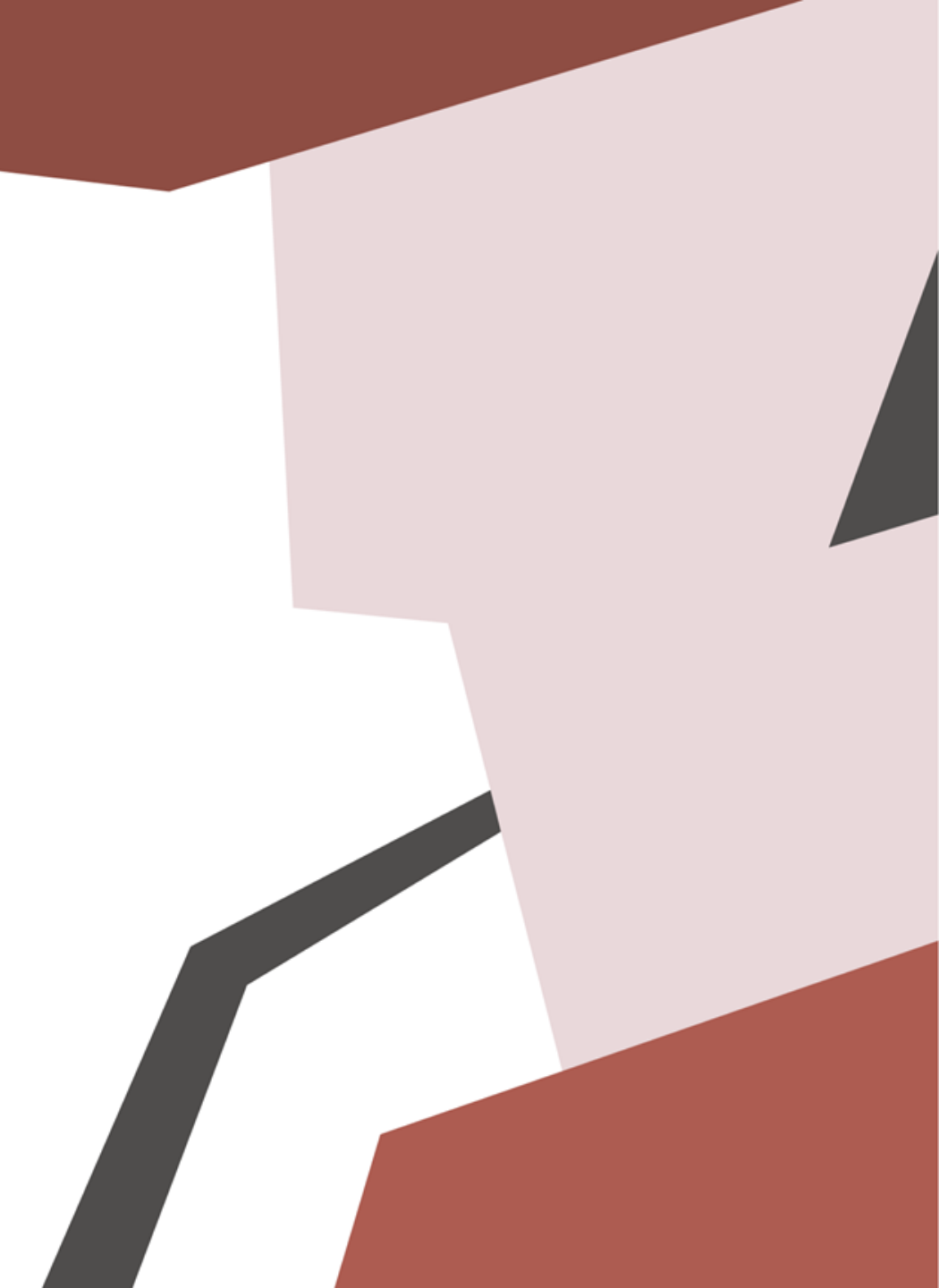
Loginom – аналитическая платформа

Алексей Арустамов

BaseGroup Labs



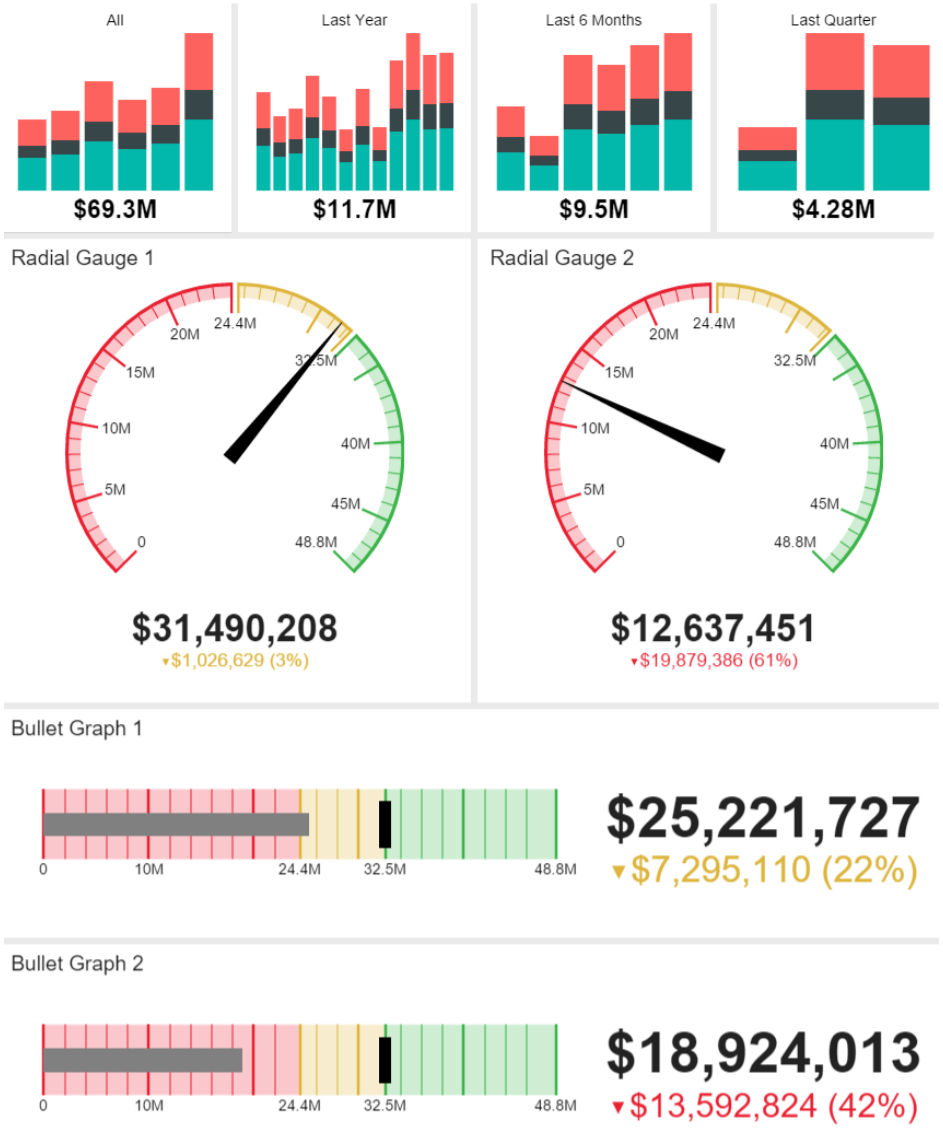
Data driven…

everything:

данные – основа

принятия всех

решений

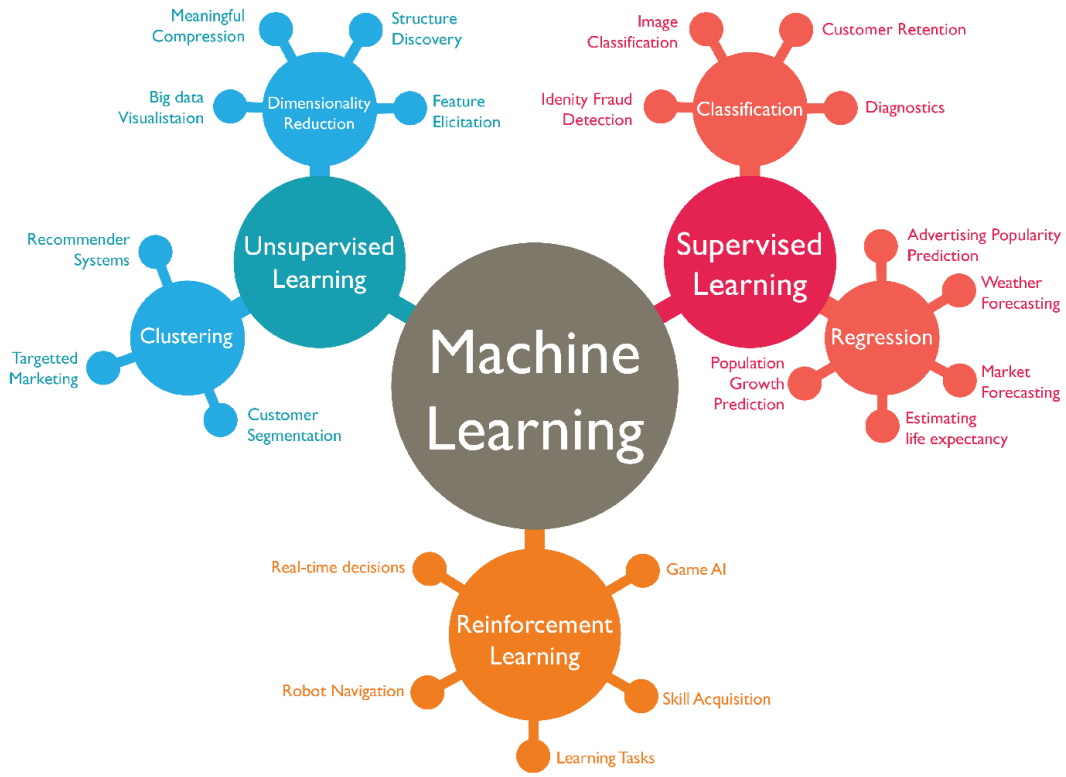


Визуализация не решает задач:

- Консолидации
- Очистки данных
- Сложных расчетов
- Прогнозирования
- Оптимизации

Машинное обучение не работает, когда:

- Мало данных
- Нет фиксации влияющих факторов
- Необходимо учесть бизнес-правила
- Данные забиты шумом



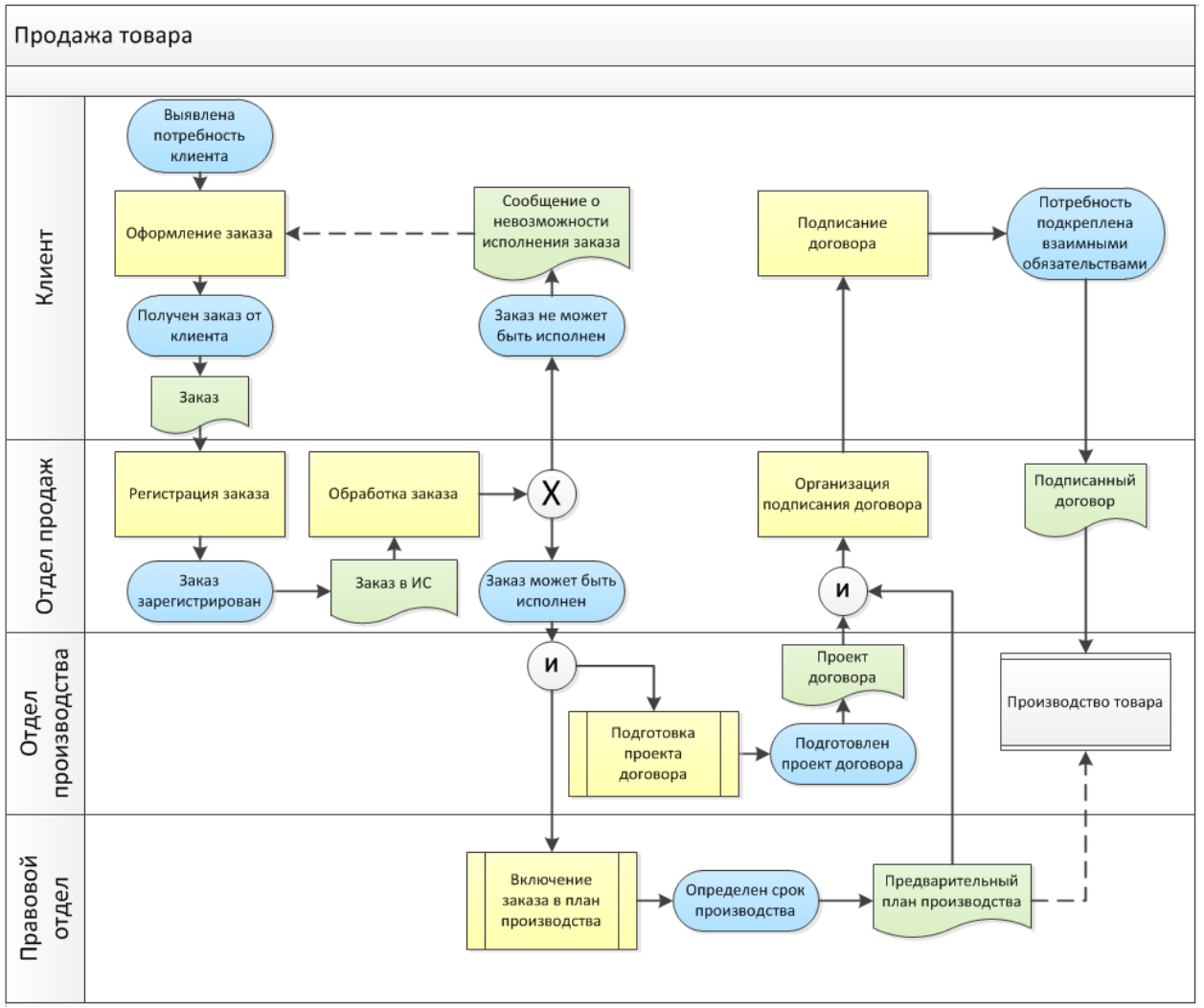
Нет единого способа, используются все подходы

Вызовы

Данные:

- Много источников
- Нет связности
- Нет унификации
- Разная структура
- Разная детализация


Дубли
Пропуски
Качество данных
Аномалии
Опечатки
Противоречия



- Учет специфики бизнеса
- Изменение бизнес-правил
- Модификация процессов

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

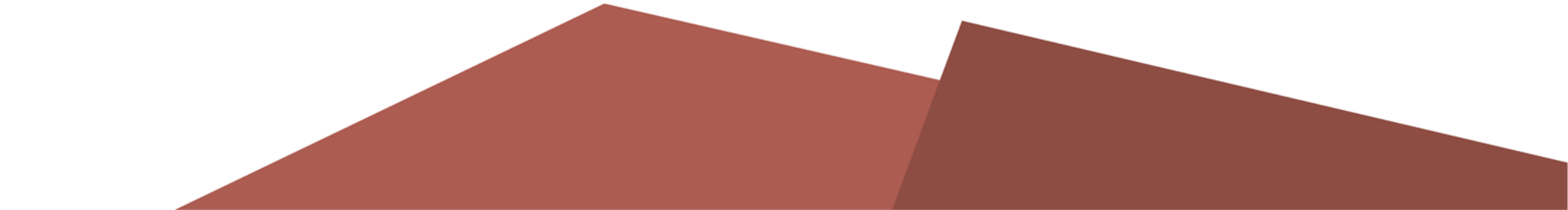
ПРЕДМЕТНАЯ ОБЛАСТЬ И SOFT SKILLS
- Понимание и интерес к бизнесу
- Интерес к данным
- Неформальное лидерство
- Хакерское мышление
- Умение решать проблемы
- Умение мыслить стратегически, проактивность, креативность, инновационный подход, готовность к сотрудничеству

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-менеджементом
- 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

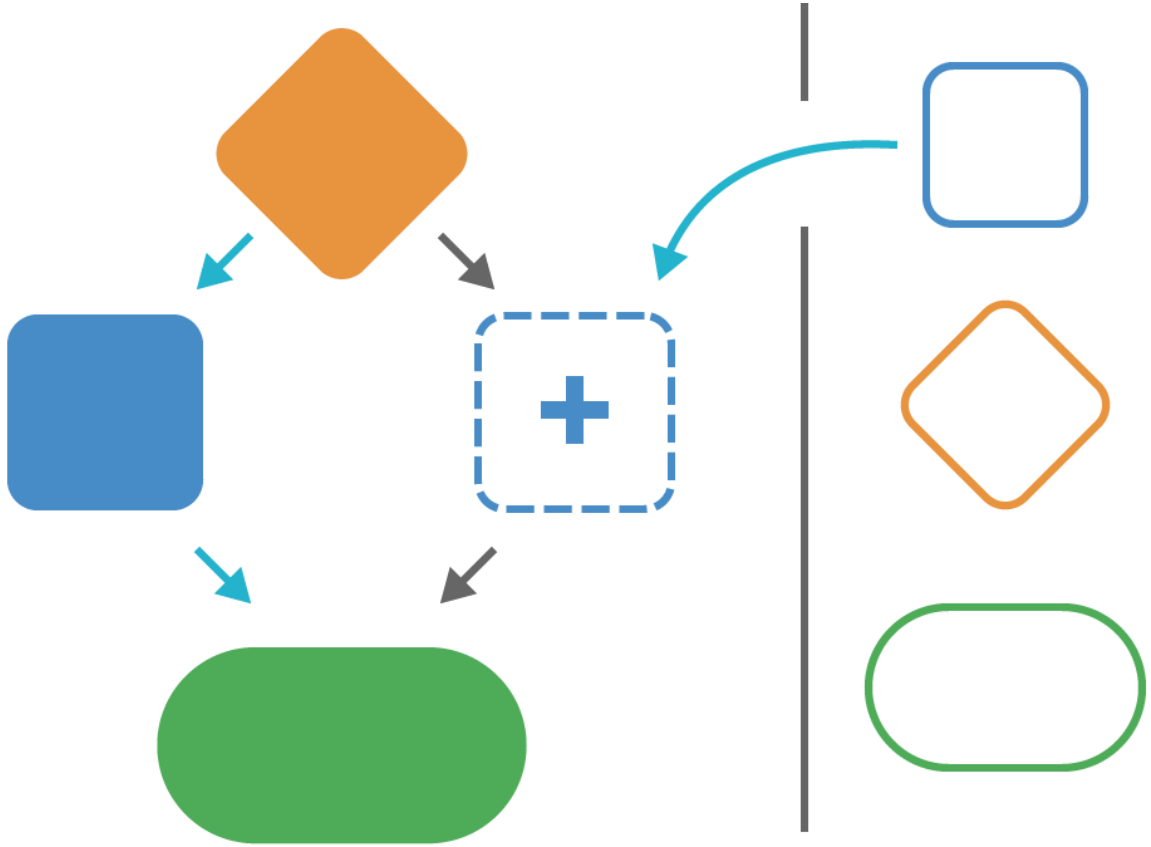
Кадровый голод

Требования



```
calculateIC<-function(gene2go, root_term) {
  library(GO.db)
  go_ic_data<-NULL
  root_list<-union(GOMFOFFSPRING[[root_term]], root_term)
  denom<-length(gene2go$GeneID[gene2go$GOTerm %in% root_list])
  for(index in 1:length(root_list)) {
    go_ic_row<-NULL
    go_term<-as.character(root_list[index])
    offspring<-union(GOMFOFFSPRING[[go_term]], go_term)
    prob<-length(gene2go$GeneID[gene2go$GOTerm %in% offspring])
    term_ic<-ic(prob, denom)
    go_term_ic<-round(term_ic, digits=2)
    go_ic_row<-c(go_term, go_term_ic, prob)
    go_ic_data<-rbind(go_ic_data, go_ic_row)
  }
  go_ic_data<-as.data.frame(go_ic_data)
  names(go_ic_data)<-c("GOTerm","IC", "Freq")
  go_ic_data
}
ic<-function(prob, total) {
  info_content<-NULL
  if(prob==0) {
    info_content<-NA
  } else {
    info_content<-(-log2(prob/total))
  }
  info_content
}
```

Минимум, а лучше отсутствие кодирования

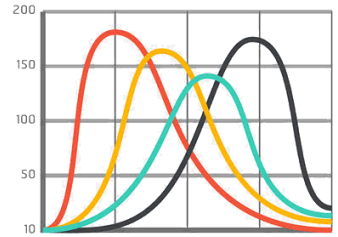


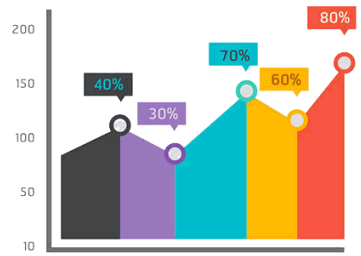
Сбор из блоков: сокрытие сложной логики и нюансов математики

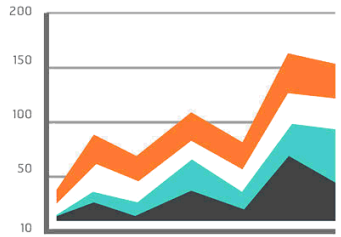
Быстрая обработка данных

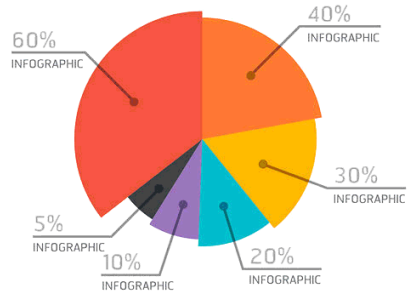
Обработка больших объемов данных

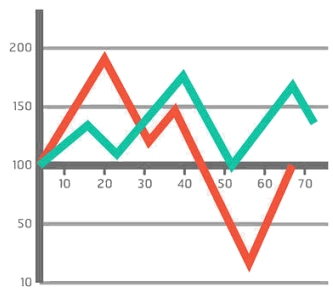
Удобная визуализация











Универсальный клиент:

- Десктоп
- Планшет
- Смартфон

Совместная работа:

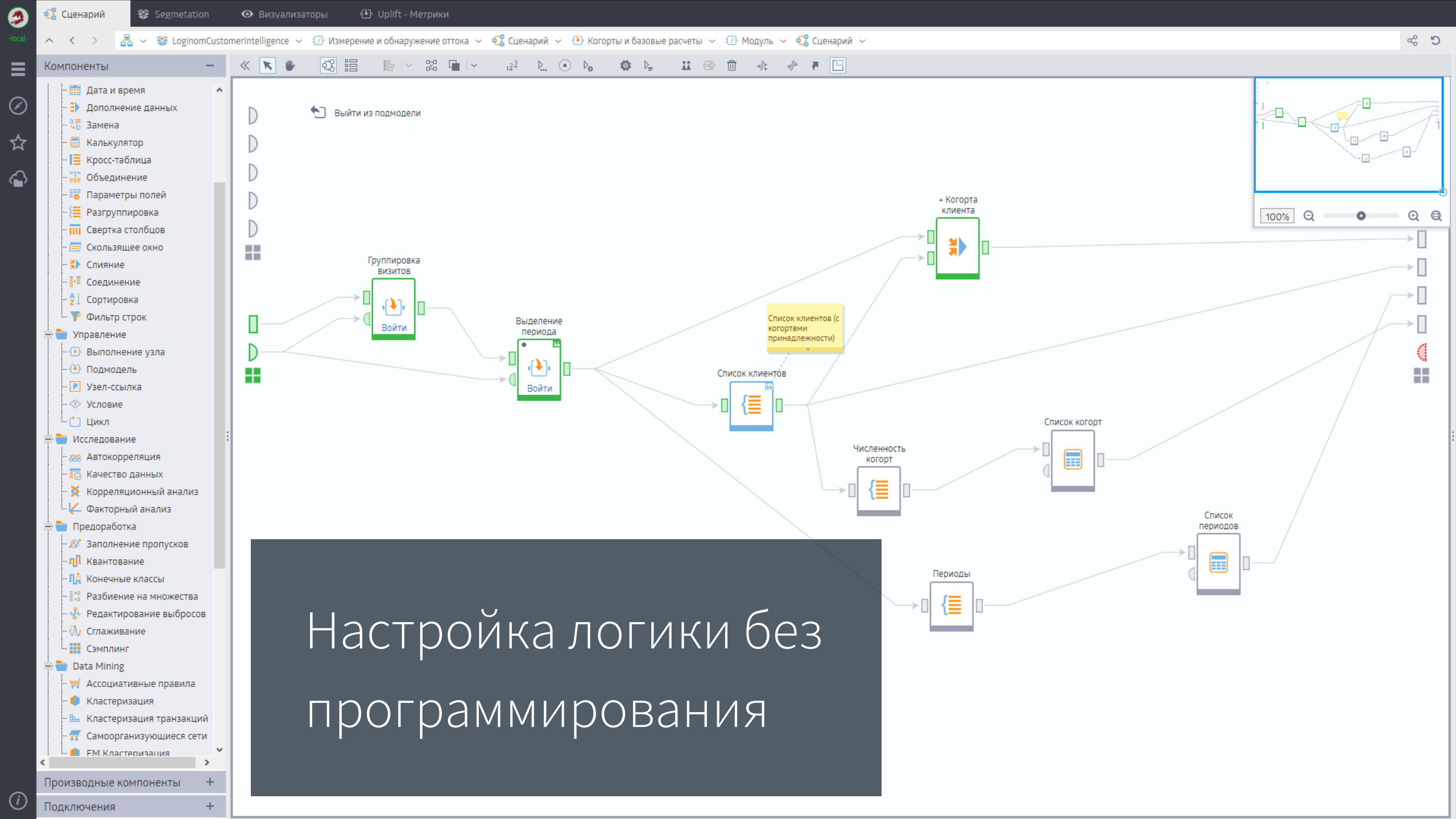
- Руководитель
- Аналитик
- Администратор
- Программист
- Пользователь

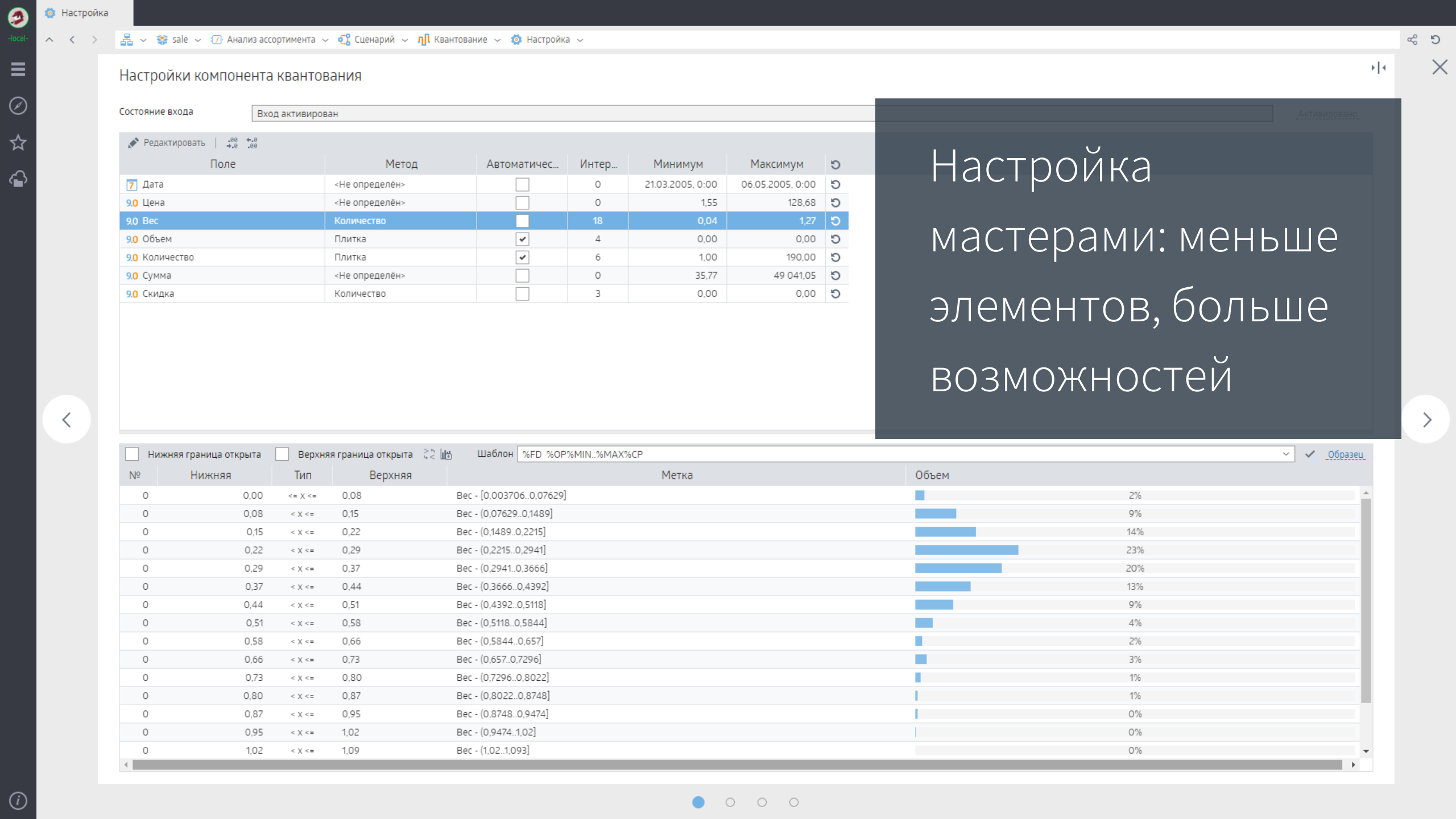
Loginom

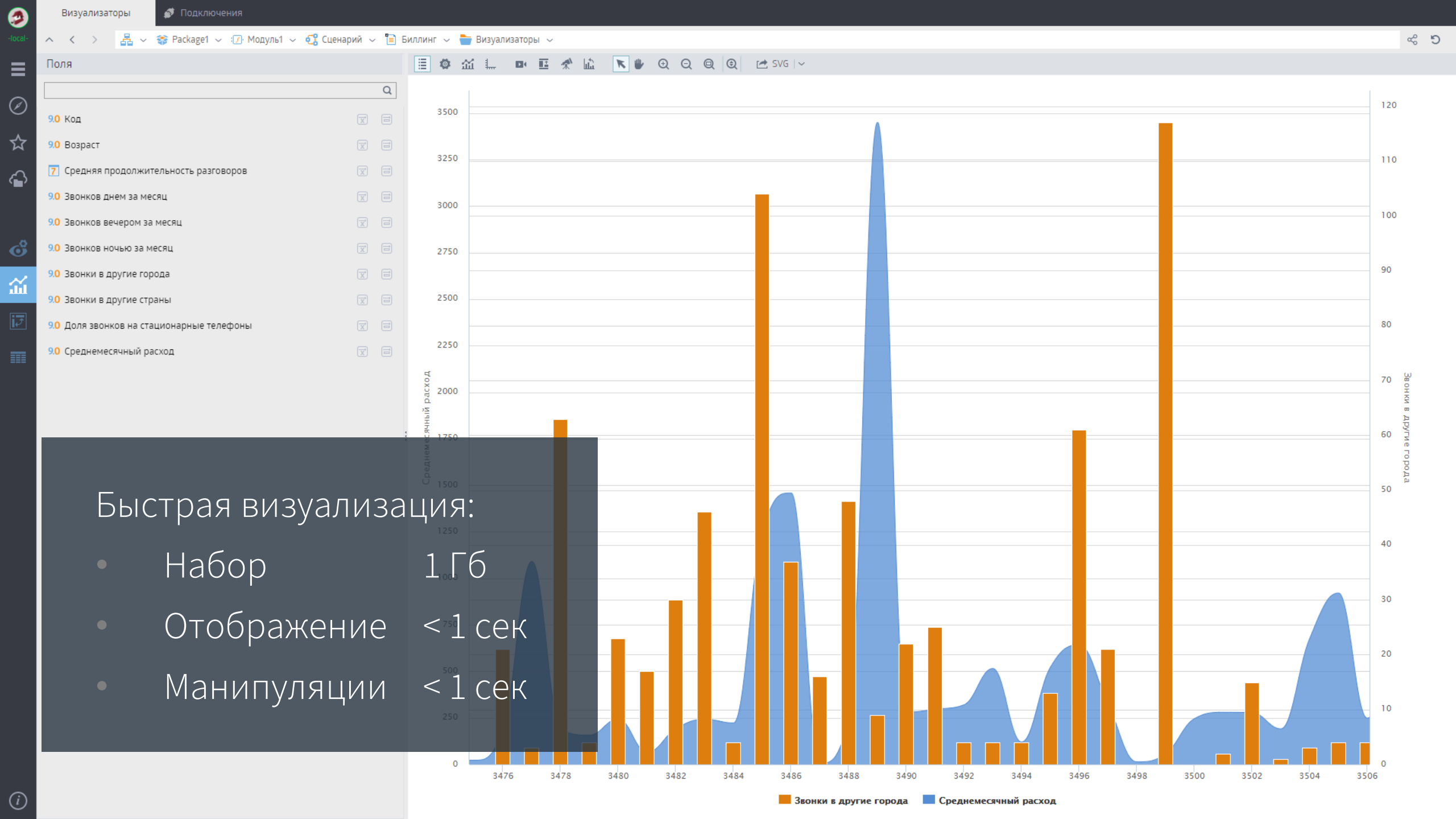
Сделать продвинутую аналитику массовой

Единая среда –
всё в браузере:

- Десктоп
- Планшет
- Телефон


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
Периоды
Список периодов
100%
Настройка логики без программирования


Настройка
sale
Анализ ассортимента
Сценарий
Квантование
Настройка
Настройки компонента квантования
Состояние входа
Вход активирован
Редактировать
Поле
Метод
Автоматичес...
Интер...
Минимум
Максимум
Дата
<Не определён>
0
21.03.2005, 0:00
06.05.2005, 0:00
Цена
<Не определён>
0
1,55
128,68
Вес
Количество
18
0,04
1,27
Объем
Плитка
4
0,00
0,00
Количество
Плитка
6
1,00
190,00
Сумма
<Не определён>
0
35,77
49 041,05
Скидка
Количество
3
0,00
0,00
Нижняя граница открыта
Верхняя граница открыта
Шаблон
%FD %OP%MIN..%MAX%CP
Образец
№
Нижняя
Тип
Верхняя
Метка
Объем
0 0,00 <= x <= 0,08 Вес - [0,003706..0,07629] 2%
0 0,08 < x <= 0,15 Вес - (0,07629..0,1489] 9%
0 0,15 < x <= 0,22 Вес - (0,1489..0,2215] 14%
0 0,22 < x <= 0,29 Вес - (0,2215..0,2941] 23%
0 0,29 < x <= 0,37 Вес - (0,2941..0,3666] 20%
0 0,37 < x <= 0,44 Вес - (0,3666..0,4392] 13%
0 0,44 < x <= 0,51 Вес - (0,4392..0,5118] 9%
0 0,51 < x <= 0,58 Вес - (0,5118..0,5844] 4%
0 0,58 < x <= 0,66 Вес - (0,5844..0,657] 2%
0 0,66 < x <= 0,73 Вес - (0,657..0,7296] 3%
0 0,73 < x <= 0,80 Вес - (0,7296..0,8022] 1%
0 0,80 < x <= 0,87 Вес - (0,8022..0,8748] 1%
0 0,87 < x <= 0,95 Вес - (0,8748..0,9474] 0%
0 0,95 < x <= 1,02 Вес - (0,9474..1,02] 0%
0 1,02 < x <= 1,09 Вес - (1,02..1,093] 0%
Настройка мастерами: меньше элементов, больше возможностей


Визуализаторы
Подключения
Package1
Модуль1
Сценарий
Биллинг
Визуализаторы
Поля
Код
Возраст
Средняя продолжительность разговоров
Звонков днем за месяц
Звонков вечером за месяц
Звонков ночью за месяц
Звонки в другие города
Звонки в другие страны
Доля звонков на стационарные телефоны
Среднемесячный расход
SVG
Среднемесячный расход
Звонки в другие города
Звонки в другие города
Среднемесячный расход
Быстрая визуализация:
Набор 1 Гб
Отображение < 1 сек
Манипуляции < 1 сек

Визуализаторы | Компоненты

Package1 | Модуль1 | Сценарий | Автозапчасти | Визуализаторы

Область фильтрации

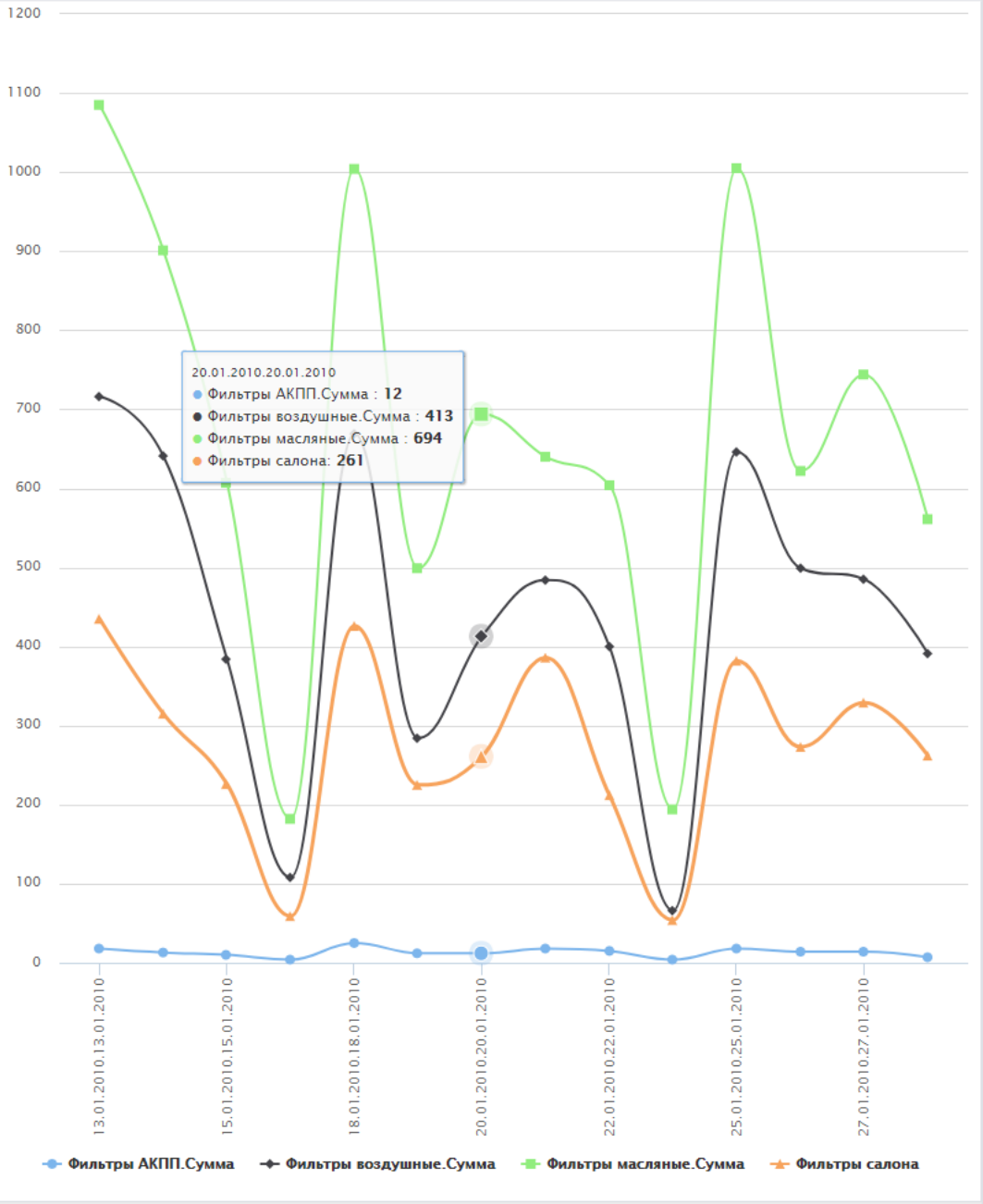
Группа товара | Факты

Дата | Код товара

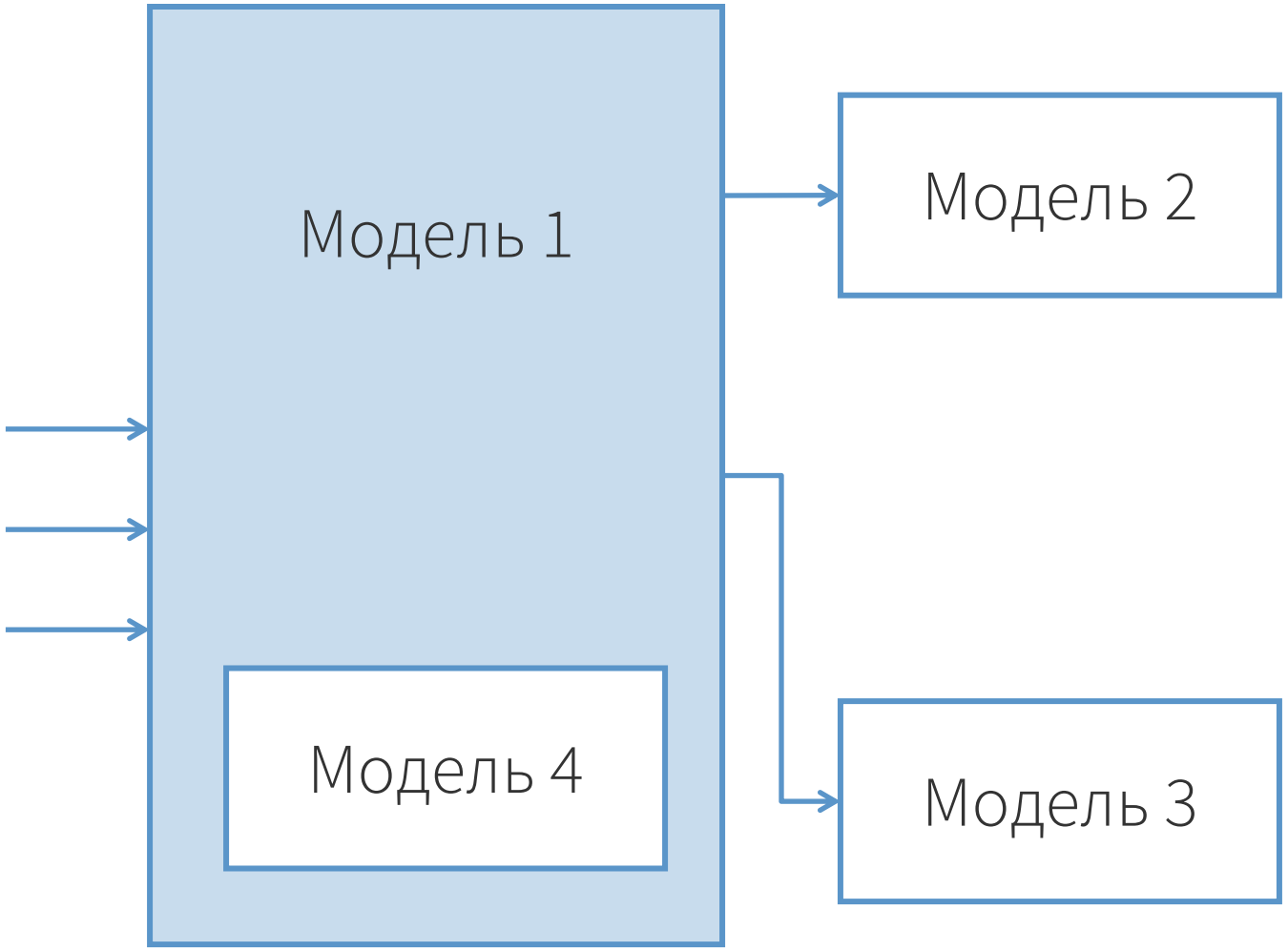
	Фильтры АКПП		Фильтры воздушные		Фильтры масляные		Фильтры салона	
	Сумма	Количест...	Сумма	Количест...	Сумма	Количест...	Сумма	Количест...
	# Колич...	Σ Сумма	# Колич...	Σ Сумма	# Колич...	Σ Сумма	# Колич...	Σ Сумма
13.01.2010	9	18,00	214	716,00	179	1 085,00	137	435,00
14.01.2010	9	13,00	191	641,00	157	901,00	116	315,00
15.01.2010	8	10,00	150	384,00	129	607,00	92	226,00
16.01.2010	4	4,00	64	108,00	61	182,00	50	59,00
18.01.2010	15	25,00	251	669,00	188	1 004,00	145	426,00
19.01.2010	9	12,00	138	284,00	119	499,00	111	225,00
20.01.2010	8	12,00	145	413,00	126	694,00	115	261,00
21.01.2010	6	18,00	175	484,00	143	640,00	127	386,00
22.01.2010	10	15,00	122	400,00	131	604,00	84	212,00
23.01.2010	2	4,00	40	66,00	52	194,00	32	54,00
25.01.2010	9	18,00	199	646,00	176	1 005,00	135	382,00
26.01.2010	7	14,00	169	499,00	151	622,00	107	273,00
27.01.2010	8	14,00	174	485,00	164	744,00	115	329,00
28.01.2010	7	7,00	140	391,00	146	561,00	102	262,00

OLAP в браузере

Сплайны | Количество.Сумма



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

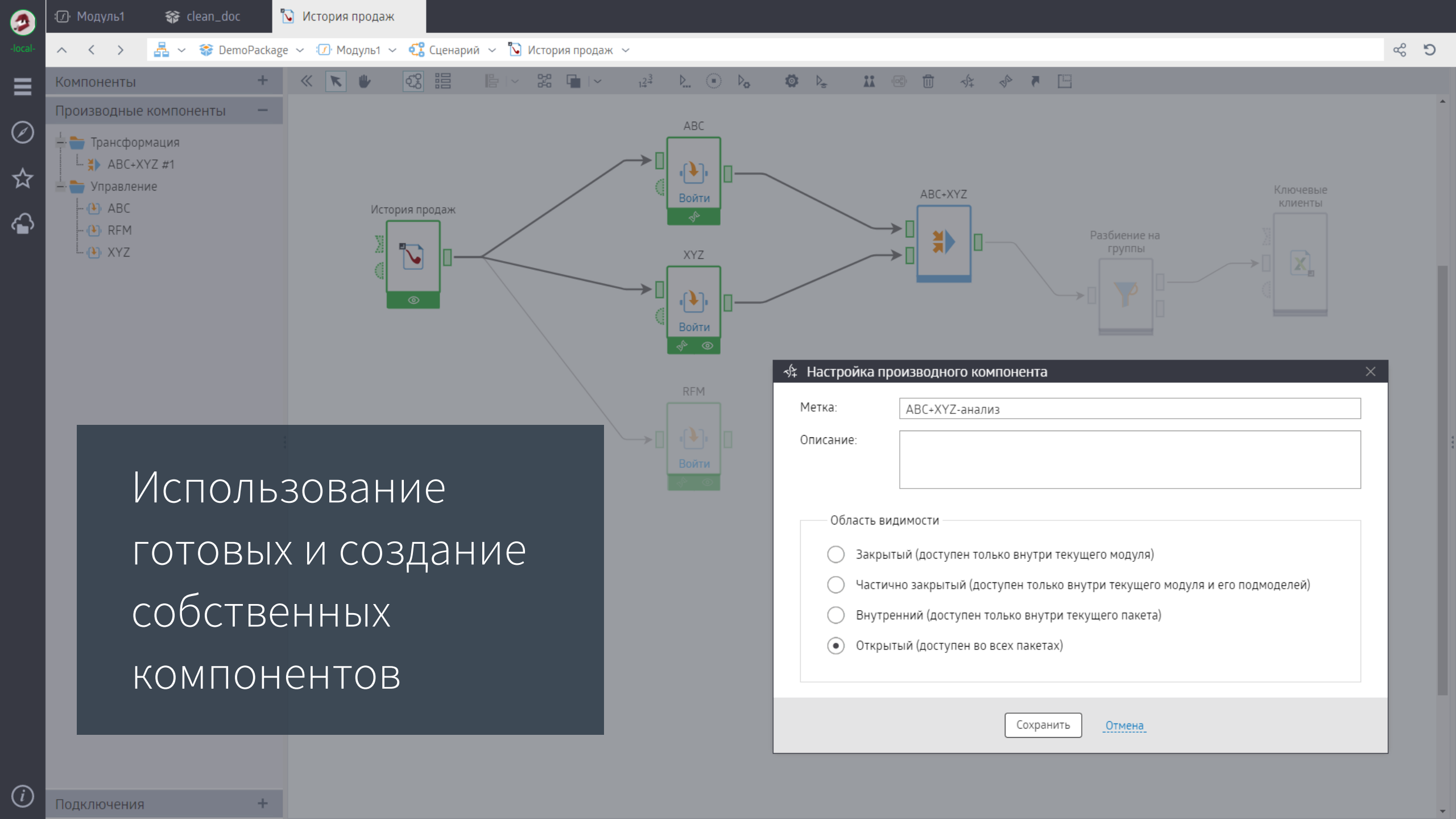


- Декомпозиция задачи
- Проектирование сверху-вниз
- Нет привязки к данным

Объектная модель

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

Варианты использования модели


Модуль1
clean_doc
История продаж
DemoPackage
Модуль1
Сценарий
История продаж
Компоненты
Производные компоненты
Трансформация
ABC+XYZ #1
Управление
ABC
RFM
XYZ
Подключения
ABC
XYZ
RFM
Войти
Войти
Войти
История продаж
ABC+XYZ
Разбиение на группы
Ключевые клиенты
Использование готовых и создание собственных компонентов
Настройка производного компонента
Метка:
ABC+XYZ-анализ
Описание:
Область видимости
Закрытый (доступен только внутри текущего модуля)
Частично закрытый (доступен только внутри текущего модуля и его подмоделей)
Внутренний (доступен только внутри текущего пакета)
Открытый (доступен во всех пакетах)
Сохранить
Отмена

Алгоритмы обработки

Трансформация:

- Слияние
- Кросс-таблицы
- Свертка столбцов
- Замена
- …

Исследование:

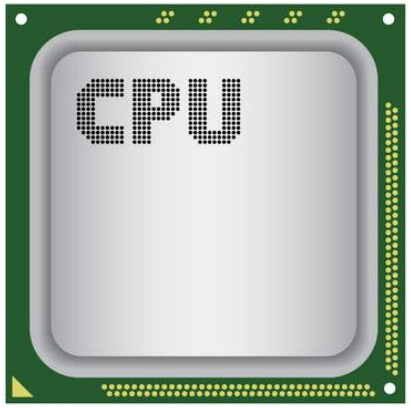
- Корреляционный анализ
- Факторный анализ
- …

Предобработка:

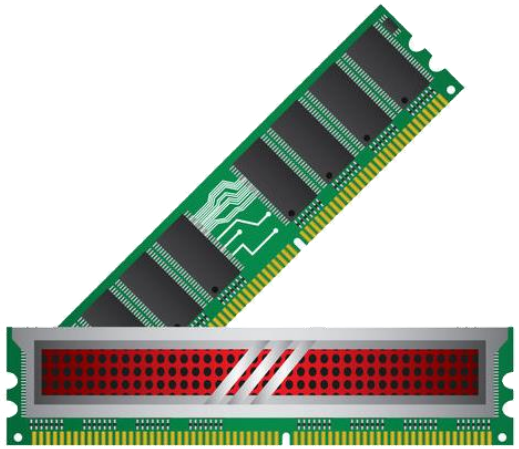
- Заполнение пропусков
- Редактирование выбросов
- Сэмплинг
- …

Data Mining:

- Кластеризация
- EM-кластеризация
- Ассоциативные правила
- …

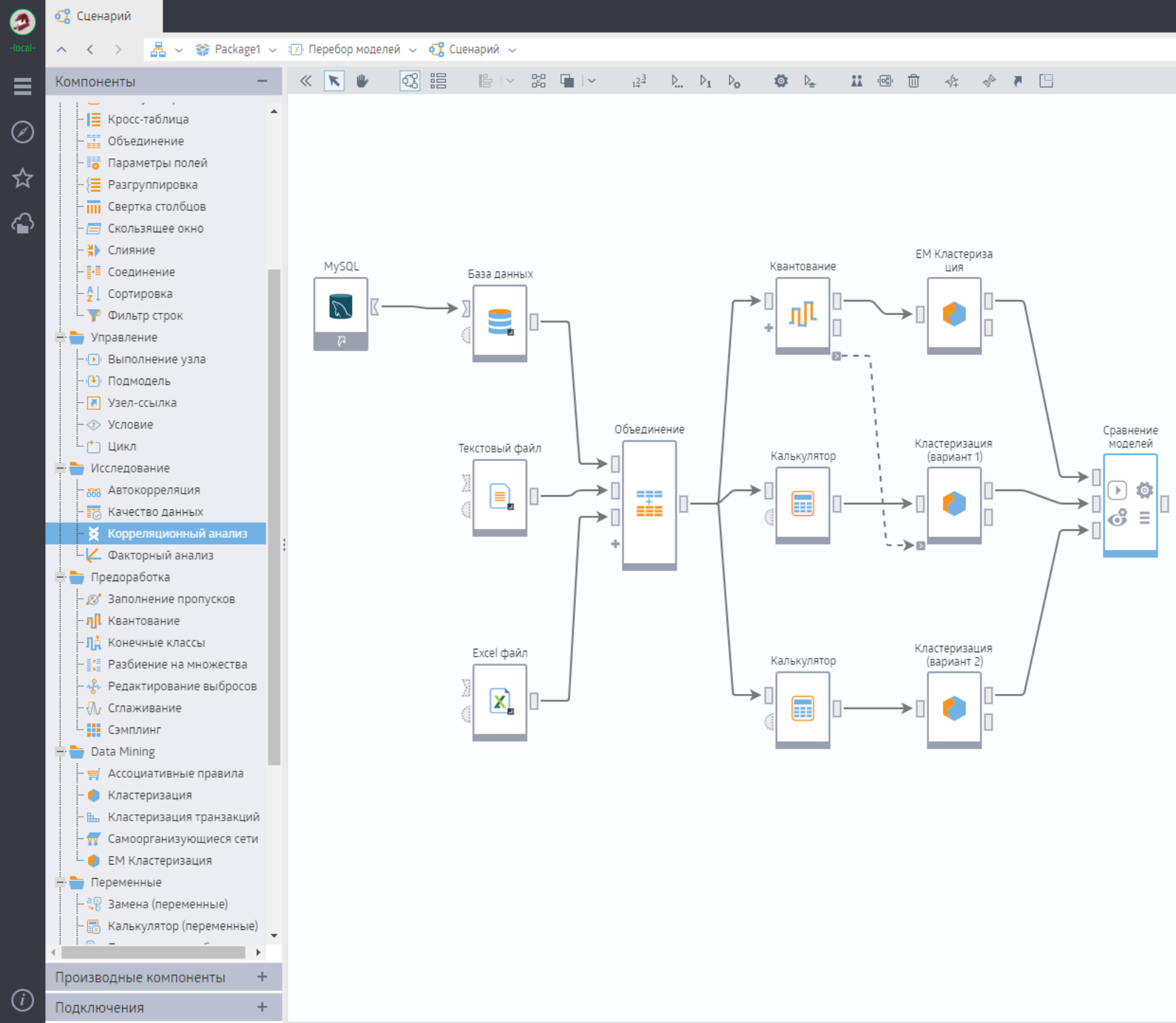


x64 – адресация
до 16 Tb RAM



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

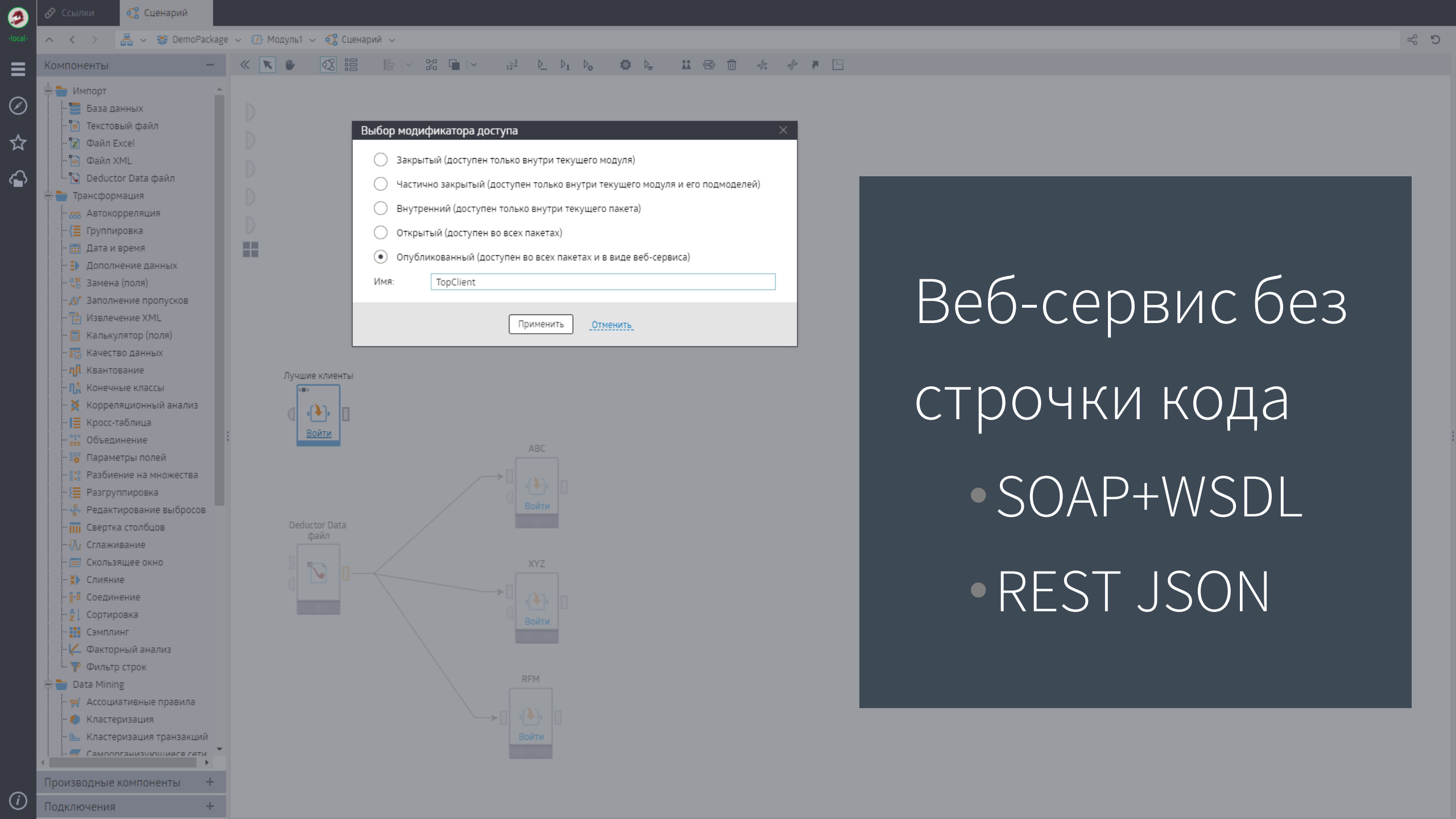
Интеграция из коробки

Доступ к данным

- Базы данных
- Файлы
- Веб-сервисы

Публикация веб-сервисов

- SOAP+WSDL
- REST JSON



Веб-сервис без строчки кода

- SOAP+WSDL
- REST JSON

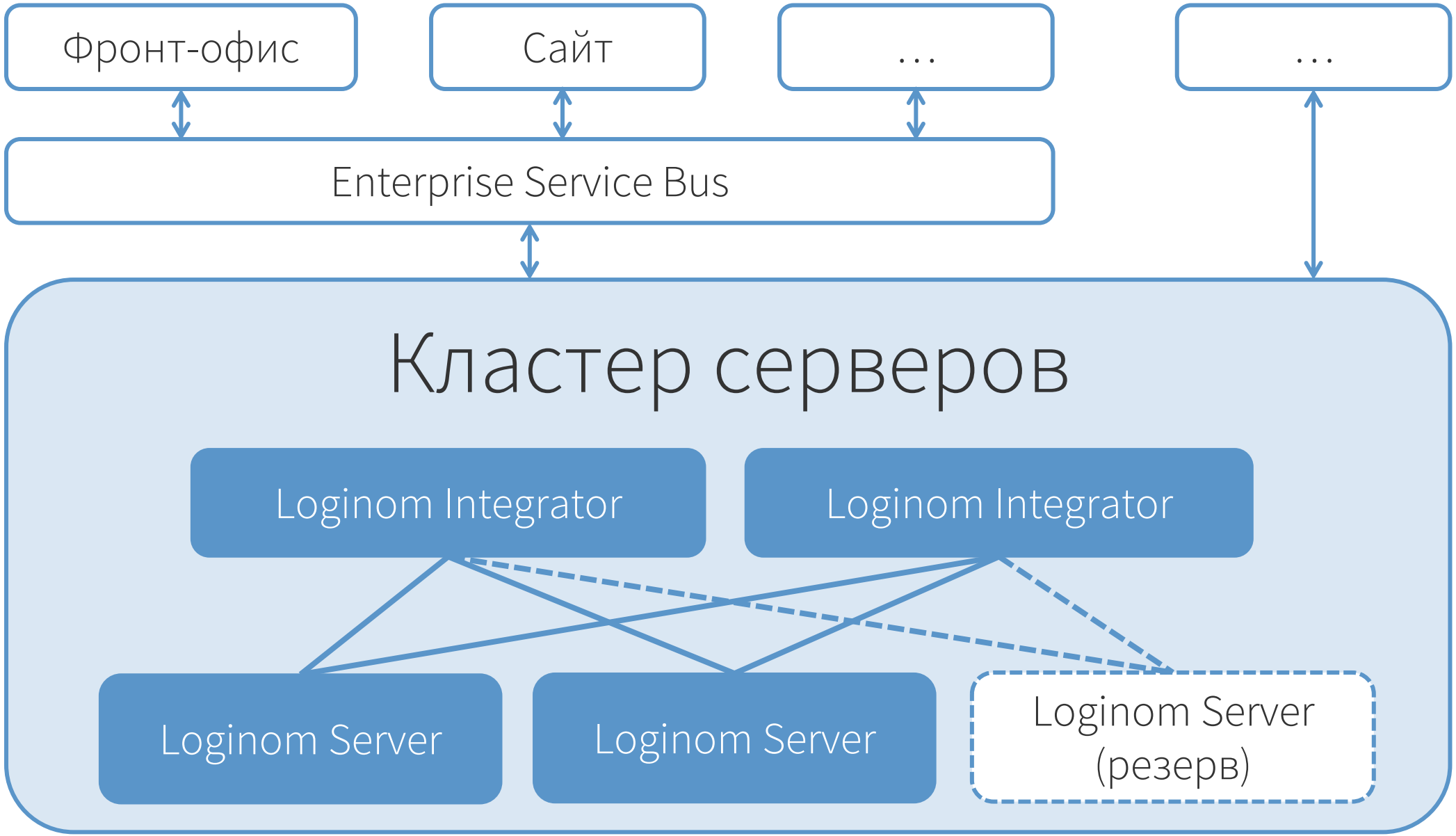
Desktop

Client-Server

Private Cloud

Public Cloud

1. Удобный вариант развертывания
2. Нет жесткой привязки к облаку
3. Минимальные требования к железу


Фронт-офис
Сайт
…
…
Enterprise Service Bus
Кластер серверов
Loginom Integrator
Loginom Integrator
Loginom Server
Loginom Server
Loginom Server
(резерв)

Преимущества кластера:

- Балансировка нагрузки
- Горячая замена серверов
- Восстановление после сбоев
- Резервные сервера в режиме ожидания

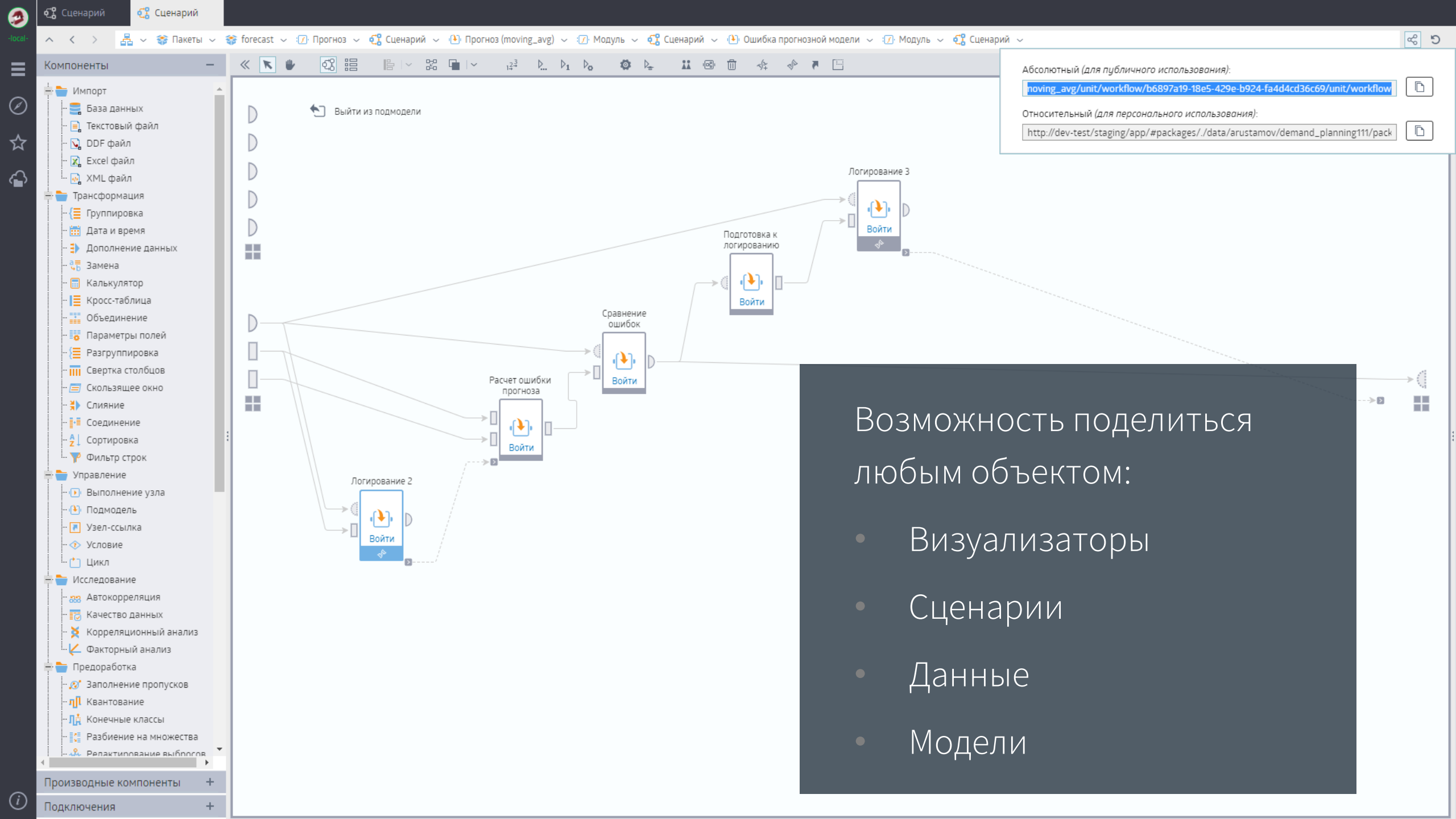
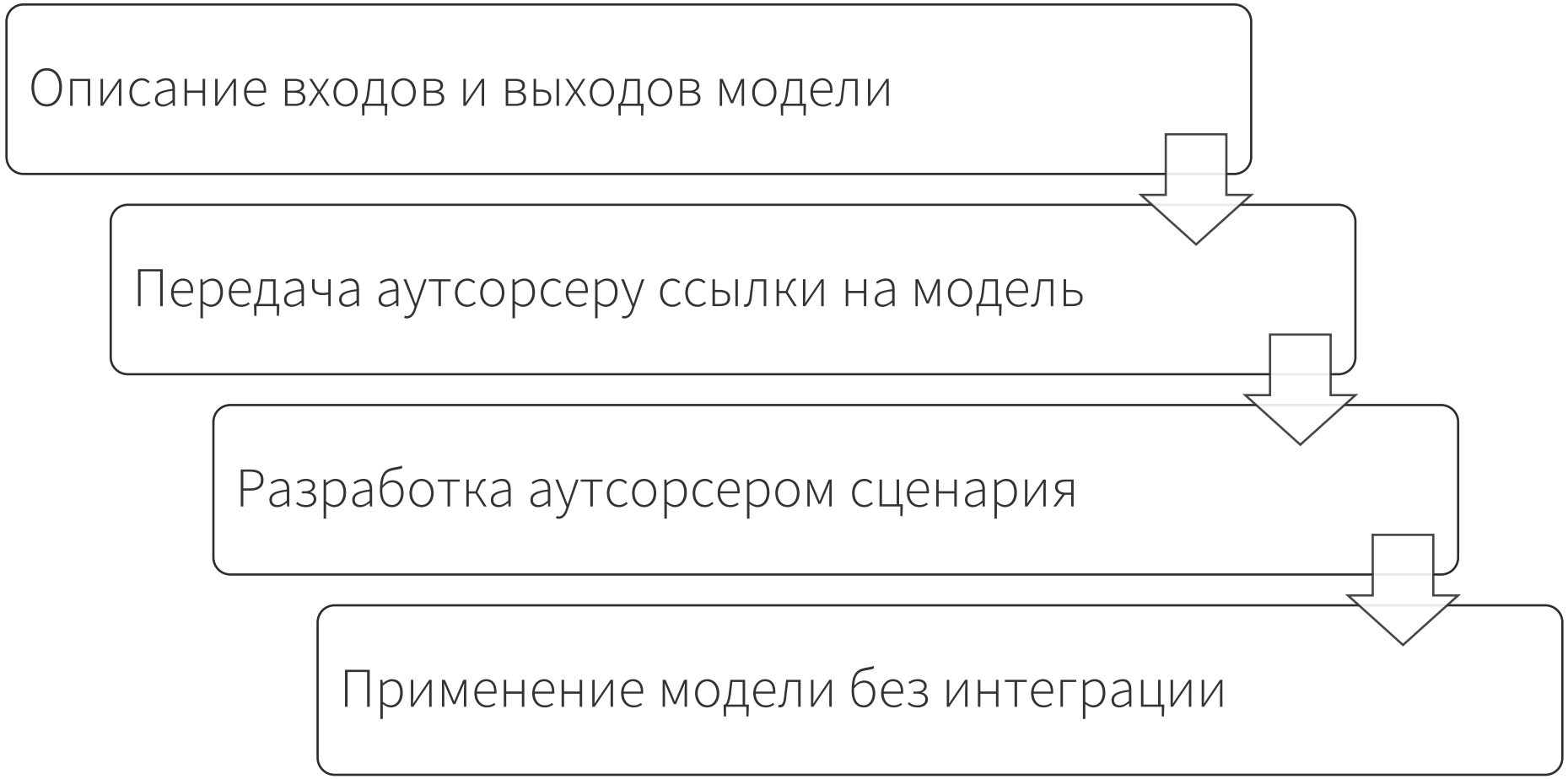
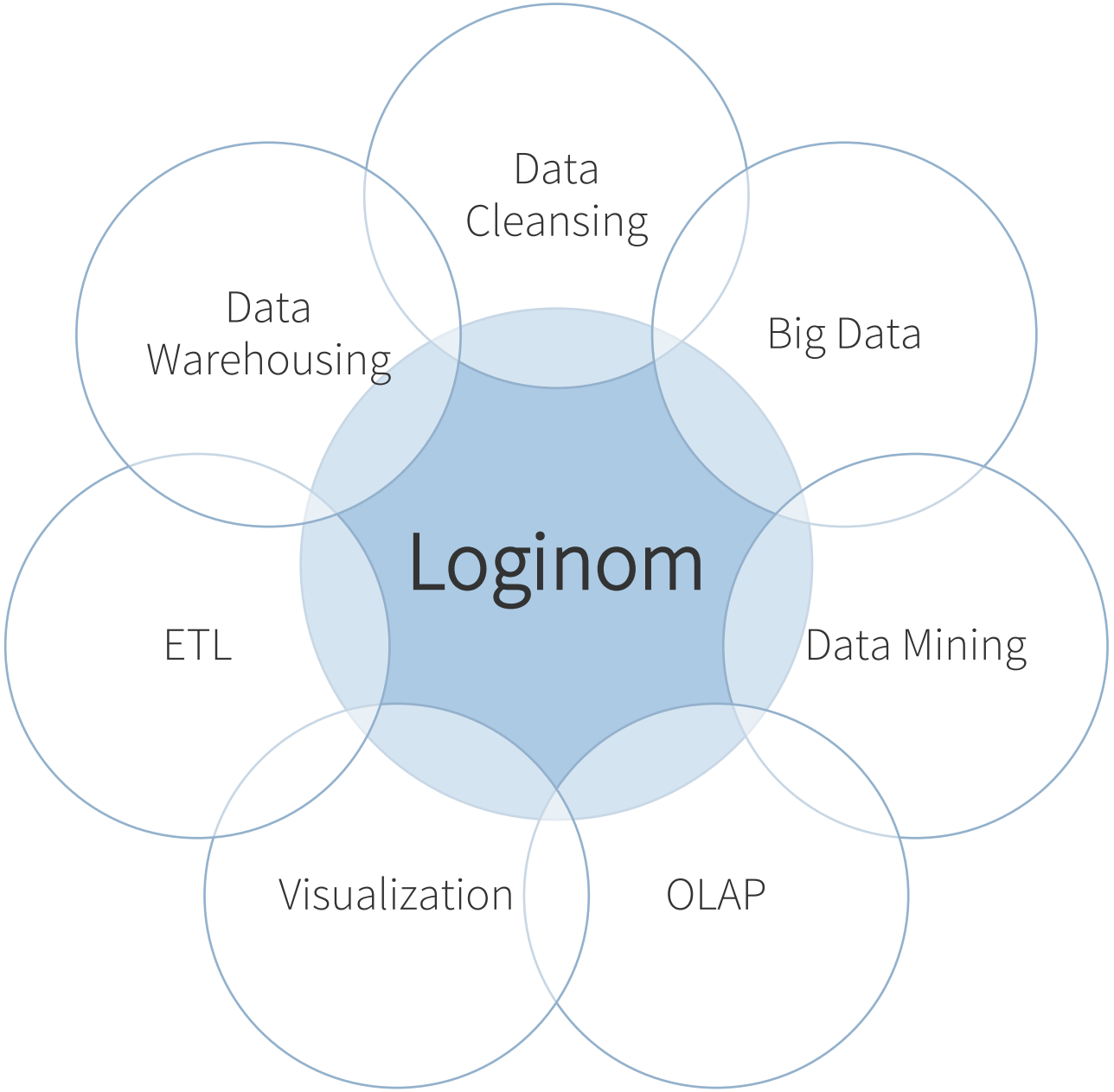

Сценарий
Сценарий
Пакеты
forecast
Прогноз
Сценарий
Прогноз (moving_avg)
Модуль
Сценарий
Ошибка прогнозной модели
Модуль
Сценарий
Компоненты
Импорт
База данных
Текстовый файл
DDF файл
Excel файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
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
Производные компоненты
Подключения
Выйти из подмодели
Абсолютный (для публичного использования):
noving_avg/unit/workflow/b6897a19-18e5-429e-b924-fa4d4cd36c69/unit/workflow
Относительный (для персонального использования):
http://dev-test/staging/app/#packages/./data/arustamov/demand_planning111/pack
Логирование 3
Войти
Подготовка к логированию
Войти
Сравнение ошибок
Войти
Расчет ошибки прогноза
Войти
Логирование 2
Войти
Возможность поделиться любым объектом:
Визуализаторы
Сценарии
Данные
Модели

Схема аутсорсинга аналитики



Loginom эффективно утилизирует аппаратные ресурсы, но способен работать и на слабых серверах



Один продукт – много применений

Loginom сделает продвинутую аналитику массовой

- Простой
- Быстрый
- Мощный
- Гибкий
- Красивый :)

basegroup.ru