Анализ больших данных на платформе Oracle

Ольга Горчинская
Директор по исследовательским проектам
Форс

ФОРС

План

- Краткое введение в Big Data
- Технологии и продукты Oracle для больших данных
- Эксперименты и проекты
- Вопросы и ответы

Введение в Big Data

ФОРС

Большие Данные

- **Сверхбольшие объемы структурированных и неструктурированных данных, с которыми трудно работат с помощью традиционных средств**
- Впервые термин появился 3 сентября 2008 года, Клиффорд Линч, редактор научного журнала Nature, в связи с возможностями для науки накопления научных данных
- Источники больших данных– интернет-документы, социальные сети, блоги, измерительные устройства, радиочастотная идентификация, устройства ауди и видеорегистрации
- Инновационные технологии сбора, хранения данных (Hadoop, MapReduce, NoSQL базы данных и др)
- Новая парадигма анализа данных

VVV
Velocity, Volume, Variety
(скорость, объем, многообразие)

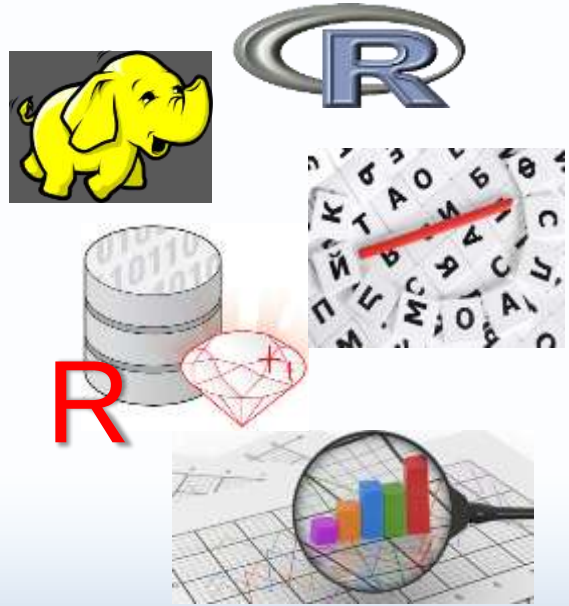
ФОРС

Новые подходы к решению бизнес-задач

Анализ статистики поисковых запросов Google показал, что примерно за несколько недель до того, как медики объявляют об эпидемии гриппа, в этом регионе отмечается всплеск поисковых запросов по его лечению. То есть, анализ запросов сообщает нам о начале эпидемии быстрее, чем официальная медицина

Технологии и инструменты

- Hadoop & MapReduce
- NoSQL базы данных
- Углубленная аналитика
  - *Статистика,*
  - *предиктивная аналитика и data mining*
  - *Лингвистическая обработка текстов*
- Инструменты класса Data Discovery



Hadoop

- Система распределенных вычислений, проект фонда Apache Software Foundation
- Свободно распространяемые функции, утилиты, framework для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов
- Разработка Hadoop началась в 2005г Дугом Каттингом (Doug Cutting) под влиянием публикации сотрудников Google (Джеффри Дин и Санжай Гемават) о вычислительной концепции MapReduce
- 2008г: в феврале Yahoo запустила кластерную поисковую машину на 10 тыс ядер под управлением Hadoop; апрель –побит мировой рекорд по стандартным тестам на сортировку (1Тб – 209 сек – 910 узлов
- Yahoo (более 4 тыс. Узлов, 15 Пбайт каждый), Facebook (около 2 тыс. узлов на 21 Пбайт) и Ebay (700 узлов на 16 Пбайт)

Общие принципы построения Hadoop систем

- Компоненты: Hadoop Common, HDFS, YARN, Hadoop MapReduce
- Большое количество (до десятков тысяч) узлов, на основе относительно дешевого оборудования
- Каждый узел является сервером и хранения и обработки
- Обработка данных ведется в массивно-параллельном режиме
- MapReduce
- Данные хранятся в нескольких копиях (обычно в трех) и отказ узла или двух не ведет к потере данных
- Система практически неограниченно масштабируется

NoSQL базы данных

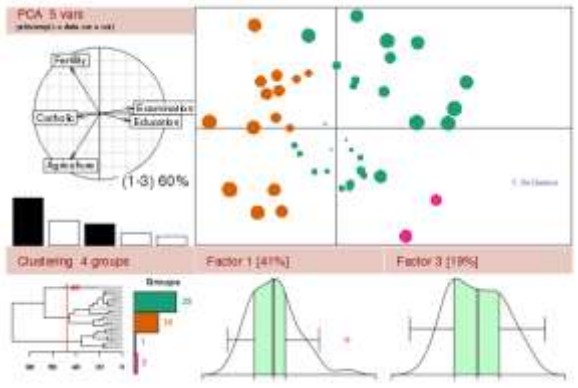
- Нереляционные СУБД
- Позволяют горизонтально масштабироваться
  - Данные хранятся в копиях (обычно 3) на разных серверах
  - Отказ узла не приводит к потере данных
  - Чтение производится с мастера или копий
- Всю логику обрабатывает приложение
- Преимущество -- скорость обработки

NoSQL базы данных

- Нет традиционной структурированной схемы
  - Часто данные хранятся в схеме Key-Value
  - Интерпретация смысла данных – в приложении
- Целостность данных поддерживается не так жестко, как в традиционных RDBMS
- Нет стандартов
  - Существует около 130 NoSQL СУБД, очень много OpenSource

Язык статистических исследований R

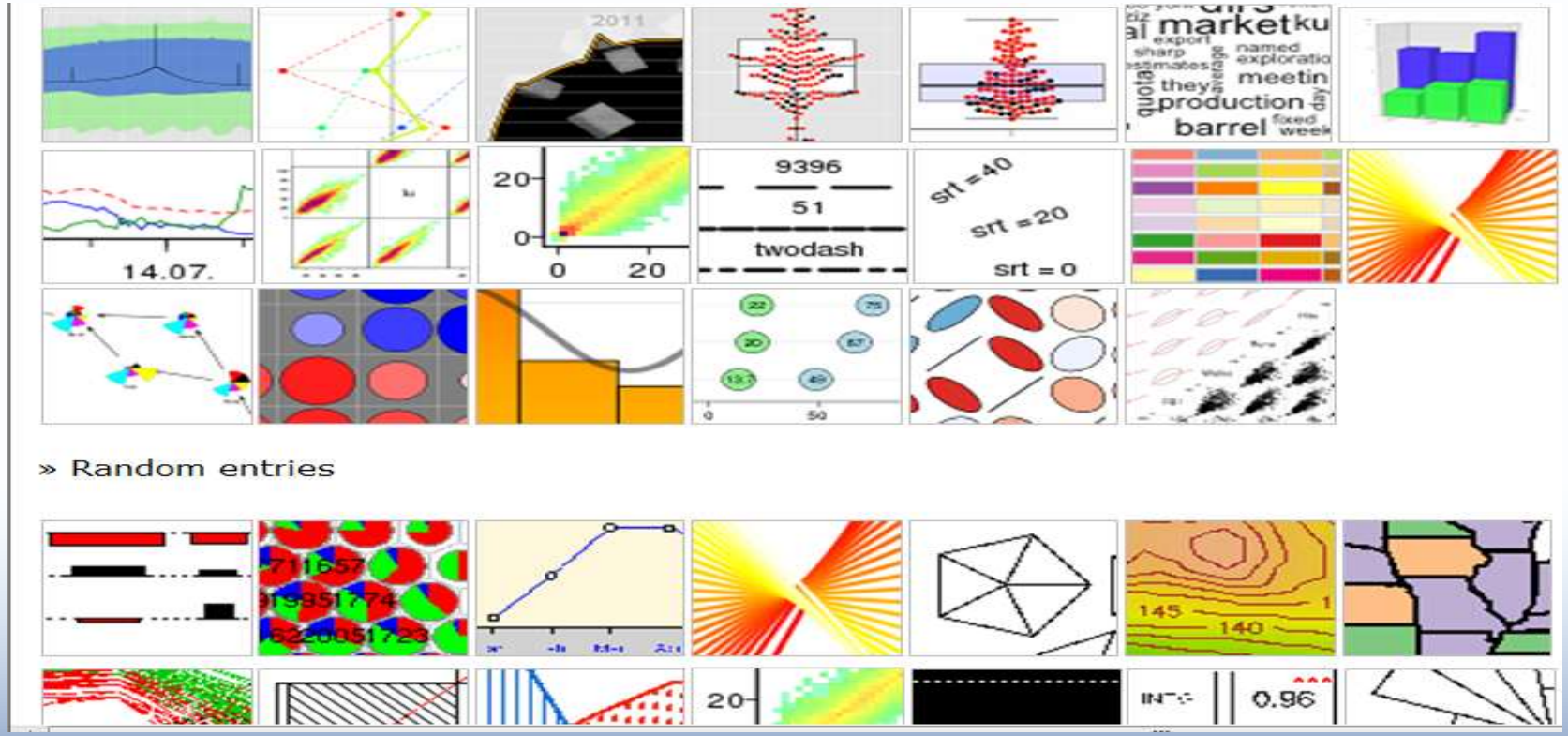
- **Open source проект**, R Foundation
- Язык для статистических исследован ий и работы с графикой (Росс Айхэк, Роберт Джентельмен, Оклендский ун-т, 1997)
- Широкий спектр различных функций (временные ряды, прогнозирование, классификация, кластеризация)
- Возможность расширения, технология разработки дополнительных пакетов участниками проекта астеризация и др) -- 2014г более 5500 пакетов
- Высокий уровень популярности



The R Project for Statistical Computing

http://www.r-project.org/

Визуализация в R



ФОРС

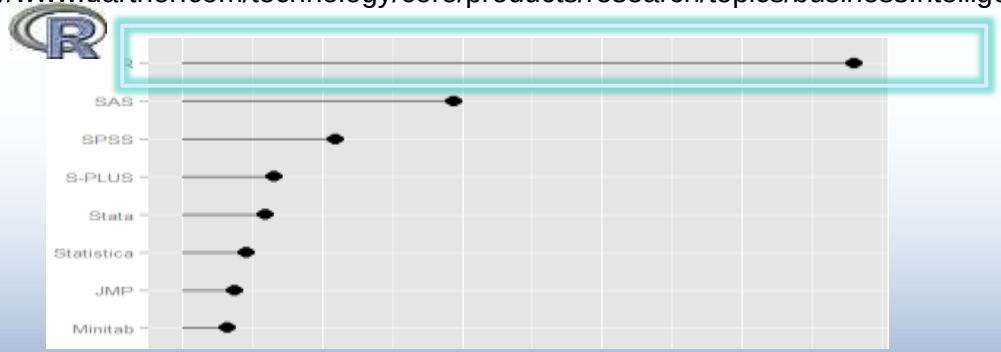
Open Source

Частично благодаря появлению концепции Big Data, бизнес-анализ(BI) остается быстро растущим рынком …. Одновременно с ростом рынка BI постоянно увеличиваются инвестиции в предиктивную аналитику; ***R является не только хорошим готовым инструментом, но и идеальной средой для исследований в области углубленной аналитики****. R ориентирован на расширения и интегрируется с инструментаим бизнес-анализа , обогащая отчеты глубокой аналитикой.*

Gartner

Hype Cycle for Analytic Applications, 2011, 30 August 2011

http://www.gartner.com/technology/core/products/research/topics/businessIntelligence.jsp



Кол-во f web site линков, которые указывают на основной сайт инструментальной среды March 19, 2011.

http://www.r4stats.com/popularity

ФОРС

Автоматическая обработка текстов

- Поддержка русского языка
- Продукты отечественных компаний
- Иногда можно обойтись без глубокой лингвистики, но почти всегда нужна морфология
- Возможности инструментов
  - *Поиск с учетом морфологии, синтаксиса, семантики*
  - *Извлечение сущностей или объектов*
  - *Извлечение фактов*
  - *Классификация, рубрикация, тональность*
  - *Определение тематики*

ФОРС

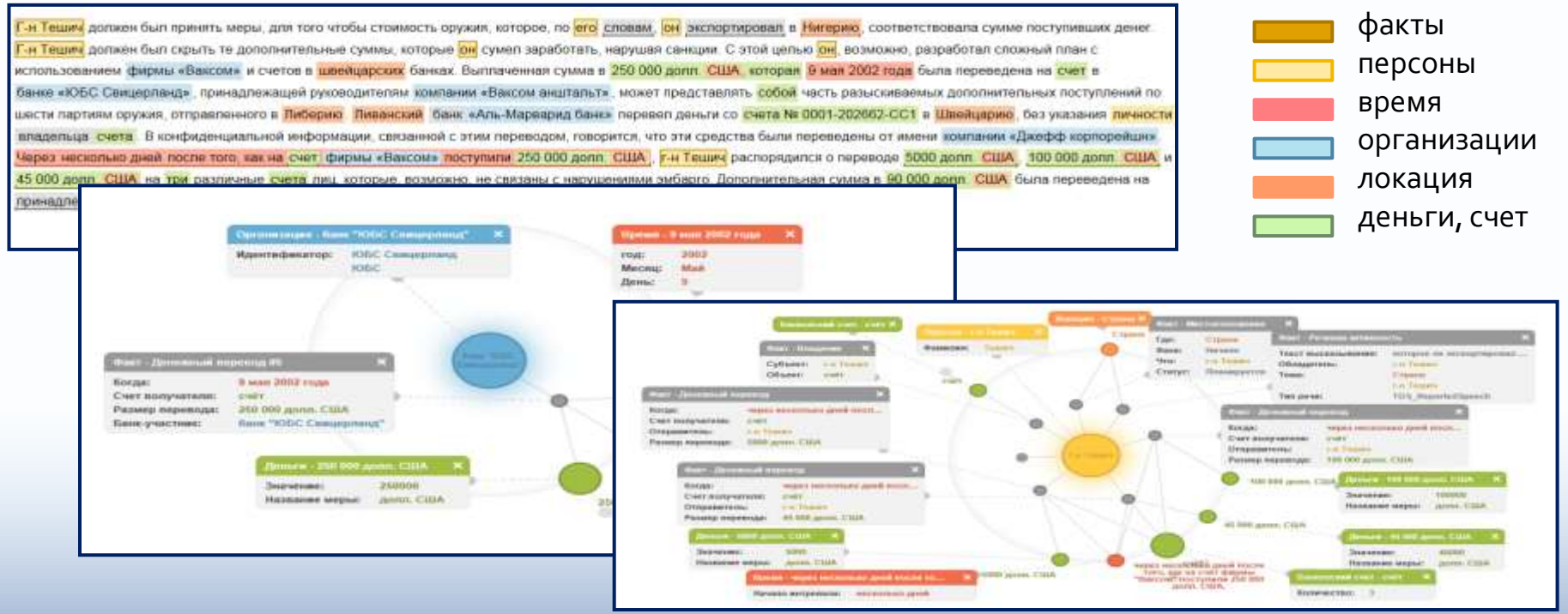
Технологии RCO (www.rco.ru)

- Содержательный портрет текста
- Упоминания персон и организаций
- Упоминания особых объектов
- Распознавание ситуаций в тексте
- Отношение к объекту в тексте
- Разбор частично-структурированного текста
- Тематическое рубрицирование текстов
- Кластеризация новостей
- Поиск документов
- Выявление заимствований и поиск похожих текстов
- …

ФОРС

ABBYY Compreno

Семантический разбор текста и визуализация



- факты
- персоны
- время
- организации
- локация
- деньги, счет

ФОРС

От бизнес-анализа к исследованию данных

Business Intelligence

- Анализ числовых показателей
- Внутренние данные

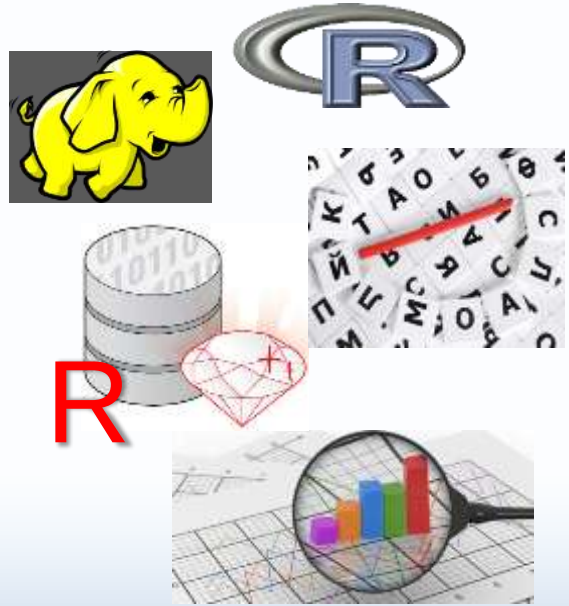
- Устойчивый состав данных
- Качество и точность данных

Data Discovery

- Преобладают поисковые операции
- «Чужие» данные
- Состав данных быстро изменяется
- Не качество, а количество

Технологии и инструменты

- Hadoop & MapReduce
- NoSQL базы данных
- Углубленная аналитика
  - *Статистика,*
  - *предиктивная аналитика и data mining*
  - *Лингвистическая обработка текстов*
- Инструменты класса Data Discovery

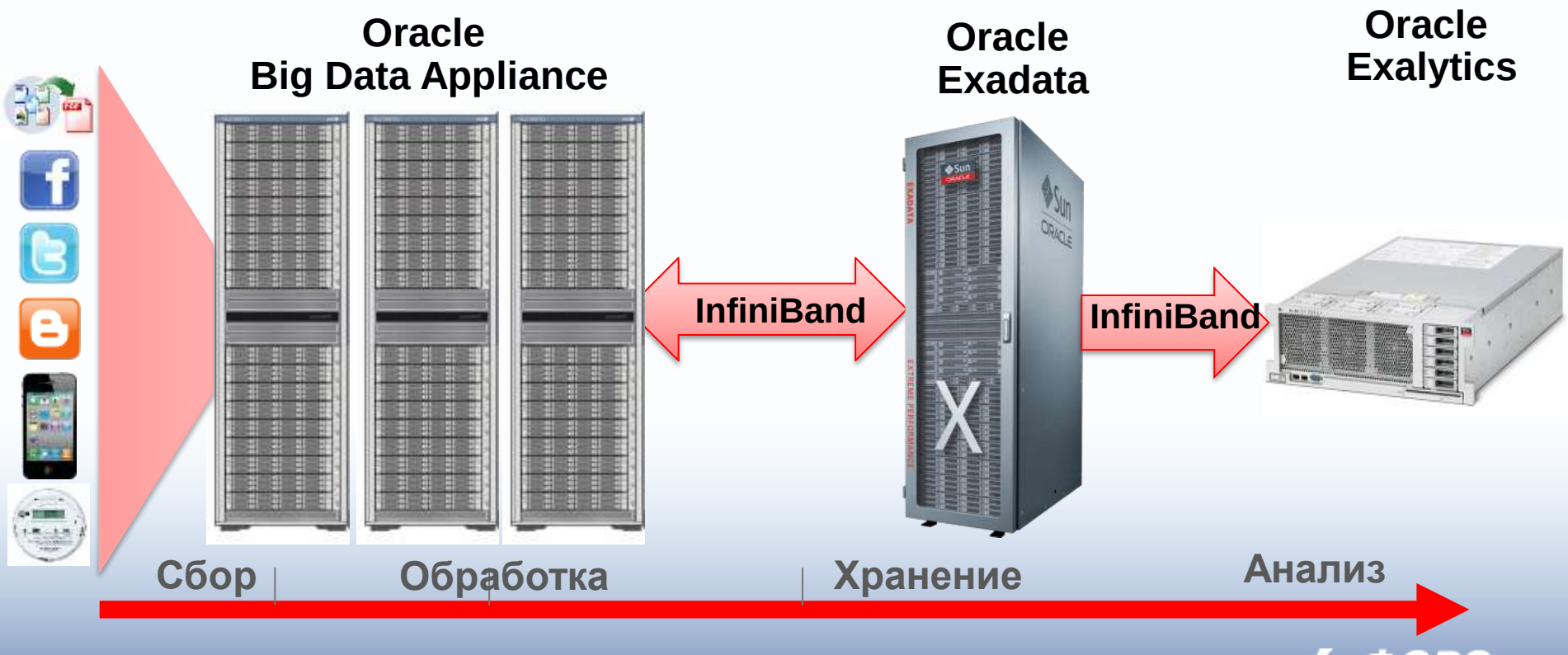


Hardware and Software
ORACLE
Engineered to Work Together

Платформа Oracle для Big Data

ФОРС

Платформа Oracle для больших данных



ФОРС

Oracle Big Data Appliance

Аппаратное обеспечение

ORACLE
BIG DATA APPLIANCE

- Кластер из 18 узлов (18 Sun X4270 M2)
- На каждом узле
  - *64 GB RAM (всего 1152GB RAM)*
  - *16 ядер Intel (всего 288 ядер)*
  - *48 ТБ дисков (всего 864TB)*
- 40 Gb p/sec InfiniBand
- 10 Gb p/sec Ethernet

ФОРС

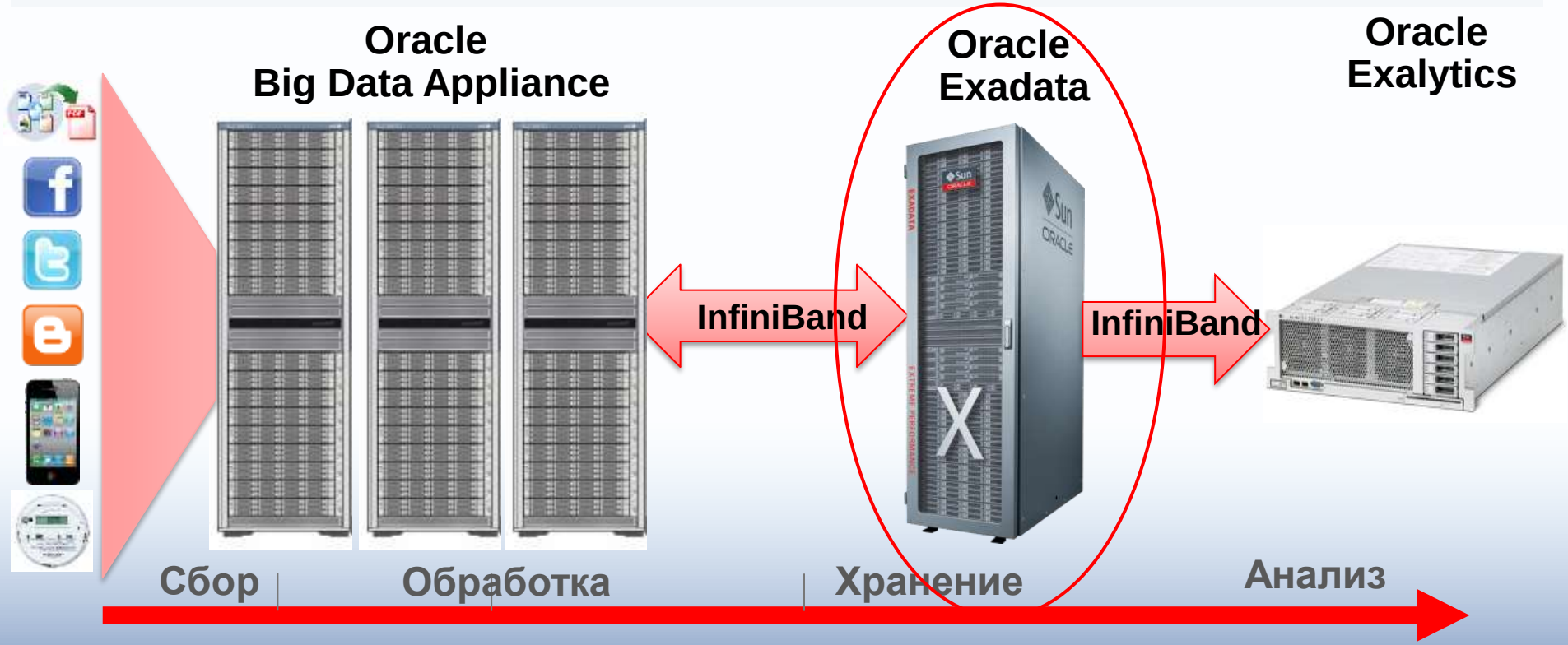
Oracle Big Data Appliance

Программное обеспечение

- Oracle Linux
- Java Hotspot VM
- Cloudera Hadoop Distribution
  - *Hadoop Core, HDFS, Hive, HBase, Zookeeper, Oozie, Mahout, Sqoop*
- R Distribution
- Oracle NoSQL Database
- Oracle Big Data Connectors

ФОРС

Платформа Oracle для больших данных



ФОРС

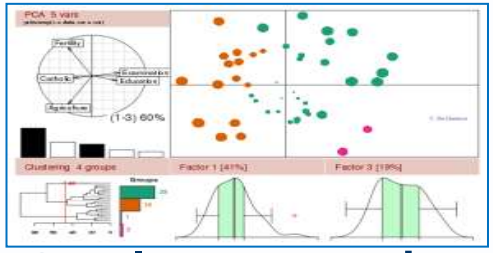
Oracle Exadata

ORACLE EXADATA

- Машина баз данных
- Построена на основе Oracle Database
- Инновации
  - *Интеллектуальная СХД (Smart Scan, InfiniBand)*
  - *Интеллектуальный PCI Flash Cash (ускорение I/O до 30 раз, сканирование снижается в 3 раза)*
  - *Гибридное колоночное сжатие (10 – 15-кратное сжатие)*

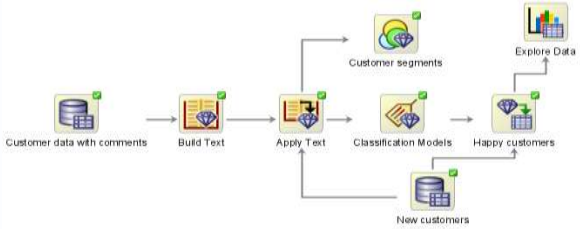
ФОРС

Oracle In-Database Advanced Analytics



Oracle R Enterprise

- Распространенный язык статистических исследований R -- open source
- Встроен в Oracle Database – R-вычисления транслируются и выполняются в Oracle Database
- R интегрирован с Hadoop & OBIEE



Oracle Data Mining

- Встроенные в базу данных процедуры data mining
- API для разработки приложений, встраивание data mining в существующие приложения и системы
- Oracle Data Miner

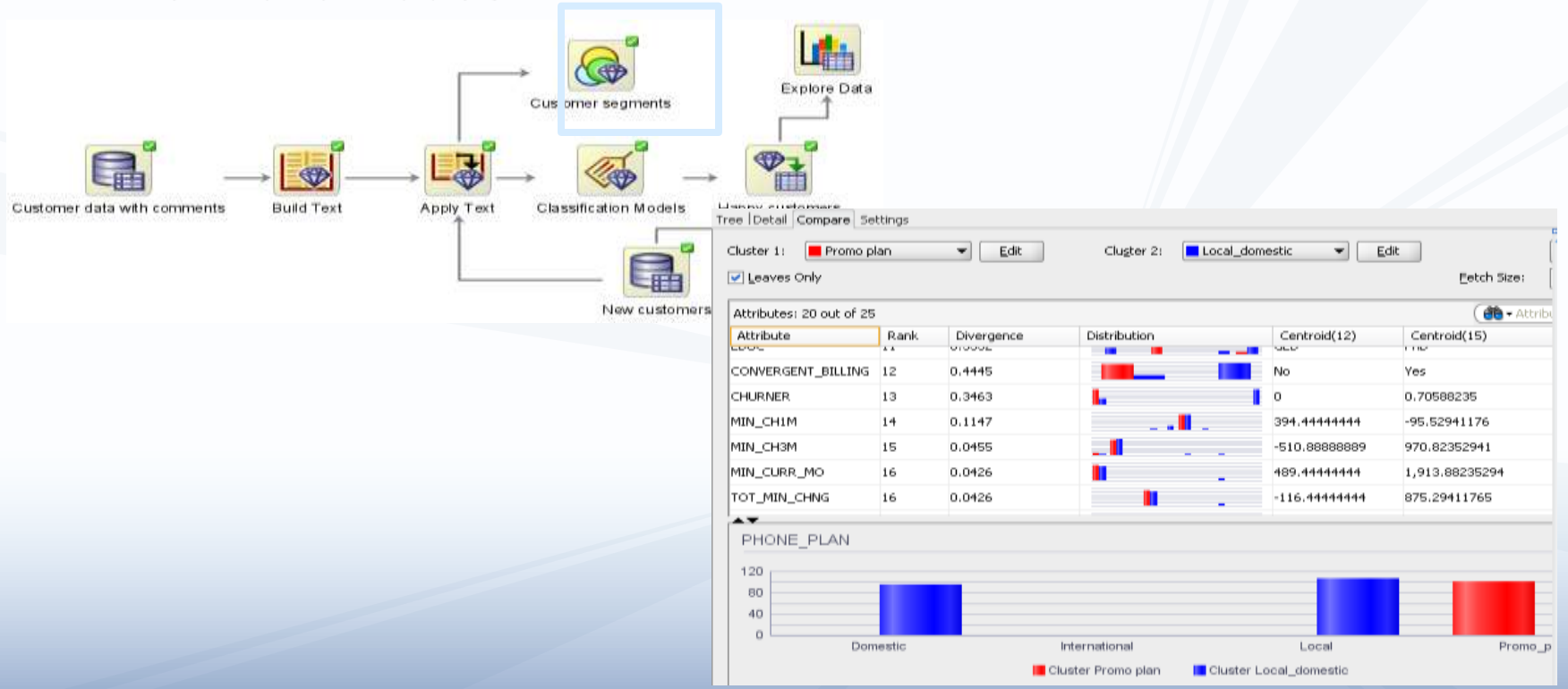
Статистика Визуализация Data & Text Mining Предиктивная аналитика

Oracle Data Mining Algorithms

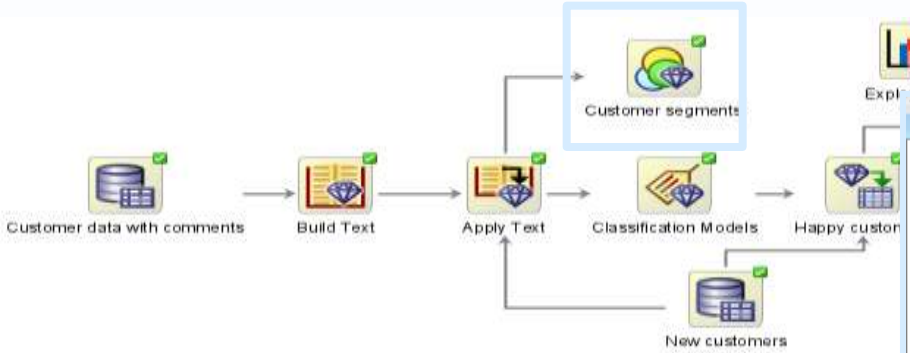
Задача	Алгоритм	Применение
Классификация 	Logistic Regression (GLM) Decision Trees Naïve Bayes Support Vector Machine	Прогнозирование принадлежности объекта к одному из заданных классов
Регрессия 	Multiple Regression (GLM) Support Vector Machine	Прогнозирование непрерывных показателей
Выявление аномалий 	One Class SVM	Отсутствие «отрицательных» примеров
Значимость атрибутов 	Minimum Description Length (MDL)	Сокращение числа атрибутов Выделение важных атрибутов
Ассоциативные правила 	Apriori	Анализ рыночной корзины Аализ связей (Link analysis)
Кластеризация 	Enhanced K-means O-cluster Expectation Maximization (EM)	Сегментация клиентской базы Text mining Gene and protein analysis
Выявление признаков 	Non-Negative Matrix Factorization Singular Value Decomposition Principal Component Analysis	Анализ текстов Уменьшение числа признаков

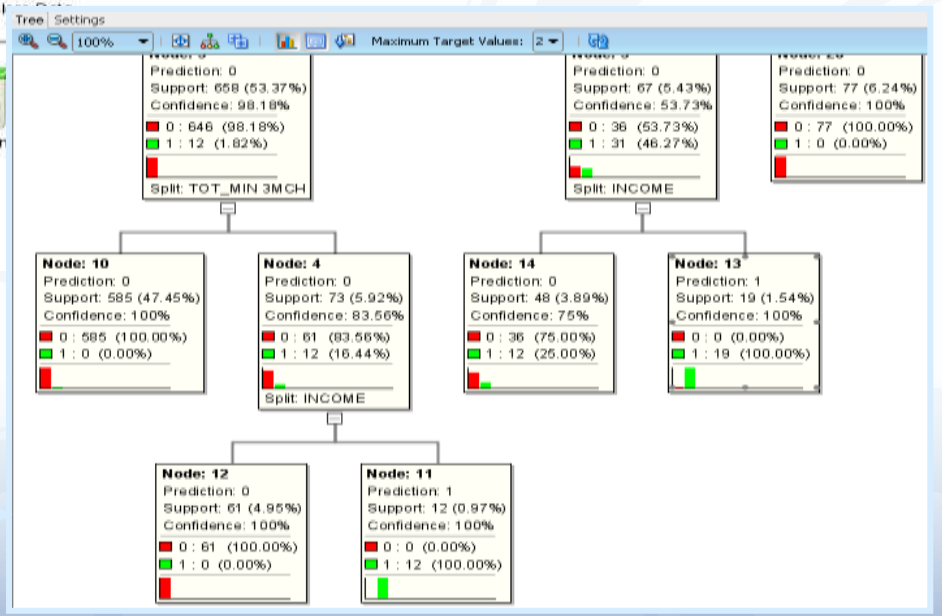
ФОРС

Модель кластеризации для сегментации клиентской базы



Построение классификационной модели для анализа лояльности клиентов

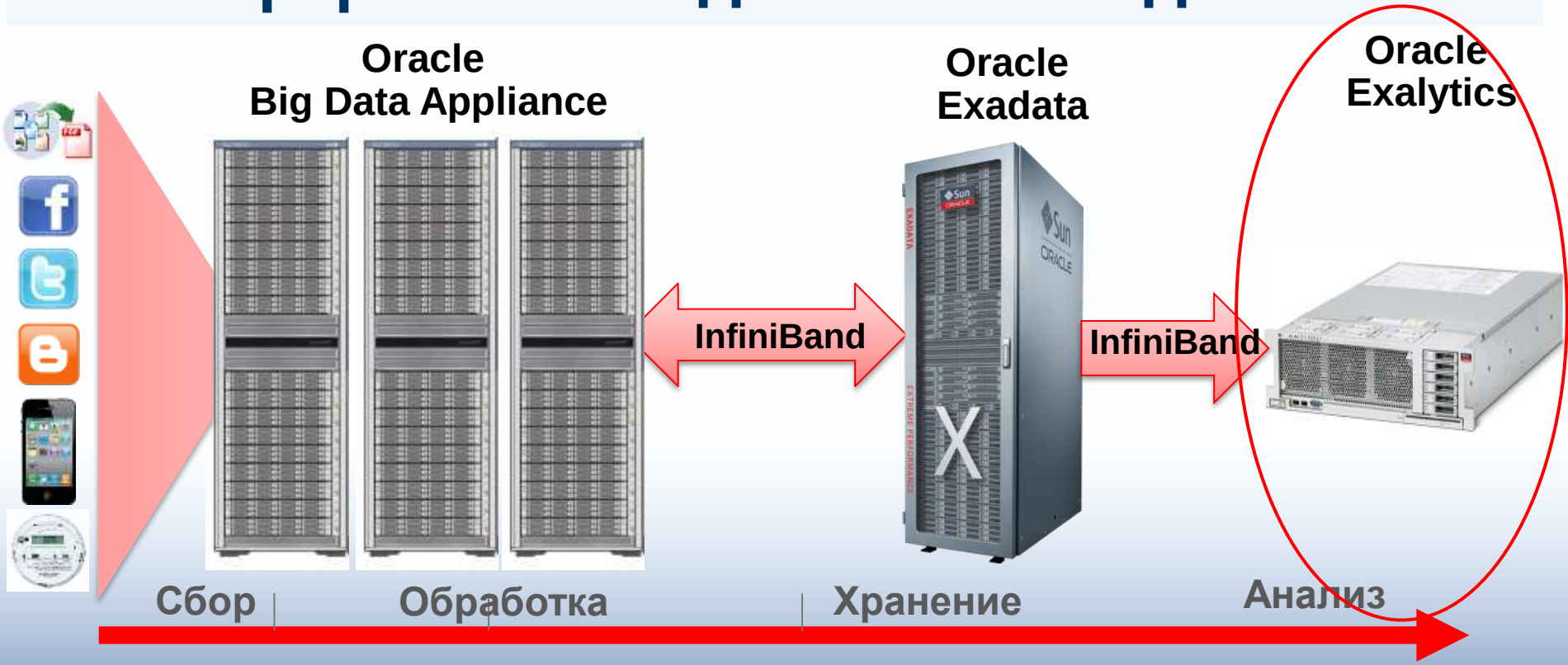




Rule | Surrogates | Target Values

```
IF   MIN_CH3M <= .5
AND  TOT_MIN 3MCH <= 70
AND  MIN_PREV_MO > 688.5
AND  INCOME <= 139500
THEN 1
```

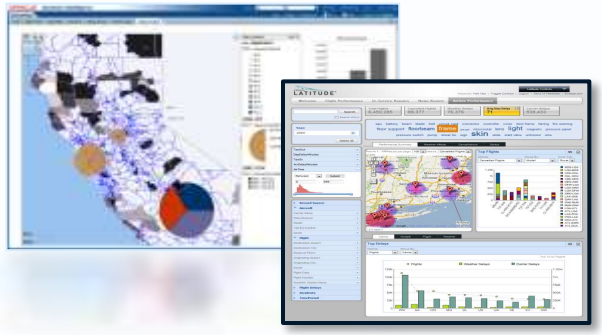
Платформа Oracle для больших данных



ФОРС

Oracle Exalytics In-Memory Machine

- Машина для анализа и визуализации
- In-Memory Analytics
- Аппаратное обеспечение
  - *2 TB RAM, 40 ядер CPU, 2,4 TB of PCI flash*
- Программное обеспечение
  - *Oracle Business Intelligence*
  - *TimesTen In-Memory Database*
  - *Oracle Essbase (OLAP сервер)*
  - ***Oracle Endeca***



ФОРС

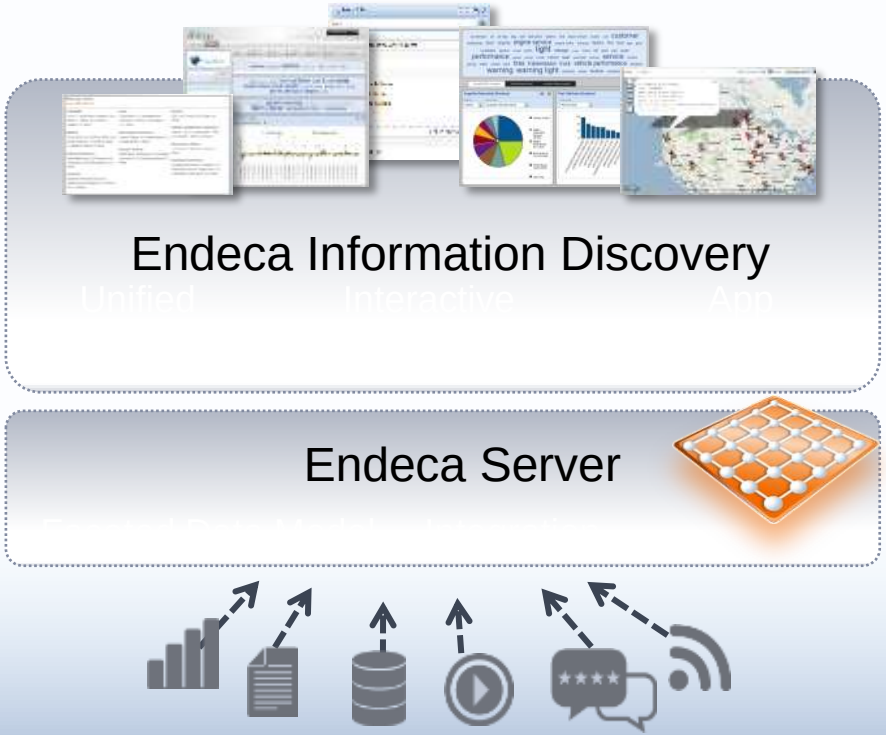
Oracle Endeca Information Discovery

- Совместный анализ структурированной и неструктурированной информации
- Интуитивное исследование
- Фасетный поиск
- Контекстный поиск, навигация, аналитика
- Динамические метаданные
- Неструктурированная информация и Text Enrichment (обогащение текста)
- In-Memory технологии

ФОРС

Oracle Endeca Information Discovery

- **Endeca Server**
  - Поисково-аналитическая база данных
- **Endeca Integrator**
  - Загрузка данных в Endeca Server
- **Endeca Information Discovery**
  - Быстрая компонентная разработка приложений для исследования данных

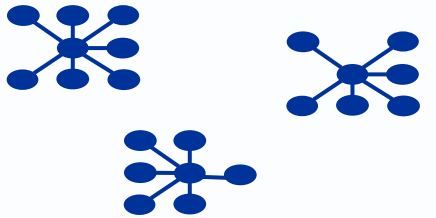


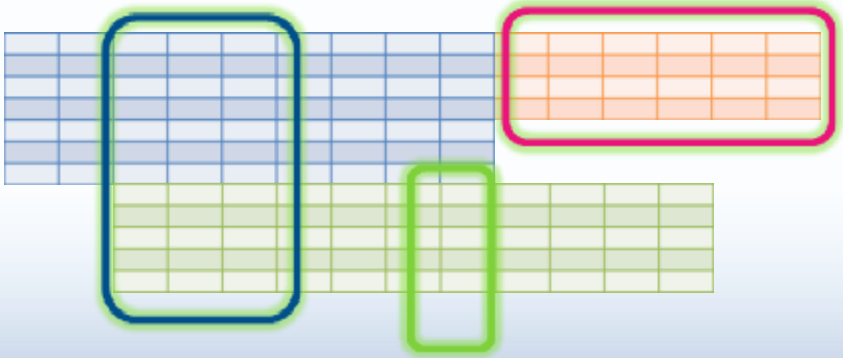
ФОРС

Endeca Server

Информационно-поисковая NoSQL база данных

- Фасетная модель данных
- Набор записей, каждая из которых имеет собственную «структуру»
  - Многозначные поля
  - Неструктурированные поля (тексты)
- Один из видов «Key Value» модели
- Модель состоит из
  - Записей
  - Атрибутов
- Каждая запись – это набор пар (атрибут, значение)
- Нет никакого разбиения на таблицы
- Нет понятия схемы данных



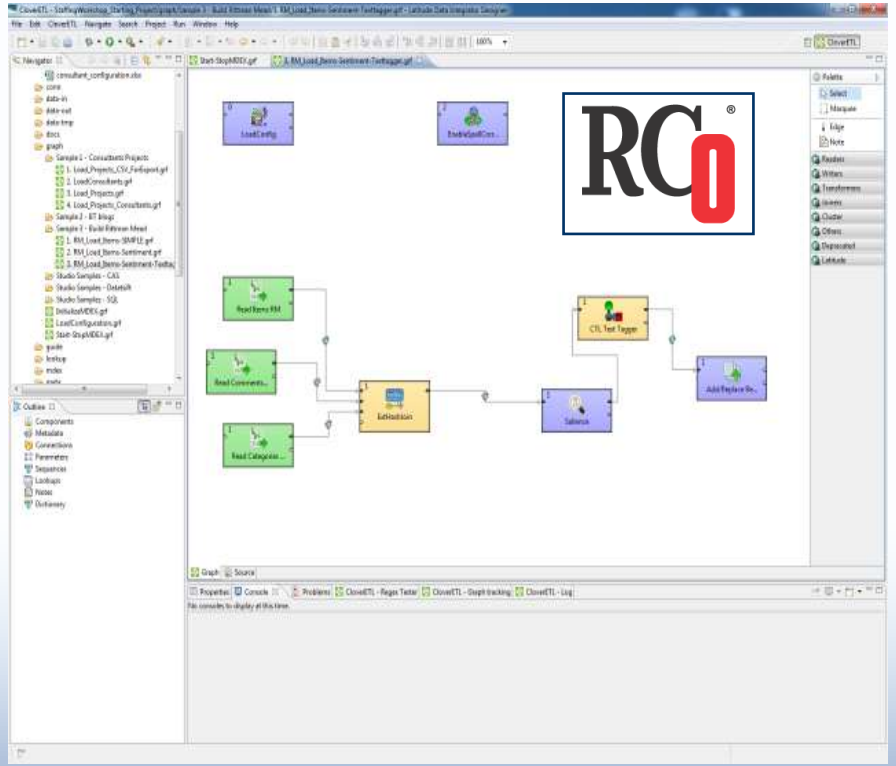


ФОРС

Endeca Integrator

Загрузка данных в Endeca Server

- ETL-среда для загрузки данных в базу данных Endeca Server
- При загрузке выполняются преобразования и извлечение из текстов фактов – упоминание персон, географических объектов, событий, определение тональности и др.
- Для семантической обработки русскоязычных текстов используются инструменты компании RCO



ФОРС

Endeca Studio

Интерфейс пользователя

- Быстрая компонентная разработка, возможность расширения
- Поиск + Фасетная навигация + Визуальный анализ
  - Поиск и выбор атрибутов в стиле вэб сайтов
  - Уточнение критерия поиска
  - Быстрое вычисление статистики
  - Текстовый поиск
- Интеграция с географическими картами
- Интерактивные исследования

ФОРС

Технологии Oracle для Больших Данных

Принятие решений

- Oracle Real-Time Decisions
- Endeca Information Discovery
- Oracle BI Foundation Suite

Поток

- Oracle Event Processing
- Apache Flume
- Oracle GoldenGate

Захват – Преобразование – Анализ

- Cloudera Hadoop
- Oracle NoSQL Database
- Oracle R Distribution

- Oracle Big Data Connectors
- Oracle Data Integrator

- Oracle Database
- Oracle Advanced Analytics
- Oracle Spatial & Graph

Эксперименты и проекты

Рынок Больших данных

- Один из самых быстрорастущих в мире
- Наиболее активные индустрии – банки, телеком и ритейл
- IDC:
  - *мировой рынок Big data будет расти в среднем на 31,7% в год и достигнет к 2016 году $23,8 млрд*
  - *ежегодно объемы хранимой информации вырастают на 40%*
- Gartner:
  - *рост объемов корпоративной информации составит 650% с 2010 по 2014 год и 85% ее будет неструктурированной*
  - *мировые расходы на big data в 2012 году - $28 млрд, а в 2013 году –$34 млрд*
  - *В 2013 году объем мирового рынка средств анализа и бизнес-аналитики вырос примерно на 8% и достиг 14,4 млрд долларов. На него практически не повлияли технологии анализа больших данных*

Большие данные в России

- Большой интерес к новому направлению, но низкий уровень востребованности
- Много экспериментов, но почти нет промышленных внедрений
- Наиболее активные индустрии
  - *Банки и финансовые организации*
  - *Телеком*
  - *Госсектор*
  - *Недвижимость*
  - *Ритейл*

ФОРС

Большие данные в российских банках

10 российских банков из топ-30 используют технологии Big Data
Cnews, январь 2014

- **Применяют технологии Big Data**
  - *Сбербанк*
  - *Газпромбанк*
  - *ВТБ 24*
  - *Альфа Банк*
  - *Райфайзенбанк*
  - *Промсвязьбанк*
  - *Ситибанк*
- **Планируют применять Big Data**
  - *ВТБ*
  - *МДМ Банк*
  - *Юникредит Банк*
  - *Россельхоз*
  - *Петрокомерц*

http://www.cnews.ru/top/2014/01/30/10_rossiyskih_bankov_iz_top30_ispolzuyut_tehnologii_big_data_558490

В финансовых организациях

- Анализ предпочтений и прогнозирование поведения клиентов на основе данных социальных сетей, блогов, документов, транзакций
  - *Увеличение кросс продаж и дополнительных продаж*
  - *Повышение уровня лояльности*
  - *Повышение потенциальной прибыльности клиентов*
- Предупреждение мошенничества
- Управление рисками невозврата кредита, мошенничества, рыночными, валютными, операционными, репутационными

Банк HSBC

- крупнейший в Европе по размеру рыночной капитализации
- Внедрена система по противодействию кредитному мошенничеству
- Результаты
  - ***эффективность этой службы повысилась в 3 раза***
  - ***точность выявления мошенничества — в 10 раз***
- В первые две недели эксплуатации семь специалистов службы безопасности HSBC выявили новые криминальные группы и схемы с общим потенциальным ущербом более $10 млн.

Тинькофф Кредитные Системы

- В банке внедрена система для анализа больших данных в режиме реального времени Причины миграции хранилища данных на технологию Big Data
  - планы по наращиванию клиентской базы
  - возросшие требования к обработке накопленной информации
- Проект выполнен за 6 месяцев
- Результаты: время решения аналитических задач сократилось минимум в 10 раз, а для некоторых – более чем в 100 раз.

Оценка проекта

«Ценность выполненного проекта заключается в развитии существующей в Банке культуры принятия решений на основе анализа информации. Умение превращать накопленные данные в знания давно является признаком конкурентоспособности Банка, а сами данные – стратегическим активом и потенциалом для будущего роста. В ближайшее время клиентами будут востребованы Банки, которые лучше понимают их поведение, привычки и максимально соответствуют им».

Вячеслав Цыганов, Вице-президент, CIO,
банк «Тинькофф Кредитные Системы»

Большие данные в российском ритейле

Cnews, 24 апреля 2014

- **Применяют технологии Big Data**
  - *Лента*
  - *Азбука вкуса*
  - *Глория Джинс*
  - *Юлмарт*
- **Планируют применять Big Data**
  - *X5 Retail Group*
  - *М. Видео*
  - *Эльдорадо*
  - *Белый ветер*

http://www.cnews.ru/news/top/index.shtml?2014/04/17/568717

ФОРС

Проекты в области больших данных

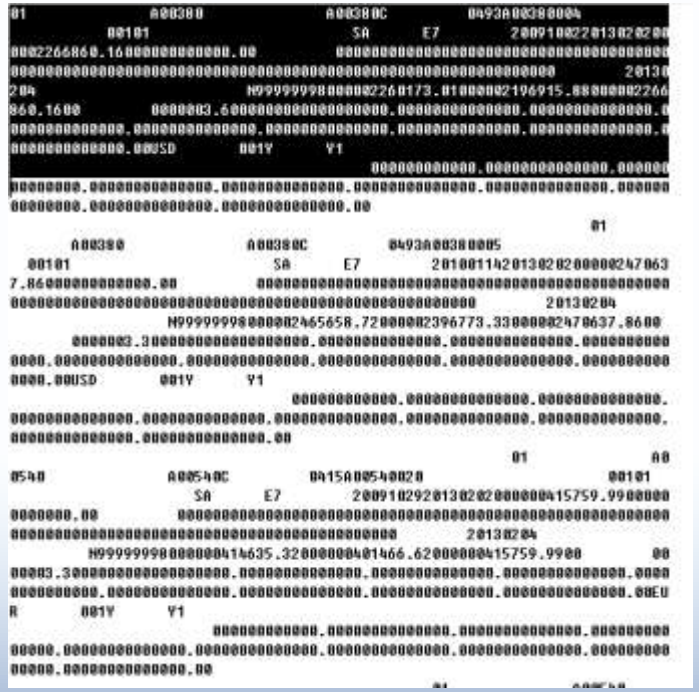
- Повышение производительности и затрат
  - *Многократное ускорение для известной задачи*
  - *Примеры – разбор входных битовых данных, предобработка для обогащения данных, выделение значимой информации в большом потоке*

- Новые данные, новая аналитика
  - *Новая задача или новые данные для известной задачи*
  - *Примеры – данные соцсетей , прогнозирование на основе интернет-запросов пользователей*

ФОРС

Проект «повышение производительности»
Пилот в российском банке

- Данные:
  - *файлы (без переноса строк) размером в несколько сотен Мб*
  - *Суммарный объем десятки Тб*
  - *Мета описание меняется раз в месяц*
- Необходимо
  - *Хранить все в течении нескольких лет*
  - *Загружать часть полей в Oracle Database (~ 50 из 1000)*
  - *Список полей для загрузки постоянно меняется*



ФОРС

Типовой проект для «новой аналитики»

- Совместный анализ текстов и «реляционных»данных
- Обработка неструктурированных текстов
  - *Интернет-источники*
  - *Внутренние архивы*

- Цели -- расширение состава данных для клиентской аналитики, смысловой поиск в неструктурированном контенте, поддержка аналитических исследований

Недвижимость и большие данные

- Индустрия недвижимости входит в top5 для Big Data
- Большой объем неструктурированных данных
- Многочисленные источники вэб-информации: уполномоченные госструктуры, открытые источники
- Постоянное непредсказуемое изменение в источниках
- Интенсивное использование методов статистики и предиктивной аналитики

ФОРС

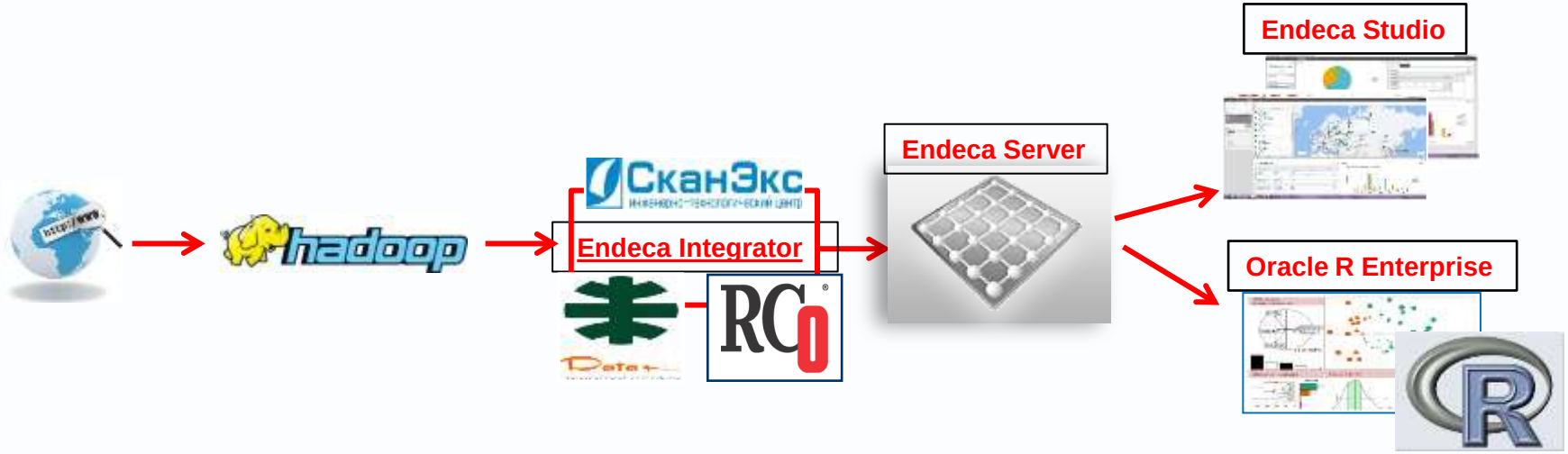
Пилотный проект по оценке недвижимости

- Проект выполнялся компанией Форс для Российского Общества оценщиков & РОСЭКО
- Функциональные возможности системы
  - *Сбор данных из интернет источников*
  - *Стандартизация и обогащение данных*
  - *Поддержка статистических исследований (регрессия, карты Кохонена, кластеризация)*
  - *Построение предиктивных моделей оценки и подбора аналогов*
- Реализация на основе технологий Oracle

РОО
РОССИЙСКОЕ ОБЩЕСТВО ОЦЕНЩИКОВ

ФОРС

Проект «Big Data для недвижимости»



Сбор информации из интернета обработка и хранение средствами Hadoop

Семантическая обработка текстов, геопозиционирование, загрузка данных

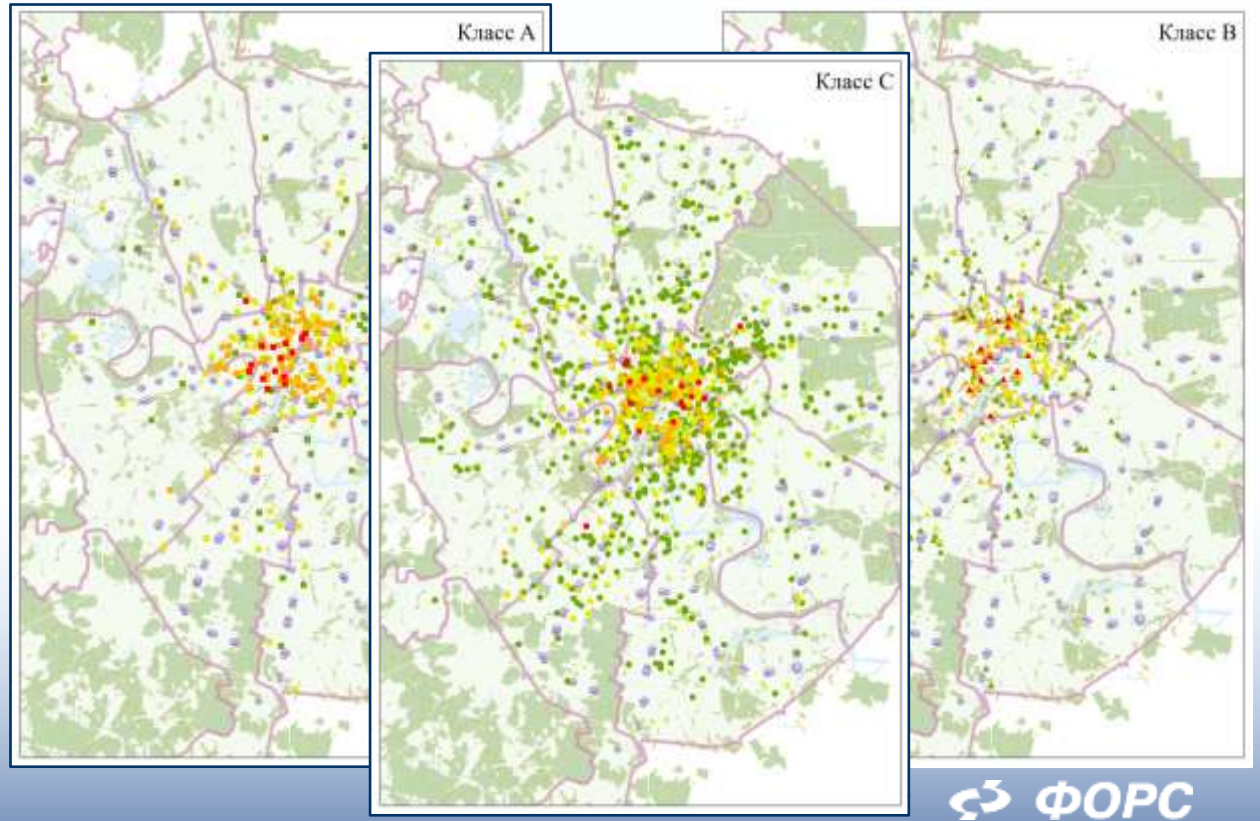
Хранение данных в информационно-поисковом сервере Endeca Server

Рабочие места пользователей: аналитиков и конечных пользователей

ФОРС

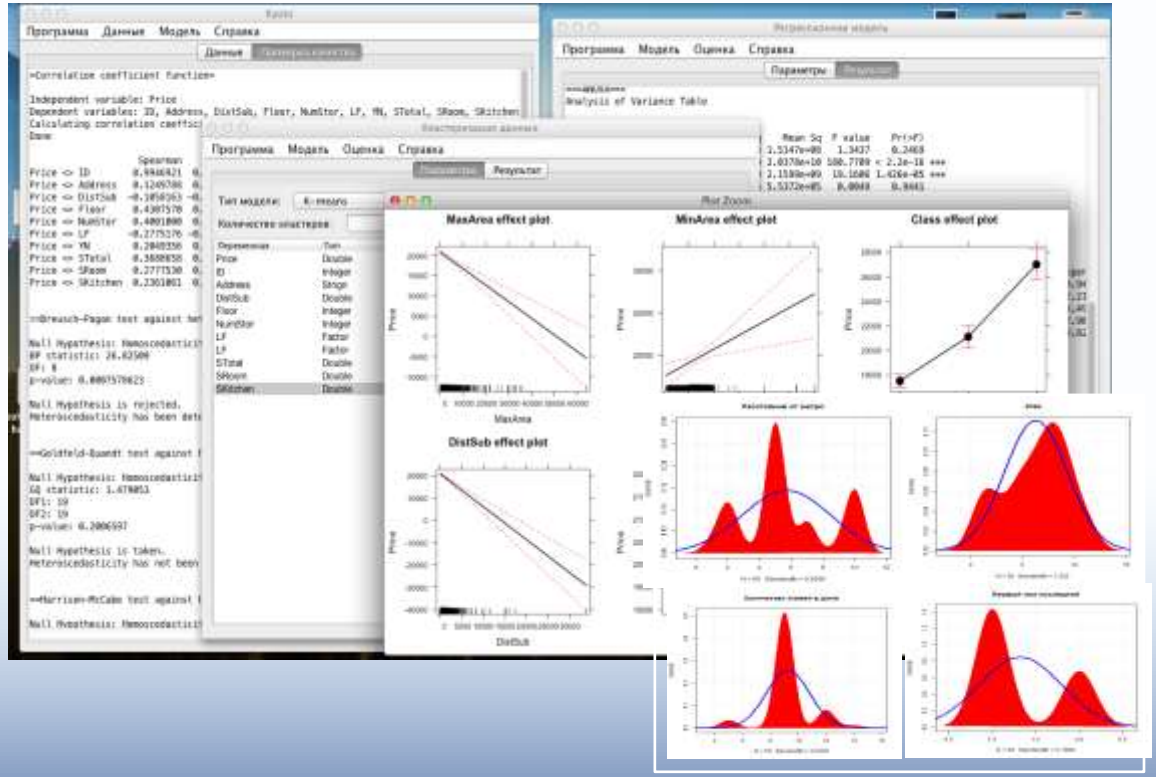
Обработка данных об адресах

- Лингвистическая обработка (RCO)
- Геопозиционирование (ArcGIS, СканЭкс)
- Получение дополнительных факторов (расстояние от центра, транспортная доступность)



ФОРС

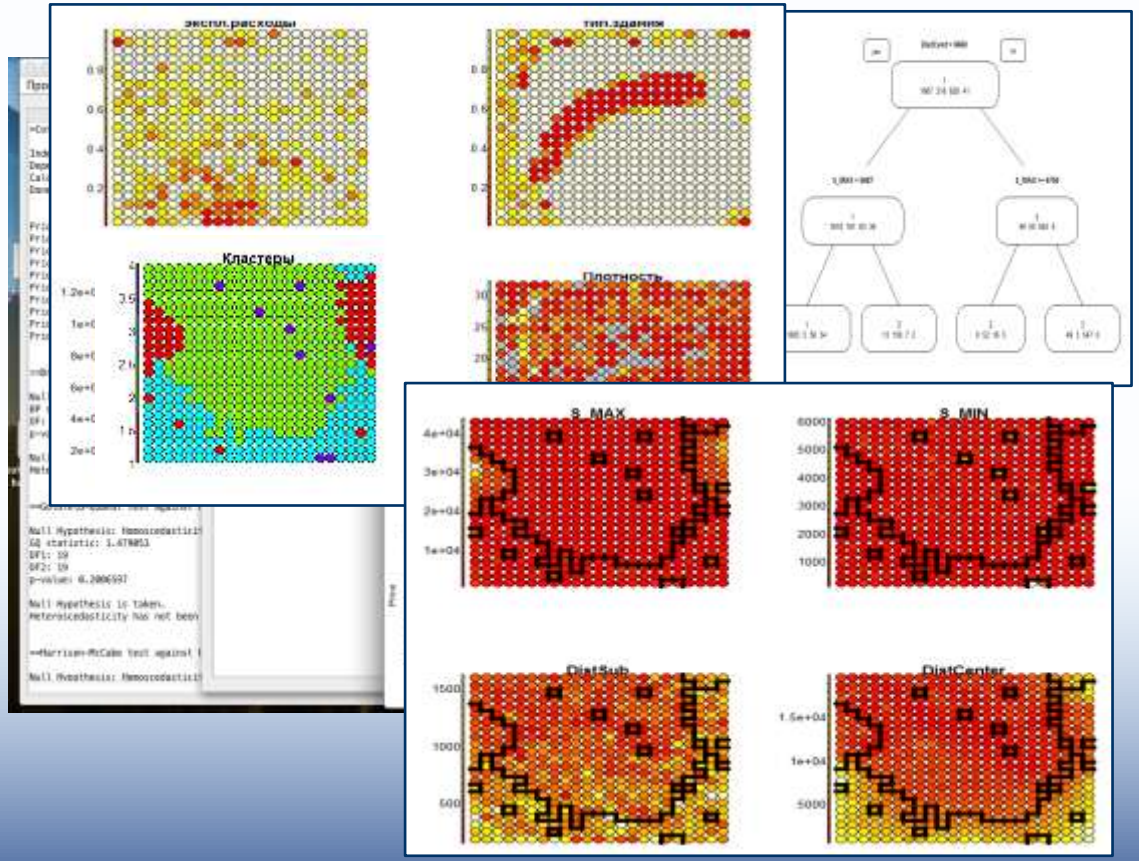
Статистическая обработка



- Оценка значимости атрибутов (корреляционный анализ)
- Гетероскедастичность (неоднородность наблюдений)
- Мультиколлинеарность (взаимозависимости между факторами)
- Чувствительность
- Тест на нормальность

ФОРС

Построение моделей

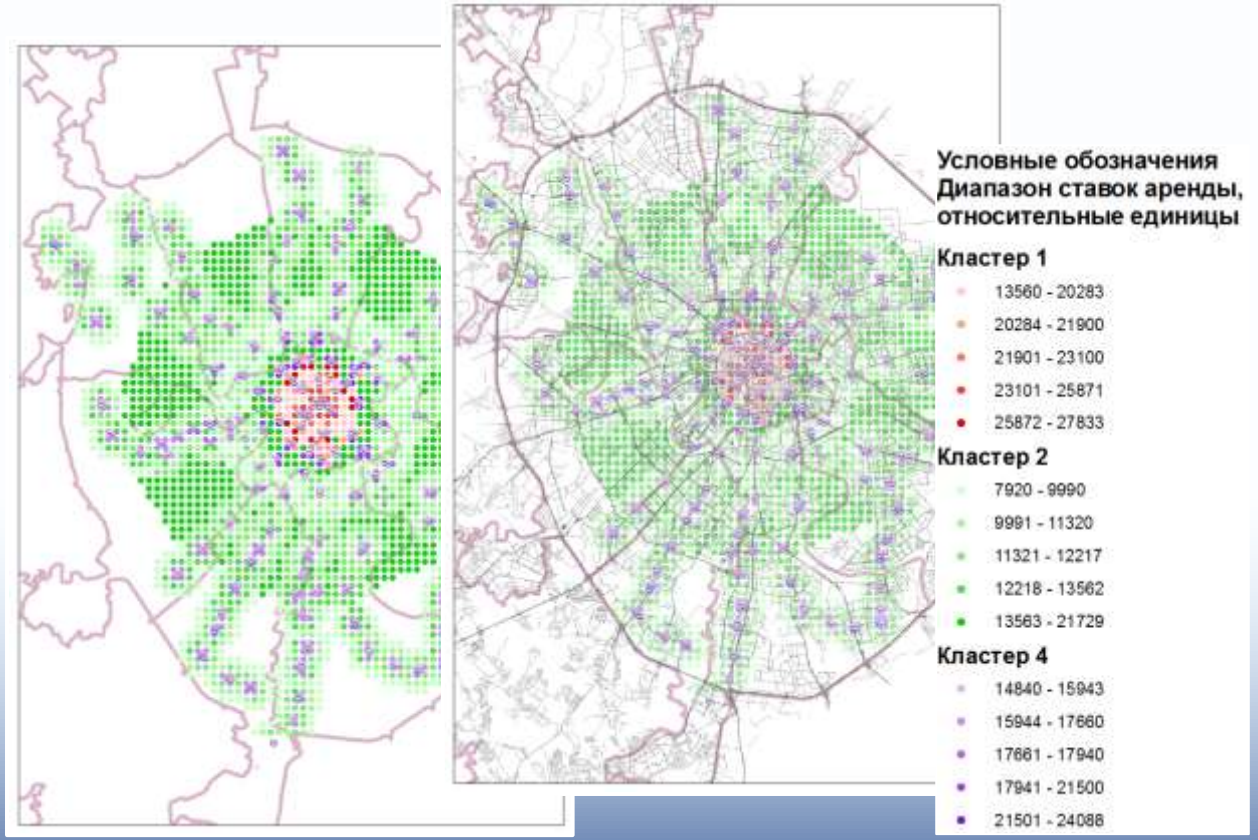


- Построение самоорганизующихся карт Кохонена
- Кластеризация
- Формирование правил

ФОРС

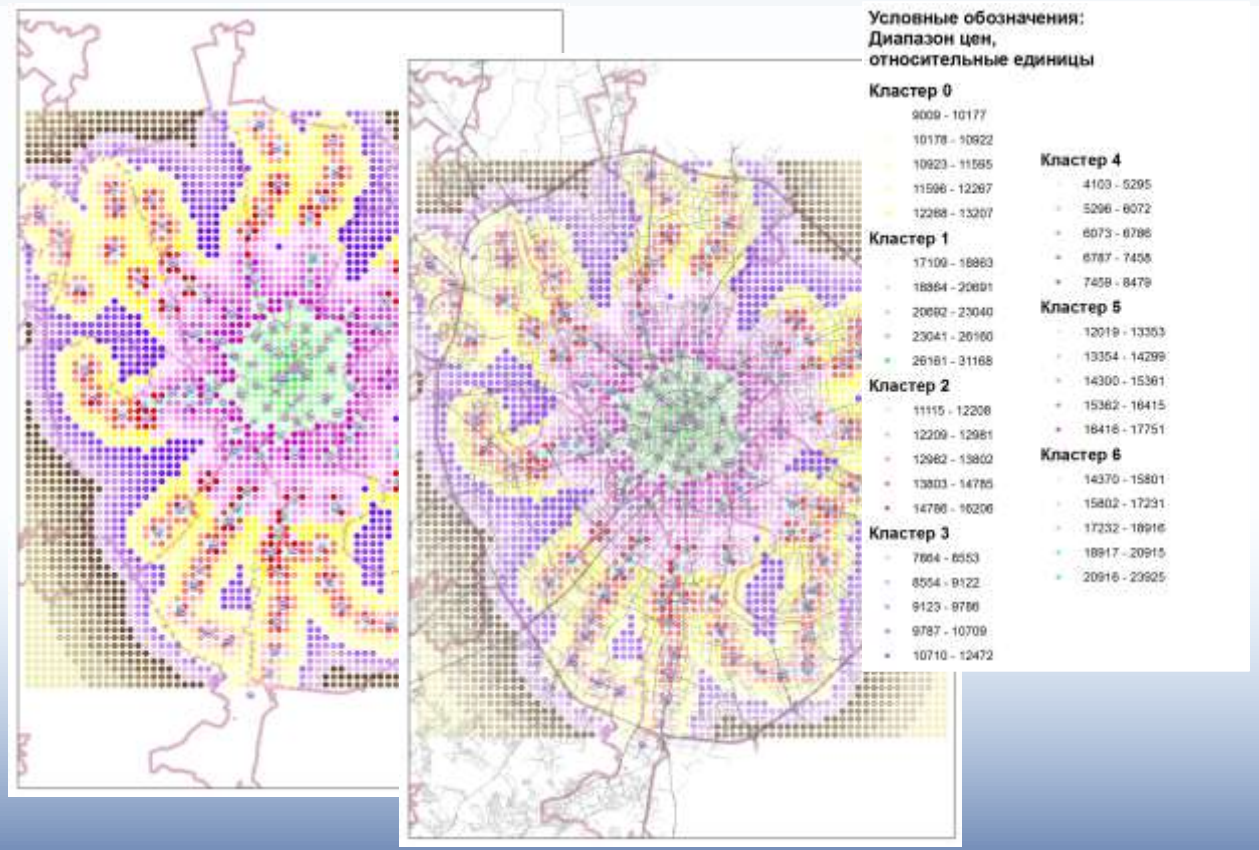
Первичное ценовое зонирование

- На основе карт Кохонена и кластеризации
- Геотрактовка
- Сопоставление полученных кластеров и с географической картов



Ценовое зонирование на основе регрессионного анализа

- Построение регрессии внутри кластера
- Применение регрессии для восстановления данных
- Окончательное ценовое зонирование



Об использовании больших данных

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Dan Ariely

Спасибо за внимание!