фото: © N-trash / www.shatterstock.com

Аналитика за 20 минут

Пётр Травкин

Технический консультант QlikTech Russia

Михаил Николаев

Исполнительный директор ЗАО «Витасофт»

Андрей Педоренко
Хочу представить участников. Круглый стол, вы видите, стоит не очень удобно, но, надеюсь, что это нам не помешает. Немножко о составе людей, которые будут принимать участие в дискуссии. Прежде всего, я хотел бы сказать, что сегодня в конференции принимают участие не только IT-шники, не только представители вендора, но и люди, которые ближе к бизнесу. Евгения Гришина, операционный директор компании «Альфастрахование», находится здесь у нас и будет участвовать в этом круглом столе. Также в нём будут участвовать Михаил Николаев, представитель компании «Витасофт», один из ведущих разработчиков этой компании, и Пётр Травкин, представитель компании «QlikTech».

Пётр Травкин
Спасибо за представление. Хотелось начать с краткой демонстрации самой системы. Мы планировали минут на тридцать, но нам сказали, что такого шикарного количества времени у нас нет, поэтому мы это сделаем за 20 минут. Мы покажем, какова система в работе: от загрузки данных до построения интерфейса. Мы пройдём те этапы, о которых Антон Порецков рассказывал на своём втором слайде.

Михаил Николаев
Добрый день, коллеги. Пару слов об источниках. У нас в примере будет три источника:

- база данных Access;
- таблица Excel;
- текстовый файл.

Какие бизнес задачи мы будем решать?
Во-первых, будем сравнивать сборы премии (или выплаты, или комиссию, или возвраты) в долях по различным разрезам: по продуктам, по каналам, по типам агентов, по типам страхователей.
Вторая задача - это анализ динамики – как менялась брутто премия (или выплаты, или любой другой показатель) во времени: по годам, кварталам, месяцам.
Третья задача - анализ сезонности сборов премии по кварталам и месяцам.
Четвёртая задача - анализ доходности агентов - также очень интересная вещь.
И, наконец, детализация. Любая выборка может быть детализирована.
С чего все начинается? Создаём новое приложение, начинаем с чистого листа (рис. 1).

Рисунок 1.

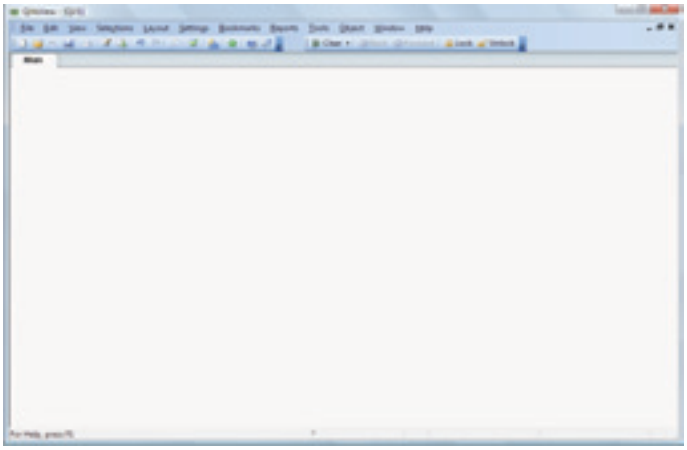
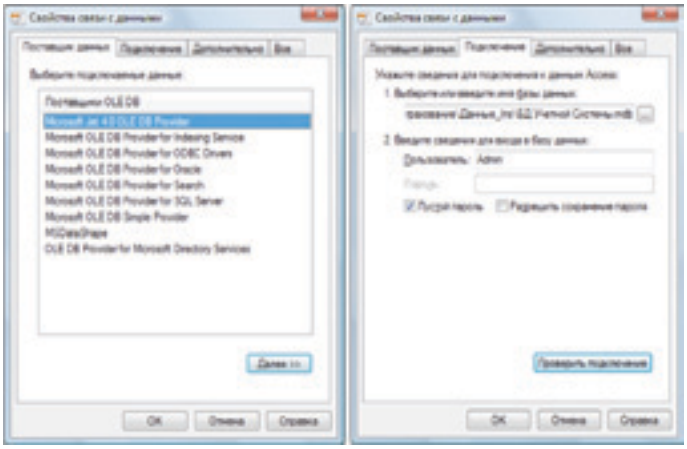


Рисунок 2.



С чего все начинается? Создаём новое приложение, начинаем с чистого листа.

Создание ETL (процесса Извлечения данных из источников, их Трансформации и Загрузки) состоит из трёх этапов работы мастера.

Первый этап - подключаемся к источникам данных.

Второй этап - выбираем, какие таблицы и какие поля будем переносить.

Третий этап - собственно сама загрузка.

Пару слов о содержимом источника. Источник - это выгрузка из типовой учётной системы страховой компании. В выгрузке данные представлены уже в виде конкретных показателей: брутто премия, выплаты, возвраты, комиссия, регрессы и несколько производных показателей. Все показатели представлены в следующих аналитических разрезах: продукт (2 уровня иерархии), канал продаж, страхователи, агенты, и календарный разрез - дата проводки. Для ускорения демонстрации мы ограничили количество аналитических разрезов, в реальной жизни их больше.

Подключаемся к источнику с помощью мастера (рис. 2).

Подключение готово.

Дальше мы выбираем, с какими таблицами и полями мы будем работать:

показатель, календарь, продукт, канал продаж, и, собственно, таблица фактов (рис. 3).

В таблице фактов в данном примере порядка 3-х миллионов записей. Из Access-источника мы всё выбрали. Мастер создал скрипты (рис. 4).

Теперь подключаемся к Excel-источнику, чтобы взять оттуда данные о страхователях.

Рисунок 3.

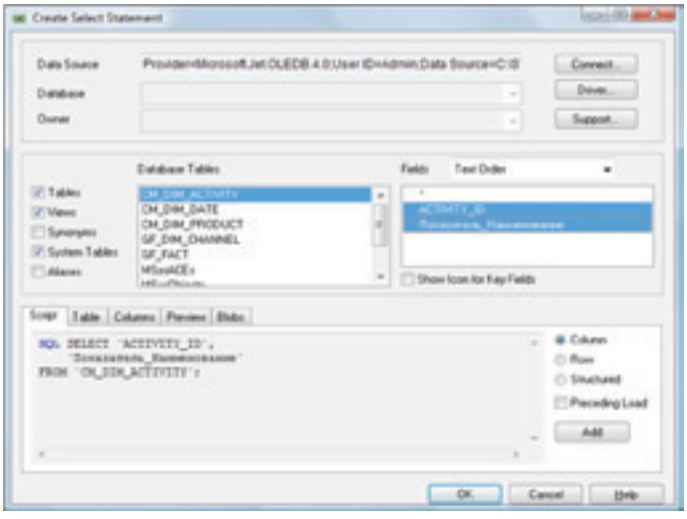


Рисунок 4.

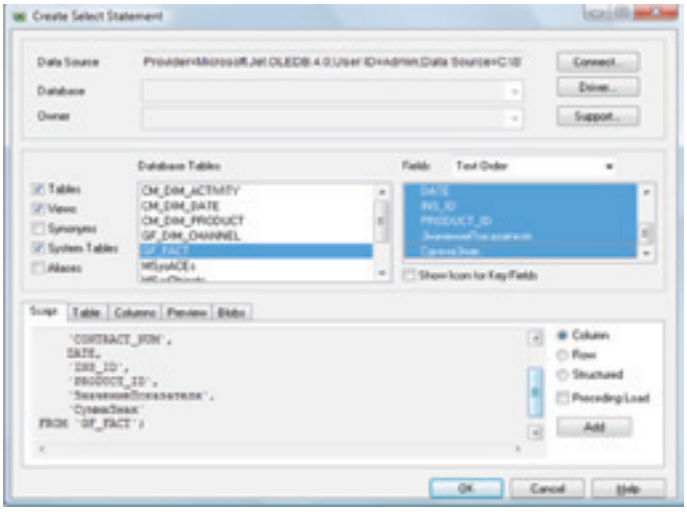


Рисунок 5.

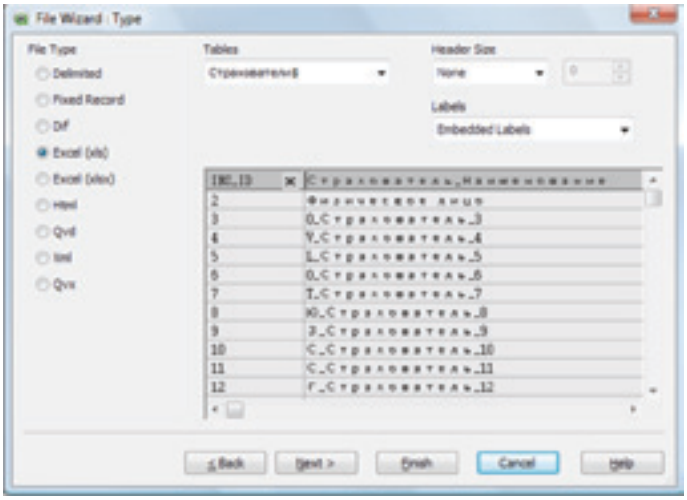
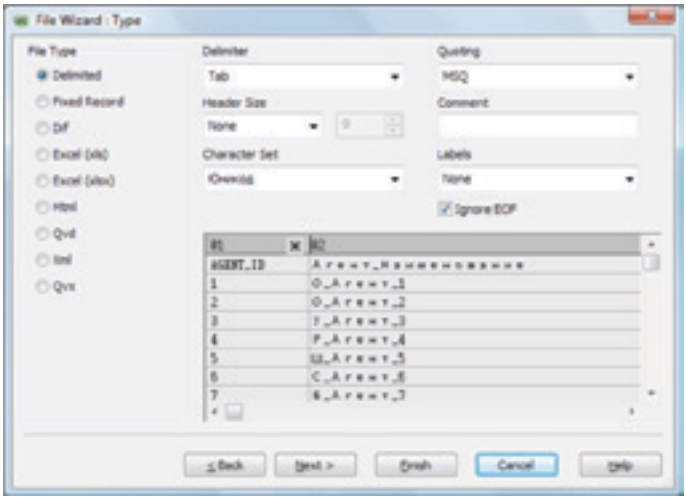


Рисунок 6.



Объёмы данных - всё, что позволяет оперативная память. Были у нас проекты с сотнями миллионов записей - всё работает.

Вот мастер сам определил, что это Excel, сам определил, из каких колонок что брать - скрипты готовы (рис. 5).

Теперь из текстового файла нужно получить данные об агентах.

Мастер сам определяет, что у нас является разделителем в текстовом файле. Задаём, какие поля нужны (рис. 6).

Всё, теперь все скрипты готовы. На языке скрипта можно делать трансформации, в том числе и довольно сложные. Здесь, для примера, мы вычислим месяц по дате, вычислим год + месяц, год + квартал, потому что интересно в этих разрезах анализировать динамику изменения показателей (рис. 7).

В реальных проектах могут быть достаточно сложные вычисления. Здесь можно делать процедуры очистки данных, сохранять промежуточные результаты расчётов на диск, это бывает очень полезно.

Ещё для порядка переименуем поле: было «Канал1» - будет «Канал».

Третий этап - это сама загрузка, запускаем её. Посмотрите, здесь 3 миллиона фактов, загрузка идёт достаточно быстро, это на слабом ноутбуке, а реально на сервере это происходит гораздо быстрее (рис. 8). Объёмы данных - всё, что позволяет оперативная память. Были у нас проекты с сотнями миллионов записей - всё работает.

Данные загрузились. Вот список полей, которые нам доступны. Мы можем их отбирать, можем выводить (рис. 9).

Сразу же можем фильтровать данные по выведенным полям. Вывели поле даты в ListBox, и сразу можно фильтровать по датам и месяцам (рис. 10).

Рисунок 7.

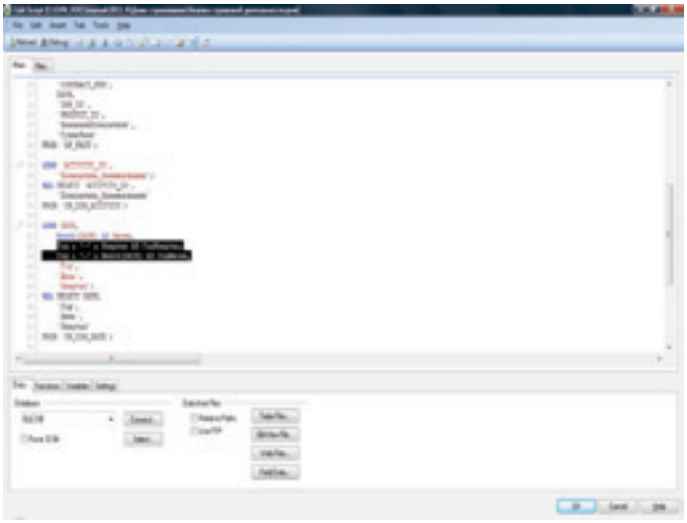


Рисунок 8.

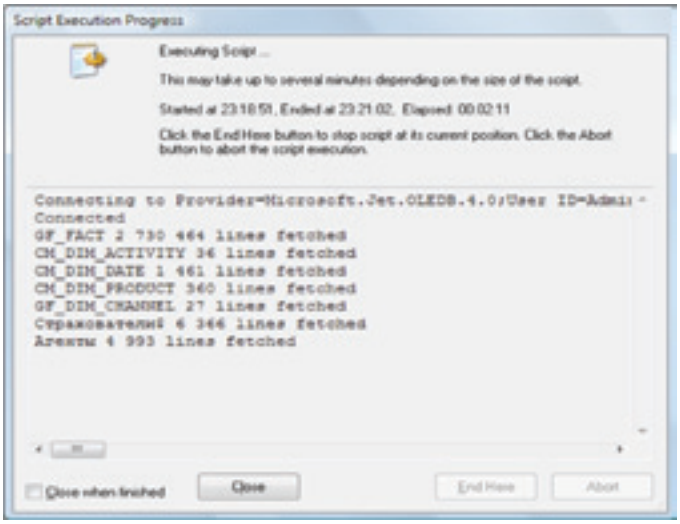


Рисунок 9.

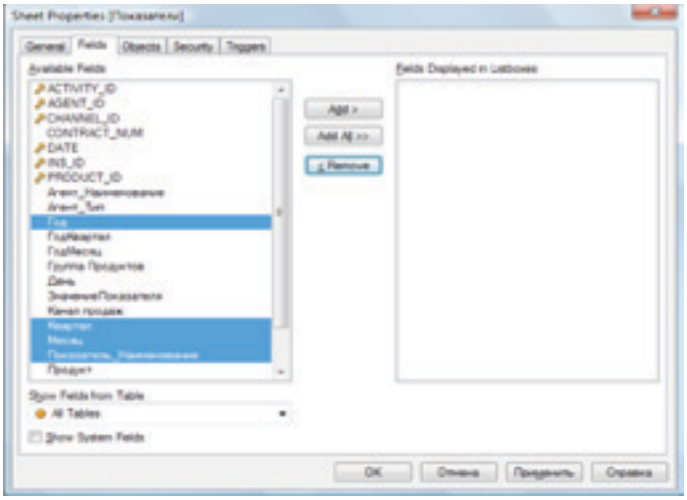
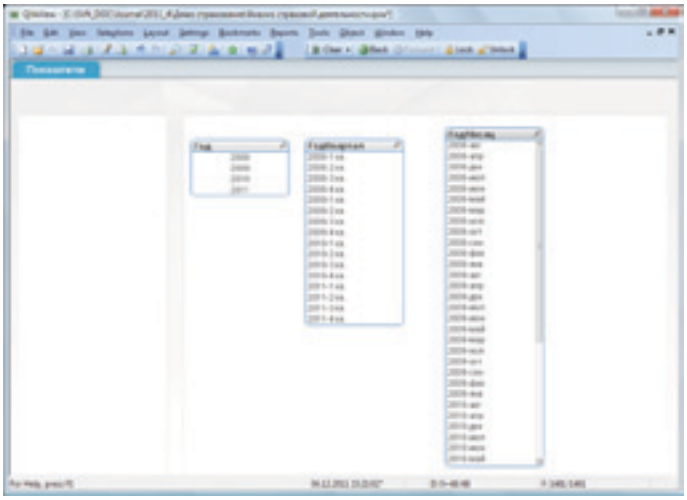


Рисунок 10.



Автоматически осуществилась связка по ключевым полям: вот таблица фактов, и вот по ключевым полям она связывается с таблицами измерений (рис. 11).
Начнём оформлять дизайн приложения. Применяем «тему». Далее, у нас есть заготовленный дизайнерский шаблон и для ускорения, чтобы не терять времени, мы скопируем сюда страницу с отбором данных. Отбор дат сверху, а сбоку у нас фильтрация по другим аналитическим разрезам.

Начнём решать бизнес задачи.
Первая бизнес задача - это анализ сборов, выплат, любых других показателей в долях по различным разрезам.

Андрей Педоренко
Михаил, можно одну секунду, хотелось бы акцентировать внимание. До сих пор, условно говоря, это была IT-шная часть. У нас в компании за это отвечает IT. За ETL, за его построение, за очистку данных - это наша ответственность. Вот то, что сейчас Миша рассказывает - это уже рабочее место пользователя. Пользователь сам может всё остальное выстраивать, по большому счёту.

Михаил Николаев
Вы видите, всё достаточно просто. Человек, прошедший 3-дневное обучение, может это легко делать.
Выберем тип диаграммы - это круговая диаграмма.
Выбираем измерение. Здесь мы хотим дать возможность пользователю оперативно переключать аналитические разрезы, для чего создаём циклическую группу измерений. Начнём с группы продуктов, потом продукт, далее канал продаж, затем тип агента, тип страхователя. Вот в рамках этой группы измерений мы будем анализировать факты (рис. 12).
Задаём, что мы суммируем. Мы суммируем значение показателя. Круговая диаграмма готова (рис. 13).
А сейчас небольшие операции с дизайном: чтобы диаграмма выглядела максимально наглядно. Сделаем также, чтобы отображались значения процентов (рис. 14).
Итак, мы видим расклад по группам продуктов. Нужно выбрать конкретный показатель, выбираем «Брутто-премию на-

Рисунок 11.

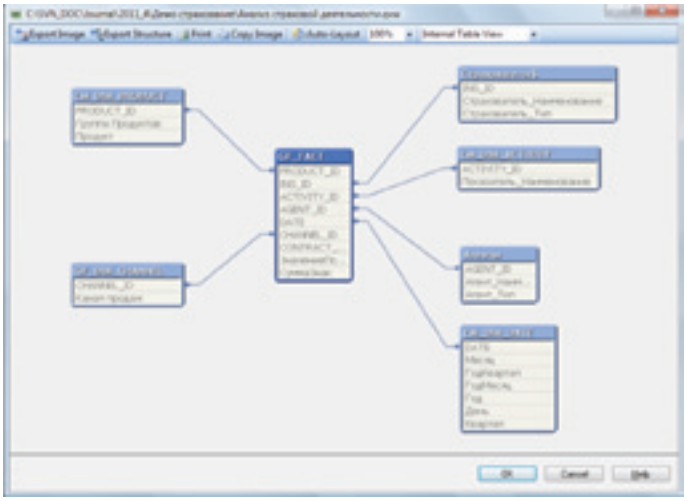


Рисунок 12.

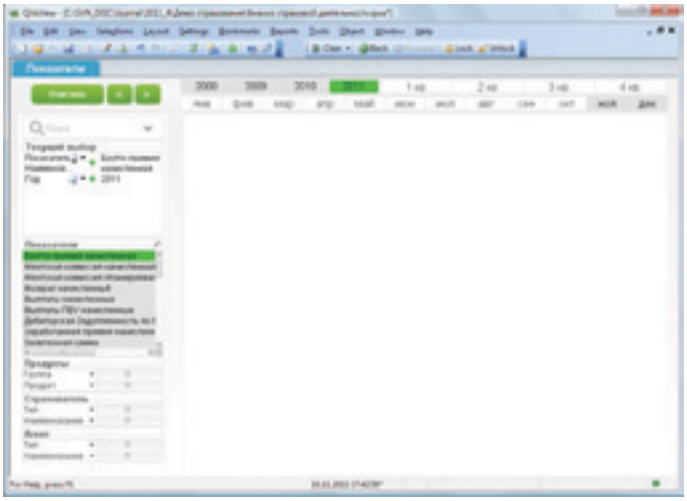


Рисунок 13.

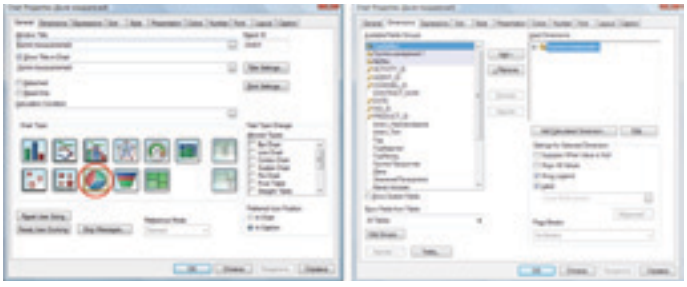


Рисунок 14.

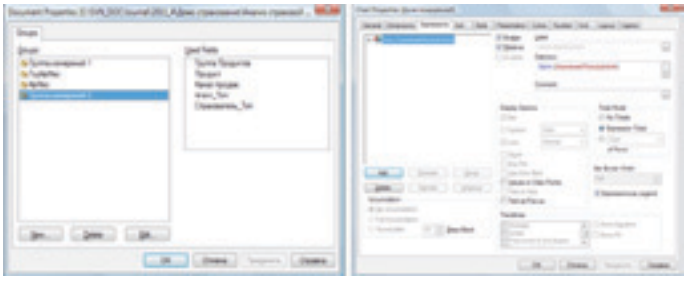
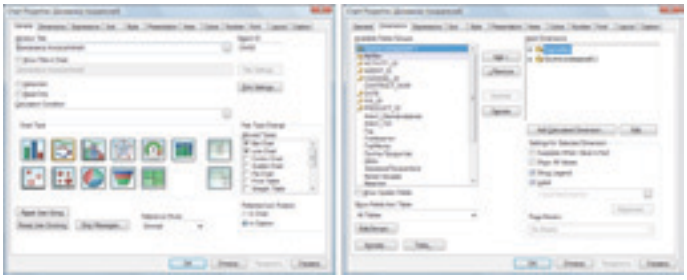


Рисунок 15.



численную» и видим, как распределяется премия по группам продуктов. Выберем год 2011, и мы видим, как распределена по группам продуктов премия, начисленная в 2011 году (рис. 16).

Григорий Жуков
Помните, я вначале говорил про три цвета: зелёный, серый, белый. Здесь наглядно видно: зелёным цветом отражается выбранное значение, белым цветом - связанные с нашим выбором значения, а серым цветом - несвязанные значения. Это ещё одно из преимуществ нашей системы, что цвет показывает информацию за пределами вашего выбора. И здесь вы легко можете перескакивать с зелёного на белое, с белого на серое (рис. 17).

Антон Порецков
А вот сейчас вы видите, что у нас в 2011 году премия начисленная была по всем группам продуктов, но её не было по «ВЗР» и по «Фин. Рискам» - они серые. То есть, в этом периоде, по этому показателю у нас не было ни одной транзакции по тем, кто выезжал. По всем остальным – были.

Михаил Николаев
Это и есть ассоциативная модель.

Андрей Педоренко
Коллеги, единственное, у вас уже 10 минут прошло, имейте в виду.

Михаил Николаев
Нам осталось не так уж и много сделать. Это была первая бизнес задача.

Следующая бизнес задача - это динамика.
Мы хотим исследовать во временных разрезах показатели (премию, выплаты, комиссию, и т.д.). Выбираем тип диаграммы, у нас будет два варианта: график или гистограмма. Дадим возможность пользователю переключать тип графика - это тоже очень удобно. Дальше мы создаем новую циклическую группу - это год, квартал, месяц: для трёх типов временных разрезов. Анализируем то же самое: сумму значения показателя. У нас по-

Рисунок 16.

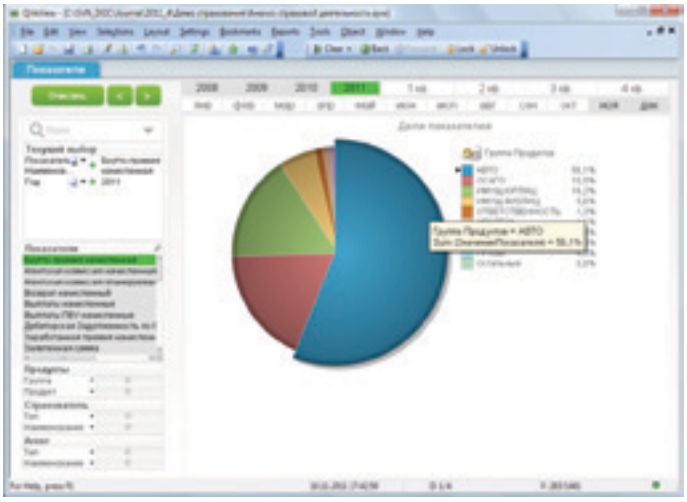


Рисунок 17.

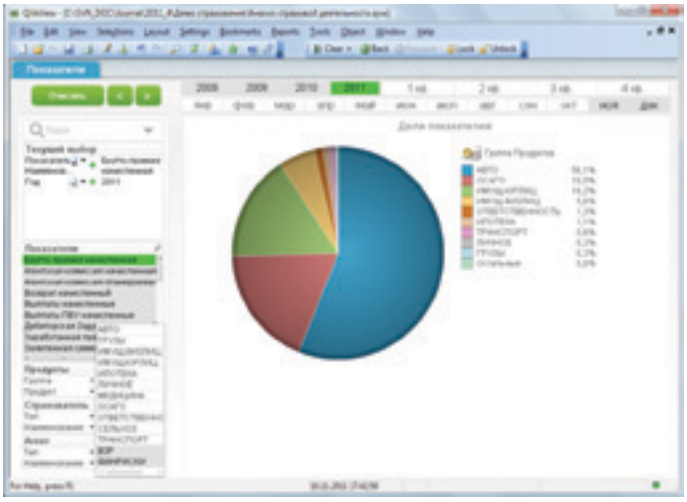


Рисунок 18.

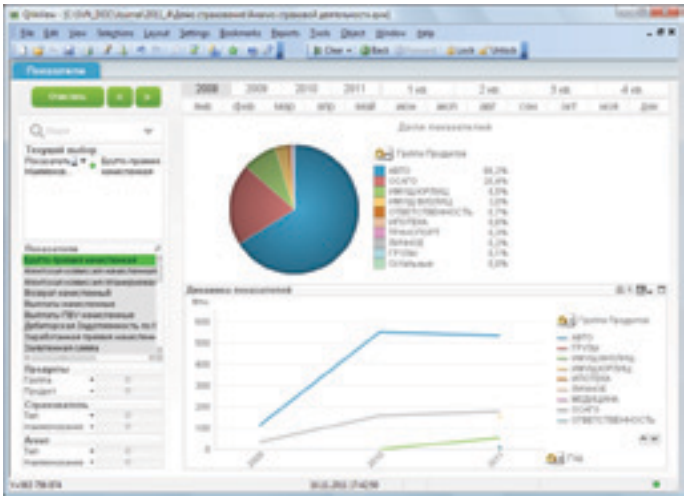


Рисунок 19.

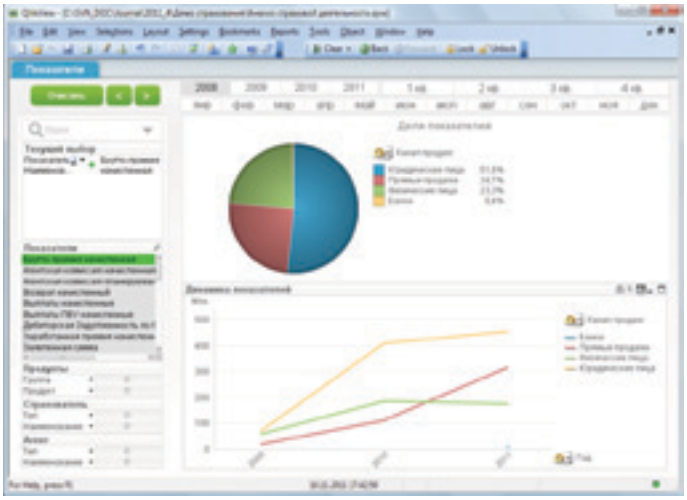


Рисунок 20.

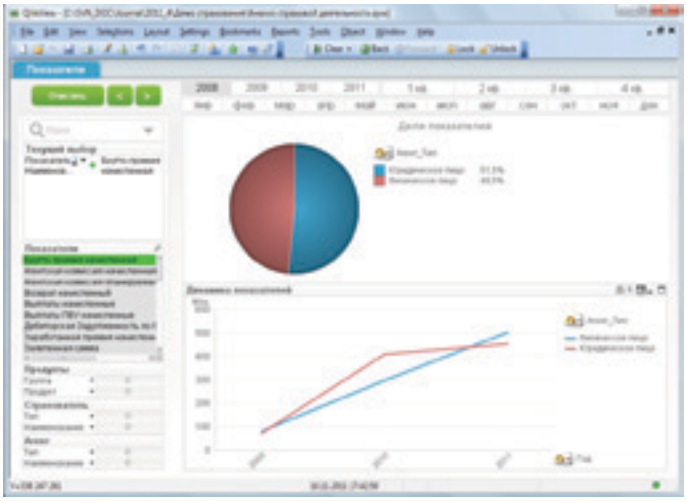


Рисунок 21.

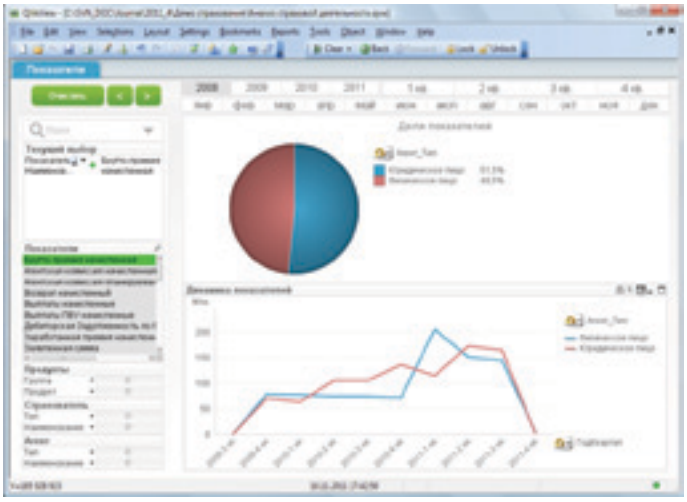


Рисунок 22.

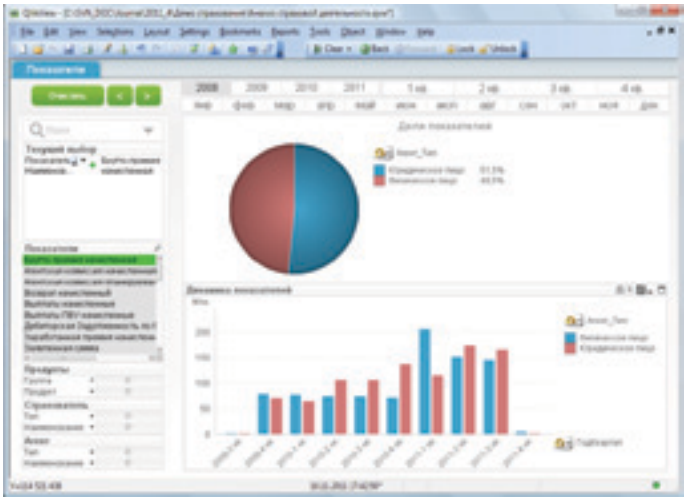


Рисунок 23.

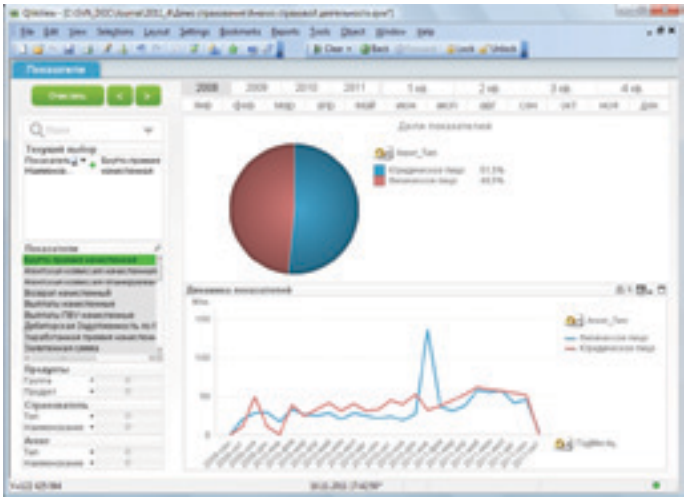
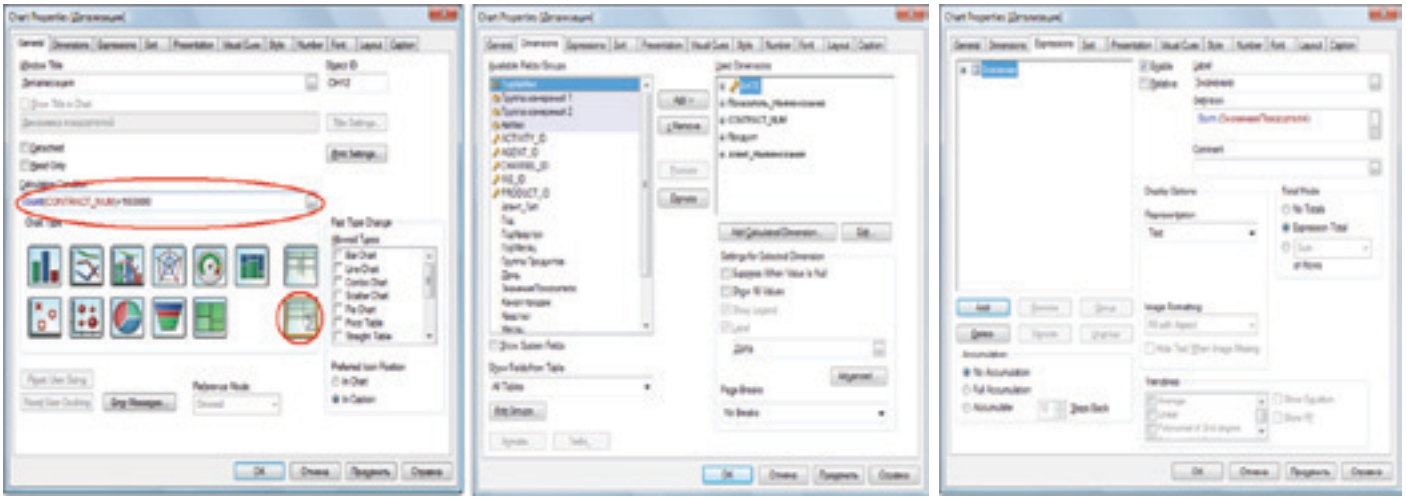


Рисунок 24.



лучается динамика, в данном случае, по группе продуктов и каждым цветом отображается своя группа продуктов.

Вот Пётр показывает (рис. 15), как пользователь может поменять измерение в циклической группе: по каналам, по типу агента, по типу страхователя. Сразу обе диаграммы переключаются по тем разрезам, которые мы выбрали. То есть, пользователь может оперативно переключать возможность анализа по различным разрезам.

Сейчас у нас динамика по годам. Теперь во второй циклической группе мы выбираем квартальный разрез, и видим квартальную динамику премии. Таким же образом можно видеть и месячную динамику. Меняем тип диаграммы - и вот у нас гистограмма, то есть, пользователь может оперативно поменять «график» на «гистограмму» и обратно.

Фильтрация данных непосредственно на графике - это очень удобно! Мы можем фильтровать данные прямо на графике и выделять интересующие нас области: сразу фильтрация идет по выбранным на графике периодам (рис. 18-23).

Следующая бизнес задача - это получение детализации.

Интересно, обнаружив какую-то закономерность или аномалию, посмотреть на неё более пристально. И конечно нам тут уже нужно спуститься до уровня конкретных договоров, конкретных транзакций. Для этого мы создаём простую таблицу. Говорим, что мы хотим показывать не более 100 тысяч записей. Как правило, нужно ограничивать для пользователей возможность выгружать большие объёмы детальных данных. Во-первых, детализация большого количества строк замедляет работу приложения, а во-вторых - есть соображения безопасности.

Мы задаём условие на количество выводимых элементов: оно не может превышать 100 тысяч записей. И пишем текст сообщения: «сделайте выбор, отражается не более100 тыс. записей».

Выведем следующие измерения: номер договора, наименование показателя, дату, далее выведем продукт, выведем агента, и достаточно. Мы суммируем значение показателя. Таким образом, мы показали, что у нас в детализации есть (рис. 24).

Рисунок 25.

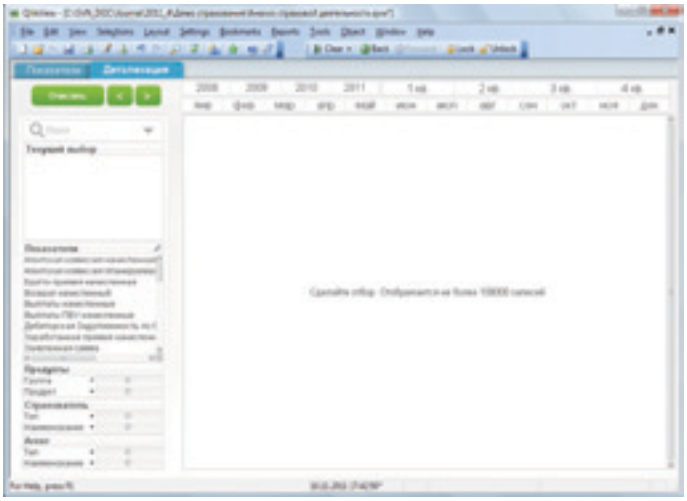


Рисунок 26.

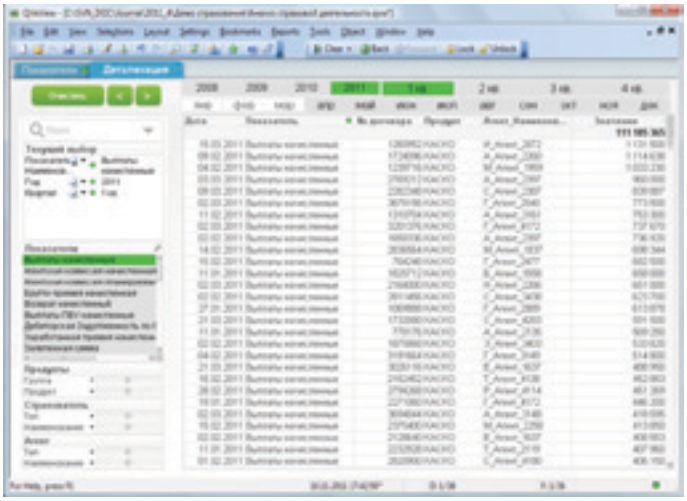


Рисунок 27.

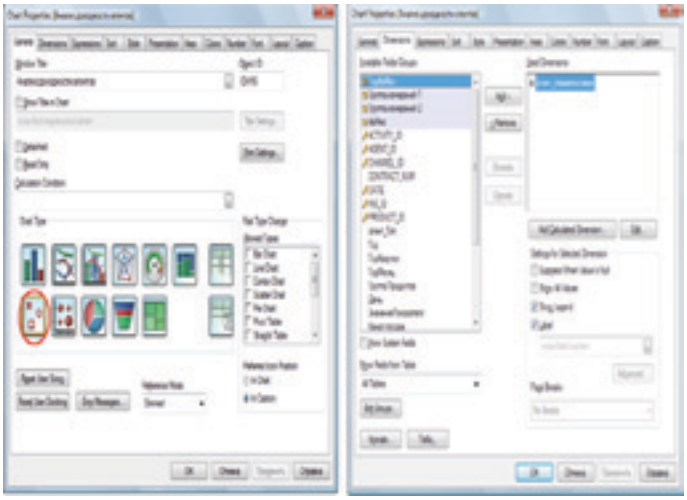
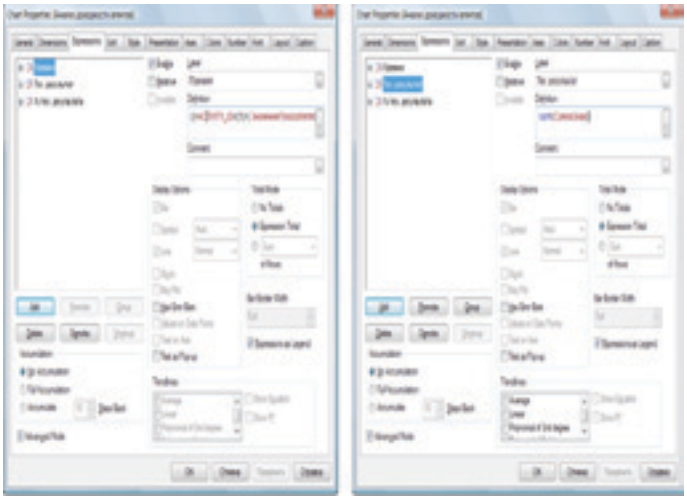


Рисунок 28.



Вопрос из зала

А 100 тысяч по какому критерию, просто первые 100 тысяч или там какие-то критерии есть?

Михаил Николаев

Нет. Если у нас получилось более 100 тысяч строк, мы просто не отображаем ничего. Нужно сделать выборку, чтобы уменьшить количество записей. Данные в таблице появятся только после того, как выборка уменьшит количество до порогового значения.

Сделаем какой-нибудь отбор, например, первый квартал 2011 года, а всё равно много данных получается. Выбираем показатель (выплаты) — количество записей уменьшилось, и мы видим детализацию. Дальше пользователь может задать сортировку по сумме и увидеть аномалию — например, самую крупную выплату (рис. 25-26).

Следующие бизнес задачи - это уже более продвинутый ана-

Рисунок 29.

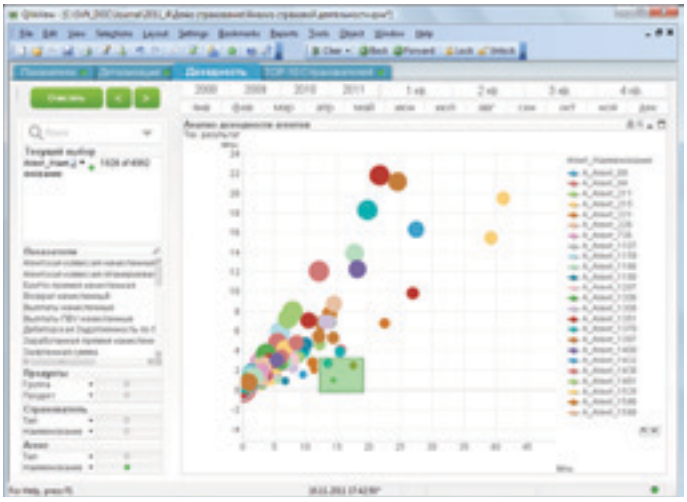


Рисунок 30.

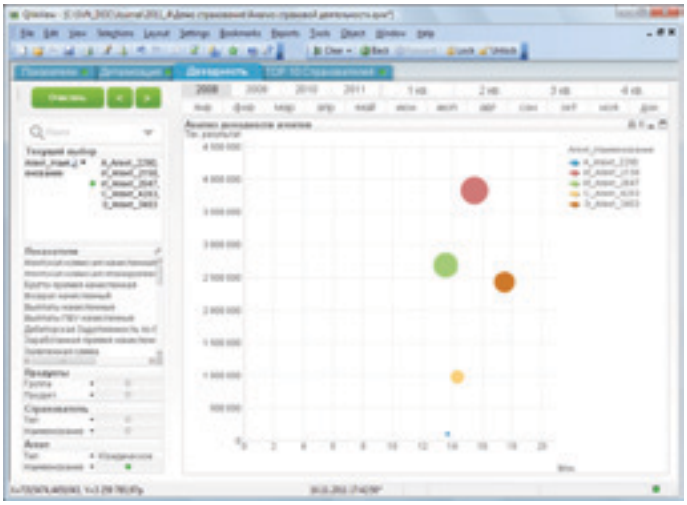


Рисунок 31.

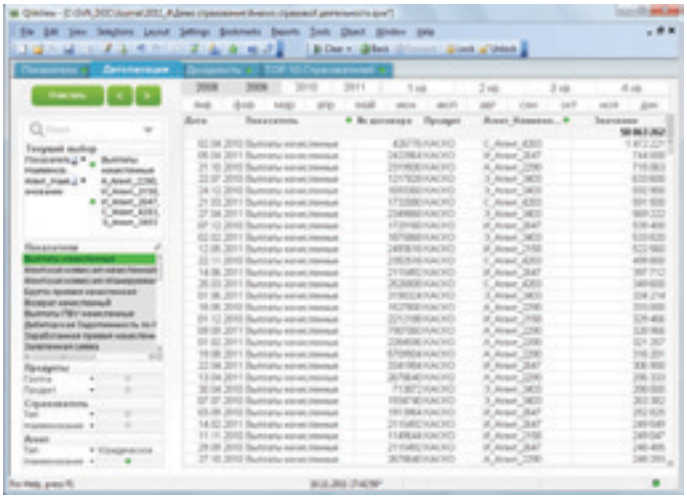
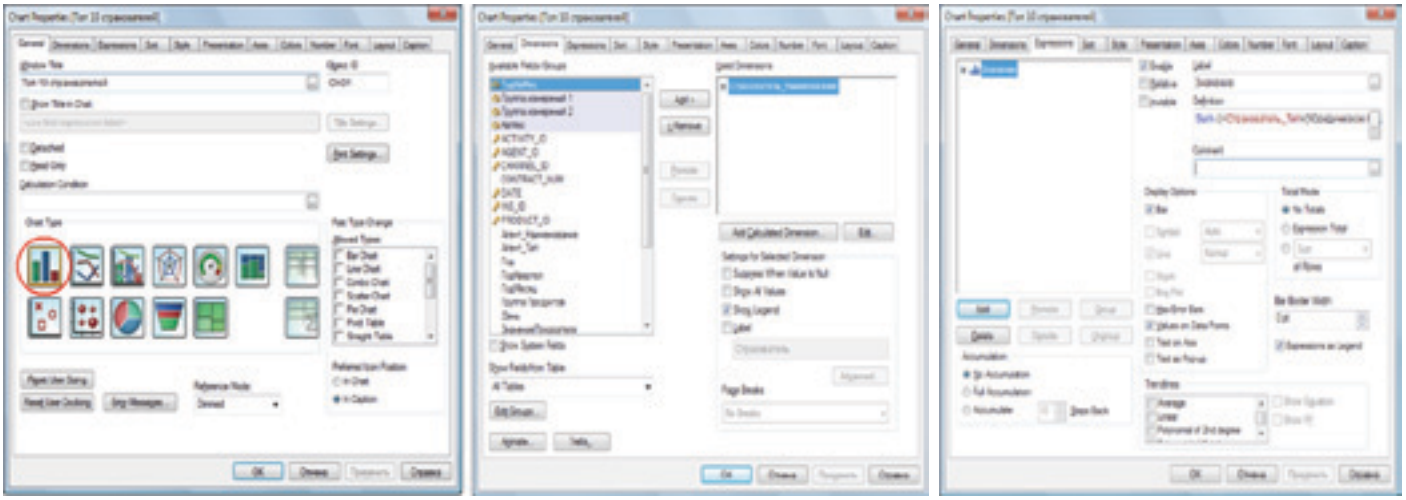


Рисунок 32.



лиз, и там будет как раз интересно обнаружить что-то, и увидеть детализацию.
Следующая бизнес задача - это анализ доходности.
Допустим, есть «плохие» агенты, есть «хорошие». Агенты могут приносить много премии, но при этом выплаты у них могут зашкаливать. Здесь мы увидим, насколько доходен каждый агент. Что мы посчитаем? Мы будем считать брутто премию начисленную, здесь мы явно поставим условие по показателю: код показателя, равный единице - это брутто премия начисленная. Премия у нас будет по оси «X», по оси «Y» у нас будет технический результат. Технический результат - это премия минус выплаты, минус возвраты и минус комиссия. То есть нужно сложить показатели, имеющие знак «+» и вычесть показатели, имеющие знак «-». Это будет просто сумма знаковой суммы: автоматически всё считается с правильным знаком. Третий фактор - это отношение технического результата к премии. Это у нас будет диаметр этого кружочка. В качестве измерения надо выбрать «Агента» (27-28).
У нас будет график, в котором каждому агенту будет соответствовать кружок на графике. По оси «X» отложена премия, собранная агентом, по оси «Y» - отложен его технический результат, а диаметр кружка - это отношение технического результата к премии. Чем отношение больше, тем агент результативней, чем меньше, тем он хуже. Соответственно, у нас есть луч от начала координат, идущий под определённым углом, все, что ниже это не очень хороший результат, все, что выше это хорошие агенты.
Выделяем область на графике, автоматически будут выбраны только выделенные агенты, и анализируем их более подробно. Вот у нас конкретные агенты, у них можно посмотреть детализацию. Переключились на страницу детализации - вот эти агенты. Выделим, допустим, выплаты, и увидим выплаты конкретных агентов, то есть вот они - нехорошие агенты (рис. 29-31).

Пётр Травкин

При этом мы ещё видим, что при переходе с вкладки на вкладку фильтр у нас сохраняется и мы, таким образом, данные можем отфильтровать на одной закладке, а потом перейти на другую, и там

Рисунок 33.

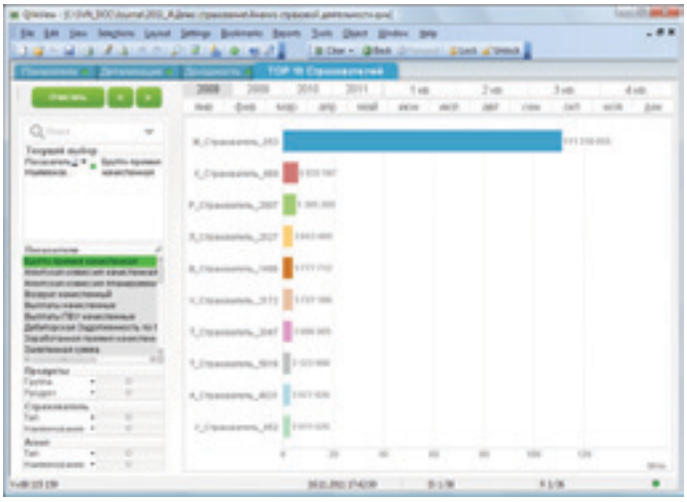


Рисунок 34.

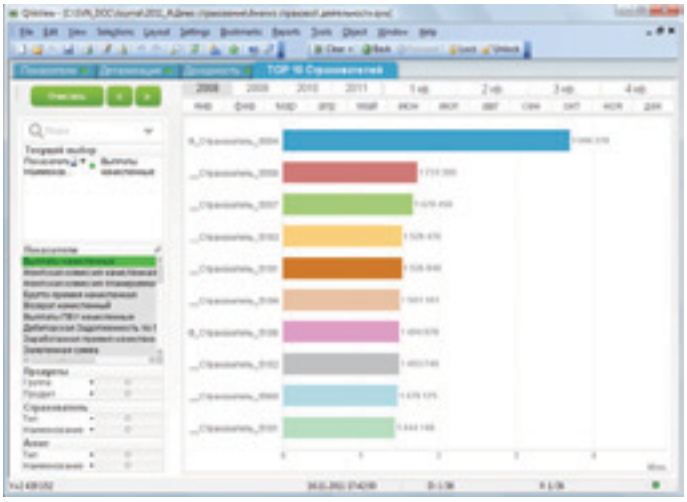


Рисунок 35.

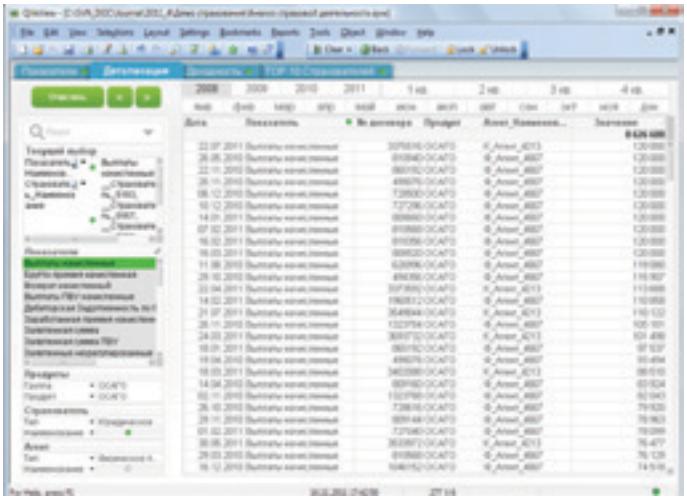
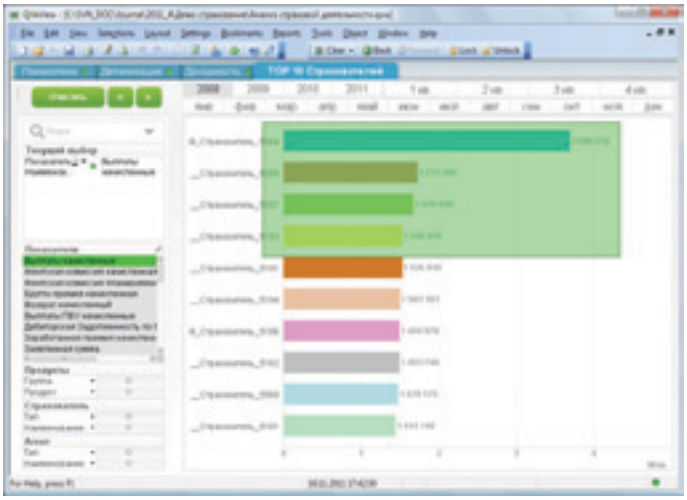


Рисунок 36.



уже посмотреть совершенно другую информацию.

Михаил Николаев

Фильтрация действует везде по всему документу, это очень удобно. Сделали отбор на одной вкладке, а на другой вкладке уже видим отобранные значения детально. Вот такая интересная вещь. Вроде бы очень простая модель, но уже даже из очень простой модели можно интересные закономерности наблюдать.

Следующая бизнес-задача - это топ 10 страхователей.

Интересно рассмотреть какие страхователи являются лидерами. Создаём новую диаграмму, выбираем тип графика, чтобы у нас была гистограмма, далее выбираем измерение «страхователь-наименование», суммируем значение показателя и говорим, что будем исследовать только «Юр. Лиц»: ставим условие, что тип страхователя – это юридическое лицо (рис. 32). И этот анализ покажет нам лидеров. Выбрали показатель «брутто премия». Видим 10 первых страхователей по премии.

Рисунок 37.

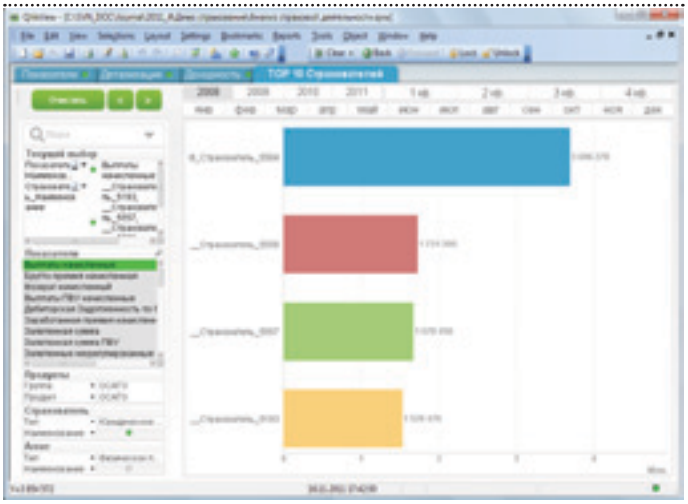


Рисунок 38.

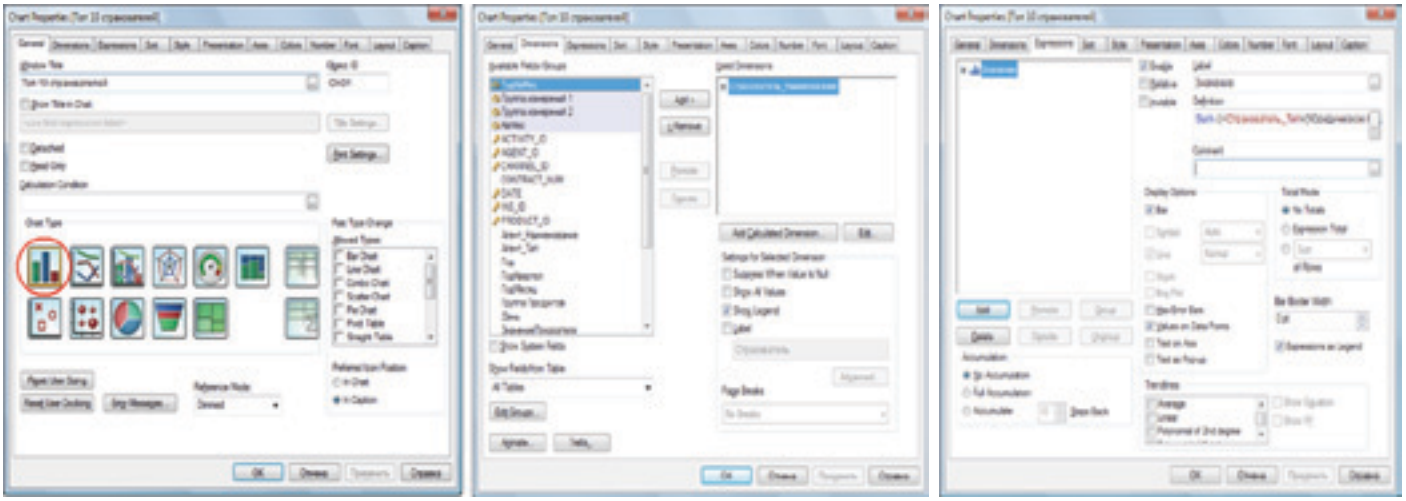


Рисунок 39.

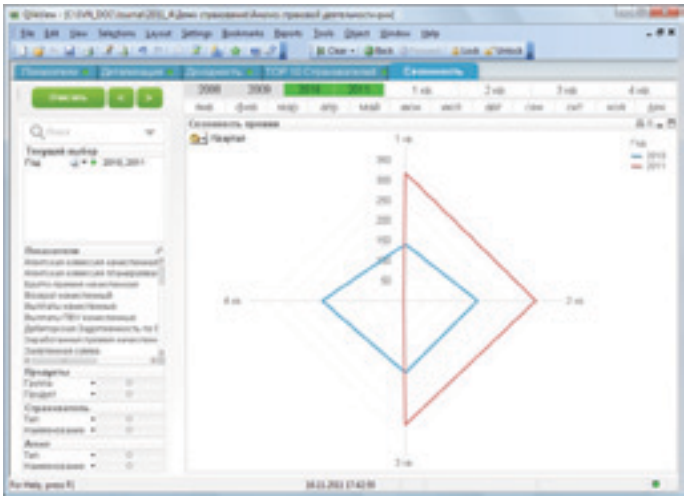
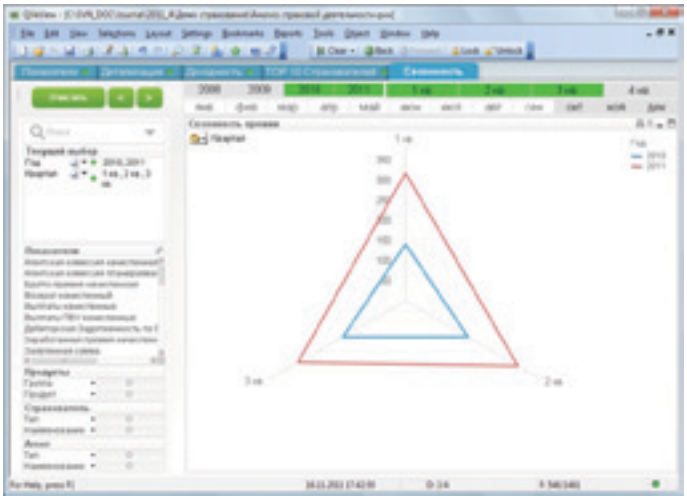


Рисунок 40.



А теперь выберем выплаты, и видим, что это другие страхователи. Теперь можно более подробно посмотреть, что это за страхователи и что это за выплаты. Опять выделяем область на графике и возвращаемся на вкладку детализации. Вот эти выплаты, то, что интересно. То есть увидели, какую-то аномалию, закономерность, сделали фильтрацию по графику, и тут же видим детализацию. На детализации мы тоже можем осуществлять выбор (рис. 33-36).

Ещё одна любопытная вещь - это поиск, например, по какому-нибудь разрезу. Вводим несколько символов имени агента и видим всех агентов, содержащих в своём имени введённый набор символов (рис. 37).

Последняя бизнес задача — анализ сезонности премии.

Создаём диаграмму - вот такая паутинистая диаграмма будет. По полярным осям будет временное измерение, здесь мы будем анализировать данные по месяцам и по кварталам. Выбираем возможность переключаться. Суммируем значение показателя, при том, что сам показатель - это премия начисленная (рис. 38).

Мы сразу одновременно на графике можем видеть сезонность 2011 года и 2010 года, вот она отличается. Правда за 2011 у нас нет последнего квартала, и если мы выберем только первые три квартала, мы уже увидим более адекватные цифры. И можно оперативно переключаться в рамках циклической группы: по кварталам или по месяцам (рис. 39-42).

Коллеги, ещё хочется сказать, что у нас есть на стенде более подробный пример, с большим количеством аналитических разрезов, там и пролонгация есть, и всякие другие интересные разрезы. Если есть интерес - подходите, можем посмотреть какие-то конкретные бизнес примеры.

Спасибо.

Андрей Педоренко

Коллеги спасибо. На мой взгляд, это уникальный опыт, впервые, наверное, на конференции. Демонстрация создания приложения в режиме реального времени. И при этом коллеги полностью уложились в заявленное время.

Рисунок 41.

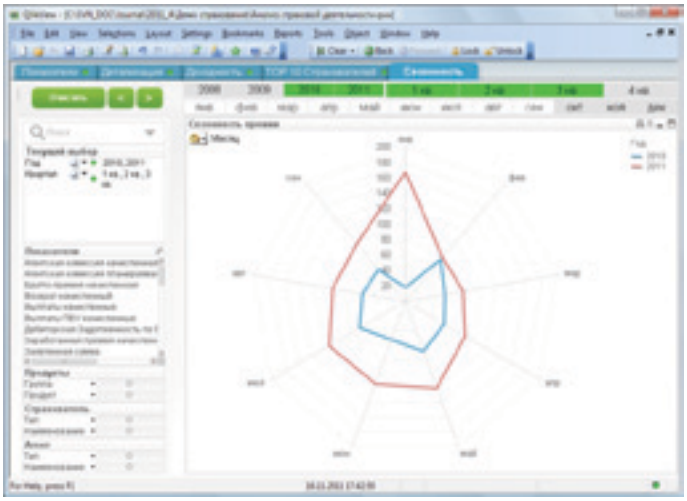
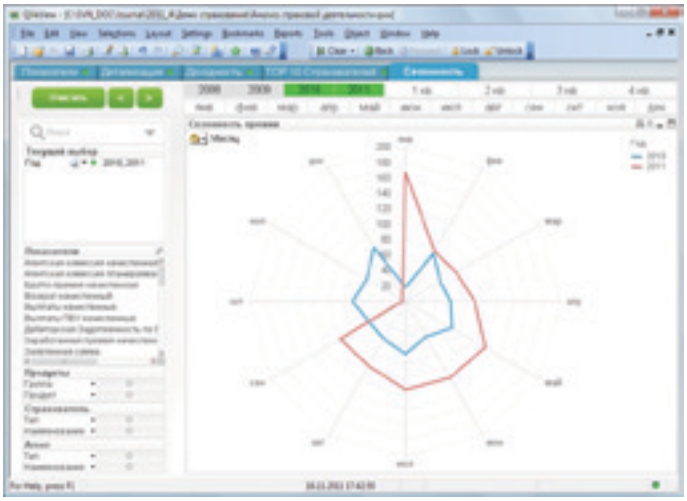



Рисунок 42.