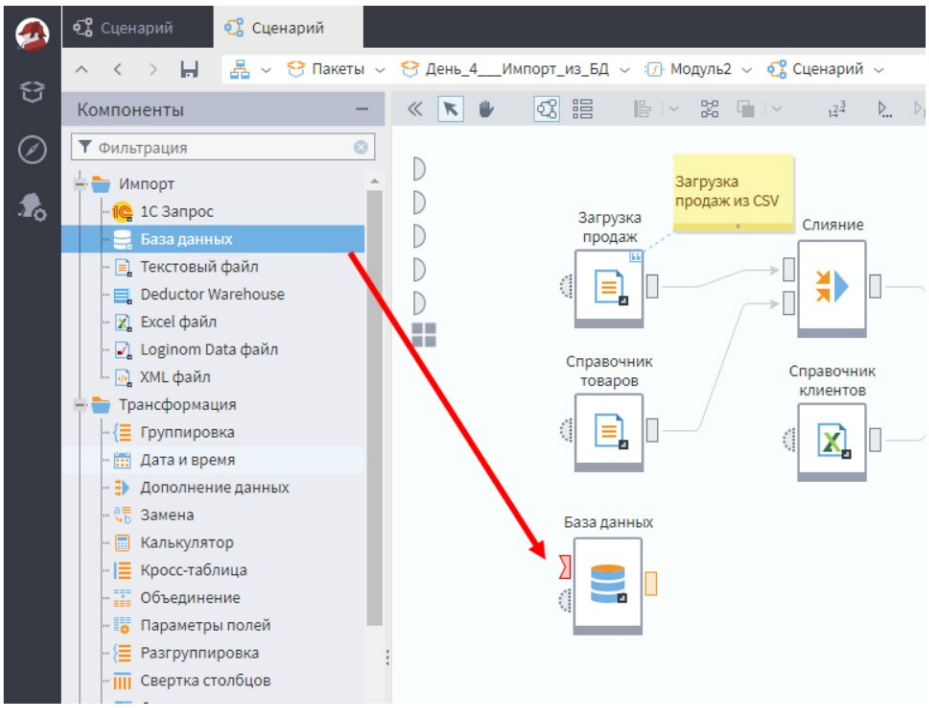
Важно! Если вы не сделали практику в теме «Импорт данных из Excel и CSV», то предварительно требуется открыть решение задания предыдущего дня, скачав его ниже.

ZIPРешение практического задания. День 3.lgp



Импорт из базы данных

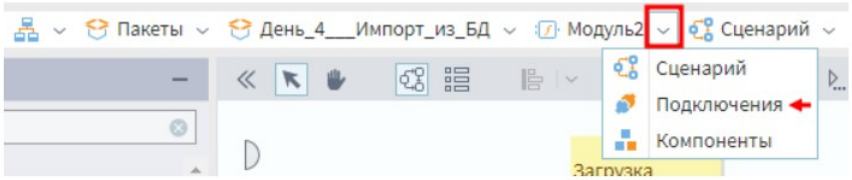
Возможно, вы уже заметили наличие в разделе Импорт узла «База данных». Перетащите его в сценарий.



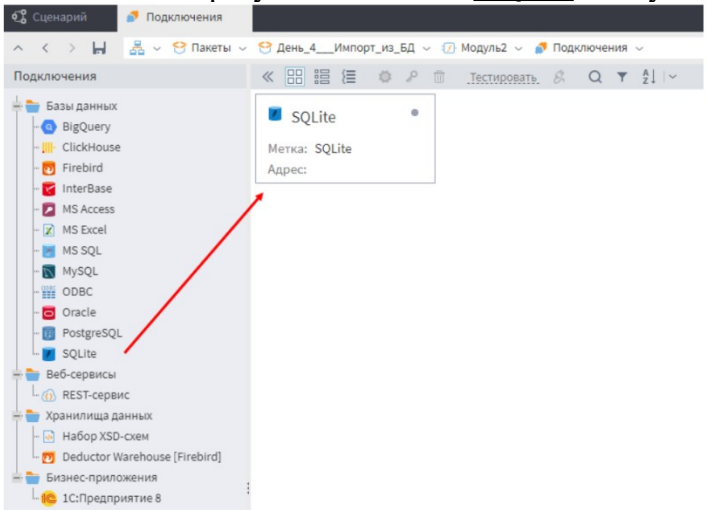
Импорт из базы данных отличается от работы с файлами. На узле имеется обязательный входной порт непонятной формы. Это порт настроек подключения, но откуда возьмутся эти настройки? Мы создадим свой узел подключения к БД.

Помните, говорилось что каждый модуль содержит в себе 3 раздела: Сценарий, Подключения и Компоненты? Вот в раздел Подключения и надо попасть.

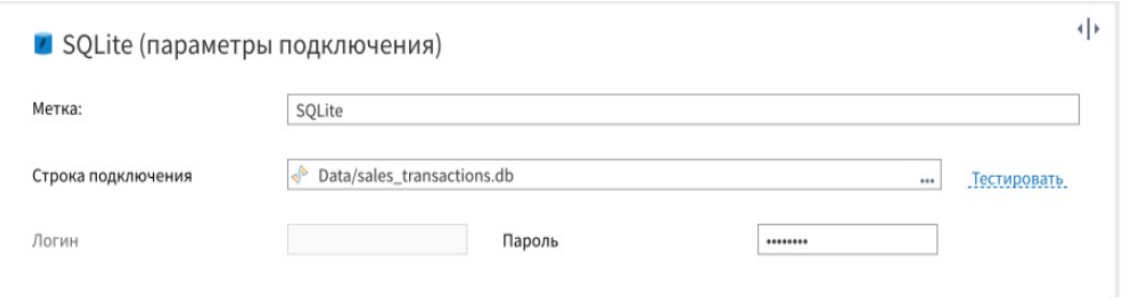
Проще всего это сделать через панель навигации, щелкнув по стрелочке у названия модуля.



Раздел подключений содержит список доступных драйверов для коннекта к базам и другим источникам. Нас интересует база данных SQLite. Ее нужно перетащить в рабочую область.



В открывшемся окне настроек надо указать путь до файла базы в папке Data и нажать Тестировать. Должно появиться сообщение об успешном подключении. Все прочие настройки можно оставить как есть.



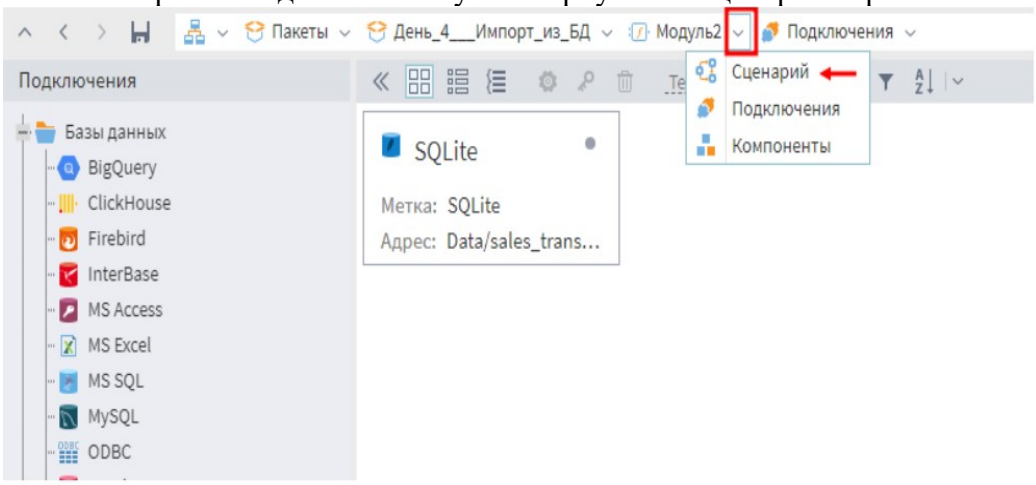
В случае коннекта к реальным базам скорее всего придется указывать реквизиты подключения и пароли. Если, конечно, рабочая база компании не SQLite, которая хранится на локальном компьютере без пароля ;)

Интересный факт: в Loginom есть встроенная СУБД SQLite. Это значит, что вы можете средствами Loginom создавать SQLite-базы, а также использовать любые поддерживаемые этой базой SQL-запросы.

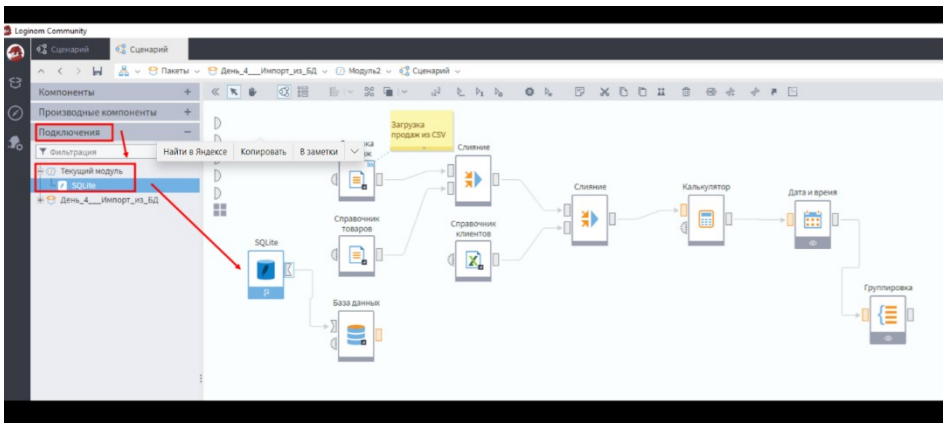
Обычно возможности выполнения запросов к базам из внешних источников определяются особенностями драйвера (коннектора), через который идет подключение. Скажем, подключение через ODBC к MSSQL может оказаться функциональнее чем «родной» коннектор аналитической системы. Например, «родной» коннектор может разрешать запросы только типа SELECT, в то время как ODBC — INSERT, TRUNCATE и множество других функций.

И в этом плане коннект между Loginom и SQLite — самый что ни на есть «родной». Он не требует установки дополнительного клиента или драйвера. Это значит, что в сценариях при необходимости можно использовать все операторы SQL в рамках возможностей SQLite без установки сторонней СУБД. Для повышения быстродействия SQLite-база может быть развернута в оперативной памяти.

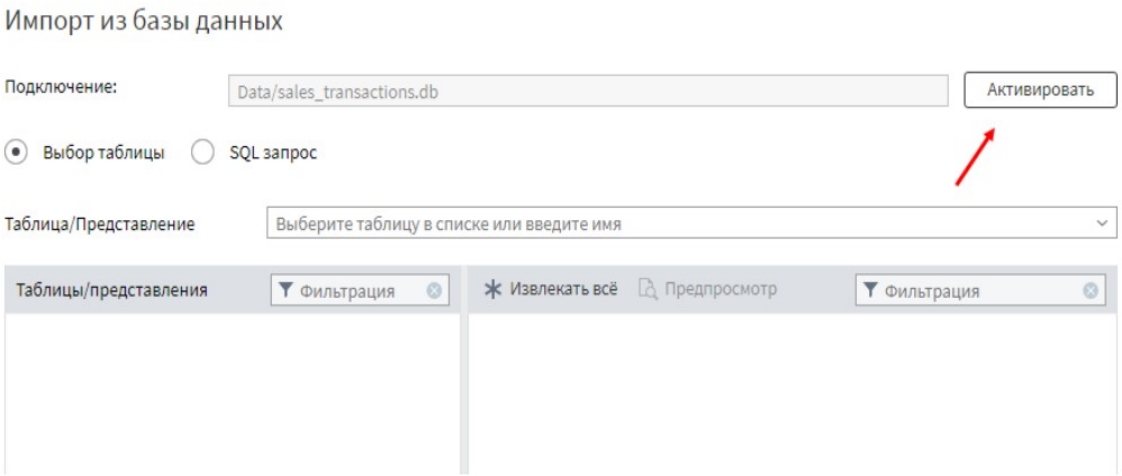
После настройки подключения нужно вернуться в сценарий через панель навигации.



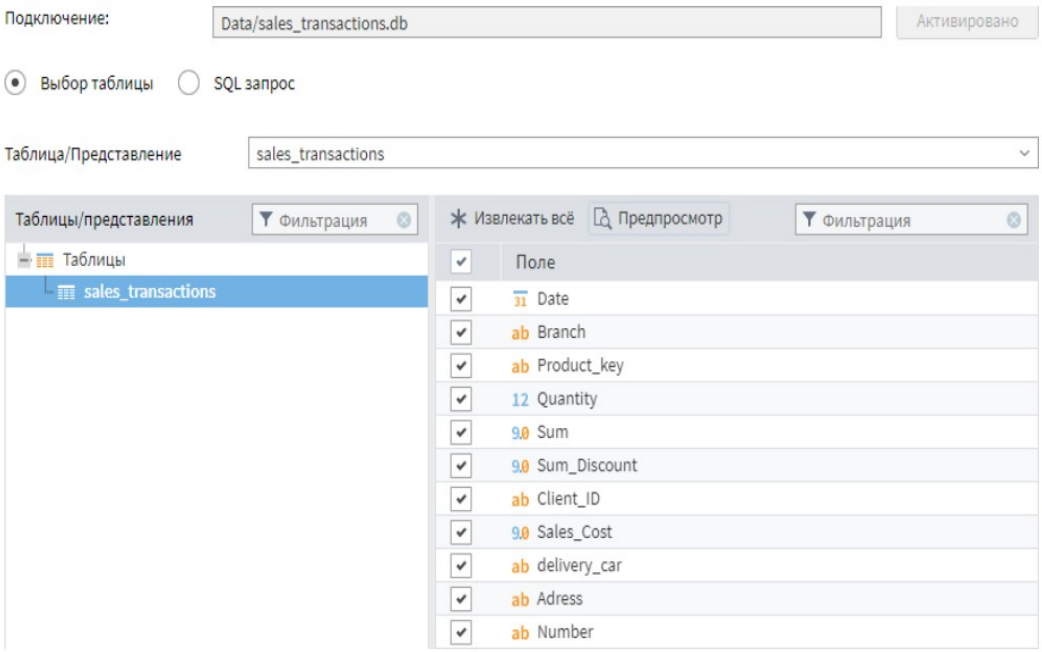
В левой области экрана надо развернуть раздел «Подключения», в нем — подраздел «Текущий модуль», перетащить созданное подключение в сценарий и соединить его с узлом импорта из базы данных.



Далее нужно запустить мастер настройки импорта из БД. Если узел подключения SQLite не был активирован, области с названиями таблиц и полей будут пустыми. Активировать подключение можно, не выходя из настроек узла импорта, через соответствующую кнопку.

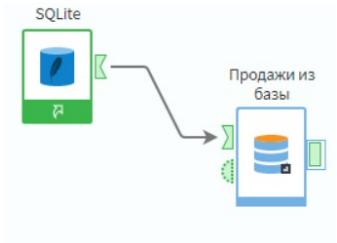


После активации в настройках импорта из БД станут видны доступные таблицы, и можно отметить желаемые для импорта поля. Нам нужны все поля.



Интересный факт: загрузку данных и другие операции с базой можно выполнять не через выбор таблицы и списка полей, а посредством SQL-запроса. В этот запрос можно подставлять значения, которые берутся из списка переменных, приходящих на вход узлу импорта из БД.

Активируем импорт из БД и получаем на выходе таблицу с данными.



Набор данных

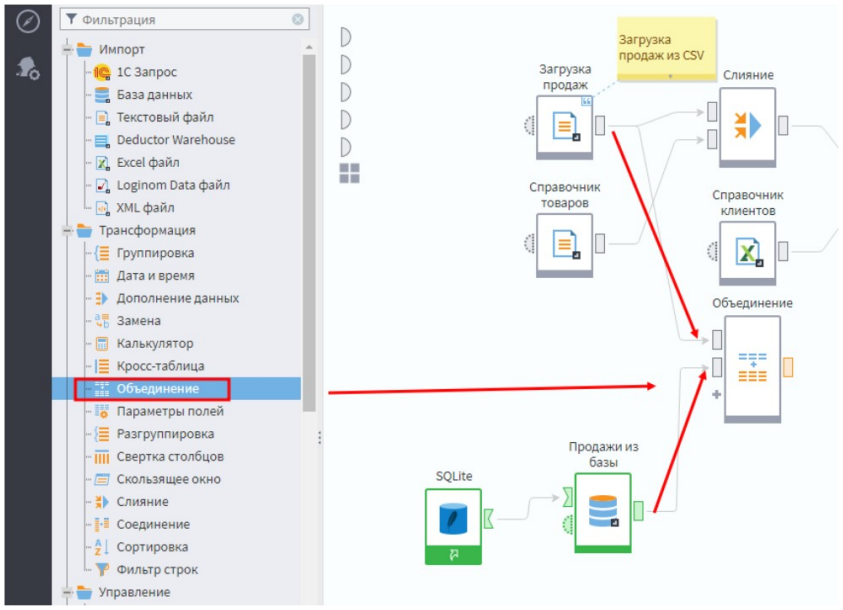
#	Date	Branch	Product_k...	Quantity	Sum	Sum_Disc
1	01.02.2022, 00:00	ТЗ-3 (Курский)	8186_2_192	5	111,55	
2	01.02.2022, 00:00	ТЗ-3 (Курский)	7170_2_192	5	111,55	
3	01.02.2022, 00:00	ТЗ-3 (Курский)	7171_2_192	5	111,55	
4	01.02.2022, 00:00	ТЗ-3 (Курский)	8087_2_192	5	111,55	
5	01.02.2022, 00:00	ТЗ-3 (Курский)	8187_2_192	5	111,55	
6	01.02.2022, 00:00	ТЗ-3 (Курский)	7172_2_192	5	111,55	

Теперь предстоит соединить их с данными из файла.

Объединение таблиц

Набор полей в данных из БД, их типы и содержание совпадают с таблицей из csv-файла (нам очень повезло ;)). Объединение однородных таблиц делается с помощью узла Объединение из группы узлов Трансформация. Было бы логично сначала сформировать объединенную таблицу продаж, а потом соединить ее со справочником.

Для этого надо добавить узел Объединение в сценарий, подключить в первый порт таблицу продаж из csv-файла, а во второй порт — таблицу продаж из БД.



Обратите внимание, хотя выход из cvs-импорта продаж подключился к узлу Объединение, его связь с узлом Слияние не пропала. Один выход может быть соединен с несколькими входами, но один вход соединяется только с одним выходом. Так можно использовать один набор данных в нескольких ветках сценария.

На узле Объединение под вторым портом есть знак +. Это значит, что данный обработчик позволяет добавлять опциональные порты для множественного входа данных. Таким образом, в одном узле можно объединить больше 2-х таблиц.

Зайдите в настройки узла Объединение. Здесь можно сопоставить для таблицы из первого входа наименования полей из остальных входов. Можно отметить галочку в шапке списка, чтобы Loginom нашел совпадения автоматически. Либо можно задавать сопоставление вручную.

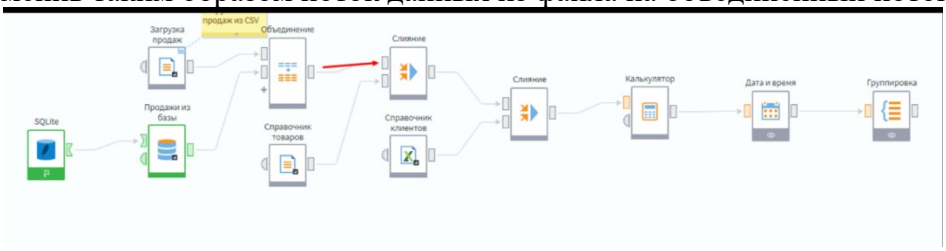
Так или иначе, результат должен получиться вот таким:

Объединение

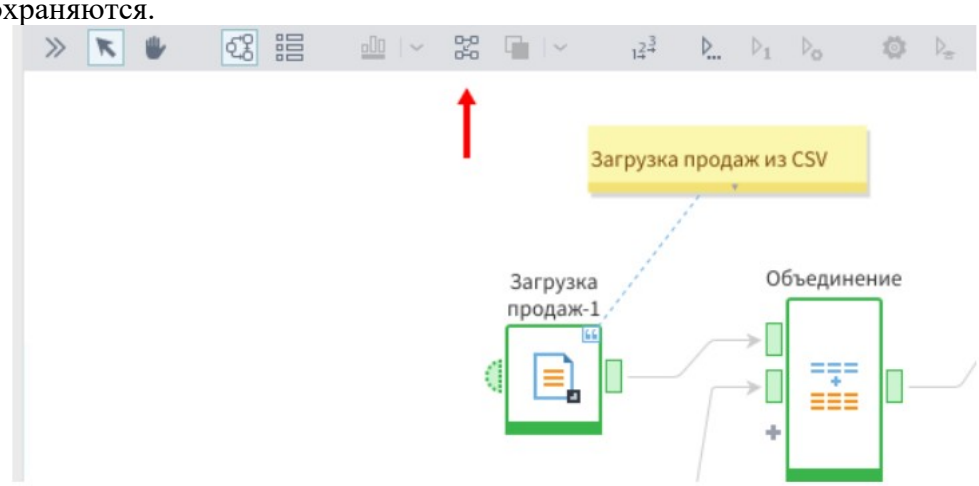
№	Главная таблица		Присоединяемая таблица
1	Дата покупки	✓	Date
2	Product_key	✓	Product_key
3	Филиал	✓	Branch
4	Количество	✓	Quantity
5	Сумма покупки	✓	Sum
6	Сумма скидки	✓	Sum_Discount
7	Client_ID	✓	Client_ID
8	Себестоимость	✓	Sales_Cost
9	Машина доставки	✓	delivery_car
10	Адрес	✓	Adress
11	Номер чека	✓	Number

В первый порт подавалась таблица из csv-файла с размеченными метками полей на русском языке, поэтому именно так поля и будут называться в итоговой объединенной таблице.

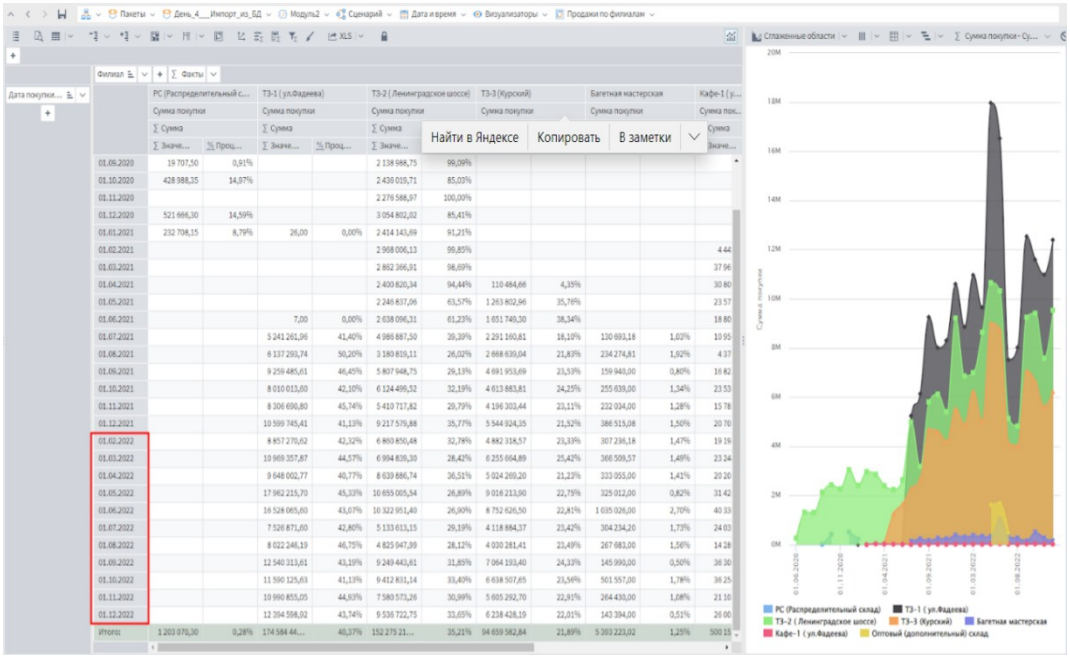
Затем нужно подключить выходной порт Объединения к первому порту первого узла Слияние, заменив таким образом поток данных из файла на объединенный поток данных из файла и БД.



Чтобы удобно расположить элементы сценария, воспользуйтесь кнопкой автоматического упорядочивания узлов на панели управления. Узлы будут расставлены в порядке их выполнения. При этом объекты вы все равно сможете перемещать для улучшения читабельности — позиции узлов сохраняются.



Активируйте весь сценарий. Откройте созданный в предыдущем занятии визуализатор. Вы должны увидеть, что добавились месяцы из 2022 года.



Задание.

Чтобы пользователи получили как можно больше практических кейсов работы, в Loginom существуют демопримеры. Если вы — продвинутый аналитик, смело листайте материалы, там много необычных лайфхаков и тонкостей работы с платформой. Если начинающий — welcome во вложенный блок каждого демопримера для уточнения деталей. Это позволяет пользователям рациональнее воспринимать информацию, исходя из уровня знаний каждого.

Все примеры можно скачать бесплатно или посмотреть демонстрацию в онлайн-режиме. Сценарии спроектированы на тестовых наборах данных, при этом их можно изменить под свои требования.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1.Охарактеризуйте метод «Деревьев решений».
2.Охарактеризуйте особенности регрессионного анализа в методах ИАД.
3.Охарактеризуйте модели временных рядов с запаздываниями.
4.Охарактеризуйте метод «Ближайшего соседа».
5.Охарактеризуйте метод поиска правила.
6.Охарактеризуйте метод кластеризации.
7.Охарактеризуйте метод классификации.
8.Охарактеризуйте метод дискриминации.
9.Какие различия в целях и алгоритмах статистического и интеллектуального подходов.

Лабораторная работа № 4.
Тема: Сегментация клиентов и построение финансового портрета контрагентов разных типов

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

Любой человек, знакомый с маркетингом, понимает, насколько важно знать о своем клиенте как можно больше. Информация о его предпочтениях, особенностях поведения, пожизненной ценности, склонности к оттоку и любых поведенческих характеристиках способна в разы увеличить доходы компании или снизить издержки.

Можно попробовать собрать информацию о поведении клиентов за счет опросов или маркетинговых исследований, но это сложно, долго, дорого и ненадежно. Намного лучше извлекать инсайты из истории покупок. Покупки расскажут о клиенте очень много, даже то, что он сам про себя не знает.

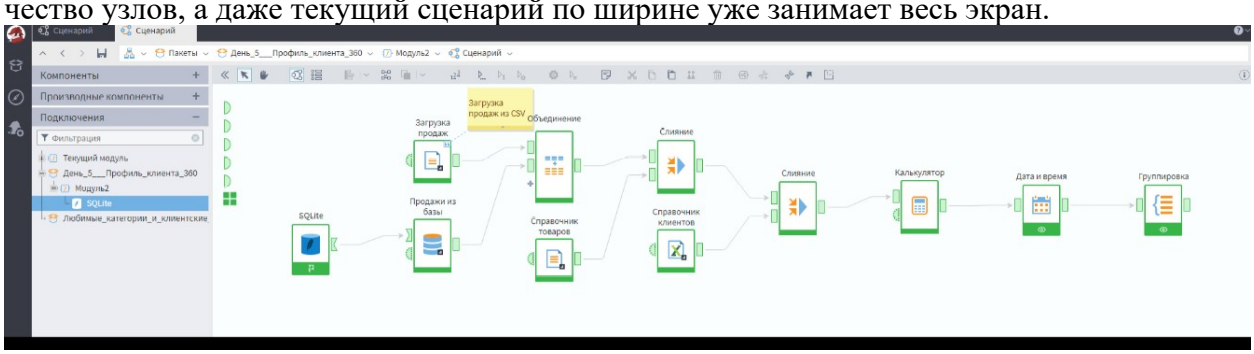
Поэтому наша цель — создать таблицу с полезными поведенческими характеристиками клиентов, выходящими за рамки того, что можно получить при помощи простых отчетов. Очистка данных позволит увидеть реальную картинку, свободную от шума, мусора и случайных событий.

Важно! Если вы не сделали практику в теме «Импорт из баз данных на примере SQLite», то предварительно требуется открыть решение задания предыдущего дня, скачав его ниже.

RARРешение практического задания. День 4.rar

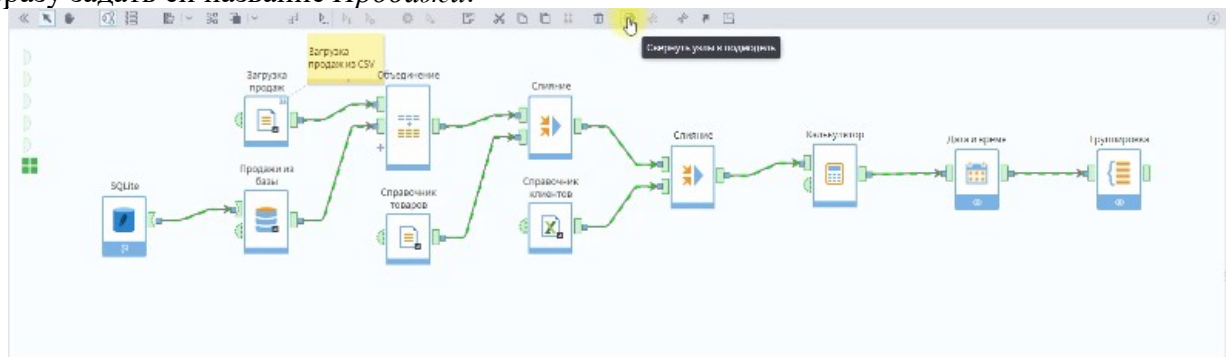
Наращиваем сценарий

Сценарий должен быть более функциональным. Для этого нужно добавить приличное количество узлов, а даже текущий сценарий по ширине уже занимает весь экран.

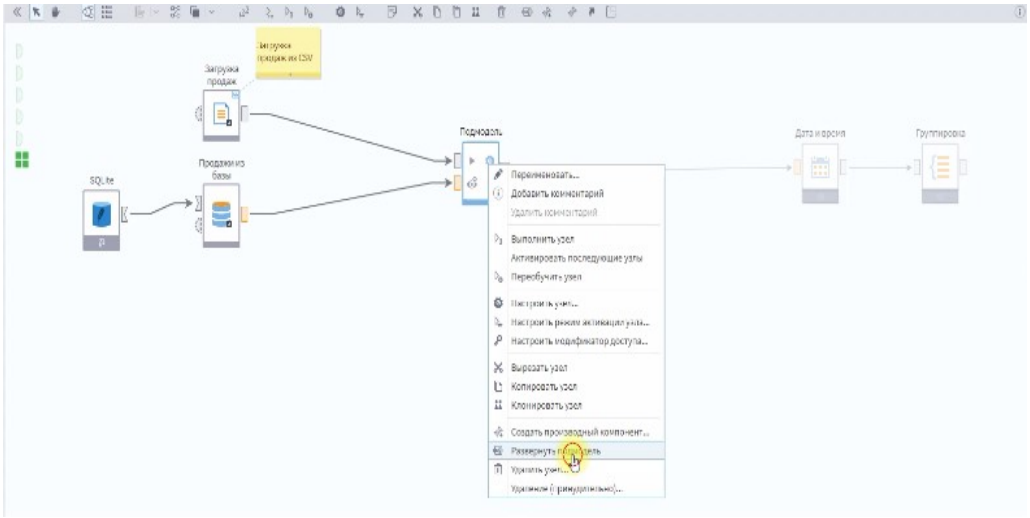


Можно продолжать достраивать его новыми узлами справа, технических ограничений нет. Но нужно принимать во внимание, что когда сценарий занимает больше одного экрана в ширину, работать с ним неудобно.

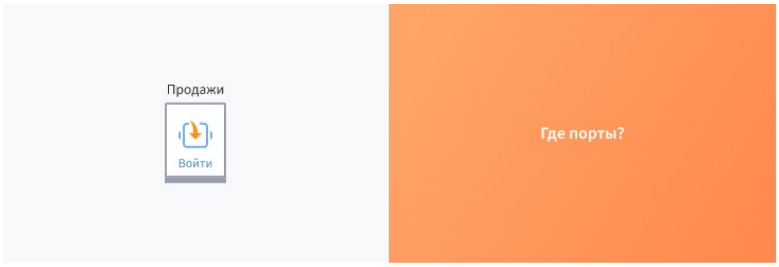
Поэтому начнем практику с оптимизации рабочего пространства. В предыдущие дни рассказывалось, что в Loginom существуют узлы, содержащие в себе другие узлы — это *подмодели*. Для большего удобства будет лучше, если мы свернем имеющийся сценарий в такую подмодель. Можно сразу задать ей название *Продажи*.



Если вы промахнулись при выделении всех узлов, и в подмодель свернулась только часть сценария, то вам нужно развернуть полученную подмодель и повторить процесс уже со всеми узлами.

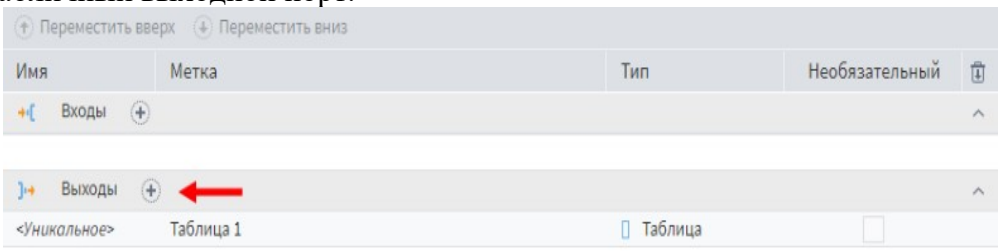


В отношении подмоделей действуют те же правила ввода-вывода данных, что и у обычных узлов. Хотите ввести данные в подмодель – нужны входные порты. Хотите вывести — нужны выходные порты. А в результате свертывания получилась подмодель без портов, словно чемодан без ручки.



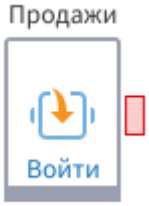
Давайте разберемся. Входные порты в этой подмодели не нужны — данные зарождаются у нее внутри через узлы импорта. А вот выход нужен, т.к. требуется производить дальнейшую обработку этих данных. Чтобы решить проблему, зайдем в настройки узла подмодели.

Здесь можно указать перечень входных и выходных портов, а также их типы. Сейчас нужен один табличный выходной порт.



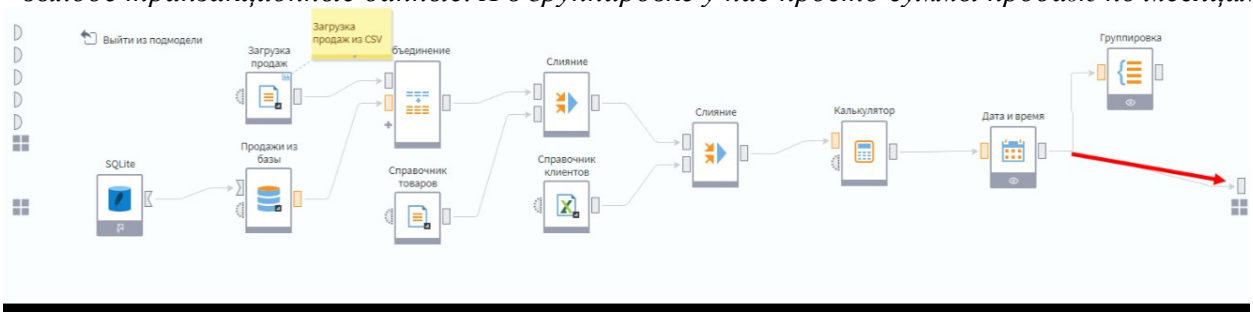
При создании нескольких портов на вход/выход обязательно давайте им понятные названия. Это упростит работу со сценарием и сэкономит в будущем много времени. Имя показывается во всплывающей подсказке при наведении мышки на порт. Правильное название позволит сразу понять назначение данных без погружения в подмодель.

Порт появился, но он красный. Причина в том, что не определено, какие данные на него выводятся.

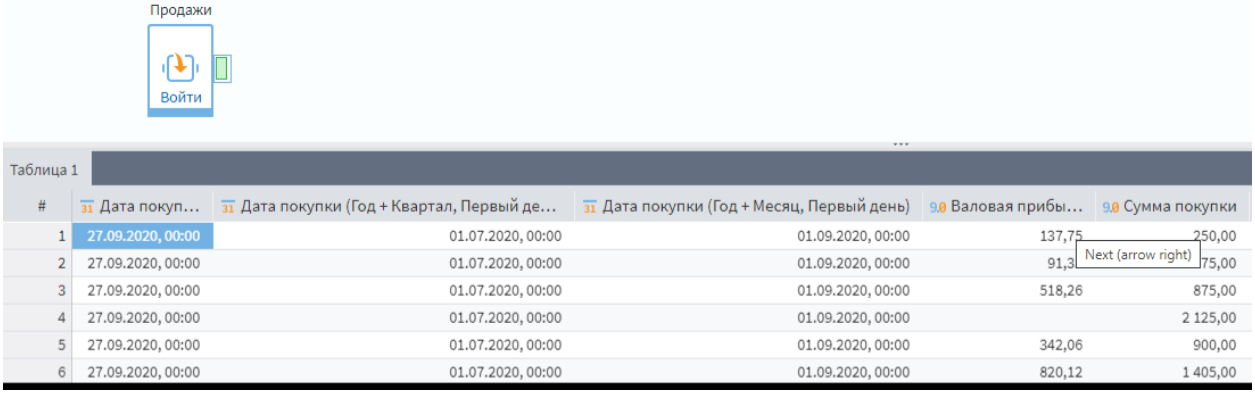


Зайдите внутрь подмодели и протяните связь между портом узла *Дата и время* и выходным портом подмодели.

*Если вы сделали задание со звездочкой из дня 3, обратите внимание, что на выход надо подавать данные не с последнего узла **Группировка**, а из узла **Дата и время**, потому что нам нужны на выходе транзакционные данные. А в группировке у нас просто суммы продаж по месяцам.*



Теперь при активации узла подмодели спустя некоторое время на выходном порту будут данные с объединенной историей продаж.



#	Дата покуп…	Дата покупки (Год + Квартал, Первый де…	Дата покупки (Год + Месяц, Первый день)	Валовая прибы…	Сумма покупки
1	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00	137,75	250,00
2	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00	91,3	75,00
3	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00	518,26	875,00
4	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00		2 125,00
5	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00	342,06	900,00
6	27.09.2020, 00:00	01.07.2020, 00:00	01.09.2020, 00:00	820,12	1 405,00

Next (arrow right)

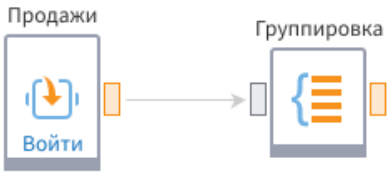
Формирование обогащенного справочника клиентов

Как сказано выше, задача — получить справочник клиентов с расширенными аналитическими признаками, отсутствующими в сводном отчета на исходных данных.

Благодаря ему можно определить проблемы и точки роста в клиентской базе. В нем будут количественные и качественные атрибуты. Часть из них будет считаться на всех доступных данных, а другая — на данных за определенный период.

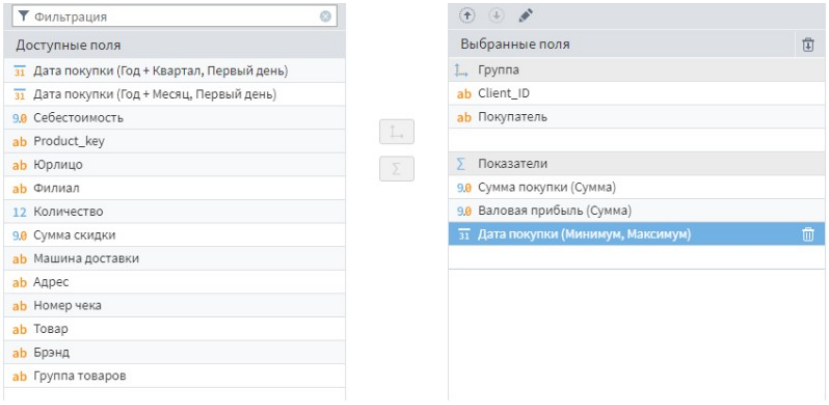
Давайте же начнем!

В качестве основы этого супер-справочника понадобится список всех клиентов и их идентификаторов. Заодно можно добавить в него несколько аналитических признаков на основе всех доступных данных. Удобнее всего сделать это с помощью узла *Группировка* из набора узлов *Трансформация*. Перенесем этот компонент в сценарий после подмодели.



Зайдем в настройки группировки. Этот узел позволяет вернуть агрегированные значения полей показателей в разрезе группируемых полей. Добавьте в группируемые поля *Client_ID* как уникальный идентификатор и ключевое поле, если нам понадобится что-то связать с этой таблицей. Также добавьте в группы поле *Покупатель*, чтобы у нас было под рукой человекопонятное название клиента.

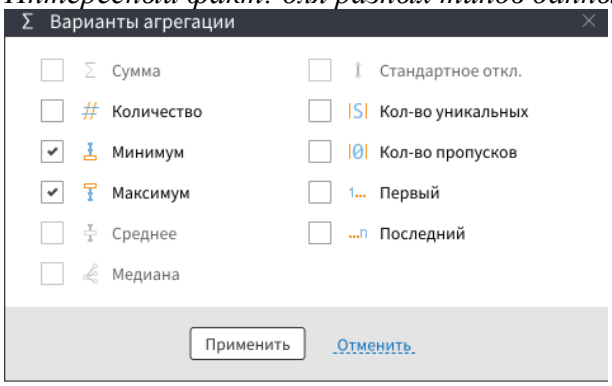
В показатели добавьте *Сумма покупки, Валовая прибыль, Дата покупки*.



Для *Суммы покупки* и *Валовой прибыли* Loginom автоматически предложит тип агрегации *Сумма*, и это нас устраивает. А вот *Дату покупки* надо поднастроить. Дважды кликните по показателю *Дата покупки*.

Откроется окно настроек, в котором надо задать способ агрегации по полю. Можно выбрать несколько вариантов. Нас интересует *Минимум* (дата первой покупки) и *Максимум* (дата последней покупки).

Интересный факт: для разных типов данных доступны разные способы агрегации.



Назовите этот узел *LTV (Life time value)* и активируйте его. На выходе должна быть таблица **3184 строки**. Можно проверить это, дважды кликнув по выходному порту *Группировки*.

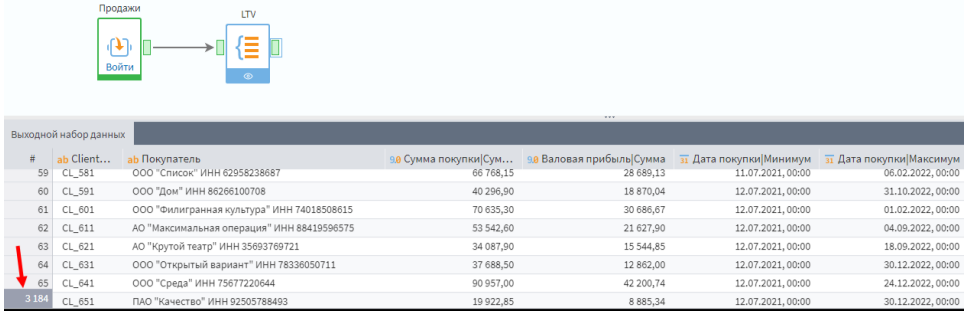
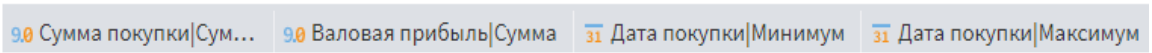


Рис.5 просмотр результатов группировки

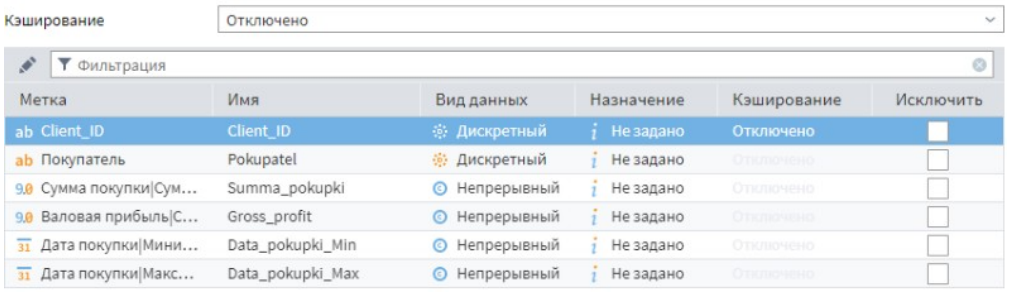
Переименование полей, исключение полей из сценария

Выглядит неплохо, но у полей не очень красивые метки. Давайте это исправим.



В Loginom существует множество способов изменить названия и метки полей в ходе сценария. Самый простой из них — использовать обработчик *Параметры полей* из набора *Трансформация*.

Добавьте его в сценарий и заведите с него выход из LTV-группировки. Зайдите в настройки узла. Откроется список для настройки полей.



Дважды кликните по полю *Суммы покупки*. В появившемся окне настройки можно изменить метку, имя и тип данных поля. Задайте полям следующие метки:

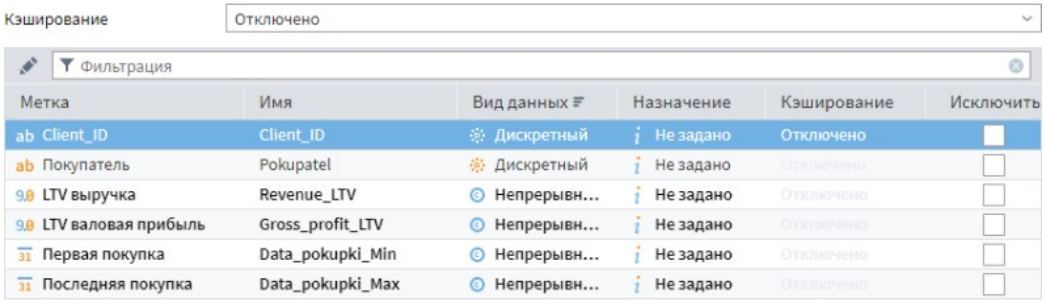
Сумма покупки|Сумма — имя *Revenue_LTV*, метка *LTV выручка*;

Валовая прибыль|Сумма — имя *Gross_profit_LTV*, метка *LTV валовая прибыль*;

Дата покупки|Минимум — имя *Data_pokupki_Min*, метка *Первая покупка*;

Дата покупки|Максимум — имя *Data_pokupki_Max*, метка *Последняя покупка* (Крайняя покупка, если в душе вы пилот).

Итог будет выглядеть так:

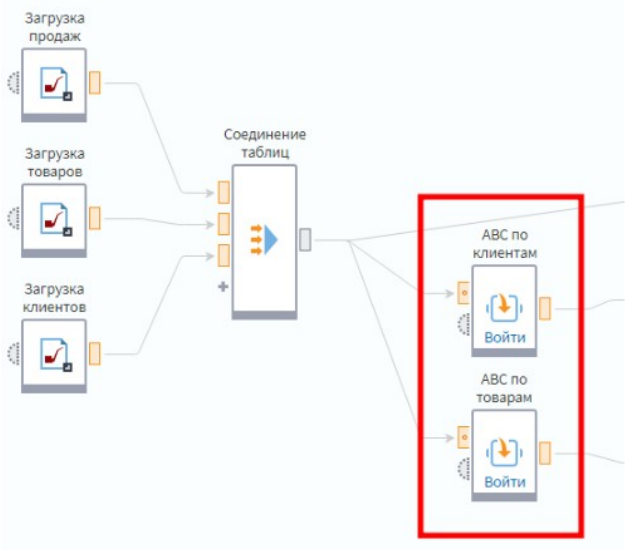


Галочкой ***Исключить*** справа можно выбрать поля, которые не будут передаваться на выход узла. Сейчас это не нужно, но в будущем может потребоваться.

Производные компоненты и внешние библиотеки

Мы подготовили набор с самыми базовыми характеристиками клиентов. Однако в реальности такие таблицы содержат десятки полей. Чтобы немного срезать путь и заодно попробовать очень важный функционал Loginom, обратимся к механике производных компонентов.

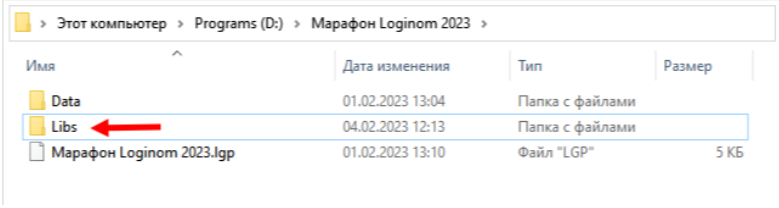
Помните подмодели для ABC-анализа во втором дне марафона? Там эти узлы были частью сценария, и использовалось копирование, чтобы создать 2 ветки ABC-анализа: по товарам и по клиентам. Т.е. каждая подмодель была отдельной сущностью, их можно было независимо редактировать, а применение в других местах было возможно только через механику «копировать-вставить».



Механика производных компонентов позволяет использовать в текущем сценарии ссылки на подмодели, которые находятся в другом модуле или даже в другом пакете. Для этого нужно добавить ссылку на пакет, содержащий подмодели, которые мы хотим подключить.

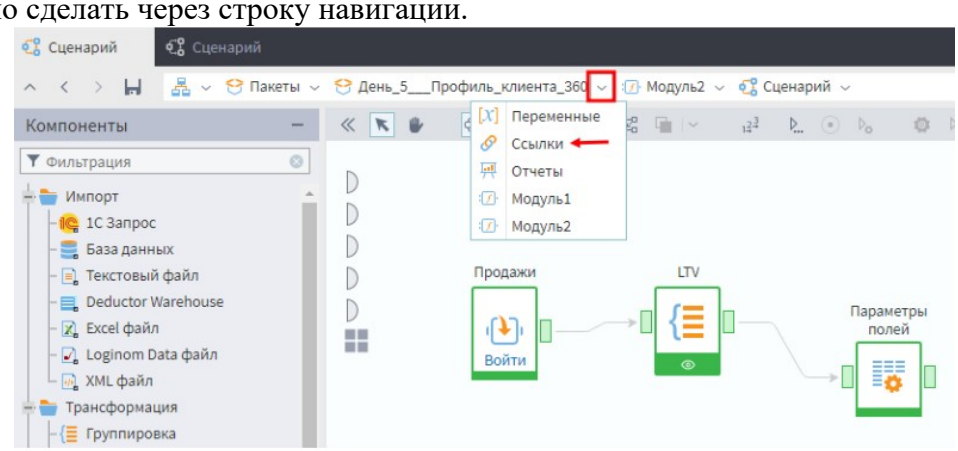
Создайте в рабочей папке марафона папку **Libs**. Скопируйте в нее пакет ниже.

LGPЛюбимые категории и клиентские показатели.lgp

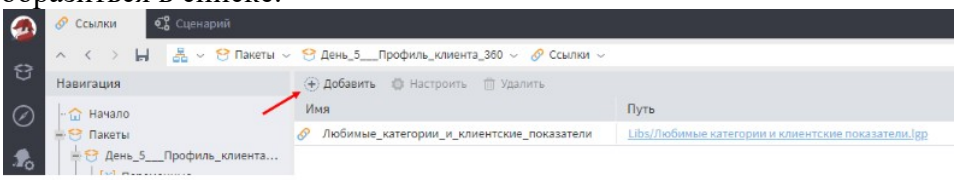


Заводить отдельную папку для пакетов-библиотек компонентов — это хорошая практика.

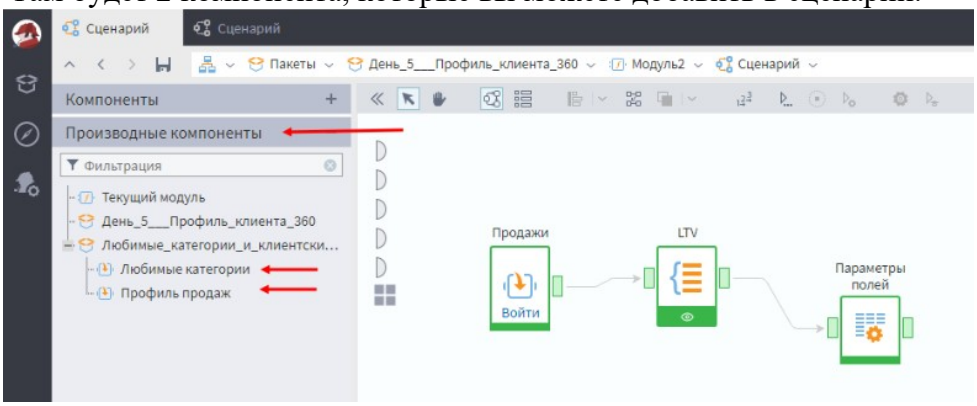
Давайте подключим этот пакет. Для этого зайдите в раздел *Ссылки* рабочего пакета. Это удобно сделать через строку навигации.



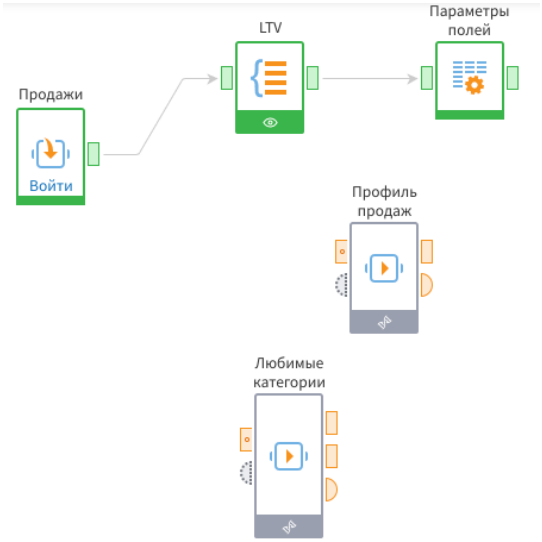
Нажмите на кнопку *Добавить*, укажите путь к пакету в папке *Libs*. После добавления он должен отобразиться в списке.



Вернитесь обратно в сценарий. В левой секции с библиотекой компонентов выберите раздел ***Производные компоненты***. Разверните в нем раздел *Любимые категории и клиентские характеристики*. Там будет 2 компонента, которые вы можете добавить в сценарий.



Добавьте оба узла в сценарий.



В целом узлы производных компонентов работают аналогично обычным узлам. У них есть входы и выходы. Логика действий узлов определяется их внутренним сценарием. Настройки узлов обычно происходят через порт переменных.

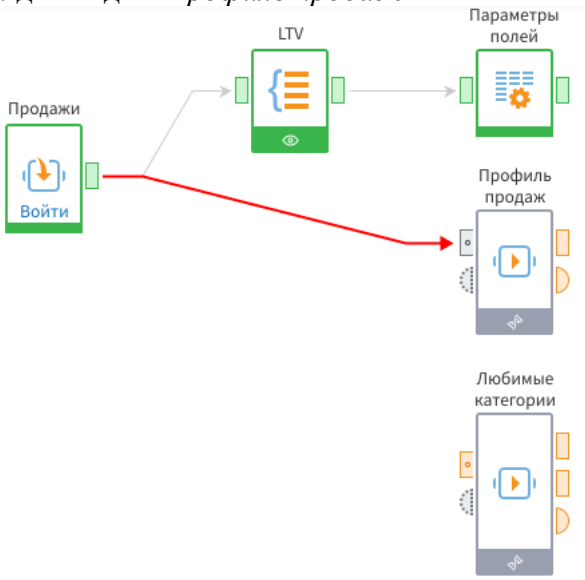
Чаще всего узлы производных компонентов имеют входной порт с отключенной автосинхронизацией (точка на порту). Это значит, что в сценарий внутри узла поступают не любые поля, а только заранее определенные, вокруг которых построена внутренняя логика узла. Согласитесь, это не совсем логично, если в узлы с предопределенной логикой работы подаются любые поля без разбора.

Профиль продаж

Узел профиль продаж рассчитывает ряд характеристик по клиенту в разрезе клиентов за определенный период:

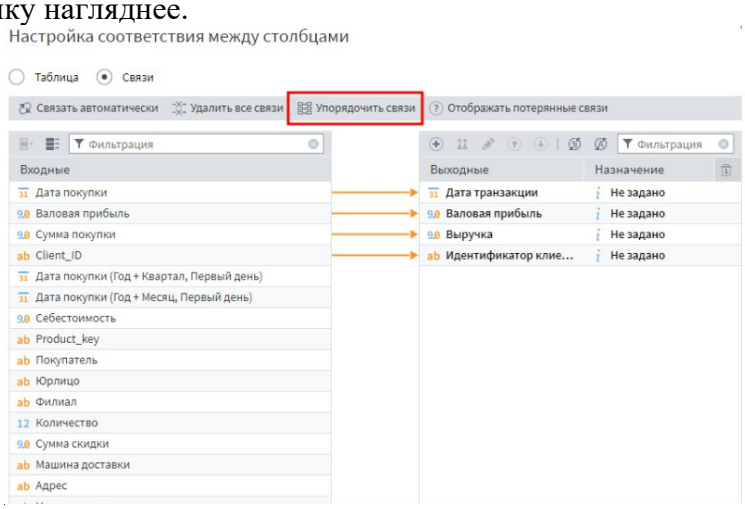
- Периодичность покупок;
- Средний чек;
- Количество покупок;
- Выручка;
- Валовая прибыль;
- % валовой прибыли.

Вначале нужно подать данные в подмодель. Протяните связь между выходом из подмодели *Продажи* до входа в *Профиль продаж*.

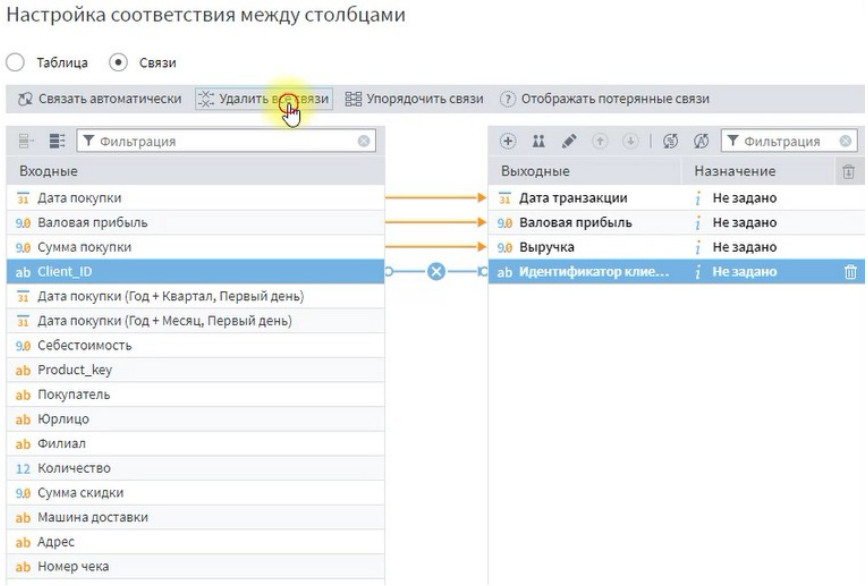


Дважды кликните ЛКМ по входному порту данных у *Профиля продаж*. Будут отображены 4 поля. Это те поля, которые предопределены для использования в подмодели.

Для большей ясности включите режим отображения *Связи* в переключателе над таблицей. Отобразится 2 списка: поля таблицы, которую вы подали на вход в левой части, и поля, которые ожидает на вход подмодель в правой части. Нажмите кнопку ***Упорядочить связи***, чтобы сделать картинку нагляднее.



Loginom старается определить соответствие полей, и если оно выглядит так же, как на картинке выше, — значит, все верно. Иначе нужно удалить неправильные связи и задать правильные. Как это сделать, показано на видео.



Активируйте узел после настройки связей. На выходе будет таблица с показателями, рассчитанными в разрезе идентификаторов клиентов.

Дважды кликните по входному порту переменных (полукруглый порт) подмодели *Профиль продаж*. Здесь задаются настройки производного компонента.

В этом порту можно настроить 2 переменные:

Первая — дата сегментации, которая определяет, на какую дату надо рассчитать параметры. Подмодель сделана так, что если дата не задана, то берется самая последняя дата из таблицы.

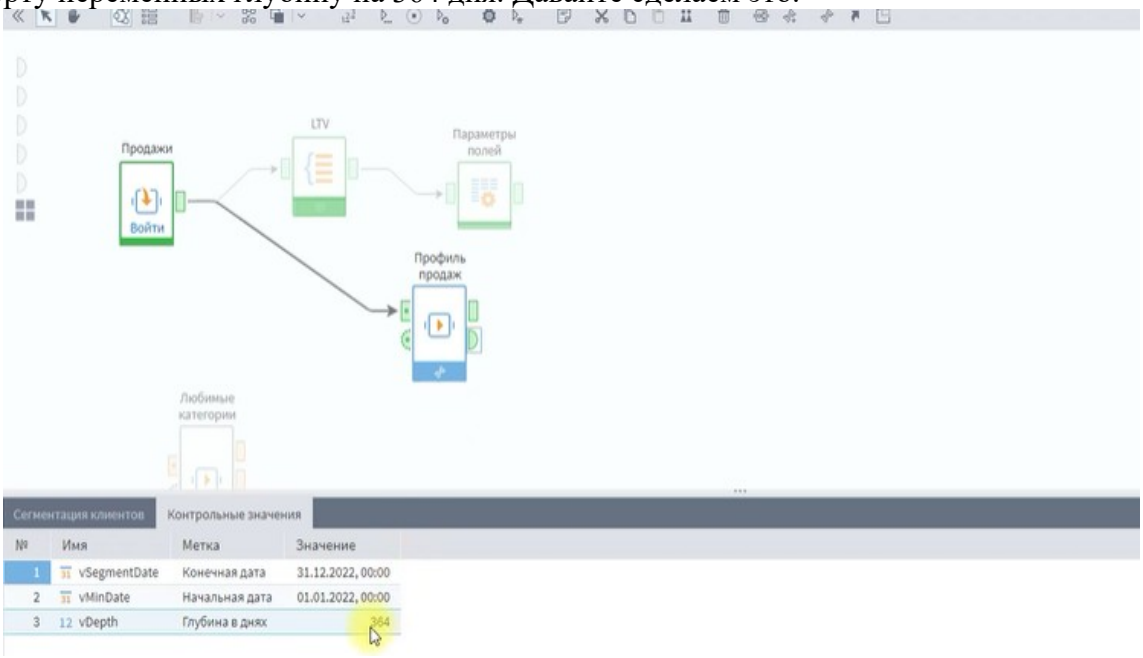
Глубина в днях — за сколько дней от даты сегментации возьмутся данные. Таким образом, можно рассчитывать параметры не по всем данным, а только по интересующему периоду активности. Например, аналитика может интересовать поведение клиентов только за последний год, т.к. предыдущая история не актуальна.

Посмотрите на выходной порт переменных в подмодели *Профиль продаж* (при необходимости, активируйте ее повторно).

В такие порты могут выводиться как переменные для использования в других узлах, так и справочная информация для понимания, что происходило внутри подмодели.

Здесь видно, что **дата сегментации** определилась как 31.12.2022 — это конечная дата в нашем наборе данных. А **начальная дата** (vMinDate) определяется как 31.12.2021. Эта переменная создана внутри подмодели, как разница конечной даты и глубины сегментации.

Если хочется, чтобы начальная дата была 01.01.2022, то это можно сделать, изменив на входном порту переменных глубину на 364 дня. Давайте сделаем это.



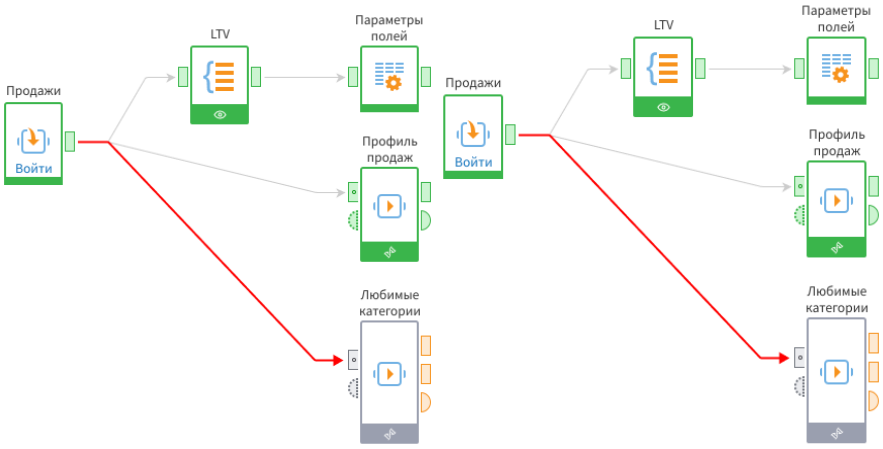
Имена полей, выходящих из подмодели, имеют метку ***_pp***

Теперь, настроим вторую подмодель. **Но возникает вопрос: переменные всегда надо задавать вручную?**

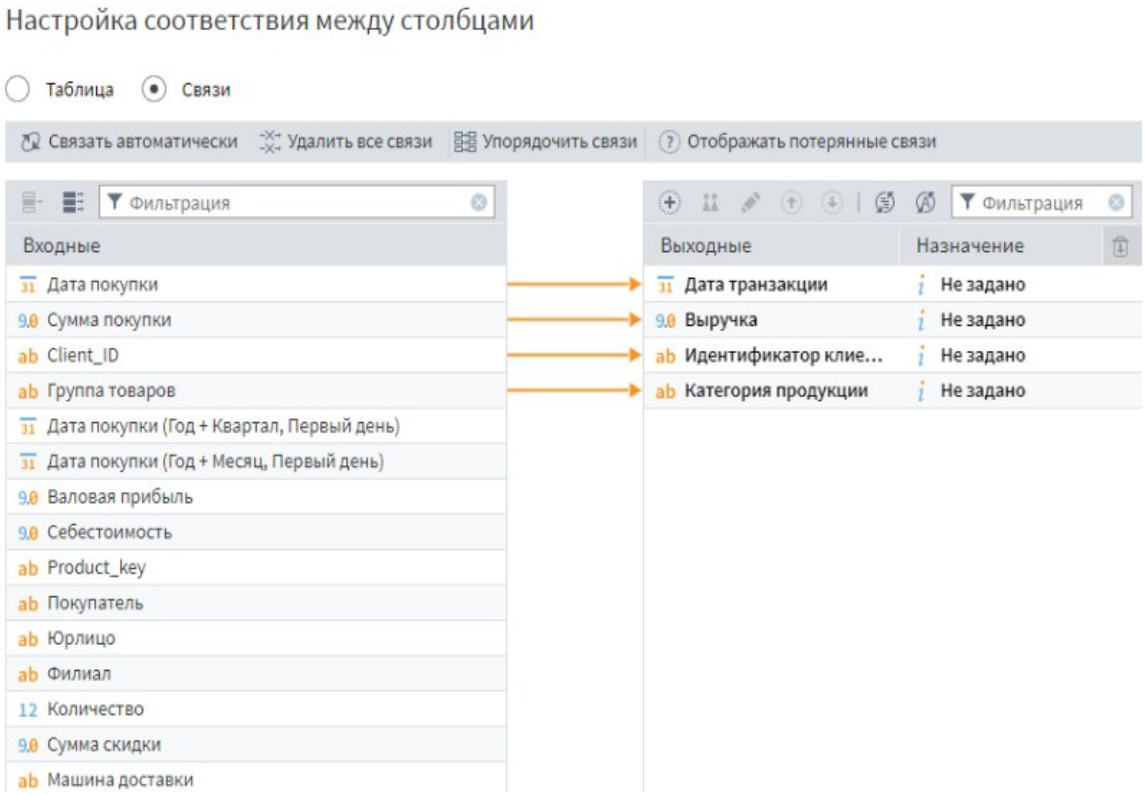
Ответ: нет, можно реализовать динамическое определение переменных, например, через узел *Калькулятор переменных*. В нем задаются значения переменных с помощью любых функций, которые можно передать на вход подмодели.

Любимые категории

Эта подмодель вернет список любимых категорий в разрезе по клиентам. Подключите связь от подмодели *Продажи* к входному порту узла *Любимые категории*.



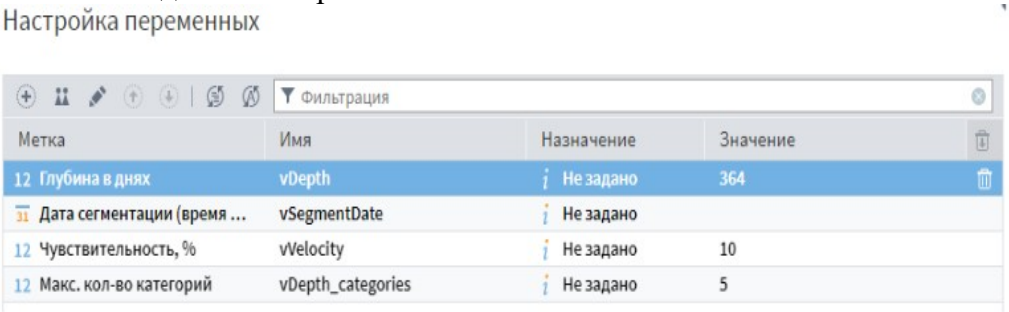
По двойному клику ЛКМ по входному порту данных на узле *Любимые категории* будет предложена настройка порта. Надо убедиться, что поля сопоставлены как на картинке.



Кстати, если в поле *Категория продукции* подать, например, поле *Бренд* — то получится определение любимых брендов по клиентам. Так это сработает и с товаром. Но пока остановимся на категориях.

Зайдем в настройки входного порта переменных. Первые 2 нам уже знакомы — это дата построения сегментации, которую можно оставить пустой, и глубина сегментации в днях, которую нужно выставить как 364.

А вот последние 2 настройки — новенькие.



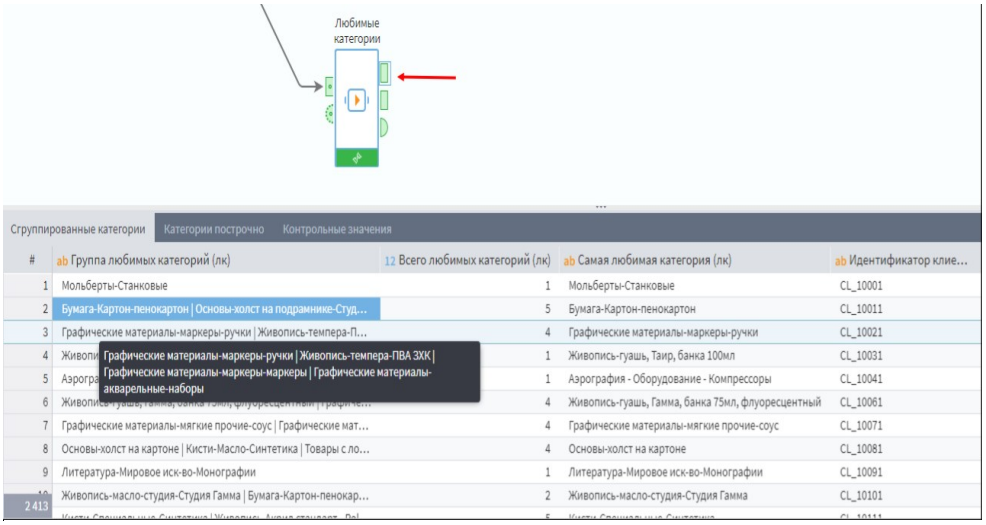
Переменная ***Чувствительность, %*** определяет, начиная с какой доли оборота клиента за период категория будет считаться любимой. Т.е. в любимые попадут только те категории, доля оборота по которым не меньше указанного числа.

Макс. кол-во категорий — сколько любимых категорий может быть. Категории выбираются в порядке убывания оборота. В целом эти настройки устраивают, но можно с ними поэкспериментировать.

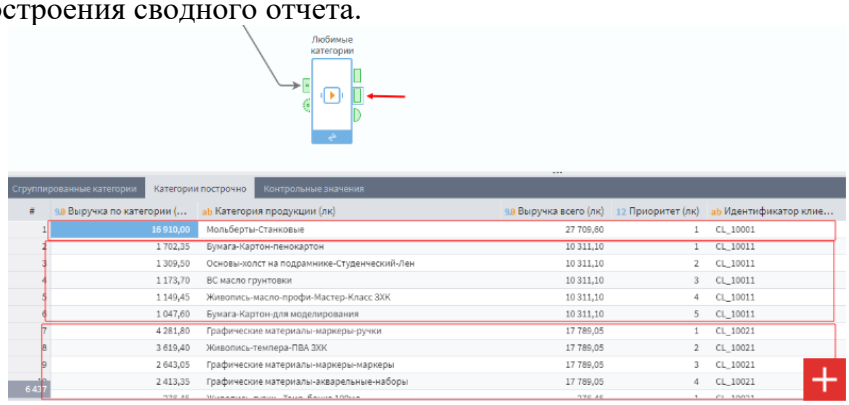
Активируйте узел. Обратите внимание, что из подмодели *Любимые категории* есть 2 табличных выхода. Это пример того, что одна подмодель может возвращать несколько разных структур данных.

Так, из первого порта данные выходят в свернутом по клиентам виде: одна строка соответствует одному клиенту, а любимые категории свернуты в одну строку через разделитель. При этом категории расположены по убыванию «любимости».

Такой формат удобен для использования в обогащении справочника, т.к. на каждого клиента приходится одна строка данных, и при объединении с основным справочником строки не задвоятся.



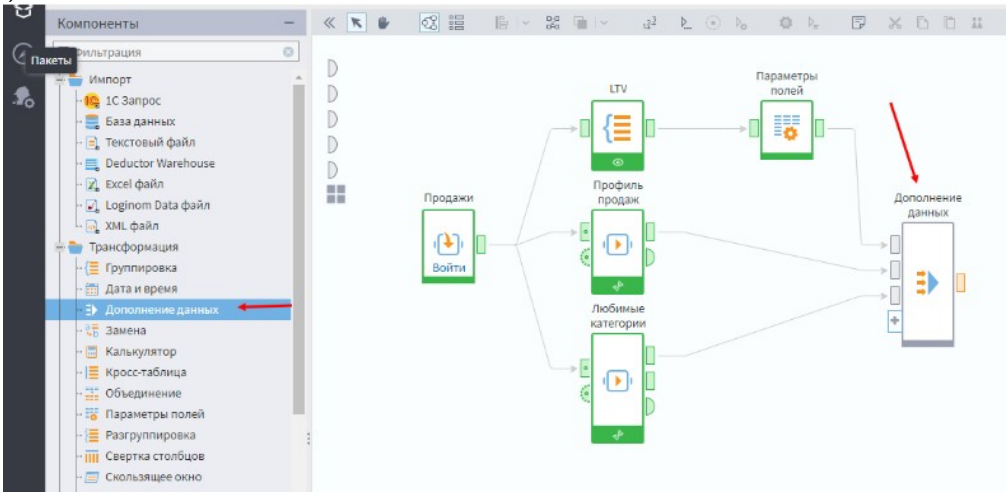
Из второго порта данных выходит массив, развернутый по строкам. Такой формат может быть удобен для передачи системам, которым требуется структурированный формат данных. Или для построения сводного отчета.



Имена полей, выходящих из подмодели, имеют метку ***_fc***.

Единый клиентский справочник

Формирование клиентских характеристик происходит в 3-х ветках сценария. Надо соединить их в единый справочник. Ранее аналогичная задача решалась за счет использования нескольких узлов *Слияние*. В этот раз будет использоваться компонент *Дополнение данных* из группы *Трансформация*.



Он позволяет соединить сразу несколько таблиц, но по сравнению с узлом *Слияние*, у него есть ряд ограничений:

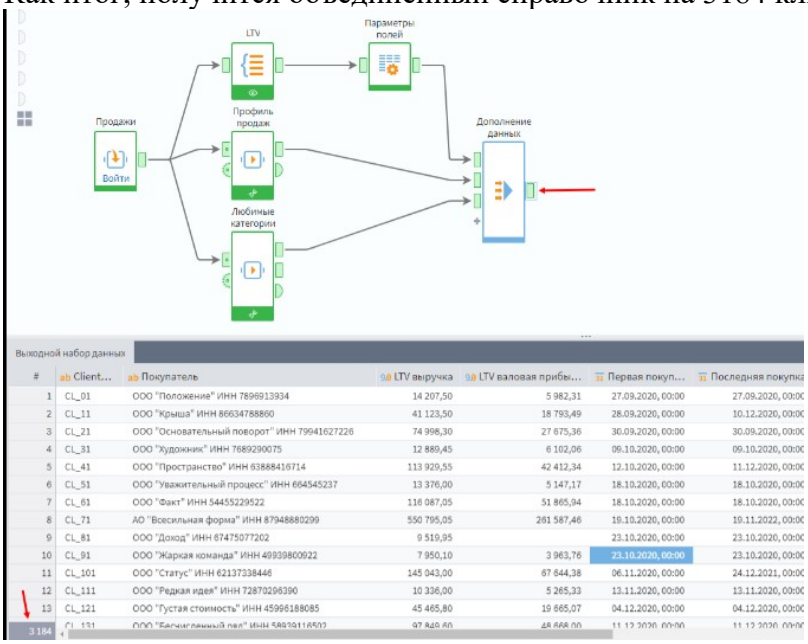
Соединение работает только по сценарию Left join.

Последовательное соединение не работает. Т.е. нельзя присоединить к первой таблице вторую по ключу между ними, а ко второй присоединить третью по ключу во второй таблице.

Этот узел нужно использовать, когда требуется сделать Left join’ы к одной таблице нескольких других таблиц. В настройках узла выставим присоединение к ***Client_ID*** полей Идентификатор клиента из 2-х других таблиц.

№	Главная таблица	Присоединяемая таблица	Присоединяемая таблица 2
1	ab Client_ID	ab Идентификатор клиента	ab Идентификатор клиента
2	ab Покупатель	Не выбрано	Не выбрано
3	9.0 LTV выручка	Не выбрано	Не выбрано
4	9.0 LTV валовая при...	Не выбрано	Не выбрано
5	31 Первая покупка	Не выбрано	Не выбрано
6	31 Последняя поку...	Не выбрано	Не выбрано

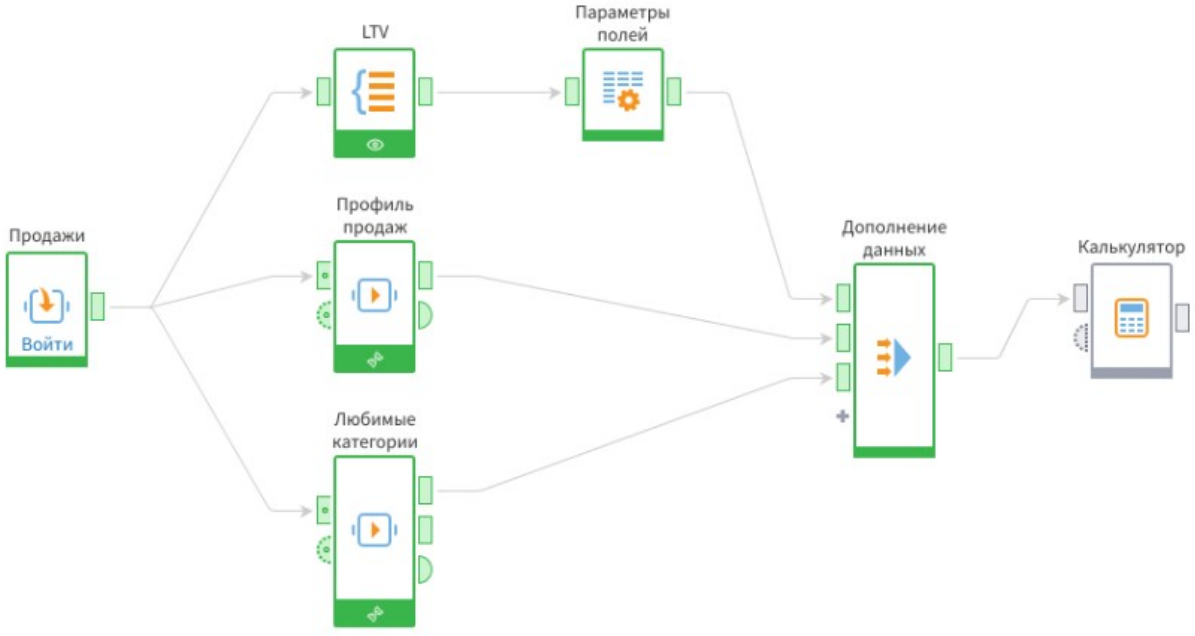
Как итог, получится объединенный справочник на 3184 клиента.



Создание дополнительных признаков

У нас появился обогащенный клиентский справочник с большим количеством параметров. Однако анализировать голые числа неудобно. Есть смысл создать несколько дополнительных аналитических признаков, которые можно использовать как разрезы для отчетов и фильтры для быстрых отборов разных зон интереса в клиентской базе.

Для этого добавим в сценарий *Калькулятор* после блока *Дополнение данных*.



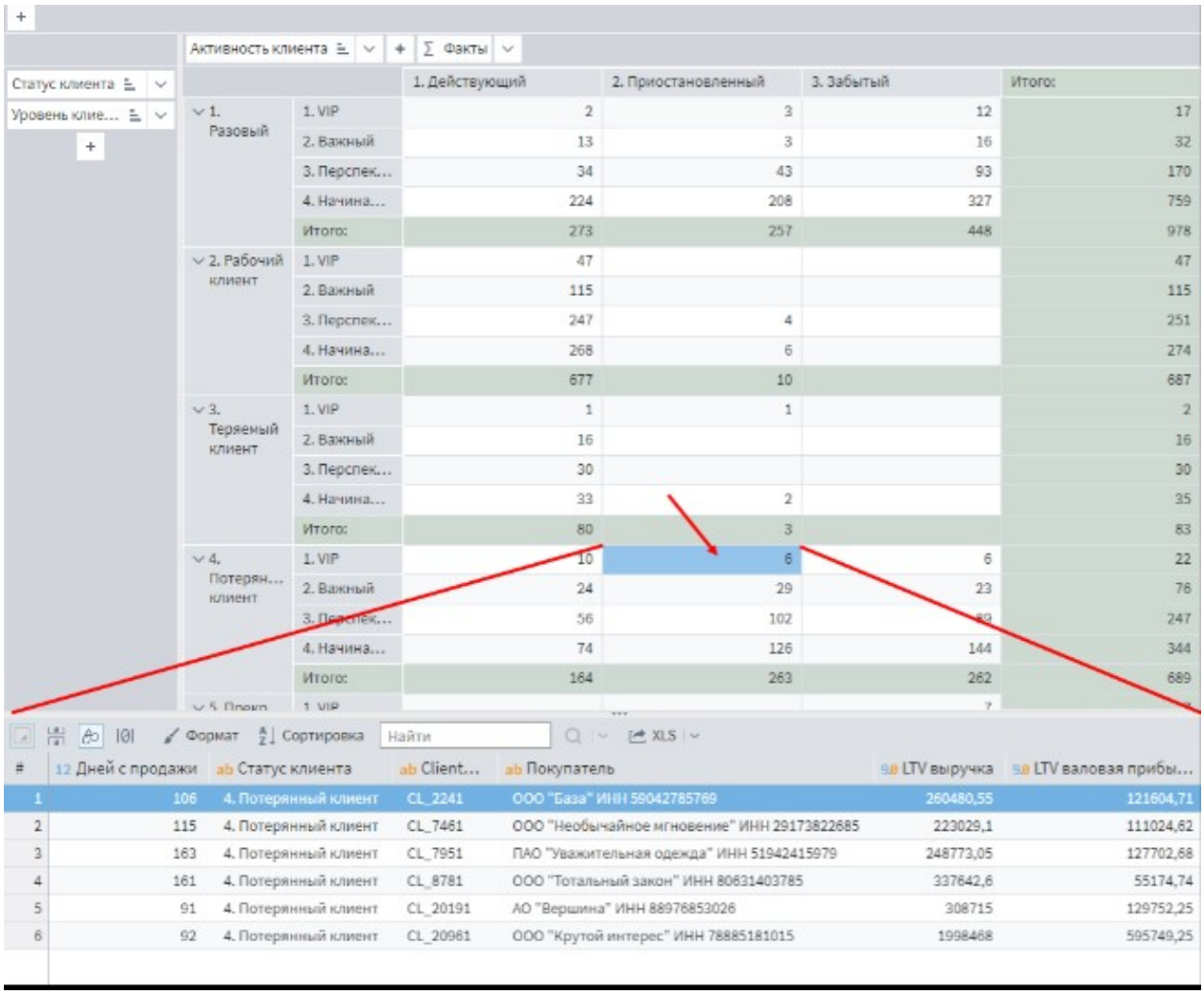
Дней с последних продаж

Для начала в настройках Калькулятора, создадим поле ***Дней с продажи***. Дадим ему имя *Days_from_sale*. Т.к. в наборе есть поле *Data_pokupki_Max*, то формула ***today()-Data_pokupki_Max*** вернет количество дней, которые прошли с последней продажи до сегодняшнего дня.

Но т.к. у нас статичный учебный датасет, сделаем дату условного сегодняшнего дня тоже статичной. Чтобы дальнейшие изменения в данных происходили от наших действий, а не потому что часики тикают.

Для этого можно использовать функцию ***EncodeDate***. Она трансформирует значение года, месяца и числа в дату. Очень полезная формула, когда надо сделать распознавание данных в дату, или задать ее статичное значение.

В результате формула дней с последней продажи будет выглядеть следующим образом: ***EncodeDate(2022,12,31)-Data_pokupki_Max***.



Активность клиента: Факты

Статус клиента	Уровень клие...	1. Действующий	2. Приостановленный	3. Забытый	Итого:
1. Разовый	1. VIP	2	3	12	17
	2. Важный	13	3	16	32
	3. Перспек...	34	43	93	170
	4. Начина...	224	208	327	759
	Итого:	273	257	448	978
2. Рабочий клиент	1. VIP	47			47
	2. Важный	115			115
	3. Перспек...	247	4		251
	4. Начина...	268	6		274
	Итого:	677	10		687
3. Теряемый клиент	1. VIP	1	1		2
	2. Важный	16			16
	3. Перспек...	30			30
	4. Начина...	33	2		35
	Итого:	80	3		83
4. Потерян... клиент	1. VIP	10	6	6	22
	2. Важный	24	29	23	76
	3. Перспек...	56	102	89	247
	4. Начина...	74	126	144	344
	Итого:	164	263	262	689

Формат | Сортировка | Найти | XLS

#	Дней с продажи	Статус клиента	Client...	Покупатель	LTV выручка	LTV валовая прибы...
1	106	4. Потерянный клиент	CL_2241	ООО "База" ИНН 59042785769	260480,55	121604,71
2	115	4. Потерянный клиент	CL_7461	ООО "Необычайное мгновение" ИНН 29173822685	223029,1	111024,62
3	163	4. Потерянный клиент	CL_7951	ПАО "Уважительная одежда" ИНН 51942415979	248773,05	127702,68
4	161	4. Потерянный клиент	CL_8781	ООО "Тотальный закон" ИНН 80631403785	337642,6	55174,74
5	91	4. Потерянный клиент	CL_20191	АО "Вершина" ИНН 88976853026	308715	129752,25
6	92	4. Потерянный клиент	CL_20961	ООО "Крутой интерес" ИНН 78885181015	1998468	595749,25

Клиентский статус

Теперь внедрим параметр, который будет определять качество клиентского поведения. Нужно понимать, находится ли клиент в своем стандартном графике закупок или начинает уходить. Самый простой способ это выяснить — посчитать разницу между средней периодичностью покупок и количеством дней с последней продажи.

Для этого создадим текстовое поле ***Client_status*** (Статус клиента). Способов расчета статусов много, но в этом примере будет использоваться следующий:

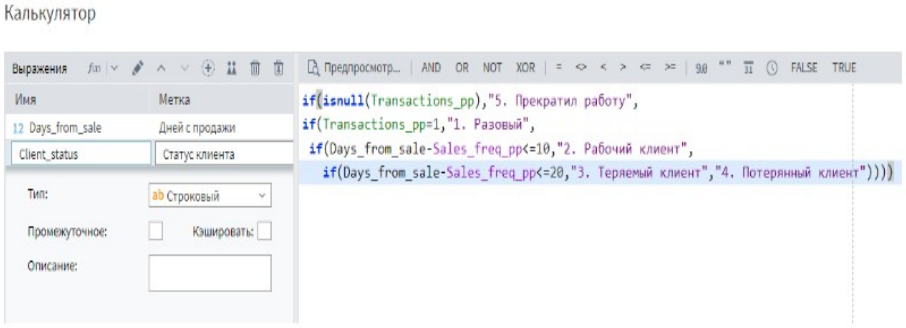
Клиентов, у которых за отчетный период нет продаж, отмечать как прекративших работу.

Клиентов, у которых за отчетный период только одна покупка (поле ***Transactions_pp***), считать разовыми.

Если разница между средней периодичностью и давностью последней покупки условно небольшая, то считать таких клиентов рабочими.

Если разница между периодичностью и давностью последней покупки увеличивается, считать таких клиентами теряемыми. Этот сегмент будет требовать срочного внимания менеджеров.

Всех, у кого давность последней продажи превысила периодичность больше заданной границы, считать потерянными клиентами.



Формула будет иметь следующий вид:

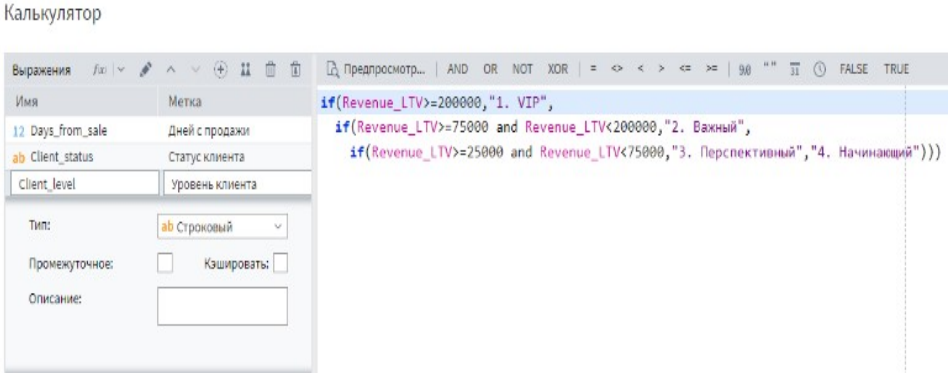
if(isnull(Transactions_pp),"5. Прекратил работу", if(Transactions_pp=1,"1. Разовый", if(Days_from_sale-Sales_freq_pp<=10,"2. Рабочий клиент", if(Days_from_sale-Sales_freq_pp<=20,"3. Теряемый клиент","4. Потерянный клиент"))))

Обратите внимание, что в формуле есть ссылка на ранее вычисленное поле *Days_from_sale*. Это сделано, чтобы упростить синтаксис выражения.

Уровень клиента

Теперь понятно, кто из клиентов работает с нами в обычном режиме, а кто уходит. Но чтобы планировать какие-то действия с базой потребуется некая система приоритетов — в первую очередь стоит работать с самыми важными покупателями.

Очень удобно определить это на основе количества денег, которые клиент принес в компанию. Это можно сделать на основе поля ***Revenue_LTV***. Создадим поле *Client_level* (Уровень клиента).



Формула будет такая:

if(Revenue_LTV>=200000,"1. VIP", if(Revenue_LTV>=75000 and Revenue_LTV<200000,"2. Важный", if(Revenue_LTV>=25000 and Revenue_LTV<75000,"3. Перспективный","4. Начинающий")))

Актуальность клиента

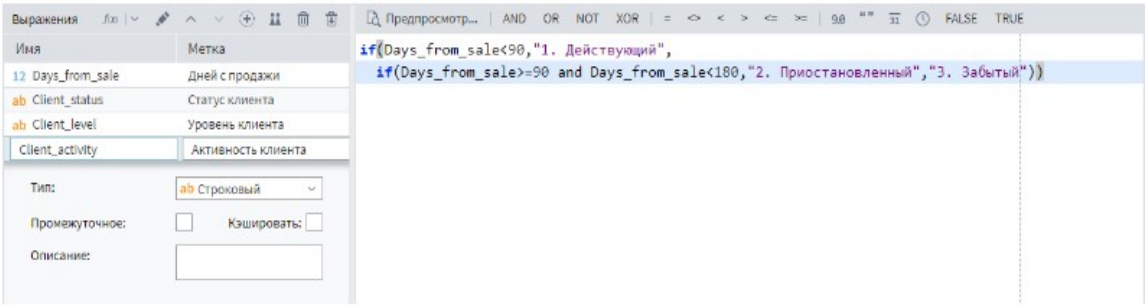
В заключение понадобится еще один параметр, который поможет лучше определять приоритетность взаимодействия с клиентами. Это классификация на основе давности последней продажи.

Согласитесь, что намного перспективнее контактировать с клиентами, которые покупали что-то в предыдущие пару месяцев, чем с теми, кто в последний раз что-то брал более года назад.

Логика будет простая: если покупал до 90 дней, то клиент действующий, если до 180, то приостановленный, а если больше 180 — значит забытый. Формула получается такая:

if(Days_from_sale<90,"1. Действующий", if(Days_from_sale>=90 and Days_from_sale<180,"2. Приостановленный","3. Забытый"))

Калькулятор



Заключение и домашнее задание

Мы проделали большую работу. Теперь есть справочник клиентов, обогащенный статистическими и управленческими атрибутами. С таким массивом можно не просто смотреть на исторические данные, но и принимать решения и действовать на опережение для устранении рисков, например, потерь клиентов.

Чтобы это стало возможным, остался последний штрих — сделать отчет, позволяющий находить точки интереса.

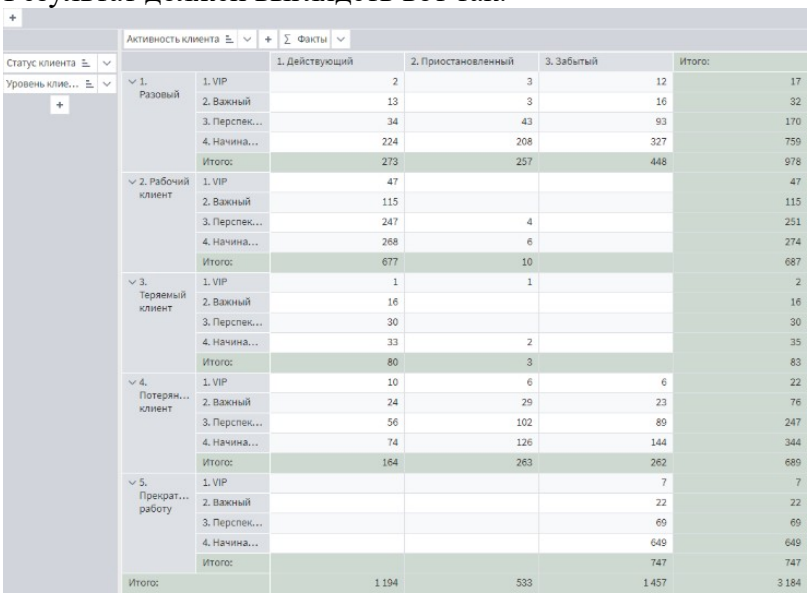
Задание.

Задание №1. Сделать отчет по сегментам клиентской базы.

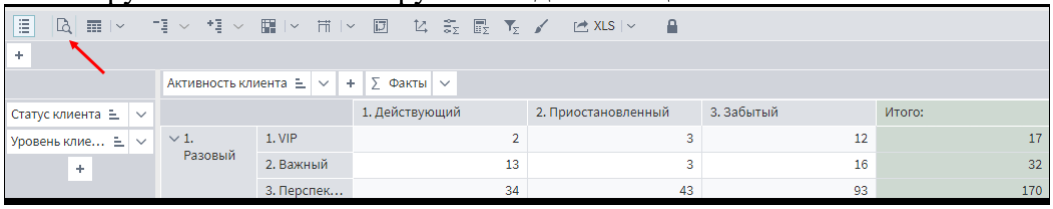
Добавьте визуализатор *Куб* в финальный узел *Калькулятор* сценария и назовите его ***Сегментация клиентов***.

Добавьте в строки разрезы *Статус клиента* еще поле *Уровень клиента*, в столбцы — *Активность клиента*, а в качестве показателей — количество уникальных значений поля ***Client_ID***.

Результат должен выглядеть вот так.



Активируйте на панели инструментов детализацию.



Покликайте по разным ячейкам таблицы и посмотрите, как строки, образующие данное число, выводятся внизу. Как несложно догадаться, так можно легко получить различные списки клиентов с потенциалом на реанимацию или допродажи.

В таблице также содержится информация о товарных предпочтениях клиентов, что поможет завязать диалог или составить предложение заранее. Эта информация может быть выгружена в Excel как в ручном, так и в автоматическом режиме. Либо передана в другие системы.

Задание №2. Оценить финансовую привлекательность.

Создайте второй визуализатор *Куб* в узле *Калькулятор* и назовите его ***Портрет клиента***. В качестве разрезов строк добавьте поле ***Уровень клиента***. В качестве показателей используйте:

Сумма по полю *Выручка (пп)*, и долю выручки по *Уровню клиента*.

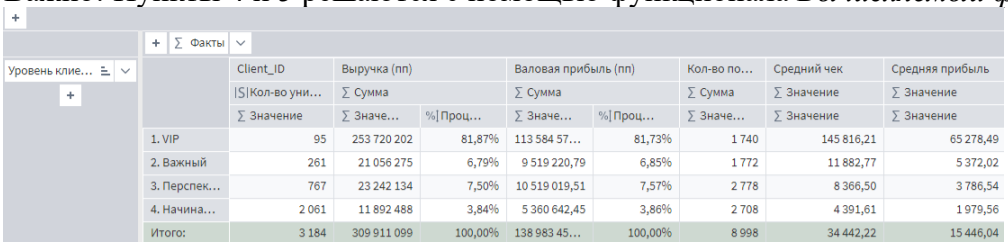
Сумма по полю *Валовая прибыль (пп)*, и долю по *Уровню клиента*.

Количество покупок, как сумму по полю *Кол-во покупок (пп)*.

Вычисляемый факт *Средний чек*, как Выручку, деленную на количество покупок.

Вычисляемый факт *Средняя прибыль*, как Валовая прибыль, деленная на количество покупок.

Важно! Пункты 4 и 5 решаются с помощью функционала *Вычисляемый факт*.



Уровень клие…	Client_ID	Выручка (пп)		Валовая прибыль (пп)		Кол-во по…	Средний чек	Средняя прибыль
	|S|Кол-во уни…	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значение	∑ Значе…	%|Проц…	∑ Значе…	%|Проц…	∑ Значе…	∑ Значение	∑ Значение
1. VIP	95	253 720 202	81,87%	113 584 57…	81,73%	1 740	145 816,21	65 278,49
2. Важный	261	21 056 275	6,79%	9 519 220,79	6,85%	1 772	11 882,77	5 372,02
3. Перспек…	767	23 242 134	7,50%	10 519 019,51	7,57%	2 778	8 366,50	3 786,54
4. Начина…	2 061	11 892 488	3,84%	5 360 642,45	3,86%	2 708	4 391,61	1 979,56
Итого:	3 184	309 911 099	100,00%	138 983 45…	100,00%	8 998	34 442,22	15 446,04

Эту таблицу можно использовать для того, чтобы определять, что клиент представляет из себя в финансовом эквиваленте. Видите ли, вы сейчас что-то подозрительное здесь?

Такую информацию надо использовать для задач финансового моделирования и прогнозирования. Поэтому цель подобных отчетов — не просто показать, что есть данные, а выдать некую картину, которая будет максимально приближена к действительности. Потому что неверные оценки приведут к заниженным/завышенным прогнозам и планам, создадут неадекватные финмодели и пустят компанию по ложному пути.

В отличие от отчетов, просто визуализирующих факты, здесь нужно чистить данные не только технически, но и по смыслу. График продаж показывает в прошлом месяце рекордные прибыли — это факт. Но будет ли такая динамика роста всегда? Или вчера закупился один клиент, который раз в 2 года делает мега-закупку, и компании в своих планах нужно ориентироваться на другой тип клиентов?

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1.Охарактеризуйте метод «Деревьев решений».

2.Охарактеризуйте особенности регрессионного анализа в методах ИАД.

3.Охарактеризуйте модели временных рядов с запаздываниями.

4.Охарактеризуйте метод «Ближайшего соседа».

5.Охарактеризуйте метод поиска правила.

6.Охарактеризуйте метод кластеризации.

7.Охарактеризуйте метод классификации.

8.Охарактеризуйте метод дискриминации.

9.Какие различия в целях и алгоритмах статистического и интеллектуального подходов.

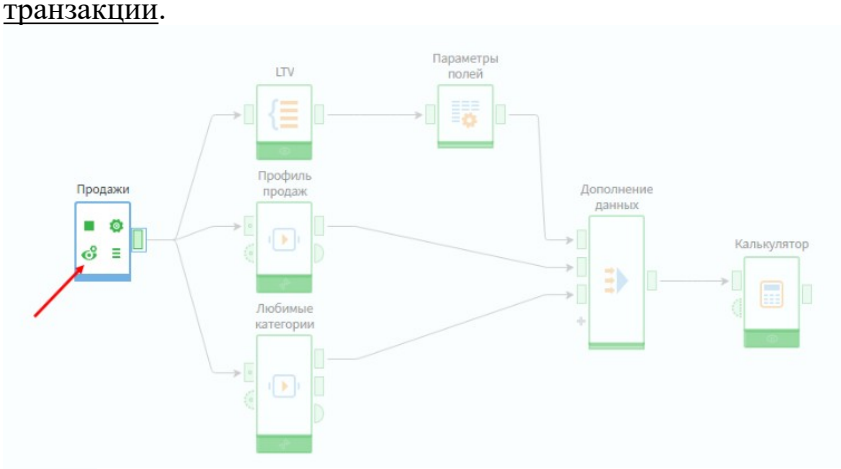
Лабораторная работа № 5.
Тема: Анализ качества данных.

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

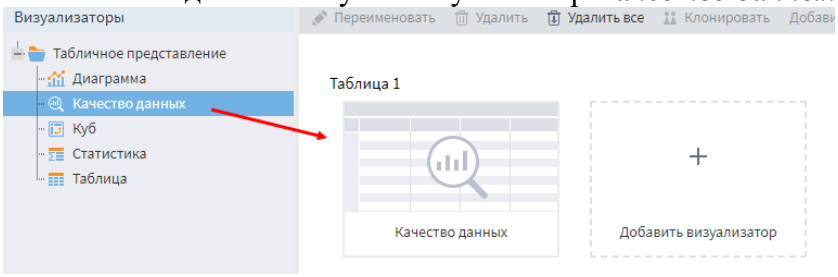
Формируемые компетенции или их части: ОПК-8

Теоретическая часть

В начале необходимо зайти в настройки визуализатора узла *Продажи*, в котором загружены транзакции.



Затем добавить в узел визуализатор *Качество данных*.



После открытия визуализатора будет отображен список полей, качество которых требуется проверить. Т.к. вначале нет предположений в каких именно колонках могут быть проблемы — лучше выбрать все. Выбор оцениваемых показателей задается в соответствующем окне.

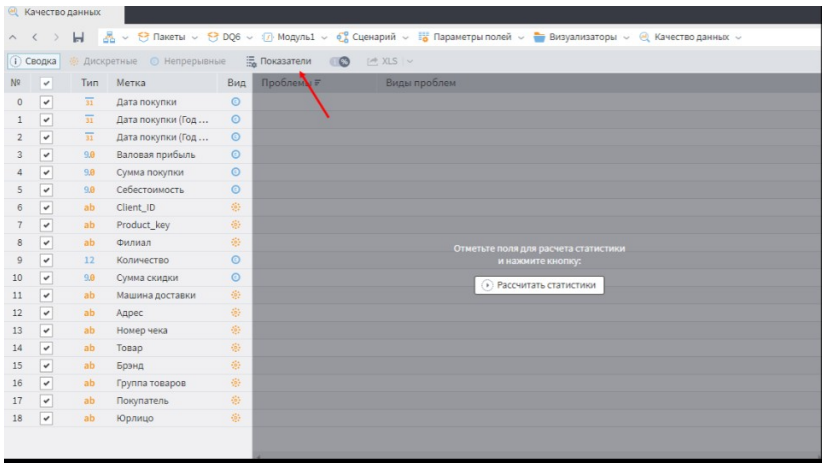
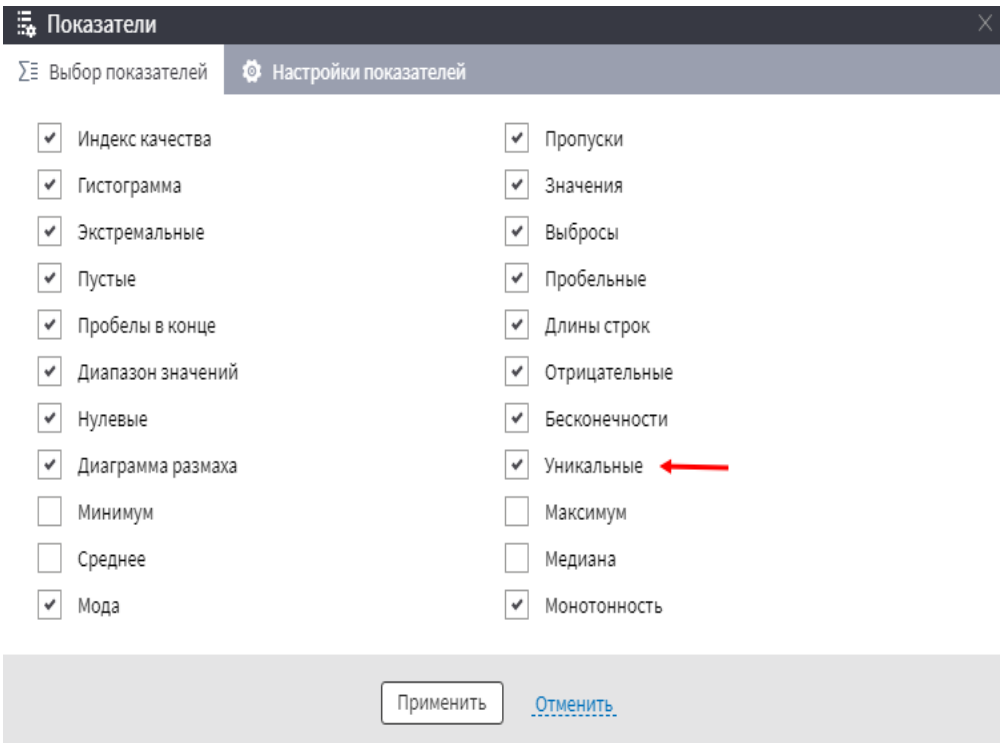


Рис.6 настройка показателей качества данных

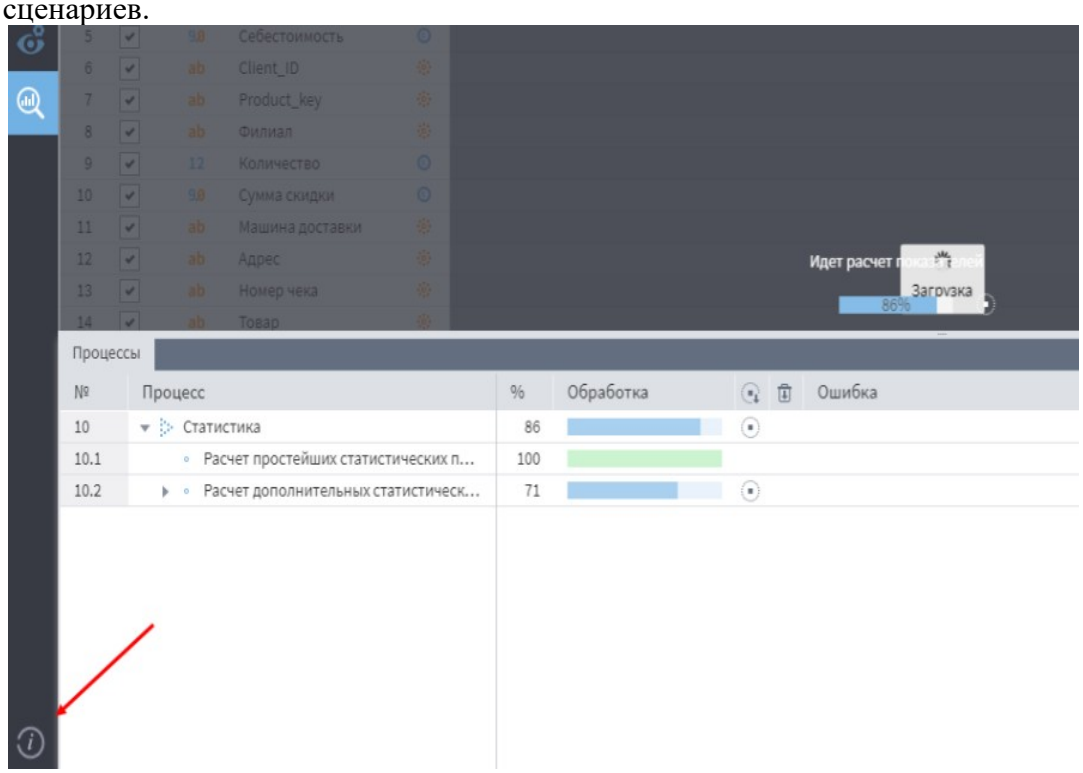
В окне настроек можно выбрать, какие проверки и статистические данные будут рассчитаны в полях. Надо включить еще одну опцию — *Уникальные*.



После нажатия кнопки *Рассчитать статистики* запустится обработка, которая может занять определенное время.

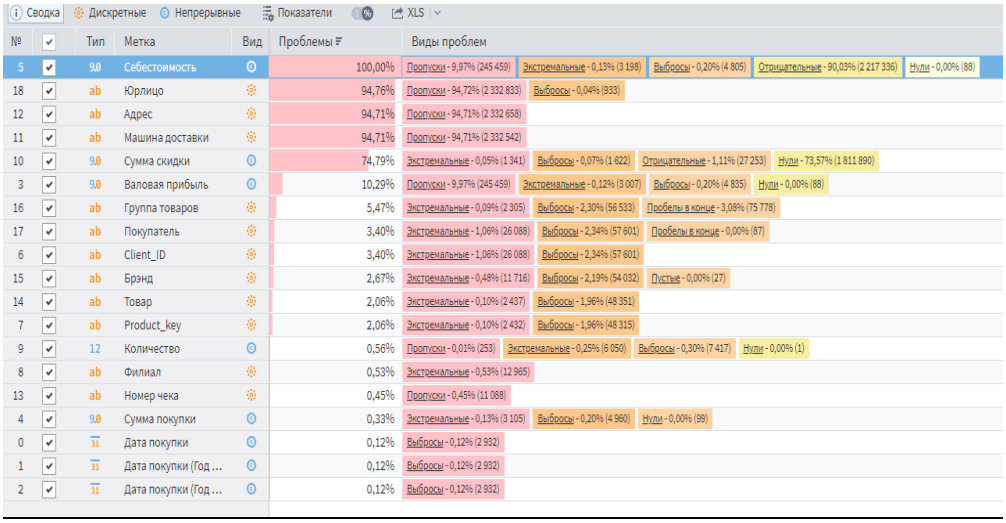
Для контроля процесса выполнения можно открыть панель процессов, нажав на букву ***i*** в левом нижнем углу. В сложных подмоделях бывает непросто оценить время, необходимое для завершения расчетов, а на панели процессов будут видны все выполняемые операции.

Кстати, во время выполнения расчетов интерфейс программы не блокируется. Можно продолжать работать с данными уже активированных узлов или достраивать сценарий. Loginom распараллеливает обработку насколько это возможно, что обеспечивает высокую скорость выполнения сценариев.



Сводные результаты проверки

По окончании расчетов будет отображена сводная информация по всем найденным проблемам.



№	Тип	Метка	Вид	Проблемы	Виды проблем
5	9.0	Себестоимость		100,00%	Пропуски - 9,97% (245 459) · Экстремальные - 0,13% (3 198) · Выбросы - 0,20% (4 805) · Отрицательные - 90,03% (2 217 336) · Нули - 0,00% (88)
18	ab	Юрлицо		94,76%	Пропуски - 94,72% (2 332 833) · Выбросы - 0,04% (933)
12	ab	Адрес		94,71%	Пропуски - 94,71% (2 332 658)
11	ab	Машина доставки		94,71%	Пропуски - 94,71% (2 332 542)
10	9.0	Сумма скидки		74,79%	Экстремальные - 0,05% (1 341) · Выбросы - 0,07% (1 622) · Отрицательные - 1,11% (27 253) · Нули - 73,57% (1 811 890)
3	9.0	Валовая прибыль		10,29%	Пропуски - 9,97% (245 459) · Экстремальные - 0,12% (3 007) · Выбросы - 0,20% (4 835) · Нули - 0,00% (88)
16	ab	Группа товаров		5,47%	Экстремальные - 0,09% (2 305) · Выбросы - 2,30% (56 533) · Пробелы в конце - 3,08% (75 778)
17	ab	Покупатель		3,40%	Экстремальные - 1,06% (26 088) · Выбросы - 2,34% (57 601) · Пробелы в конце - 0,00% (87)
6	ab	Client_ID		3,40%	Экстремальные - 1,06% (26 088) · Выбросы - 2,34% (57 601)
15	ab	Брэнд		2,67%	Экстремальные - 0,48% (11 716) · Выбросы - 2,19% (54 032) · Пустые - 0,00% (27)
14	ab	Товар		2,06%	Экстремальные - 0,10% (2 437) · Выбросы - 1,96% (48 351)
7	ab	Product_key		2,06%	Экстремальные - 0,10% (2 432) · Выбросы - 1,96% (48 315)
9	12	Количество		0,56%	Пропуски - 0,01% (253) · Экстремальные - 0,25% (6 050) · Выбросы - 0,30% (7 417) · Нули - 0,00% (1)
8	ab	Филиал		0,53%	Экстремальные - 0,53% (12 965)
13	ab	Номер чека		0,45%	Пропуски - 0,45% (11 088)
4	9.0	Сумма покупки		0,33%	Экстремальные - 0,13% (3 105) · Выбросы - 0,20% (4 960) · Нули - 0,00% (99)
0	31	Дата покупки		0,12%	Выбросы - 0,12% (2 932)
1	31	Дата покупки (Год ...		0,12%	Выбросы - 0,12% (2 932)
2	31	Дата покупки (Год ...		0,12%	Выбросы - 0,12% (2 932)

Система выполнила указанные проверки и разместила поля в порядке убывания количества проблем.

Пропуски

Пропуски — количество null-значений в полях. Null — это не пробел или ноль, а **отсутствие данных**. Если в текстовом поле есть непечатный символ, например пробел или табуляция, то он не является null-ом, т.е. пропуском, хотя на экране выглядит так же.

Является ли пропуск проблемой или нет — зависит от назначения поля. При работе с атрибутом, присутствие которого опционально, это может и не быть проблемой, хотя на пропуски и в этом случае все равно стоит обратить внимание.

Но наличие null-ов в полях, используемых для расчета показателей, почти всегда создает трудности. В анализируемой выборке в поле *Себестоимость* 9.97% пропущенных данных.

Причин такой ситуации может быть много. Например, при приемке товара не заносилась информация о стоимости закупки, или в процессе импорта данных был неверно настроен формат дробного числа.

В ранее подготовленном сценарии валовая прибыль считалась как разница между выручкой и себестоимостью. Наличие пропусков в поле *Себестоимость* привело к неверным цифрам.

При выполнении математических операций с числом и пустым (null) значением возвращается пустое значение. Можно обратить внимание, что процент пропусков в поле *Валовая прибыль* такой же, как в поле *Себестоимость*. А значит в отчетах получается заниженная прибыль.

Много пропусков в полях *Юрлицо, Адрес, Машина доставки*. Забегая вперед можно сказать, что это корректно с технической точки зрения, т.е. ошибки нет. Однако подобная картина является указанием на то, что анализируется весьма неоднородный массив данных с точки зрения процессов, которые его формируют.

Выбросы и экстремальные значения

Выбросы и экстремальные значения показывают наличие значений, статистически выбивающихся из конкретного поля. Является ли это проблемой, зависит от того, как планируется использовать эти данные.

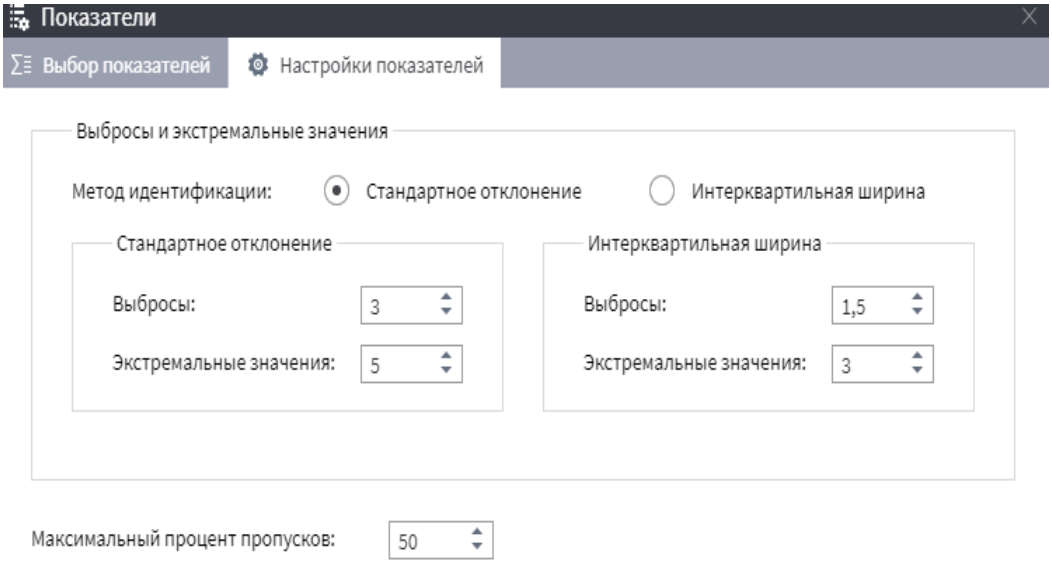
При необходимости визуализировать отчет по фактическим событиям — проблем нет. Если в среднем были продажи на 1 млн рублей в месяц, а однажды была отгрузка на 10 млн рублей крупному заказчику, то именно это и требуется показать в отчетах, потому что все так и было.

Но если эти данные надо использовать для моделирования, прогнозирования, расчета статистически значимых показателей, выбросы и экстремальные значения могут испортить картину. Например, завысить средний чек у клиентов до уровня VIP и создать предпосылки для неверных выводов при планировании.

По смыслу выбросы и экстремальные значения довольно близки:

- **Выброс** — значение, выбивающееся из общего ряда;
- **Экстремальное значение** — это очень большой выброс.

Правила определения выбросов и экстремальных значений определяются в настройках параметров.



Пробелы в конце

Пробелы в конце — тип ошибок, который трудно обнаружить на глаз. Например, определить разницу между названиями «*Компания Орион»* и «*Компания Орион »*, если значения не взяты в кавычки, практически нереально. Loginom позволяет подсвечивать подобные ситуации.

При этом под пробелами подразумевается не только собственно пробел, который находится на клавиатуре. Но и целый набор пробельных символов, вроде переноса строк или табуляции.

Обычно наличие таких значений является ошибкой ввода данных в учетные системы. Это может создать проблемы, когда они будут связываться со значениям из другой системы, но уже без пробелов. Для программы это будет 2 разные строки, и при фильтре по «*Компания Орион»* в выборку не попадут значения, связанные с *«Компания Орион »*.

Для исправление конкретно этой проблемы в калькуляторе есть функция *Trim()*, обрезающая открывающие и закрывающие пробелы в строке.

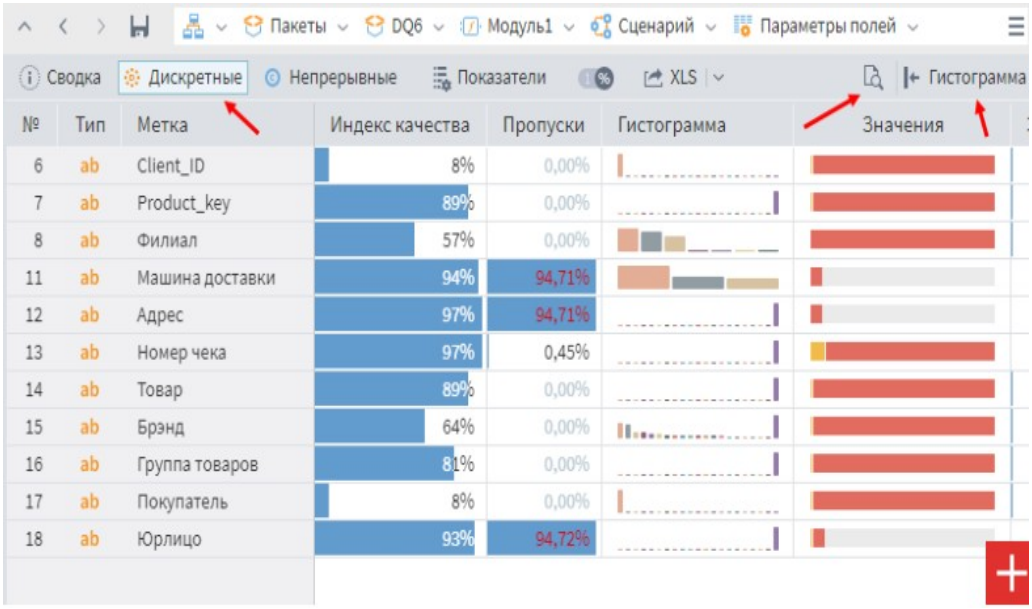
Более подробно о всех типах проблем можно прочитать в справке Loginom.

Анализ дискретных полей

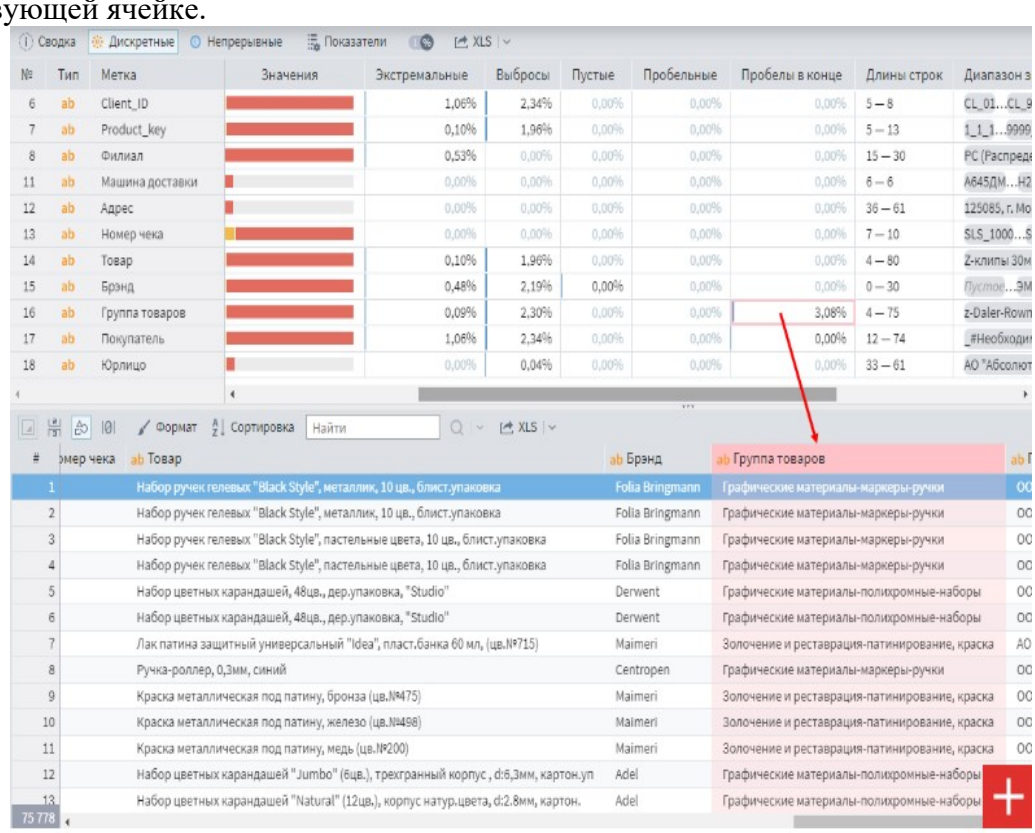
Помимо сводки есть более детальное представление статистики по отдельным полям в зависимости от вида данных.

В визуализаторе *Качество данных* нужно перейти на вкладку *Дискретные*. Дискретными являются текстовые и логические поля, но могут быть и числовые или дата/время, если это будет явно задано.

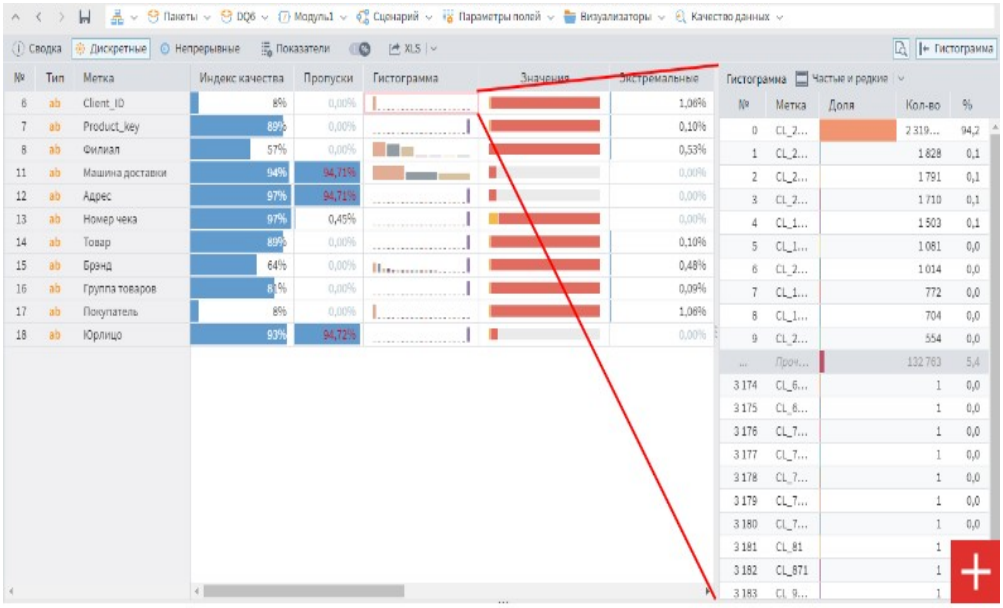
Затем необходимо активировать кнопки *Детализация* и *Гистограмма* в правом верхнем углу.



С помощью детализации можно увидеть конкретные проблемные значения, кликнув по соответствующей ячейке.



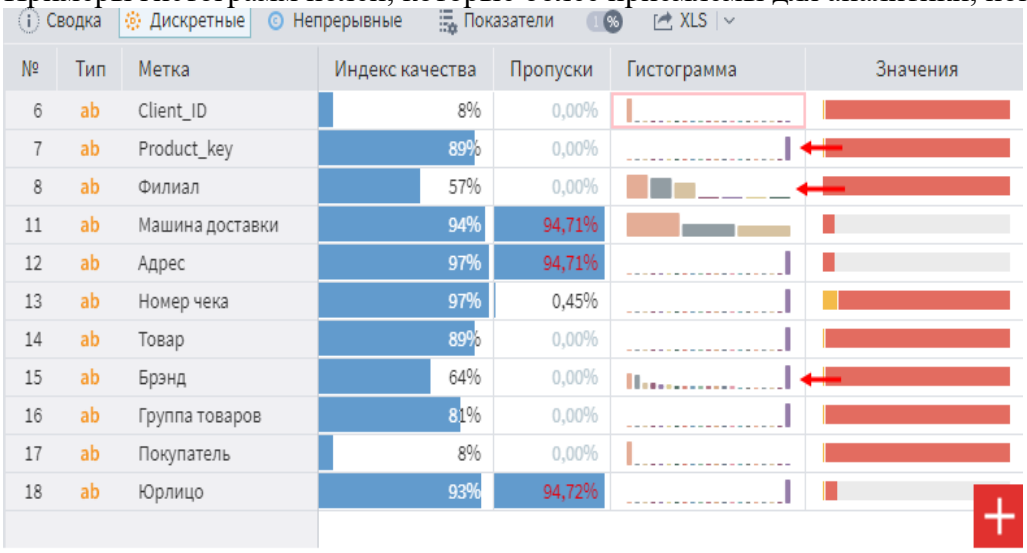
Гистограмма показывает частоту, с которой значения встречаются в поле. Она является хорошим индикатором на предмет потенциальных проблем, которые поле может привнести в аналитику.



Когда для дискретного поля гистограмма выглядит так, как на картинке выше, это означает, что большинство значений поля имеет одно значение, а следовательно:

1. Использовать его как аналитический признак может быть нецелесообразно, т.к. недостаточно разнообразия данных, чтобы делать выводы.
2. Возможно, в датасете намешаны данные из разных по смыслу процессов.

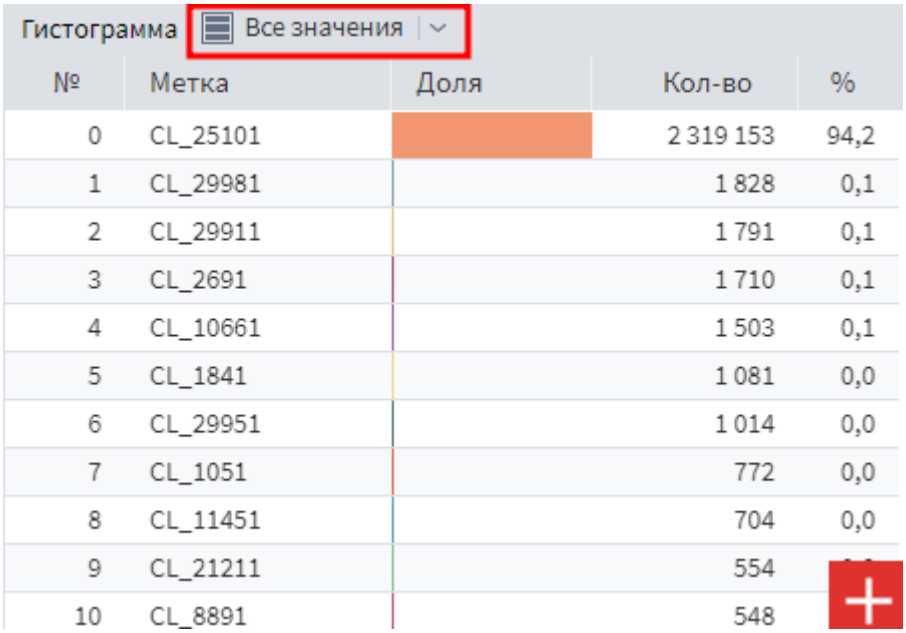
Примеры гистограмм полей, которые более приемлемы для аналитики, показаны стрелками.



Варианты следующие:

1. Все столбцы маленькие, а последний большой — это значит, что в поле нет значений, явно выбивающихся из общей частотности. Большой последний столбец — это прочие, т.е. значения, не вошедшие в топ первых по частоте.
2. Все столбцы или их существенная часть хорошо различима на гистограмме. Значит значения представлены достаточно равномерно.
3. Столбцы убывающей высоты и большая последняя колонка — комбинация вариантов 1 и 2.

Кстати, в детализированной гистограмме (справа) можно переключать режимы отображения, например, показать все уникальные значения поля. Это простой способ быстро посмотреть значения по любому полю таблицы.


Гистограмма Все значения

№	Метка	Доля	Кол-во	%
0	CL_25101		2 319 153	94,2
1	CL_29981		1 828	0,1
2	CL_29911		1 791	0,1
3	CL_2691		1 710	0,1
4	CL_10661		1 503	0,1
5	CL_1841		1 081	0,0
6	CL_29951		1 014	0,0
7	CL_1051		772	0,0
8	CL_11451		704	0,0
9	CL_21211		554	
10	CL_8891		548	

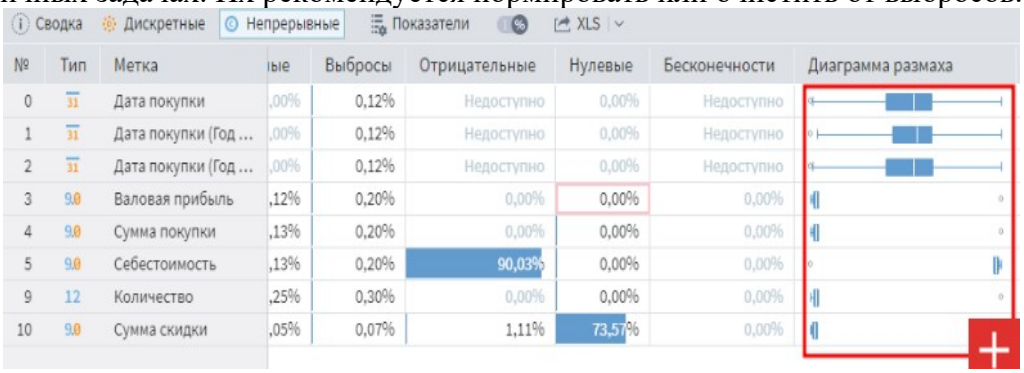
Анализ непрерывных полей

Для непрерывных полей гистограммы строятся по диапазонам значений.

Для экспресс-анализа имеет смысл обратить внимание на диаграмму размаха. Чем более эта диаграмма симметрична и центрирована, тем более равномерно распределены значения поля. А значит, тем выше однородность процессов, формирующих эти данные.

Пример подозрительных диаграмм можно посмотреть в строках, начиная с четвертой. Так выглядят диаграммы с экстремальными выбросами.

Такие поля чаще всего не подходят для использования в моделировании, сегментации и аналогичных задачах. Их рекомендуется нормировать или очистить от выбросов.


Сводка | Дискретные | Непрерывные | Показатели | XLS

№	Тип	Метка	ые	Выбросы	Отрицательные	Нулевые	Бесконечности	Диаграмма размаха
0	31	Дата покупки	,00%	0,12%	Недоступно	0,00%	Недоступно	
1	31	Дата покупки (Год ...	,00%	0,12%	Недоступно	0,00%	Недоступно	
2	31	Дата покупки (Год ...	,00%	0,12%	Недоступно	0,00%	Недоступно	
3	9.0	Валовая прибыль	,12%	0,20%	0,00%	0,00%	0,00%	
4	9.0	Сумма покупки	,13%	0,20%	0,00%	0,00%	0,00%	
5	9.0	Себестоимость	,13%	0,20%	90,03%	0,00%	0,00%	
9	12	Количество	,25%	0,30%	0,00%	0,00%	0,00%	
10	9.0	Сумма скидки	,05%	0,07%	1,11%	73,57%	0,00%	

Задание.

Визуализатор *Качество данных* — отличный инструмент для экспресс-аудита данных, особенно на этапе разведочного анализа. Однако, помимо технических проблем, нужно понимать, насколько текущий массив данных подходит для решения бизнес-задач.

Наша задача — **построить клиентскую аналитику оптовых продаж**, а также финансовый портрет клиентов в разрезе их уровней. Технически, мы ее уже решили. А теперь, с помощью узла *Качество данных* проведите поиск проблем, которые ставят под сомнение достоверность и качество результатов, которые мы получили.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1. Охарактеризуйте метод «Деревьев решений».
2. Охарактеризуйте особенности регрессионного анализа в методах ИАД.
3. Охарактеризуйте модели временных рядов с запаздываниями.
4. Охарактеризуйте метод «Ближайшего соседа».

5. Охарактеризуйте метод поиска правила.
6. Охарактеризуйте метод кластеризации.
7. Охарактеризуйте метод классификации.
8. Охарактеризуйте метод дискриминации.
9. Какие различия в целях и алгоритмах статистического и интеллектуального подходов.

Лабораторная работа № 6.
Тема: Дублирование данных

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

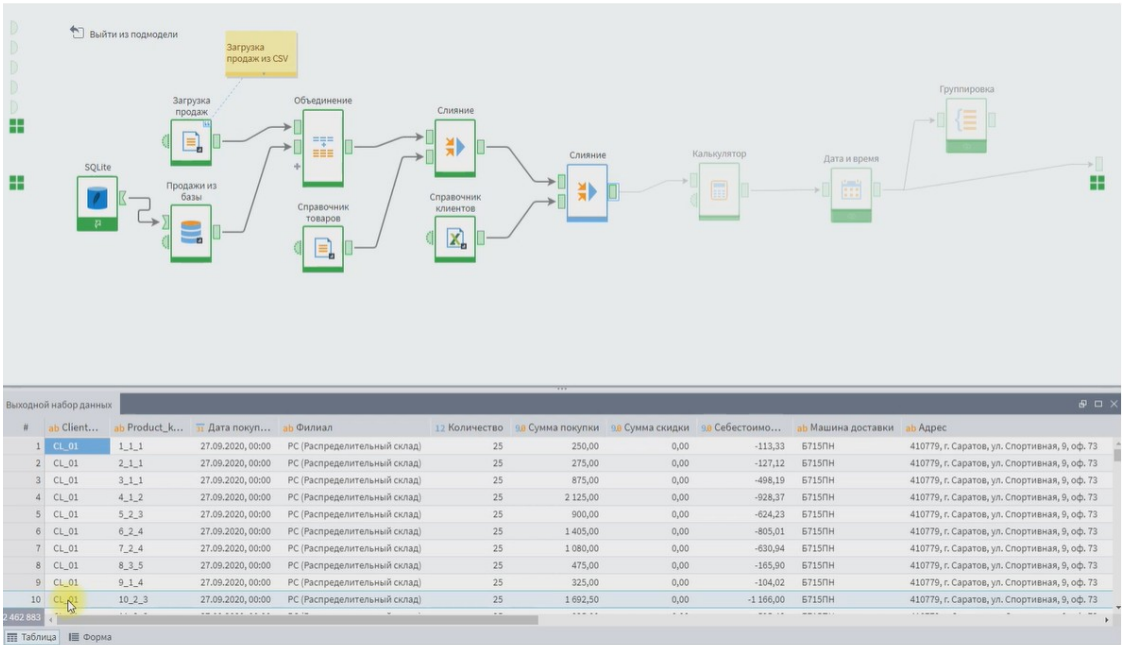
Теоретическая часть

Дублирование данных при слиянии

Зайдите в подмодель *Продажи* и проверьте, увеличивается ли количество строк после слияний. Если да — значит в таблице возникли дубли.

Если соединение таблиц выполняется в Loginom, то все просто — можно сравнить количество записей на выходных портах узлов до соединения и после.

Зайдите в подмодель *Продажи* и проверьте, увеличивается ли количество строк после слияний. Если да — значит в таблице возникли дубли.



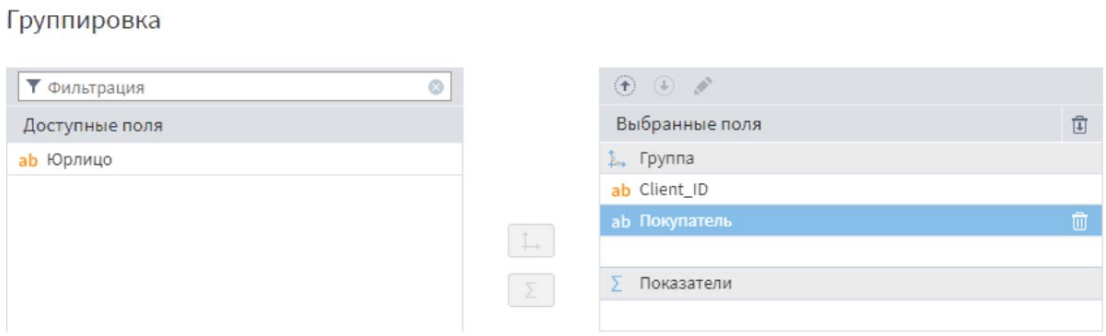
#	Client…	Product_k…	Дата покуп…	Филиал	Количество	Сумма покупки	Сумма скидки	Себестоимо…	Машина доставки	Адрес
1	CL_01	1_1_1	27.09.2020, 00:00	РС (Распределительный склад)	25	250,00	0,00	-113,33	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
2	CL_01	2_1_1	27.09.2020, 00:00	РС (Распределительный склад)	25	275,00	0,00	-127,12	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
3	CL_01	3_1_1	27.09.2020, 00:00	РС (Распределительный склад)	25	875,00	0,00	-498,19	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
4	CL_01	4_1_2	27.09.2020, 00:00	РС (Распределительный склад)	25	2 125,00	0,00	-928,37	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
5	CL_01	5_2_3	27.09.2020, 00:00	РС (Распределительный склад)	25	900,00	0,00	-624,23	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
6	CL_01	6_2_4	27.09.2020, 00:00	РС (Распределительный склад)	25	1 405,00	0,00	-805,01	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
7	CL_01	7_2_4	27.09.2020, 00:00	РС (Распределительный склад)	25	1 080,00	0,00	-630,94	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
8	CL_01	8_3_5	27.09.2020, 00:00	РС (Распределительный склад)	25	475,00	0,00	-165,90	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
9	CL_01	9_1_4	27.09.2020, 00:00	РС (Распределительный склад)	25	325,00	0,00	-104,02	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73
10	CL_01	10_2_3	27.09.2020, 00:00	РС (Распределительный склад)	25	1 692,50	0,00	-1 166,00	Б715ПН	410779, г. Саратов, ул. Спортивная, 9, оф. 73

Как видно по ролику, количество строк увеличивается после присоединения справочника клиентов. Почему это происходит?

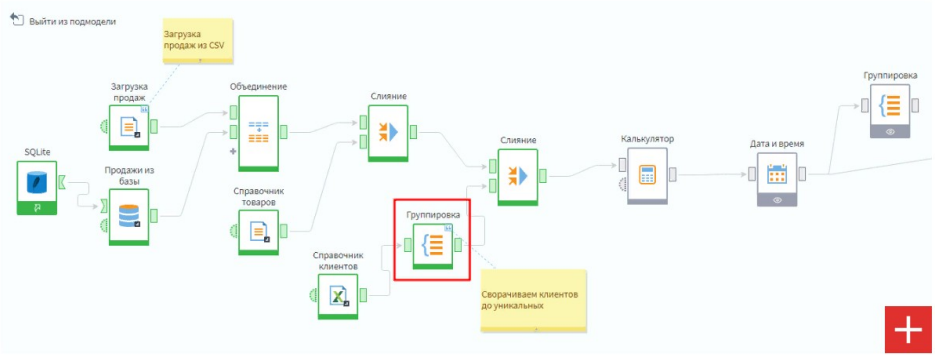
Если изучить справочник клиентов, то можно обнаружить, что встречаются контрагенты у которых несколько юрлиц, а связываются таблицы по идентификатору клиента. При слиянии каждый клиент с несколькими юридическими лицами дублирует количество своих продаж столько раз, сколько у него юрлиц.

#	ab Client...	ab Покупатель	ab Юрлицо
1	CL_01	ООО "Положение" ИНН 7896913934	ООО "Положение" ИНН 7896913934 ЮРЛИЦО №1
2	CL_10001	ООО "Решительная скорость" ИНН 96234274341	ООО "Решительная скорость" ИНН 96234274341 ЮРЛИЦО №1
3	CL_1001	ООО "Управление" ИНН 66912194025	ООО "Управление" ИНН 66912194025 ЮРЛИЦО №1
4	CL_10011	ПАО "Час" ИНН 13746866995	ПАО "Час" ИНН 13746866995 ЮРЛИЦО №1
5	CL_10021	ООО "Зал" ИНН 84426324283	ООО "Зал" ИНН 84426324283 ЮРЛИЦО №1
6	CL_10031	ООО "Фактор" ИНН 84921320774	ООО "Фактор" ИНН 84921320774 ЮРЛИЦО №1
7	CL_10041	ООО "Правительство" ИНН 39913231628	ООО "Правительство" ИНН 39913231628 ЮРЛИЦО №1
8	CL_10041	ООО "Правительство" ИНН 39913231628	ООО "Правительство" ИНН 39913231628 ЮРЛИЦО №3
9	CL_10041	ООО "Правительство" ИНН 39913231628	ООО "Правительство" ИНН 39913231628 ЮРЛИЦО №4
10	CL_10051	ПАО "Массовый круг" ИНН 61958913507	ПАО "Массовый круг" ИНН 61958913507 ЮРЛИЦО №1
3 838	CL_10061	ООО "Наибольшое право" ИНН 7940585529	ООО "Наибольшое право" ИНН 7940585529 ЮРЛИЦО №1

Это проблему можно решить с помощью узла *Группировка*, группируя только уникальные комбинации тех полей, которые требуются. При этом поля в область показателей можно не добавлять.



Надо добавить эту группировку после загрузки справочника клиентов, но до слияния с продажами. Полем *Юрлицо* жертвуем, потому что оно непригодно для аналитики с этим массивом данных.



Добавление группировки в сценарий

Можно убедиться, что после группировки при слиянии продаж с клиентами количество строк в таблице не увеличивается.

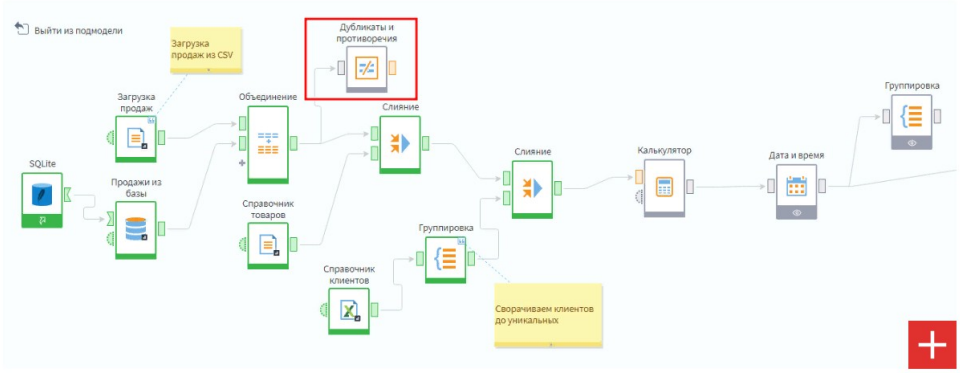
Поиск дублей во входных наборах

Следующий вопрос: как быть с таблицами, которые загружаются в готовом виде? Ранее импортировались транзакции из CSV и базы данных без понимания истории их возникновения.

Таблицы могут быть получены аналитиком с пониманием логики их создания. Как следствие, у пользователя больше доверия к таким данным. Но часто аналитики работают с выгрузками, сделанными неизвестным программистом, который уже несколько лет не работает в компании.

Лучшее, что можно сделать в этой ситуации, отталкиваться от подхода «доверяй, но проверяй». Если ошибки не будут обнаружены — хорошо, проверка повысит уровень доверия к анализируемой информации.

Надо добавить в сценарий узел ***Дубликаты и противоречия*** из группы ***Исследование***. Затем подать на вход объединенную таблицу транзакций до присоединения справочников.

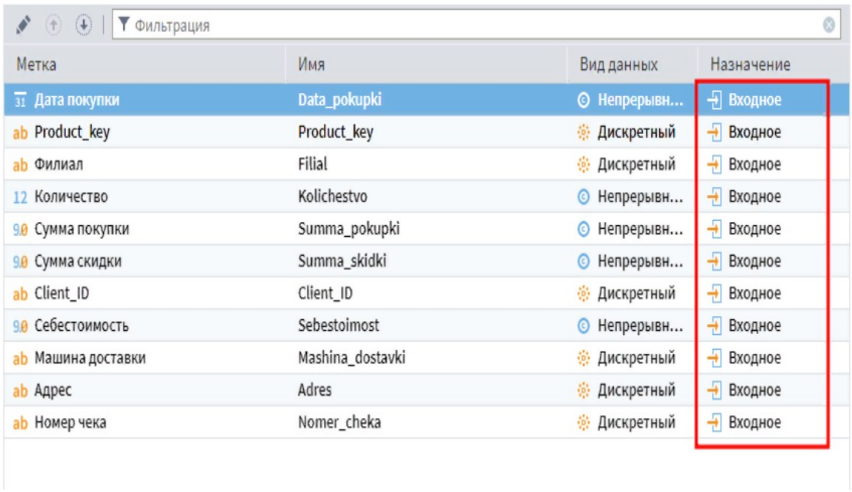


Добавление узла поиска дублей и противоречий

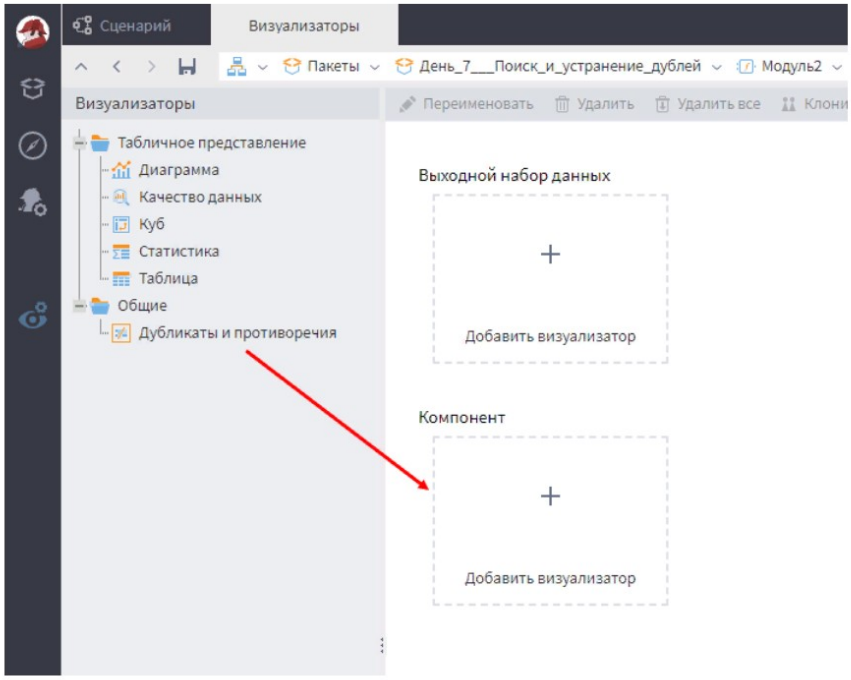
После запуска настройки узла видно, что в таблице транзакций нет поля, которое позволяет однозначно идентифицировать транзакцию, т.е. нет первичного ключа. Поле *Номер чека* не подходит, потому что оно дублируется, если в чеке несколько позиций.

Поэтому можно исходить из того, что в таблице не должно быть ни одной дублирующейся строки.

Для поиска дублирующихся строк надо все поля в настройках узла *Дубликаты и противоречия* определить как входные.

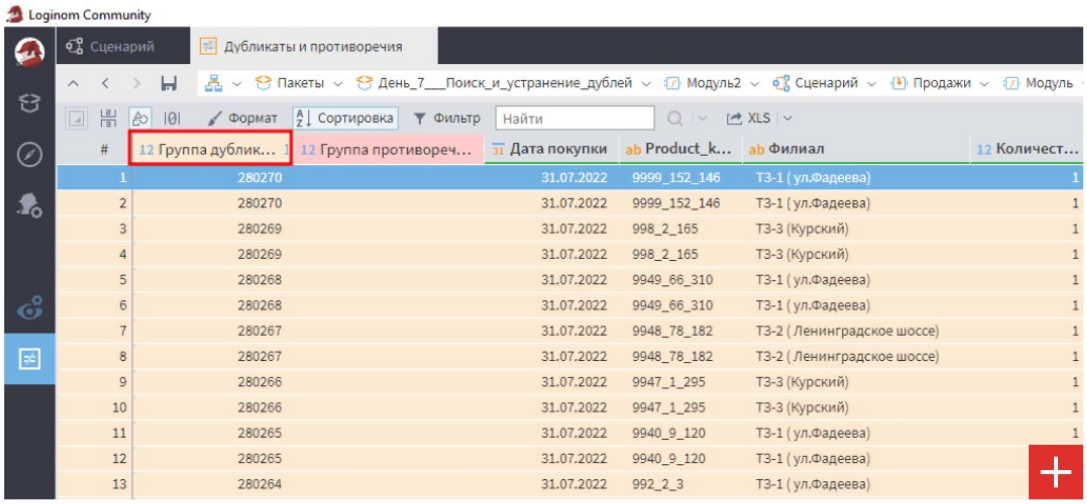


После активации узла нужно зайти в настройку его визуализаторов. Некоторые обработчики имеют свои особенные визуализаторы, и *Дубликаты и противоречия* — один из них. Надо перетащить визуализатор в область компонентов.



В визуализаторе *Дубликаты и противоречия* требуется отсортировать по убыванию группы дубликатов. Нужные строки отобразятся вверху таблицы.

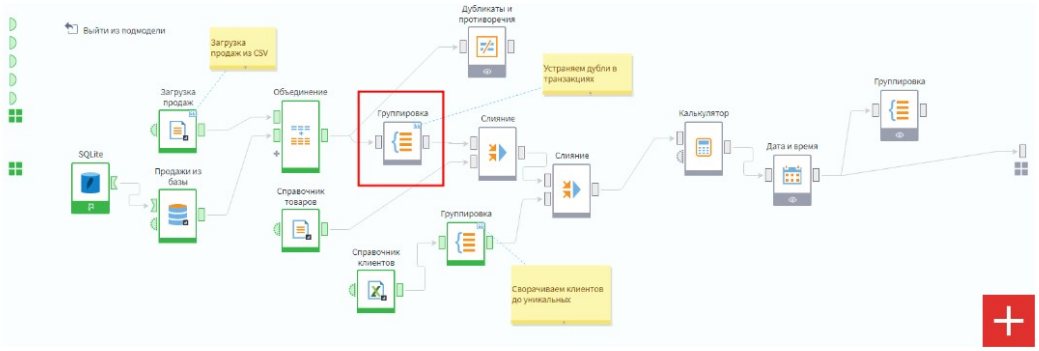
Как видно на скриншоте, выявлено 280 270 групп дублей! Неплохой сюрприз от создателя таблицы



#	Группа дублик...	Группа противореч...	Дата покупки	Product_k...	Филиал	Количест...
1	280270		31.07.2022	9999_152_146	ТЭ-1 (ул.Фадеева)	1
2	280270		31.07.2022	9999_152_146	ТЭ-1 (ул.Фадеева)	1
3	280269		31.07.2022	998_2_165	ТЭ-3 (Курский)	1
4	280269		31.07.2022	998_2_165	ТЭ-3 (Курский)	1
5	280268		31.07.2022	9949_66_310	ТЭ-1 (ул.Фадеева)	1
6	280268		31.07.2022	9949_66_310	ТЭ-1 (ул.Фадеева)	1
7	280267		31.07.2022	9948_78_182	ТЭ-2 (Ленинградское шоссе)	1
8	280267		31.07.2022	9948_78_182	ТЭ-2 (Ленинградское шоссе)	1
9	280266		31.07.2022	9947_1_295	ТЭ-3 (Курский)	1
10	280266		31.07.2022	9947_1_295	ТЭ-3 (Курский)	1
11	280265		31.07.2022	9940_9_120	ТЭ-1 (ул.Фадеева)	1
12	280265		31.07.2022	9940_9_120	ТЭ-1 (ул.Фадеева)	
13	280264		31.07.2022	992_2_3	ТЭ-1 (ул.Фадеева)	

Проблемы надо решать. Есть 2 варианта:

1. **Социальная инженерия**. Нужно найти разработчика некорректной таблицы/выгрузки и добиться того, чтобы он переписал запрос. Это правильный подход, особенно если полученные данные используются не только для аналитики, но и в других системах и процессах.
2. **Исправить проблему на стороне Loginom**. Для этого потребуется группировка. Это и будет домашним заданием — свернуть дублирующиеся транзакции с помощью группировки по всем полям транзакций.



Исключение дублей

Возникает вопрос: насколько корректно для исключения дублей просто брать и группировать все поля транзакций?

Здесь потребуется понимание логики формирования документов. Если клиенту внесли несколько экземпляров одного и тоже же товара в один чек, то записи должны схлопнуться, а такие показатели, как *Количество* и *Сумма покупки* сложиться. Каждый товар в чеке должен быть уникальным.

Таким образом, возникновение 2-х одинаковых строк в этой таблице невозможно. Следовательно, повторяющиеся строки — это дубли, а не фактические продажи.

В реальных проектах надо убедиться, что логика формирования строк в таблице продаж такая, как описана выше. Только в этом случае можно использовать подобную группировку для исключения дублей.

Поиск противоречий в данных

В данных помимо дублей может быть и другой тип ошибок — противоречия. Он близок к дублям, но есть и отличия. Именно поэтому обработчик называется *Дубликаты и противоречия*!

Поиск противоречий — это такой вариант сканирования данных, когда обнаруживаются строки, у которых входные поля дублируются, а выходные — отличаются.

Проще всего это объяснить на примере.

Товар (Входное)	Категория (Выходное)	Дубликат	Противоречие
Фломастер	Товары для рисования	1	1
Фломастер	Детские игрушки	1	1
Карандаш	Товары для рисования	2	
Карандаш	Товары для рисования	2	

В таблице поле *Товар* указано как входное, а поле *Категория* – как выходное. Т.е. для всех товаров с одинаковыми названием должна быть одна и так же категория.

Как видно, это справедливо для карандашей, все они относятся к категории товаров для рисования. Но в случае фломастеров имеются противоречия: одна запись отнесена к категории товаров для рисования, а вторая — детских игрушек. Такого быть не должно, данные противоречивы.

В общем случае входных и выходных полей может быть несколько. Таким образом, можно перефразировать так: противоречие — это когда одинаковому набору значений входных полей соответствует разные значения выходных.

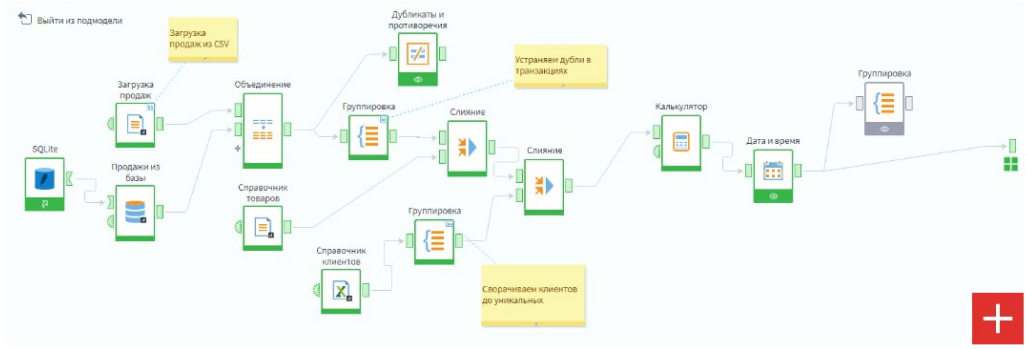
Противоречия можно использовать не только для поиска технических ошибок, но и для выявления проблем в бизнес-логике.

Задание.

Задание №1. Устраните с помощью группировки всех полей дубли в таблице транзакций. В итоге из подмодели *Продажи* должно выходить 2 155 970 строк.

Задание №2. С помощью перенастройки узла *Дубликаты и противоречия* найдите все случаи, когда в один день на один адрес приезжали разные машины доставки. Будем считать, что это неэффективно с точки зрения логистики, поэтому такие случаи нужно выявлять для разбора.

Пример структуры итогового сценария:



Итоговый сценарий

Может возникнуть вопрос: почему узел *Дубликаты и противоречия* висит как тупиковая ветка, а не стоит перед/после группировки? Ответ: для того, чтобы построение итоговой таблицы продаж не дожидалось выполнения проверки на дубликаты.

Посмотрите, как устранение дубликатов повлияет на портреты клиентов. Вот какой был профиль до их исключения

Факты

	Выручка (пп)		Валовая прибыль (пп)		Кол-во покупок (пп)	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%|Проц...	∑ Значе...	%|Проц...	∑ Значение	∑ Значение	∑ Значение
1. VIP	253 720 202	81,87%	113 584 57...	81,73%	1 740	145 816,21	65 278,49
2. Важный	21 056 275	6,79%	9 519 220,79	6,85%	1 772	11 882,77	5 372,02
3. Перспек...	23 242 134	7,50%	10 519 019,51	7,57%	2 778	8 366,50	3 786,54
4. Начина...	11 892 488	3,84%	5 360 642,45	3,86%	2 708	4 391,61	1 979,56
Итого:	309 911 099	100,00%	138 983 45...	100,00%	8 998	34 442,22	15 446,04

А вот как стало после.

Факты

	Выручка (пп)		Валовая прибыль (пп)		Кол-во покупок (пп)	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%|Проц...	∑ Значе...	%|Проц...	∑ Значение	∑ Значение	∑ Значение
1. VIP	196 244 612	81,73%	88 039 147,24	81,59%	1 429	137 330,03	61 608,92
2. Важный	12 846 264	5,35%	5 794 241,92	5,37%	1 313	9 783,90	4 412,98
3. Перспек...	18 147 709	7,56%	8 267 434,26	7,66%	2 858	6 349,79	2 892,73
4. Начина...	12 878 607	5,36%	5 803 369,78	5,38%	3 398	3 790,06	1 707,88
Итого:	240 117 192	100,00%	107 904 19...	100,00%	8 998	26 685,62	11 992,02

Ожидаемо, что значения выручки и валовой прибыли снизились. Но интереснее то, что в меньшую сторону изменились значения среднего чека и средней прибыли.

Почему так важен именно средний чек и средняя прибыль? Потому что эти показатели отражают качество клиентской базы. До устранения дублей значение данных характеристик было завышенным. Реальность оказалась суровее.

Слишком оптимистичные показатели создавали ошибочные предпосылки для дальнейших выводов по развитию клиентской базы.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1. Охарактеризуйте генетические алгоритмы.
2. Охарактеризуйте нейросетевые методы анализа.
3. Охарактеризуйте методы для анализа нечетких множеств.
4. Перечислите основные направления эволюционного моделирования и приведите основные факторы, определяющие неизбежность эволюции.
5. В чем особенности эволюционного программирования? Приведите основные шаги обобщенного алгоритма эволюционного программирования.
6. Охарактеризуйте метод эволюционных стратегий. В чем его отличие от эволюционного программирования и от генетических алгоритмов?
7. Применение эволюционных вычислений в ИИС.
8. Какие алгоритмы называют генетическими? Сформулируйте основные особенности генетических алгоритмов.
9. Охарактеризуйте простой генетический алгоритм. Приведите пример.

Лабораторная работа № 7.
Тема: Фиктивные данные.

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

Смысловая фиктивность

При анализе транзакций легко заметить, что **95% продаж** приходится на покупателя ***Частное лицо***. Такая ситуация возникает, когда продажи через контрольно-кассовую технику привязываются к виртуальному клиенту, олицетворяющему всех клиентов-физлиц.

Трюк с виртуальной компанией используется очень часто. Например, так могут оформляться перемещения товаров между подразделениями, отгрузка рекламной или акционной продукции, списание брака и прочее.

Гистограмма | Все значения

№	Метка	Доля	Кол-во	%
0	Частное лицо		2 319 153	94,2
1	0_Администратор Курский		1 828	0,1
2	0_Администратор Речного		1 791	0,1
3	ООО "Команда" ИНН 53146401966		1 710	0,1
4	АО "Бурная трава" ИНН 97892907130		1 503	0,1
5	ООО "Управление" ИНН 15228944452		1 081	0,0
6	0_Администратор Фадеева		1 014	0,0
7	ПАО "Настоящая реформа" ИНН 24946103725		772	0,0
8	ООО "Ослепительная комиссия" ИНН 26414489758		704	0,0
9	ООО "Редкая сила" ИНН 85476084426		554	0,0
10	ООО "Буйное колено" ИНН 55401732908		548	0,0
11	АО "Добрая игра" ИНН 84426779602		547	0,0
12	ООО "Жизненный костюм" ИНН 94231125853		540	0,0
13	ООО "Необходимость" ИНН 48135559850		512	0,0
14	ООО "Художник" ИНН 25920802362		490	0,0
15	ООО "Яркое море" ИНН 72995798164		469	0,0
16	ООО "Стойкий город" ИНН 31172036083		458	0,0

Однако сквозная задача марафона посвящена анализу клиентской базы по **оптовым продажам**. Поэтому наличие подобного виртуального покупателя в отчете ***Портрет клиента*** искажает статистику.

По объемам продажи *Частное лицо* попадает в категорию VIP-клиентов. Более того — его можно считать чуть ли не единственным VIP-клиентом. Он перетягивает на себя почти всю выручку, завышает средний чек и пессимизирует вклад в продажи других покупателей.

Факты

Уровень клие...	Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%|Проц...	∑ Значе...	%|Проц...	∑ Значе...	∑ Значение	∑ Значение
1. VIP	196 244 612	81,73%	88 039 147,24	81,59%	1 429	137 330,03	61 608,92
2. Важный	12 846 264	5,35%	5 794 241,92	5,37%	1 313	9 783,90	4 412,98
3. Перспек...	18 147 709	7,56%	8 267 434,26	7,66%	2 858	6 349,79	2 892,73
4. Начина...	12 878 607	5,36%	5 803 369,78	5,38%	3 398	3 790,06	1 707,88
Итого:	240 117 192	100,00%	107 904 19...	100,00%	8 998	26 685,62	11 992,02

Если строить финансовую модель, опираясь на такую структуру VIP-клиентов, то будут сделаны неверные выводы: фиктивный клиент объединен с реальными, а продажи в розницу с оптовыми. Полагаться на такие цифры нельзя.

В широком смысле фиктивные данные — это информация, которая не подходит для выбранного сценария обработки. При анализе в других разрезах фейковые данные могут не представлять проблемы.

Например, при фокусе на товарах, а не клиентах, можно игнорировать информацию о покупателе. Не важно, продан товар частному лицу или организации, следовательно, не принципиально – указан реальный или фиктивный покупатель.

Техническая фиктивность

Ценность данных определяется тем бизнес-процессом, который их породил. Например, информация о продажах частным и юридическим лицам существенно не отличается — она одинаковая. Но анализировать ее с целью улучшения процесса можно, только разделяя продажи по направлениям.

Однако помимо «естественных» бизнес-процессов в компаниях существуют технические, искажающие наборы данных. Например, бесплатная выдача продукции по промо-акциям проводится как продажа с нулевой суммой.

Логика простая: товар отгружен, и его количество на складе уменьшилось, при этом оплата не предполагается, поэтому сумма продажи равна нулю. Все корректно как с точки зрения логистики, так и финансов. Однако такие отгрузки в отчетах будут фигурировать как обычные продажи и, как следствие, снизят сумму среднего чека.

Если взглянуть на наименования клиентов, то тоже можно обнаружить интересные экземпляры, вроде *_#Необходима перерегистрация*. Что это такое? Имя клиента заменено на нечто техническое? Баг в учетной системе? Загружено не то поле из источника?

В реальности причины могут быть разными и до истины придется докапываться. В нашем примере условимся, что это некие неопределенные клиенты и другой информации о них получить невозможно.

№	Метка	Доля	Кол-во	%
268	ООО "Живая сеть" ИНН 73487241228		114	0,0
269	ООО "Неземной воздух" ИНН 6752604348		114	0,0
270	ООО "Огромное правило" ИНН 9175590440		114	0,0
271	ООО "Стоимость" ИНН 82764377259		114	0,0
272	ООО "Тотальный масштаб" ИНН 8472688047		114	0,0
273	ООО "Ход" ИНН 12988939111		114	0,0
274	ООО "Всемерная стоимость" ИНН 47930481335		113	0,0
275	ООО "Неиссякаемая деталь" ИНН 7356817205		112	0,0
276	_#Необходима перерегистрация 10007986		110	0,0
277	ООО "Исключительный магазин" ИНН 98432645832		110	0,0
278	ООО "Кухня" ИНН 82627860210		110	0,0
279	АО "Прямой шаг" ИНН 85559318538		109	0,0
280	ООО "Черный ответ" ИНН 99139767502		109	0,0
281	_Необходима перерегистрация 10006287		108	0,0
282	АО "Помощь" ИНН 23453141431		108	0,0
283	ООО "Убедительное внимание" ИНН 48362406341		108	0,0
284	ООО "Черный подход" ИНН 14010932107		108	0,0
285	ПАО "Направление" ИНН 51589412496		108	0,0
286	АО "Голос" ИНН 28249585013		107	0,0
287	ООО "Век" ИНН 77802526164		107	0,0
288	ООО "Ошеломляющая экономика" ИНН 36484356888		107	0,0

Еще один источник фиктивных данных о продажах — внутренние перемещения, которые отражают товарооборот между подразделениями.

В финансовых отчетах, например о прибылях и убытках, такие данные имели бы смысл, но в анализе продаж они явно лишние. Дело в том, что эти записи не отражают сути взаимоотношений с реальными клиентами.

Гистограмма | Все значения

№	Метка	Доля	Кол-во	%
0	Частное лицо		2 319 153	94,2
1	0_Администратор Курский		1 828	0,1
2	0_Администратор Речного		1 791	0,1
3	ООО "Команда" ИНН 53146401966		1 710	0,1
4	АО "Бурная трава" ИНН 97892907130		1 503	0,1
5	ООО "Управление" ИНН 15228944452		1 081	0,0
6	0_Администратор Фадеева		1 014	0,0
7	ПАО "Настоящая реформа" ИНН 24946103725		772	0,0
8	ООО "Ослепительная комиссия" ИНН 26414489758		704	0,0
9	ООО "Редкая сила" ИНН 85476084426		554	0,0
10	ООО "Буйное колено" ИНН 55401732908		548	0,0
11	АО "Добрая игра" ИНН 84426779602		547	0,0
12	ООО "Жизненный костюм" ИНН 94231125853		540	0,0
13	ООО "Необходимость" ИНН 48135559850		512	0,0

Нередко такие особенности можно обнаружить только погрузившись в данные. Поэтому надо активно использовать визуализацию и элементы разведочного анализа:

1. Просмотр данных в табличном виде;
2. Построение диаграмм в различных разрезах;
3. Изучение распределений: гистограммы, ящики с усами и т.п.;
4. Фильтрация по вхождению подозрительных символов или слов.

Сценарии работы с фиктивными данными

По большому счету, сценариев работы с фиктивными данными два:

1. **Исключать** фильтрацией на уровне загрузки/обработки данных, чтобы они не искажали расчеты;
2. **Помечать** дополнительными аналитическими признаками для быстрой фильтрации в отчетах.

Исключение при помощи фильтрации нужно, когда фиктивные данные не должны учитываться, т.к. искажают вычисления. Например, в нашем случае не нужны продажи частным лицам. Транзакции с нулевой суммой тоже можно убрать. Так же как и перемещения между подразделениями, которые по сути не являются продажами.

С другой стороны, в данных есть некие неопределенные клиенты, «требующие перерегистрации». Такой метод таск-менеджмента использовала бухгалтерия, чтобы не забыть выполнить необходимые проверки по клиенту.

Это не самый лучший способ отметки записей, чтобы не забыть внести правки, но, возможно, остальные варианты были менее удобными. Желательно иметь возможность мониторить наличие таких клиентов, быстро включать/исключать их из отчета. Поэтому лучше всего их не убирать из данных, а создать дополнительный признак в виде поля, которое будет фильтром в отчете.

Часть 2

Представьте, что вы идете по улице и находите купюру 5 000 рублей. Вокруг — никого. Вы беспрепятственно забираете деньги себе. Надо ли отныне планировать жизнь с расчетом на то, что каждый день на дороге будут лежать 5 000 рублей? Вовсе нет, потому что это редкое событие :(

При принятии решений на основе данных нужно учитывать, что в анализируемом потоке могут попадаться уникальные события, например, продажи редких товаров. Но почему это должно кого-то беспокоить? Ведь имеется хоть и редкое, но реальное событие.

На самом деле все не так просто, т.к. наличие в ассортименте редко покупаемых товарных позиций приводит к проблемам.

Во-первых, бизнес может нести непропорционально большие затраты на инфраструктуру для хранения подобной продукции. Место на складе занято товарами, которые покупают раз в год, а его можно было бы использовать для продукции с высокой оборачиваемостью.

Во-вторых, процесс работы с такими товарами может быть сложнее, чем с ходовыми SKU. Т.к. их количество невелико, нужно отслеживать наличие, отдельно заказывать, формировать партии. Проблема не только в сложности бизнес-процесса, но и в том, что это отнимает ресурсы от работы с популярными позициями.

В-третьих, прогнозы и статистические расчеты на данных, в которых встречаются редкие значения, как правило, получаются менее точными.

Редкие события

По тексту выше можно подумать, что редкие значения — разновидность фейковых данных, но это не так.

Фейковые данные искажают картину из-за того, что эта информация недостоверная (например, несуществующие клиенты) или не имеющая отношения к изучаемому процессу (например, покупки частными лицами при изучении оптовых продаж).

С редкими событиями все не так. Они отражают достоверную информацию и относятся к изучаемому процессу, но низкая частота событий не позволяет их использовать для формирования надежных выводов. Это связано с тем, что каждое редкое событие — уникальное, а следовательно трудно прогнозируемое.

Для работы с ними нужно придумать способ превращения уникального события в обычное, например, за счет объединения нескольких редких событий в одно не столь уникальное.

История аптечной сети. *Ассортимент даже маленькой аптеки — тысячи позиций. Значительная его часть – достаточно ходовые наименования. Однако встречаются и редкие медикаменты. Можно их исключить из ассортимента, но при таком уровне конкуренции, как в этом бизнесе, когда в любом доме по 2-3 аптеки, важен каждый клиент, а значит и покупатель редких лекарств.*

Решение было найдено: компания открыла «Аптеку редких лекарств», в которую в основном завозились редкие позиции. В результате туда приходили покупатели не только с окрестных домов, но и со всей округи. Таким образом, был консолидирован спрос на редкое со всего города.

Ведь если редкое случается часто, оно перестает быть редким! Кроме того, аптека смогла улучшить условия закупок, потому что теперь работала с более крупными партиями.

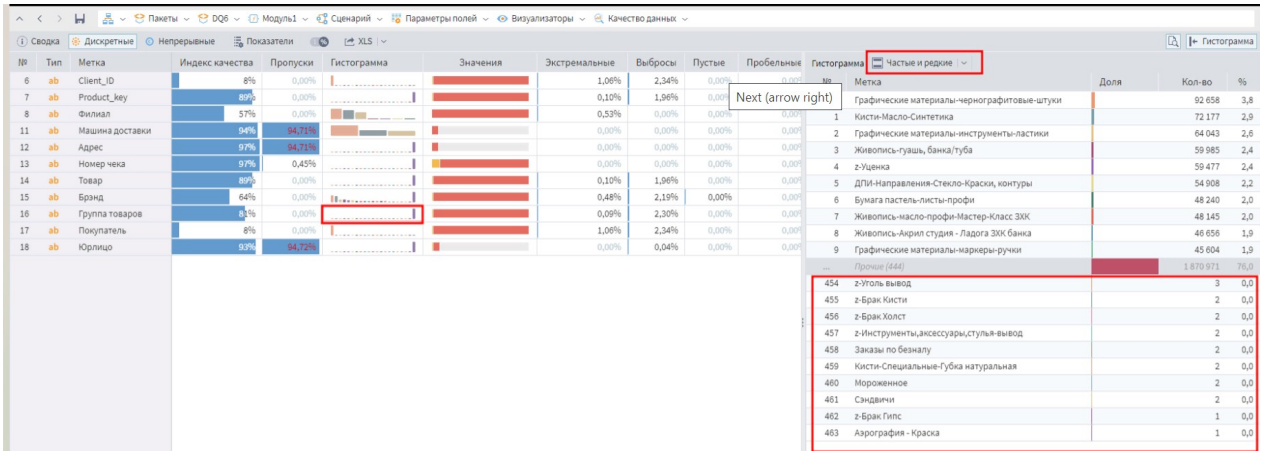
Устойчивый бизнес базируется на регулярных процессах. В идеале хочется продавать ходовые товары стабильно покупающим клиентам, а любое редкое событие нарушает ритмичность процесса и повышает издержки.

Сегментация клиентов с учетом редких позиций

В начале необходимо понять, сколько клиентов приобретает редкие позиции, и как это влияет на продажи. Нужно выделить редкие товары и посчитать, какой процент от закупаемого ассортимента у каждого клиента они составляют.

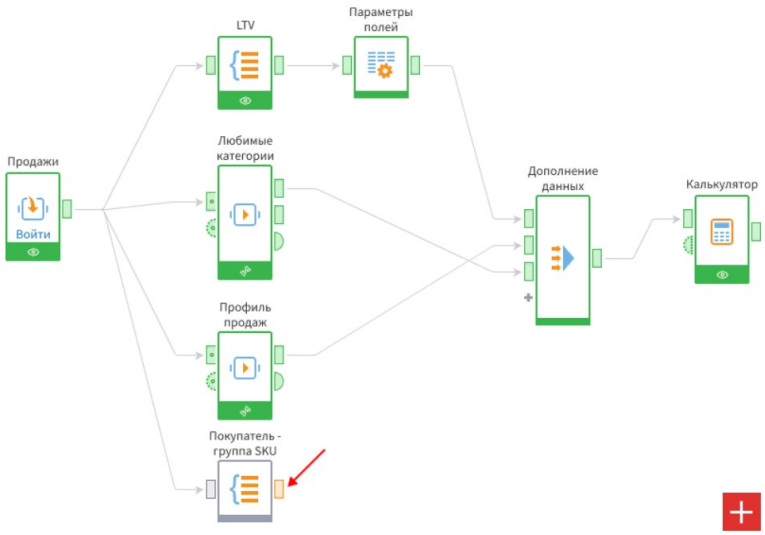
Самый быстрый способ проверить наличие редких значений — посмотреть детализацию *Гистограммы* в визуализаторе *Качество данных*.

Начать стоит с поля *Группа товаров*. Если редко продаются все SKU из целой группы, то это тревожный сигнал.

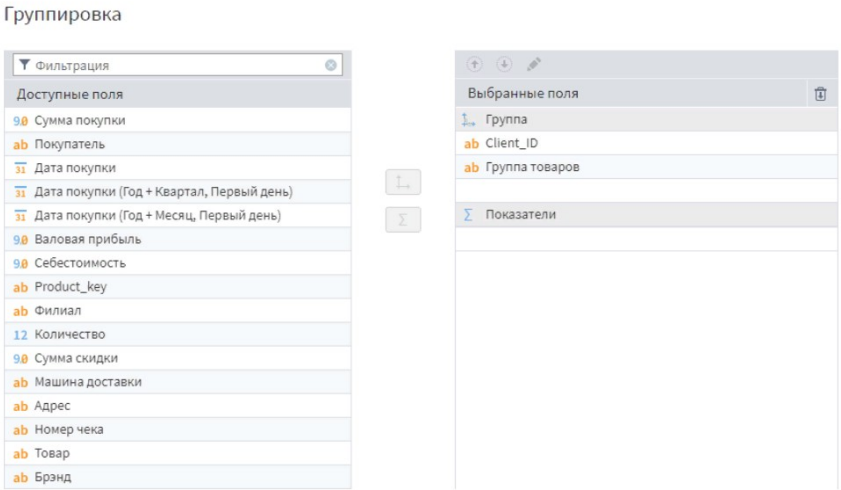


Как видно по гистограмме, редко продаваемые группы встречаются. Товары из некоторых групп были приобретены всего пару раз за несколько лет! Теперь можно оценить, какие клиенты часто покупали редкие позиции.

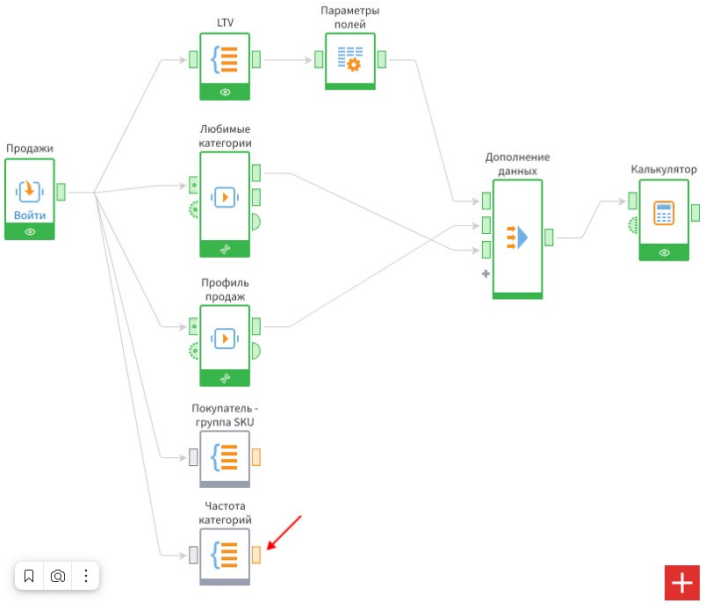
Для этого понадобится сформировать список групп товаров. Его можно получить с помощью узла ***Группировка***. Заведем в него выход из подмодели *Продажи*.



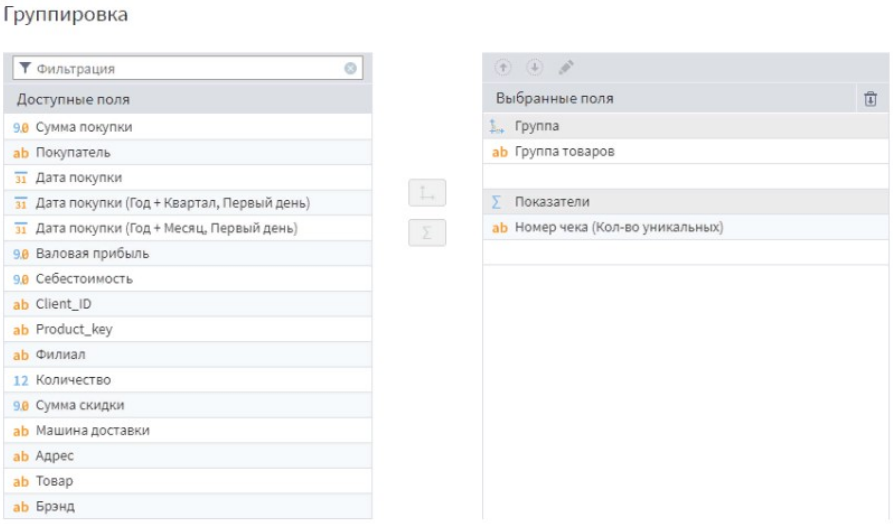
В настройках укажем группы *Client_ID* и *Группа товаров*.



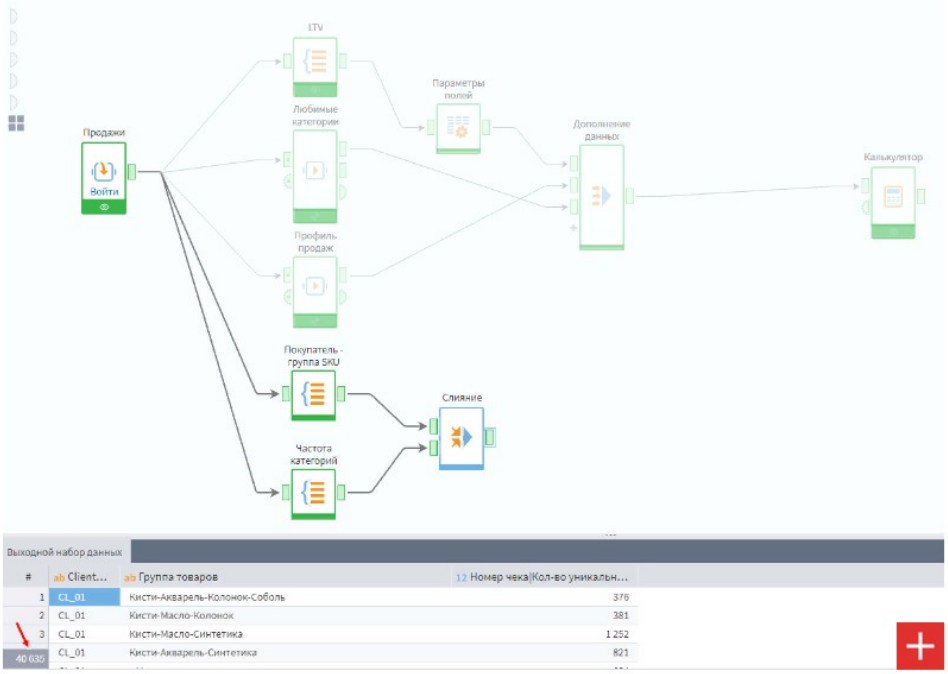
Помимо этого нужно рассчитать, сколько раз каждая категория товаров продавалась **всем клиентам**.



Для подсчета нужно использовать еще одну группировку, в которой будет подсчитано количество уникальных номеров чеков в разрезе товарных групп.

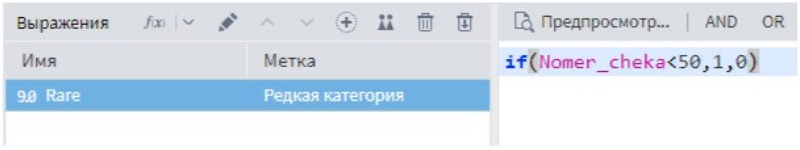


Информацию из этих узлов необходимо соединить при помощи компонента ***Слияние*** по полю *Товарная группа*. Получится таблица, где напротив каждой приобретенной клиентом категории будет общее число продаж этой группы за все время.

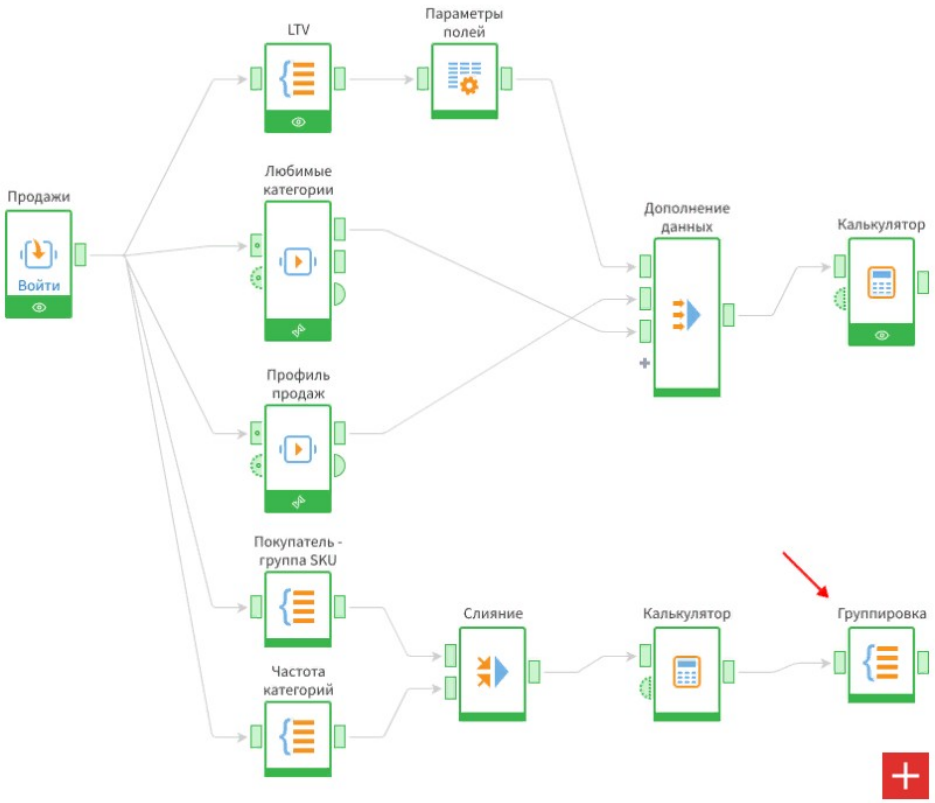


Для расчета процентного соотношения редких и нередких товаров в ассортименте клиента нужно использовать ***Калькулятор***.

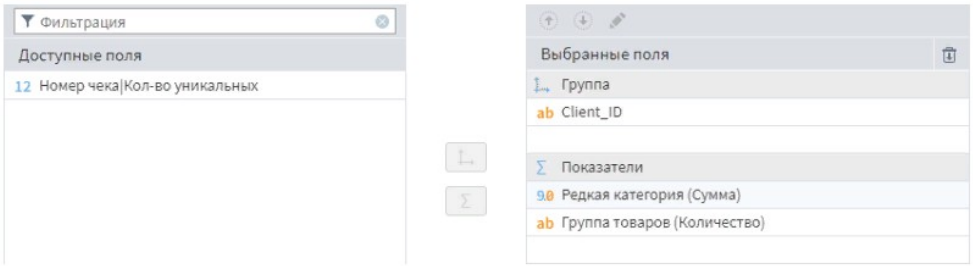
Надо создать числовое поле ***Rare***, которое будет содержать 1, если количество общих продаж категории меньше 50, и 0, если больше.



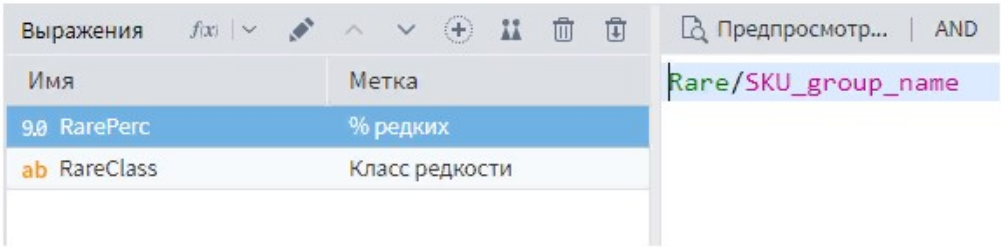
Далее с помощью *Группировки* можно подсчитать количество категорий, которые купил клиент, и сколько из них — редкие



Настройки *Группировки* выглядят так: в разрезе ИД клиента мы считаем сумму по полю ***Редкая категория*** (там 1 или 0) и общее количество купленных товарных групп.

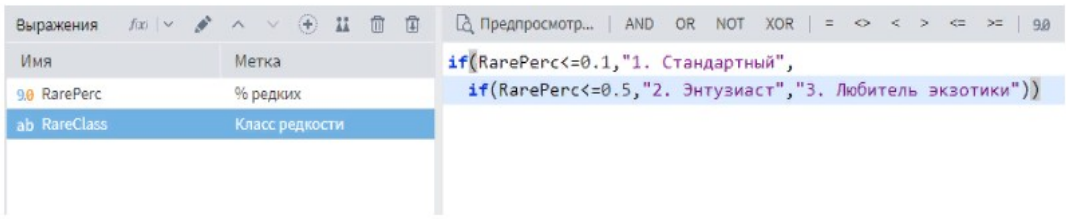


Осталось преобразовать вычисления в аналитический признак. Для этого надо добавить еще один *Калькулятор* после *Группировки*, а в нем посчитаем 2 поля: ***% редких*** — как количество редких деленное на общее количество, и ***Класс редкости*** — текстовое описание.

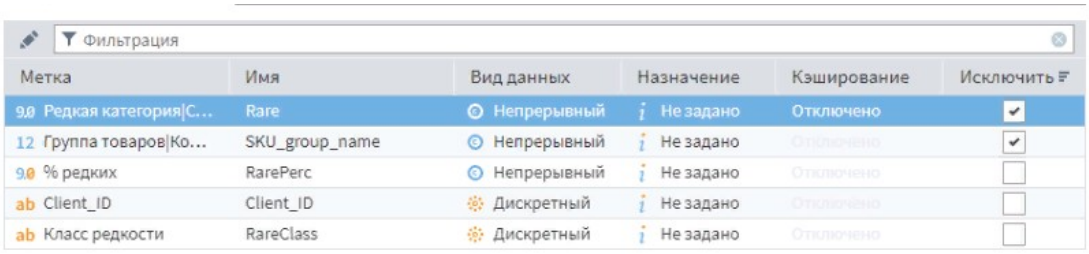


Текстовое описание редкости формируется следующим образом:

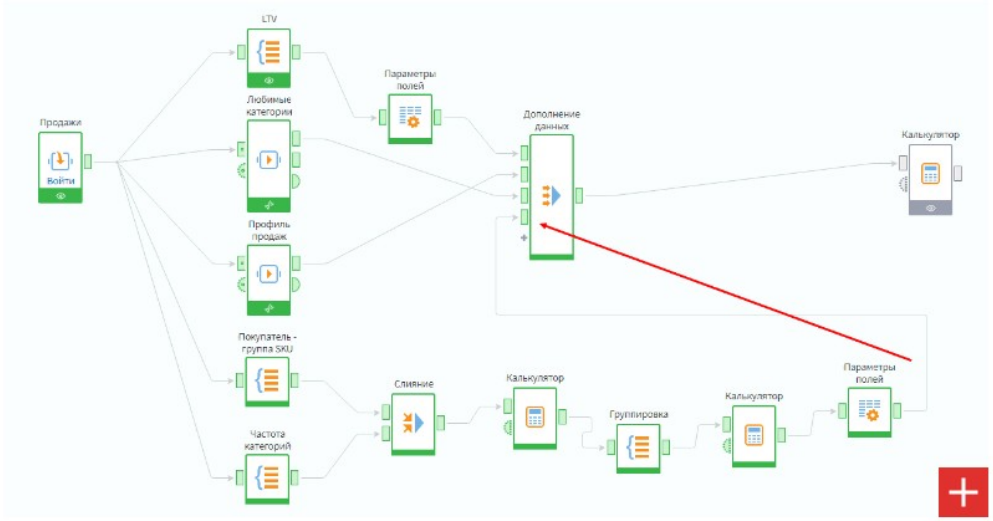
if(RarePerc<=0.1,"1. Стандартный", if(RarePerc<=0.5,"2. Энтузиаст","3. Любитель экзотики"))



Следующим шагом стоит избавиться от вспомогательных полей, которые далее не требуются. Для этого можно добавить после *Калькулятора* узел ***Параметры полей*** и исключить поля ***Rare*** и ***SKU_group_name***.



Получившуюся таблицу надо связать с узлом *Дополнение данных*. Таким образом в результирующей таблице появится еще один аналитический атрибут.

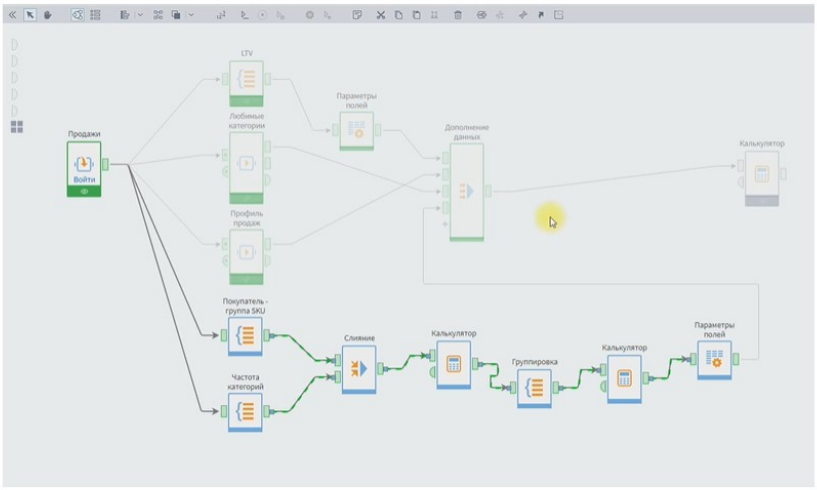


При настройке присоединяемой таблицы нужно указать, что связывание происходит по полю *Client_ID*.

Дополнение данных

№	Главная таблица	таблица		Присоединяемая таблица 2		Присоединяемая таблица 3
1	ab Client_ID	о клиента	✔	ab Идентификатор клиента	✔	ab Client_ID
2	ab Покупатель			Не выбрано		Не выбрано
3	90 LTV выручка			Не выбрано		Не выбрано
4	90 LTV валовая при...			Не выбрано		Не выбрано
5	вая покупка			Не выбрано		Не выбрано
6	31 Последняя поку...			Не выбрано		Не выбрано

Полученную цепочку действий рекомендуется свернуть в подмодель, чтобы схема была более читаемой.



При добавлении нового атрибута ***Класс редкости*** в отчет *Портрет клиента* выявляются интересные детали.

У клиентов всех сегментов имеются покупки позиций разных уровней редкости. Для редких товаров характерен самый высокий средний чек, но более или менее существенную долю прибыли в 5.25% любители экзотики дают только для VIP-класса. Во всех остальных случаях эта доля не превышает 2%.

Учитывая, что количество VIP-клиентов невелико, стоит оценить, какую реальную нагрузку они оказывают на отдел закупок, и как это соотносится с общей эффективностью.

Тип клиента | Факты | Уровень клие... | Класс редкости

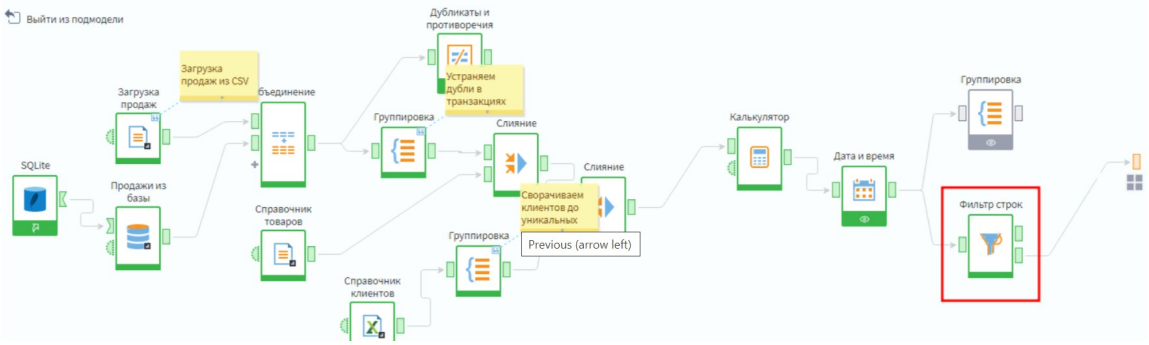
		Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
		∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
		∑ Значе...	%∑ Процент п...	∑ Значе...	%∑ Проц...	∑ Значе...	∑ Значение	∑ Значение
1. VIP	1. Стандартный	3 279 774	5,71%	1 383 761,31	5,40%	154	21 297,23	8 985,46
	2. Энтузиаст	6 849 982	11,93%	3 011 843,48	11,76%	426	16 079,77	7 070,06
	3. Любитель экзотики	3 431 499	→ 5,97%	1 344 581,05	5,25%	9	381 277,67 ←	149 397,89
	Итого:	13 561 255	23,61%	5 740 185,84	22,42%	589	23 024,20	9 745,65
2. Важный	1. Стандартный	5 796 084	10,09%	2 577 911,87	10,07%	588	9 857,29	4 384,20
	2. Энтузиаст	6 185 114	10,77%	2 841 707,27	11,10%	716	8 638,43	3 968,86
	3. Любитель экзотики	865 066	→ 1,51%	374 622,78	1,46%	9	96 118,44 ←	41 624,75
	Итого:	12 846 264	22,37%	5 794 241,92	22,63%	1 313	9 783,90	4 412,98
3. Перспект...	1. Стандартный	11 624 874	20,24%	5 300 689,51	20,70%	2 021	5 752,04	2 622,81
	2. Энтузиаст	5 884 598	10,25%	2 654 970,67	10,37%	822	7 158,88	3 229,89
	3. Любитель экзотики	638 237	→ 1,11%	311 774,08	1,22%	14	45 588,36 ←	22 269,58
	Итого:	18 147 709	31,60%	8 267 434,26	32,29%	2 857	6 352,02	2 893,75
4. Начинаю...	1. Стандартный	9 147 141	15,93%	4 123 400,40	16,10%	2 568	3 561,97	1 605,69
	2. Энтузиаст	3 545 070	6,17%	1 599 855,15	6,25%	798	4 442,44	2 004,83
	3. Любитель экзотики	186 396	→ 0,32%	80 114,23	0,31%	32	5 824,88 ←	2 503,57
	Итого:	12 878 607	22,42%	5 803 369,78	22,66%	3 398	3 790,06	
Итого:		57 433 835	100,00%	25 605 231,80	100,00%	8 157	7 041,05	

Теперь есть возможность оценить характеристики только тех клиентов, которые закупают самый ходовой ассортимент, и понять, чем они отличаются от любителей экзотики.

Задание.

Задание №1. С помощью узла ***Фильтр***, исключите из транзакций значения, в которых сумма продажи равна 0 или в поле *Покупатель* содержится фраза *0_Администратор* или поле Покупатель содержит *Частное лицо*.

Рекомендуемое расположение фильтра — перед выходом из подмодели ***Продажи***.

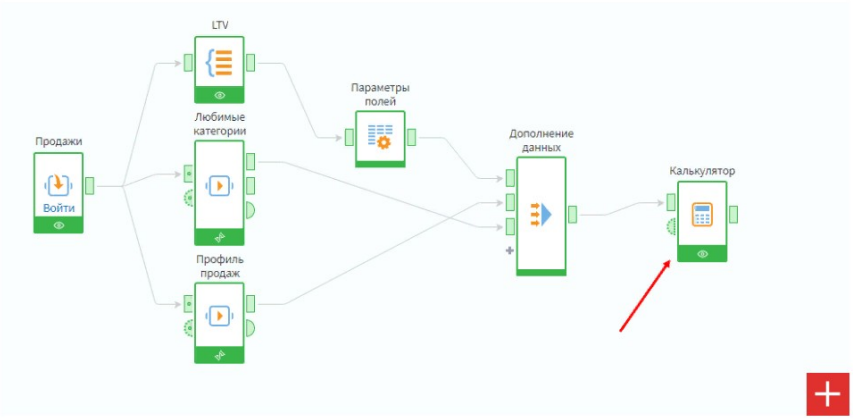


Обратите внимание, что используется не первый выход фильтра, а второй, куда попали записи, не удовлетворяющие условию.

Чаще всего проще прописать условие в утвердительном варианте *Найди строки где сумма = 0*, а покупатель содержит *0_Администратор* или *Частное лицо*, а на выходе из узла взять те строки, которые **не удовлетворяют** этим условиям.

Должно получиться 98 893 строки. Количество записей сократилось с 2.5 миллионов более, чем в 25 раз.

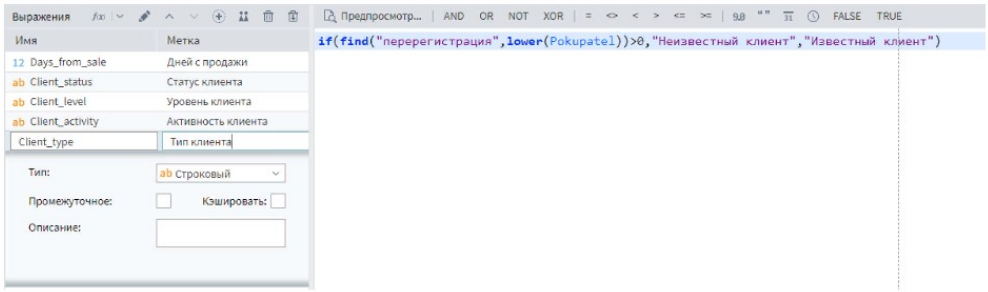
Задание №2. Введите дополнительный аналитический признак для фильтрации в отчете клиентов, которым «необходима перерегистрация». Это будет удобно сделать в узле *Калькулятор*, в котором можно сформировать дополнительные признаки клиентской базы.



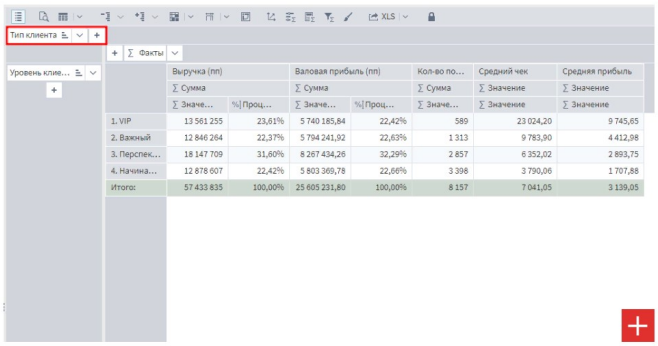
Дополнительный признак клиента

Для замены понадобится проверить, есть ли в названии клиента подстрока *Перерегистрация*. Это можно сделать функцией ***Find***. Она возвращает номер символа строки, с которого начинается искомое слово. Таким образом, если *Find* выдаст значение больше 0 — значит слово было найдено.

Обратите внимание, что поле ***Pokupatel*** обрабатывается функцией ***Lower***, чтобы проверяемая строка содержала только строчные буквы.



Полученное поле нужно добавить как фильтр в визуализаторы внутри *Калькулятора*. Теперь можно оценить, какой процент продаж приходится на подобных клиентов.



Куб с показателями продаж

Что в итоге

В конце можно оценить, как повлияла работа с фиктивными данными на итоговые отчеты.

Отчет ***Клиентская матрица*** затронут несильно, т.к. в основном рассчитывает количество покупателей. Поэтому изменения небольшие: исчез клиент *Частное лицо* и пара внутренних подразделений.

А вот в *Портрете клиента* все намного интереснее. Вот что было до работы с фиктивными данными.

	Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%∣Проц...	∑ Значе...	%∣Проц...	∑ Значе...	∑ Значение	∑ Значение
1. VIP	196 244 612	81,73%	88 039 147,24	81,59%	1 429	137 330,03	61 608,92
2. Важный	12 846 264	5,35%	5 794 241,92	5,37%	1 313	9 783,90	4 412,98
3. Перспек...	18 147 709	7,56%	8 267 434,26	7,66%	2 858	6 349,79	2 892,73
4. Начина...	12 878 607	5,36%	5 803 369,78	5,38%	3 398	3 790,06	1
Итого:	240 117 192	100,00%	107 904 19...	100,00%	8 998	26 685,62	11

До исключения фейковых данных

А вот что стало после.

	Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%∣Проц...	∑ Значе...	%∣Проц...	∑ Значе...	∑ Значение	∑ Значение
1. VIP	13 561 255	23,61%	5 740 185,84	22,42%	589	23 024,20	9 745,65
2. Важный	12 846 264	22,37%	5 794 241,92	22,63%	1 313	9 783,90	4 412,98
3. Перспек...	18 147 709	31,60%	8 267 434,26	32,29%	2 857	6 352,02	2 893,75
4. Начина...	12 878 607	22,42%	5 803 369,78	22,66%	3 398	3 790,06	1
Итого:	57 433 835	100,00%	25 605 231,80	100,00%	8 157	7 041,05	3

После исключения фейковых данных

Оказывается, VIP-клиенты приносят не 81.73% выручки, а 23.61%. И средний чек у них не 137 330, а 23 024 рубля. Теперь финансовая значимость этого сегмента выглядит иначе, поэтому приоритеты работы могут серьезно измениться.

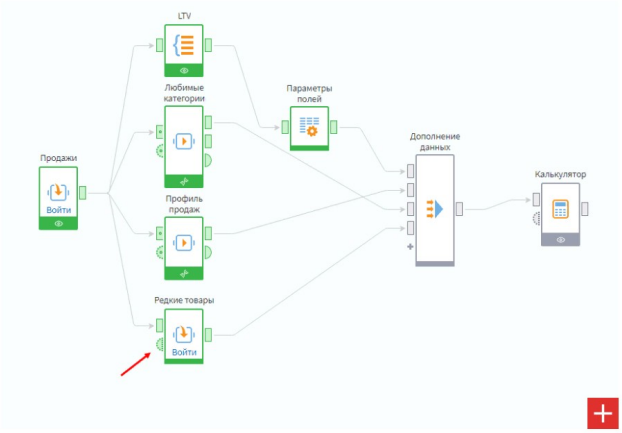
Кейс с разделением продаж на опт и розницу простой, но это хороший повод задать вопрос: достаточно ли одноуровневого разбиения на 3 сегмента? Может быть стоит разбить на подсегменты, требующие отдельного изучения?

Оптовые продажи могут включать в себя совершенно разные категории клиентов. Например, государственные закупки, ритейловые сети, малый опт и т.д. Скорее всего объединение их в одну группу — не лучшая идея.

Не всегда эти разделения доступны в виде готовых справочных данных. Только опыт экспертов позволит все разложить по полочкам. А если своих знаний особенностей бизнеса недостаточно, лучше всего обратиться к опытному сотруднику и оцифровать его экспертизу в Loginom.

Задание для техно-энтузиастов. Создайте на подмодели *Редкие товары* входной порт переменных. Создайте в нем переменную типа целое число, назвав ее *Порог редкости*.

Передайте переменную на вход калькулятора, в котором проставляется признак 1 или 0 в зависимости от редкости категории. Измените формулу так, чтобы вместо константного значения 50 шло сравнение с переменной.



Теперь можно настраивать уровень определения редкости, не заходя в подмодель.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1.Опишите операторы репродукции и кроссинговера в простом генетическом алгоритме. Приведите примеры.

2.Фундаментальная теорема генетического алгоритма.

3.Приведите пример применения фундаментальной теоремы генетического алгоритма.

Лабораторная работа № 8.
Тема: Пропуски.

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

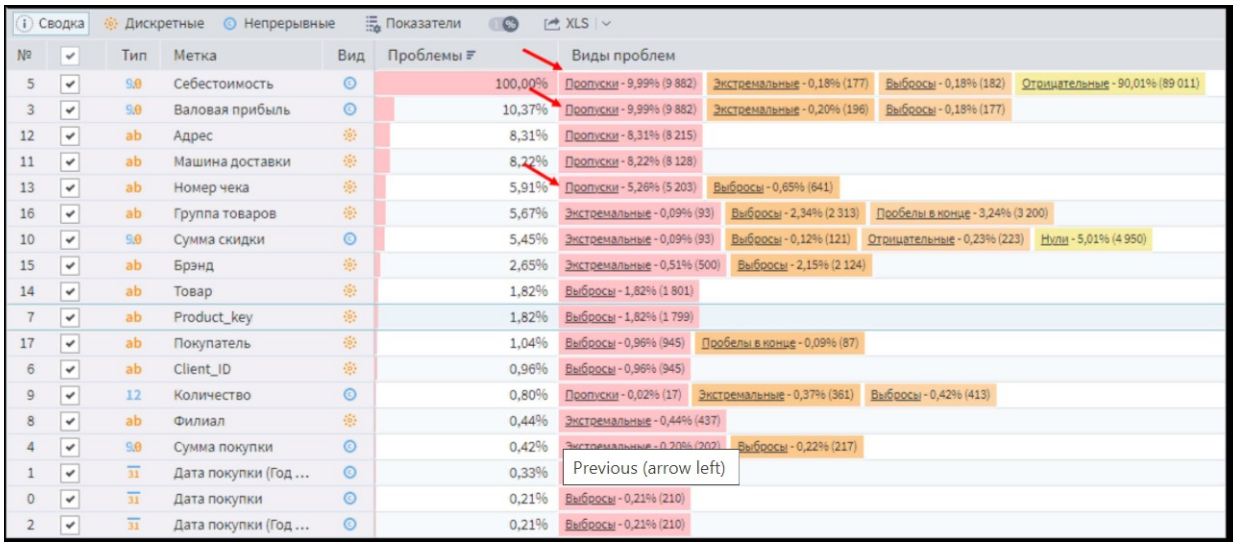
Теоретическая часть

В предыдущие дни проблемы в данных обычно заключались в том, что их больше, чем нужно: дубли или фейковые данные избыточны для анализа.

Сегодня будет рассмотрена обратная проблема — пропуски, т.е. отсутствие данных. Их можно разделить на 2 вида:

1. Отсутствие значений в **полях**;
2. Пропуски **строк** в таблицах.

Первый тип проблем легко обнаружить с помощью визуализатора ***Качество данны***. Если просмотреть данные по клиентским транзакция, то наличие пропусков сразу бросится в глаза.



Реакция на проблему зависит от целей анализа. Иногда пропуски можно игнорировать. Например, на отсутствие значений в полях адреса и машин доставки можно не обращать внимание, если не интересуют вопросы транспортировки.

А вот пропуски в себестоимости игнорировать нельзя. Особенно если учесть, что анализируется *Валовая прибыль*, являющаяся суммой между между *Суммой продаж* и *Себестоимостью*. Себестоимость задается отрицательной.

Если в арифметических расчетах имеется null-значение, то результат тоже будет пустым. Как следствие для некоторых записей валовая прибыль отсутствует, и данный показатель по всей базе получится заниженным.

Также стоит обратить внимание на пропуски в поле *Номер чека*. Оно используется для расчета количества транзакций, что, в свою очередь, учитывается при вычислении среднего чека. Таким образом, получается заниженное количество транзакций и завышенное значение среднего чека.

Возникновение пропусков в полях

Причин обычно две: технические проблемы или брак бизнес-процесса.

Технические причины

К возникновению пропусков в номере чека привела типичная техническая проблема. В компании используется контрольно-кассовое оборудование разных производителей. Данные с них попадают в единую базу. Часть борудования формировала номер чека как число, а часть – как микс числа и текстовых символов.

При этом в консолидированной базе под хранение номера чека отведено числовое поле. В результате часть значений, которые удалось конвертировать в число, попали в базу, а часть — стала пустым (null) значением.

Ошибки бизнес-процесса

Еще одной частой причиной пропусков является человеческий фактор, усугубленный сложным бизнес-процессом.

Например, сотрудник, принимающий товар от поставщиков, должен прописывать его закупочную стоимость в учетной системе, но регулярно этого не делает. Причины могут быть разные: не успевает или забывает, лень, нет корректных данных на момент ввода, анализ себестоимости вне его зоны ответственности...

Поэтому контроль качества вводимых данных лучше всего автоматизировать. В частности, с забывчивостью можно бороться настройкой обязательных полей в учетных системах. Правда, это не спасает от фейковых данных, когда в обязательное поле *Телефон* вписывают *+123456789*.

Человеческий фактор нельзя игнорировать. Его влияние слишком велико. Значит, все, что может собираться и проверяться автоматически, должно быть автоматизировано:

- Контактные данные нужно перебрасывать из форм на сайтах.

- Обязательные поля должны контролироваться в момент ввода в учетной системе.
- Вносимые данные желательно оценивать на адекватность, например, выход за границы диапазонов.
- Поля типа телефонов или адресов необходимо нормализовать.

Далеко не всегда есть возможность исправить проблемы задним числом, например, заполнить пропущенные значения. Борьбу за качество данных рекомендуется начинать на этапе их ввода.

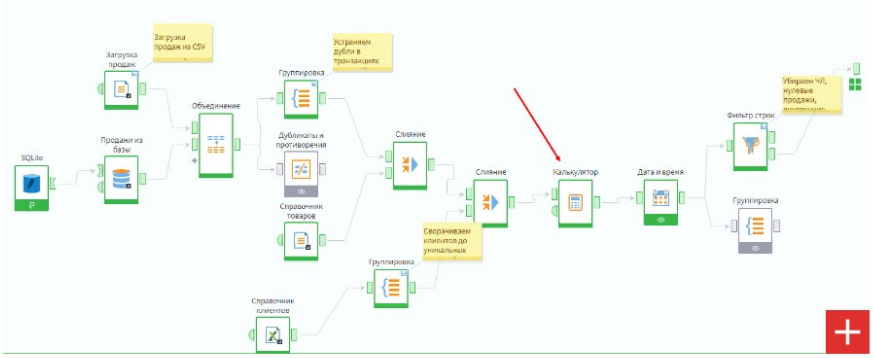
Заполнение пропусков

Восстановление идентификаторов

Вначале продумаем, как заполнить пропуски в номерах чеков. В идеале хотелось бы восстановить реальную информацию из систем учета, но это не всегда возможно. Поэтому попробуем заполнить эти пропуски при помощи сценариев в Loginom.

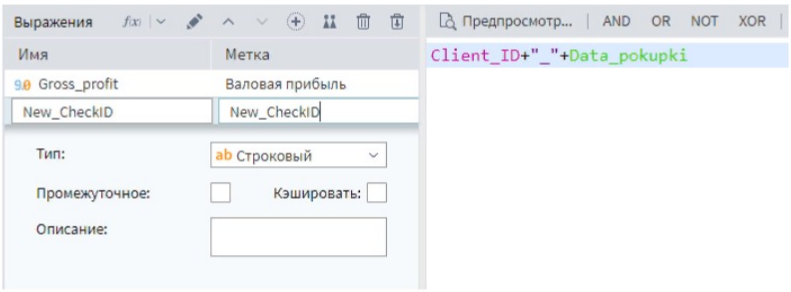
Так как анализируются продажи оптовым клиентам, то можно отталкиваться от предположения, что каждому контрагенту в день осуществляется не более одной отгрузки. Если отгрузок несколько, то на практике довольно часто их объединяют в одну.

Следовательно, пустой идентификатор чека можно заменить на комбинацию полей ***Client_ID*** и ***Даты продажи***. Это можно сделать, используя компонент ***Калькулятор*** в подмодели *Продажи*.



Необходимо добавить текстовое поле *New_CheckID*. В качестве формулы задать объединение двух полей через разделитель _.

Составные поля рекомендуется соединять через разделитель. Это не только повышает читаемость значения, но и позволяет избежать коллизий, например, когда соединение ***11*** и ***1***, дает такой же результат, как соединение ***1*** и ***11***.

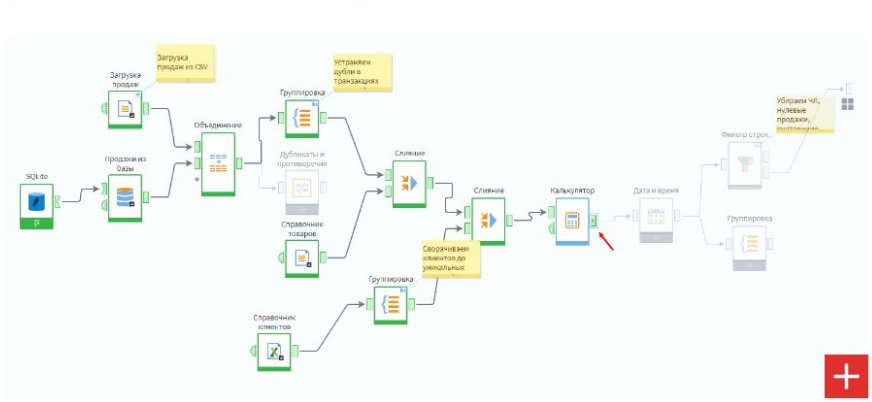


Данное действие сформировало новое поле, а не заменило старое. Не хотелось бы теперь искать места по всему сценарию, где нужно внести исправления, чтобы все работало как надо.

Чтобы подменить значение одного поля другим, требуется настроить соответствующую связь на выходе из узла. Порядок действий следующий:

1. Зайти в настройки выходного порта.
2. Переключиться в режим отображения связи.
3. Удалить в правой части выходное поле *New_CheckID*, т.к. оно не нужно в дальнейших расчетах.
4. Удалить связь между полями *Номер чека* в калькуляторе и *Номер чека* на выходе из порта.
5. Протянуть связь от *New_CheckID* из *Калькулятора* в поле *Номер чека* на выходе из порта.

После этих действий на выходном порту узла появится точка. Это значит, что в узле отключена автосинхронизация, и теперь перечень полей, выходящих из *Калькулятора*, управляется вручную. Так, если во входной таблице появится новое поле, то чтобы оно появилось на выходе, нужно будет его туда самостоятельно добавить.



Отключенная автосинхронизация

Восстановление себестоимости

Следующий вопрос — заполнение пропусков по полю *Себестоимость*.

Самое простое решение, приходящее в голову, — отфильтровать все записи с пустым значением данного поля. Проблема в том, что тогда исключится слишком много строк, что повлияет на другие показатели.

Если при анализе реальных данных отбрасывать все строки с пустыми значениями в любом поле, то скорее всего анализировать будет нечего. Пропуски в данных встречаются слишком часто.

Следовательно, надо каким-то образом «восстановить» цену закупки. Конечно, для этого придется принять некоторые разумные допущения. Они не смогут гарантировать 100% корректность, но позволят получить адекватный портрет клиента, чего в данном случае достаточно.

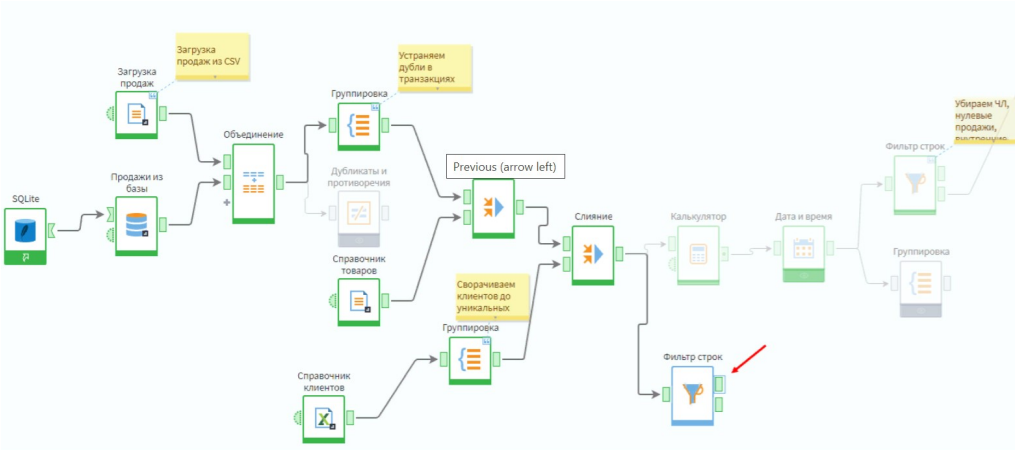
Если оставить как есть, то в отчетах будут заниженные данные по валовой прибыли и неверное понимание ценности клиентов. Так как пропуски есть примерно в 10% транзакций, то расчет себестоимости, основанный на разумных допущениях, – меньшее зло по сравнению с отсутствием у десятой части продаж данных по прибыли.

Подобные действия надо документировать, чтобы пользователи отчетов осознавали, что имеют дело с не совсем точной информацией.

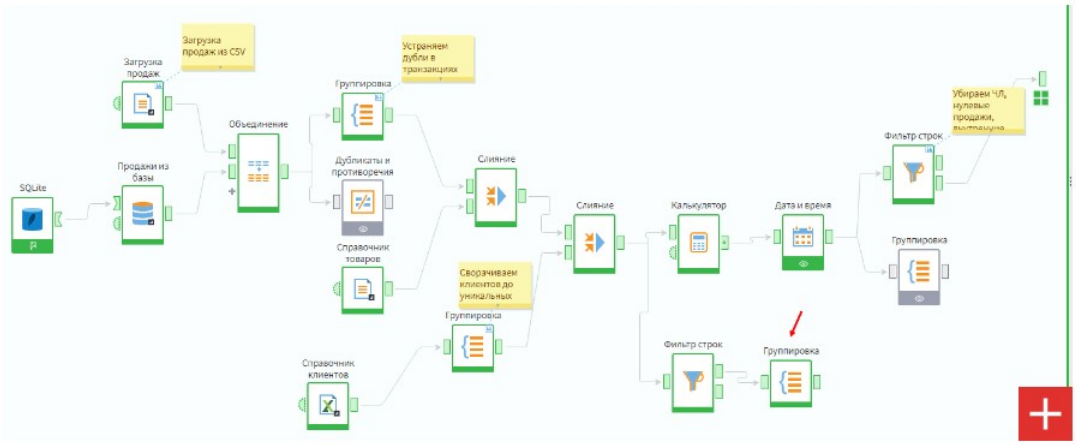
«Восстановить» себестоимость можно с помощью статистики. Например, отталкиваться от средней наценки на все позиции или конкретную товарную группу. Можно построить и более сложные модели, но для экономии времени предлагается реализовать следующий сценарий: рассчитать медианный процент прибыли для всех товаров и использовать его, когда реальная себестоимость не известна.

Расчет медианной наценки

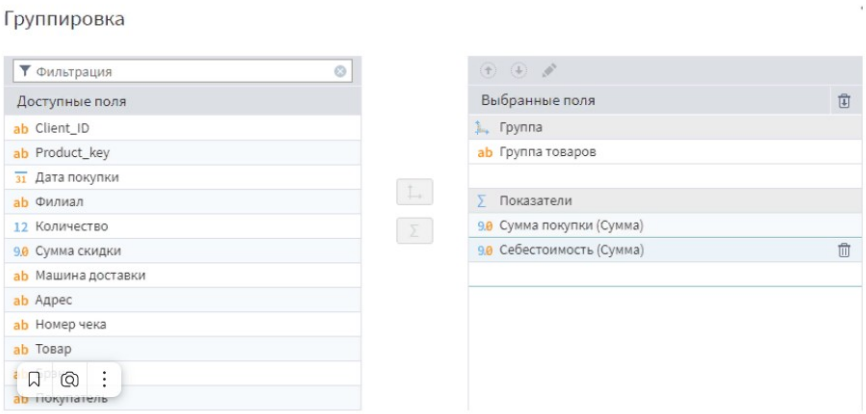
Надо в подмодели *Продажи* добавить фильтр по условию ***Поле себестоимость — не пустое***.



Далее сгруппировать данные строк, **где себестоимость не пустая**, по товарным группам, просуммировав поля *Себестоимость* и *Сумма покупки*.



Группировка по товарным категориям



Следующий шаг — добавить еще один калькулятор, в котором будет рассчитано поле *Profit_perc* по формуле ***(Summa_pokupki+Sebestoimost)/Summa_pokupki***.

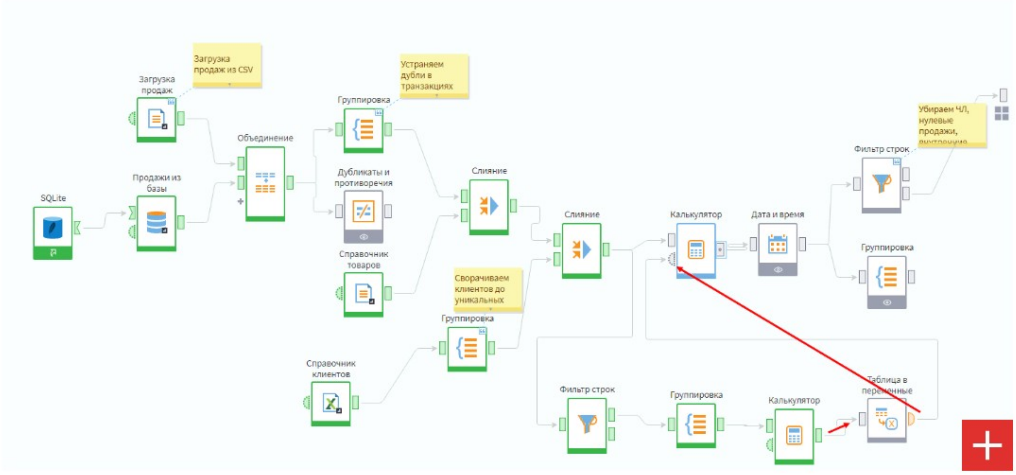
Получится таблица с процентом прибыли по каждой категории товаров.

#	Группа товаров	Сумма покупки|Сум…	Себестоимость|Сумма	% прибы…
176	ДПИ-ВС-основы	28 694,43	-13 752,40	0,52
177	ДПИ-Контуры-Наборы	178 249,39	-89 050,06	0,50
178	ДПИ-Направления-Керамика-Наборы	281 033,66	-144 137,28	0,49
179	Основы-холст на подрамнике-Базовый	12 090 560,15	-6 007 229,79	0,50
180	ДПИ-Направления-Каллиграфия-тушь, чернила	149 101,49	-74 465,47	0,50
181	Литература-Практические руководства по живописи и рисунку-…	1 463 079,04	-730 987,98	0,50
182	Графические материалы-черно-графитовые-механические/цанг…	1 311 820,01	-655 342,54	
464	Сумки-Чехлы для подрамников	1 041 473,67	-528 818,54	

Процент наценки по группам товаров

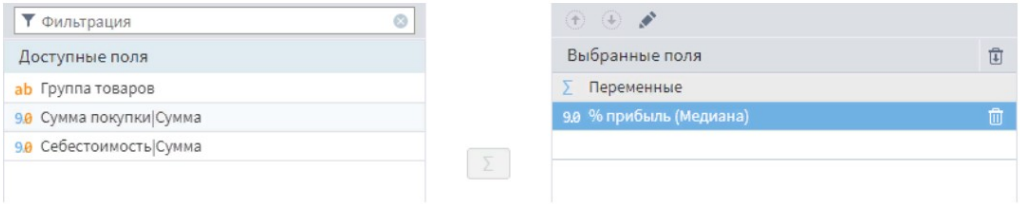
Пометка. Значение последнего столбца ***% прибыли*** *по факту является долей.*

Далее нужно рассчитать медианный процент наценки по группам. Для этого можно использовать обработчик ***Таблица в переменные***, передав полученный результат на вход *Калькулятора*.

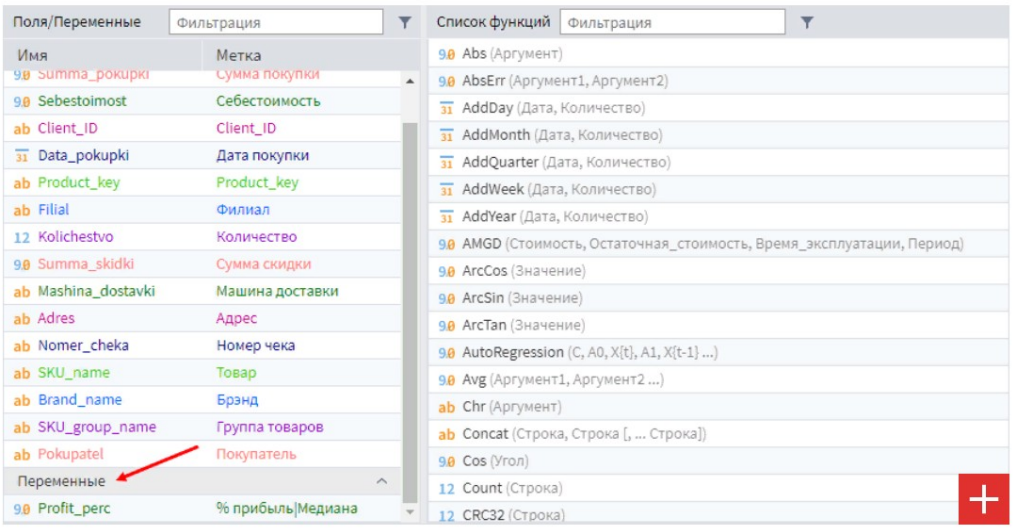


Расчет медианной наценки по группам товаров

Логика этого узла похожа на обработчик *Группировк*, но без возможностей указания групп. Из каждого поля рассчитывается одно или несколько агрегированных значений, которые передаются дальше по сценарию как переменные. Нужно указать агрегацию *медиана* для поля *% прибыли*.

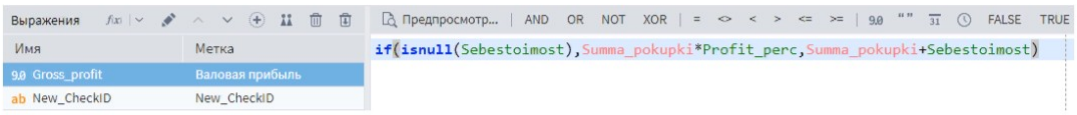


Теперь медианный процент прибыли доступен как переменная в *Калькуляторе*.

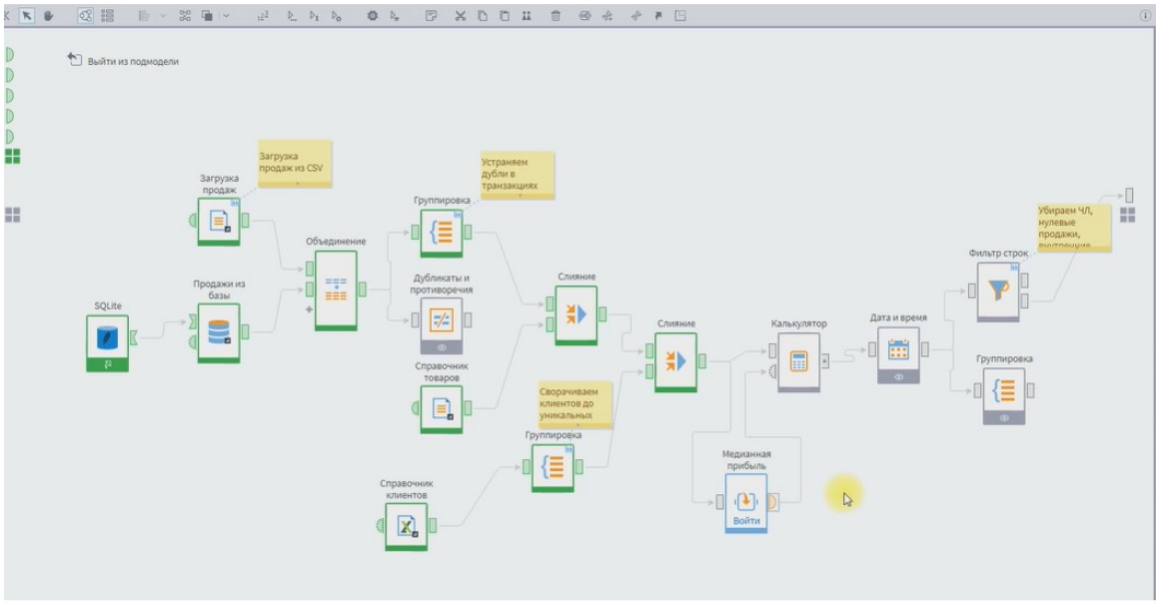


Расчитанная наценка в калькуляторе

Далее нужно модифицировать формулу для поля ***Gross_profit*** так, чтобы в случае пустого поля ***Sebestoimost*** прибыль считалась как выручка, умноженная на медианный процент прибыли. В ином случае – как сумма выручки и себестоимости (заданной отрицательным числом).



Для того, чтобы сценарий стал более компактным, нужно свернуть узлы, в которых рассчитывалась медианная наценка в подмодель. Для этого надо выделить их по очереди, прокликав ЛКМ с зажатым Ctrl, а затем выбрать на панели инструментов действие *Свернуть в подмодель*.



Задание.

Бонусное задание. Создайте дополнительное поле расчета себестоимости, в котором при пустом значении поля *Sebestoimost* будет происходить вычисление себестоимости через вычитание из

валовой прибыли суммы продажи (в исходных данных себестоимость идет со знаком минус по стандартам финансовой отчетности).

Последний шаг — настройка выхода из узла, чтобы новое поле себестоимости передавалось в старое по аналогии с номером чека.

мониторинга корректности загрузок: анализ логов, контроль завершения задач, сравнение данных в первоисточнике и хранилище данных и прочее.

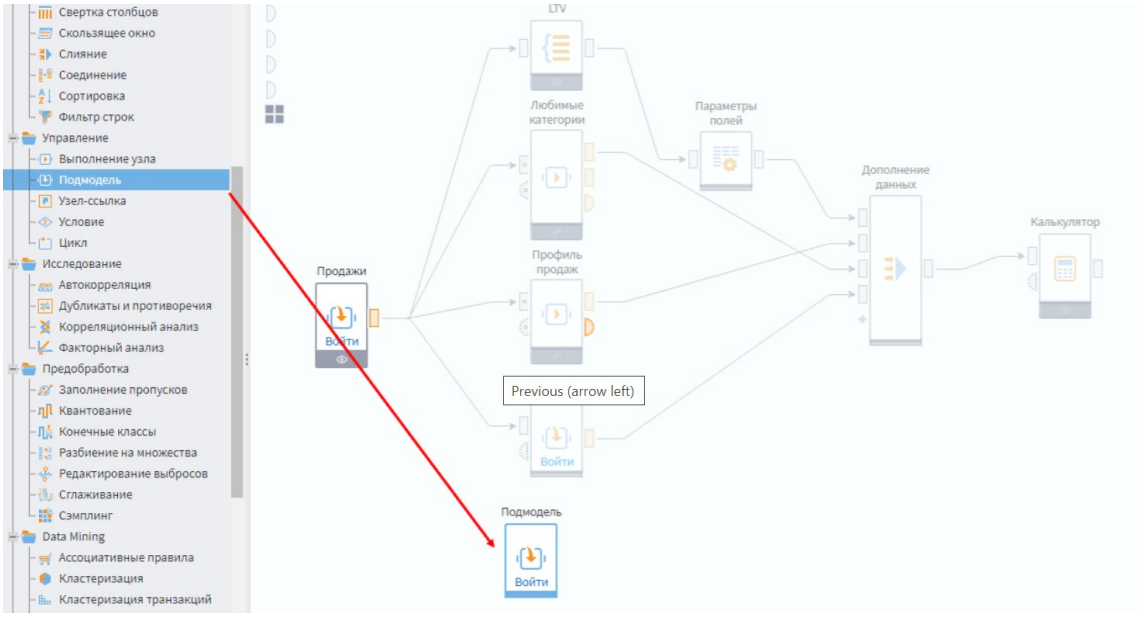
Одна из самых распространенных проблем такого рода — не прогруженные записи. Чаще всего продажи осуществляются каждый рабочий день, и отсутствие хотя бы одной отгрузки в какой-то день — надежный индикатор проблем, возникших в ETL-процессе. В Loginom нет готового обработчика для выявления таких проблем, но его можно реализовать средствами платформы.

Создание переиспользуемой подмодели

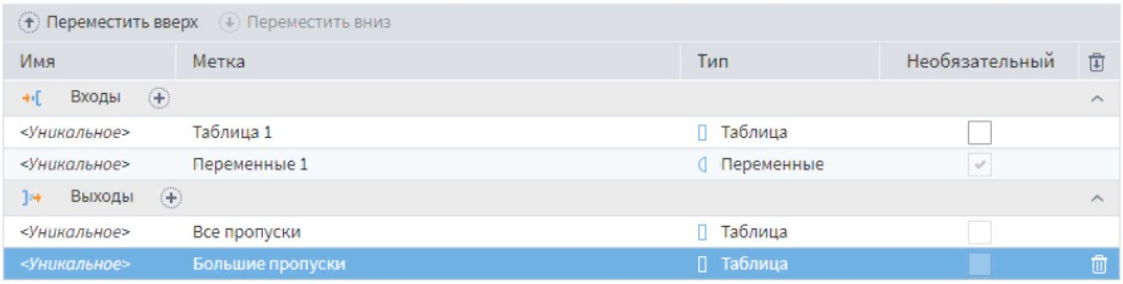
До этого момента подмодели использовались для оптимизации рабочего пространства за счет сворачивания больших фрагментов сценария. Но сейчас будет спроектирована подмодель, которая может переиспользоваться как готовый компонент в других сценариях.

Подмодель будет принимать на вход поле с датой, а на выход возвращать список пропущенных периодов. При этом особо длительные пропущенные периоды будут выводиться отдельно.

Добавьте компонент ***Подмодель*** в область сценария, где создается клиентский портрет, и зайдите в настройки узла.

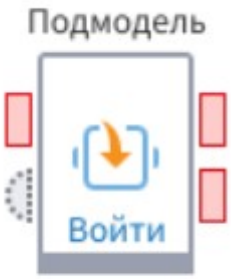


Нужно создать на входе 1 табличный порт и 1 порт переменных, а на выходе — 2 табличных порта.



При создании портов, особенно если их несколько, надо давать им понятные названия. Иначе в будущем такую подмодель будет сложно поддерживать самому и переиспользовать коллегам. Название порта помогает разобраться, какие данные туда подаются.

Результат:

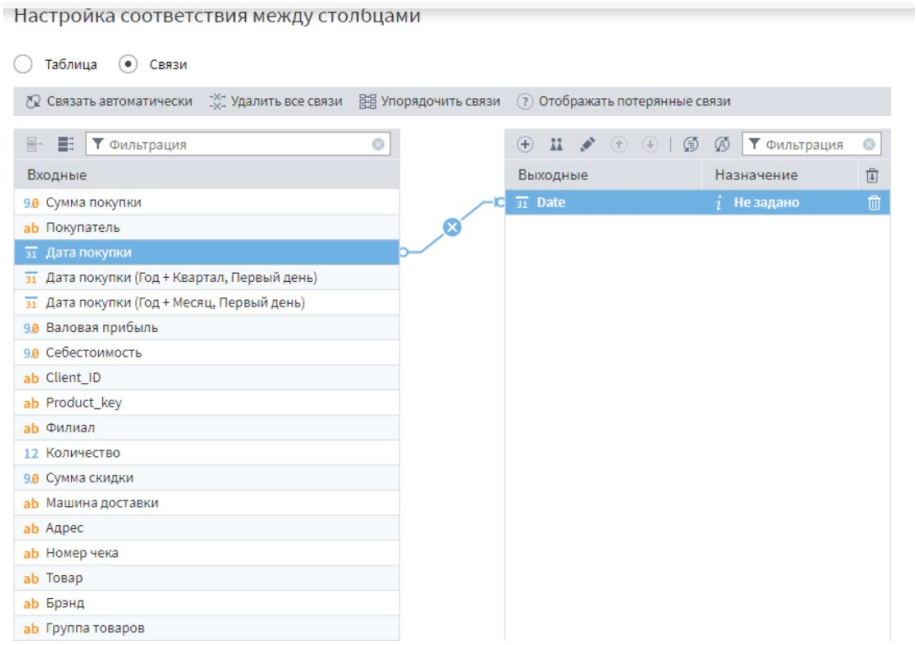


Т.к. подмодель рассчитана на решение конкретной задачи (работа с данными определенного типа), стоит задать перечень полей и переменных на входных портах.

В табличном порту нужно отключить автосинхронизацию и создать поле ***Date*** с типом *Дата*.

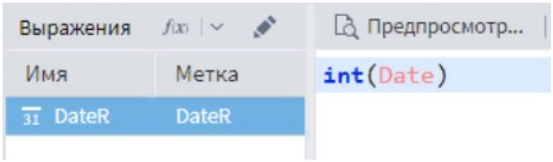
В порту переменных надо отключить автосинхронизацию и создать переменную с типом целое число и значением по умолчанию — 10.

На вход подмодели надо подать данные из узла *Продажи*, а в настройках входного порта данных задать соответствие между полем ***Дата покупки*** и ***Date***, переключив отображение в режим *Связи*.

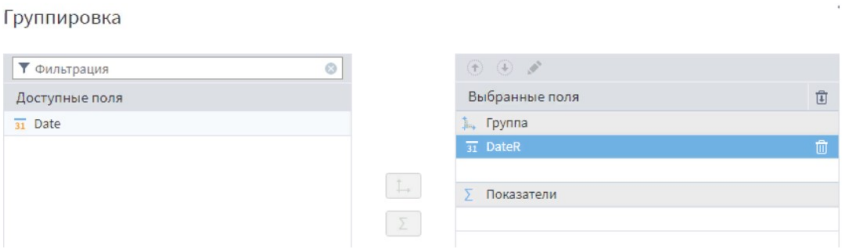


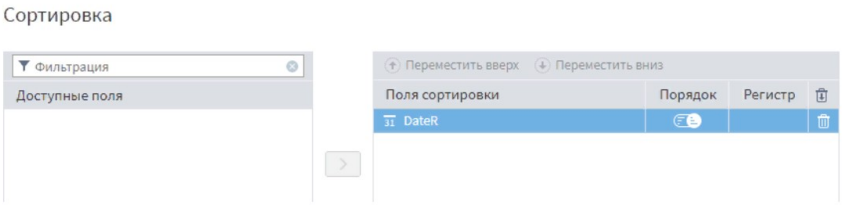
В подмодели определены входы, теперь надо построить сценарий от входа до выхода. Для этого нужно сформировать набор уникальных дат и отсортировать их по возрастанию.

Т.к. заранее неизвестно, содержатся в поле только даты или это поле с датой и временем, надо с помощью функции ***int()*** округлить значения. В полях этого типа целая часть числа отвечает за дату, а дробная — за время.

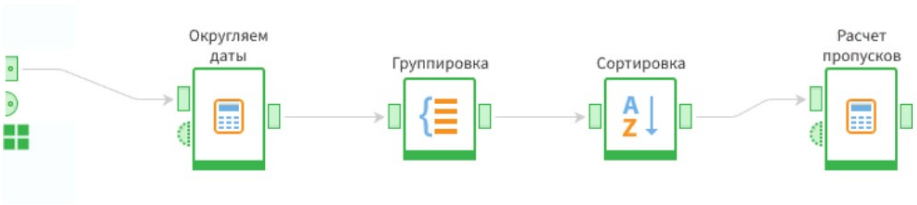


Дальше при помощи *Группировки* нужно сформировать перечень уникальных дат и отсортировать по возрастанию.





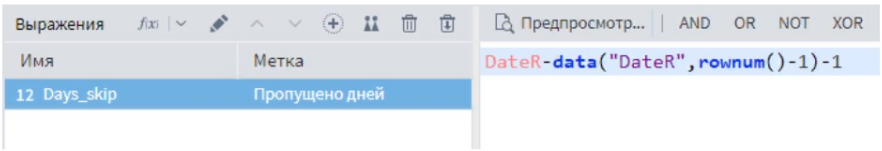
Потом рассчитать, есть ли пропуски. Для этого на вход последнего узла подать ранее отсортированную таблицу.



Следующий шаг — рассчитать, сколько дней между соседними строками. Для этого используется функция ***Data()***.

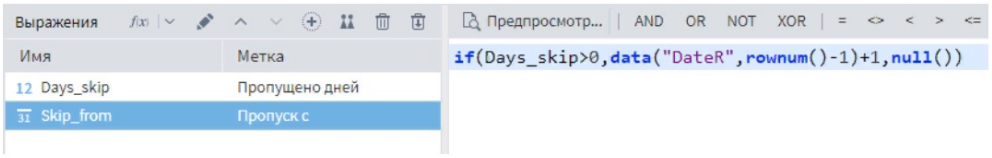
Первым аргументом (в двойных кавычках) идет название поля, из которого берутся данные. Вторым аргументом — номер строки. Функция ***RowNum()*** возвращает номер текущей строки, следовательно, вычитание единицы дает номер предыдущей строки.

Дополнительно нужно вычесть единицу, т.к. между текущим и следующим днем всегда разница в 1 день, а требуется найти дни между которыми разница больше, чем один пропущенный день.

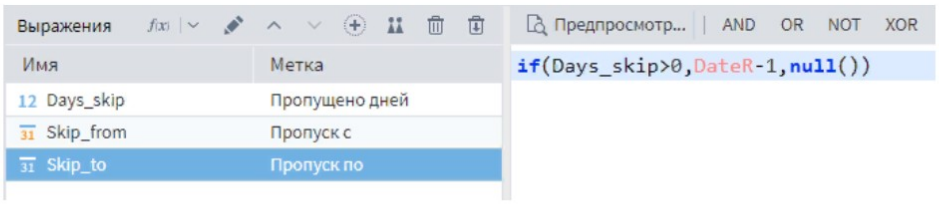


Далее надо добавить пару полей, которые будут показывать нам пропущенные периоды.

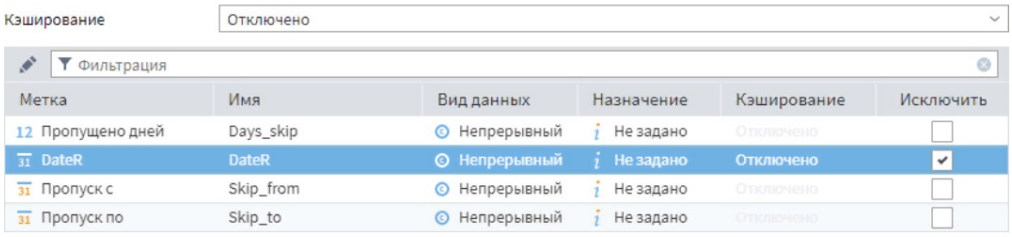
- Поле *Пропуск с* считается по формуле ***если пропущено дней больше нуля, то берем предыдущую дату +1 день. Иначе — пустое значение***.



- Поле *Пропуск по* вычисляется по правилу ***если пропущено дней больше нуля, тогда берем текущую дату минус 1. Иначе — пустое значение***.



При помощи обработчика *Параметры полей* можно убрать поле *DateR*, т.к. на выходе интересуют только пропущенные интервалы.



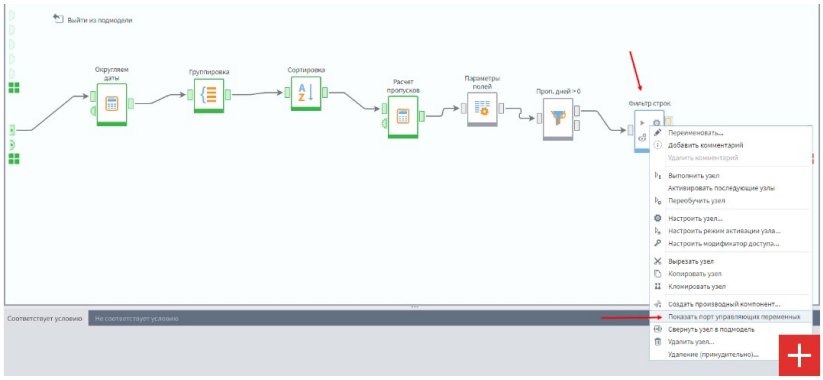
После исключения поля *DateR* нужно добавить узел *Фильтрации* с условием ***Пропущено дней > 0***.

Фильтрация записей

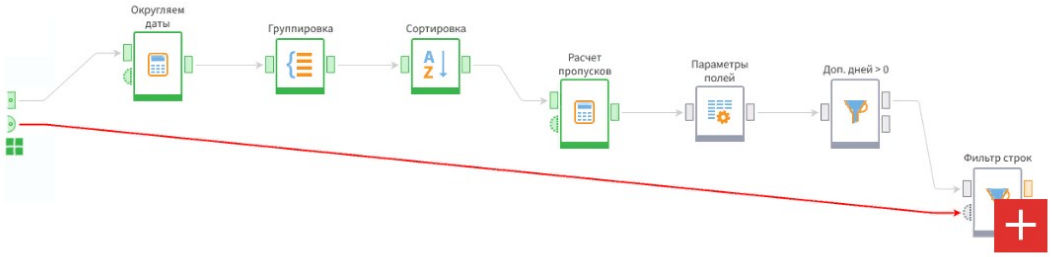
Далее надо добавить еще один *Фильтр*, принимающий данные с узла *Параметры полей* и показать в данном узле порт переменных, который по умолчанию скрыт. Для этого надо кликнуть по

последнему узлу фильтрации ПКМ и в меню выбрать пункт ***Отобразить порт управляющих переменных***.

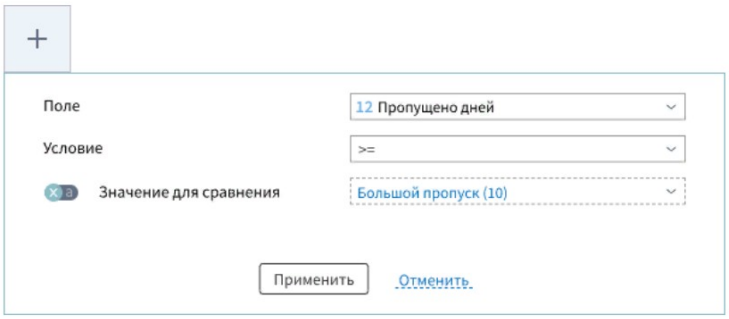
Некоторые узлы имеют скрытый порт переменных, который при необходимости можно отобразить описанным выше образом.



На этот порт переменных нужно подать данные с порта переменных подмодели.

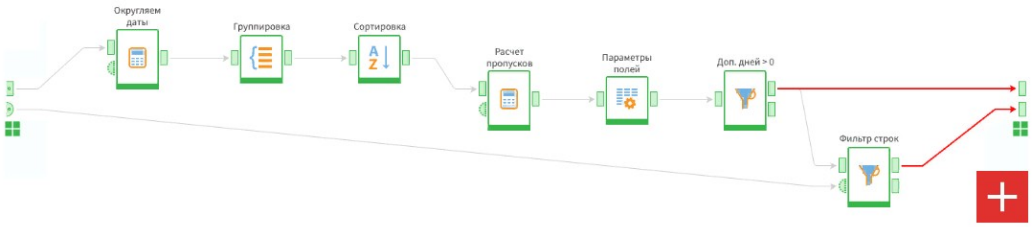


Теперь при открытии мастера настройки фильтра надо активировать соответствующий переключатель напротив параметра и выбрать нужную переменную.



После постановки условия фильтрации, будут отбираться записи, в которых длина пропуска больше или равна значению из переменной.

Финальный этап — отправить выход с первого фильтра на первый входной порт подмодели, а со второго фильтра — на второй выходной порт. Теперь в них будет выводиться полный список пропусков и отдельно большие пропуски. Критерий большого пропуска определяется через переменную на входе в подмодель.



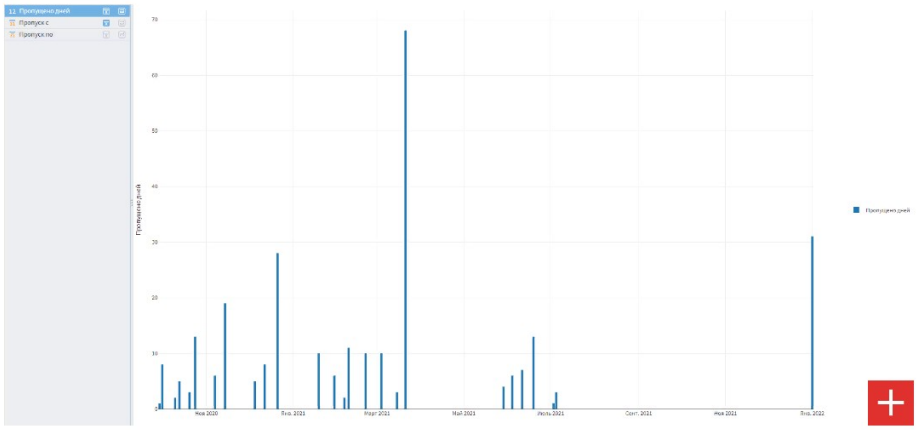
Вывод данных из подмодели

Таким образом был создан свой первый компонент, пригодный для переиспользования. Как его использовать, можно посмотреть в других пакетах и в инструкции по производным компонентам после занятия.

Визуализация пропуска дат

Для визуализации полученной таблицы надо добавить в подмодель определения пропусков визуализатор *Диаграмма*.

На ось X требуется перетащить поле *Пропуск с*, а поле *Пропущено дней* разместить в центре визуализатора, а затем выбрать вариант отображения ***Столбчатая диаграмма***.



Визуализация пропущенных дней

По графику видно, что пропусков немало. В реальной жизни надо разбираться, в чем причина такого плохого качества. Вопросов может быть много:

- Правильно ли отработал импорт в Loginom?
- Правильно ли отработал экспорт из источника (например в файл)?
- В эти дни действительно не было продаж или пропуски возникли из-за технического сбоя?
- Можно ли восстановить потерянные данные?

Реальность обычно сурова: чем больше времени прошло с момента возникновения ошибки, тем сложнее ее исправить. Поэтому пропуски нужно мониторить на постоянной основе.

В нашем случае пропуски 2021 года восстановить не получится. Часть из них возникла из-за технических ошибок, а где-то продажи действительно отсутствовали.

Аудит пропусков наводит на мысль, что не стоит строить прогнозирующую и моделирующую аналитику на этих данных. Лучше считать клиентские портреты по данным за последний год.

Пропуски в относительно свежих данных чаще всего можно восстановить, но все равно нужно понять причину их возникновения, чтобы исправить брак. Чаще всего это ошибки в работе ETL-процесса. Поэтому предположим, что в новой базе ниже эта проблема устранена.

DBsales_transactions.db

Для исправления ситуации нужно заменить файл базы в папке ***Data*** на новый, и повторно импортировать данные. После этого надо еще раз просмотреть график пропусков в январе 2022.

Заключение

Работа с пропусками — деликатная тема. В зависимости от сценария анализа и критичности пропусков можно попытаться их восстановить, но иногда приходится оставлять как есть.

Независимо от выбранного подхода нужно проверять наличие пропусков и оповещать об этом коллег. Даже если данные так и остались с пропусками, наличие информации о том, что таковые в исходных таблицах имеются, позволяет делать более корректные выводы.

Часто решение проблемы исключения пропусков возможно на стороне систем учета и автоматизации. Нельзя рассчитывать, что задачу можно решить только за счет хитрых алгоритмов.

Как итог работы с пропусками можно заметить, что **средняя прибыль и средний чек в портрете клиентов изменились**. Вот значения до восстановления пропусков.

+ ∑ Факты

	Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%| Процент п...	∑ Значе...	%| Проц...	∑ Значе...	∑ Значение	∑ Значение
> 1. VIP	13 561 255	23,61%	5 740 185,84	22,42%	589	23 024,20	9 745,65
> 2. Важный	12 846 264	22,37%	5 794 241,92	22,63%	1 313	9 783,90	4 412,98
> 3. Перспективный	18 147 709	31,60%	8 267 434,26	32,29%	2 857	6 352,02	2 893,75
> 4. Начинающий	12 878 607	22,42%	5 803 369,78	22,66%	3 398	3 790,06	1
Итого:	57 433 835	100,00%	25 605 231,80	100,00%	8 157	7 041,05	3

Портрет клиентов до восстановления пропусков

А вот, что получилось после.

+ ∑ Факты

	Выручка (пп)		Валовая прибыль (пп)		Кол-во по...	Средний чек	Средняя прибыль
	∑ Сумма		∑ Сумма		∑ Сумма	∑ Значение	∑ Значение
	∑ Значе...	%| Процент п...	∑ Значе...	%| Проц...	∑ Значе...	∑ Значение	∑ Значение
> 1. VIP	17 218 041	26,95%	8 596 244,02	26,89%	649	26 530,11	13 245,37
> 2. Важный	14 494 470	22,69%	7 251 288,64	22,68%	1 619	8 952,73	4 478,87
> 3. Перспективный	19 116 581	29,92%	9 610 093,75	30,06%	3 050	6 267,73	3 150,85
> 4. Начинающий	13 056 469	20,44%	6 515 660,16	20,38%	3 514	3 715,56	1
Итого:	63 885 561	100,00%	31 973 286,58	100,00%	8 832	7 233,42	3

Портрет клиентов после восстановления пропусков

Средняя прибыль с продажи в категории VIP неплохо подросла, как и в сегменте перспективных клиентов. А значит дальнейшее планирование по работе с этими сегментами может быть пересмотрено.

Общий средний чек вырос, потому что добавлены пропущенные транзакции за январь 2022. Без этой операции средний чек бы снизился, т.к. при неизменной сумме продаж, увеличилось количество транзакций за счет восстановления номеров продаж.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

1.Сформулируйте прикладную экономическую или управленческую оптимизационную задачу и опишите ее решение с применением генетического алгоритма.

2.Классифицирующие системы Холланда.

3.Перечислите основные этапы технологии генетического программирования.

Лабораторная работа № 9.

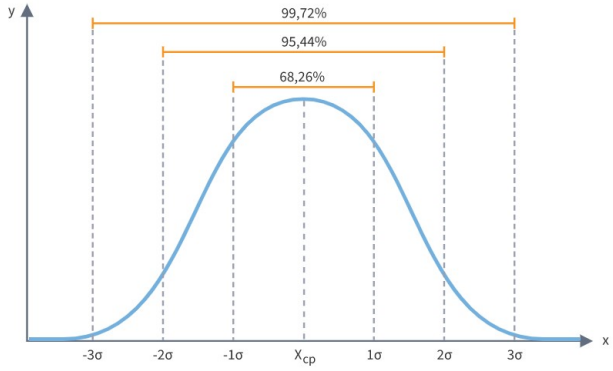
Тема: Паранормальные явления в работе с данными.

Цель работы. Формирование знаний и навыков работы в среде интеллектуального анализа данных.

Формируемые компетенции или их части: ОПК-8

Теоретическая часть

Для ответа на вопрос, что является выбросом, необходимо определиться с тем, как распределены значения анализируемого процесса. Одним из частых вариантов распределения является нормальное (Гауссово) распределение, которое выглядит следующим образом.



При таком распределении большая часть значений (95.44%) находится в диапазоне +/- двух стандартных отклонений. Выбросами обычно считаются значения в диапазоне от 3 до 5 стандартных отклонений, а экстремальными значениями – то, что превышает 5 стандартных отклонений.

Так в чем проблема с выбросами? Чем плох, например, резкий всплеск продаж? Это же как выиграть в лотерею! Однако в реальности все не так просто.

Выброс, с точки зрения регулярного менеджмента, нарушает прогнозируемый ход событий. Скорее всего процессы, приведшие к такому результату, отличаются от того, под что заточена организация. Отклонения в любую сторону создают проблемы основному направлению бизнеса.

Пример из жизни. *Оптовая компания торгует канцелярскими товарами и отгружает каждый день в среднем 10 палет продукции десяткам покупателей. В результате многолетней работы сложился отлаженный процесс — компания работает стабильно без простоев и перегрузок.*

Вдруг появляется новый клиент — огромная розничная сеть, которой за раз требуется отгрузить сотню палет. Казалось бы, отличная сделка, но по ходу выполнения заказа выясняется, что:

1. *Большая отгрузка привела к тому, что за день опустела половина склада, и снизился уровень сервиса для основных клиентов.*
2. *Людей в штате недостаточно для сборки заказа в срок, как следствие — переработки и недовольство.*
3. *Стабильные бизнес-процессы нарушились, что привело к кассовым разрывам, увеличению брака при сборке партии и прочим негативным последствиям.*

Конечно, этот пример гротескный. Не каждое экстремальное явление приводит к негативным последствиям. Но общий смысл не меняется: если в данных (в нашем случае в продажах) есть выбросы, значит реализовался сценарий, выбивающийся из стабильного бизнес-процесса.

Поэтому компании стремятся строить работу так, чтобы минимизировать резкие отклонения в любую сторону. Если же они возникают часто, то проблемы стараются локализовать:

1. Выделить отдельное направление деятельности с другими бизнес-процессами.
2. Оптимизировать работу, например, договариваться с крупными клиентами заранее или согласовать специальные регламенты.

3. Построить логику работы с клиентами так, чтобы поощрять за действия, повышающие стабильность процессов, и «наказывать» за то, что их ухудшает.

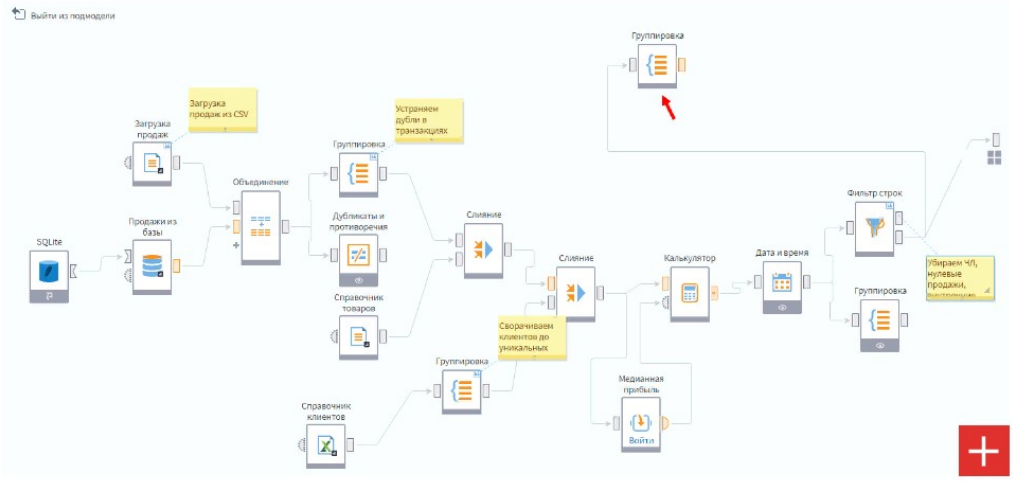
Если, как в примере выше, встречается один выброс, то его легко обнаружить. Сложнее найти их в большом наборе данных. Для этого нужно использовать соответствующие алгоритмы.-

Выбросы в неупорядоченных значениях

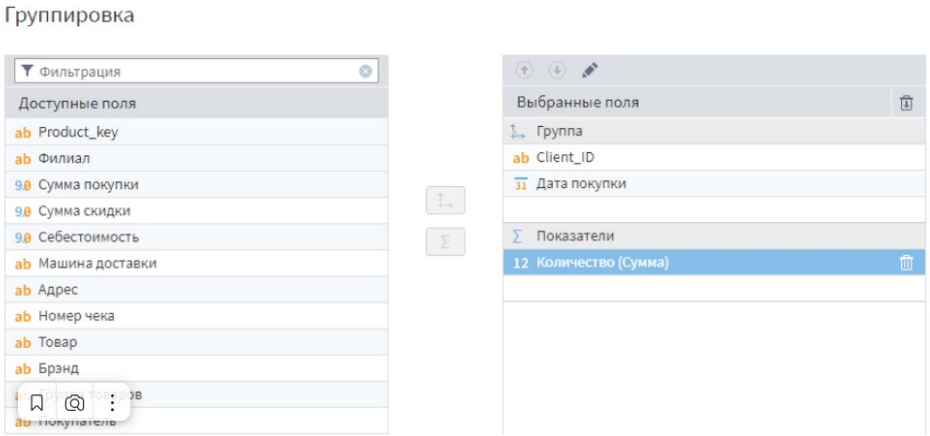
В визуализаторе *Качество данных* отображаются выбросы в полях. Выделив ячейку, можно просмотреть все строки, где они содержатся.

Однако было бы удобно иметь в таблице продаж дополнительный аналитический признак, позволяющий увидеть аномалии в отчетах не по каждой транзакции, а в разрезе дней. Для чего требуется правильно подготовить данные

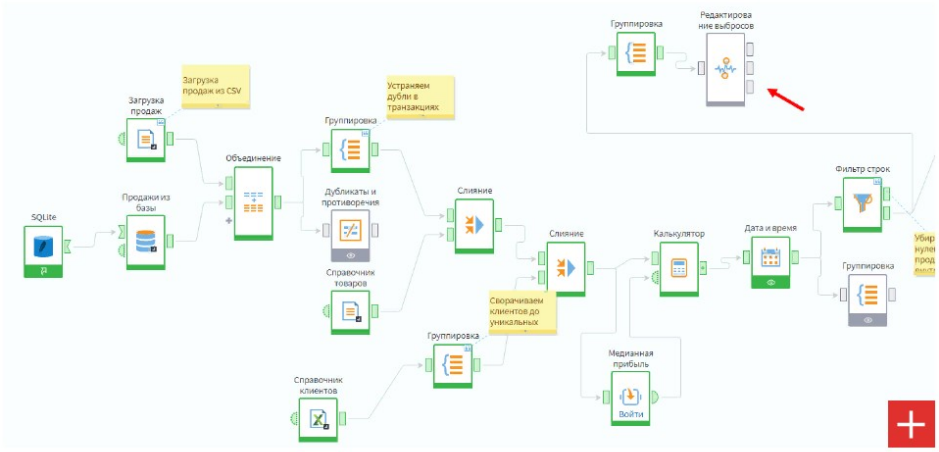
В подмодели *Продажи* нужно добавить узел ***Группировка*** после фильтра, с которого подаются данные на выходной порт узла.



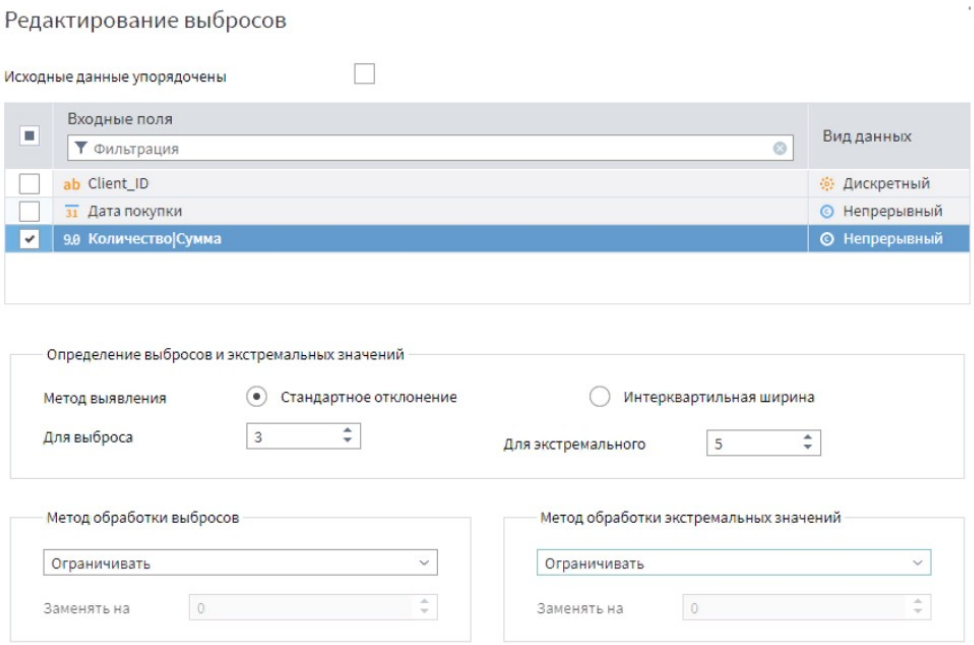
Далее надо настроить обработчик так, чтобы данные группировались по полям *Client_ID* и *Дата покупки* и суммировались по полю *Количество*.



После этого узла нужно добавить компонент ***Редактирование выбросов*** из группы *Предобработка*.



Искать выбросы надо по полю *Количество*, его требуется отметить галочкой. В открывшейся внизу панели настроек задаются способы обнаружения выбросов и методы их обработки.



Методы и параметры определения выбросов надо оставить без изменений. На практике может потребоваться подобрать границы экспериментальным путем, но начинать стоит со стандартных значений.

Вариант обработки *Ограничивать* срезает выбросы. Это полезно, когда нужно сузить разброс данных в задачах моделирования или прогнозирования. Остальные опции — разные варианты замены.

Важно! *Необходимо понимать, что в ассортименте имеются товары разного формата. 100 карандашей — это не то же самое, что и 100 глобусов с точки зрения логистики или финансов. Но при решении учебной задачи не будем обращать на это внимание.*

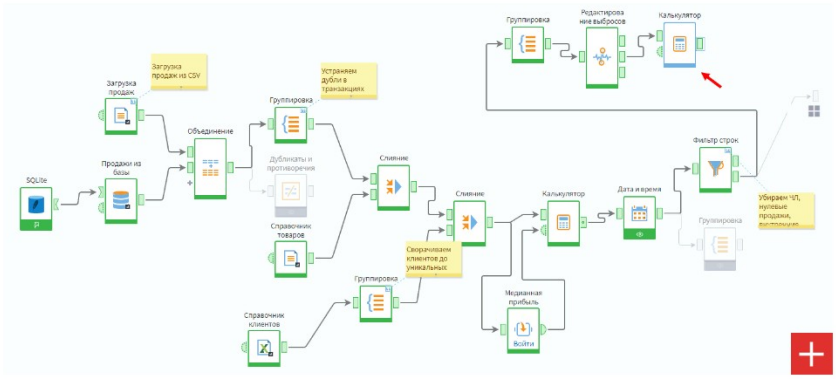
Правильнее искать выбросы среди однородных объектов. Такую разбивку можно сделать вручную, разделив данные на несколько наборов и обработав каждый в отдельности. Другой способ — автоматизировать расчет для каждой категории в отдельности с помощью Циклов. Ознакомиться с этим можно в инструкции.

На выходе из узла редактирования выбросов имеется 3 порта:

- выходной набор;
- выбросы;
- экстремальные значения.

Во второй и третий порт выводятся значения, которые являются нетипичными для этого датасета. Их надо отметить аналитическим признаком и присоединить к транзакциям.

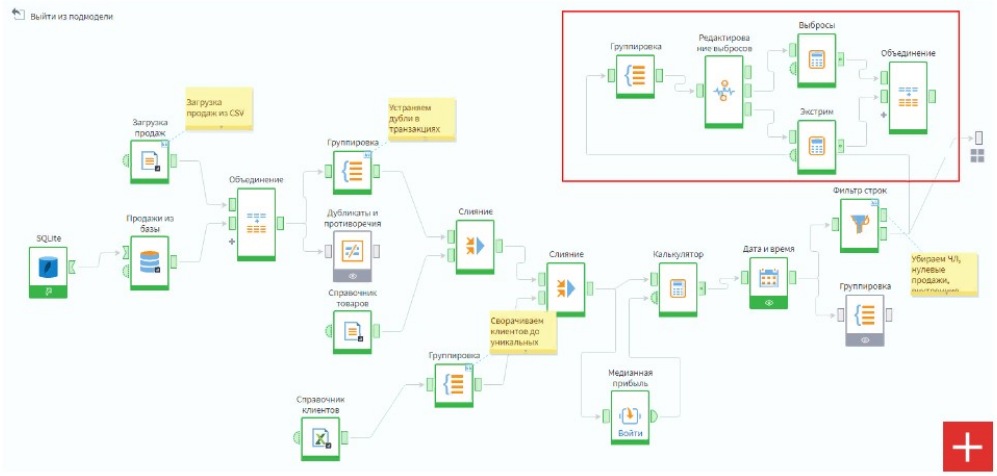
Для этого нужно подать данные со второго выхода на узел *Калькулятор*, добавить текстовое поле ***Ejection*** (метка *Тип выброса*) и значением ***Выброс***.



Аналогично сделать для третьего порта, только в значении поля прописать ***Экстемальное***.

На выходном порту каждого калькулятора надо отключить поле ***Количество***, т.к. оно не понадобится при слиянии с основной таблицей продаж.

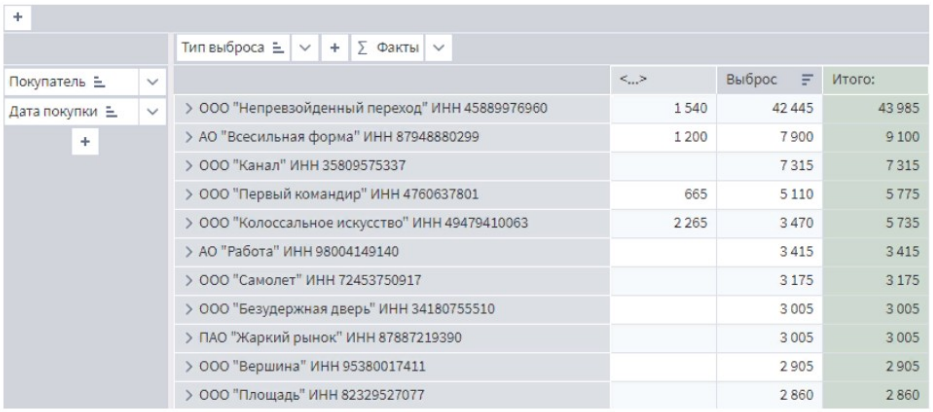
Обе ветки нужно объединить при помощи обработчика ***Объединение***, не забывая настроить совпадение полей.



Затем нужно добавить в сценарий узел ***Слияние*** и подать на первый порт данные по транзакциям, а на второй порт – выход из узла ***Объединение***, связав по полю *ИД клиента* и *Дата покупки*.

Для экономии места подготовленные узлы желательно свернуть в подмодель как продемонстрировано в ролике.

Теперь при построении *Куба* по продажам данные разбиваются на записи с выбросами и без них. Эту информацию можно детализировать до дней и транзакций.



Важный момент — выбросы надо рассчитать на очищенных и отфильтрованных данных, т.е. убрать дубли, заполнить пропуски, разобраться с фейками и т.п.

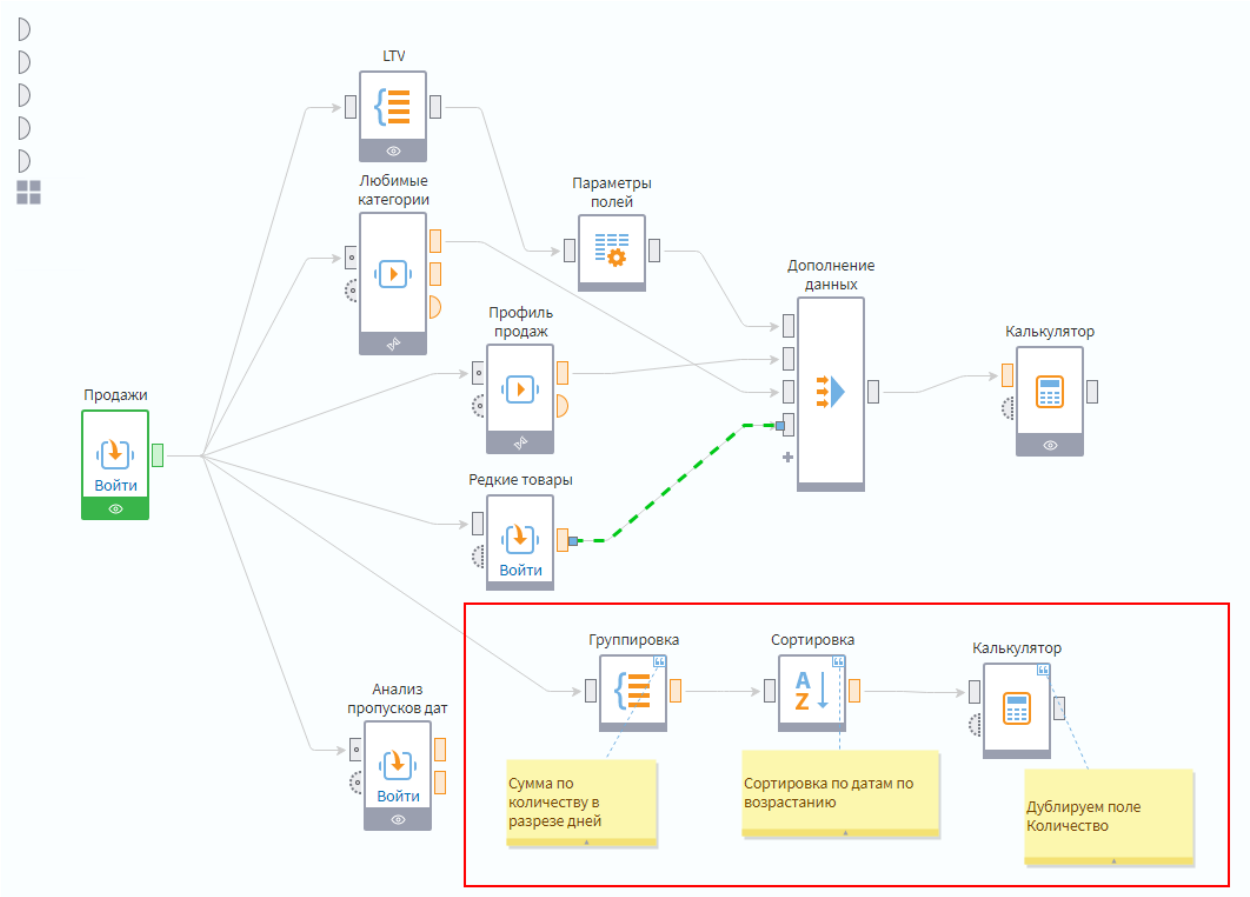
Выбросы в упорядоченных данных

В ранее подготовленном сценарии не принимался во внимание порядок следования записей. Однако если анализируются, например, временные ряды с ежедневными продажами, то упорядоченность должна учитываться при определении выбросов.

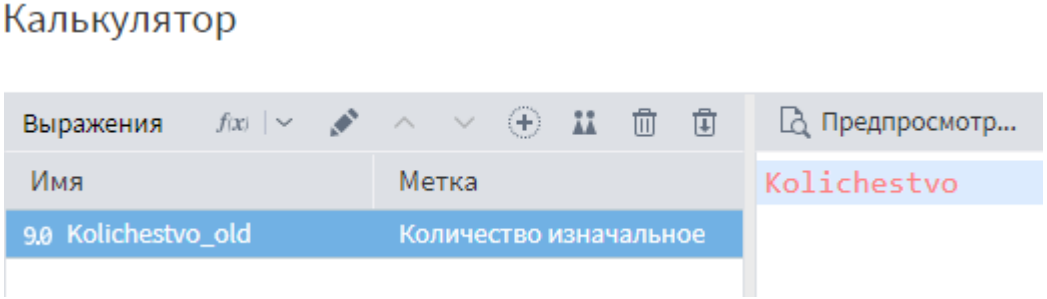
Это может потребоваться при анализе отдачи от маркетинговых мероприятий. Если взять за основу идею, что в обычной ситуации отгрузки изменяются плавно, то всплеск продаж после рекламы можно рассматривать как экономический эффект от ее проведения. При этом отсутствие всплеска, а значит и выбросов, говорит о том, что реклама не принесла пользы.

Для этого в сценарий надо добавить узлы *Группировки* количества продаж по дням, а затем *Сортировку* по возрастанию дат.

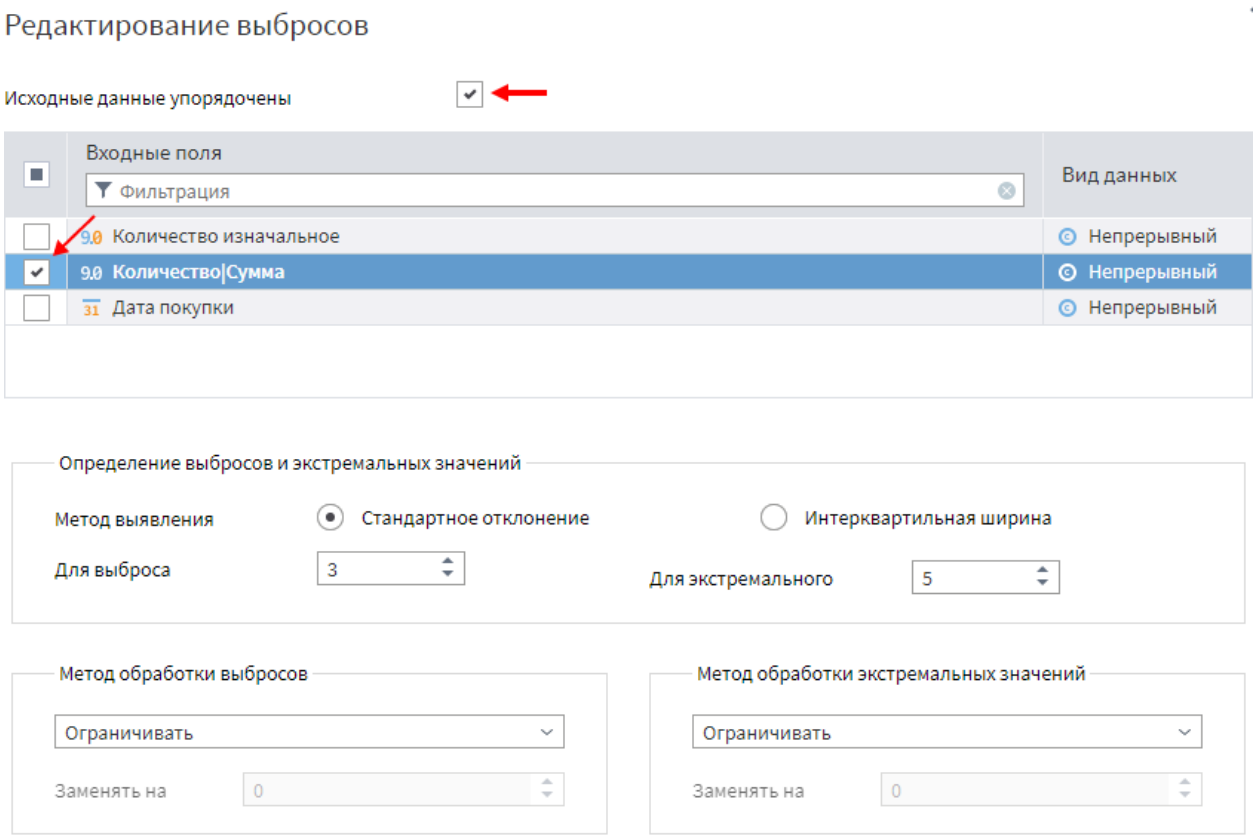
Важный момент — выбросы надо рассчитать на очищенных и отфильтрованных данных, т.е. убрать дубли, заполнить пропуски, разобраться с фейками и т.п.



Затем добавить узел *Калькулятор*, где создать поле ***Изначальное количество***, содержащее значение поля ***Количество***.



Потом добавить узел *Редактирование выбросов* и настроить его следующим образом, обязательно указав, что данные упорядочены.



Для просмотра полученных результатов нужно в первый порт обработчика *Редактирование выбросов* добавить визуализатор *Диаграмма* и разместить ***Даты покупки*** на ось X, а поля ***Количество|Сумма*** и ***Количество изначальное*** – в центр диаграммы.

Поле *Количество|Сумма* содержит данные с отредактированными выбросами. Сравнив его с графиком по полю *Количество изначальное* можно оценить размер всплеска.

Сопоставив графики с календарем мероприятий или датами проведения рекламных кампаний, можно примерно оценить, был ли от этих действий экономический эффект.

Заключение

Анализ выбросов не требует долгих расчетов или сложной математики. Этот вид аналитики выполняется просто и в то же время позволяет легко делать практические выводы.

Любые отклонения — это ситуации, на которые стоит обратить внимание. Они могут указать как на проблемные зоны бизнеса, так и на точки роста. Как всегда самые интересные результаты получаются на стыке дата-аналитики и бизнес-экспертизы.

Loginom сокращает время на выполнение рутинных операций с данными. Платформа позволяет автоматически выявить выбросы, т.е. события, на которые надо обратить внимание, а знание предметной области – понять причинно-следственные связи.

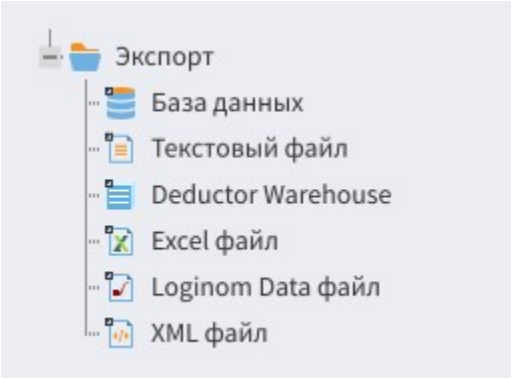
Полученные данные можно просматривать не только визуализаторами, встроенными в Loginom. Их можно выгрузить в различные приемники данных, такие как системы учета или специализированные BI продукты.

Однако недостаточно правильно подготовить таблицы и сделать видимым то, что ранее скрывалось за информационным шумом. Чтобы выполненная работа принесла максимум пользы, нужно разобраться еще с одним аспектом.

Конечной целью аналитики является принятие решений. Следовательно, рассчитанные метрики, показатели, прогнозы должны быть доставлены до «потребителей» — лиц, принимающих решения.

Не все сотрудники захотят и смогут работать в Loginom. Основным инструментом, к примеру, клиентского менеджера является не аналитическая платформа, а CRM-система, где он хотел бы видеть всю необходимую информацию.

Для передачи информации в другие системы в Loginom имеется набор узлов группы ***Экспорт***. Кроме того, можно воспользоваться механизмами интеграции при помощи веб-сервисов: REST или SOAP.



Данные, сформированные в Loginom, могут быть выгружены в сторонние системы, в том числе для нужд пользователей, не работающих с платформой и не имеющих лицензии не нее.

Бизнес-процессы на основе данных

Один из сценариев применения ранее рассчитанных данных для принятия решений может быть следующим:

1. Раз в месяц профили клиентов пересчитываются и сохраняются в базе.
2. Текущие показатели сравниваются с данными предыдущего месяца.
3. Если отклонение показателей больше определенного порога, ответственным сотрудникам высылаются оповещения.

Идеальным вариантом было бы отображать непосредственно в CRM-системе статусы и историю их изменений, а в случае необходимости автоматически запускать бизнес-процессы реагирования. Такие интеграции требуют времени и ресурсов на разработку, поэтому можно начать с простой, но действенной механики — оповещения в месседжерах.

Пример автоматической отправки уведомлений из сценария Loginom в Telergam при помощи чат-бота описан в статье «Уведомления из сценариев Loginom при помощи Telegram». Для этого необходимо зарегистрировать в месседжере чат-бот и воспользоваться готовым компонентом.

Настроить сценарий можно таким образом, чтобы сотрудник не просто прочитал сообщение о проблеме, но и перешел по ссылке в отчет с подготовленными данными.

Аналогичная механика может работать при взаимодействии с другими web-сервисами, что позволяет использовать Loginom как систему поддержки принятия решений.

Loginom идеально подходит в качестве движка («мозга») подобных систем за счет способности обрабатывать большие объемы данных, объединять информацию из множества источников, создавать/изменять/расширять без программирования логику принятия решений.

Кроме того, можно настроить внешнее управление сценариями за счет использования переменных или настроечных таблиц.

Выгрузка данных в другие аналитические системы

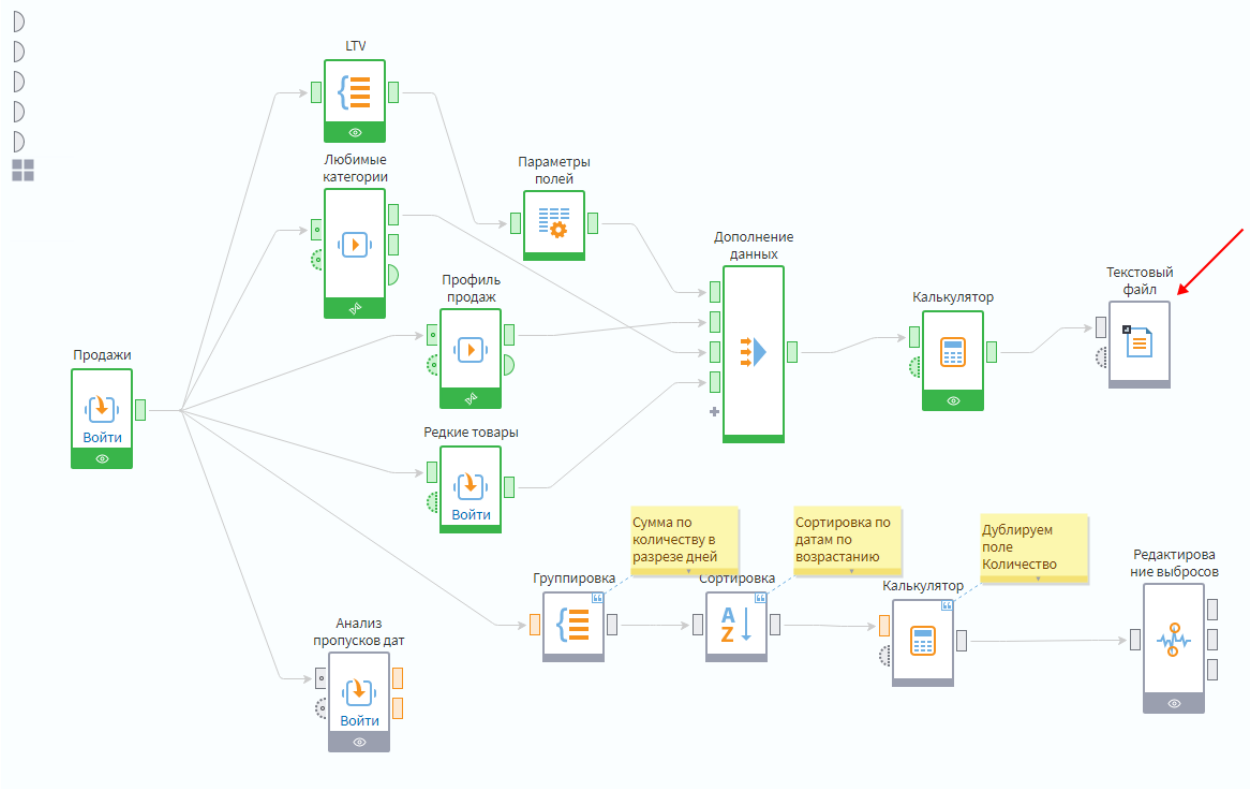
Куб – удобный визуализатор, но он не покрывает всех потребностей. Часто требуется изучить данные из множества таблиц, строить дашборды, просматривать графики на смартфоне. Для этого существуют специализированные BI-продукты.

Если возможностей встроенных в Loginom визуализаторов недостаточно, можно подготовленные и очищенные данные экспортировать в сторонние системы.

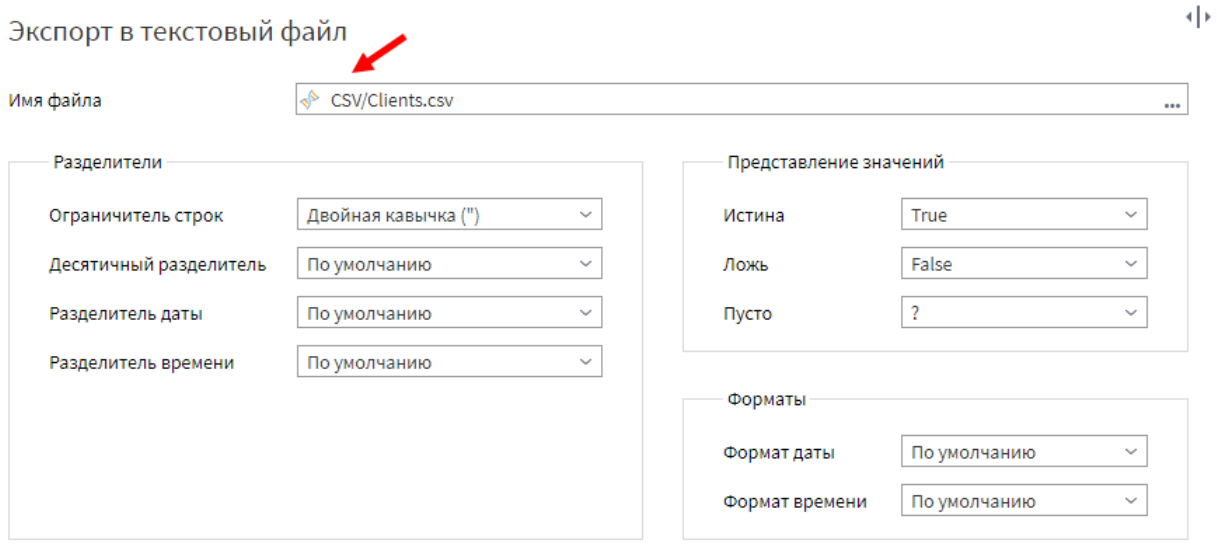
Наиболее надежный и универсальный механизм обмена — загрузка таблиц в базу. Чаще всего ETL-процессы завершаются экспортом подготовленных данных в БД, витрину или хранилище данных. Но для простоты можно рассмотреть другой распространенный вариант интеграции — обмен при помощи csv-файлов.

В качестве BI-инструмента будет продемонстрирован сервис визуализации Yandex DataLens, предоставляемый бесплатно пользователям Яндекс.Облако.

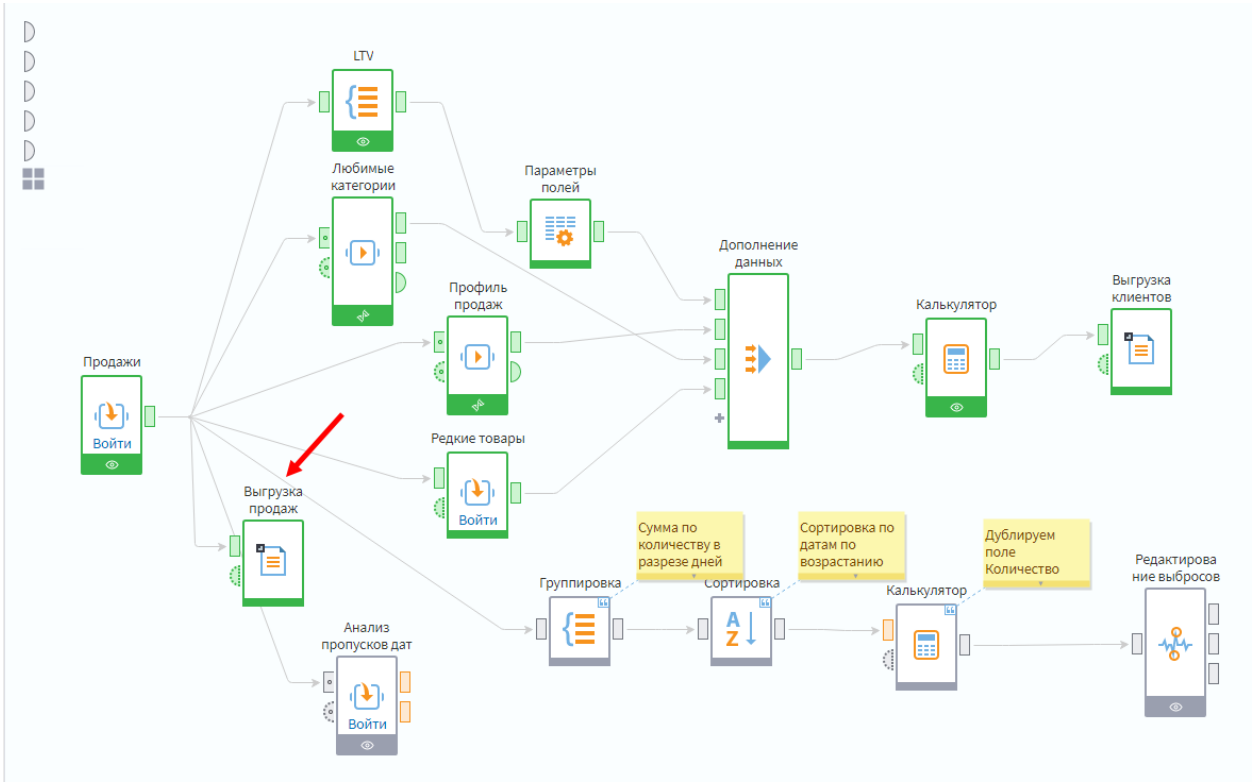
Для передачи данных в DataLens надо добавить узел экспорта в текстовый файл.



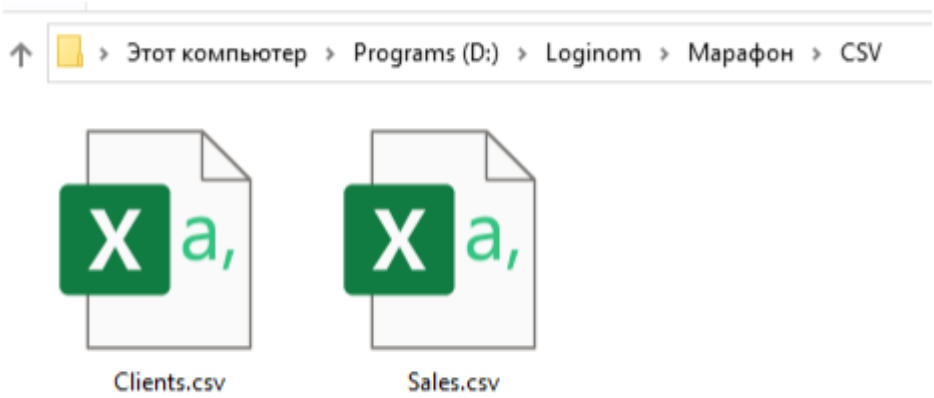
Необходимо создать в рабочей папке проекта каталог CSV и задать имя файла *Clients.csv*.



Аналогичные действия надо выполнить для выгрузки данных из подмодели *Продажи*. Только для этого файла нужно задать имя *Sales.csv*.



В итоге в целевой папке появятся 2 файла, которые будут перезаписываться при каждой активации узла экспорта.



Кстати, при наличии в IT отделе разработчиков, способных писать код, например в 1С, можно настроить загрузку по расписанию данных из файлов в бухгалтерскую систему. Это рабочий сценарий интеграции, особенно на первое время.

Визуализация в Yandex Datalens

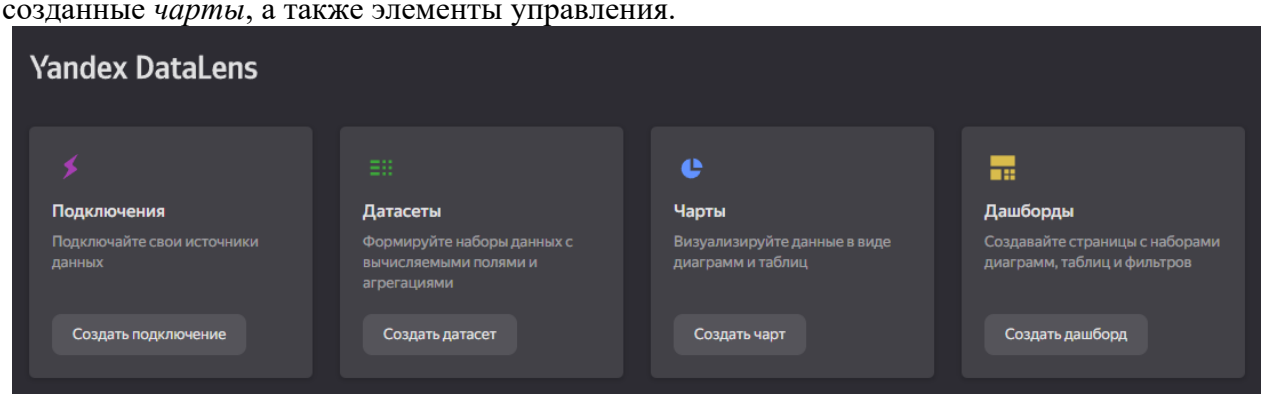
Вначале надо зарегистрироваться в сервисе Yandex DataLens, нажав на кнопку ниже. Сервис предоставляется бесплатно. Без ограничений на количество пользователей и запросов.

Далее процесс построения дашбордов в YDL таков:

1. Создать подключения к данным. Это могут быть коннекты к базам данных или загруженные файлы.
2. Из подключенных данных нужно собрать Датасет. Это модель данных, т.е. набор таблиц из подключений с прописанными связями между ними. Таким образом, данные из разных источников могут быть объединены в одну структуру.

3. На основе *Датасета* создать *Чарты* — отдельные визуализации, оперирующие данными из датасета как единым целым. Иначе говоря, на одном чарте можно собрать данные из одной таблицы в разрезе другой.

4. Последний шаг — создание *дашбордов*, т.е. рабочих областей, на которые добавляются ранее созданные *чарты*, а также элементы управления.



>Необходимо все описанное выше выполнить для выгруженных файлов.

Вначале надо создать подключение с типом *File*, загрузив в него файлы продаж и клиентов из папки *CSV*.

Далее создать *Датасет*, перетащив в рабочую область таблицу *Sales*, а после нее — *Client*s.

В DataLens модели данных строятся по топологии *Снежинка*. Первая таблица — центральная. К ней присоединяются таблицы второго уровня через Join. Тип Join'а можно задать, если кликнуть по настройке соединения между наборами данных. К таблицам второго уровня могут присоединяться таблицы третьего уровня и т.д.

В настройках полей можно отключить дублирующиеся поля (*Client_ID* и *Покупатель* повторяются), а полям показателей задать способы агрегации по умолчанию.

Датасет нужно сохранить под названием ***Loginom***.

А затем создать второй чарт со списком клиентов и рассчитанными для них параметрами, вроде группы любимых категорий и статуса. Здесь тоже рекомендуется поэкспериментировать с настройками для выбора наиболее удобного способа представления данных/

Задание.

Создание дашборда, в котором размещено два ранее настроенных чарта.

Если добавить селектор по полю Тип выброса и Месяц покупки, то можно будет оперативно выявлять клиентов по признаку, генерировали ли они аномальные продажи за период.

Как видно, совместное использование платформы Loginom и BI-систем позволяет взять лучшее из двух миров. В Loginom это возможность повысить качество данных, обогатить аналитическими атрибутами, рассчитать сложные метрики, построить прогнозы, а затем, загрузив в специализированную систему, доставить информацию лицам, принимающим решения в наиболее удобном для них виде.

Содержание отчета: Отчёт по выполненной работе.

Контрольные вопросы (перечень вопросов по теме, на которые студент обязан знать ответы) и /или тестовые задания

4.Сформулируйте прикладную экономическую или управленческую оптимизационную задачу и опишите ее решение с применением генетического алгоритма.

5.Классифицирующие системы Холланда.

6.Перечислите основные этапы технологии генетического программирования.

Список рекомендуемой литературы

Основная литература:

1. Нестеров, С. А. Интеллектуальный анализ данных средствами MS SQL Server 2008 / С.А. Нестеров. - 2-е изд., испр. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 338 с. : ил. - http://biblioclub.ru/. - Библиогр. в кн
2. Пальмов, С.В. Интеллектуальный анализ данных Электронный ресурс : учебное пособие / С.В. Пальмов. - Самара : Поволжский государственный университет телекоммуникаций и информатики, 2017. - 127 с. - Книга находится в базовой версии ЭБС IPRbooks.
3. Управление данными : учебник / Ю.Ю. Громов, О.Г. Иванова, А.В. Яковлев, В.Г. Однолько ; Министерство образования и науки Российской Федерации ; Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тамбовский государственный технический университет». - Тамбов : Издательство ФГБОУ ВПО «ТГТУ», 2015. - 192 с. : ил., табл., схем. - http://biblioclub.ru/. - Библиогр. в кн. - ISBN 978-5-8265-1385-9

Дополнительная литература:

1 Васюков, О. Г. Управление данными : учебно-методическое пособие / О.Г. Васюков ; Министерство образования и науки РФ ; Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Самарский государственный архитектурно-строительный университет». - Самара : Самарский государственный архитектурно-строительный университет, 2014. - 161 с. : табл., ил. - http://biblioclub.ru/. - Библиогр. в кн. - ISBN 978-5-9585-0608-8

2 Козлов, А. Ю. Статистический анализ данных в MS EXCEL : учеб. поосбие / А. Ю. Козлов, В. С. Мхитарян, В. Ф. Шишов. - М. : ИНФРА-М, 2012. - 320 с. - (Высшее образование). - Гриф: Рек. УМО. - ISBN 978-5-16-004579-5

3 Мельниченко, А. С. Математическая статистика и анализ данных Электронный ресурс : Учебное пособие / А. С. Мельниченко. - Математическая статистика и анализ данных,2019-09-01. - Москва : Издательский Дом МИСиС, 2018. - 45 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-906953-62-9

Интернет-ресурсы:

1. Официальный сайт библиотеки ФГАОУ ВО СКФУ Режим доступа: http://cataloq.ncstu.ru| cataloq –.
2. Информационная справочная система ГАРАНТ.РУ // Режим доступа: http://www.garant.ru/
3. Информационная справочная система КонсультантПлюс. // Режим доступа: http://www.consultant.ru
4. Инфраструктура научно-исследовательских данных — платформа доступа к данным для научных исследований // Режим доступа: https://www.data-in.ru
5. Профессиональное сообщество специалистов по обработке данных и машинному обучению // Режим доступа: https://www.kaggle.com/
6. Профессиональная база данных Росстата // Режим доступа: Росстат — Базы данных (rosstat.gov.ru)