

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

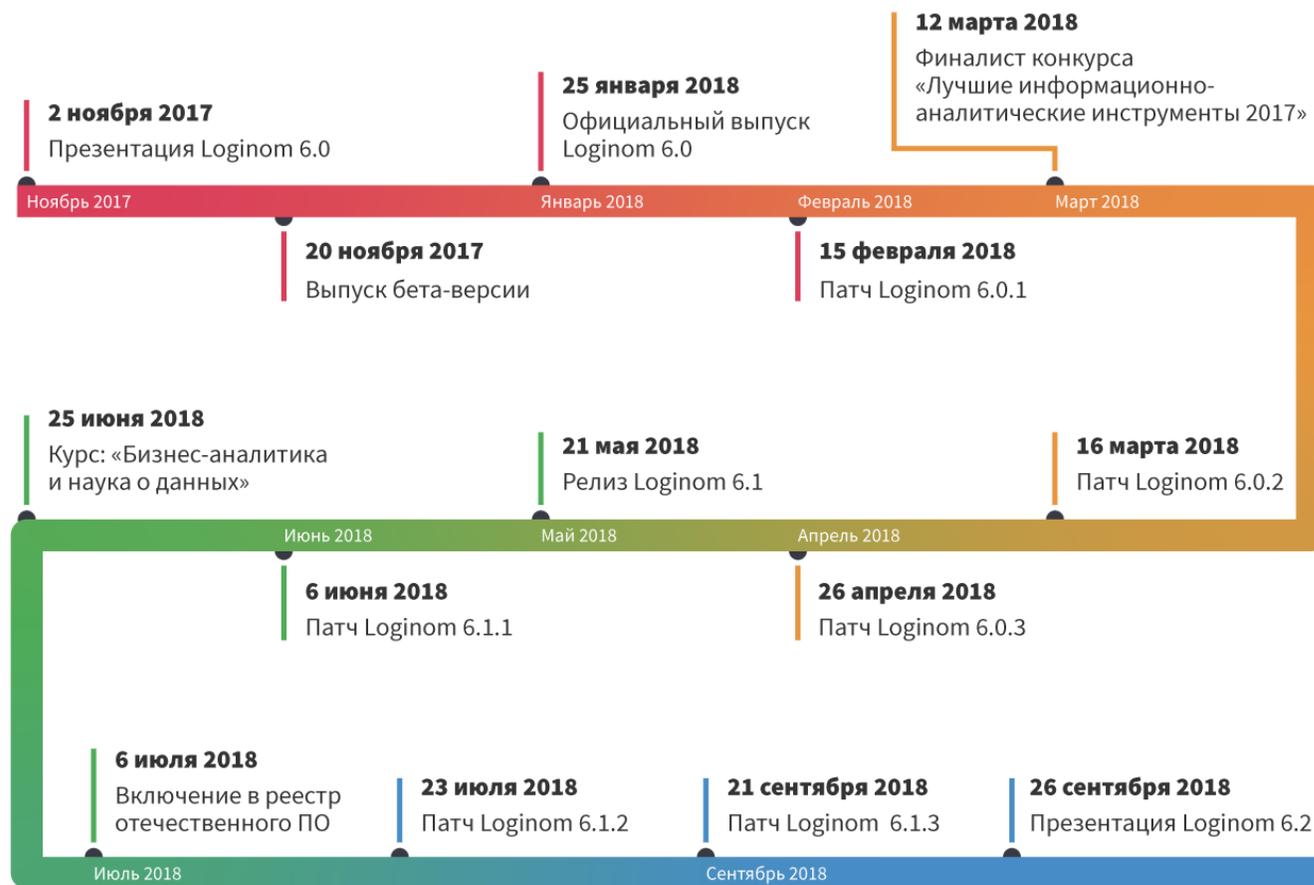
- Аналитическая платформа Deductor – предшественник Loginom
- Книга «Бизнес-аналитика: от данных к знаниям» (2008-2013 гг.)





Сделать
анализ данных
массовой
технологией

Развитие платформы Loginom



BIG DATA ФЕСТИВАЛЬ



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

INFRASTRUCTURE

HADOOP ON-PREMISE
 cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

HADOOP IN THE CLOUD
 aws, Microsoft Azure, Google Cloud, IBM InfoSphere BigInsights, Treasure Data, Duoble, Altiscale, CAZENA, CenturyLink

STREAMING / IN-MEMORY
 aws, databricks, stream, confluent, GridGain, ORACLE, dataArtisans, hazelcast, TERRACOTTA, KX, FASTDATA, WallarooLABS

ANALYTICS

DATA ANALYST PLATFORMS
 Microsoft, pentaho, alteryx, Digital Reasoning, GUAVUS, AYASDI, ATTIVO, Datameer, Quid, Incorta, interana, ClearStory, Origami, ENDOR, MODE, Bottlenose, switchboard

DATA SCIENCE PLATFORMS
 IBM, KNIME, dataiku, DOMINO, rapidminer, CONTINUUM ANALYTICS, ALGORITHMIA, DATAWATCH, ANDESS, SAS

APPLICATIONS - ENTERPRISE

SALES
 Einstein, CHORUS, INSIDESALES.COM, conversica, GONG, clari, aviso, tact.ai, fuse|machines, TROOPS

MARKETING - B2B
 RADIUS, App Annie, EVERSTRING, Lattice, MINTIGO, sense, tubular, ENGAGIO, mpo

MARKETING - B2C
 zeta, bloomreach, SendGrid, BlueYonder [PERSADO], kahuna, ACTIONIQ, SAILTHRU, BLUECORE, QUANTIFIND, mparticle, Ampero, amperity, TEALIUM, Simon, Lytics

CUSTOMER SERVICE
 MEDALLIA, zendesk, CLARABRIDGE, Gainsight, NGDATA, DigitalGenius, afiniti, AUTOMAT, frame.ai, msgai, INTERCOM, Ca#Desk

NoSQL DATABASES
 Google Cloud, aws, ORACLE, Microsoft Azure, mongoDB, MarkLogic, EROSPIKE, DATASTAX, ArangoDB, Couchbase, redislabs, SCYLLA

NewSQL DATABASES
 SAP HANA, Clustring, Pivotal, nuodb, Cockroach LABS, Cloud Spanner, MEMSQL, influxdata, MariaDB, TIMESCALE, citusdata, splice, paradigm4, Trafalgar, TIDB

GRAPH DBs
 neo4j, Amazon Neptune, IBM, ORACLE, OrientDB, InfiniteGraph, Objectivity

MPP DBs
 TERADATA, VERTICA, IBM Data Warehouse Systems, Action, Kognitio, Exasol, dremio

CLOUD EDW
 aws, Google Cloud, Microsoft Azure, Pivotal, snowflake

BI PLATFORMS
 Microsoft, aws, wave Analytics, looker, THOUGHTSPOT, AT SCALE, GoodData, Information Builders, MicroStrategy, birst

VISUALIZATION
 +tableau, Google Cloud, celonis, Qlik, Periscope Data, ZEPL, COMDATA, plotly, CHARTIO, TOUCAN TOCO

MACHINE LEARNING
 Azure Machine Learning, aws, Google Cloud, H2O, DataRobot, gamalon, ELEMENT, VISENZE, VERSIVE, deepsense, bonsai

HUMAN CAPITAL
 HireVue, entelo, hiQ, GIGSTER, textIQ, RESTLESS BANDIT, Wade&Wendy, Stella, Cxstree, pymetrics, mya, uncommon

LEGAL
 RAVEL, QSeal, Everlaw, JUDICATA, EBREVIA, IRONCLAD, PREMANTION, ROSS, Casetext

FINANCE
 naplan, ZUORO, SAP/SAHANA, TRADESHIFF

ENTERPRISE PRODUCTIVITY
 slack, ORACLE, lumiatu, DIFFBOT, clara, talla, butter.ai, Kasisto

BACK OFFICE AUTOMATION
 UiPath, blueprism, AppZen, WorkFusion

SECURITY
 CYLANCE, zscaler, STARKPath, illumio, CODE42, CipherCloud, Darktrace, ANOMALI, ThreatMetrix, VECTRA, cyberession, Guardian Analytics, DATAVISOR, sift science, SIGNIFYD, SentinelOne, SecurityScorecard, socure, ASEA1 security, BlueTalon, Recorded Future, feedzai, cybex, sparkcognition, IronNet Cybersecurity

DATA TRANSFORMATION
 talend, pentaho, alteryx, TRIFACTA, tamr, Paxata, StreamSets, UNIFI

DATA INTEGRATION
 SAP Data Services, Informatica, MuleSoft, snapLogic, TEALIUM, enigma, podium data, Segment, aloomo, xplenty, ZALONI, Stitch, import.io, Infoworks, ATTUNITY

DATA GOVERNANCE
 Informatica, IBM, SailPoint, McAfee Skyhigh Security Cloud, collibra, Alation, Waterline Data, KIMUTA, OKERA

MGMT / MONITORING
 aws, New Relic, octio, rubrik, APPDYNAMICS, dynatrace, vmware, splunk, SignalFx, druvo, Moogsoft, unravel, pagerduty, Numerify, Anodot

COMPUTER VISION
 Microsoft Azure, Amazon Rekognition, Cloud Vision API, clarifai, EVER AI, deepomatic, twentybn, neurala

HORIZONTAL AI
 IBM Watson, Cortana, Face++, sentient, Voyager Labs, vicarious, Affectiva, PROPHESEE, CognitiveScale, Numenta, PETUUM, nalogics, CURIOUS AI, OSARO, BLUE VISION

SPEECH & NLP
 Google Cloud, twilio, amazon alexa, narrative science, semanticmachines, Mobvoi, Eigen Technologies, SoundHound Inc., voltera, NIJANCE, snips, Mindfield, cortico.ai, ysecp, mauboo, GraspSpace

ADVERTISING
 AppNexus, criteo, xAd, Integral Ad Science, ORACLE, MOAT, OpenX, dataroma, theTradeDesk, algorithms, distillery, LiveIntent, TAPAD, dataxu, gumgum, Cppier, DYNAMIC YIELD, weldmco

EDUCATION
 Lullishuo, KNEWTON, Clever, Cleclara, kidaptive, PANORAMA, KINOWA, gradescope

GOVERNMENT
 OPENGOV, mark43, EN FiscalNote, GRIDSMART, LiveStories, Passport, SmartProcure, STREETLIGHTDATA, OpenDataSoft

REAL ESTATE
 REDFIN, Opendoor, VTS, CREDIF, reonomy, COMPSTAK, CAPE

FINANCE - INVESTING
 KENSHC, Dataminr, Quantopian, ADDEPAR, NUMERAL, ISENTIUM, ALGORIZ, RavenPack, PAGAYA

FINANCE - LENDING
 ondeck, Affirm,拍拍贷, JIANPU.AI, Kreditech, AVANT, TALA, Finance, Upstart, INSIKT, Upgrade, 100Credit, WeLab, Weeshi, TrueAccord, MoneyLion, Active.AI, aire, cignifi

INSURANCE
 Metromile, Lemonade, CYENCE, Shift Technology, TRACTABLE

STORAGE
 aws, Google Cloud, Microsoft Azure, IBM, PURE STORAGE, ALLUXIO, nimblestorage, Qubole, COHERITY

CLUSTER SVCS
 aws, kubernetes, doctor, MESOSPHERE, Core

APP DEV
 Lightbend, Keen IO, rainforest, Upwork, appen, floure, eight, scale

CROWD-SOURCING
 amazon, mechanical Turk, Upwork, appen, floure, eight, scale

HARDWARE
 Google, TPU, arm, GRAPHCORE, intel AI, MYTHIC, NVIDIA, eSensar, Movius, BLAZINGDR

GPU DBs
 kinetica, MAEP, SOREM, MYTHIC, NVIDIA, eSensar, Movius, BLAZINGDR

SEARCH
 elasticsearch, ENDECA, COVEO, Lucidworks, ATTIVO, swiftype, algolia, alphaSense, MAANA, kibana

LOG ANALYTICS
 splunk, sumologic

SOCIAL ANALYTICS
 Hootsuite, sprinklr, NETBASE, synthesio, tracx, similarWeb

WEB / MOBILE / COMMERCE ANALYTICS
 Google Analytics, mixpanel, AMPITUDE, sumall, Airtable, RESCI, SIGOPT, granify, custora

HEALTHCARE
 flatiron, Clover, Xyruus, HealthTap, METABIOTA, Gingerio, Glow, babylon, 3D Med, zebra, PathAI, ovia, TEMPUS, patientslikeme, AiCure, RECURSION, prognos, enlitic, imago, Qventus, BAYLABS, ARTERYS, LUKA MEDX, MAGEN, Kang Health, PAIGE, DATAVAN, INNOVACOR, LeonTouS

LIFE SCIENCES
 Zandime, color, Carbonize, BenevolentAI, verily, WuXiNextCODE, ZEPHYR HEALTH, freonome, CLEAR PATH, CISCAN, Clear Labs, PILOT.AI, NIO, OPTIMUS, moovit, nexar, CITRINE, twoSTAR, Atomwise, deep phenomics, SOPHIA, OWKIN

TRANSPORTATION
 UBER, TESLA, ZOAX, CLEARPATH, nuTonomy, drive.ai, navto, AMOTIVE, PILOT.AI, NIO, OPTIMUS, moovit, nexar, comma.ai, nectradyne, Civil Maps, German Autolabs

AGRICULTURE
 FARMERS, Granular, JOHN DEERE, BLUE RIVER, FarmersEdge, FarmLogs, TARAMIS, GAMAYA, prospera

COMMERCE
 Instacart, STITCH FIX, RetailNext, Dia & Co, HowGood, heuritech

INDUSTRIAL
 AVEVA, SIEMENS, PREDIX, Schneider, GIGLOT, UPTAKE, OSI, SMART MACHINE, TACHYUS, Alliumium, SCORTEX, OTHER: eharmony, stem, reitab, Amper, ByteDance, hoppers, celecst, BOEYER, VERDIGRIS, duetto, Unbabel, Juledeck, Second Spectrum, remesh, ASAPP

CROSS-INFRASTRUCTURE/ANALYTICS

aws, Google Cloud, sas, IOI, DATA, vmware, TIBC, TERADATA, ORACLE, NetApp, syncsort, MAPR, cloudera

OPEN SOURCE

FRAMEWORK
 Hadoop, MapReduce, YARN, TEZ, Spark, CDAP

DATA ACCESS
 presto, SLAMDATA, APACHE DRILL, cassandra, SciDB, riak, HBASE, ACCUMULO

COORDINATION
 Apache Zookeeper, Apache Ambari

STREAMING
 Spark, Flink, kafka, druid, STORM

STAT TOOLS
 gijthon, ScalaLab, SciPy, julia

AI / MACHINE LEARNING / DEEP LEARNING
 TensorFlow, theano, Caffe, Microsoft Cognitive Toolkit, OpenAI, Apache SINGA, FeatureFu, mynet, neon, DSSTNE, mlilb, DLAI, MAHOUT, Aerosolve

SEARCH
 elasticsearch, Solr, Lucene

LOGGING & MONITORING
 elasticsearch, kibana, SENTRY, logstash, Prometheus

VISUALIZATION
 BeakerX, Rodeo

COLLABORATION
 jupyter, Zepplin, ANACONDA

SECURITY
 Apache Ranger, KNOX, Sentry

DATA SOURCES & APIs

HEALTH
 Apple, VALIDIC

IOT
 GE Digital

FINANCIAL & ECONOMIC DATA
 Bloomberg, THOMSON REUTERS, DOW JONES

AIR / SPACE / SEA
 Orbital Insight, planet, SKYTECH

PEOPLE / ENTITIES
 axiomatic, experian

LOCATION INTELLIGENCE
 FOURSQUARE, Mapbox, mapbox

OTHER
 qualtrics, DATA GOV

DATA RESOURCES

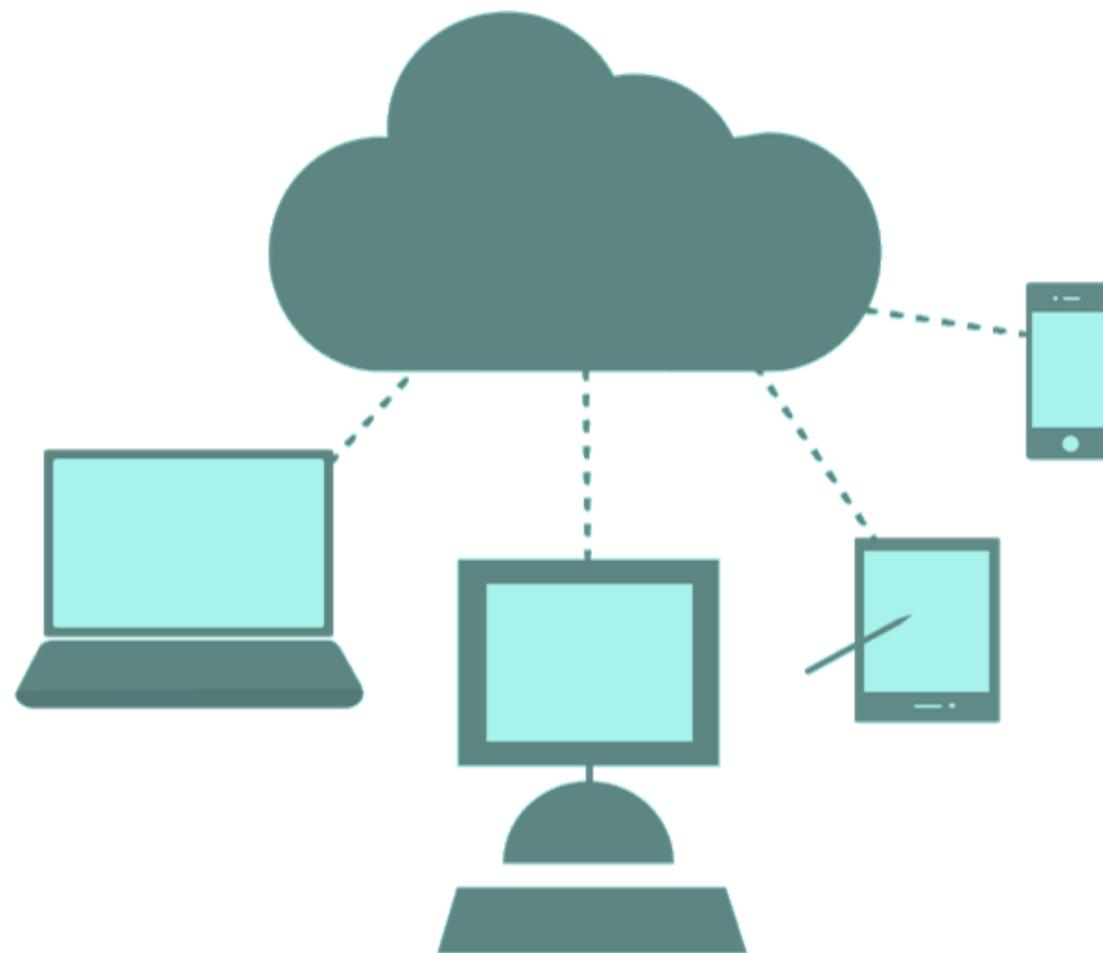
DATA SERVICES
 Palantir, LO, My Stamp

INCUBATORS & SCHOOLS
 GA, jalvanize

RESEARCH
 OpenAI, facebook, research, MIRI

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 хааS-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



Программист

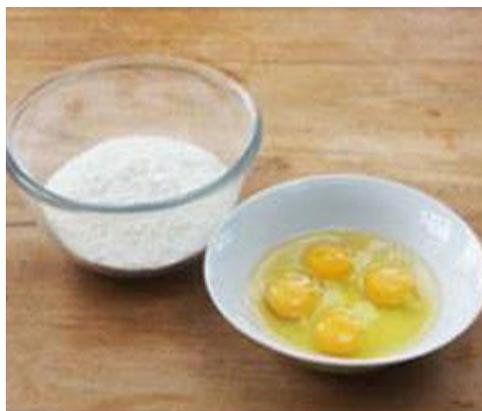
- Знает языки программирования
- Разбирается в математических методах
- Может формализовать логику

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 x
---
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 remainder is shown
below:
```{r, echo = FALSE}
smaller %>%
 ggplot(aes(carat)) +
 geom_freqpoly(binwidth = 0.01)
```
```

```
8:17 Chunk 1: setup
R Markdown

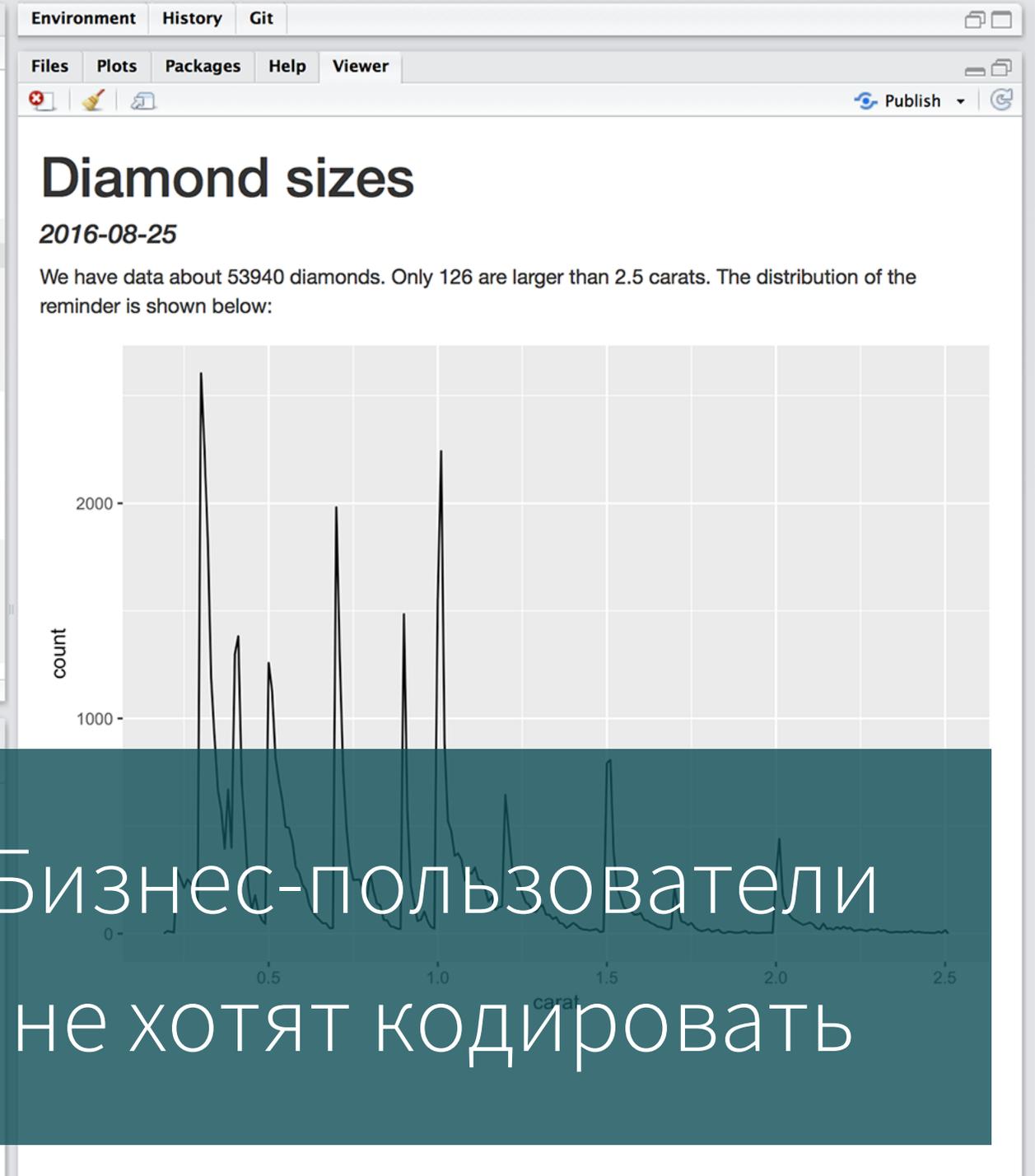
~/Documents/r4ds/r4ds/rmarkdown/
Platform: x86_64-apple-darwin13.4.0 (64-bit)

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.
```

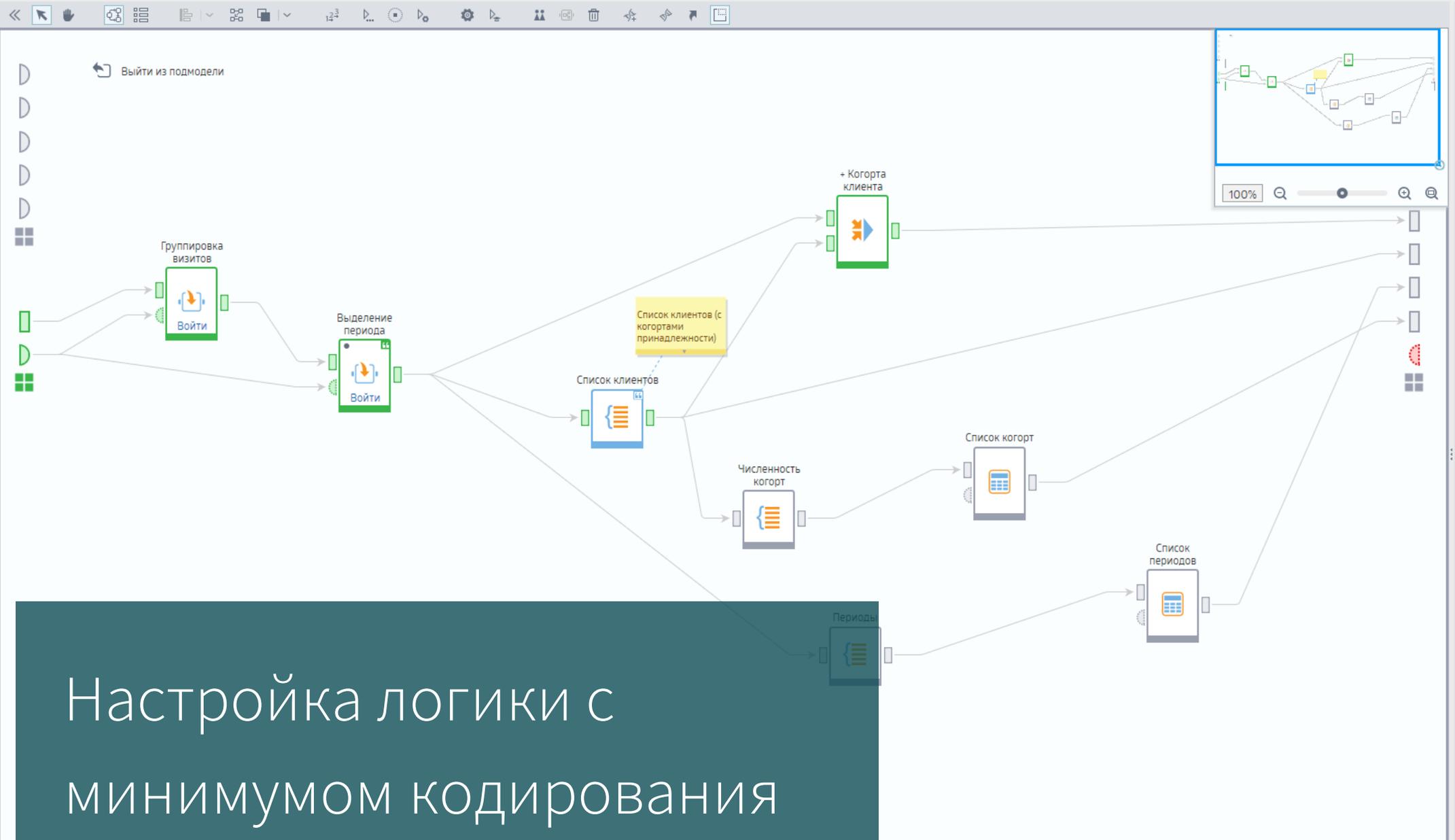


Бизнес-пользователи
не хотят кодировать

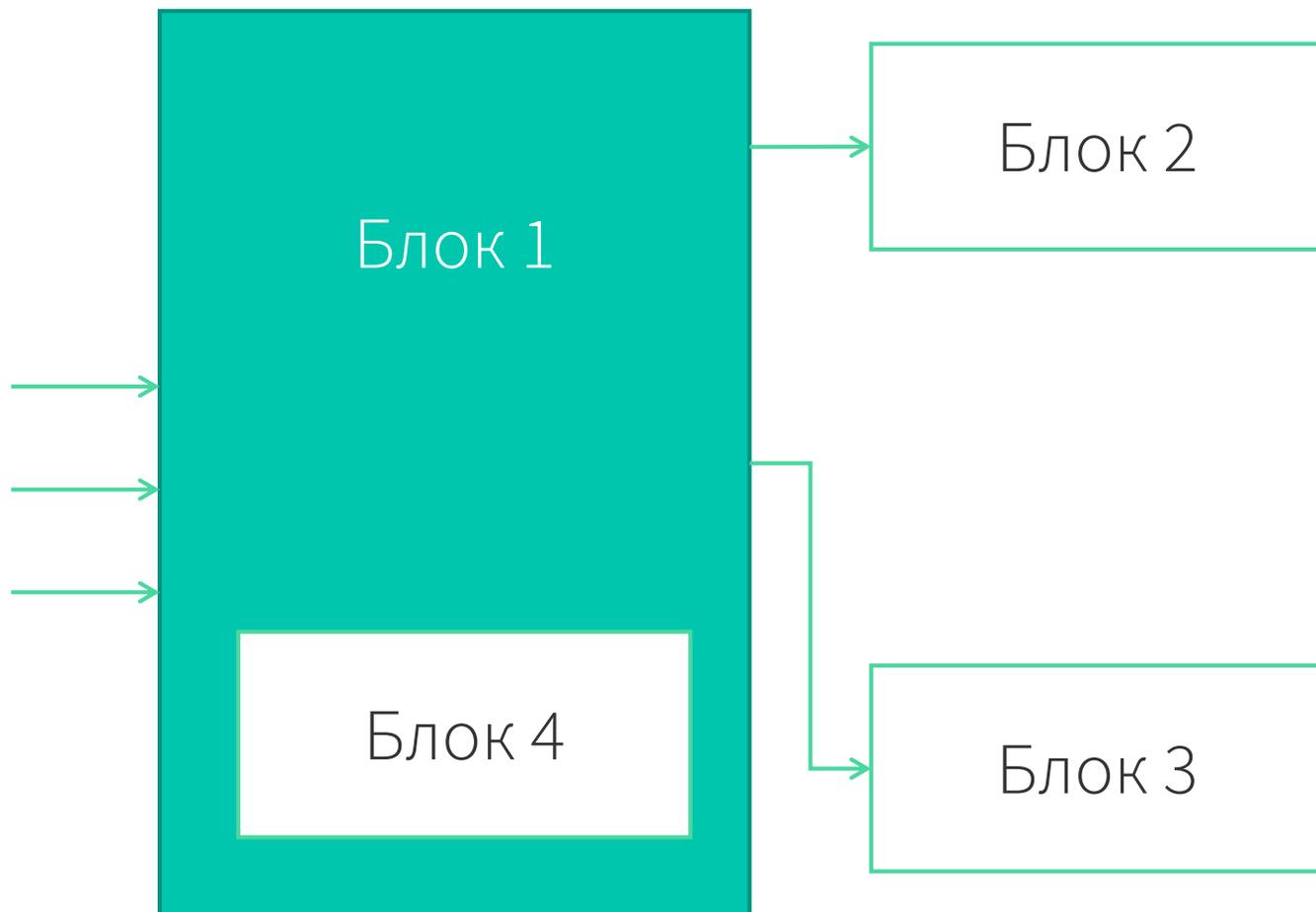
Схемы – естественное
описание логики анализа



- 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
 - FM Кластеризация
- Производные компоненты +
- Подключения +



Настройка логики с минимумом кодирования



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

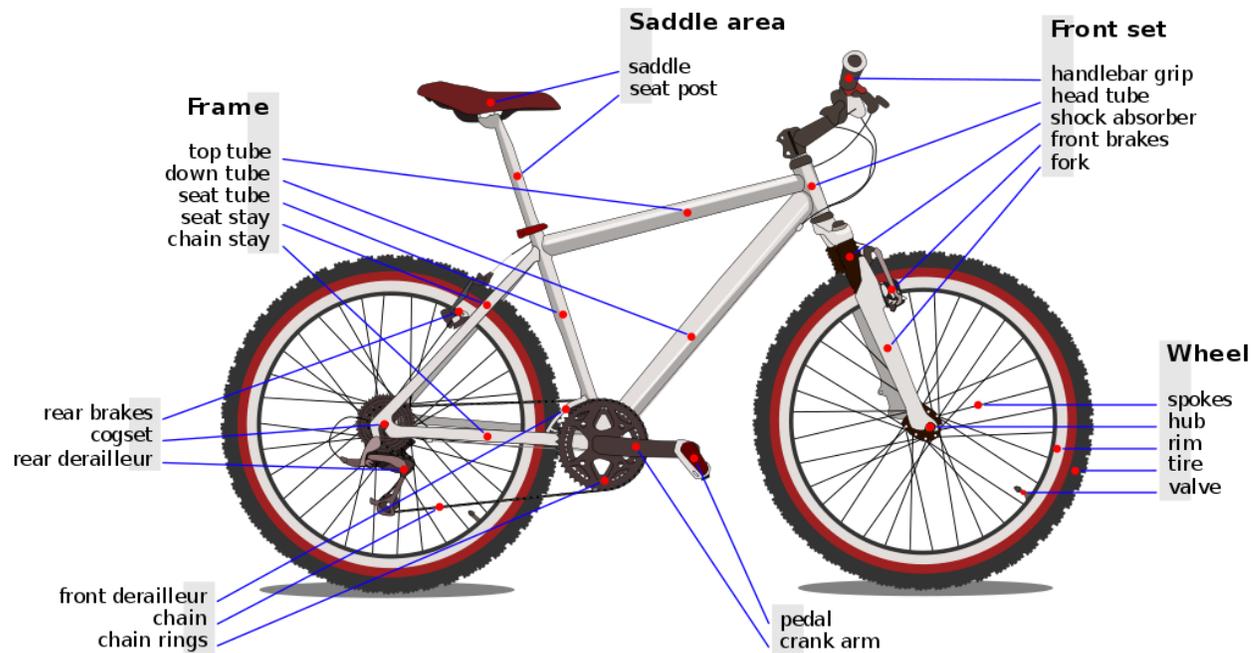
Минимум кодирования

Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms

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



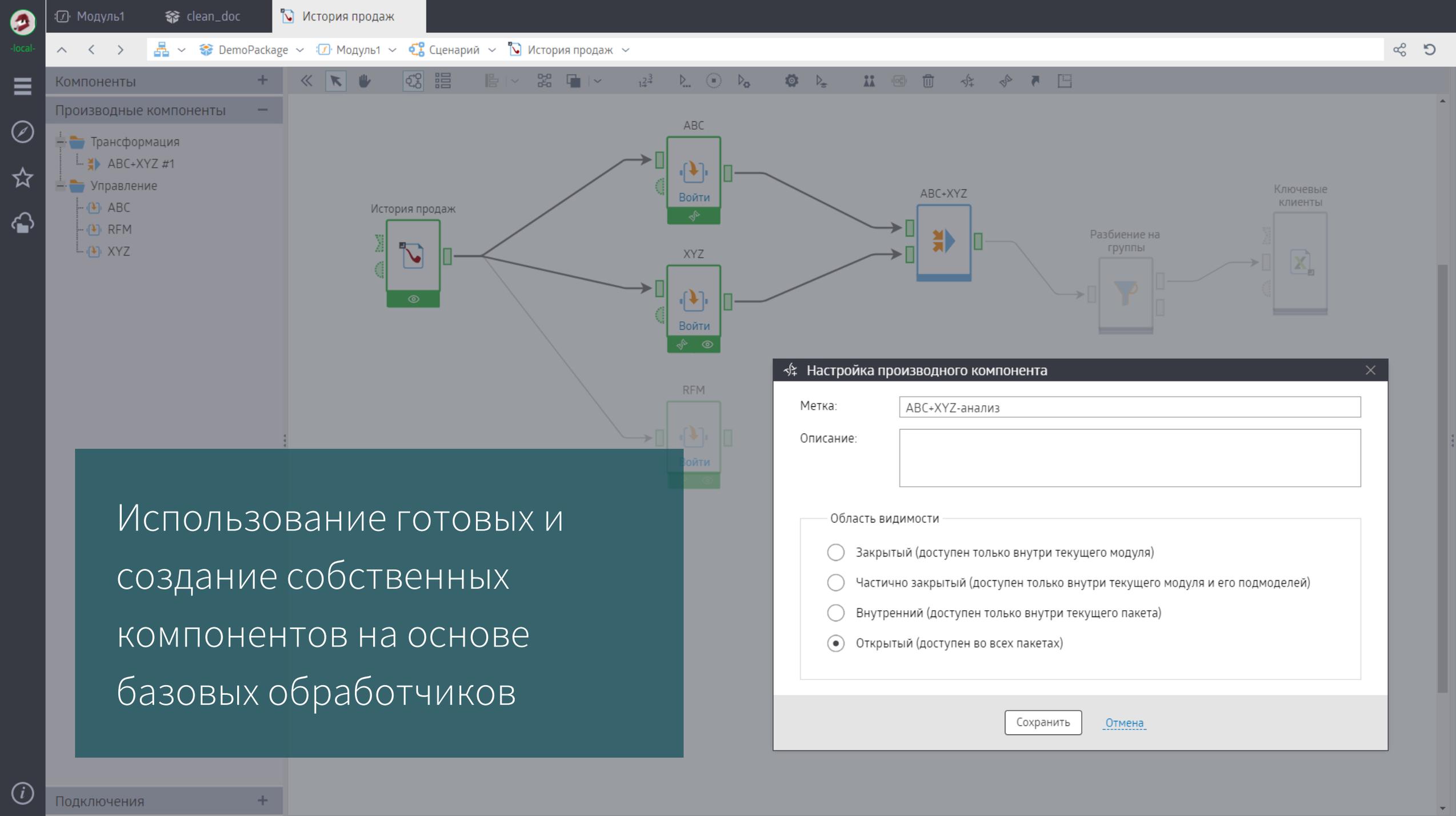
Повторное использование



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

Настройка производного компонента

Метка:

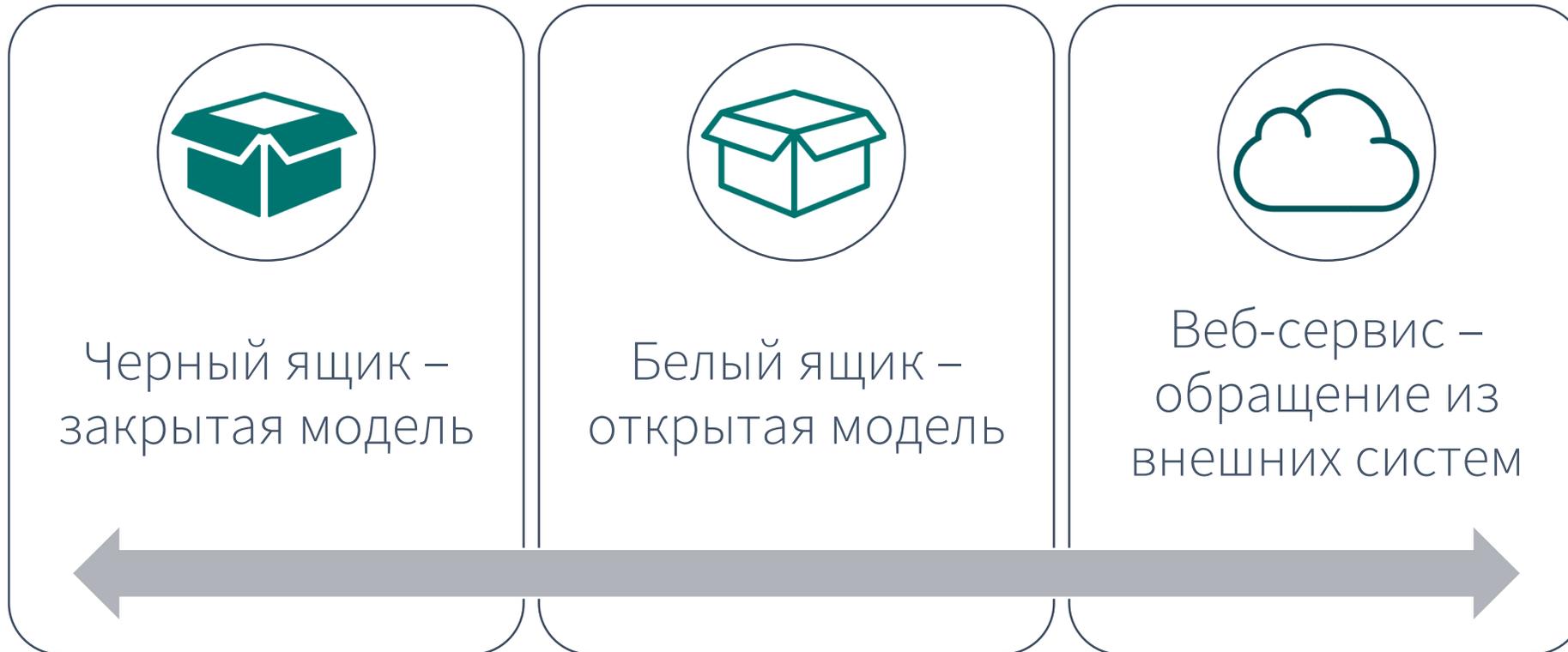
Описание:

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)

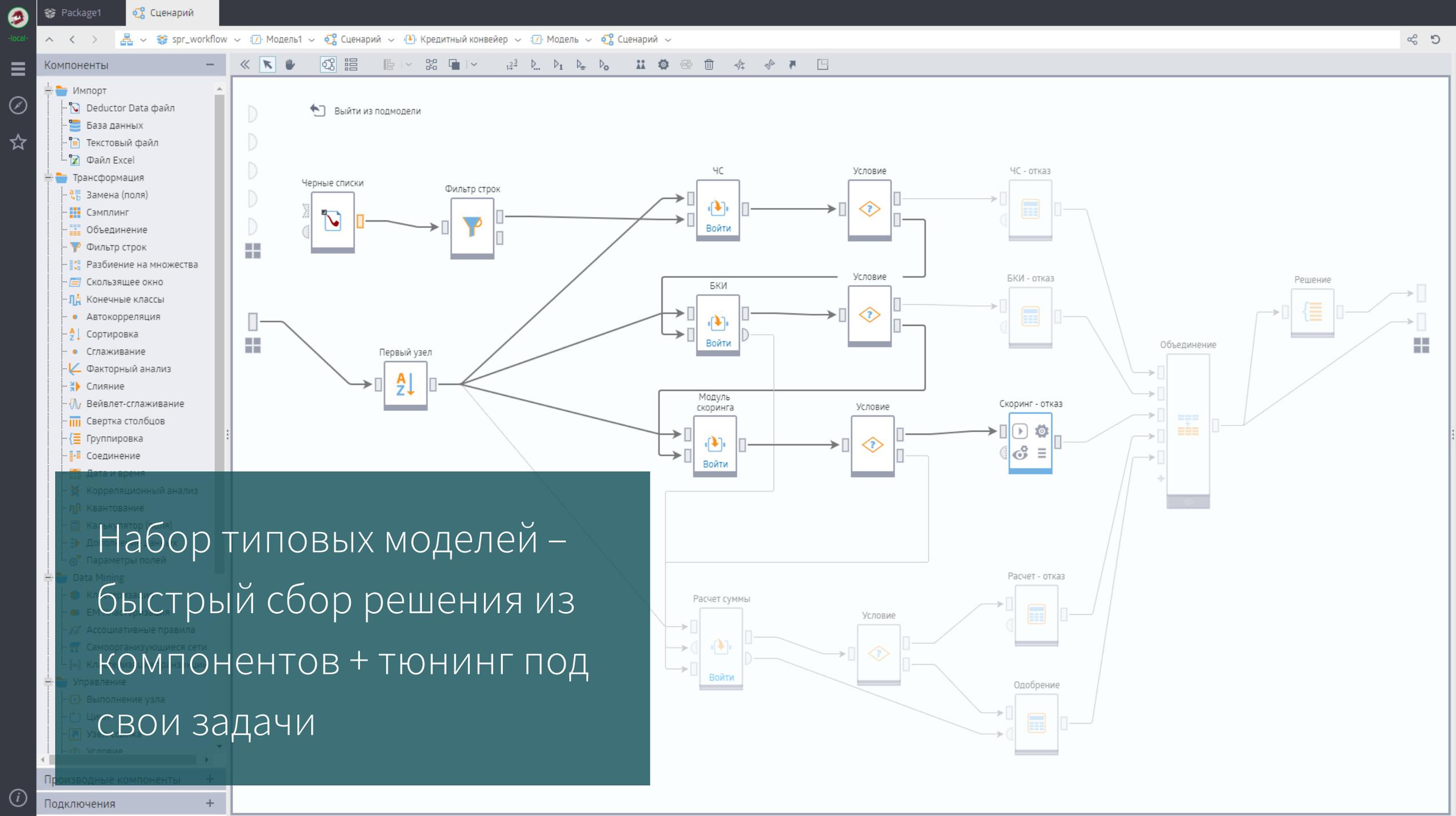
[Отмена](#)

Варианты использования компонента

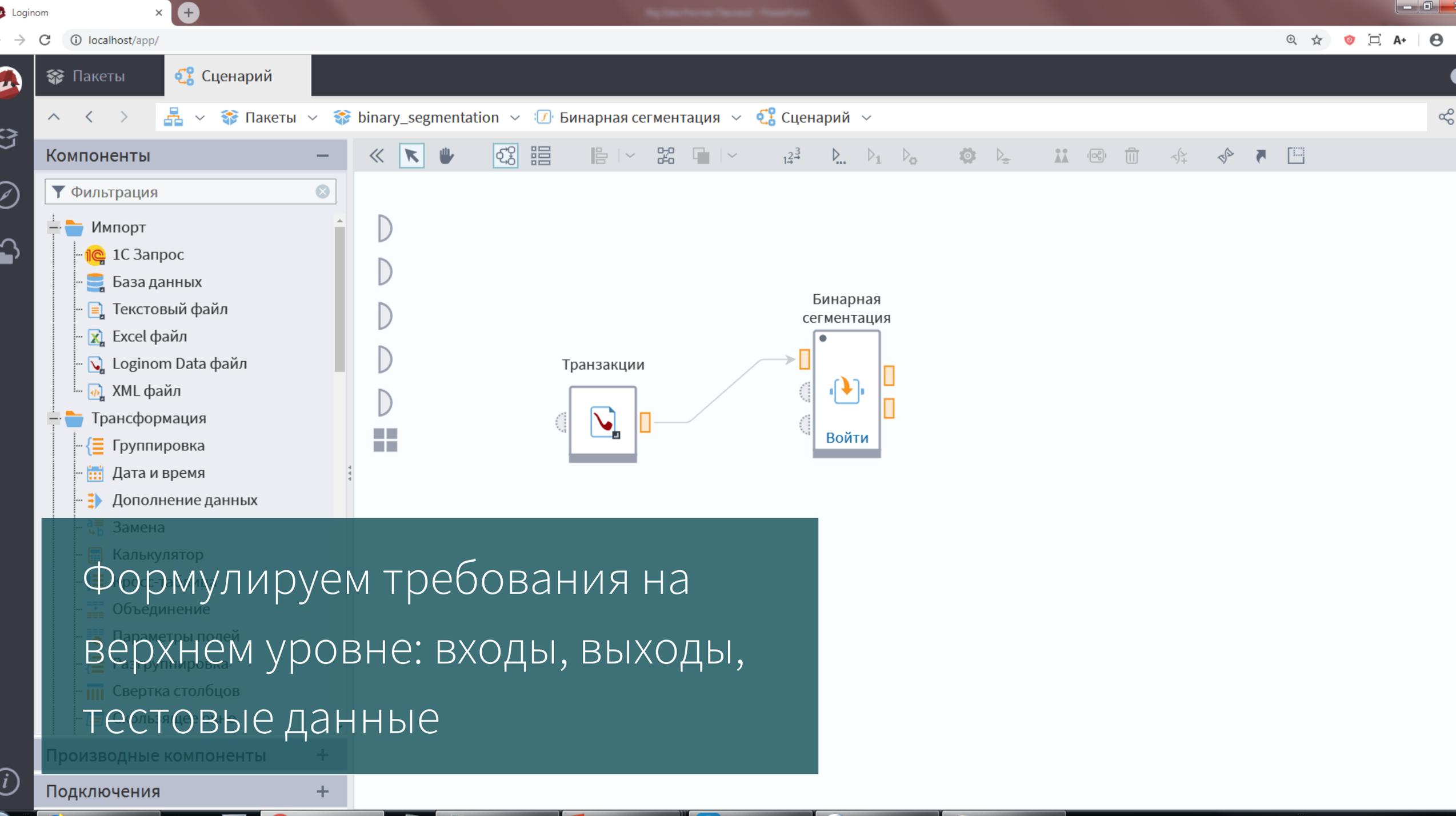


Объектная модель

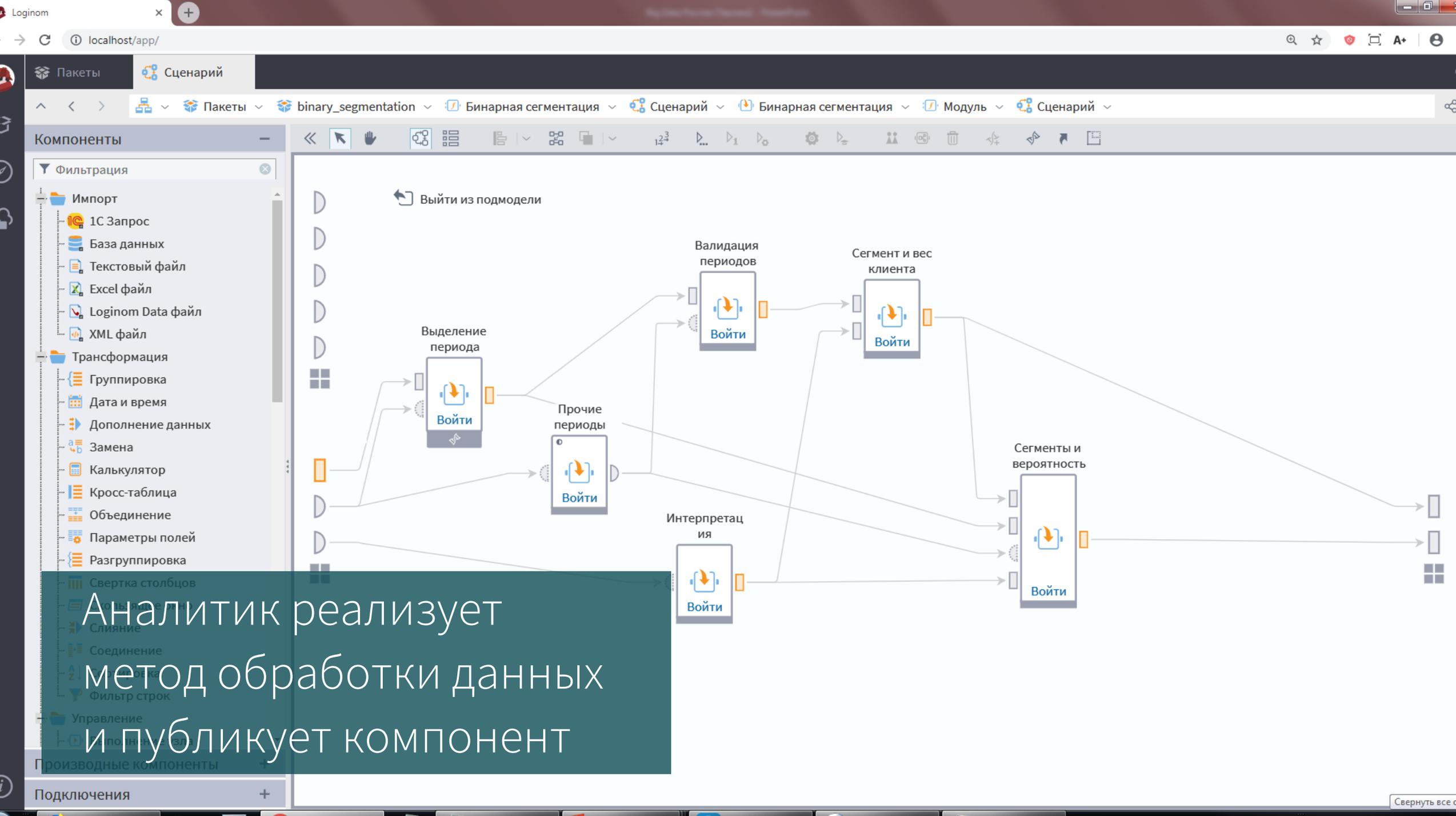
| Принцип | Назначение |
|--------------|--|
| Абстракция | Возможность оперировать блоком обработки данных как единым целым, не вдаваясь в особенности реализации |
| Инкапсуляция | Включение в модель как логики обработки, так и скрытых от внешнего пользователя данных |
| Наследование | Создание узла-наследника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



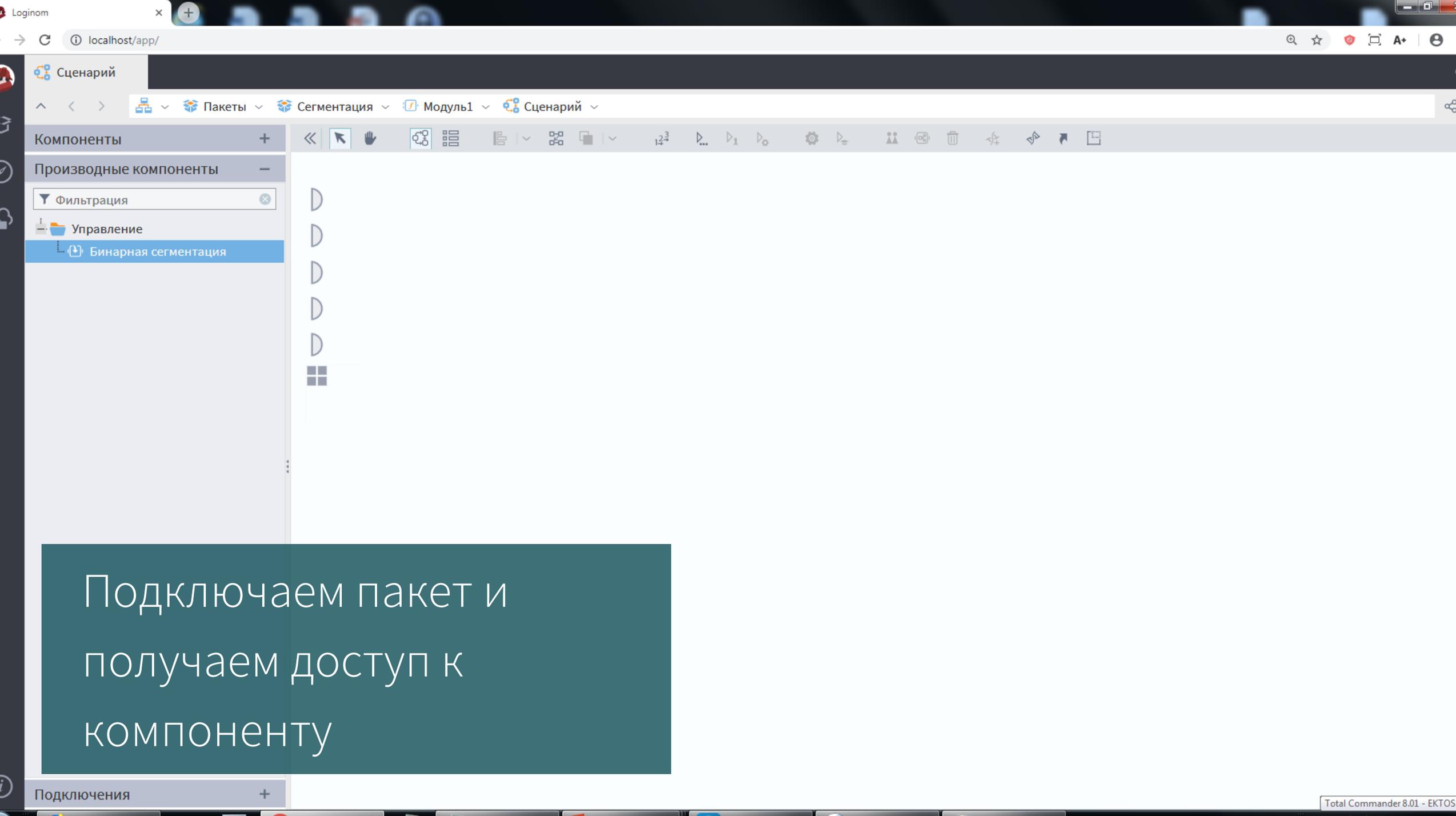
Набор типовых моделей –
быстрый сбор решения из
компонентов + тюнинг под
свои задачи



Формулируем требования на
верхнем уровне: входы, выходы,
тестовые данные



Аналитик реализует метод обработки данных и публикует компонент



Подключаем пакет и
получаем доступ к
компоненту

Подключения +

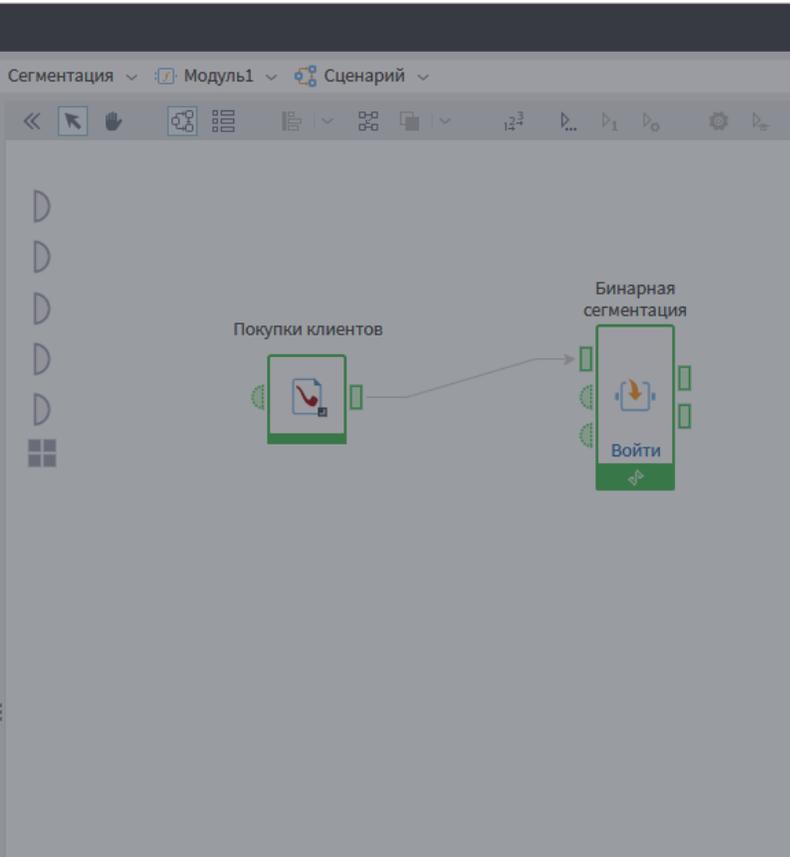
Сценарий

Производные компоненты

Фильтрация

Управление

Бинарная сегментация



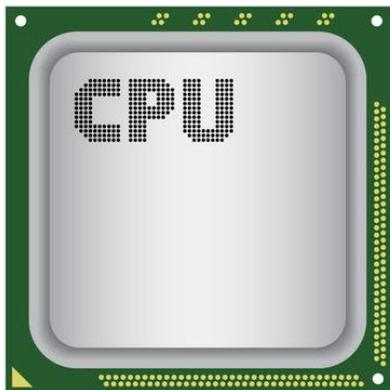
Бинарная сегментация • Набор данных • Быстрый просмотр данных

| # | ab Код сегмента | 12 Вес клиента | ab Идентификатор | ab 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 |

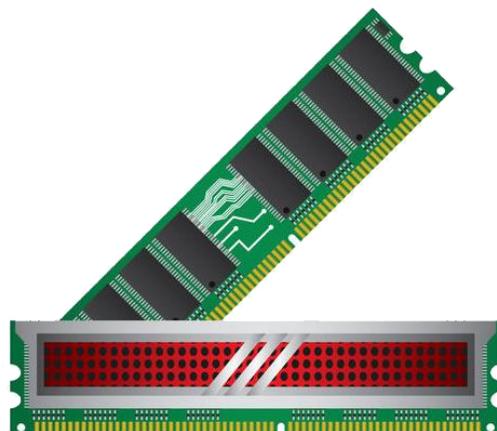
[Заккрыть](#)

Используем его на любых данных, удовлетворяющих требуемой структуре

Высокая производительность



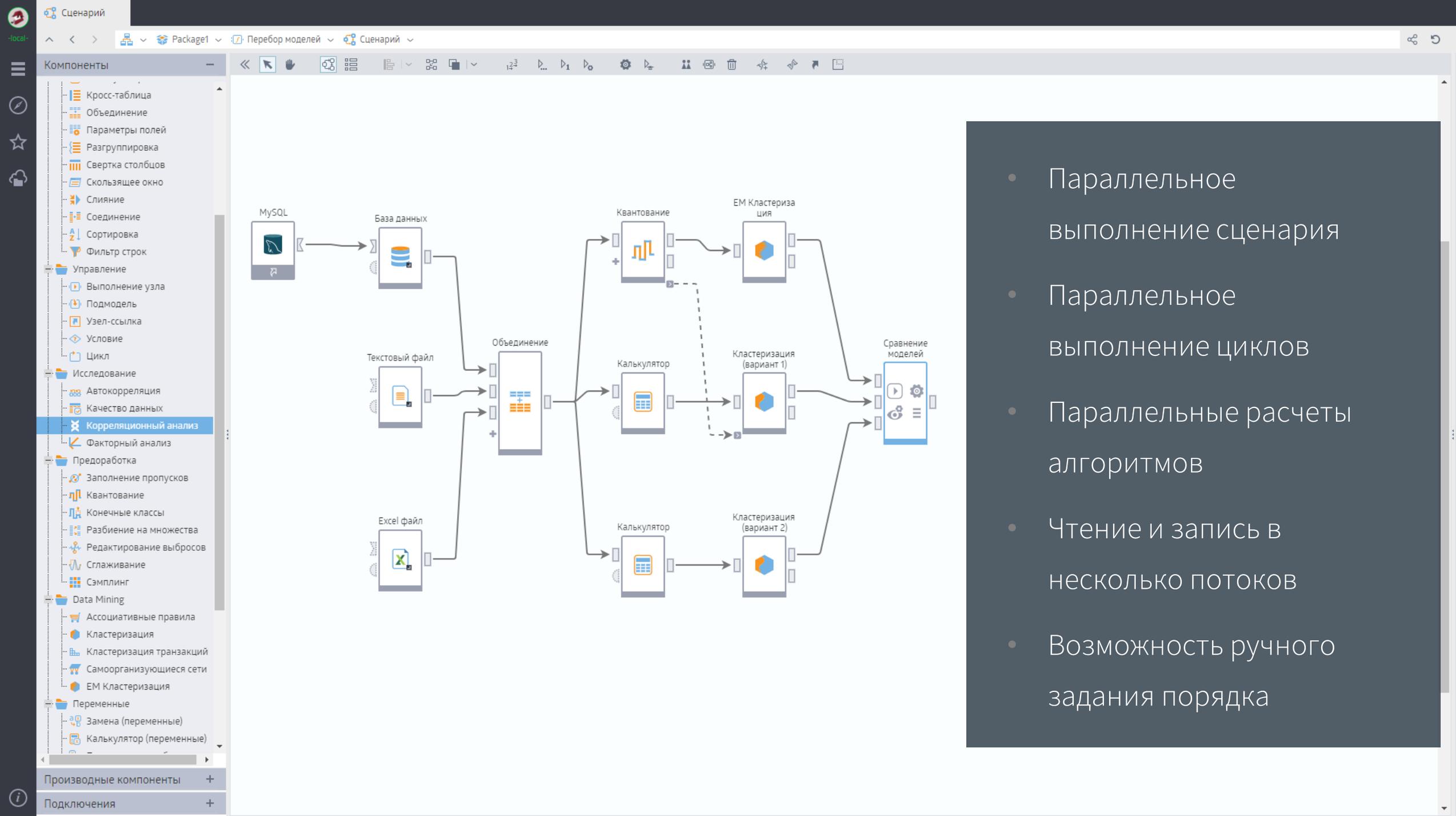
x64 – адресация
до 16 Tb RAM



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