IBM SolutionsConnect

от стратегии к практике

«Усиливая конкурентные преимущества в эпоху разумных решений»

---

4 марта 2015 года | Баку, отель “Fairmont Baku”

Производительная прогнозная аналитика с использованием специализированных OLAP-комплексов

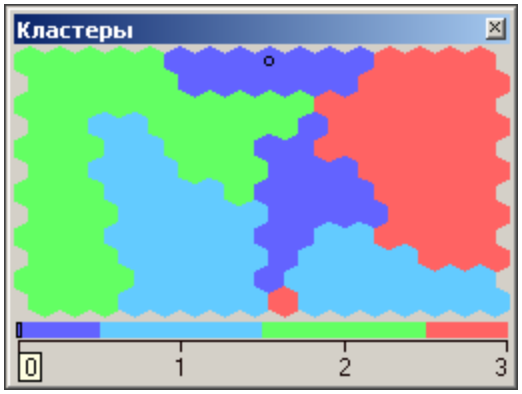
Александр Тимчур, IBM BigData Solutions

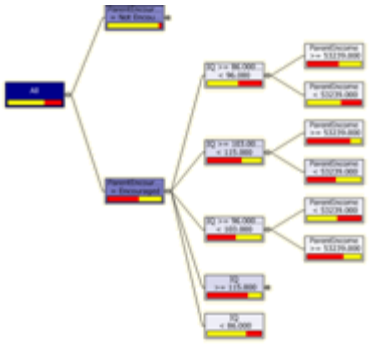
IBM

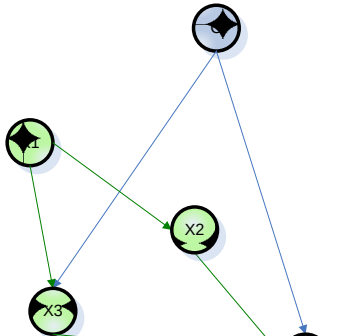
Что такое интеллектуальная аналитика?

Интеллектуальная аналитика - технология анализа информации с целью нахождения в накопленных данных ранее ***неизвестных, нетривиальных и практически полезных знаний***, необходимых для принятия оптимальных решений в различных областях человеческой деятельности

Данные → Знания → Действия → Ценность







IBM

Прогнозный или предсказательный анализ позволяет прогнозировать события

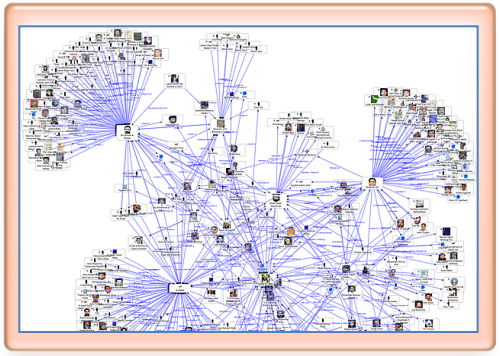
- Какова **вероятность** отклика клиента на конкретное предложение (телеком)?
- Какие клиенты **наиболее склонны** к уходу (телеком)?
- Какие **дополнительные** предложения следует **сделать** клиенту в данный момент (телеком, банки, розница)?
- Какой размер финансовых поступлений **ожидается** в следующем квартале (банки, банки, розница)?
- Какова **вероятность** возврата кредита (банки)?
- Какова **вероятность** выхода их строя оборудования в этом квартале (производство, телеком)?

Отчетность и аналитика

BI отчетность и Ad-Hoc анализ

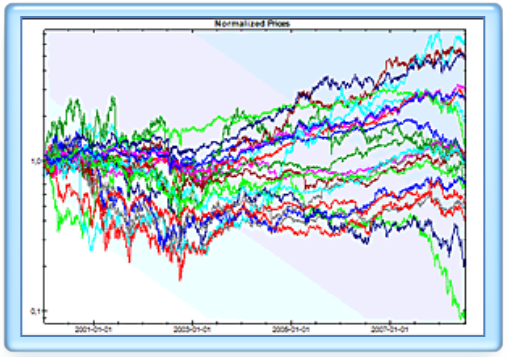
- Что случилось?
- Когда и где?
- Как много?

Предсказательная аналитика



- Что произойдет?
- Как будет влияние?

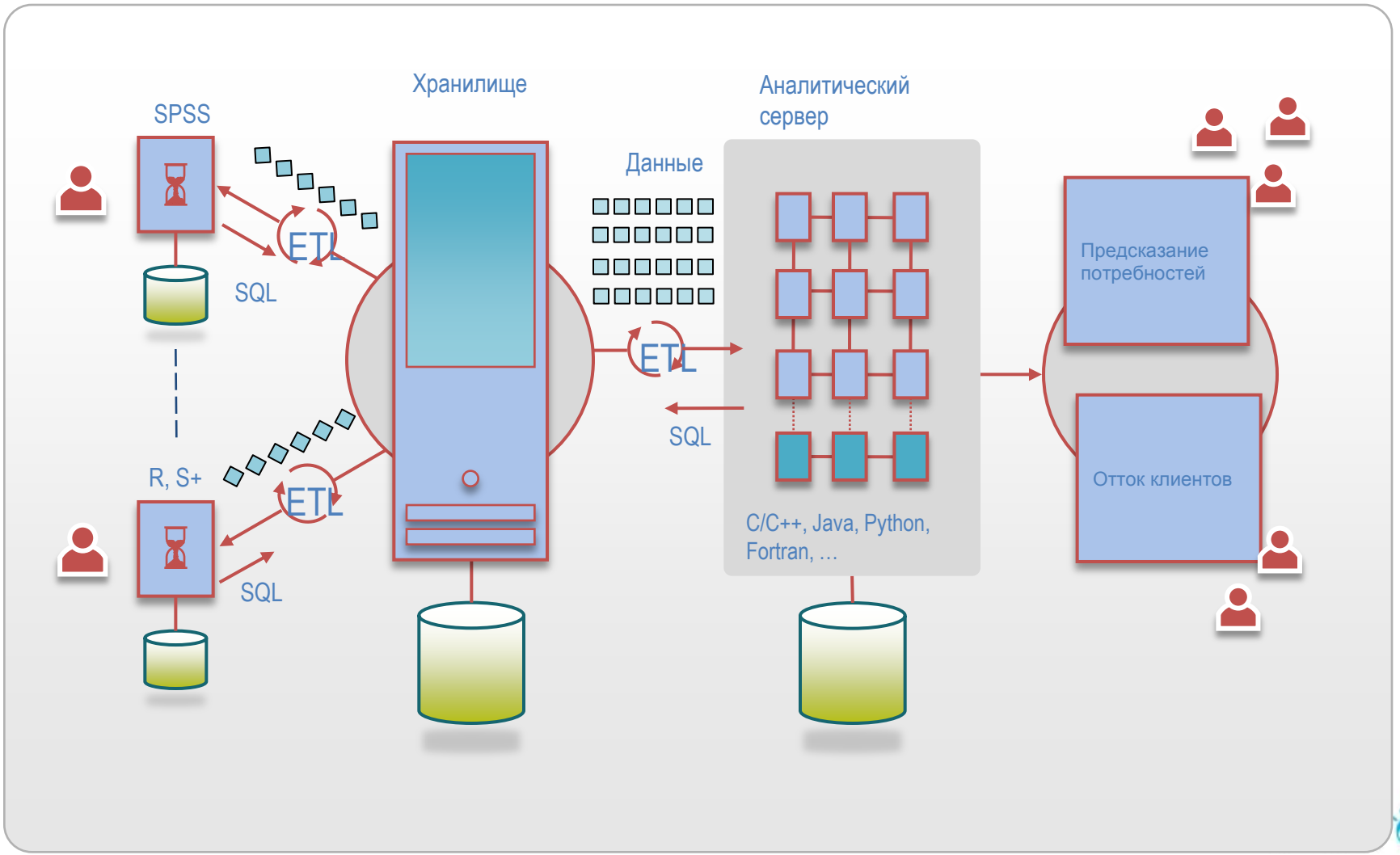
Оптимизация



- Каким будет лучший выбор?

IBM

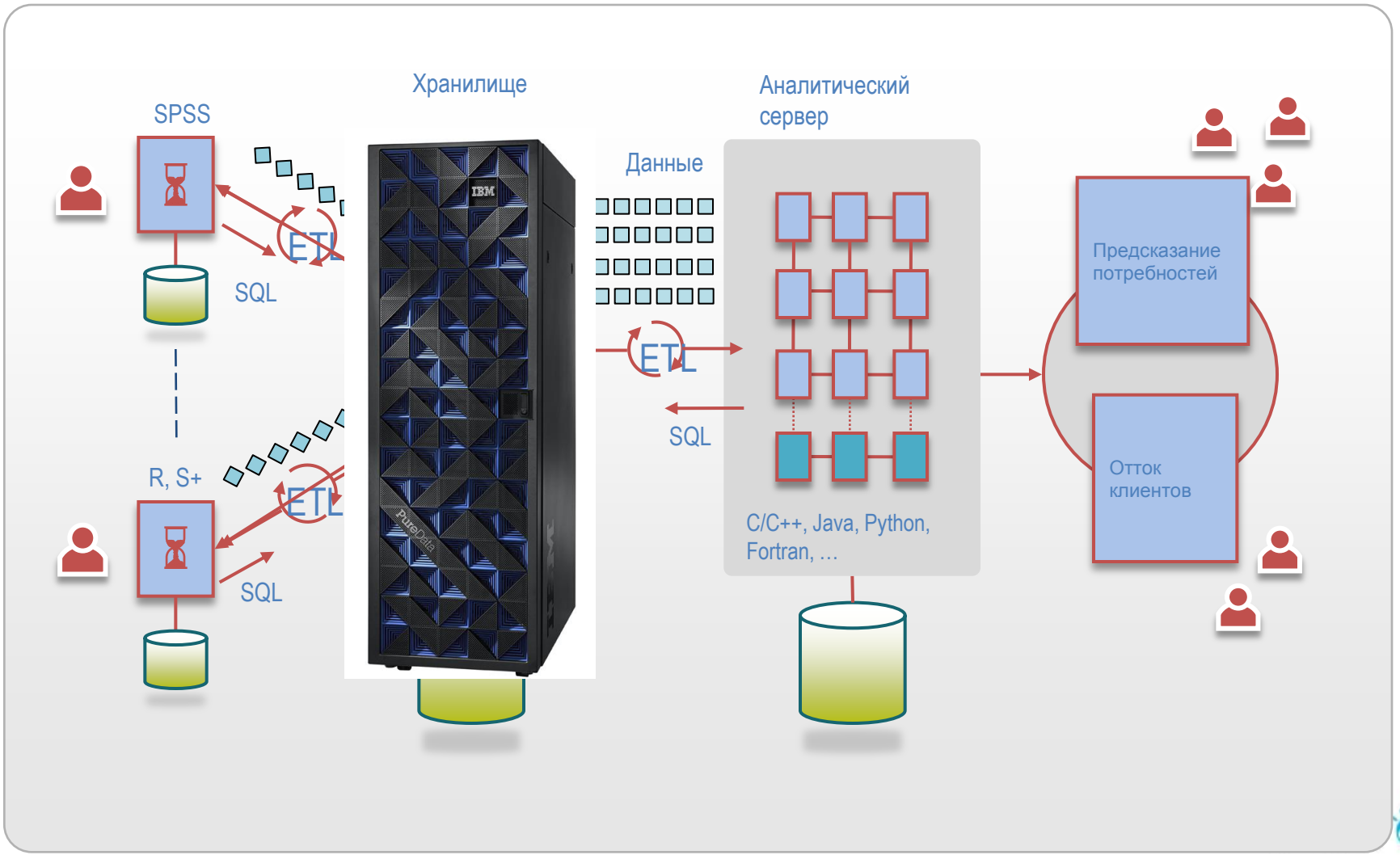
Интеллектуальная аналитика – традиционный путь



Проблемы анализа больших данных

- Благодаря большим объемам данных их перемещение на аналитические сервера для анализа становится слишком дорогим и сложным, стоимость лицензирования ПО для анализа растет. То есть **анализ** должен осуществляться **на стороне данных**.
  - Возникают требования к архитектуре, ориентированной на **производительность** анализа
  - Большинство классических алгоритмов анализа не предназначены для таких задач.
- Аналитическая платформа должна поддерживать **распределенные вычисления** на стороне данных.
- Требования к **квалификации аналитиков** растут.
  - Большинство аналитиков не обладают такими навыками.

Интеллектуальная аналитика с PureData for Analytics



IBM PureData for Analytics - интегрированный комплекс для хранилищ данных и интеллектуальной аналитики

- Оптимизированное, изначально спроектированное под **аналитику** ядро СУБД
- MPP-решение: все ресурсы (процессоры, диски) работают вместе над выполнением любого запроса
- Интегрированные СУБД, вычислительные мощности и система хранения данных
- Стандартные интерфейсы доступа

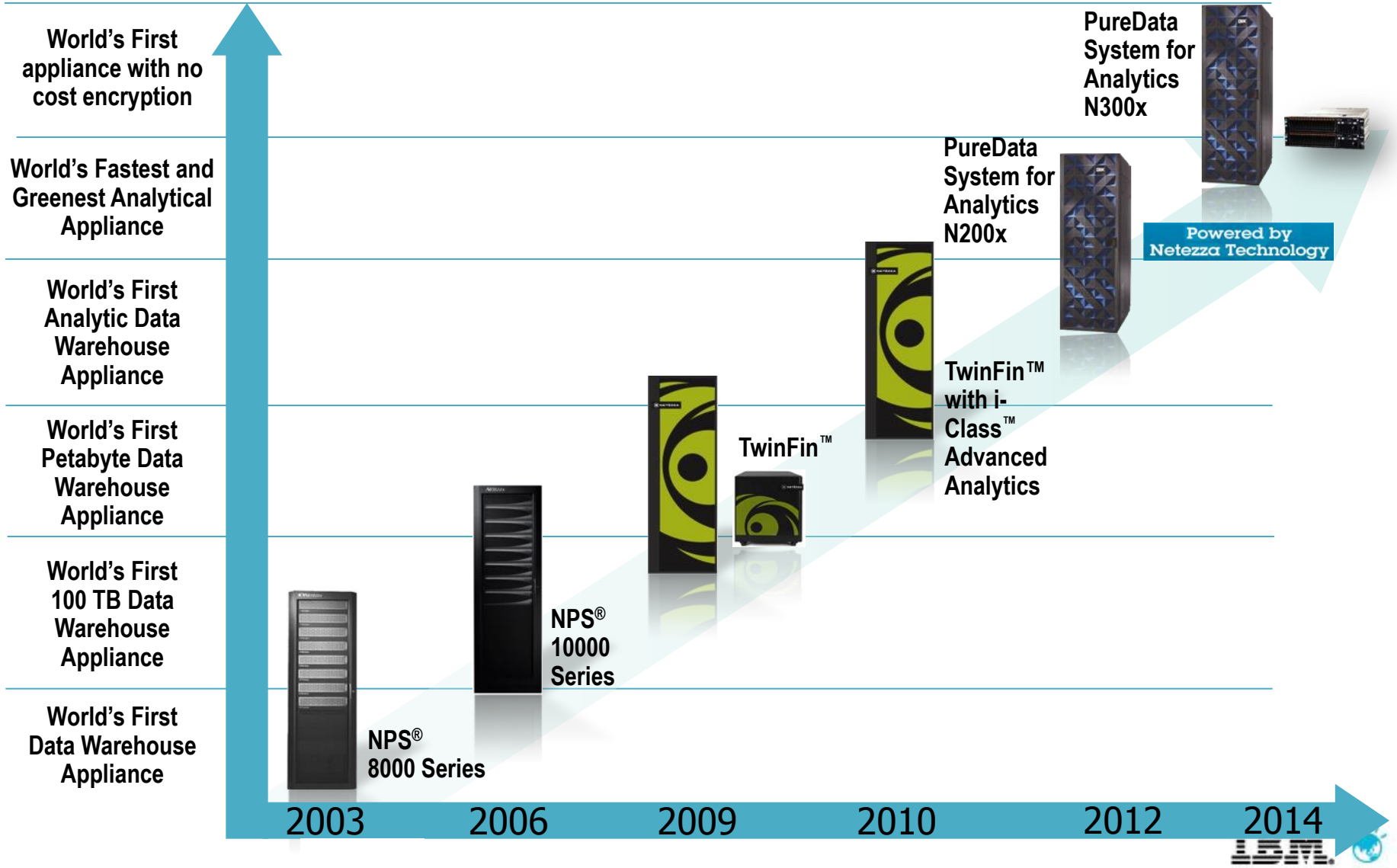
Производительность от 10 до 100 раз быстрее транзакционных СУБД

Простота - минимальное администрирование

Масштабируемость - от 16 ТБ до петабайтов данных

Высокопроизводительная **аналитика в самой СУБД** (SAS, SPSS, R, …)

IBM


World's First appliance with no cost encryption
World's Fastest and Greenest Analytical Appliance
World's First Analytic Data Warehouse Appliance
World's First Petabyte Data Warehouse Appliance
World's First 100 TB Data Warehouse Appliance
World's First Data Warehouse Appliance
NPS® 8000 Series
NPS® 10000 Series
TwinFin™
TwinFin™ with i-Class™ Advanced Analytics
PureData System for Analytics N200x
PureData System for Analytics N300x
Powered by Netezza Technology
2003
2006
2009
2010
2012
2014

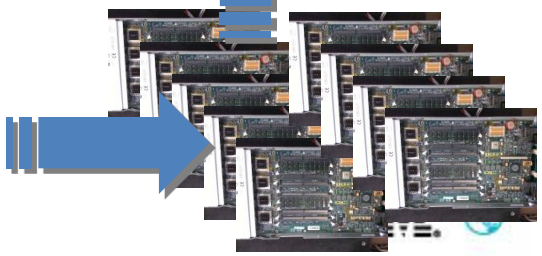
Что делает интеллектуальную аналитику на IBM PureData такой быстрой? 2 особенности.

```
select c_name, sum(o_totalpri  pric  from customer, orders
where o_orderkey in (select  rderke rom lineitem2 where
o_orderkey=l_orderkey and l_ hipdate>= 1-01-1995' and
l_shipdate<='01
c_name;" test_t
```

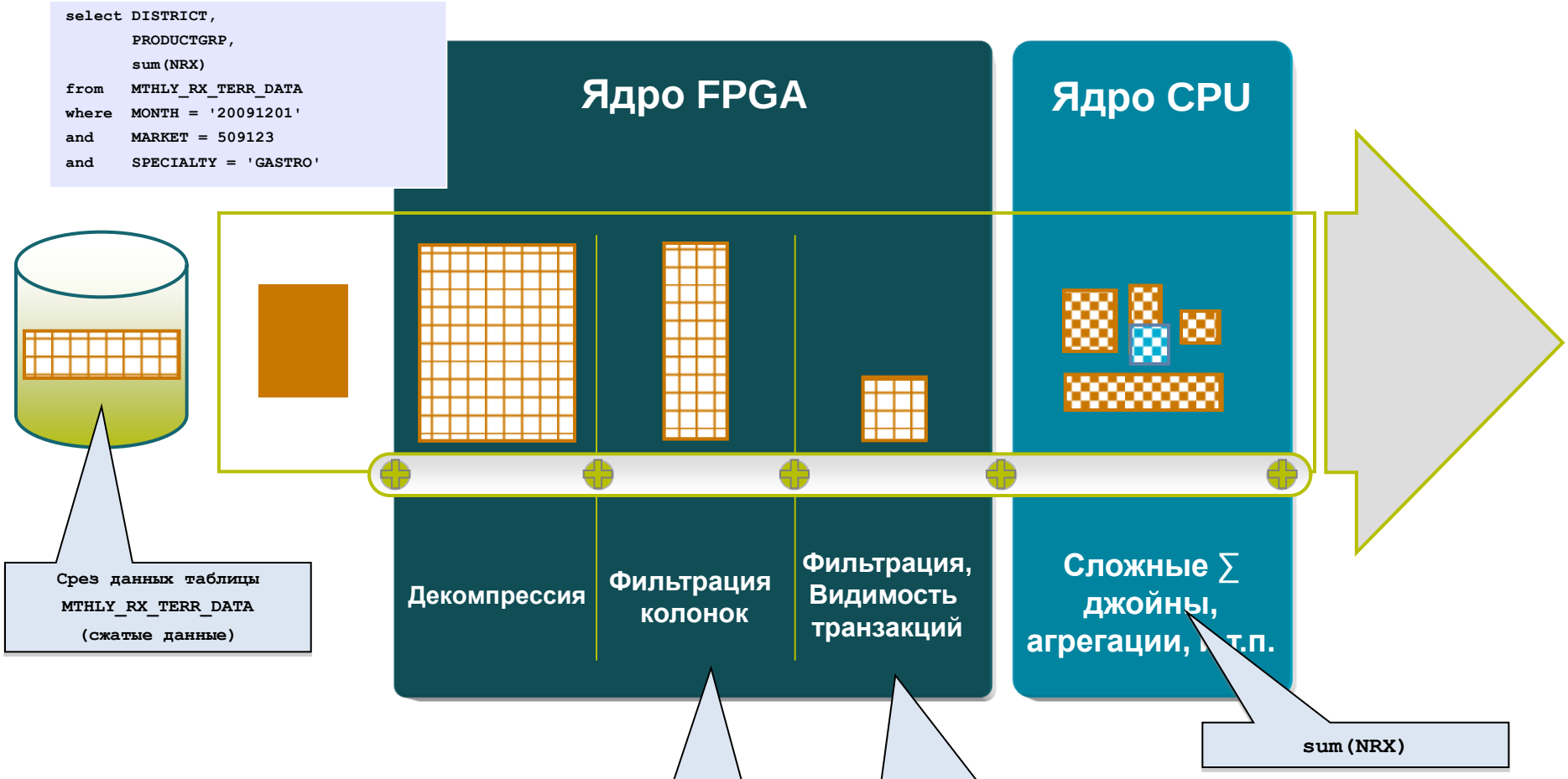
```
/********* Code *****  */

void GenPlan1(CPlan *plan, char *bufStarts,char *bufEnds,
bool lastCall) {
    //
    // Setup for next loop (nodes 00..07)
    //
    // node 00 (TScanNode)
    TScanNode *node0 = (TScanNode*)plan->m_nodeArray[0];
    // For ScanNode:
                TScan0 *Scan0 = BADPTR(TScan0*);
                CTable *tScan0 = plan->m_nodeArray[0]-
>m_result;
    char *nullsScan0P = BADPTR(char *);
    // node 01 (TRestrictNode)
    TRestrictNode *node1 = (TRestrictNode*)plan-
>m_nodeArray[1];
    // node 02 (TProject
    TProjectNode *node  (TProj tNode*)plan-
>m_nodeArray[2];
    // node 03 (TSaveTempNode)
    T
>m_no
    /
    T
    C
    CR
    // node 04 (THashNode)
…
```

```
 c_name        |   price
--------------------+-----------
 Customer#000000796 | 318356.97
 Customer#000001052 | 293680.56
 Customer#000001949 | 215280.98
 Customer#000002093 | 282531.93
 Customer#000005656 | 335297.31
 Customer#000005861 | 233691.03
 Customer#000006002 | 267000.92
 Customer#000006343 | 595819.82
 Customer#000006532 | 442254.91
….
real    0m0.552s
user    0m0.010s
sys     0m0.000s
```

```
1011010101010101010111111010101001001010101110101010010111101010001011110
1010100101011111011010010101010111010101100101010101011111010010101001010
1001010101010101010101010010101001111110101010101010100101010010101001010
1010100101001011010011111111010101010100110100101010101001010101001010101…
```



1. Архитектура комплекса PureData for Analytics



2. Встроенные в PureData алгоритмы in-database

Применение

- Управление воронкой продаж, пакетирование продуктов, кросс-продажи, управление маркетинговыми кампаниями
- Оценка ценности клиента, сегментация, удержание заказчиков
- Управление рисками, оптимизация доходов

Особенности

- **Встроенные in-database функции**
  - Дата майнинг, предсказательная аналитика, статистический анализ и гео аналитика
- **Интеграция с средствами BI и визуализации**
  - IBM Cognos, Microstrategy, Business Objects, SAS, Excel, SSRS, Kognitio, Qlikview и Tableau
- **Интеграция с инструментами статистического моделирования и скоринга моделей**
  - **IBM SPSS, SAS, Open Source R, Fuzzy Logix**
- **Возможность создания собственных in-database расширений аналитики**
  - R, Java, C, C++, Python, LUA и Perl

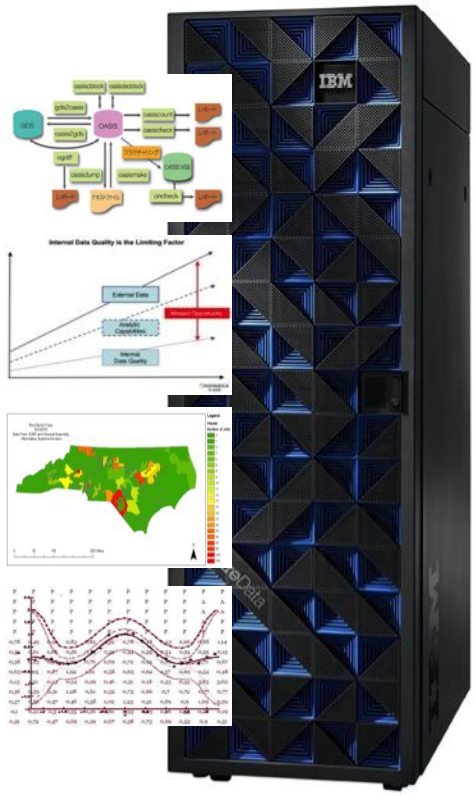
Выполнение аналитики в хранилище без перемещения данных к серверам приложений

Подготовка данных

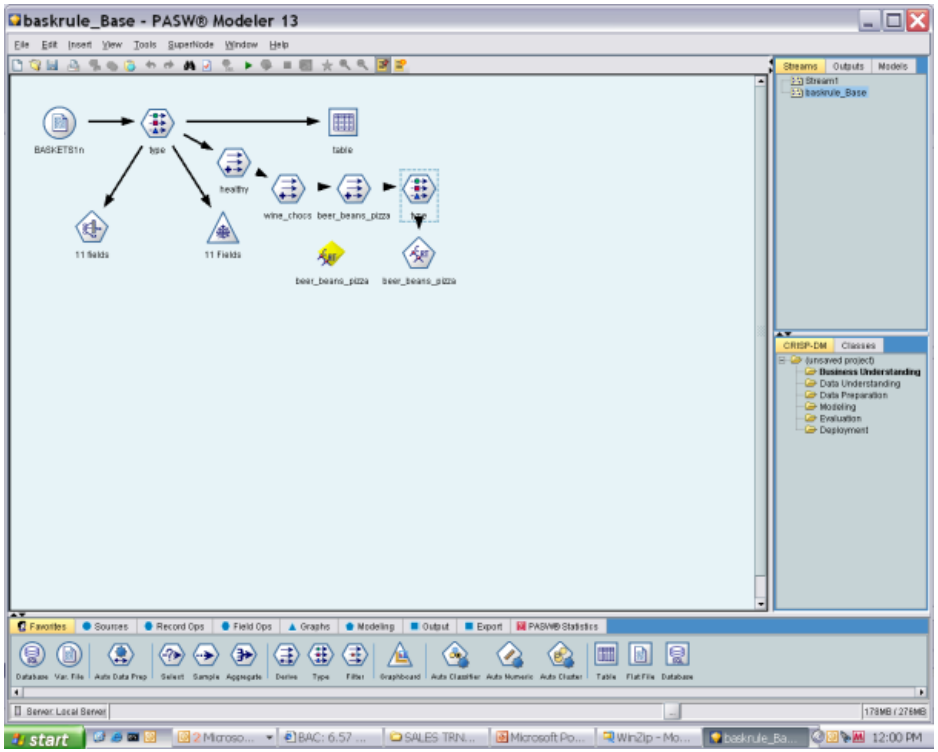
Прогнозный анализ

Гео анализ

Расширенная статистика



Интеграция IBM SPSS с комплексом PureData for Analytics



Modeler Client

Modeler Server

SPSS

Полная поддержка SPSS Modeler, включая SQL Pushback и Data Mining в PureData

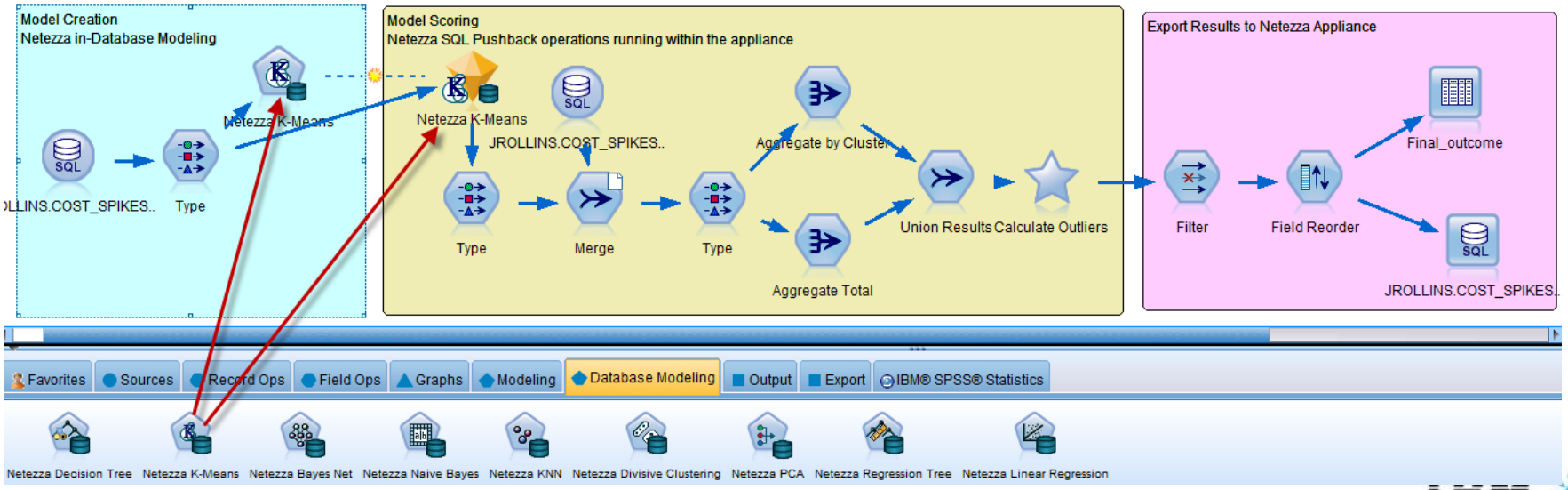
IBM

Выполнение функций SPSS совместно с PureData

```
-- Run KMEANS
call nza..KMEANS('intable=cost_spikes_maxlimit_nrm_cat, id=physician_id, k=' || cluster_max || ', maxiter=10, distance=euclidean, randseed=3, model=cost_spikes_maxlimit_nrm_cat_initial_centers,
outtable=cost_spikes_maxlimit_nrm_cat_initial_clusters') ;

-- Add initial cluster_id to Table for Reporting... Use resulting table for Reporting
call DROP_IF_EXISTS('cost_spikes_maxlimit_initial_clusters');
create table cost_spikes_maxlimit_initial_clusters as select
B.PHYSICIAN_ID,
A.CLUSTER_ID,
B.MARKET,
B.SPECIALTY,
B.PERIOD2_TOTAL_MEMBERS as TOTAL_MEMBERS,
B.PCT_CHG_TOTAL_MEMBERS,
B.PCT_CHG_TOTAL_VISITS,
B.PCT_CHG_TOTAL_CLAIMS,
B.PCT_CHG_TOTAL_UNITS,
B.PCT_CHG_TOTAL_BILLED,
.....
```

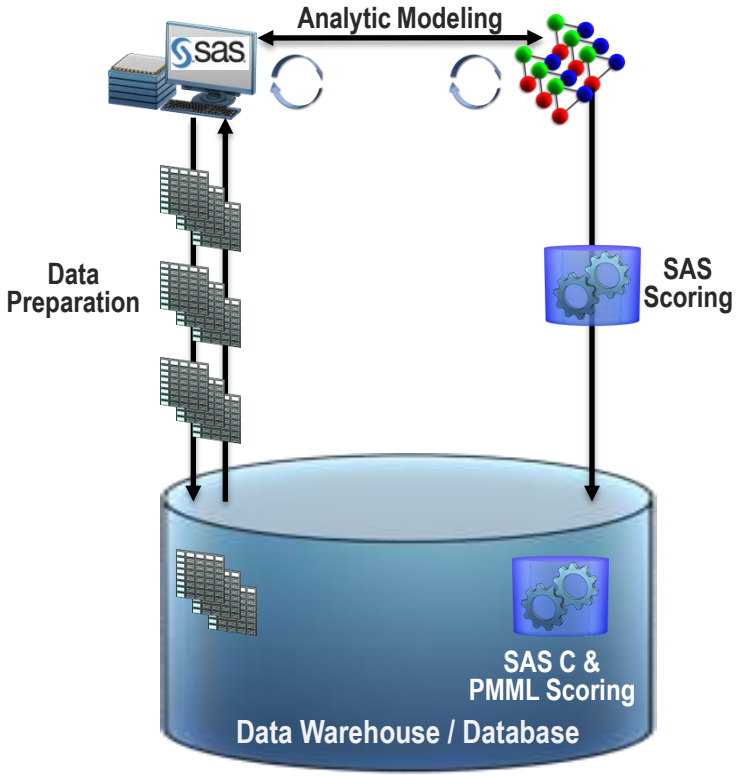
SQL-код с вызовами процедур SPSS

SPSS Modeler с механизмами in-Database Mining

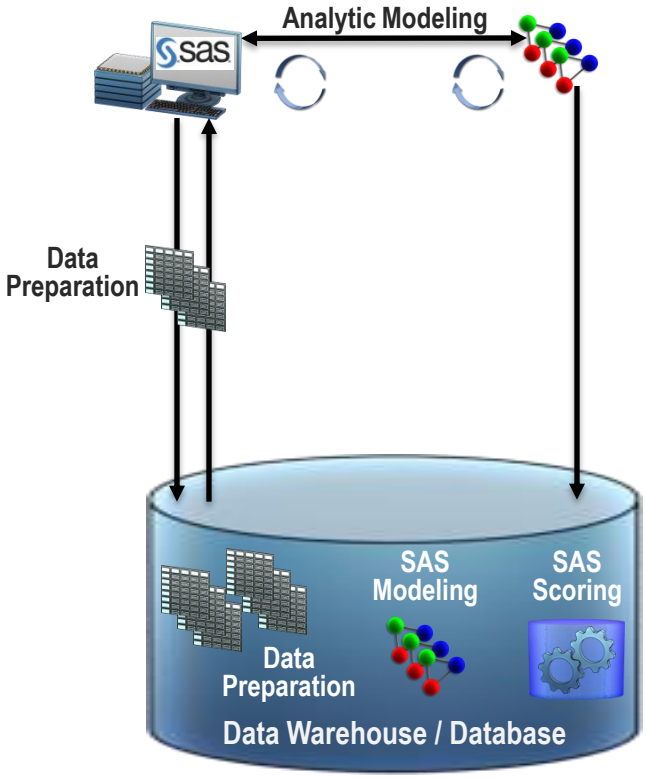


Интеграция SAS с комплексом PureData for Analytics

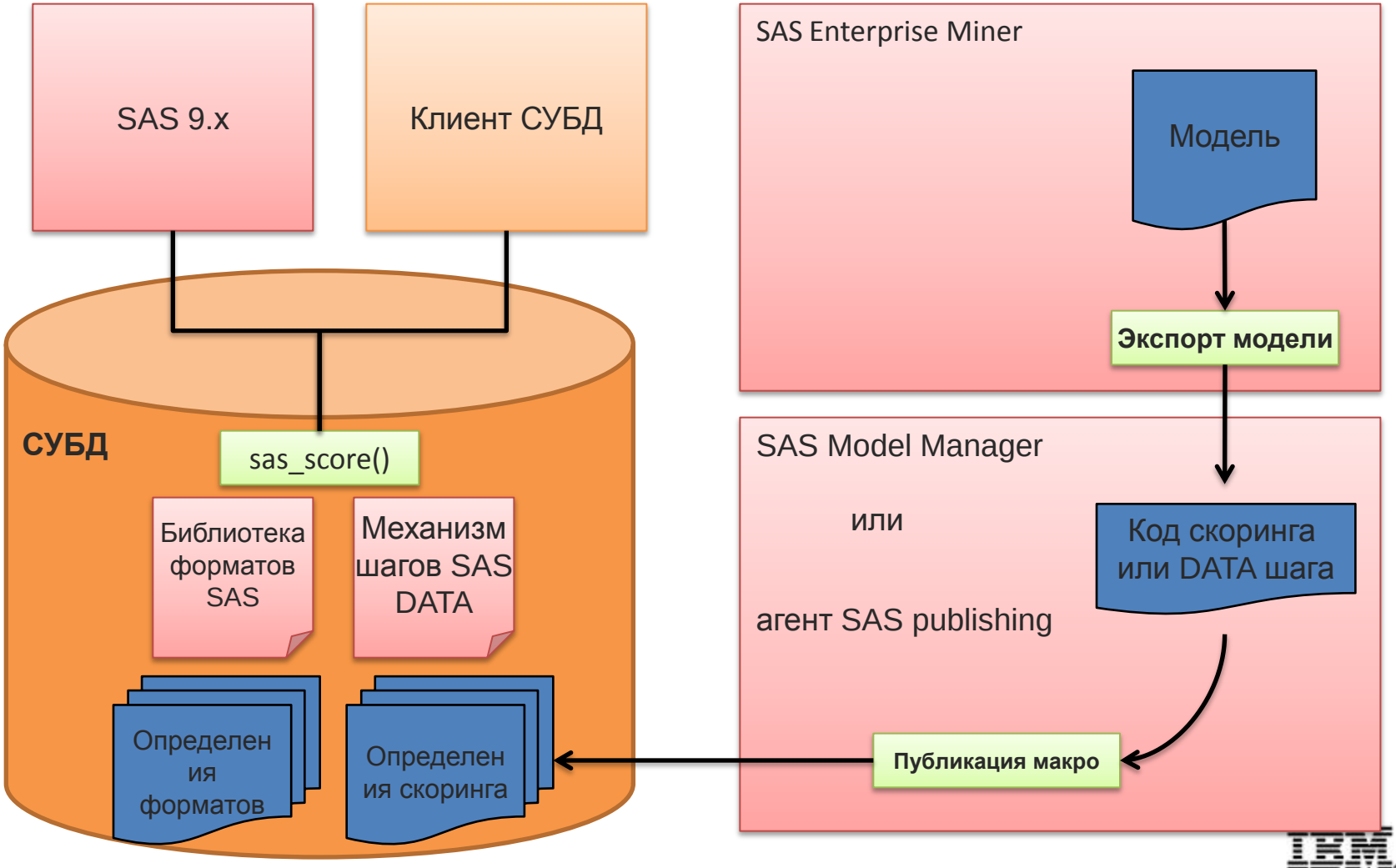
Традиционная архитектура



Архитектура In-Database



Выполнение расчетов SAS совместно с PureData



Результаты интеграции PureData for Analytics с SAS

- Более 30 совместных заказчиков используют PureData for Analytics и SAS в промышленной среде.
- Использование PureData всегда показывает более чем значительный выигрыш в проектах SAS, нежели чем ранее использованные СУБД.
- **Marriott** тестировал 2 сложных SAS-процесса с PureData и конкурирующей OLAP СУБД:
  - Первый тест занял 7 секунд в PureData и 761 секунд другой СУБД → 108 раз быстрее
  - Второй тест занял 16 секунд в PureData и 2591 секунд в другой СУБД → 161 раз быстрее
- **Epsilon** тестировал SAS-процесс с PureData (13 минут), ранее этот процесс работал на конкурирующей СУБД более 2 часов → в 9 раз быстрее
- **Premier** тестировал SAS-процесс с PureData (3 минуты) в сравнении с процессом на предыдущей системе – 75 минут, или в 25 раз быстрее.

Marriott epsilon PREMIER

IBM

Catalina Marketing – пример ускорения интеллектуальной аналитики

- Подготовка маркетинговых кампаний для US Retail, анализ 80% всех транзакций магазинов в США;
- Программы лояльности для 200 миллионов покупателей;
- Столкнулись с невозможностью повысить производительность существующей аналитической платформы SAS;
- PureData for Analytics развёрнута более чем на 100 шкафах.
- Результаты: в 10 раз больше предсказательных моделей (прямая корреляция между количеством моделей и прибылью), 2.5 петабайта обрабатывается в PureData (**500 миллиардов товарных записей**). В 70 раз больше запросов на 5-ти кратном объёме данных;
- Среднее время выполнения скоринга модели уменьшилось с **4.5 часов** до **60 секунд**.

CATALINA MARKETING

IBM

Преимущества выполнения интеллектуальной аналитики в OLAP комплексах

- Достигается наивысшая производительность выполнения скоринга аналитических моделей и быстрые результаты.
- Увеличивается точность и эффективность аналитических моделей
- Уменьшатся затраты на перемещение данных при расчетах и связанные с этим затраты.
- Нет необходимости изменять модель и код обработки моделей при переходе к обработке «больших данных», сокращаются расходы на поддержку проектов.