Николай Паклин

Демократизация Big Data

BIG DATA ФЕСТИВАЛЬ

БАНК ЦЕНТР-ИНВЕСТ

Loginom Company (ex. BaseGroup Labs)

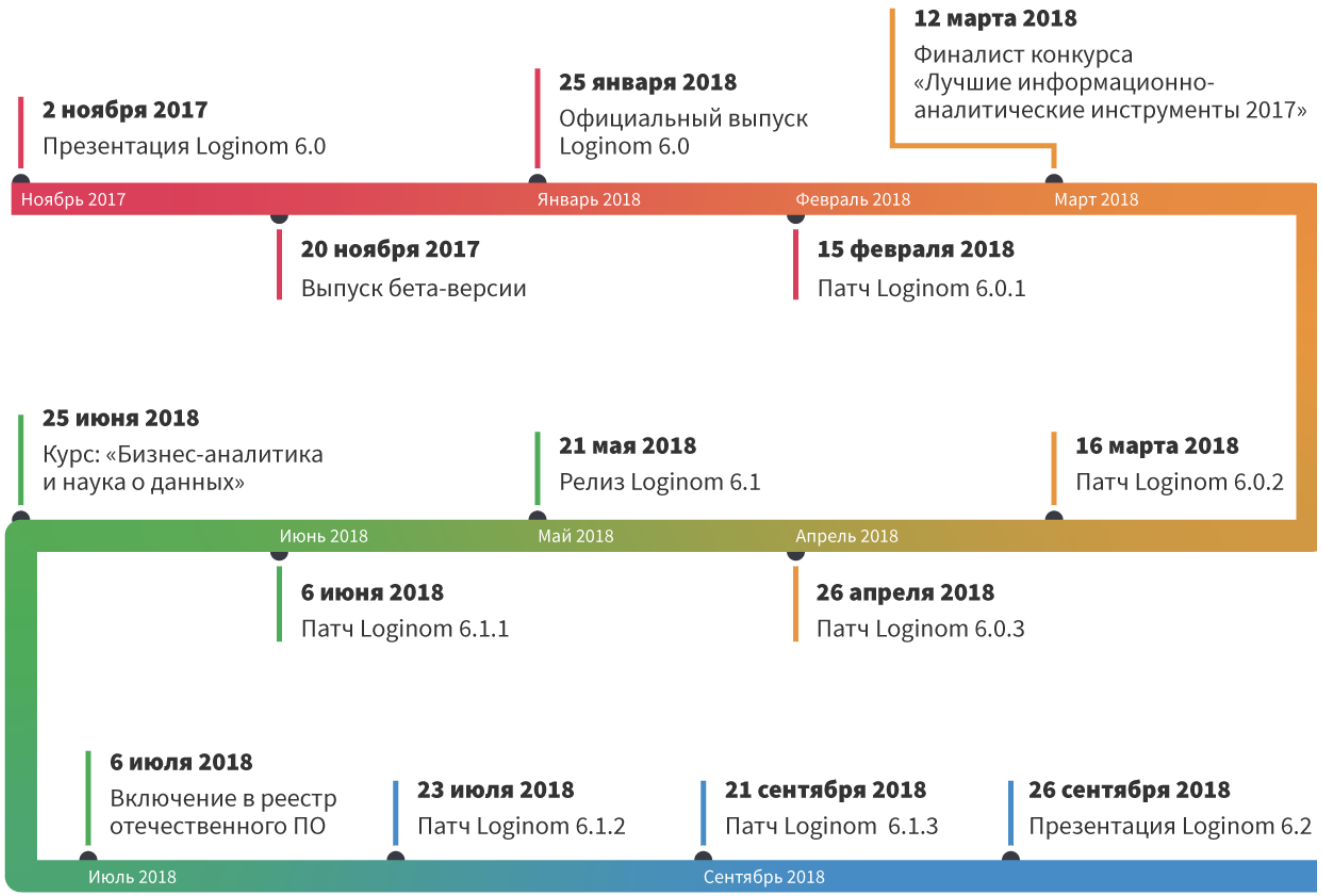
- Специализация: анализ данных
- На рынке с 1995 года
- 150+ аналитических проектов
- 100+ вузов-партнеров

- Аналитическая платформа Deductor – предшественник Loginom
- Книга «Бизнес-аналитика: от данных к знаниям» (2008-2013 гг.)

Loginom

Сделать анализ данных массовой технологией

Развитие платформы Loginom



Нет единого способа, используются все подходы

О чем доклад

О рынке инструментов для анализа данных

Как сделать анализ данных доступным массовому потребителю

Особенности рынка сегодня

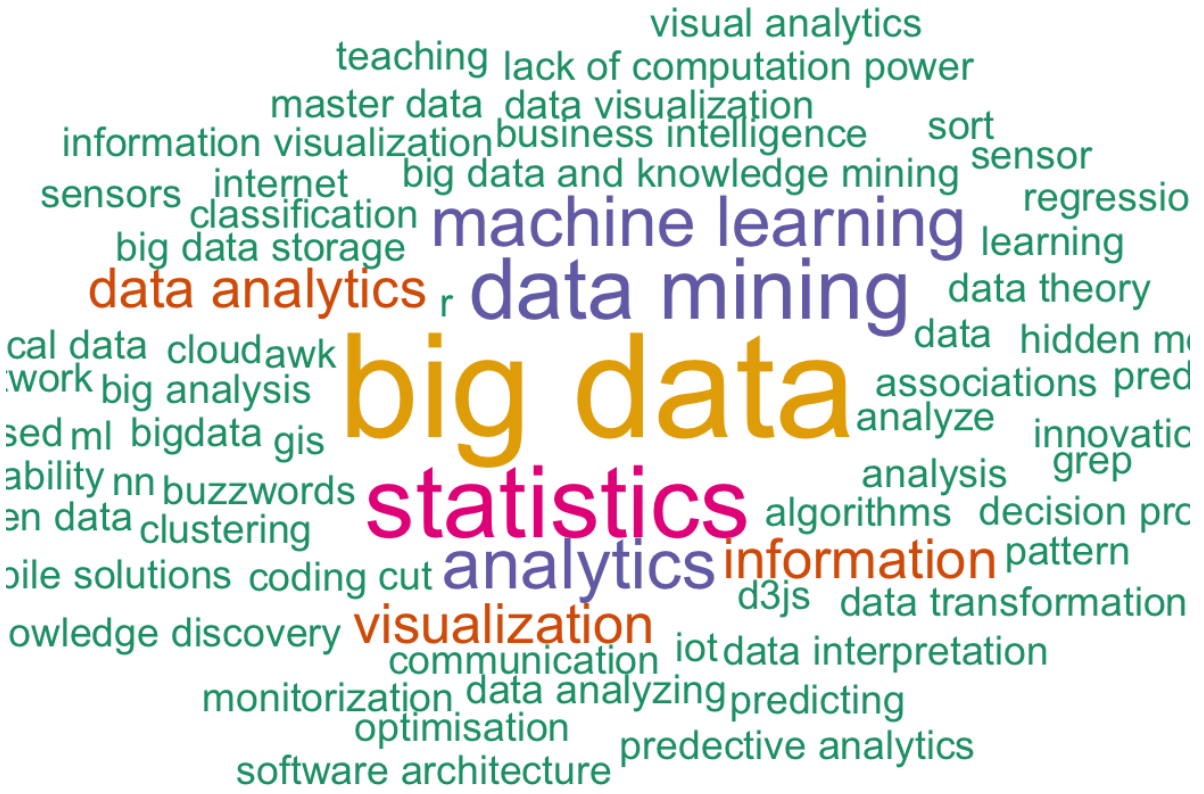
- Крупные игроки разогревают рынок
- Много обещаний и лукавств
- Нехватка кадров и высокие зарплаты
- Распространение open-source инструментов
- Рост аутсорсинга аналитики

Data driven…

everything:

данные – основа принятия всех решений

Термины разные – обещания одни и те же



Сегодня

- ~ 2 млн. исследователей данных по всему миру
- 150 тыс. свободных вакансий только в США

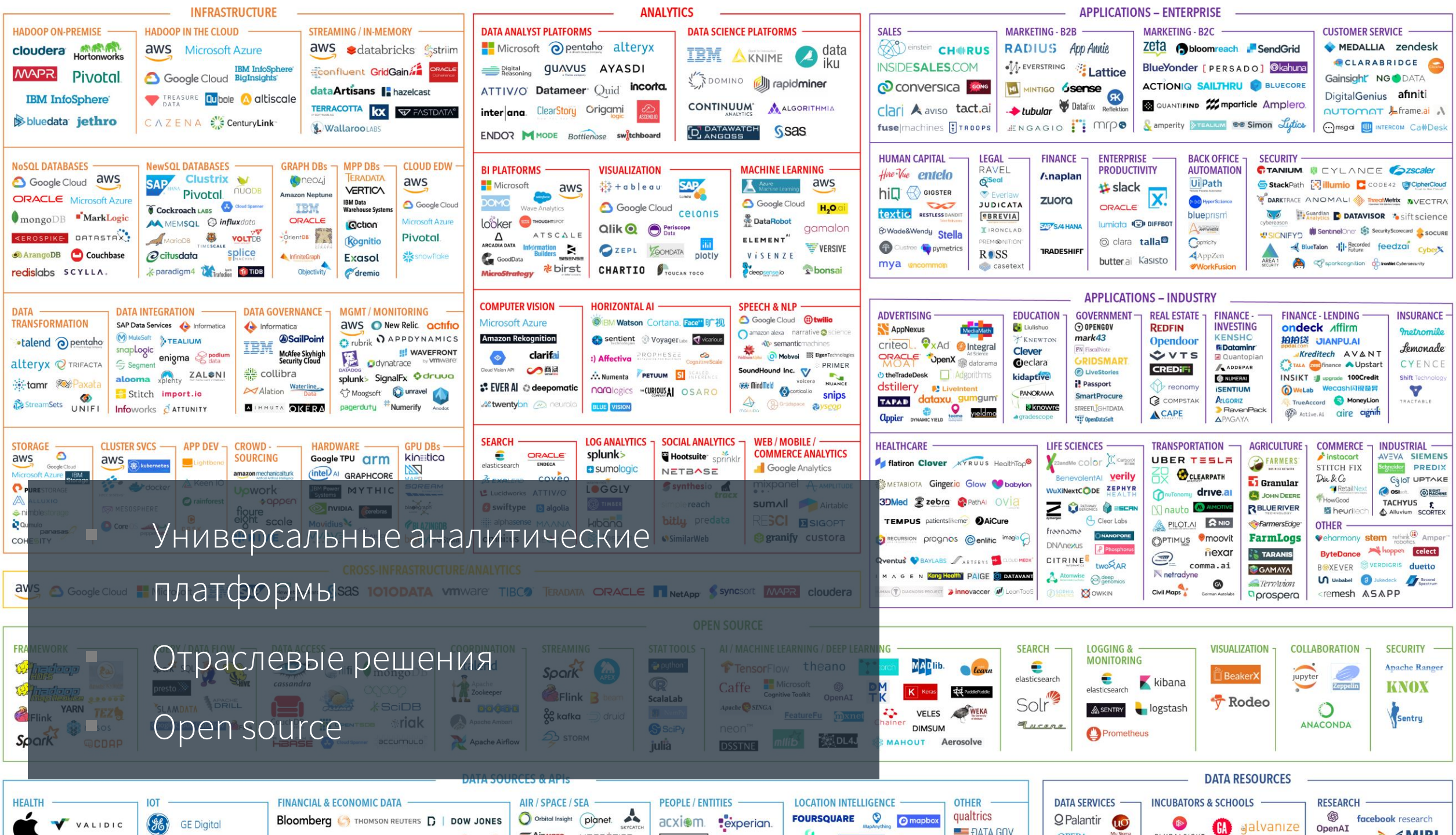
По данным kdnuggets.com 08/2018

Завтра

- Десятки миллионов исследователей данных по всему миру*
- Объем рынка РФ 28 млрд. руб. (2020 г.), мировой – 300 $ млрд.**

*Прогноз Статистического управления Минтруда США

**Оценки на Russian AI Forum 2017 и J.P. Morgan, McKinsey

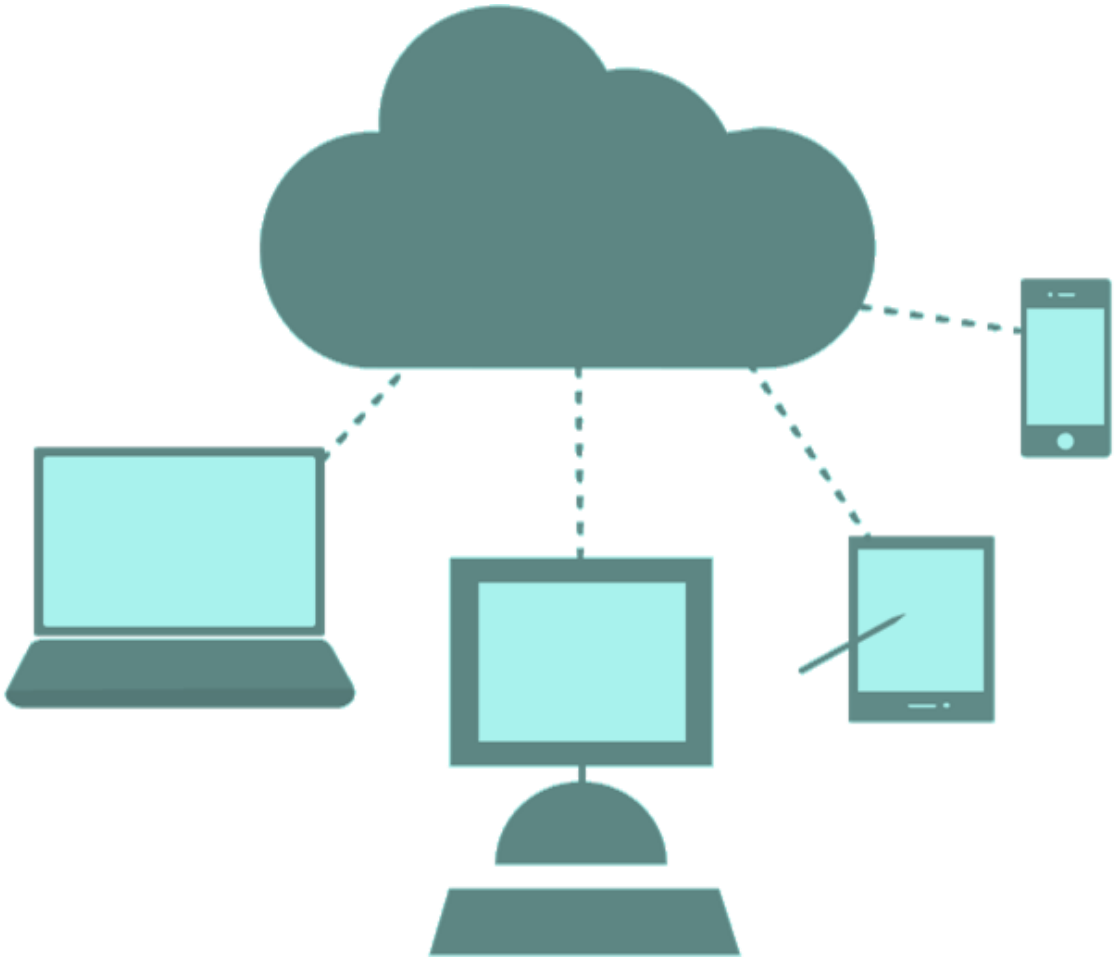


Универсальные аналитические платформы

Отраслевые решения

Open source

Аутсорсинг аналитики – растущий рынок



Как сделать аналитику массовой?

Демократизация аналитики:

1. Анализ бизнес-пользователями
2. Работа со множеством источников
3. Свободное исследование данных
4. Быстрая проверка гипотез
5. Независимость от IT департамента

СОВРЕМЕННЫЙ DATA SCIENTIST

МАТЕМАТИКА И СТАТИСТИКА
- Машинное обучение
- Статистическое моделирование
- Планирование эксперимента
- Байесовский вывод
- Обучение с учителем: деревья принятия решений, Random forests, логистическая регрессия
- Обучение без учителя: кластерный анализ, понижение размерности
- Оптимизация: градиентный спуск и варианты

ПРЕДМЕТНАЯ ОБЛАСТЬ И SOFT SKILLS
- Понимание и интерес к бизнесу
- Интерес к данным
- Неформальное лидерство
- Хакерское мышление
- Умение решать проблемы
- Умение мыслить стратегически, проактивность, креативность, инновационный подход, готовность к сотрудничеству

ПРОГРАММИРОВАНИЕ И БАЗЫ ДАННЫХ
- Базовые знания в компьютерных науках
- Скриптовый язык, например, Python
- Специализированные статистические инструменты, например, R
- Базы данных SQL и NoSQL
- Реляционная алгебра
- Параллельные системы баз данных и параллельная обработка запросов
- Понимание MapReduce Hadoop и Hive/Pig
- Опыт в xaaS-сервисах (инфраструктура-как-сервис), например, в Amazon Web Services

КОММУНИКАЦИЯ И ВИЗУАЛИЗАЦИЯ
- Умение общаться с топ-менеджментом
- Навыки сторителлинга
- Умение превратить инсайты в управленческие решения и конкретные действия
- Визуальный дизайн
- Пакеты R — ggplot, lattice
- Знание инструментов визуализации — например, Flare, D3.js, Tableau

Основная проблема – кадровый голод


Радиус окружности показывает относительный объем индустрии
Индекс легкости захвата данных
Высокий
Низкий
Коммунальные услуги
Поставщики медицинских услуг
Природные ресурсы
Информация
Производство
Финансы и страхование
Профессиональные услуги
Компьютеры и электронные изделия
Транспорт и складирование
Управление отходами
Управление компаниями
Недвижимость и аренда
Строительство
Оптовая торговля
Предоставление услуг проживания и питания
Розничная торговля
Прочие услуги
Услуги в области образования
Искусство и развлечения
Правительство
Низкий
Высокий
Сравнительный индекс потенциала применения технологии Big Data

Программист

- Знает языки программирования
- Разбирается в математических методах
- Может формализовать логику

VS

Бизнес-эксперт

- Знает предметную область
- Может поставить задачу
- Может интерпретировать результаты

Эксперт-предметник – ключевое лицо, способное извлечь из данных знания.

Данные

Информация

Визуализация

Знания

Требования к инструменту

1. Визуальное проектирование
2. Повторное использование наработок
3. Минимум кодирования
4. Высокая производительность

Визуальное проектирование

diamond-sizes.Rmd

```
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---

```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)

smaller <- diamonds %>%
  filter(carat <= 2.5)
```

We have data about `r nrow(diamonds)` diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are larger than
2.5 carats. The distribution of the reminder is shown
below:

```{r, echo = FALSE}
smaller %>%
  ggplot(aes(carat)) +
  geom_freqpoly(binwidth = 0.01)
```
```

8:17 Chunk 1: setup R Markdown

Console R Markdown

~/Documents/r4ds/r4ds/rmarkdown/

```
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Environment History Git

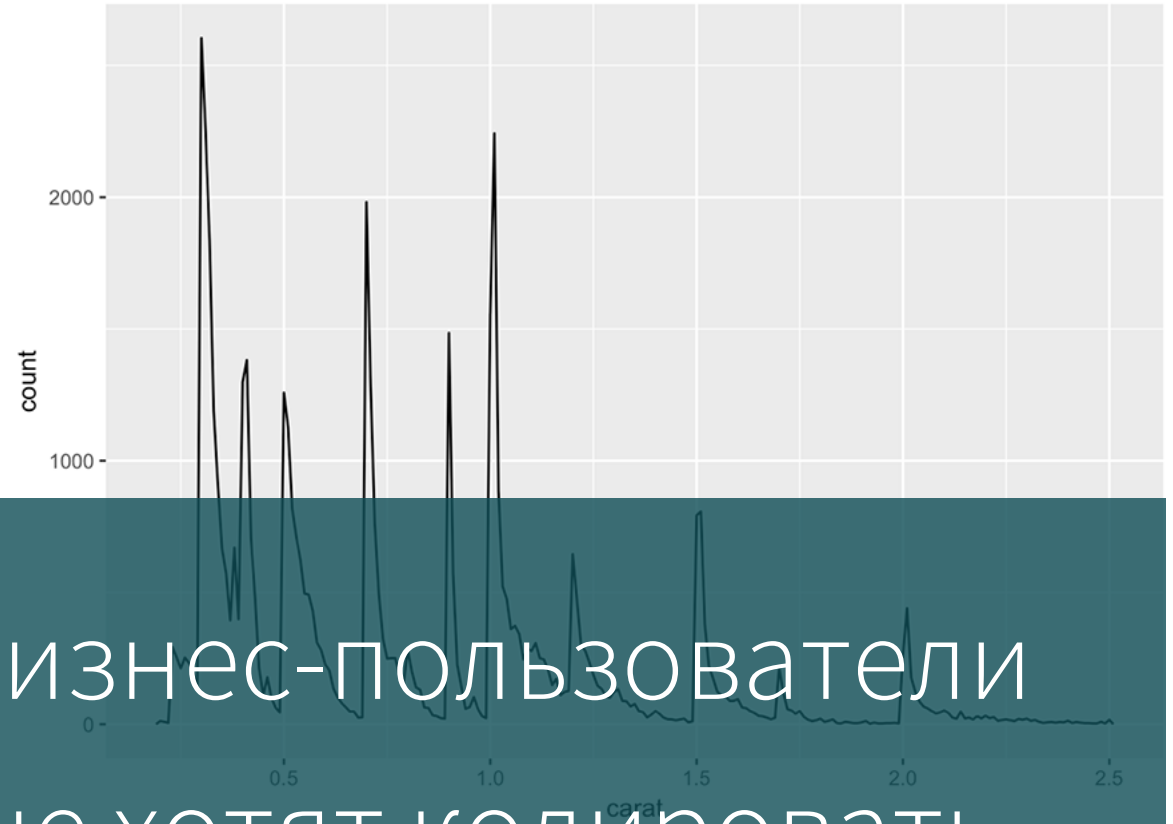
Files Plots Packages Help Viewer

Publish

Diamond sizes

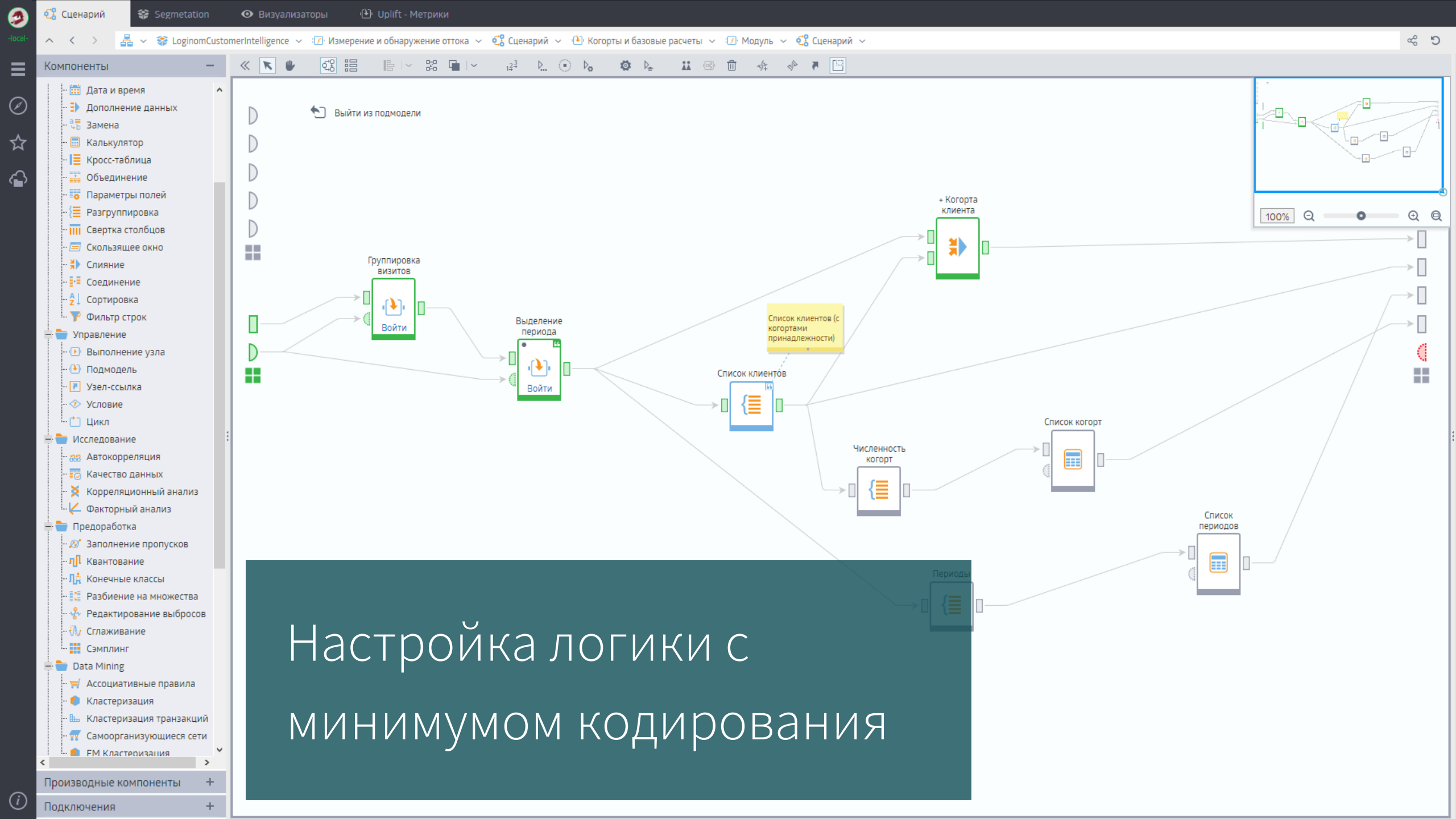
2016-08-25

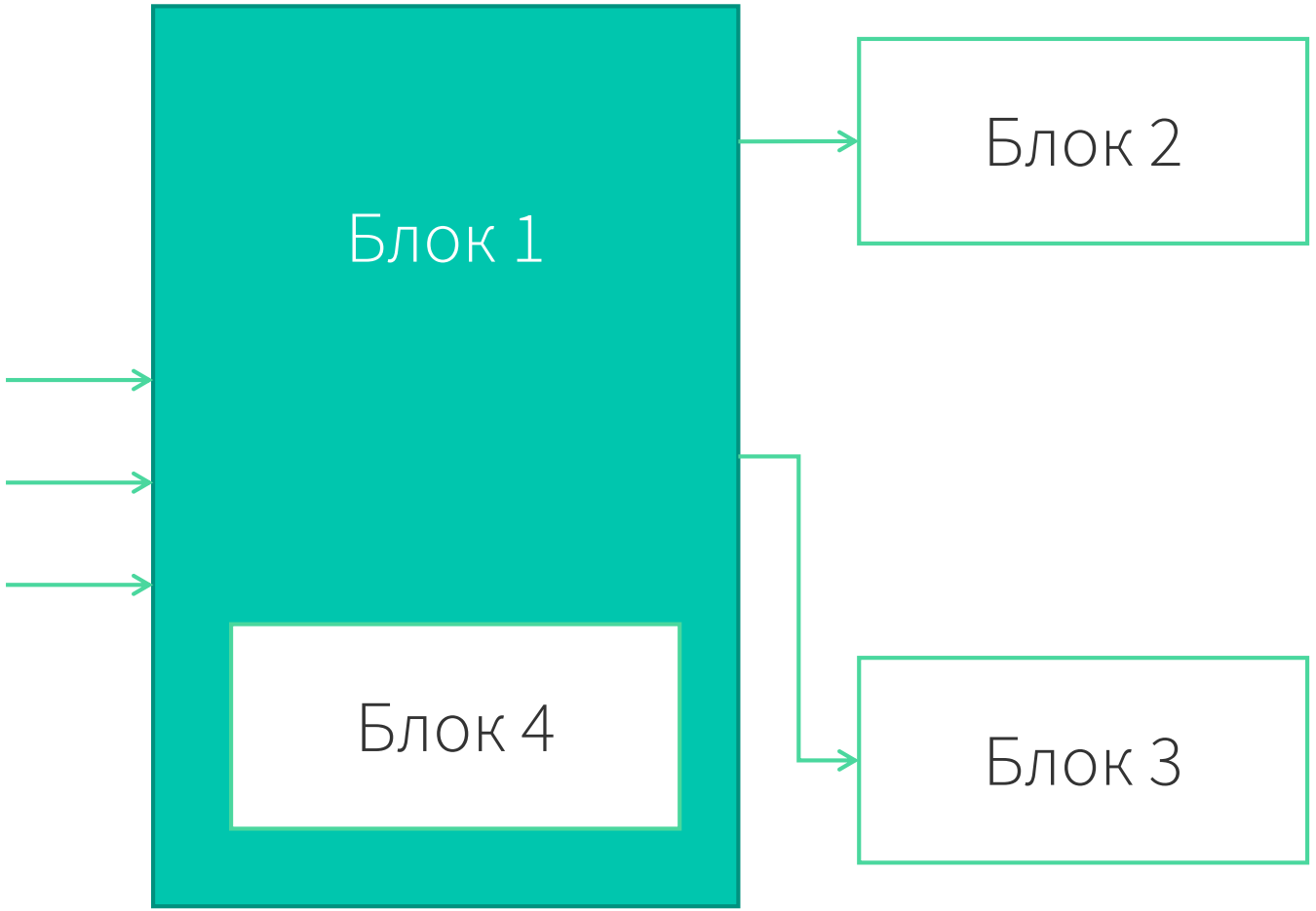
We have data about 53940 diamonds. Only 126 are larger than 2.5 carats. The distribution of the reminder is shown below:



Бизнес-пользователи не хотят кодировать

Схемы – естественное описание логики анализа


Сценарий
Segmetation
Визуализаторы
Uplift - Метрики
LoginomCustomerIntelligence
Измерение и обнаружение оттока
Сценарий
Когорты и базовые расчеты
Модуль
Сценарий
Компоненты
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Управление
Выполнение узла
Подмодель
Узел-ссылка
Условие
Цикл
Исследование
Автокорреляция
Качество данных
Корреляционный анализ
Факторный анализ
Предобработка
Заполнение пропусков
Квантование
Конечные классы
Разбиение на множества
Редактирование выбросов
Сглаживание
Сэмплинг
Data Mining
Ассоциативные правила
Кластеризация
Кластеризация транзакций
Самоорганизующиеся сети
EM Кластеризация
Производные компоненты
Подключения
Выйти из подмодели
Группировка визитов
Войти
Выделение периода
Войти
+ Когорта клиента
Список клиентов (с когортами принадлежности)
Список клиентов
Численность когорт
Список когорт
Список периодов
Периоды
100%
Настройка логики с минимумом кодирования



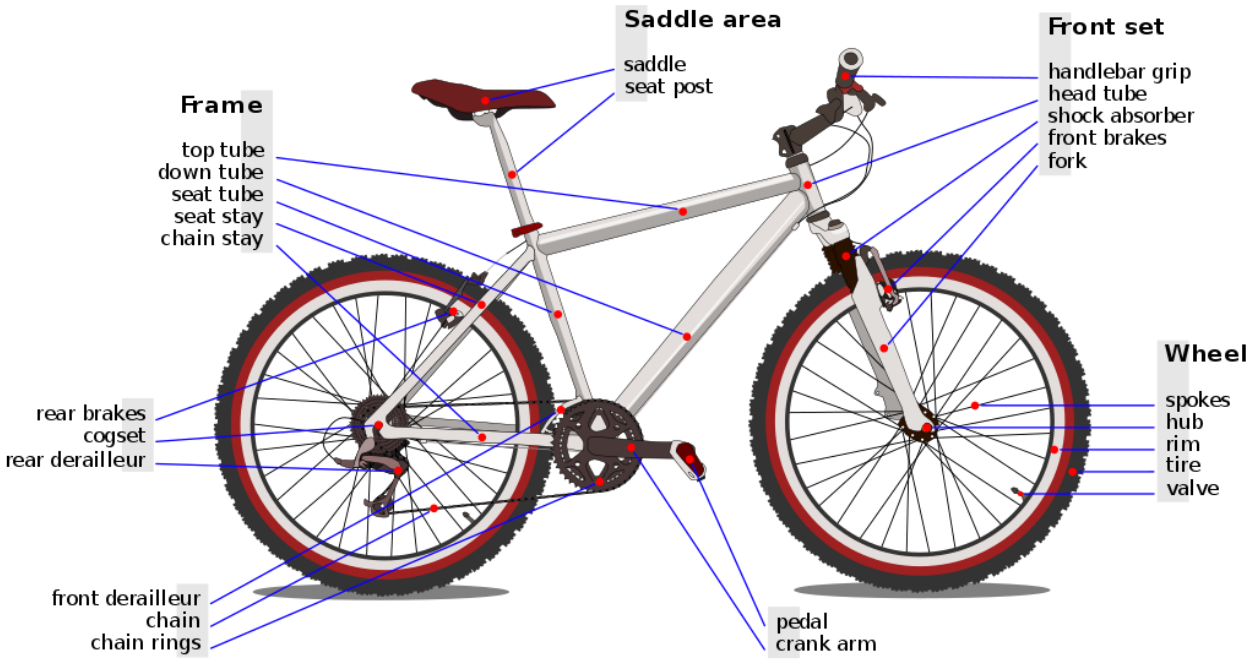
- Декомпозиция задачи
- Проектирование сверху-вниз
- Нет привязки к данным

Минимум кодирования

Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms

4 из 5 инструментов
в квадрате лидеров
это Low-code
платформы

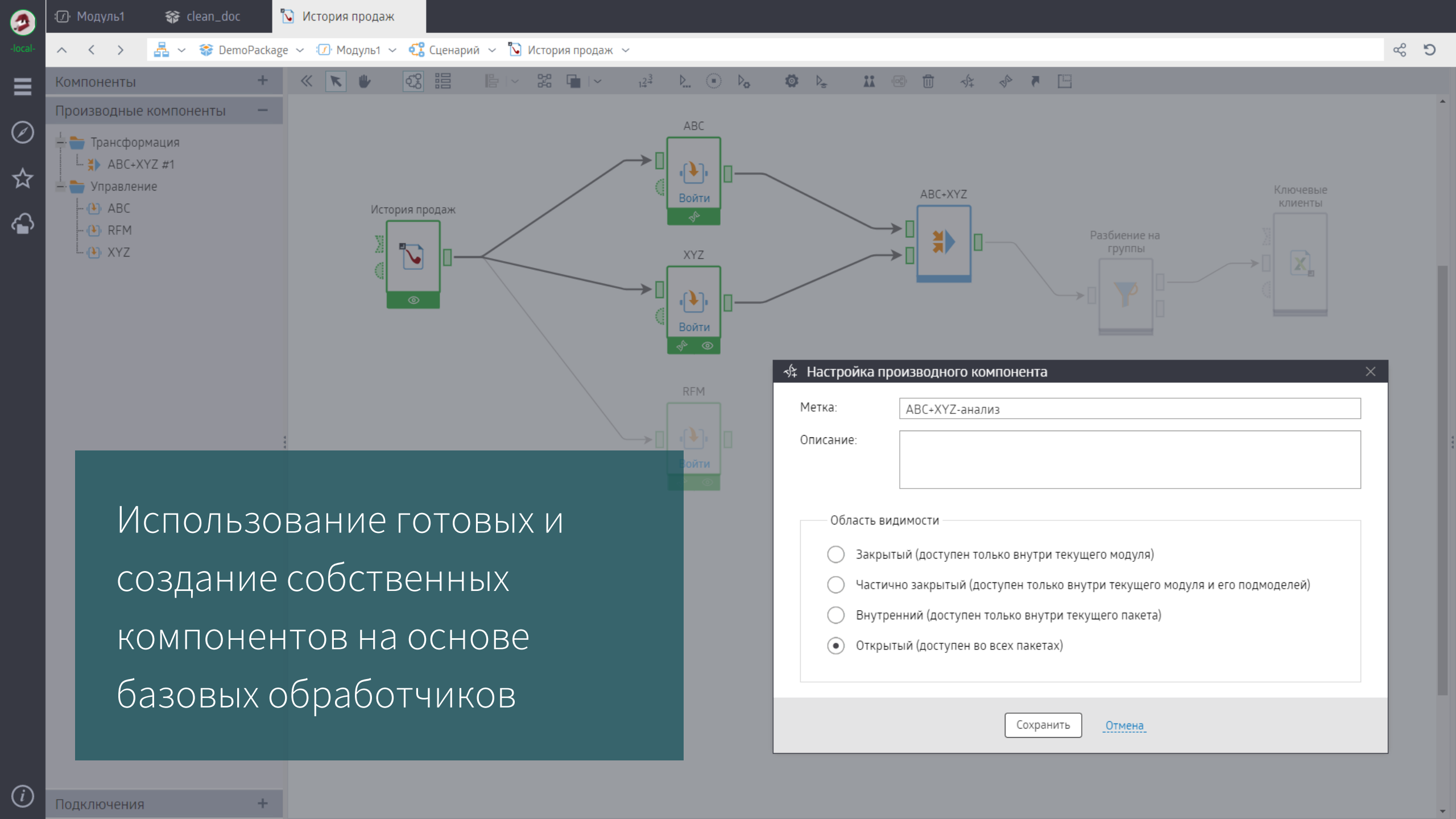
Повторное использование



В любом проекте есть типовые модули. Зачем изобретать велосипед при наличии готовых компонентов?

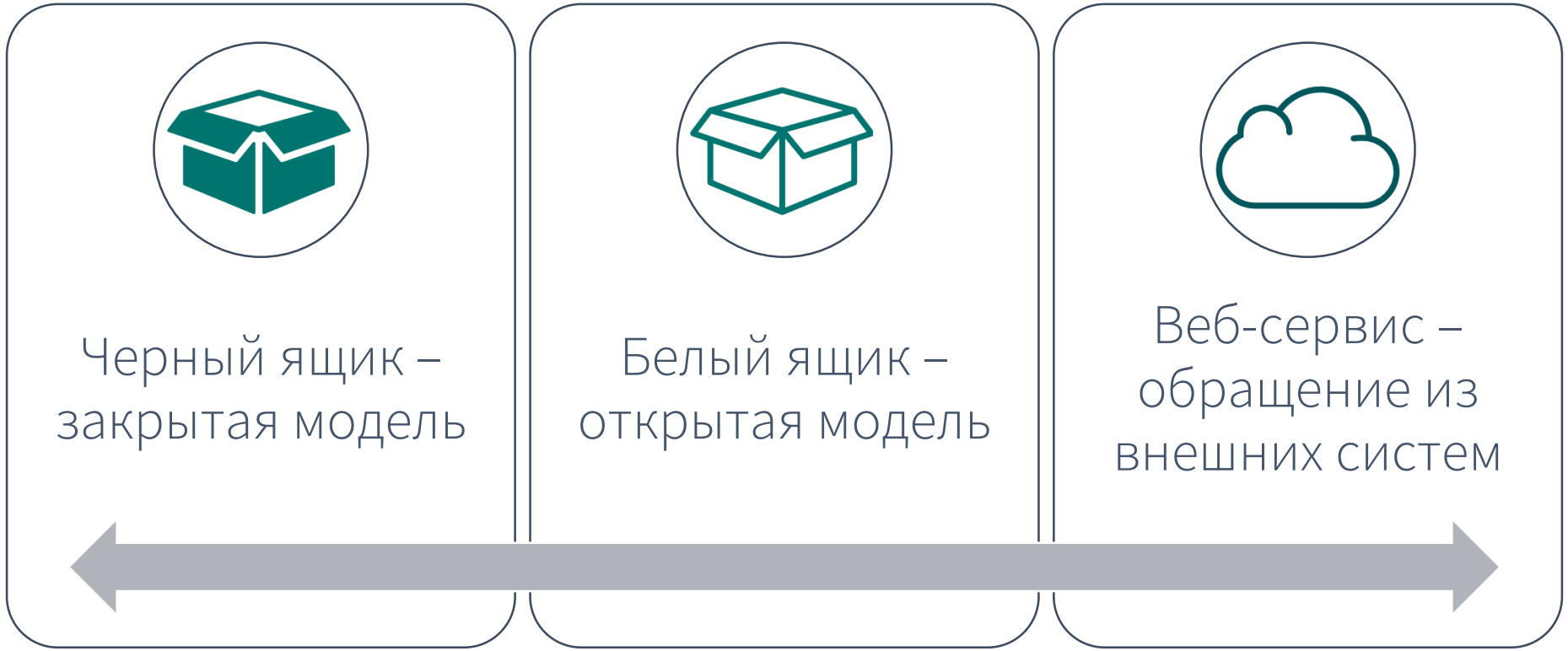
Повторное использование – способ сокращения времени проекта

1. Готовые компоненты
2. Подготовленные данные
3. Веб-сервисы



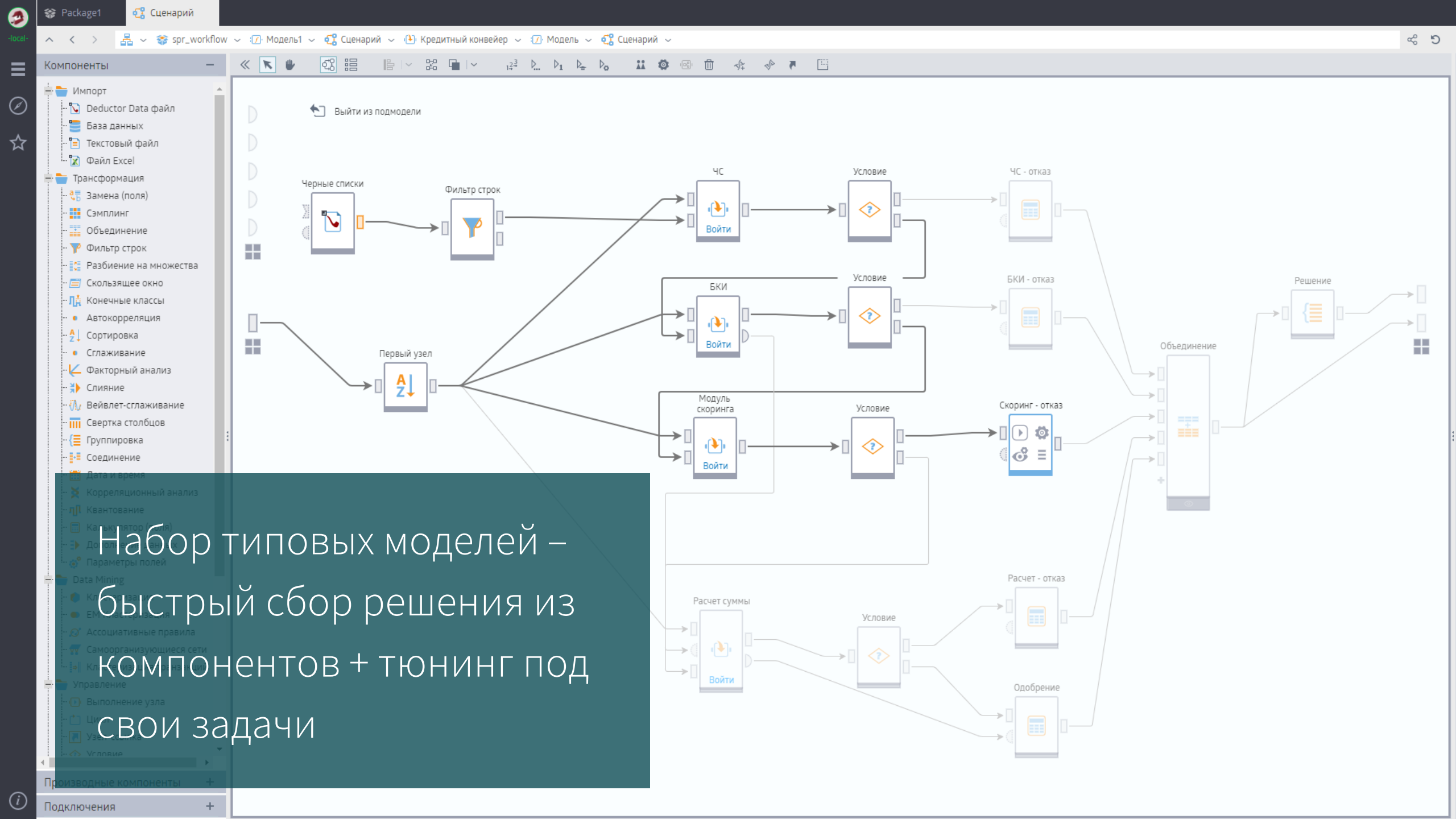
Использование готовых и создание собственных компонентов на основе базовых обработчиков

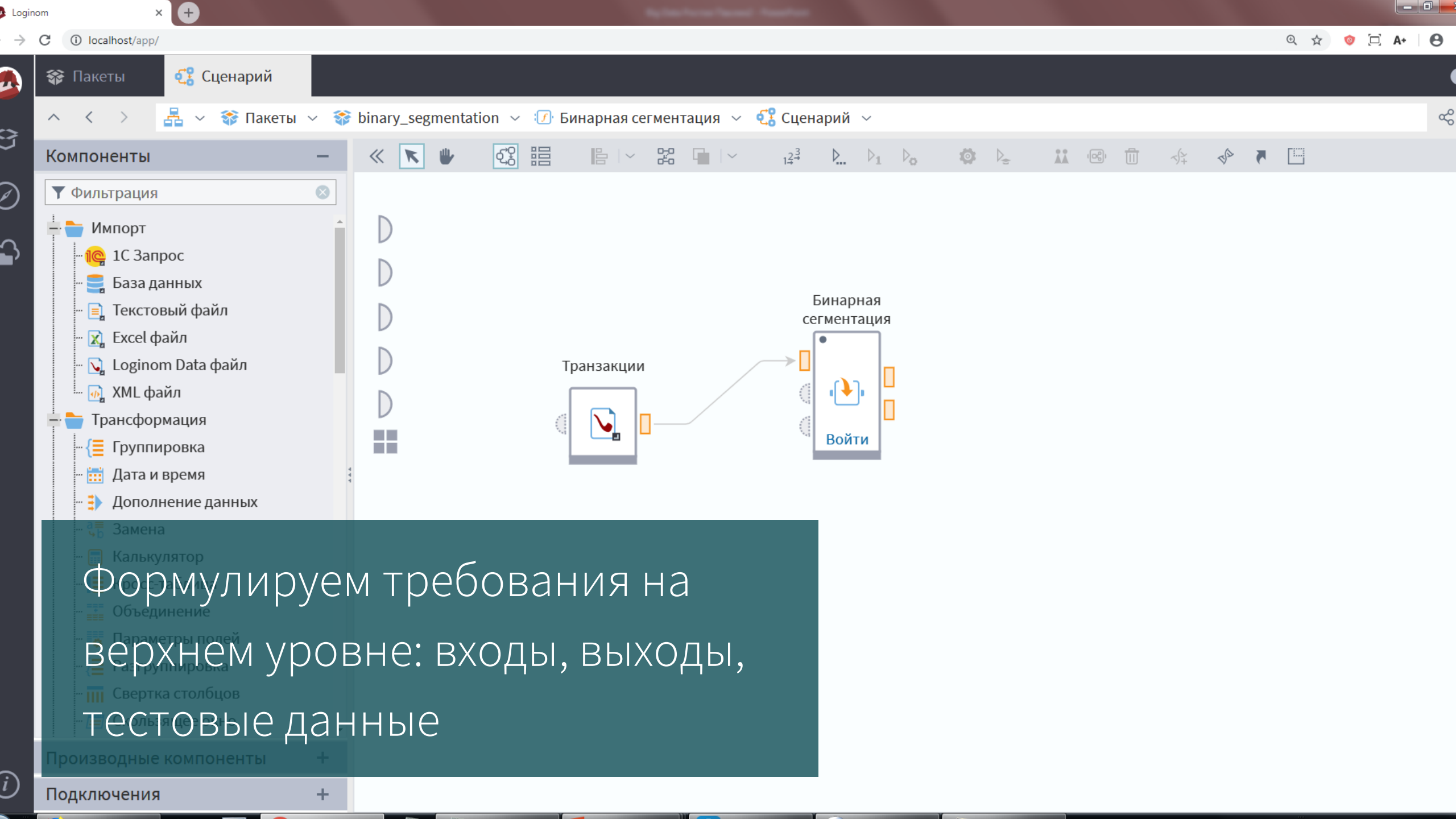
Варианты использования компонента

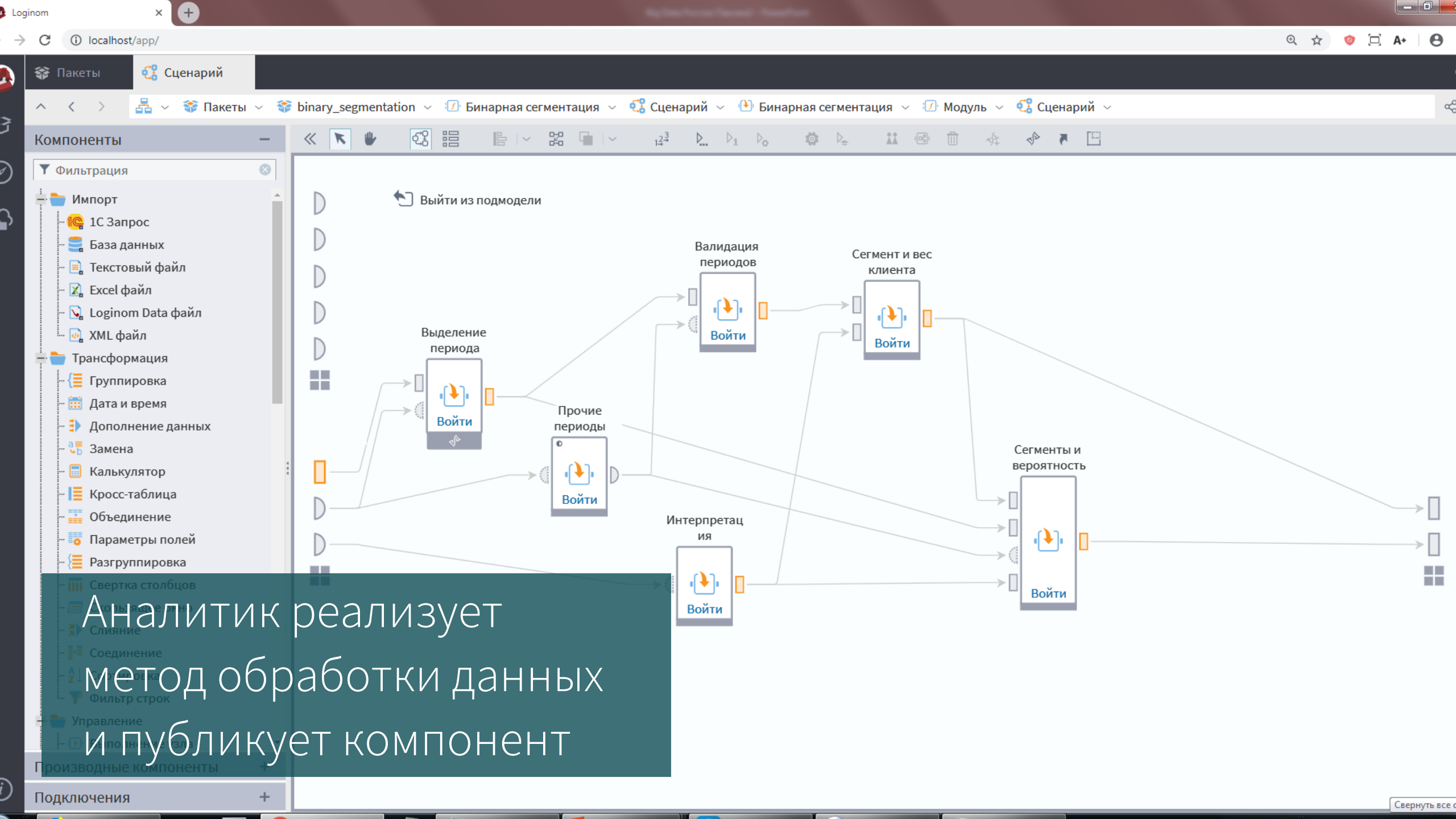


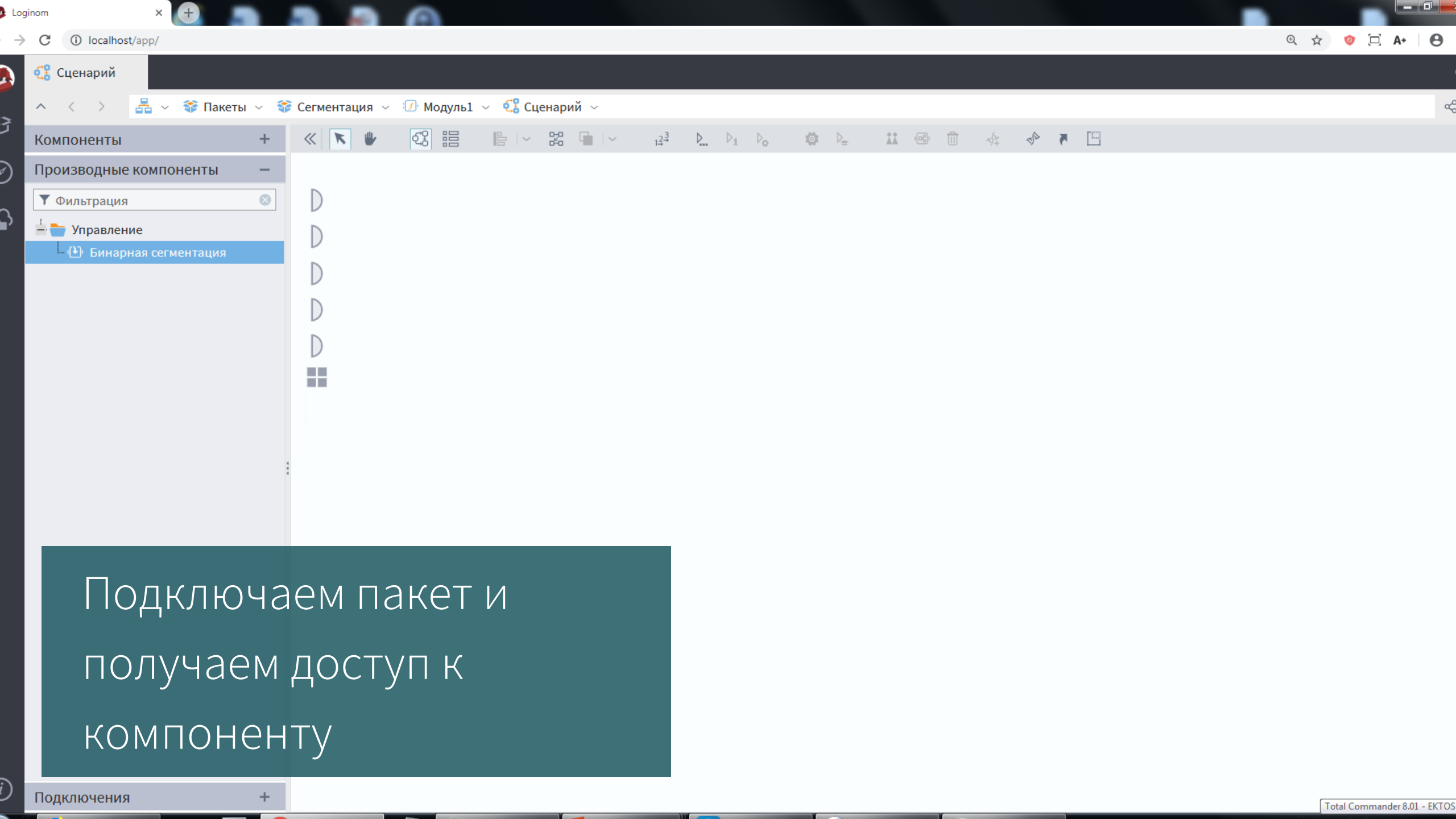
Объектная модель

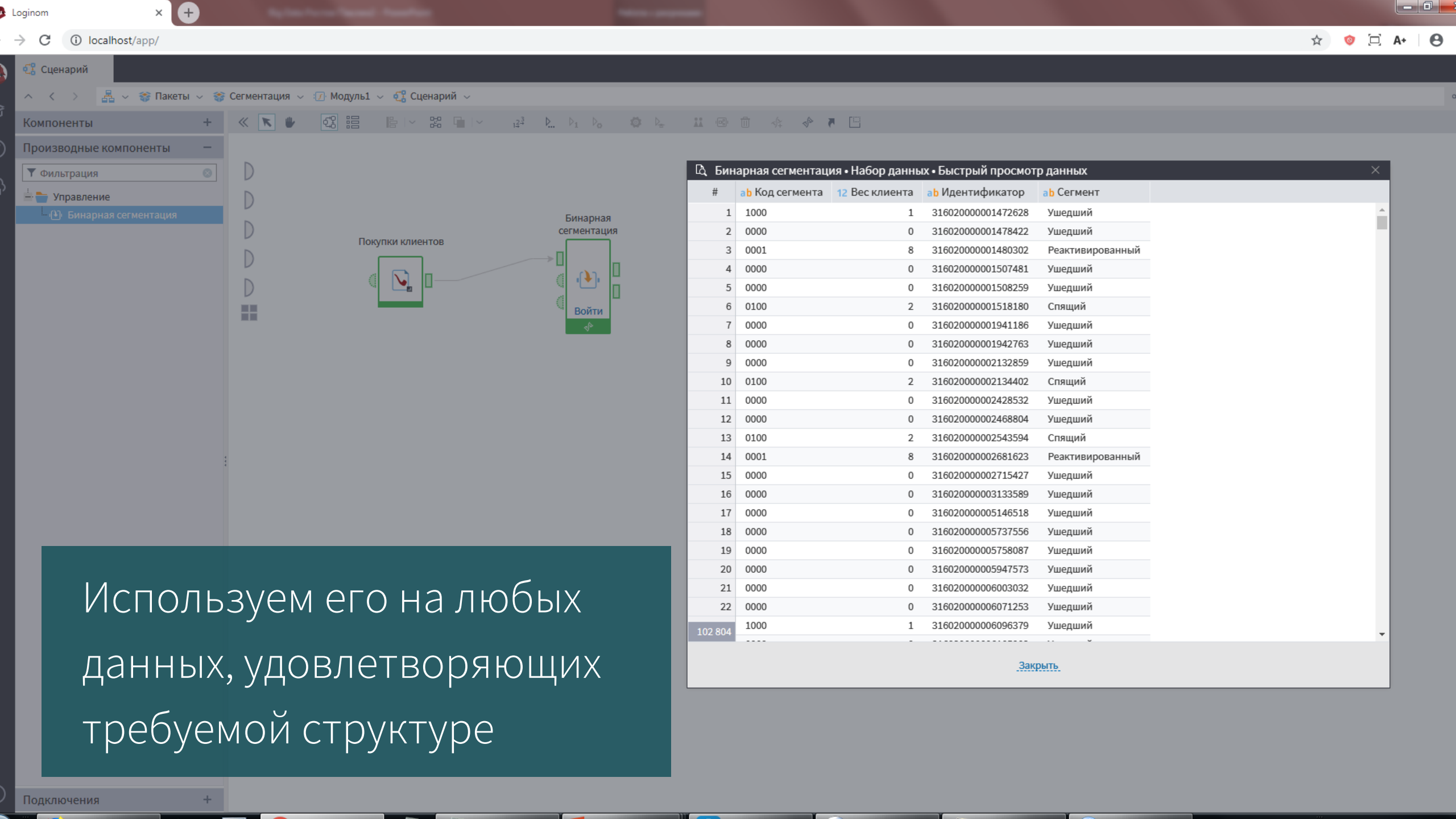
Принцип	Назначение
Абстракция	Возможность оперировать блоком обработки данных как единым целым, не вдаваясь в особенности реализации
Инкапсуляция	Включение в модель как логики обработки, так и скрытых от внешнего пользователя данных
Наследование	Создание узла-наследника на основе существующего с заимствованным у узла-родителя функционалом
Полиморфизм	Изменение в наследнике логики обработки или данных для адаптации к новому применению


Package1
Сценарий
spr_workflow
Модель1
Сценарий
Кредитный конвейер
Модель
Сценарий
Компоненты
Импорт
Deductor Data файл
База данных
Текстовый файл
Файл Excel
Трансформация
Замена (поля)
Сэмплинг
Объединение
Фильтр строк
Разбиение на множества
Скользящее окно
Конечные классы
Автокорреляция
Сортировка
Сглаживание
Факторный анализ
Слияние
Вейвлет-сглаживание
Свертка столбцов
Группировка
Соединение
Дата и время
Корреляционный анализ
Квантование
Параметры полей
Data Mining
Ассоциативные правила
Самоорганизующиеся сети
Управление
Выполнение узла
Производные компоненты
Подключения
Выйти из подмодели
Черные списки
Фильтр строк
ЧС
Войти
Условие
ЧС - отказ
БКИ
Войти
Условие
БКИ - отказ
Первый узел
Модуль скоринга
Войти
Условие
Скоринг - отказ
Объединение
Решение
Расчет суммы
Войти
Условие
Расчет - отказ
Одобрение
Набор типовых моделей – быстрый сбор решения из компонентов + тюнинг под свои задачи

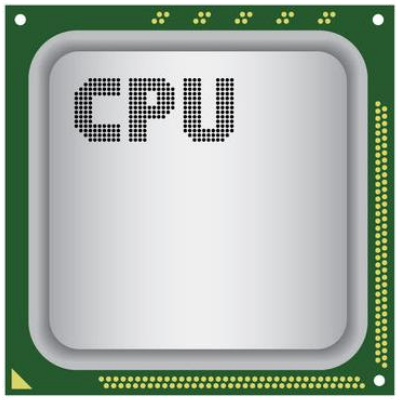

Loginom
localhost/app/
Пакеты
Сценарий
Пакеты
binary_segmentation
Бинарная сегментация
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Объединение
Свертка столбцов
Производные компоненты
Подключения
Транзакции
Бинарная сегментация
Войти
Формулируем требования на верхнем уровне: входы, выходы, тестовые данные


Loginom
localhost/app/
Пакеты
Сценарий
Пакеты
binary_segmentation
Бинарная сегментация
Сценарий
Бинарная сегментация
Модуль
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Слияние
Соединение
Фильтр строк
Управление
Производные компоненты
Подключения
Выйти из подмодели
Выделение периода
Войти
Валидация периодов
Войти
Сегмент и вес клиента
Войти
Прочие периоды
Войти
Интерпретация
Войти
Сегменты и вероятность
Войти
Аналитик реализует
метод обработки данных
и публикует компонент
Свернуть все

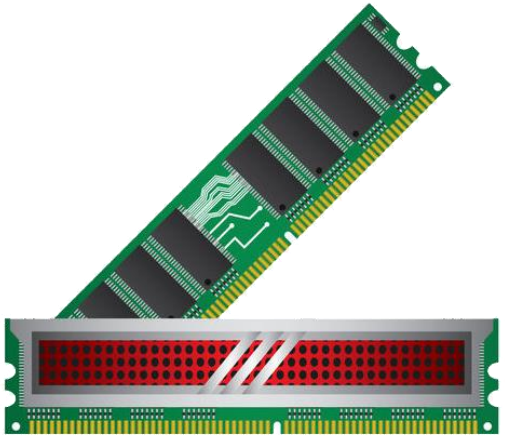

Loginom
localhost/app/
Сценарий
Пакеты
Сегментация
Модуль1
Сценарий
Компоненты
Производные компоненты
Фильтрация
Управление
Бинарная сегментация
Подключаем пакет и получаем доступ к компоненту
Подключения
Total Commander 8.01 - EKTOS


Loginom
localhost/app/
Сценарий
Пакеты
Сегментация
Модуль1
Сценарий
Компоненты
Производные компоненты
Фильтрация
Управление
Бинарная сегментация
Покупки клиентов
Бинарная сегментация
Войти
Подключения
Бинарная сегментация • Набор данных • Быстрый просмотр данных
| Код сегмента | Вес клиента | Идентификатор | Сегмент
1 | 1000 | 1 | 316020000001472628 | Ушедший
2 | 0000 | 0 | 316020000001478422 | Ушедший
3 | 0001 | 8 | 316020000001480302 | Реактивированный
4 | 0000 | 0 | 316020000001507481 | Ушедший
5 | 0000 | 0 | 316020000001508259 | Ушедший
6 | 0100 | 2 | 316020000001518180 | Спящий
7 | 0000 | 0 | 316020000001941186 | Ушедший
8 | 0000 | 0 | 316020000001942763 | Ушедший
9 | 0000 | 0 | 316020000002132859 | Ушедший
10 | 0100 | 2 | 316020000002134402 | Спящий
11 | 0000 | 0 | 316020000002428532 | Ушедший
12 | 0000 | 0 | 316020000002468804 | Ушедший
13 | 0100 | 2 | 316020000002543594 | Спящий
14 | 0001 | 8 | 316020000002681623 | Реактивированный
15 | 0000 | 0 | 316020000002715427 | Ушедший
16 | 0000 | 0 | 316020000003133589 | Ушедший
17 | 0000 | 0 | 316020000005146518 | Ушедший
18 | 0000 | 0 | 316020000005737556 | Ушедший
19 | 0000 | 0 | 316020000005758087 | Ушедший
20 | 0000 | 0 | 316020000005947573 | Ушедший
21 | 0000 | 0 | 316020000006003032 | Ушедший
22 | 0000 | 0 | 316020000006071253 | Ушедший
102 804 | 1000 | 1 | 316020000006096379 | Ушедший
Закрыть
Используем его на любых данных, удовлетворяющих требуемой структуре

Высокая производительность

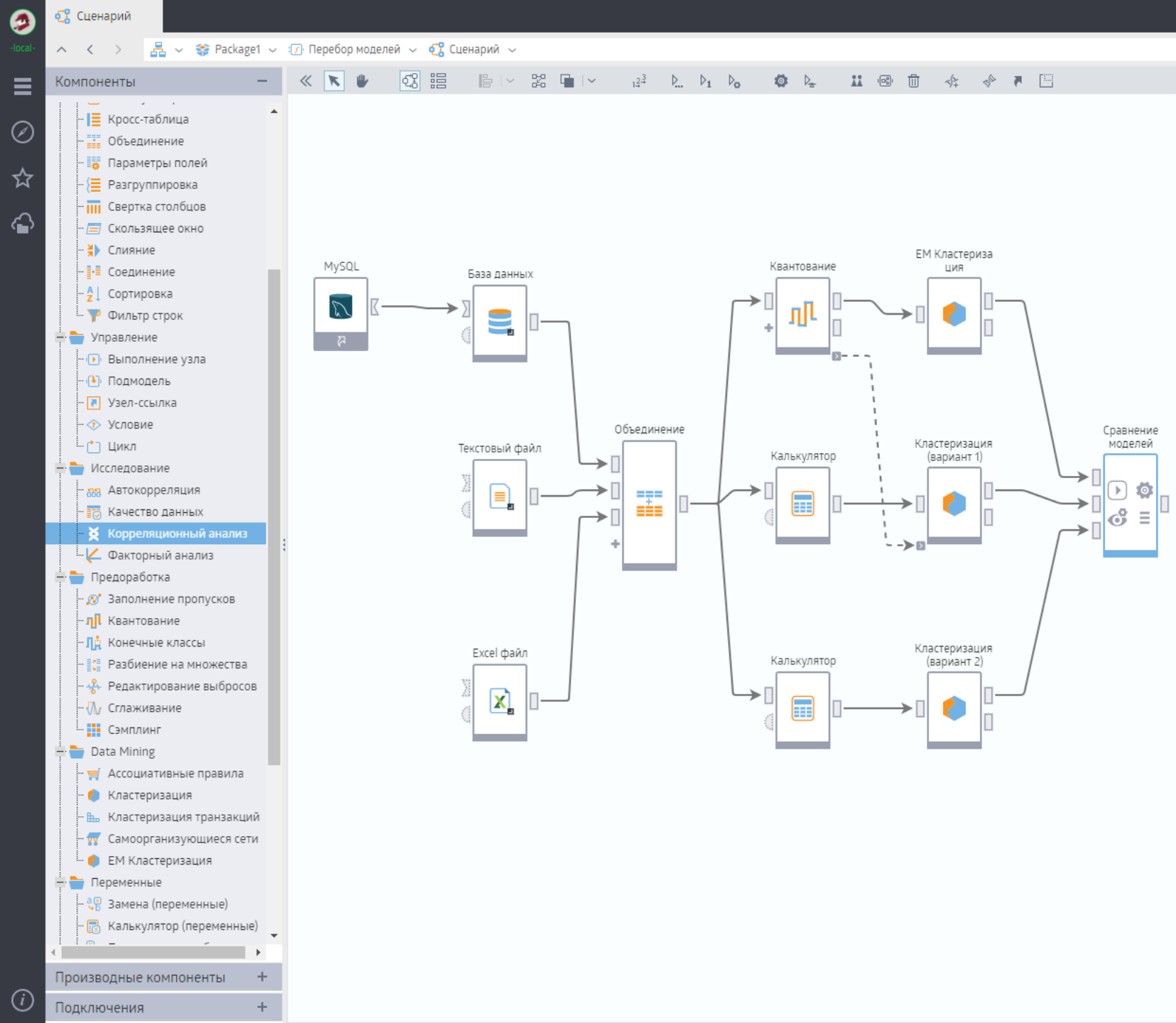


x64 – адресация
до 16 Tb RAM



Стараемся все
хранить в RAM

При недостатке
RAM кэшируем



- Параллельное выполнение сценария
- Параллельное выполнение циклов
- Параллельные расчеты алгоритмов
- Чтение и запись в несколько потоков
- Возможность ручного задания порядка

Технология	Выгода
Асинхронная обработка	Отсутствие блокировок пользовательского интерфейса при долгих операциях
Ленивые вычисления	Экономия ресурсов за счет расчетов только при необходимости

Резюме

Профессии – группы риска

Под угрозой	Условно под угрозой	Угроза вследствие повышения производительности труда		Возрастающий спрос
Оператор колл-центра	Водитель	Секретарь	Преподаватель	Специалист по МО и ИИ
Оператор склада	Курьер	Солдат	Полицейский	Data Scientist
Продавец	Пилот	Бухгалтер	Уборщик	Цифровой адвокат
Клерк-юрист	Комбайнёр	Переводчик	Промышленный рабочий	Специалист по безопасности ИИ

Self-made аналитика – основной способ демократизации аналитики. Специалист в предметной области проанализирует без программистов:

- Быстрее
- Лучше
- Дешевле

Современная платформа аналитики

Веб-интерфейс	Низкий порог входа	Широкая сфера применения
Реализация сложной логики	Повторное использование компонентов	Веб-сервисы из коробки

Академические инициативы

Студентам

1. Бесплатная версия платформы Loginom Academic
2. Электронные курсы
3. Хакатоны

Вузам

1. Действующая с 2007 года академическая программа
2. Доступ к курсам и методическим материалам
3. Коммерческие редакции Loginom

loginom.ru