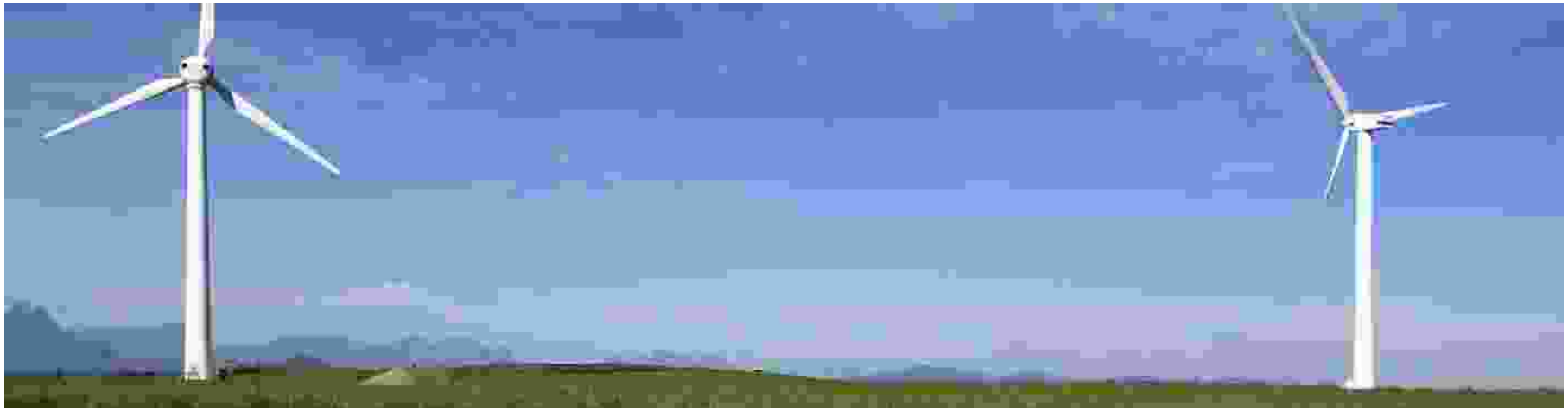
Платформа IBM Big Data

Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData



Содержание

- Большие данные – природа явления
- Продукты IBM и варианты использования
- Обзор IBM BigInsights
- Обзор IBM InfoSphere Streams
- Обзор IBM Netezza

Почему Big Data?

12+ Тб
твитов каждый день

? TBs данных
каждый день

25+ Тб
логов каждый день

30 млрд. RFID тэгов
(1.3B in 2005)

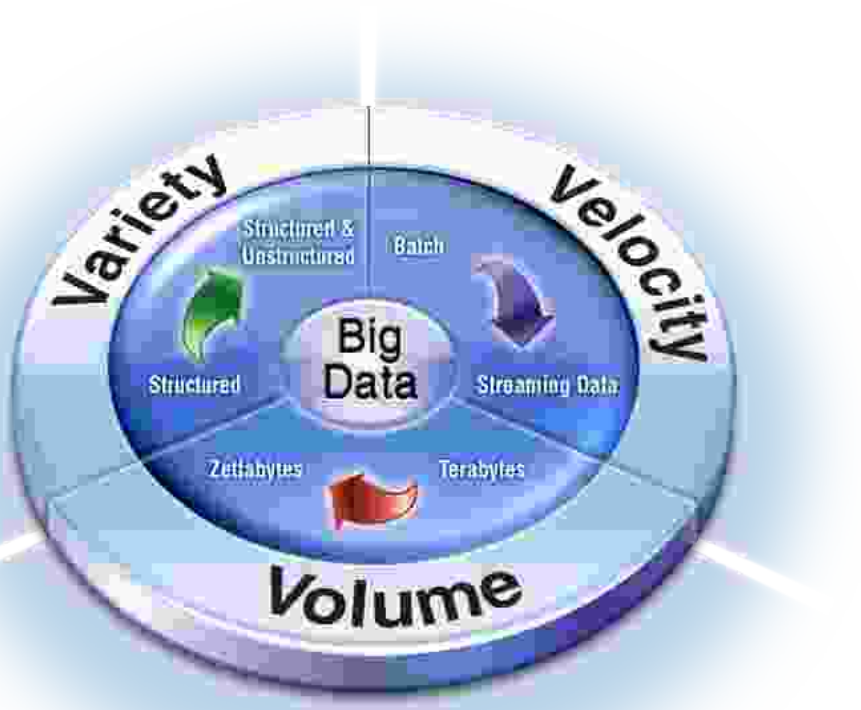
4.6 billion camera phones world wide

100ни мил. GPS устройств продается в год

2+ млрд. пользователей Интернет в 2011

76 мил. Электр. датчиков в 2009… 200M в 2014 !

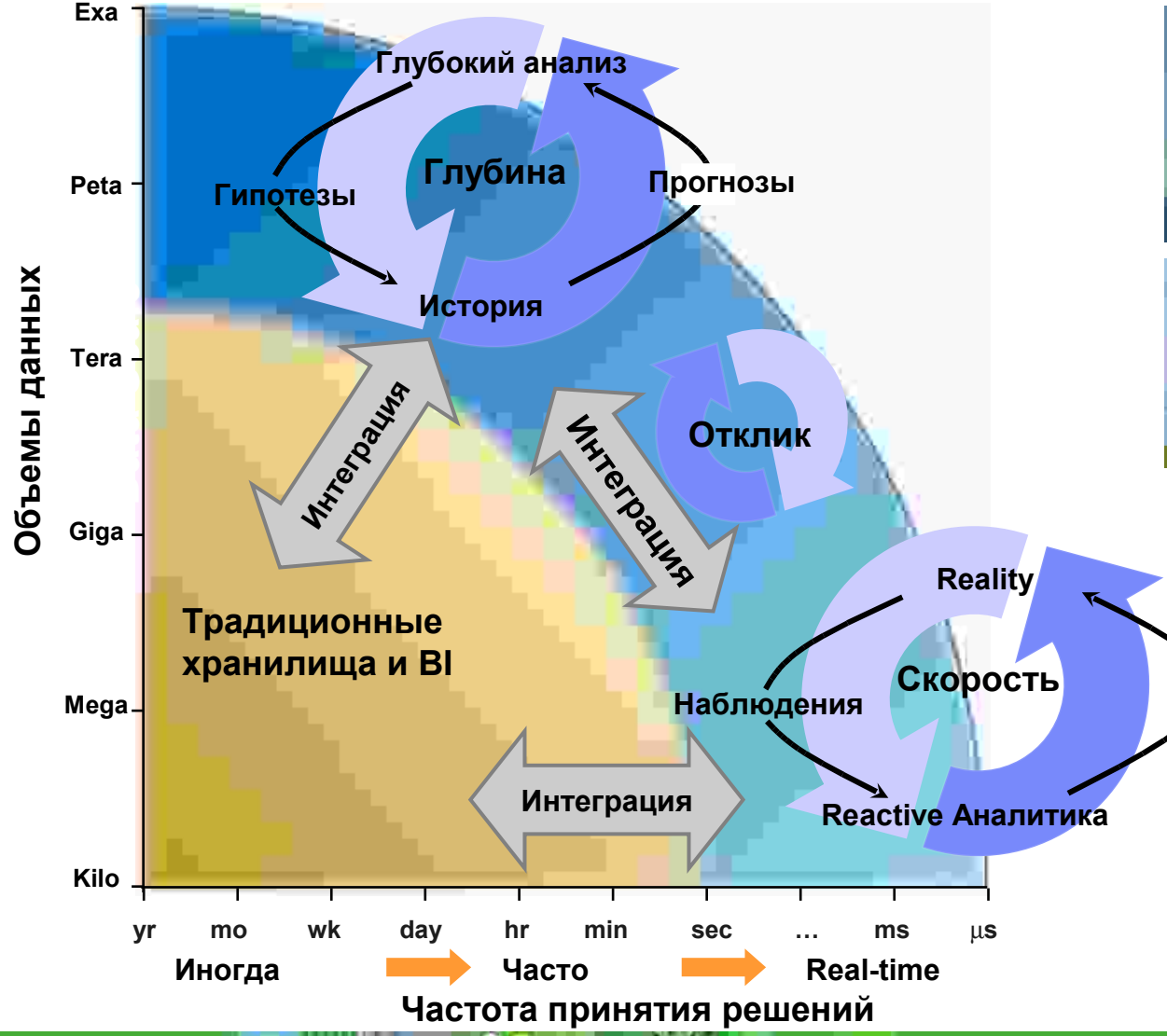
Что такое Big Data?

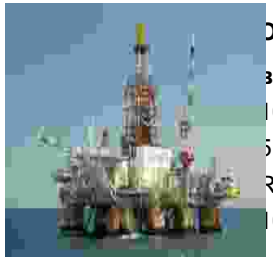


Анализ огромных объемов данных, имеющих разнообразную структуру, и поступающих с большой скоростью.

Раньше это считалось невозможным!

Максимальный эффект при использовании разных видов анализа





Оптимизация в реальном времени
100,000 обновлений в сек,
5 ms на решение
Round-trip automation
10PB for Deep Analytics

Прогнозирование
100,000 records/sec, 6B/day
10 ms/decision
6PB for Deep Analytics

DeepQA
100s GB for Deep Analytics
3 sec/decision
1 PB training corpus

Smart Traffic
250K GPS probes/sec
630K segments/sec
2 ms/decision, 4K vehicles

Платформа IBM Big Data

InfoSphere BigInsights
Hadoop-based low latency analytics for variety and volume

Hadoop

InfoSphere Information Server
High volume data integration and transformation

Information Integration

Stream Computing

InfoSphere Streams
Low Latency Analytics for streaming data

MPP Data Warehouse

IBM InfoSphere Warehouse
Large volume structured data analytics

IBM Netezza High Capacity Appliance
Queryable Archive Structured Data

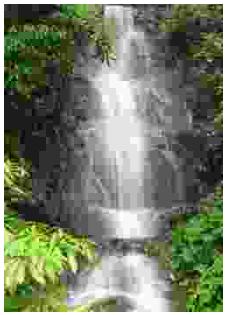
IBM Netezza 1000
BI+Ad Hoc Analytics on Structured Data

IBM Smart Analytics System
Operational Analytics on Structured Data

IBM Informix Timeseries
Time-structured analytics

Отчет Forrester Big Data Wave

Потоки и океаны информации . .



Information Streams

Высокоскоростные потоки информации

- Информация с сенсоров, инструментов и т.п
- Информация из логов и ПО для мониторинга в реальном времени
- Потоковый контент, аудио и видео
- Большой объем транзакций, например тикеры, трейды, информация о трафике.

Information BigInsights

Информация для хранения вне традиционных СУБД

- Результаты обработки потоковых данных
- Информация из социальных сетей, эл. Почты, clickstream логи, и т.п.
- Неструктурированные документы – формы, заявления, отчеты, отсканированные изображения
- Структурированная информация из смешанных источников

Netezza – A Smarter Appliance for Smarter Customers

- Big Data Parallel Computing Platform
- Big Data Warehouse for Complex Data Types
- Big Data Hadoop & NO SQL
- Big Data Business Analytics & R
- Big Data Business Intelligence
- Big Data ETL & ELT
- Big Data Geospatial

Варианты использования

Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData

Глобальная медиа-компания

Бизнес-задача

- Идентифицировать пиратский контент
- Оценить потерю выручки и проанализировать тенденции
- Мониторить социальные сети (например, Twitter, Facebook), чтобы отследить распространение пиратского контента. Время отклика критично!

Цели проекта:

- Анализировать большие объемы данных. Точный размер определить невозможно.
- Начать с данных из соц. Сетей за 1 год. При помощи анализа текстов:
  - Извлечь и квалифицировать необходимую информацию (при помощи сложного набора правил)
  - Поиск URL через которые распространяется искомая информация, ….
- Обеспечить возможность для анализа видео в будущем

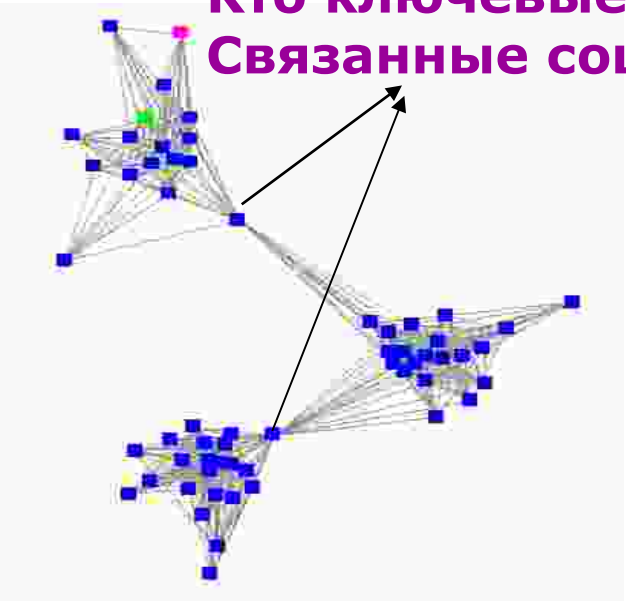
Компоненты решения:

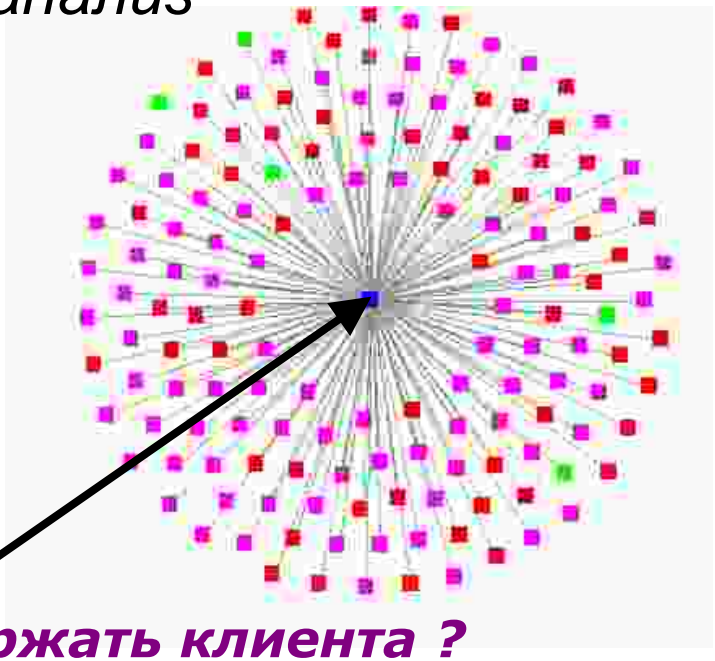
- IBM InfoSphere BigInsights Enterprise Edition:
  - Инструменты для анализа текстов
  - Custom text annotators
  - Flexible query support
  - Масштабируемость

Анализ социальных сетей непосредственно по CDR

Визуализация упрощает анализ

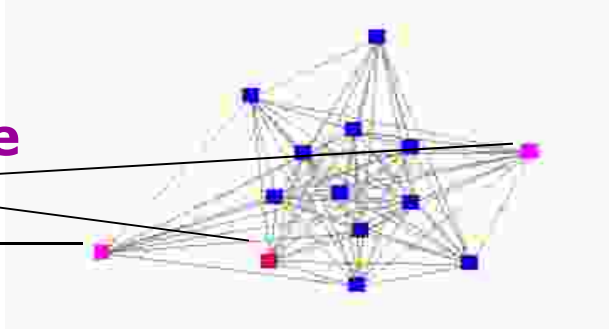
Кто ключевые клиенты?
Связанные соц. группы





Как удержать клиента ?
(и переманить его друзей?)

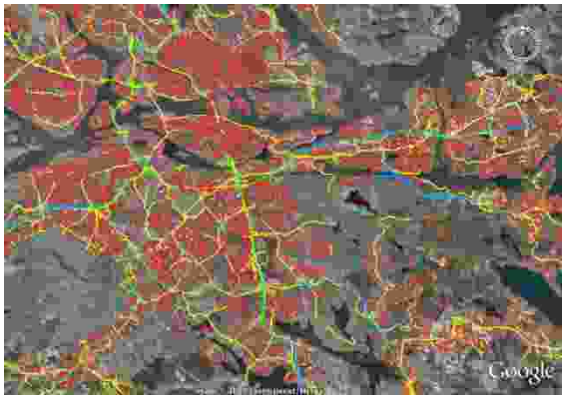
Потенциальные мишени для конверсии?

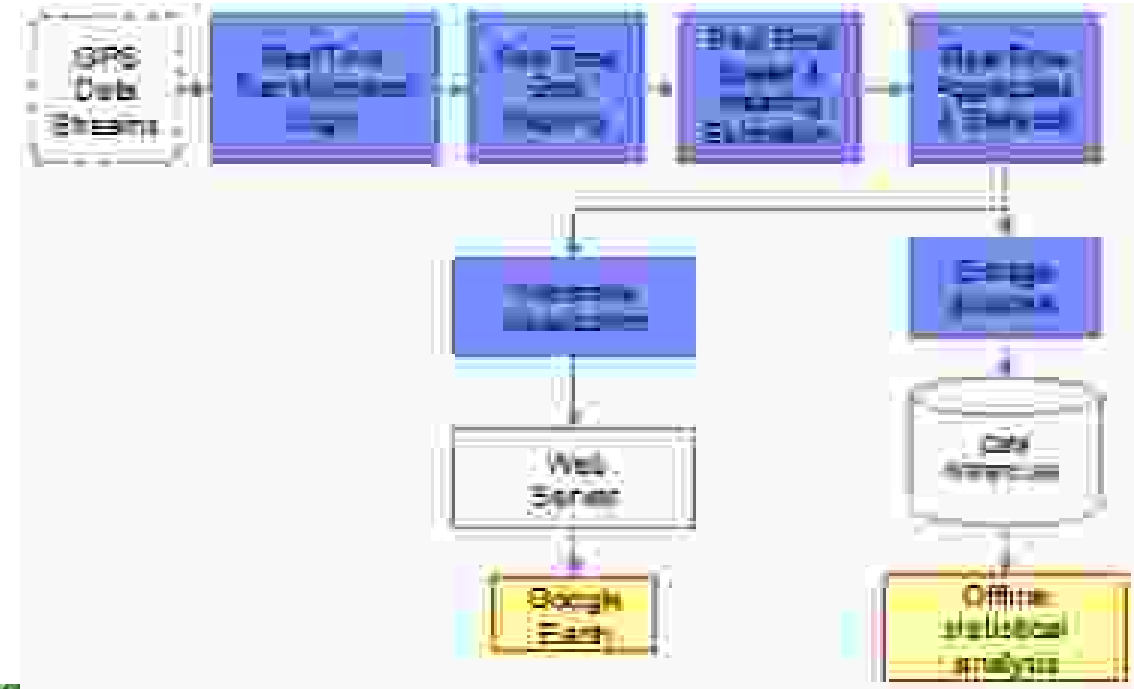


- Platinum Customer
- Regular Customer
- CDMA Competitor
- GSM Competitor

Управление дорожным трафиком

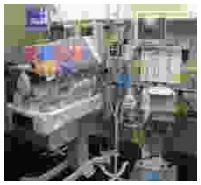
- Разрозненные потоки данных
  - GPS
  - Сотовые телефоны (отслеживание местоположения)
  - Общественный транспорт
  - Погодные и дорожные условия
  - Инциденты и дорожные происшествия
  - Время проезда мимо камер на основании распознавания номеров
  - Оптические детекторы потока машин
  - Измерение загрязнения воздуха
  - Дорожные работы
  - Неподвижные снимки с дорожных камер
- Мониторинг дорожного трафика в реальном времени
- Планирование поездок

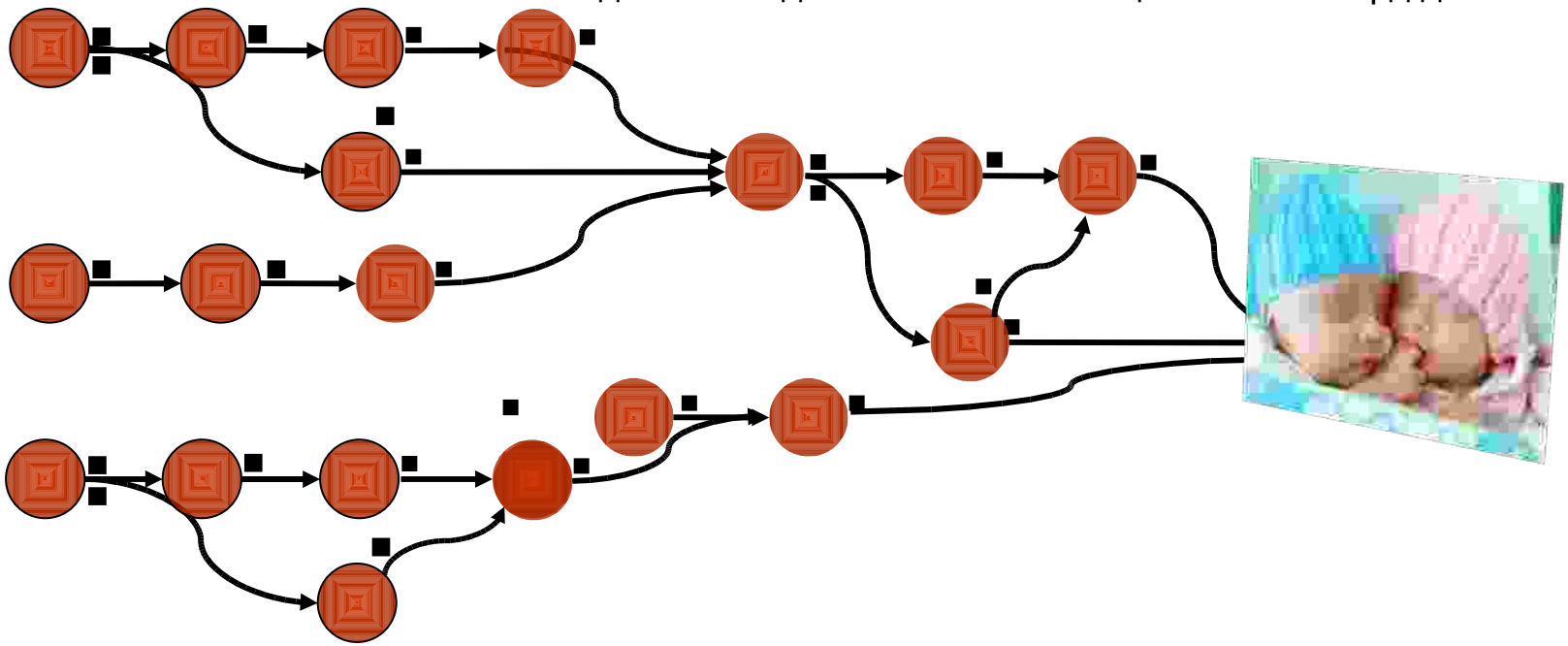




Данные в движении

- **Информация, которую нужно проанализировать каждые 30-60 минут, и которую не имеет смысла хранить более 72 часов**
- **Анализ 1000 показателей медицинской диагностической информации в секунду, динамическая модель данных**
- **Перспективы: 20% снижение смертности в контрольной группе**
  - Наблюдение 120 детей: 120тыс сообщений/сек…млрд/день





TerraEchos – Скрытое наблюдение

- Сверхсовременная система наблюдения, основана на платформе Streams
- Мониторинг акустических сигналов в реальном времени, принятие решений об уровне угрозы и адекватной реакции
- Текущий дизайн позволяет анализировать одновременно 1600 потоков сырых данных

Обзор продукта IBM BigInsights

Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData

Обзор IBM BigInsights

- IBM BigInsights – это фреймворк для параллельной обработки больших объемов неструктурированных данных
- Основан на проекте с открытым исходным кодом Apache Hadoop
- Является платформой для разработки и выполнения MapReduce заданий
- Но BigInsights – это гораздо больше чем просто Hadoop и MapReduce

Что такое MapReduce

- Подход к распределенной параллельной обработке данных
- Популяризован компанией Google в 2004 году
  - Появление публикации «MapReduce: Simplified Data Processing on Large Clusters»
- Однако использовался гораздо раннее в функциональных языках программирования
- Любое задание состоит из двух шагов – Map и Reduce, которые выполняются одновременно (Reduce принимает на вход обработанные данные из Map)
- Оба шага на вход и на выходе используют массивы (ключ, значение)
- Шаг Map принимает на вход массив (ключ, значение) и применяет к каждому элементу некую функцию.
  - **Пример: Map(sqrt, [1,4,9,16]) даст [1,2,3,4]**
- Шаг Reduce агрегирует массив при помощи некой функции.
  - **Например reduce(sum,[1,2,3,4]) даст 10**

Классический пример MapReduce задания

- Задача : посчитать количество вхождений каждого слова по группе документов
- Шаг Map:

```
map(String key, String value):
// key: имя документа
// value: содержимое документа
for each word w in value: EmitIntermediate(w, "1");
```

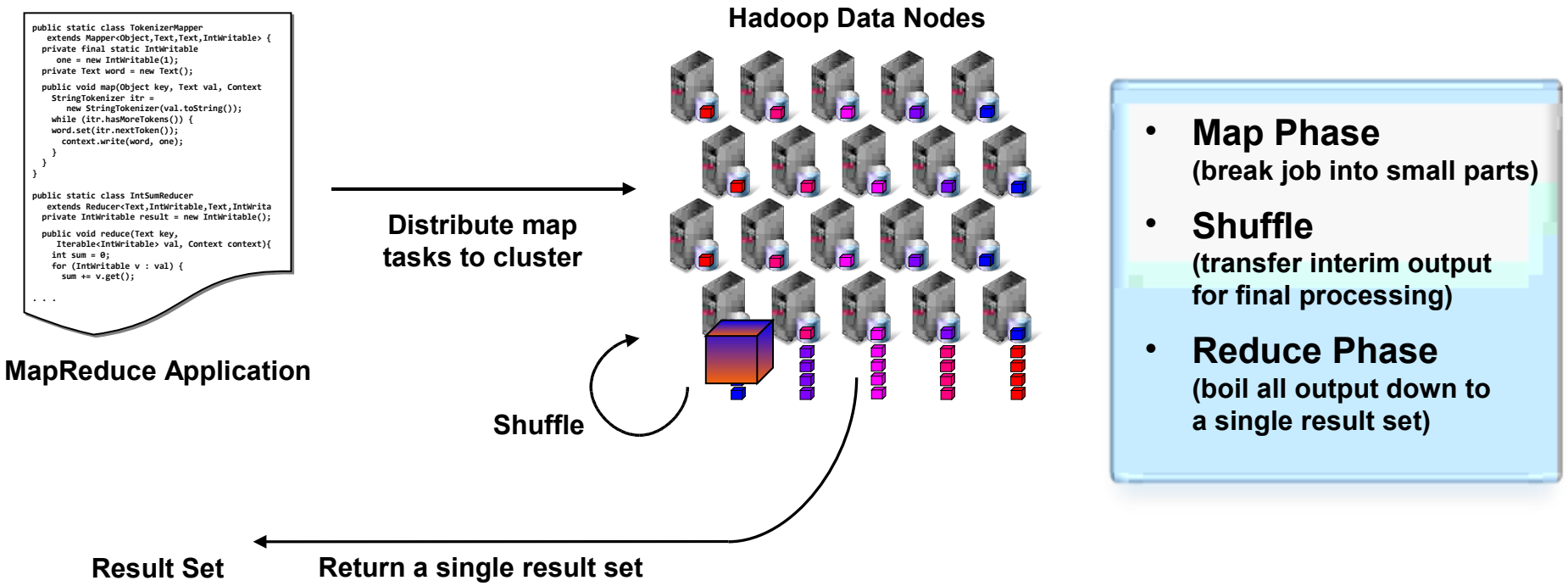
- Шаг Reduce:

```
reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values: result += ParseInt(v);
Emit(AsString(result));
```

- Вокруг этой простой модели программирования построена инфраструктура Hadoop!

Выполнение заданий MapReduce

- Вычислительная модель Hadoop
  - Данные хранятся в распределенной файловой системе на множестве недорогих машин
  - Вычисления близко к данным
  - Распределенное приложение – вычислительные ресурсы применяются там, где хранятся данные
- Масштабируемость до тысяч узлов и петабайт данных



Что такое Hadoop?

- Это имя плюшевого слоника сына основателя проекта Apache Hadoop Дага Каттинга (Doug Cutting), который на тот момент работал в Yahoo
- Инфраструктура для выполнения MapReduce заданий
- Основные компоненты это :
  - HDFS (Hadoop Distributed File System) – распределенная файловая система
  - Hadoop Mapreduce – фреймворк для выполнения MapReduce заданий
  - Hadoop Common - утилиты и библиотеки, необходимые для множества подпроектов Hadoop.
- Есть множество подпроектов для различных целей
- Высокоуровневые языки программирования и СУБД – Hive, Hbase, Pig, Cassandra
- Управление заданиями и мониторинг – Chukwa, Zookeper, Oozie
- Текстовый поиск и аналитика – Lucene
- Коннекторы и адапторы – Flume

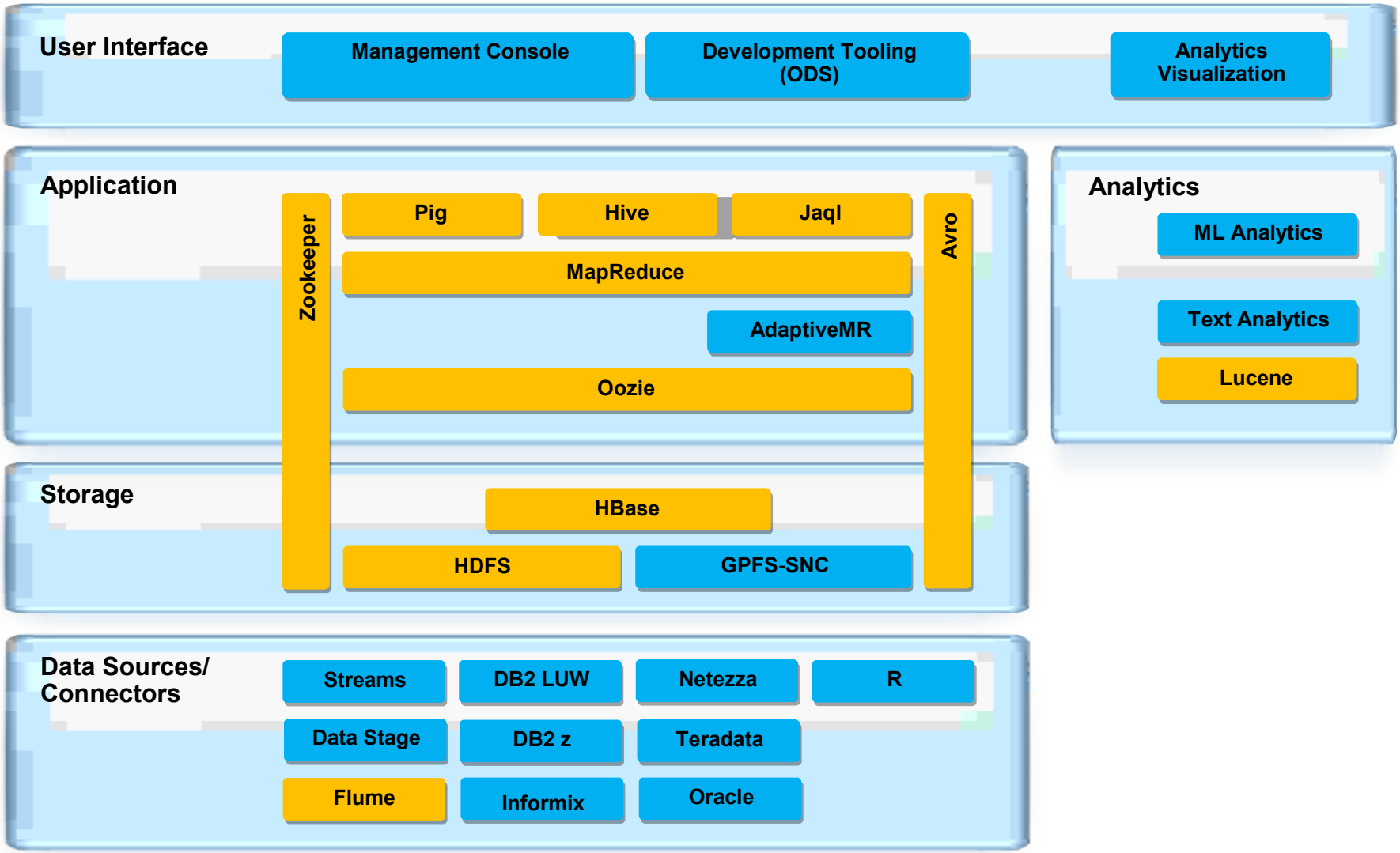
- IBM BigInsights включает в себя компоненты Hadoop и добавляет множество других!

СУБД и Hadoop – взаимнодополняющие друг друга компоненты

- Структурированные данные с известной моделью
- Записи, длинные поля, объекты, XML
- Допустимы обновления
- SQL & XQuery
- Быстрый отклик
- Потери данных недопустимы
- Безопасность и аудит
- Шифрование
- Сложные алгоритмы компрессии данных
- Дорогое промышленное аппаратное обеспечение
- 30+ лет развития технологий
- случайный доступ (индексирование)
- Большое сообщество администраторов и разработчиков, повсеместное использование

- Структурированные и неструктурированные данные
- Файлы
- Только вставка и удаление
- Hive, Pig, Jaql
- Пакетная обработка
- Потери данных могут случаться
- Пока нет
- Пока нет
- Простое пофайловое сжатие
- Недорогое аппаратное обеспечение
- 2-3 года развития технологии
- Только файловый доступ
- Используется небольшим количеством компаний

InfoSphere BigInsights – A Full Hadoop Stack



Компоненты BigInsights, которые не входят в Hadoop

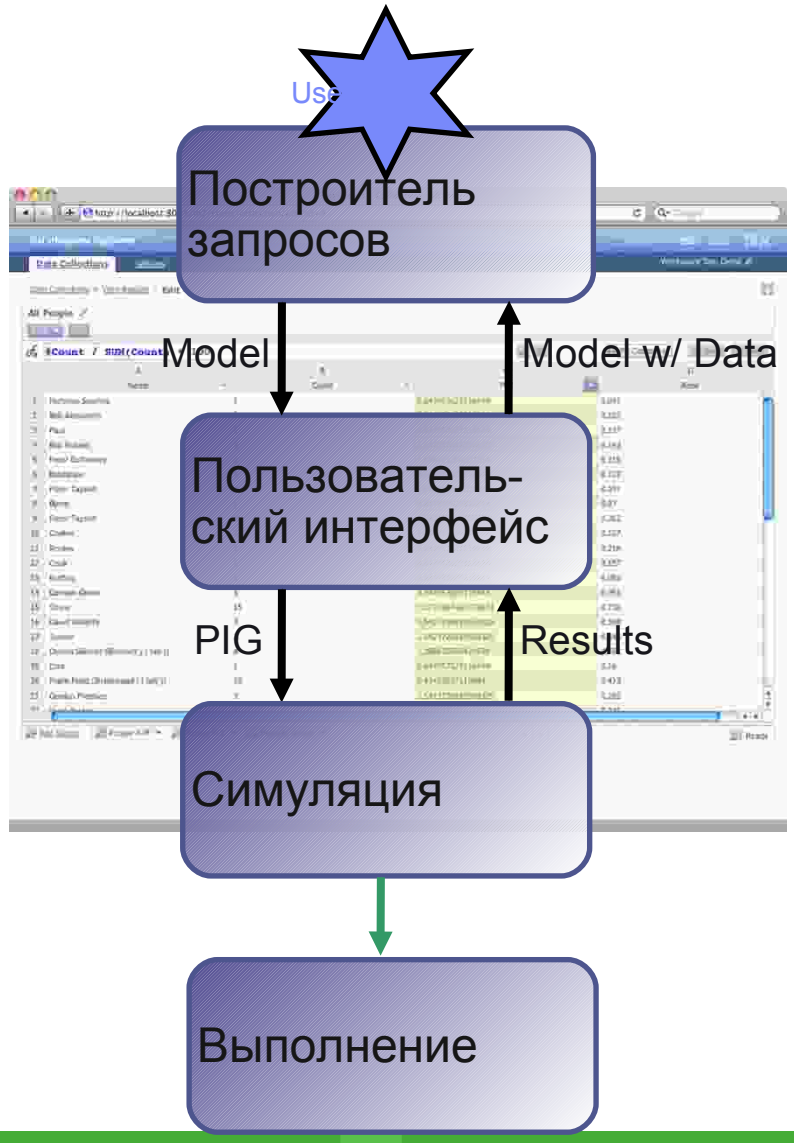
- AdaptiveMR - существенно улучшенный планировщик по сравнению с Hadoop MR
- Management Console – административная GUI-консоль
- Analitics Visualisation (Sheets) – инструмент аналитика
- Machine Learning libraries – библиотеки для маш. обучения
- Text Analytics – инструментарий для анализа текстов и язык текстовых запросов AQL
- Development Tooling – Основанные на Eclipse инструменты разработчика
- **GPFS-SNC - Более безопасная, быстрая и удобная файловая система**
- Коннекторы
  - Streams
  - DB2 LUW, z
  - Netezza
  - Teradata, Oracle, Informix
  - DataStage
  - R

Сравнение GPFS и HDFS

Файловая система	GPFS	HDFS
Надежность	Нет уязвимых компонент	Уязвимость NameNode
Целостность данных	Высокая	Возможность потерь
Масштабируемость	Тысячи узлов	Тысячи узлов
POSIX-совместимость	Полная	Ограниченая
Управление данными	Security, Backup, Replication	Limited
Производительность MapReduce	Хорошая	Хорошая
Изоляция нагрузки	На уровне диска	Нет поддержки
Производительность традиционных приложений	Хорошая	Плохая произв-ть случайных чтений и записи

Sheets Runtime Processing

- Пользователь создает коллекции данных, фильтрует их и трансформирует
- Sheets проверяет и компилирует команды пользователя
- Sheets выполняет задание на ограниченном сэмпле данных
- Пользователь выполняет задание на полном объеме данных, строит графики и диаграммы

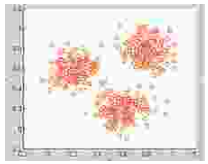


Обзор продукта IBM InfoSphere Streams

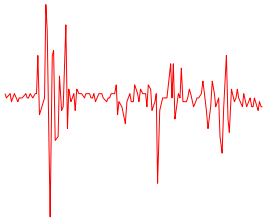
Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData

Streams анализирует любые данные



Mining in Microseconds
(включено в Streams)



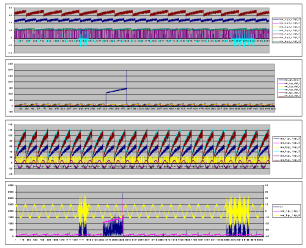
Акустика
(IBM Research)
(Open Source)

New

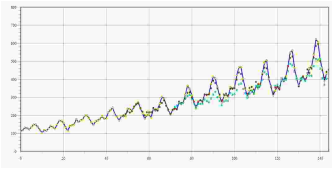
Текст
(listen, verb),

(radio, noun)

Simple & Advanced Text
(включено в Streams)



Advanced Mathematical Models
(IBM Research)



Predictive
(IBM Research)

$$\sum_{population} R(s_t, a_t)$$

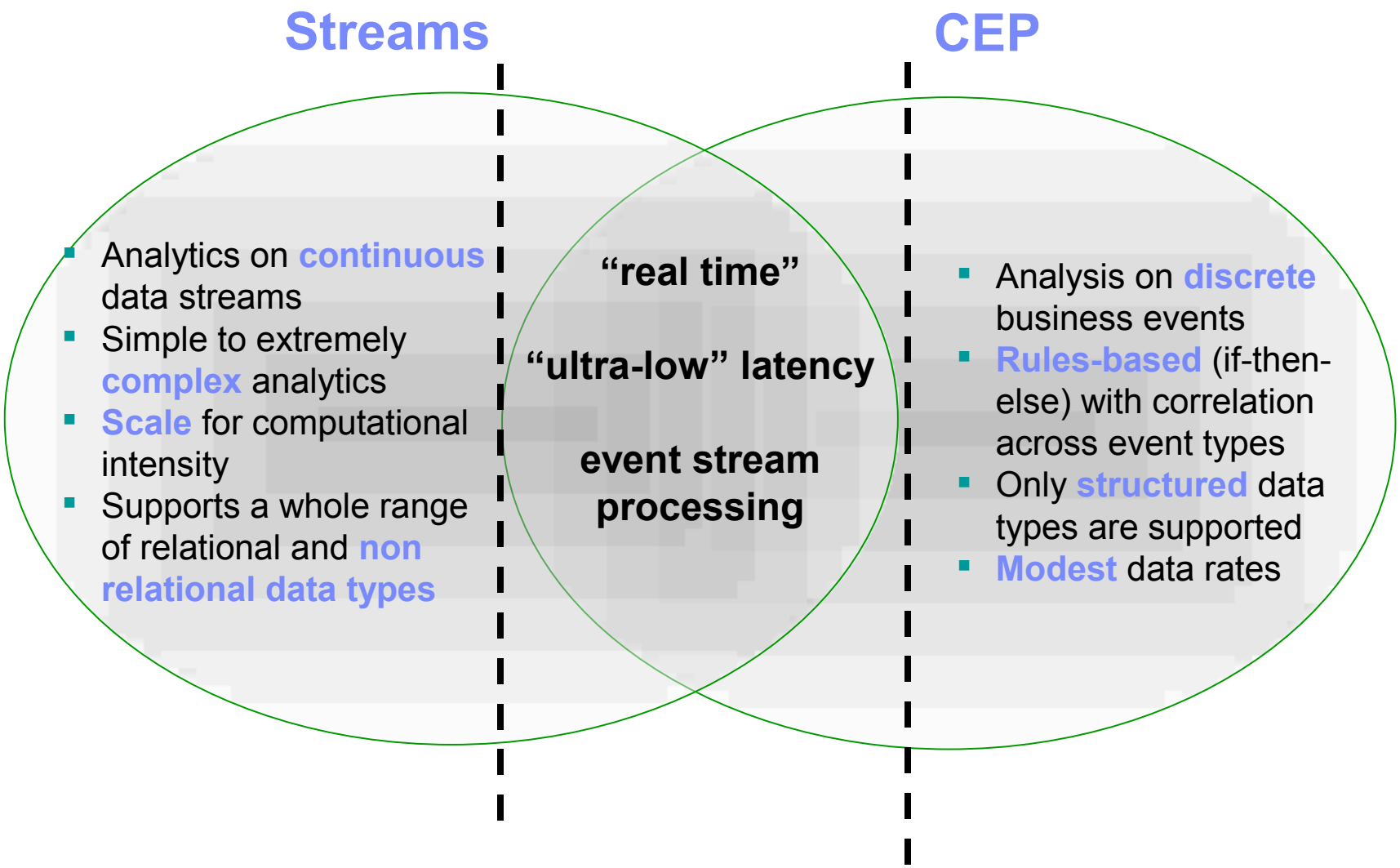
Статистика
(включено в Streams)

Geospatial
(IBM Research)

OpenCV

Фото & Видео *(Open S*

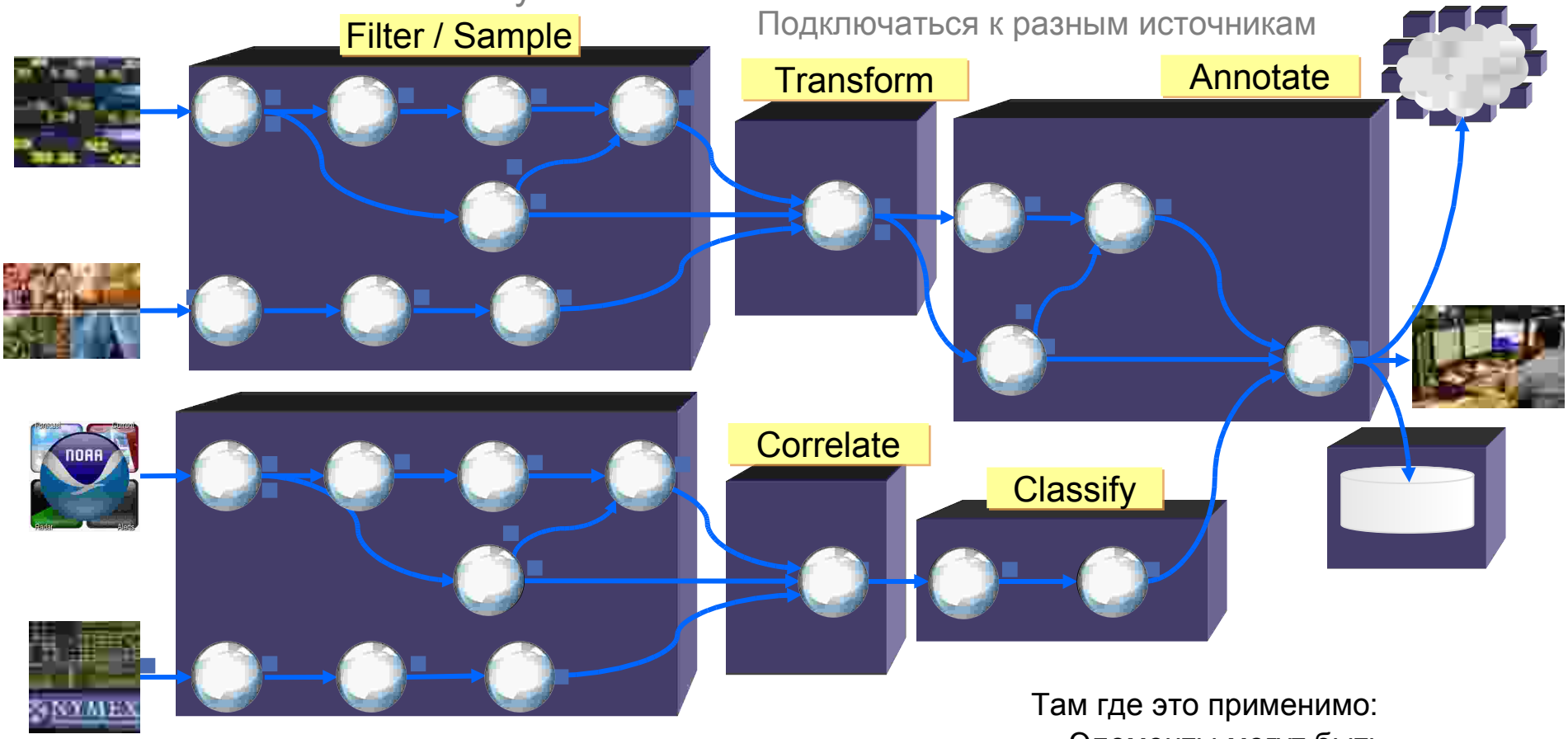
Комплексная обработка событий (CEP) и Streams



Как работает Streams

→ Continuous ingestion
→ Continuous analysis

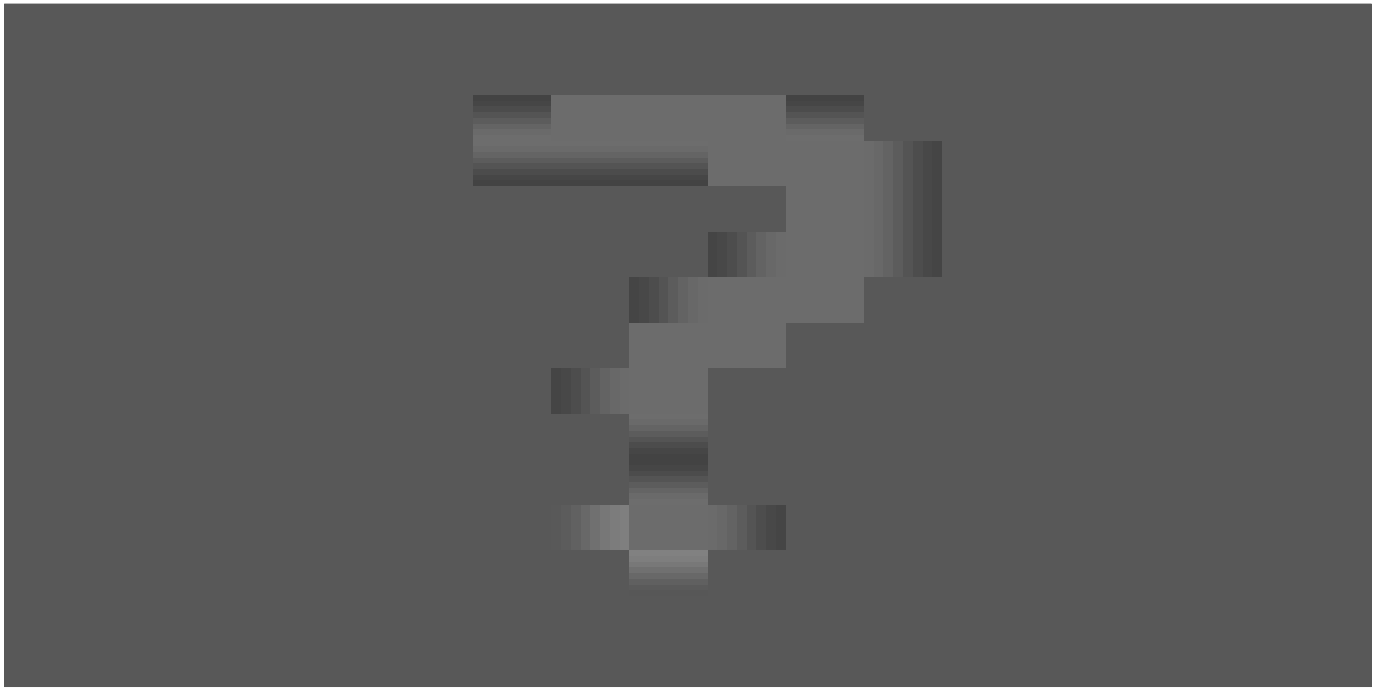
Инфраструктура Streams позволяет:
- Распределять задания между разными узлами
- Подключаться к разным источникам



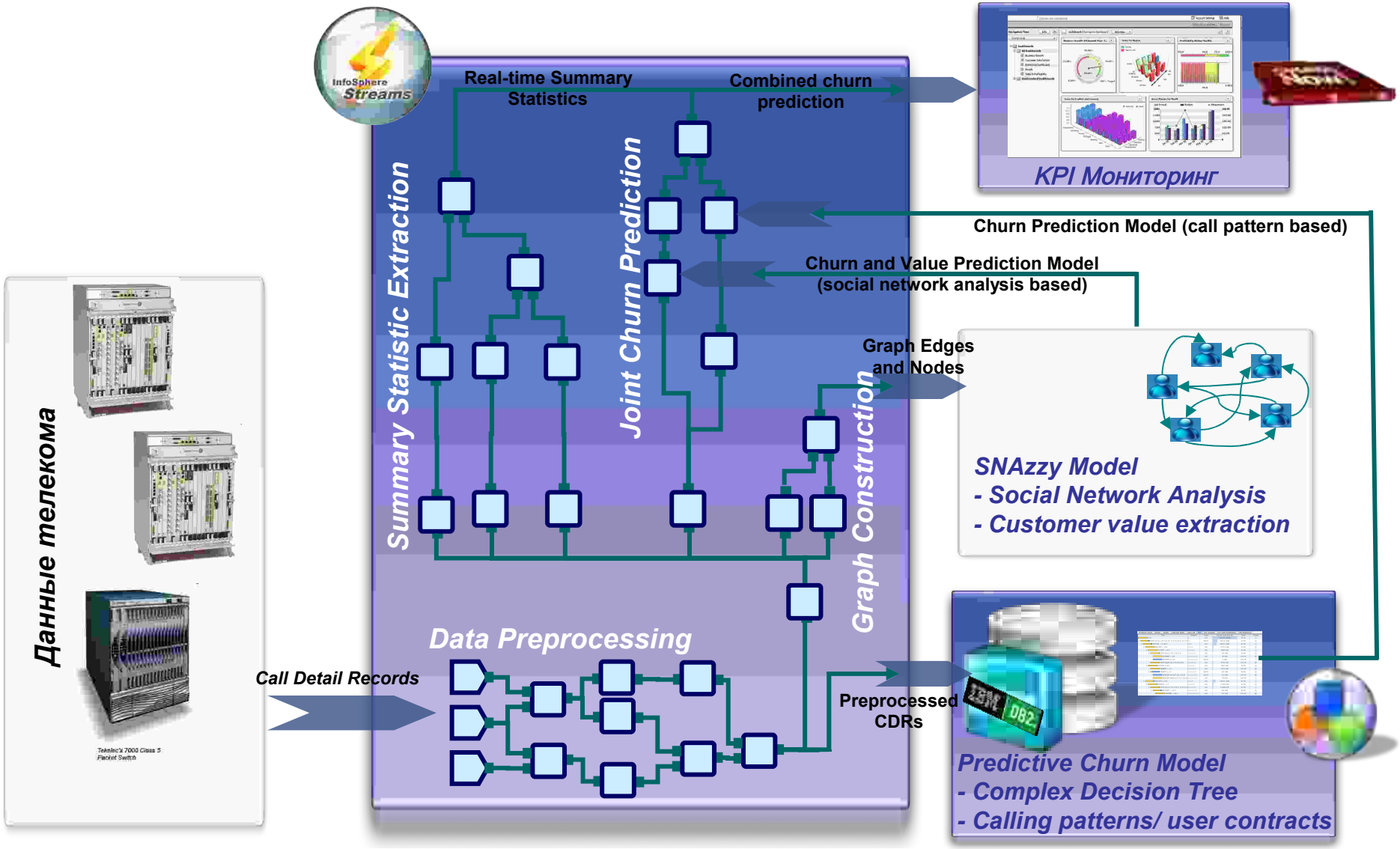
Масштабируемость за счет:
- Разбиения приложения на логические компоненты
- Распределенное выполнение компонент на разных апп. узлах

Там где это применимо:
- Элементы могут быть объединены для уменьшения отклика

Demo: Video Processing (Contour Detection)

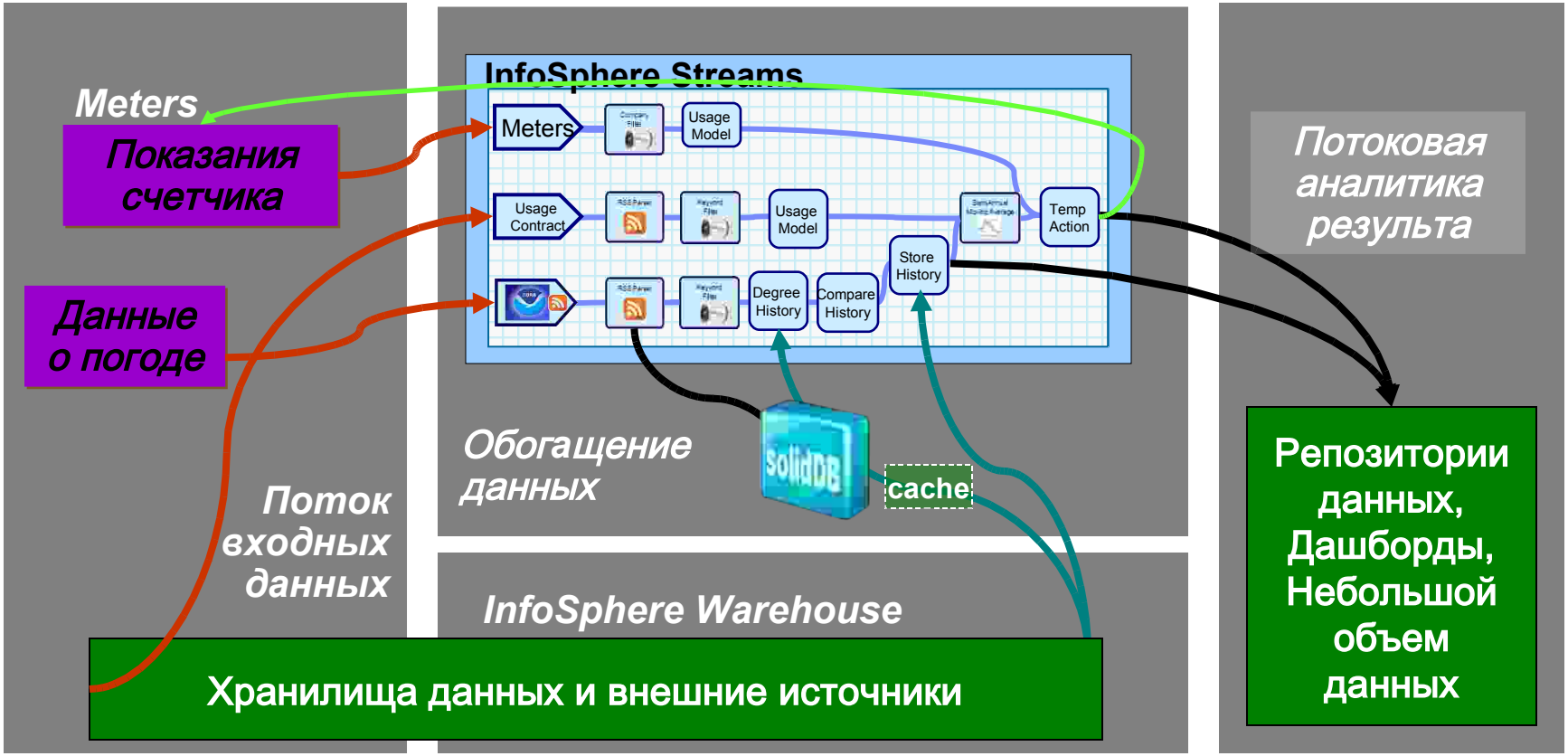


Telephony Architecture



Умные утилиты

- Анализ данных с различных источников
- Принятие решений в реальном времени



IBM InfoSphere Streams v2.0

Удобная среда разработки

- **Eclipse IDE**
- **Streams Live Graph**
- **Streams Debugger**

Распределенная среда выполнения

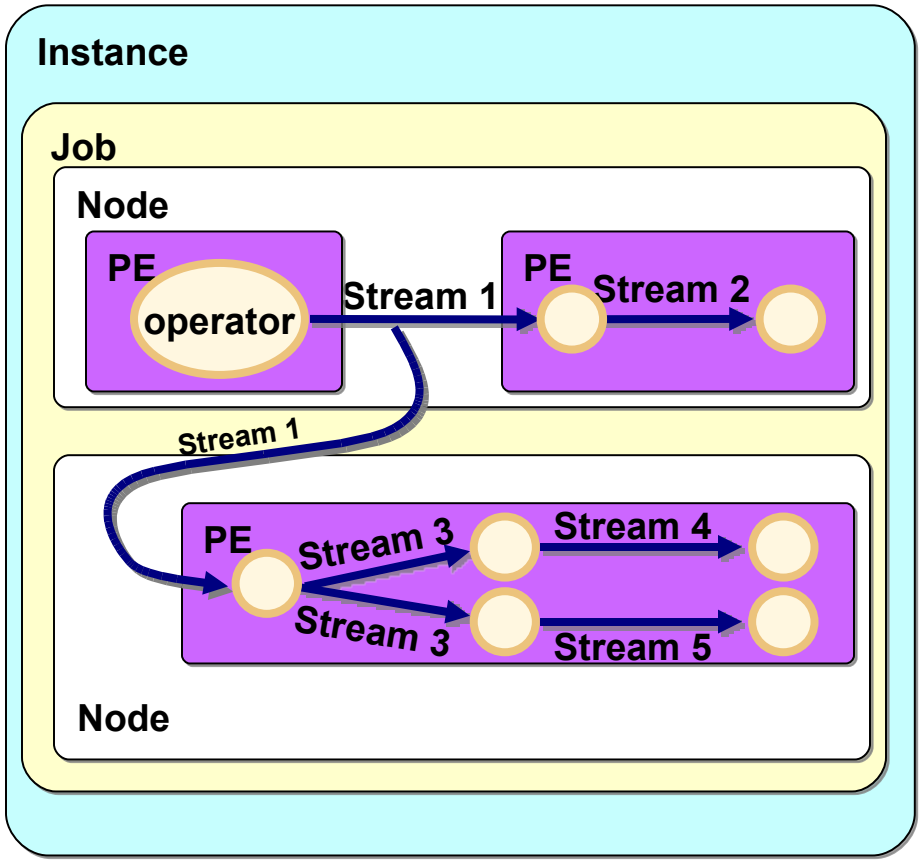
- Clustered runtime for near-limitless capacity
- RHEL v5.3 и выше
- x86 multicore hardware
- Поддержка InfiniBand

Богатый набор инструментов и адаптеров

- Database
- Mining
- Financial
- Standard
- Internet
- **Big Data (HDFS)** ***New***
- **Text** ***New***
- User-defined toolkits

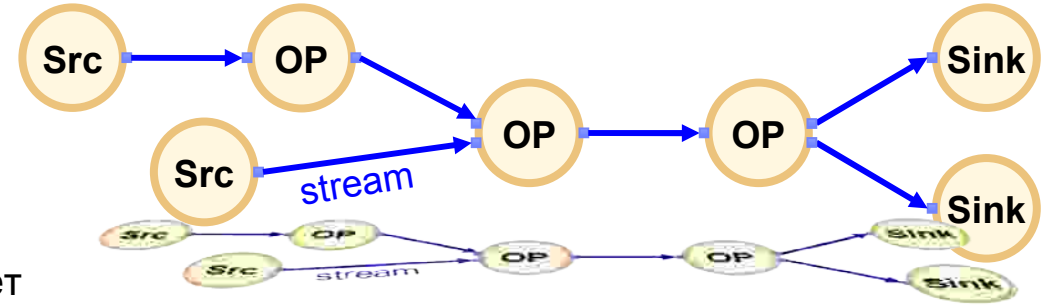
Объекты InfoSphere Streams: Исполнение

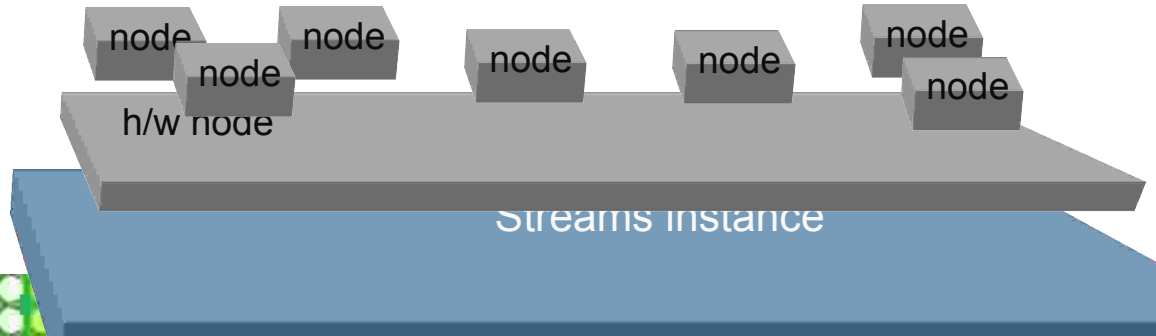
- Instance
  - Исполняемый экземпляр InfoSphere Streams на одном или нескольких хостах
  - Набор компонентов и сервисов
- Processing Element (PE)
  - Основная единица исполнения, запускаемая в инстансе Streams
- Job
  - Развернутое приложение Streams исполняемое в инстансе
  - Состоит из одного или нескольких PEs



От проектирования к развертыванию

- Граф приложения Streams:
  - Направленный (циклический) граф
  - Набор операторов
  - Подключение к потокам
- Каждое законченное приложение может быть переложено в job
- Jobs развертываются в instance
- Instance включает в себя одни node на уровне железа
- Или множество процессных node

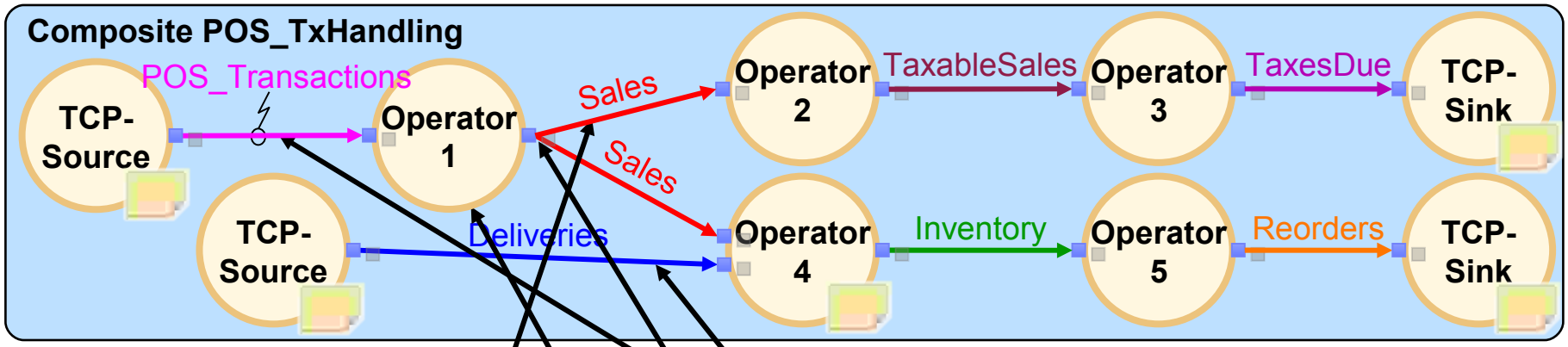




Что такое Streams Processing Language?

- Разработан для потоковой обработки
  - Определяет граф потоков данных
  - Богатый набор типов для определения атрибутов
- Декларативный язык
  - Вызовы операторов определяют входныи и выходные потоки
  - Достаточно указать имена операторов для создания графа
- Процедурный язык
  - Полноценный язык, похожий на C++/Java
  - Возможность программировать логику операторов
- Расширяемость
  - Пользовательские типы данных
  - Пользовательские функции на SPL или на C++ или Java
  - Пользовательские операторы на SPL
  - Пользовательские операторы на C++ или Java

Разработка графа потоков данных при помощи Streams




```
composite POS_TxHandling
{
    graph
        stream<…> POS_Transactions = TCPSource() {…}
        stream<…> Sales = Operator1(POS_Transactions) {…}
        stream<…> TaxableSales = Operator2(Sales) {…}
        stream<…> TaxesDue = Operator3(TaxableSales) {…}
        () as Sink1 = TCPSink(TaxesDue) {…}
        stream<…> Deliveries = TCPSource() {…}
        stream<…> Inventory = Operator4(Sales,Deliveries) {…}
        stream<…> Reorders = Operator5(Inventory) {…}
        () as Sink2 = TCPSink(Reorders) {…}
}
```

Инструменты и операторы для ускорения и упрощения разработки

Standard Toolkit

Relational Operators

Filter	Sort
Functor	Join
Punctor	Aggregate

Adapter Operators

FileSource	UDPSource
FileSink	UDPSink
DirectoryScan	Export
TCPSource	Import
TCPSink MetricsSink	

Utility Operators

Custom	Split
Beacon	DeDuplicate
Throttle	Union
Delay	ThreadedSplit
Barrier	DynamicFilter
Pair	Gate
JavaOp	

Стандартный инструментарий поставляется вместе с продуктом

Internet Toolkit

InetSource

HTTP	FTP	HTTPS
FTPS	RSS	file

Database Toolkit

ODBCAppend	ODBCEnrich
ODBCSource	SolidDBEnrich
DB2SplitDB	DB2PartitionedAppend

Поддержка: DB2 LUW, IDS, solidDB, Netezza, Oracle, SQL Server, MySQL

- Financial Toolkit
- Data Mining Toolkit
- Big Data toolkit
- Text Toolkit
- User-Defined Toolkits
  - Расширение языка

Обзор продукта IBM Netezza

Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData

Что такое аналитический комплекс?

- Оптимизированное, изначально спроектированное под аналитику ядро
- Интегрированные СУБД, СХД и серверные мощности
- От 10 до 100 раз быстрее традиционных РСУБД на задачах аналитики
- Низкая стоимость владения
- Стандартные интерфейсы (SQL, ODBC, JDBC, OLE DB)
- Минимум администрирования и настройки
- Поддержка аналитики SAS, SPSS, R и других
- “Зелёное решение”

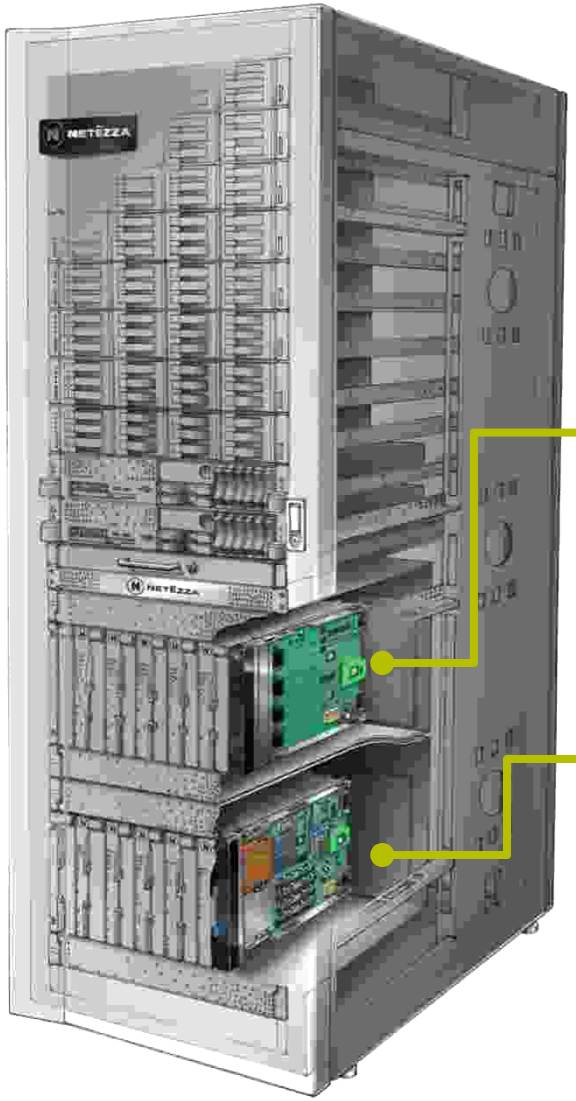
Архитектура IBM Netezza 1000

Оптимизированное «железо» + ПО

Разработано специально для высокопроизводительной аналитики, не требует настройки

Настоящий MPP

Все процессоры нагружены для максимальной скорости и производительности



Обработка потоков данных

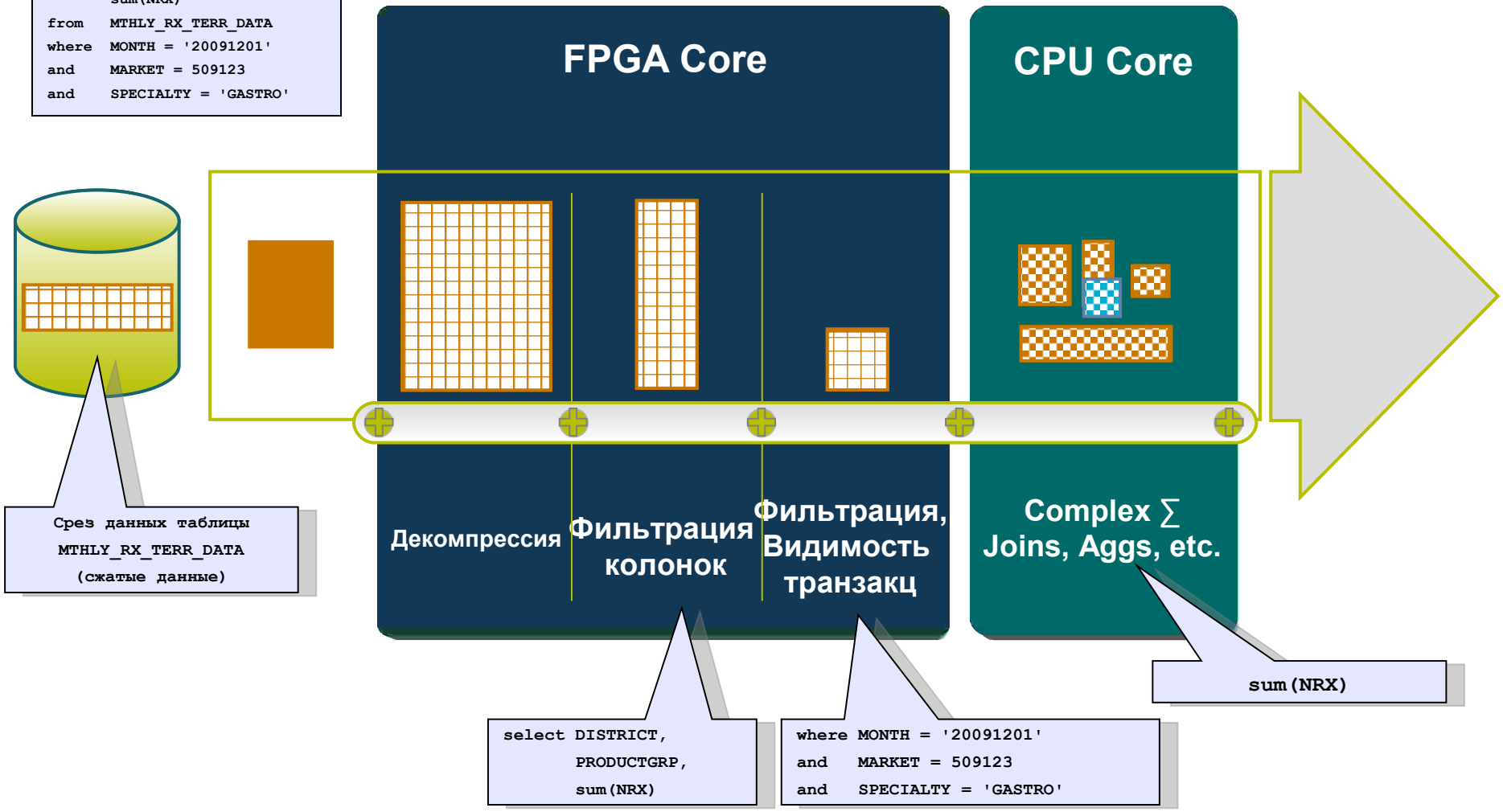
Аппаратное ускорение запросов

Сложная аналитика

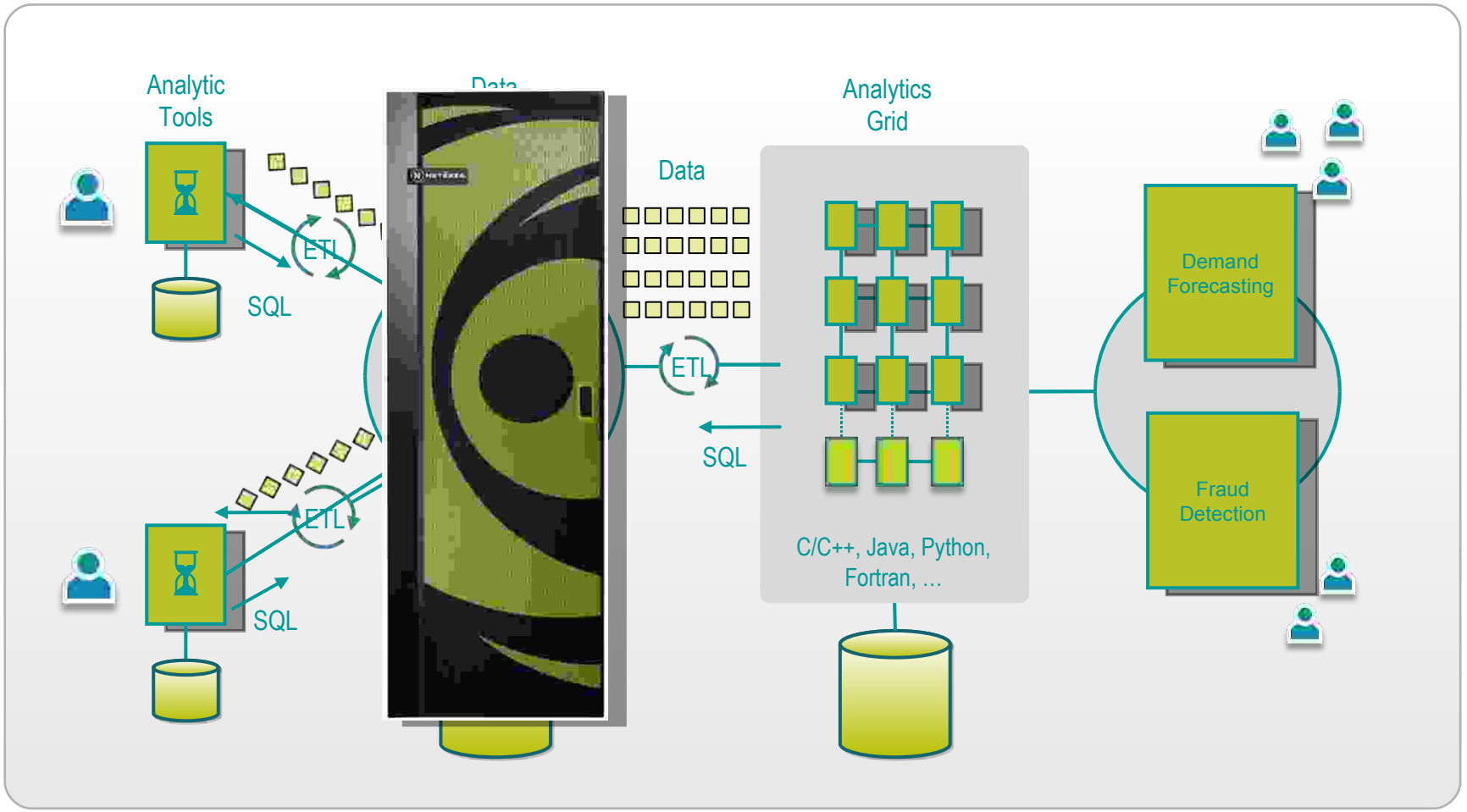
Сложная аналитика выполняется внутри СУБД

Наш секретный соус

```
select DISTRICT,
       PRODUCTGRP,
       sum(NRX)
from   MTHLY_RX_TERR_DATA
where  MONTH = '20091201'
and    MARKET = 509123
and    SPECIALTY = 'GASTRO'
```

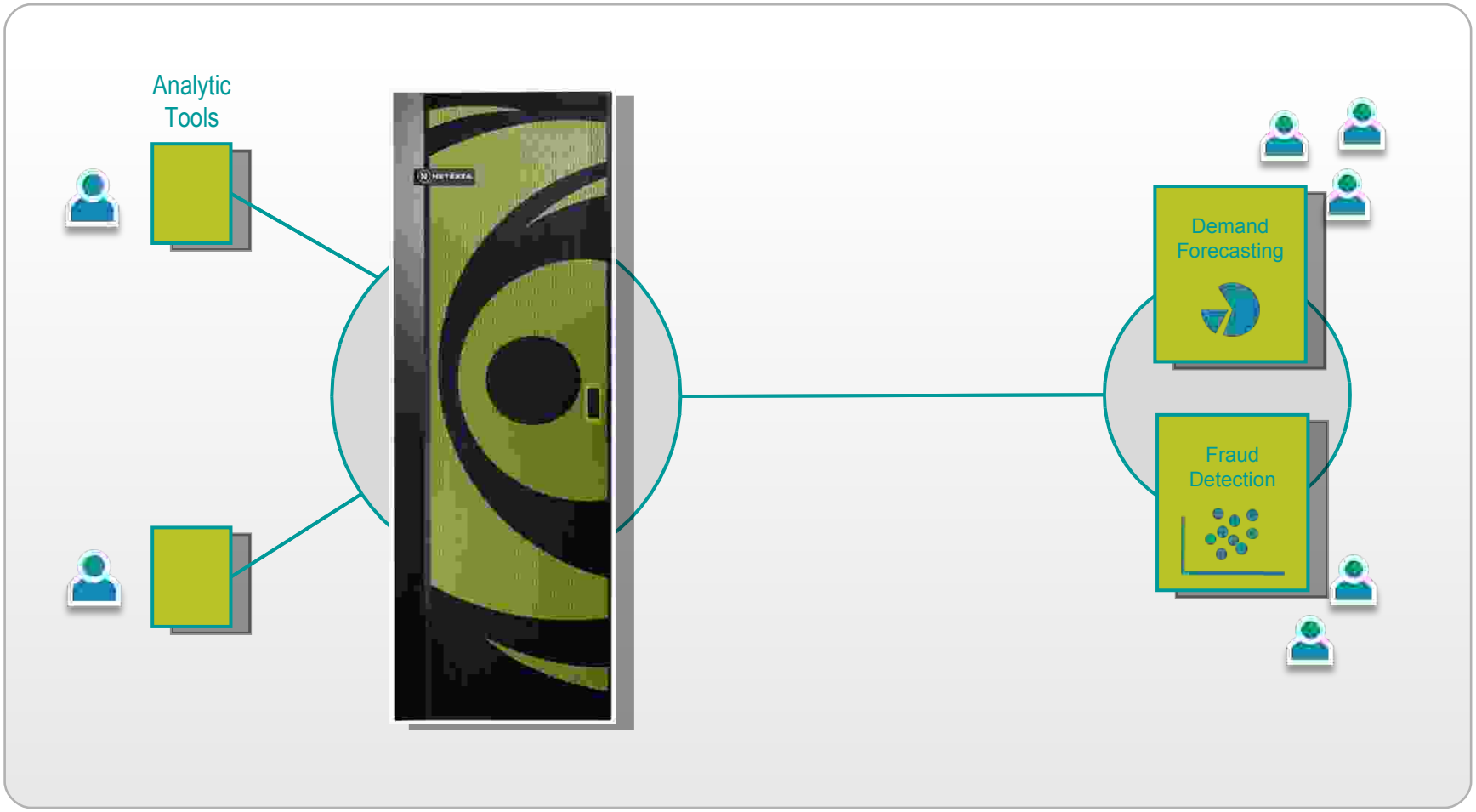


Advanced Analytics with IBM Netezza



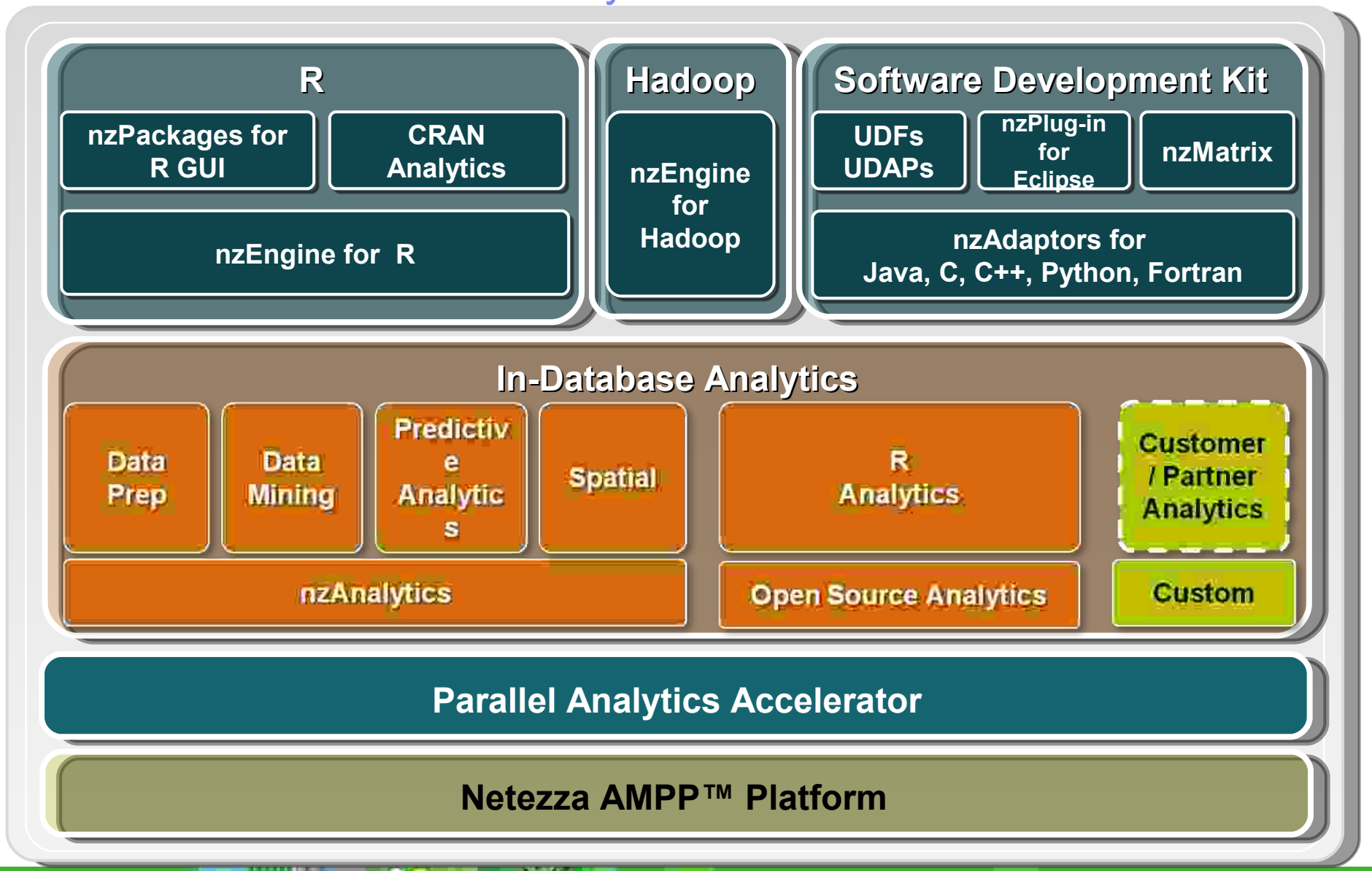
44

Advanced Analytics with IBM Netezza



45

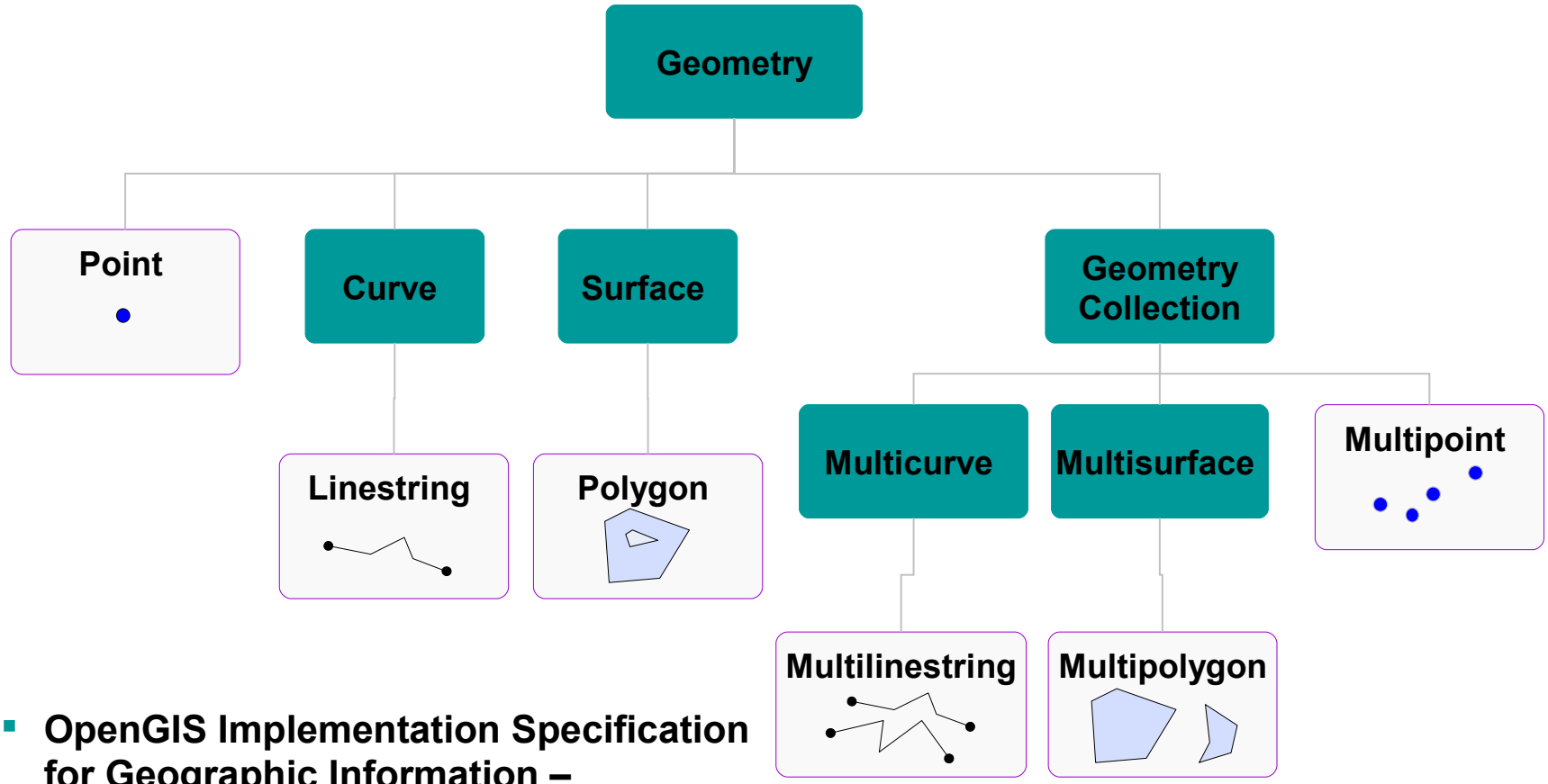
IBM Netezza In-Database Analytics



Преимущества использования MapReduce в IBM Netezza

- Нет необходимости менять код MapReduce, написанный для Hadoop
- Простота – нет необходимости управлять компонентами
  - Нет необходимости указывать количество mappers и reducers
  - Отказоустойчивая и высокоскоростная внутренняя сеть
- Надежное управление нагрузкой
  - Поддерживает смешанную нагрузку – MapReduce и SQL, множество заданий
- Соблюдение ACID
  - Детерминистические результаты, если данные добавляются в таблицу по которой выполняется запрос
- Отказоустойчивость и целостность данных
- Безопасность на уровне объектов и строк
- Управление пользователями
- Интеграция с СУБД
  - Доступны все SQL операции
  - Возможность выбора подходящей технологии для разных задач

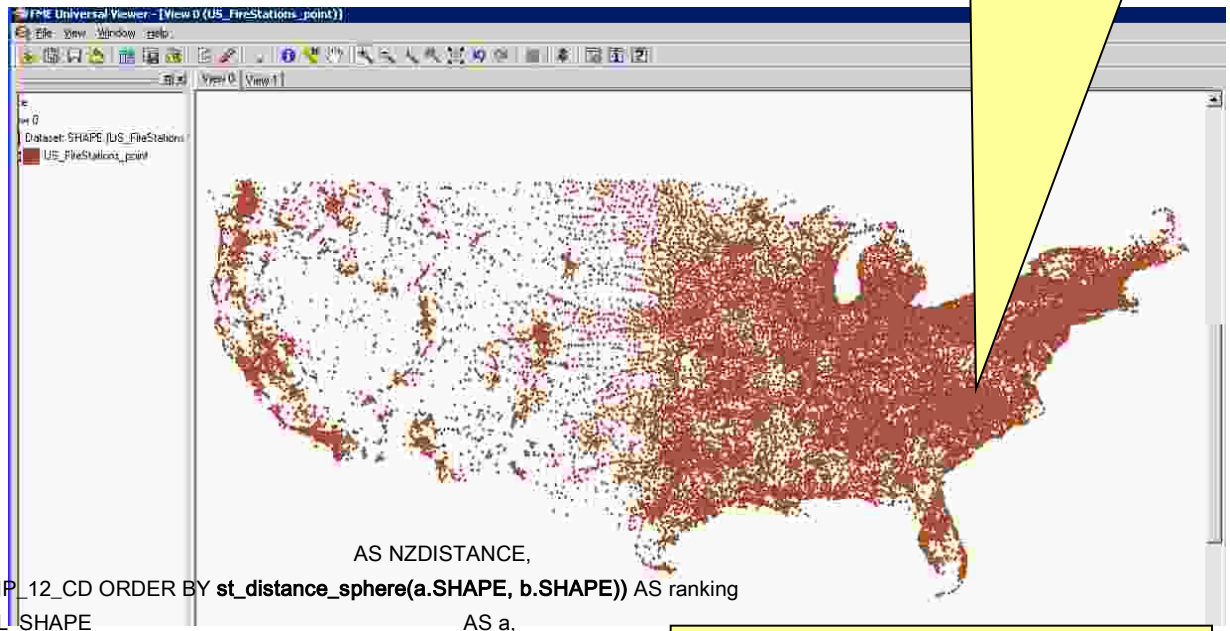
Геопространственные типы



- **OpenGIS Implementation Specification for Geographic Information – Simple Feature Access v. 1.2.0**

Netezza Spatial (5 ближайших пожарных станций)

- 5 ближайших пожарных станций к домовладению
  - 41,000 пожарных станций
  - 4,600,000 домоладений (Иллинойс).



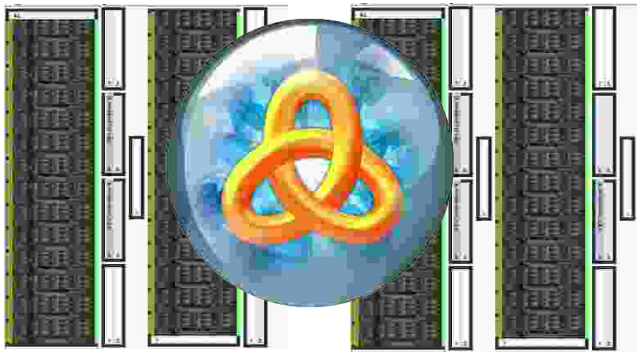

```
INSERT
INTO   ADMIN.FIRE_STATIONS_DISTANCE_NRAL
    (SELECT *
    FROM    (SELECT  a.*                                    ,
                b.STATION_ID                                ,
                b.DPTID                                     ,
                st_distance_sphere(a.SHAPE, b.SHAPE)                AS NZDISTANCE,
                row_number() over (partition BY a.SCRB_ZIP_12_CD ORDER BY st_distance_sphere(a.SHAPE, b.SHAPE)) AS ranking
            FROM    TDM_NRAL_DISTINCT_NEW_SPATIAL_SHAPE             AS a,
                US_FIRE_STATIONS_V2                             AS b
            WHERE   a.SCRB_ST_CD ='IL'
             AND    st_y(b.SHAPE) BETWEEN a.LAT_DEG-.33 AND     a.LAT_DEG+.33
             AND    st_x(b.SHAPE) BETWEEN a.LNG_DEG-.33 AND     a.LNG_DEG+.33
            ) AS subsel
    WHERE   ranking BETWEEN 1 AND    5
    );
Information: Stream execution complete, Elapsed=237.0 sec, CPU=0.24 sec
```

4 минуты – Иллинойс

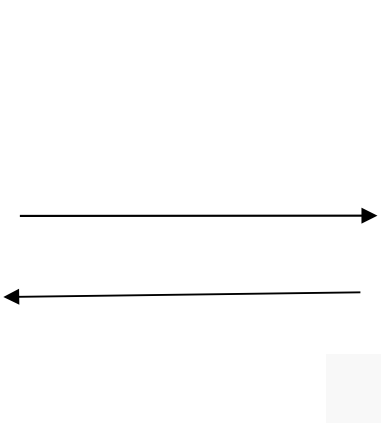
Netezza и BigInsights

- Задания BigInsights могут писать и читать в/из Netezza
  - Можно обогащать хранилище результатами из BigInsights
  - Можно использовать данные из хранилища в BigInsights

- Модуль Jaql для Netezza– включен в BigInsights

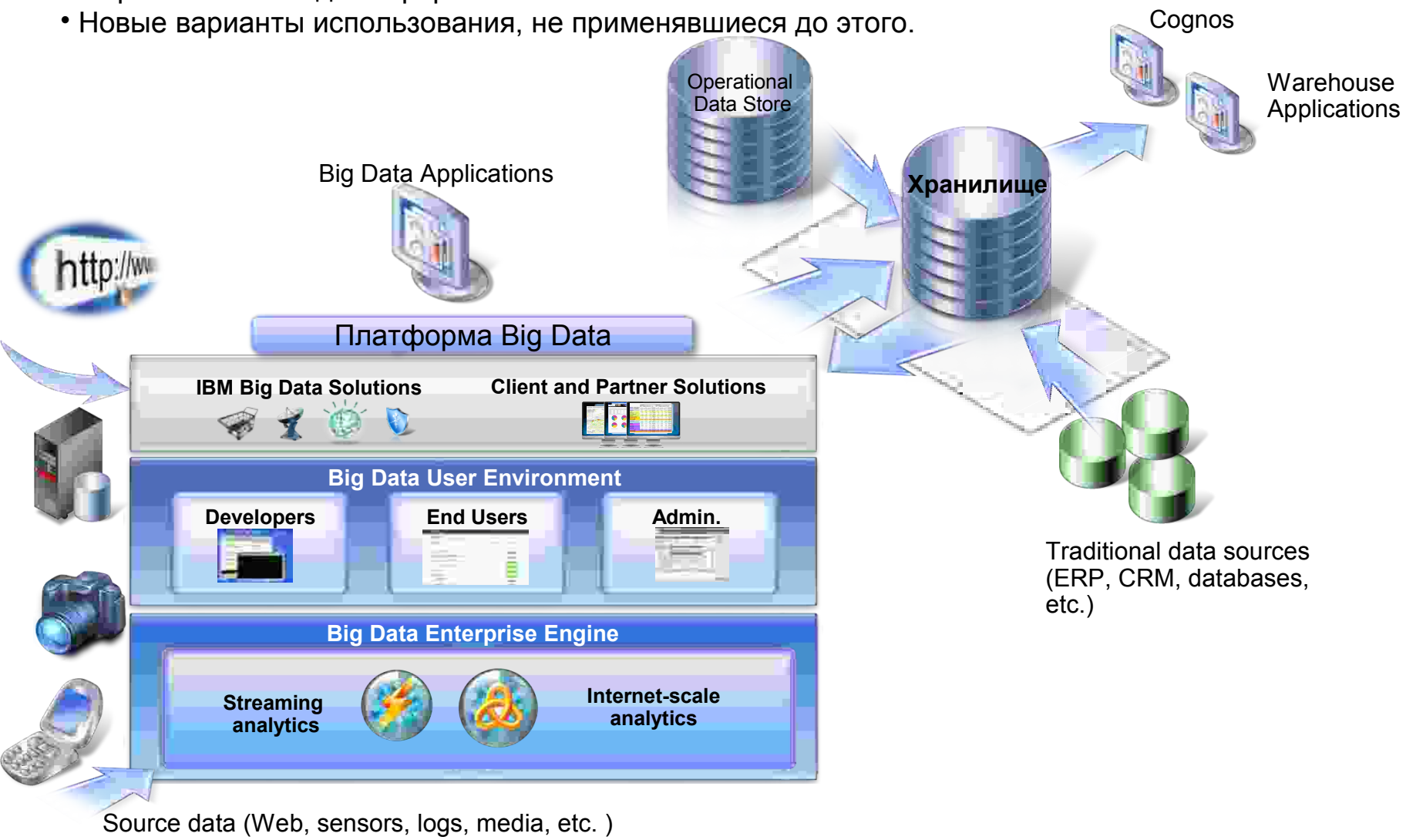


BigInsights



Big Data: интегральная часть предприятия

- Управление данными с момента их возникновения
- Обработка в исходном формате
- Новые варианты использования, не применявшиеся до этого.



В заключение

- Объемы данных растут колоссально
- Аналитика становится «всеядной», структурировать данные часто невозможно
- Измерения, от которых зависят требования к аналитике
  - Время на принятие решения
  - Объемы
  - Скорость
  - Разнообразие структуры
- Стремительно развивается инфраструктура для параллельного программирования по модели MapReduce
- Отказ от реляционных СУБД не является целесообразным
- Имеет смысл гибридный подход (СУБД + MapReduce, статистические языки и пакеты)
- Имеет смысл совместное использование СУБД + потоковая обработка Streams + MapReduce

- Платформа IBM BigData покрывает весь спектр аналитических задач

Спасибо за Ваши вопросы!

Запасные слайды

Андрей Выходцев (andrey.vykhodtsev@ru.ibm.com)

Консультант по направлениям Netezza и BigData

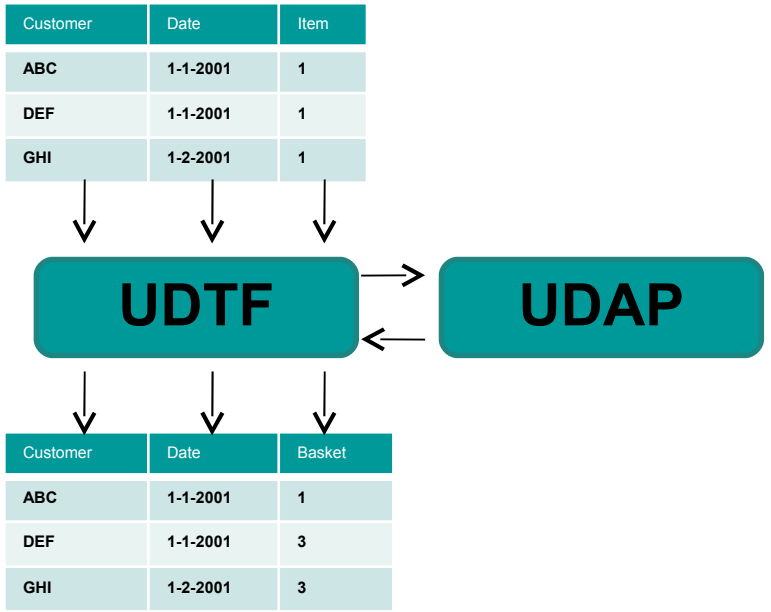
Newest Extension – User-Defined Analytic Process

- UDTF
  - Spawns the UDAP
  - Obeys the UDAP's commands
  - Becomes a server, answering UDAP's requests
    - Serves data.
    - Publishes results.

- UDAP
  - Assumes control
  - Asks the UDTF for data from the data stream.
  - Passes output back to the UDTF

- **UDAP can be written in C/C++, Java, Python, Fortran, R**

- **Run from command-line**



Parallelized In-Database Analytics

Data Prep

Data Profiling / Descriptive Statistics

Probability Density and Inverse Functions

- **Normal**
- **Fisher**
- **Exponential**
- **Uniform**
- **Weibull**
- **Wilcoxn**
- **Man-Whitney**
- **T Student**

General Diagnostic Measures

Error Calculation

- **Classification Error**
- **Mean Absolute Error**
- **Mean Squared Error**
- **Relative Absolute Error**
- **Relative Squared Error**

Statistics

Histogram and Frequency Table

- **Histogram**
- **Bivariate Frequency Table**
- **Univariate Frequency Table**

Quantiles

- **Quantiles**
- **Median**
- **Outliers**
- **Quartile**

Parametric Statistics

- **Chi-Square**
- **T Student**

Non-Parametric Statistics

- **Spearman's Rank Correlation**
- **Man-Whitney-Wilcoxn**
- **Wilcoxn**

Moments

- **Kurtosis**
- **Skewness**

Sampling

Uniform Random Sampling

- **Uniform Random Sampling Count**
- **Uniform Random Sampling Fraction**

Data Prep / Transformations

Binning and Discretization

- **Entropy Minimization**
- **Equal Width**
- **Equal Frequency**

Standardization and Normalization

Parallelized In-Database Analytics

Data Mining

Association Rules Mining

Association

- **A Priori**
- **FP-Growth**

Clustering

K-Means

Hierarchical Clustering

- **Divisive Clustering**

Feature Extraction

Dimension Reduction

- **Principal Components Analysis**

Predictive Analytics

Sample Size

One-Way ANOVA

- **Complete Randomized Design**
- **Randomized Block Design**

Regression

Linear Regression

Classification

Decision Trees

- **Entropy Decision Tree**
- **Gini Index Decision Tree**
- **Regression Tree**

Neighborhood Methods

- **K Nearest Neighbors**

Bayesian Methods

Classifier

- **Naïve Bayes**

Graphical Model

- **Bayesian Networks**

Model Testing

Error Calculation

- **Cross Validation**
- **Percentage Split**
- **Train / Test**

Spatial

Geometric Functions

Geometric Information

Geometric Object Manipulation

Geometric Analytics

Conversion

Comparison

Distance and Area

Netezza Spatial™ Functions

Accessors	Accessors	Constructors	Measures	Outputs
ST_Boundary	ST_Is3D	ST_Ellipse	ST_Area	ST_AsBinary
ST_CoordDim	ST_IsClosed	ST_Point	ST_Centroid	ST_AsKML
ST_Dimension	ST_IsEmpty		ST_Distance	ST_AsText
ST_EndPoint	ST_IsMeasured		ST_Length	ST_GeomFromText
ST_Envelope	ST_IsRing		ST_Perimeter	ST_GeomFromWKB
ST_ExteriorRing	ST_IsSimple		ST_PointOnSurface	ST_WKBToSQL
ST_GeometryN	ST_M			ST_WKTToSQL
ST_GeometryType	ST_MaxM			
ST_GeometryTypeID	ST_MaxX			
ST_GrandMBR	ST_MaxY			
ST_InteriorRingN	ST_MaxZ			
ST_NumGeometries	ST_MBR			
ST_NumInteriorRings	ST_MinM			
ST_NumPoints	ST_MinX			
	ST_MinY			
	ST_MinZ			
	…			

Processing	Referencing	Relationships
ST_Buffer	ST_LocateAlong	ST_Contains
ST_ConvexHull	ST_LocateBetween	ST_Crosses
ST_Difference		ST_Disjoint
ST_Expand		ST_DWithin
ST_Intersection		ST_Equals
ST_SymDifference		ST_Intersects
ST_Union		ST_MBRIntersects
		ST_Overlaps
		ST_Relate
		ST_Touches
		ST_Within

Open Source Analytics

R Analytics

Horizontal

- Bayesian
- Cluster
- Distributions
- Graphics
- Graphical Models
- Machine Learning
- Multivariate
- Natural Language Processing
- Optimization
- Robust Statistical Metrics
- Spatial
- Survival Analysis
- Time Series

Vertical

- Econometrics
- Experimental Design
- Computational Physics
- Clinical Trials
- Environmetrics
- Finance
- Genetics
- Medical Imaging
- Pharmacokinetics
- Phylogenetics
- Psychometrics
- Social Sciences

nzMatrix – Parallelized Linear Algebra package

Matrix Operations
arallel Basic Linear Algebra
asic Linear Algebra
inear Equations
east Squares
igenvalues & Eigenvectors
ingular Value Decomposition

atrix Factorization & Inversion

atrix Element Scalar Functions

atrix Reduction Functions

- Accessible from
  - R
  - Python
  - Java
  - etc.
- Via
  - ODBC
  - Stored Procedures