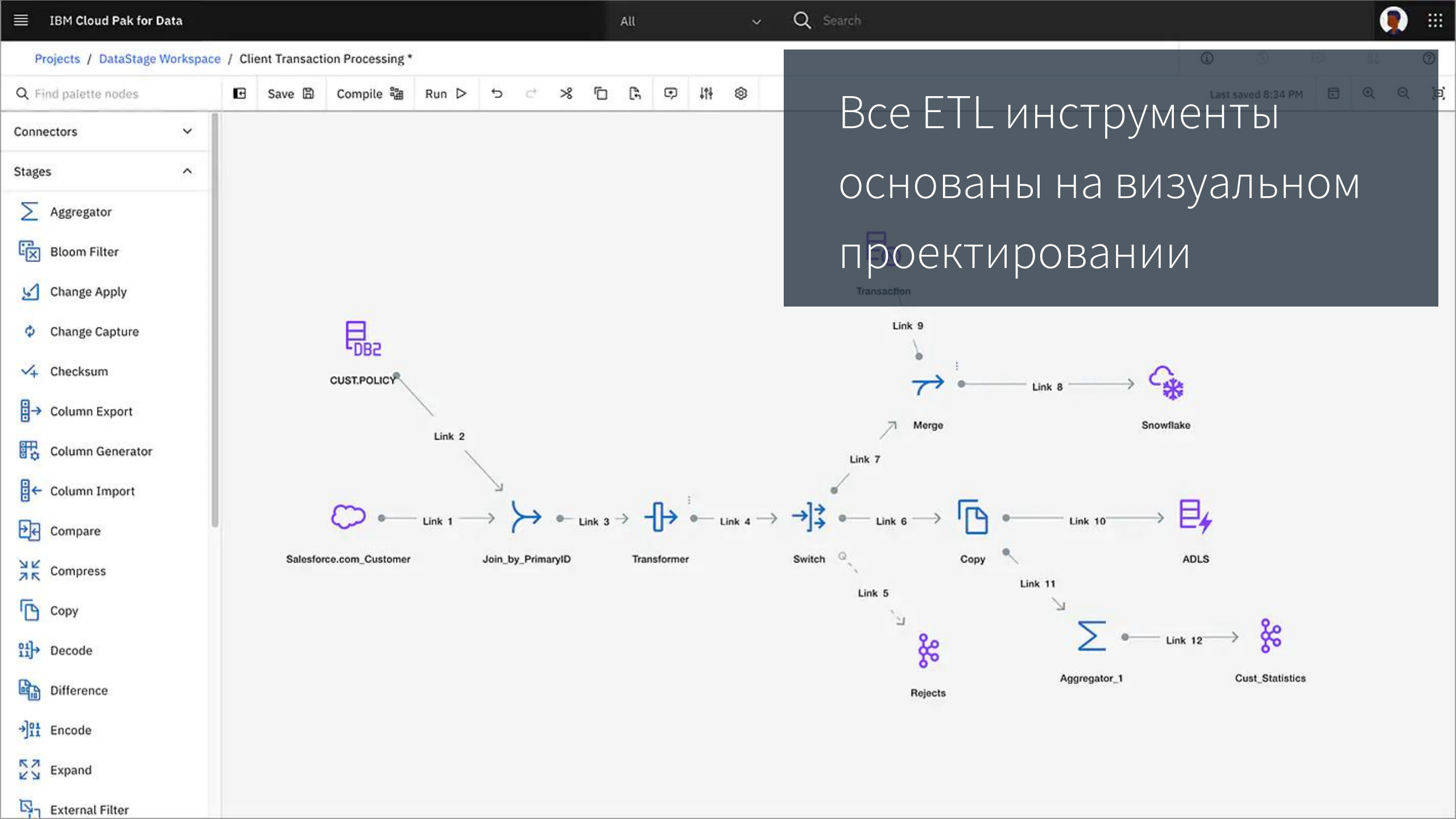
Быстрый ETL для PostgreSQL

Арустамов Алексей, Loginom Company

Extract, Transform, Load

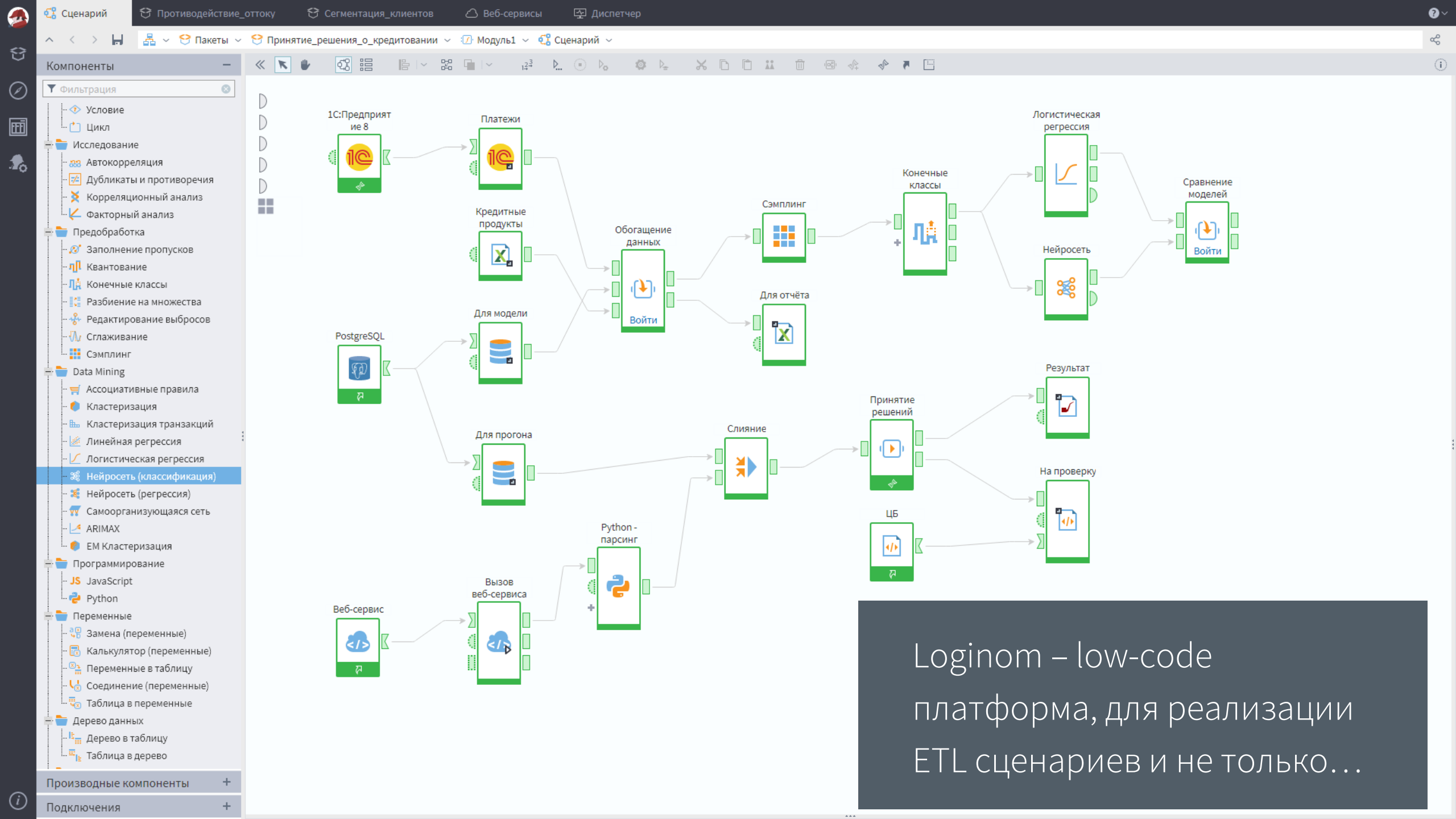
Разработчики баз данных с развитыми экосистемами предлагают инструменты для реализации ETL-процессов:

- Oracle Data Integrator
- Microsoft SQL Server Integration Services
- IBM DataStage

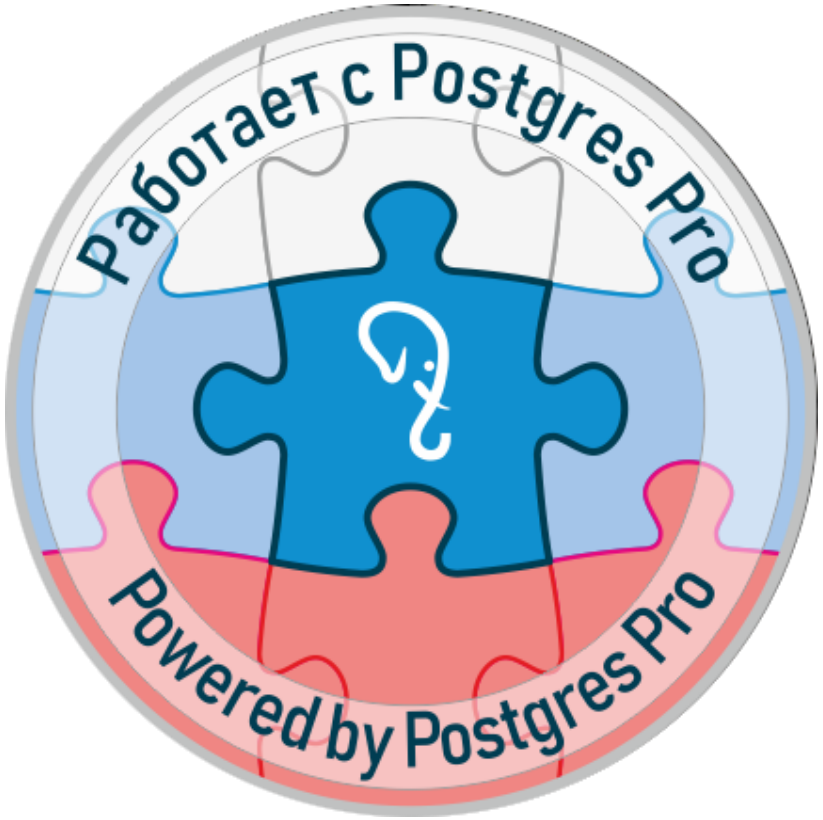

IBM Cloud Pak for Data
All
Search
Projects / DataStage Workspace / Client Transaction Processing *
Find palette nodes
Save
Compile
Run
Last saved 8:34 PM
Connectors
Stages
Aggregator
Bloom Filter
Change Apply
Change Capture
Checksum
Column Export
Column Generator
Column Import
Compare
Compress
Copy
Decode
Difference
Encode
Expand
External Filter
Все ETL инструменты основаны на визуальном проектировании
DB2
CUST.POLICY
Link 2
Link 9
Merge
Link 8
Snowflake
Link 7
Link 1
Link 3
Link 4
Link 6
Link 10
Salesforce.com_Customer
Join_by_PrimaryID
Transformer
Switch
Copy
ADLS
Link 5
Link 11
Rejects
Aggregator_1
Link 12
Cust_Statistics

Визуальный ETL

Реализация ETL-процесса, не требующая кодирования и (обязательного) знания SQL упрощает работу с базой данных, позволяя решать большинство задач пользователям без специальной подготовки. Self-service избавляет разработчиков от рутины.



Loginom – low-code платформа, для реализации ETL сценариев и не только…



СУБД **Postgres Pro Standard** и **Postgres Pro Enterprise** технологически полностью совместимы с аналитической low-code платформой **Loginom**

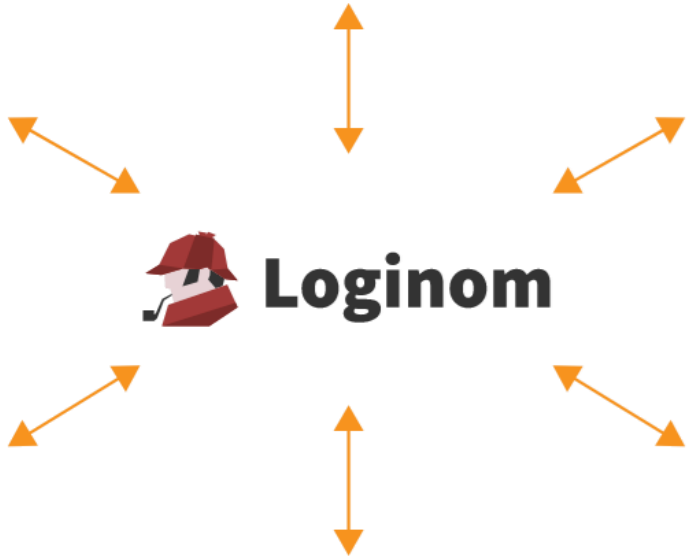
Извлечение и загрузка

Files

Excel, CSV, XML, Loginom Data File

Databases

Oracle, MS SQL, PostgreSQL, ClickHouse, BigQuery, MySQL...



ODBC sources

Hive, MongoDB, Amazon Redshift, SQL Azure...

Business applications

Creatio, Tableau

Web services

SOAP (WSDL), REST (JSON)

Data warehouses

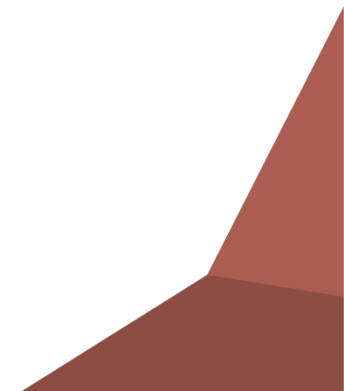
Oracle, MS SQL, Firebird

Обработка и трансформация

- Grouping
- Cross Table
- Union
- Join
- Row Filter
- Correlation Analysis
- Duplicates Detection
- Eliminate Outliers
- Sampling
- Imputation
- Clustering
- Logistic Regression
- Neural Network (Classification)
- ARIMAX
- XML Generation
- Transaction Clustering
- Smoothing
- 49+ components

Способы повышения скорости

1. Прямой доступ без драйверов
2. Чтение/запись пачками
3. Пул подключений
4. Параллелизм
5. In-memory обработка
6. Ленивые вычисления
7. Паттерн MapReduce



Прямой доступ к PostgreSQL

1. Не требуется установка клиента
2. Поддержка СУБД версий от 8.0 до 14.4
3. Аутентификация
    1. До версии Loginom 6.5 – md5
    2. С версии Loginom 6.6 – md5 + scram-sha-256
4. Поддерживается Greenplum и Arenadata

Чтение/запись пачками

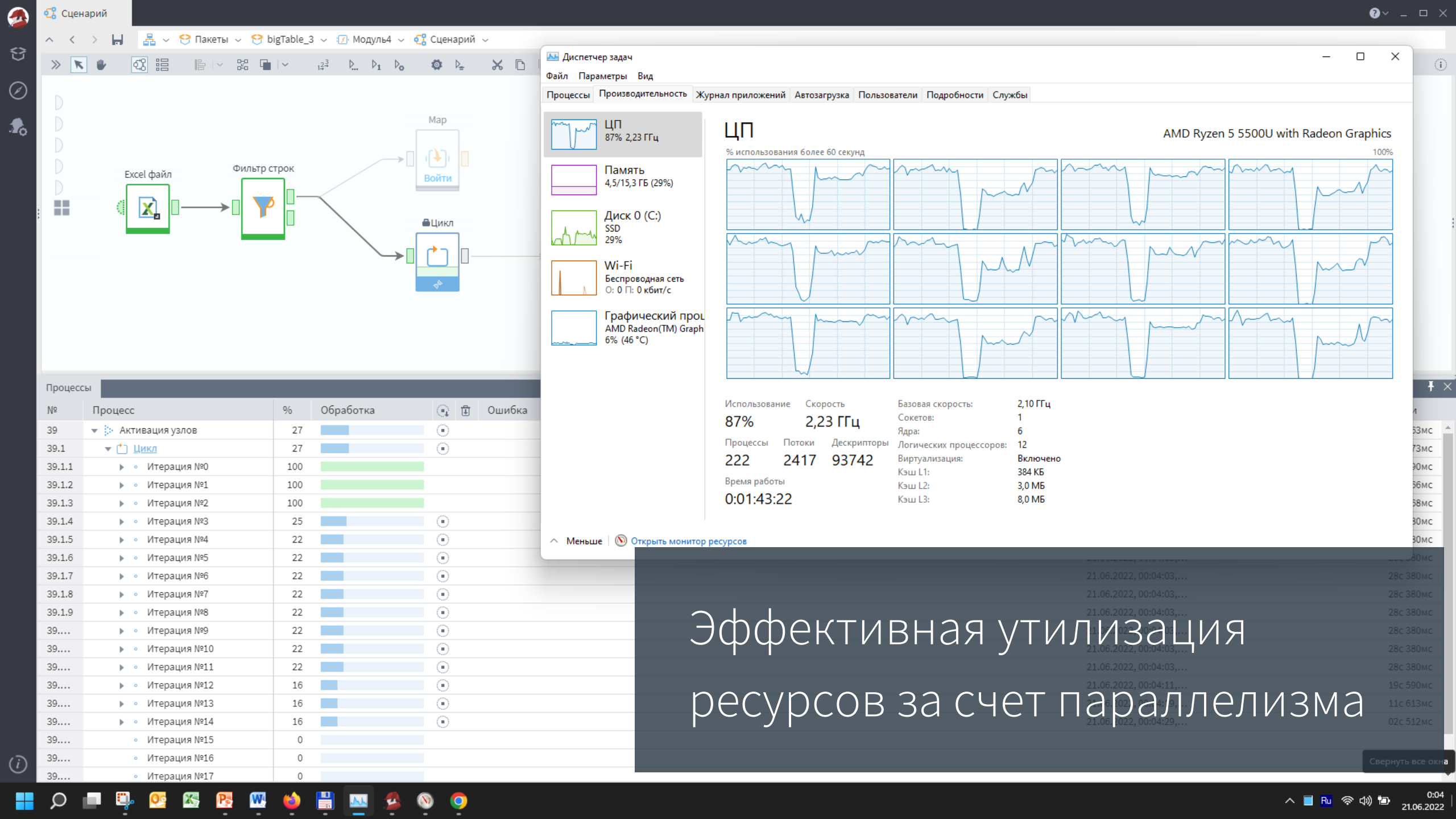
1. Чтение и запись производится пачками

Операция	Стандарт (сек.)	Пакетный (сек.)
Insert	346.7	1.69
Update	334.4	4.59
Delete	373.7	2.05

2. Размер пачки – 1000 строк

Пул подключений к PostgreSQL

1. Пул подключений уменьшает время затрачиваемое на повторное подключение. Подключение держится после завершения выполнения сценария, чтобы ускорить пакетное выполнение
2. При отсутствии подключения оно создается, по завершении выполнения запросов подключение возвращается в пул, при следующем обращении ищем есть ли в пуле подключение с такими же параметрами и используем его.
3. Уменьшается количество создаваемых подключений. Есть ограничение на количество одновременных подключений , по умолчанию 100. Если сценарий выполняется последовательно то одно подключение, если параллельно, то столько подключений, сколько параллельных потоков.

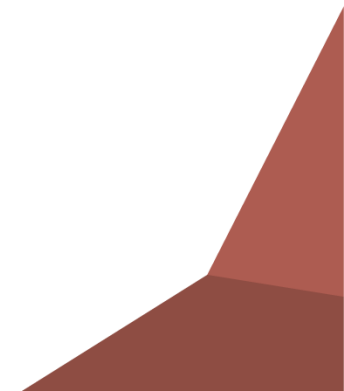

Сценарий
Пакеты
bigTable_3
Модуль4
Сценарий
Excel файл
Фильтр строк
Map
Войти
Цикл
Диспетчер задач
Файл
Параметры
Вид
Процессы
Производительность
Журнал приложений
Автозагрузка
Пользователи
Подробности
Службы
ЦП
87% 2,23 ГГц
Память
4,5/15,3 ГБ (29%)
Диск 0 (C:)
SSD
29%
Wi-Fi
Беспроводная сеть
О: 0 П: 0 кбит/с
Графический проц
AMD Radeon(TM) Graph
6% (46 °C)
ЦП
AMD Ryzen 5 5500U with Radeon Graphics
% использования более 60 секунд
100%
Использование
87%
Скорость
2,23 ГГц
Процессы
222
Потоки
2417
Дескрипторы
93742
Время работы
0:01:43:22
Базовая скорость: 2,10 ГГц
Сокетов: 1
Ядра: 6
Логических процессоров: 12
Виртуализация: Включено
Кэш L1: 384 КБ
Кэш L2: 3,0 МБ
Кэш L3: 8,0 МБ
Меньше
Открыть монитор ресурсов
Процессы
№
Процесс
%
Обработка
Ошибка
39 Активация узлов 27
39.1 Цикл 27
39.1.1 Итерация №0 100
39.1.2 Итерация №1 100
39.1.3 Итерация №2 100
39.1.4 Итерация №3 25
39.1.5 Итерация №4 22
39.1.6 Итерация №5 22
39.1.7 Итерация №6 22
39.1.8 Итерация №7 22
39.1.9 Итерация №8 22
39.... Итерация №9 22
39.... Итерация №10 22
39.... Итерация №11 22
39.... Итерация №12 16
39.... Итерация №13 16
39.... Итерация №14 16
39.... Итерация №15 0
39.... Итерация №16 0
39.... Итерация №17 0
Эффективная утилизация
ресурсов за счет параллелизма
Свернуть все окна
0:04
21.06.2022

In-memory

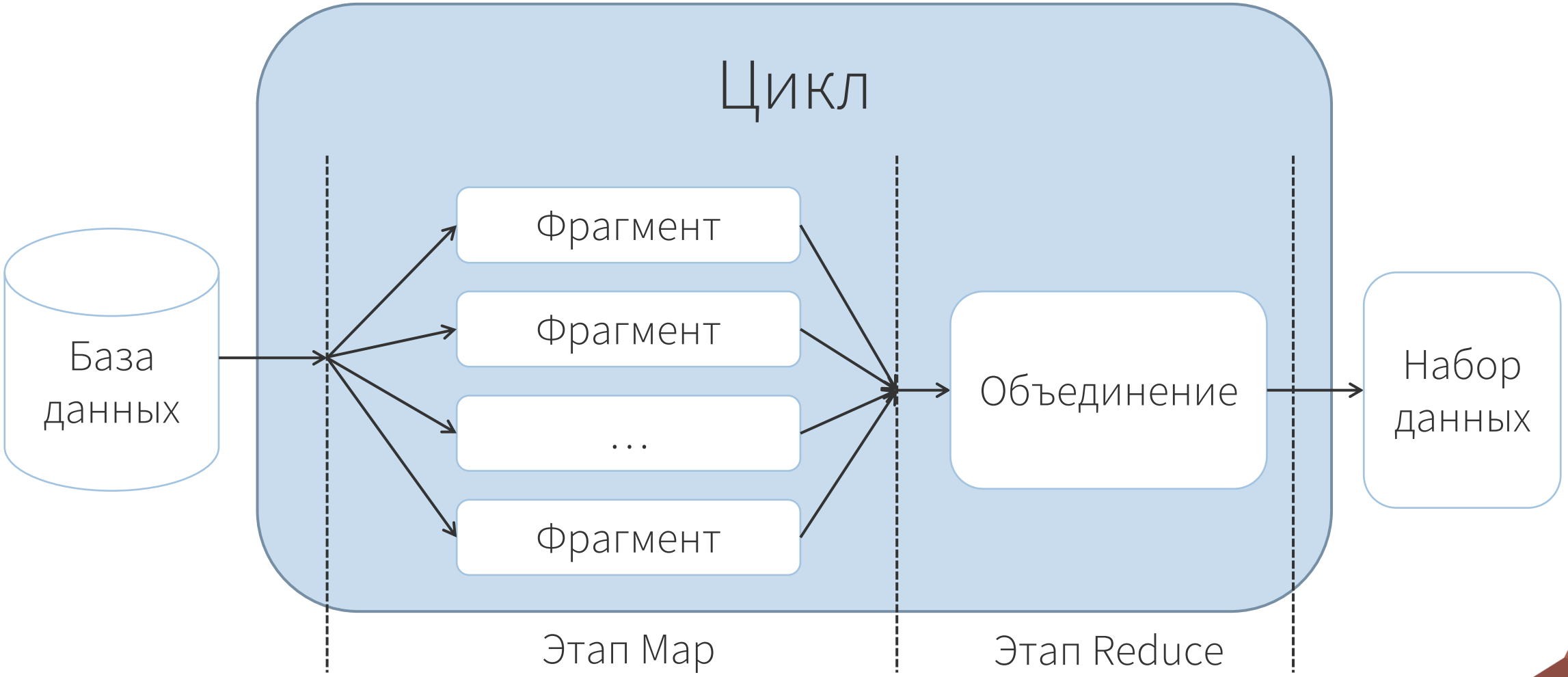
1. Платформа пытается все данные держать в памяти
2. Оптимизировано хранение наборов в памяти, в частности, по умолчанию хранятся только уникальные данные
3. При необходимости можно задать точки кэширования данных

Ленивые вычисления

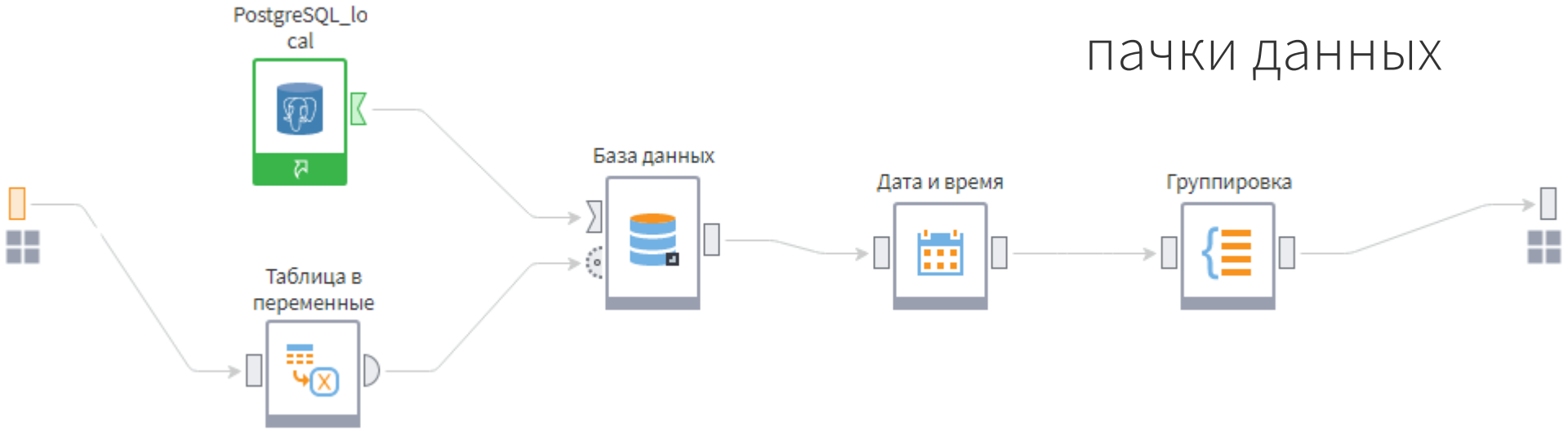
1. Вытягивать данные, а не толкать
2. Не читать пока не будет запрошено
3. Не считать пока не будет запрошено
4. Не показывать пока не будет запрошено



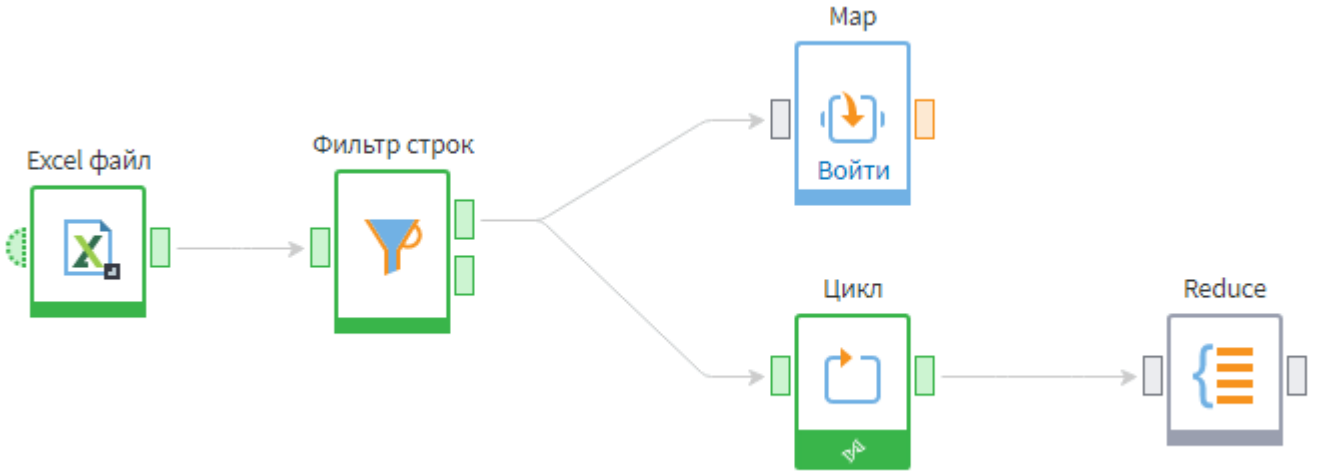
MapReduce в многопроцессорных системах



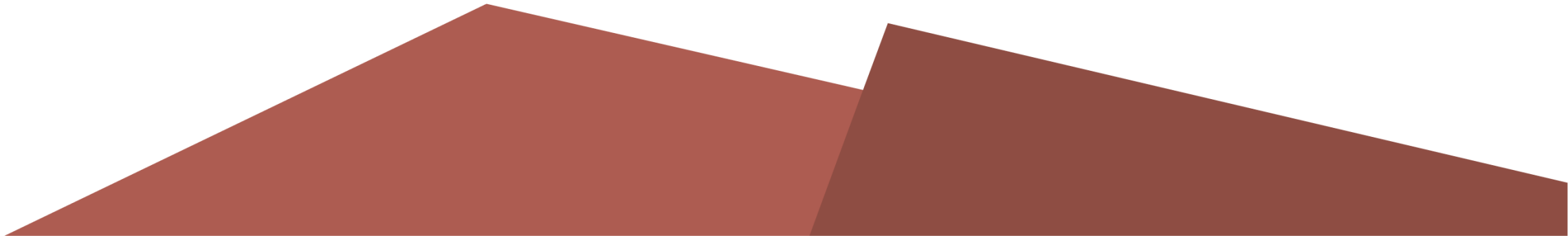
Map – обработка одной пачки данных

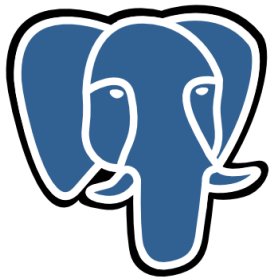


Выполнение в параллельном цикле и сбор итога



Демо реализации MapReduce

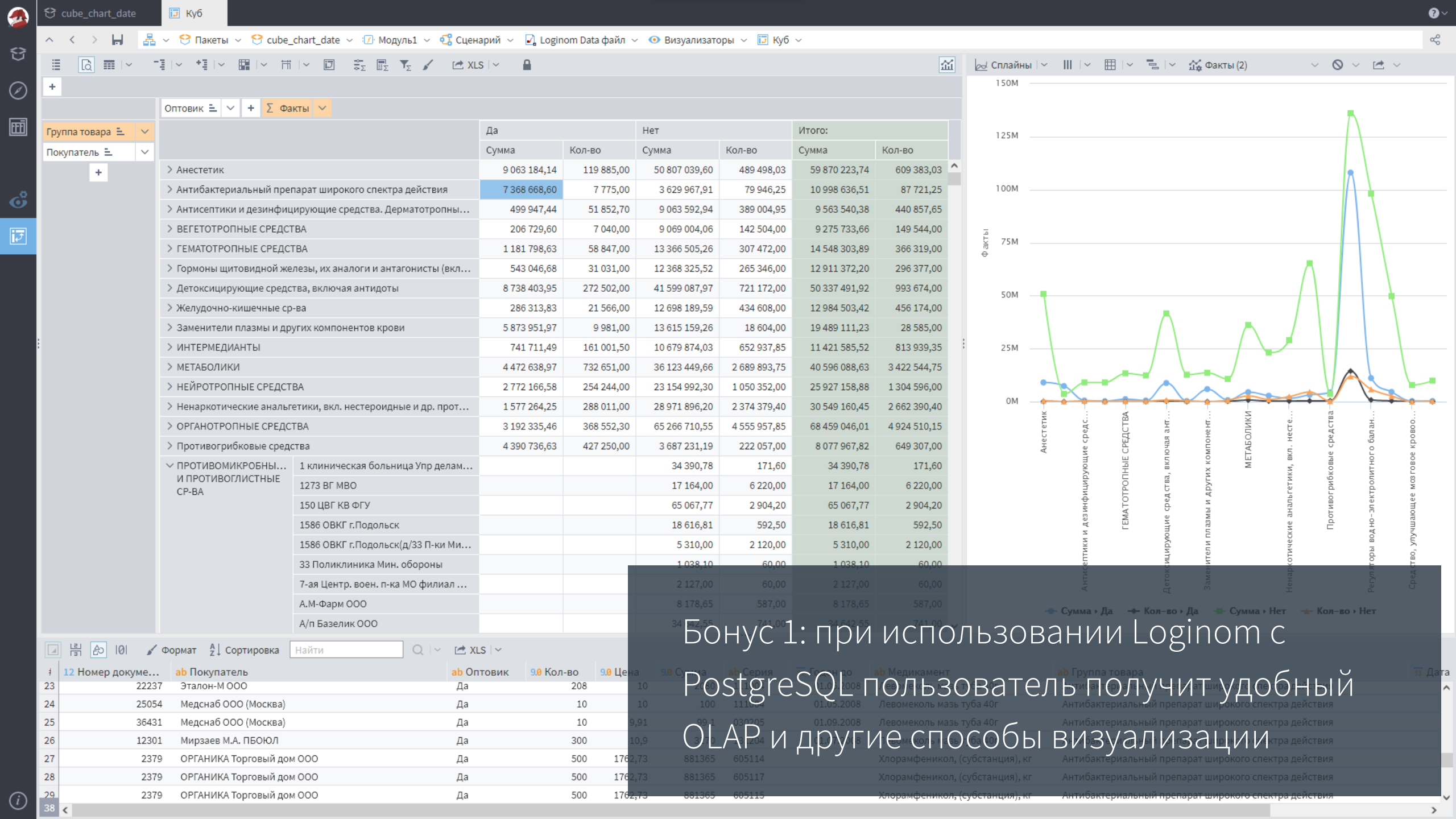




PostgreSQL

Loginom

Не только хранение данных и обработка запросов, но и сложный ETL + продвинутая аналитика



Бонус 1: при использовании Loginom с PostgreSQL пользователь получит удобный OLAP и другие способы визуализации

Сравнение скоринговых карт | Качество бинарной классификации

Пакеты | lsm_demo | Модуль1 | Сценарий | Моделирование и оценка качества | Модуль | Сценарий | Логистическая регрессия | Визуализаторы | Качество бинарной классификации

Выбор диаграммы

- (•) ROC-кривая
- ( ) PR-кривая
- ( ) Базовые показатели
- ( ) Диаграмма точности
- ( ) Диаграмма равновесия
- ( ) % распознанных событий
- ( ) Диаграмма роста
- ( ) Диаграмма отклика
- ( ) Диаграмма выигрыша

[x] Кумулятивная

10 диапазонов

Множества

[x] Обучающее [x] Тестовое

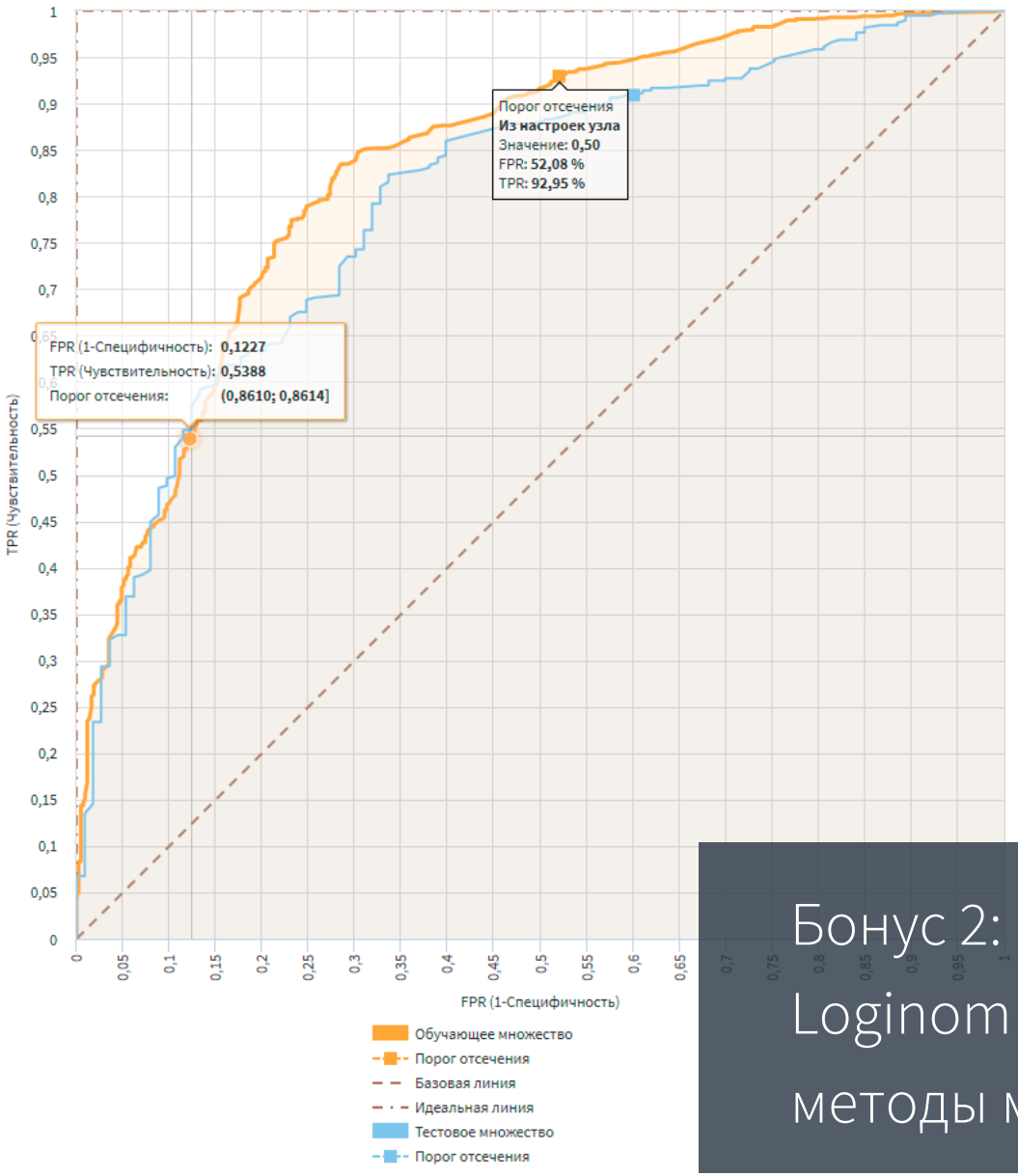
Порог отсечения

Из настроек узла

Значение порога:

ROC-кривая

Событие: Статус = хороший



Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,8362	0,8010
AUC PR	0,9424	0,9275
Коэффициент Джини	0,6724	0,6019
KS	55,2314	48,9691
Порог отсечения: Из настроек узла		
Значение	0,5000	0,5000
TPR (Чувствительность)	0,9295	0,9091
TNR (Специфичность)	0,4792	0,3982
FPR (1-Специфичность)	0,5208	0,6018
PPV	0,8658	0,8373
F1 Score	0,8965	0,8717
MCC	0,4604	0,3505

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	1 561	432	
Событие	1 451	225	1 676
Не-событие	110	207	317
Тестовое	385	113	
Событие	350	68	418
Не-событие	35	45	80

Распознано

Обучающее	83,19%
Тестовое	79,32%

Бонус 2: при использовании Loginom с PostgreSQL доступны методы машинного обучения

loginom.ru