Oracle для анализа Больших Данных

Новые решения и практика внедрения

Ольга Горчинская
Oracle EE&CIS

ORACLE

План

- Oracle Exalytics -- аналитика в оперативной памяти
- Oracle R Enterprise – статистический анализ и визуализация для Больших Данных
- Endeca – платформа для систем исследования неструктурированной информации

План

- **Oracle Exalytics -- аналитика в оперативной памяти**
- Oracle R Enterprise – статистический анализ и визуализация для Больших Данных
- Endeca – платформа для систем исследования неструктурированной информации

Oracle Exalytics Business Intelligence Machine

- Программно-аппаратный комплекс для бизнес-анализа
- Экстремальная производительность
- Неограниченные возможности визуализации и анализа

Oracle Exa* - решения

ORACLE
EXADATA

Хранилища данных и консолидация баз данных

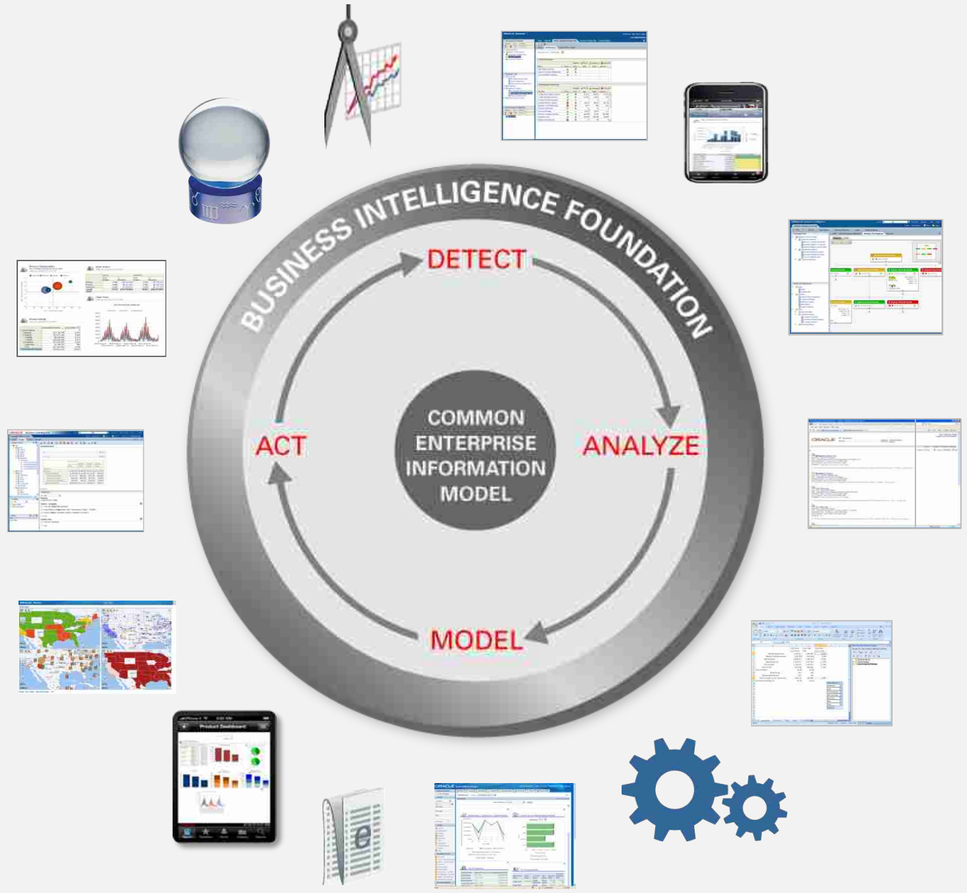
ORACLE
EXALOGIC

Облачные вычисления и консолидация приложений

ORACLE
EXALYTICS

Бизнес-анализ и EPM приложения

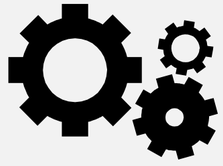
Oracle Exalytics – основные компоненты



Oracle Business Intelligence Suite – специальная редакция для Exalytics

TimesTen for Exalytics

Memory Optimized Essbase



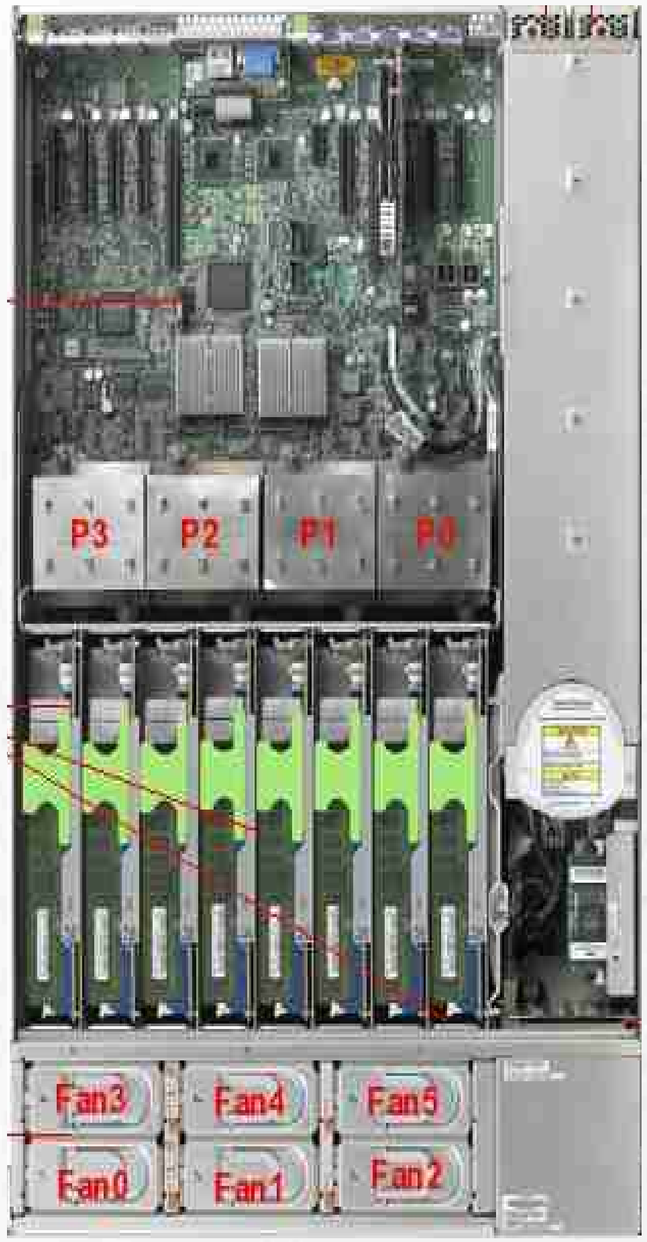
Adaptive In-Memory Tools

Программное обеспечение для *In-Memory* аналитики

1 TB RAM
40 Processing Cores
High Speed Networking

Аппаратный комплекс для In-Memory аналитики

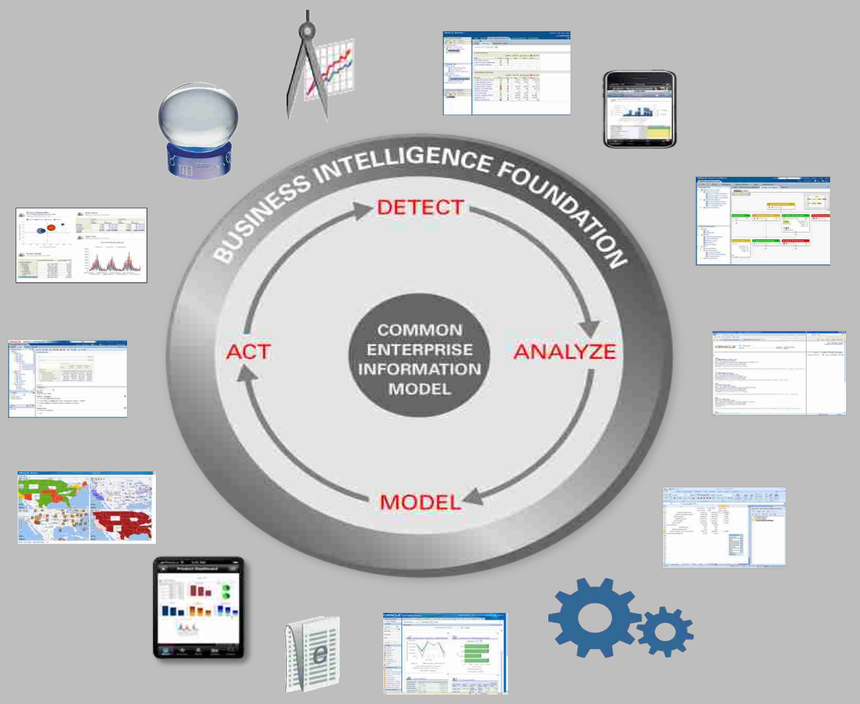
Аппаратное обеспечение



- **Оперативная память**
  1 TB RAM, 1033 MHz
- **Процессоры**
  4 Intel® Xeon® E7-4870, 40 cores
- **Сетевые интерфейсы**
  40 Gbps InfiniBand – 2 ports
  10 Gbps Ethernet – 2 ports
  1 Gbps Ethernet – 4 ports
- **Дисковая память**
  3.6 TB HDD Capacity

Программное обеспечение

- Oracle Business Intelligence
- Oracle Essbase
- Oracle TimesTen for Exalytics
- Адаптивные in-memory акселераторы



Oracle Business Intelligence

Essbase

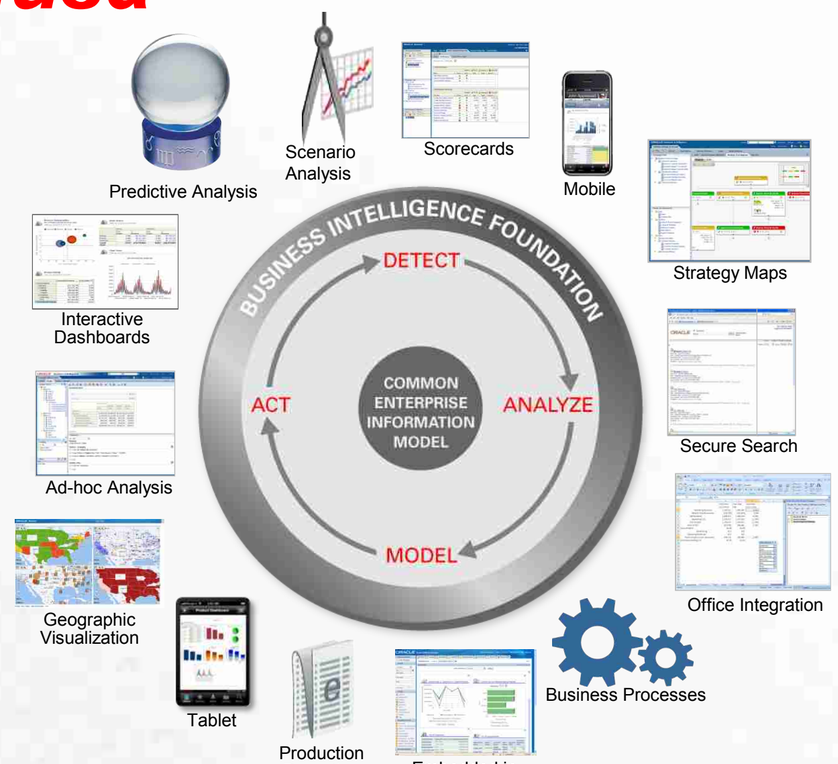
TimesTen for Analytics

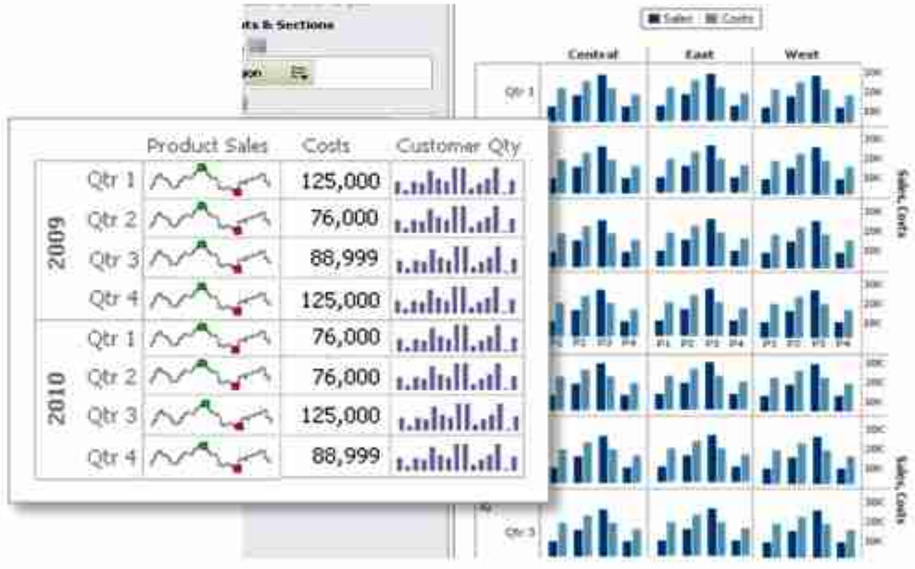
Адаптивные in-memory акселераторы

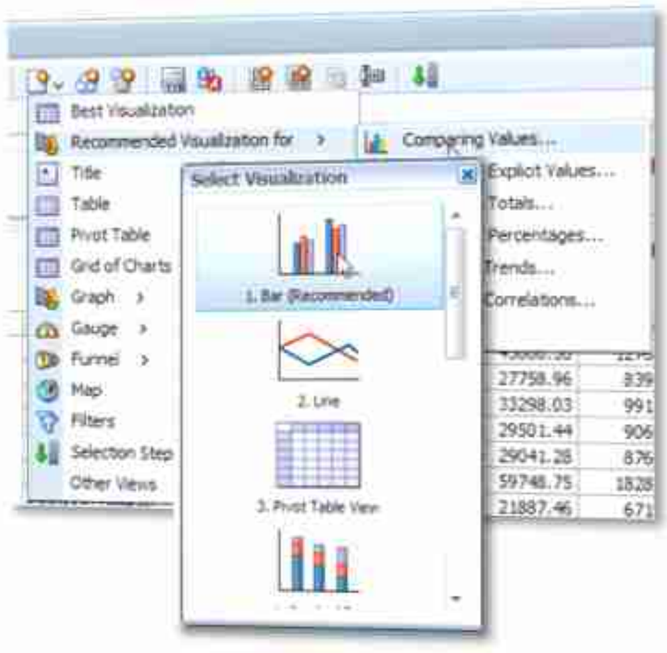
Oracle Business Intelligence

Единая платформа бизнес-анализа

- Единые метаданные для всех видов и стилей анализа
- Высокая степень интерактивности
- Инновационная визуализация
- Встроенные сценарии
- Активная аналитика







Oracle Essbase

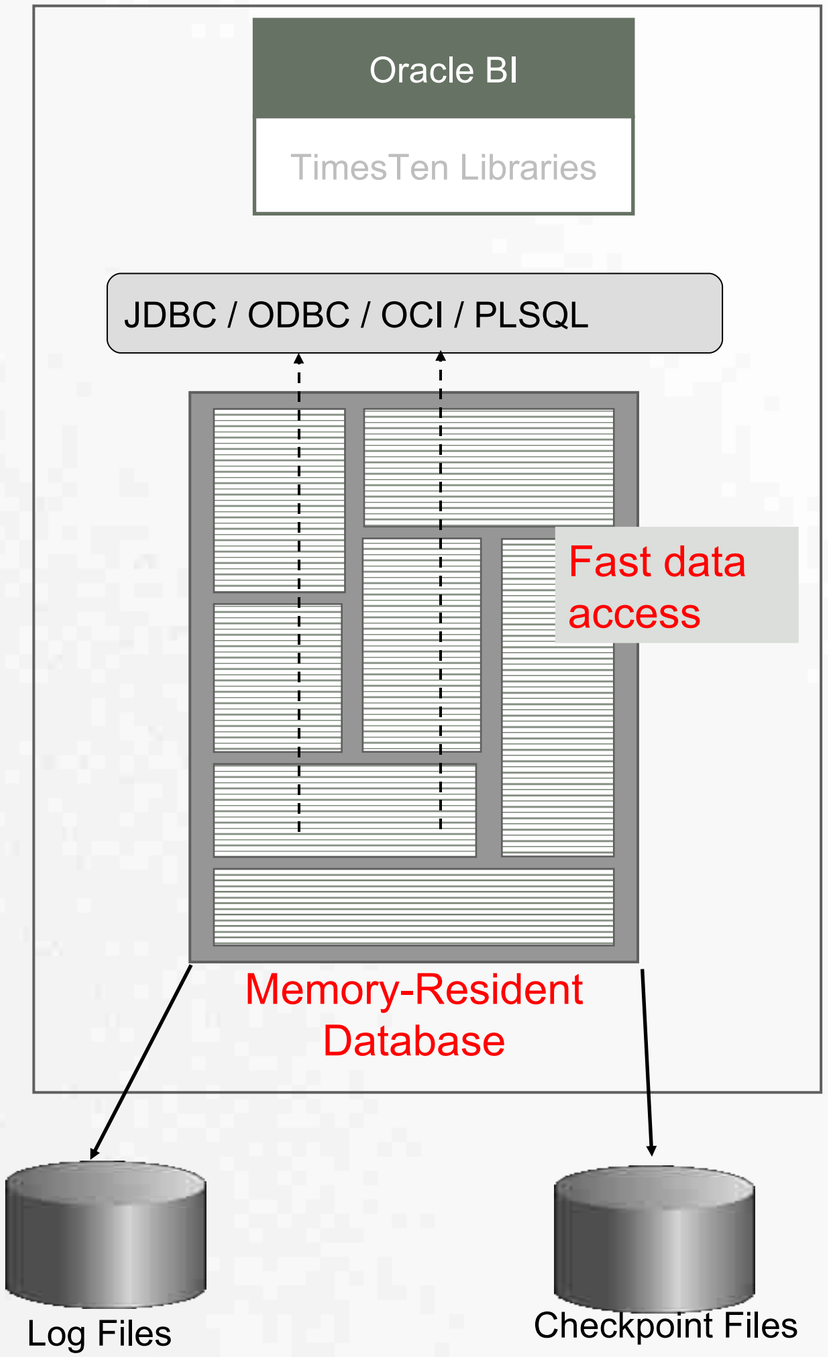
Многомерный сервер

- Универсальный OLAP-сервер для хранения, обработки и представления информации
- Высокая производительность
- Моделирование сложной аналитики
- Основа систем бюджетирования и планирования (Hyperion Planning)

Times Ten for Exalytics

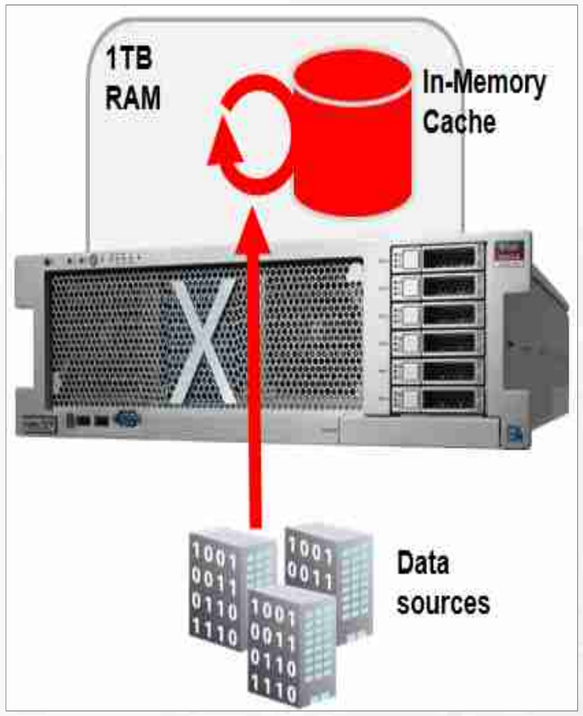
На основе Oracle TimesTen In-Memory Database

- 100% In-memory RDBMS
  - База данных в оперативной памяти
- Высокая производительность
  - Снижение времени отклика
  - Высокая пропускная способность
- Сохранение в дисковой памяти
  - Транзакции и контрольные точки копируются для постоянного хранения
- Колоночная компрессия
  - Сжатие от 5 до 10 раз
  - Аналитические алгоритмы работают непосредственно с компрессированными данными
- Аналитические функции
  - Эффективное выполнение аналитических функций
  - Разгрузка BI-сервера



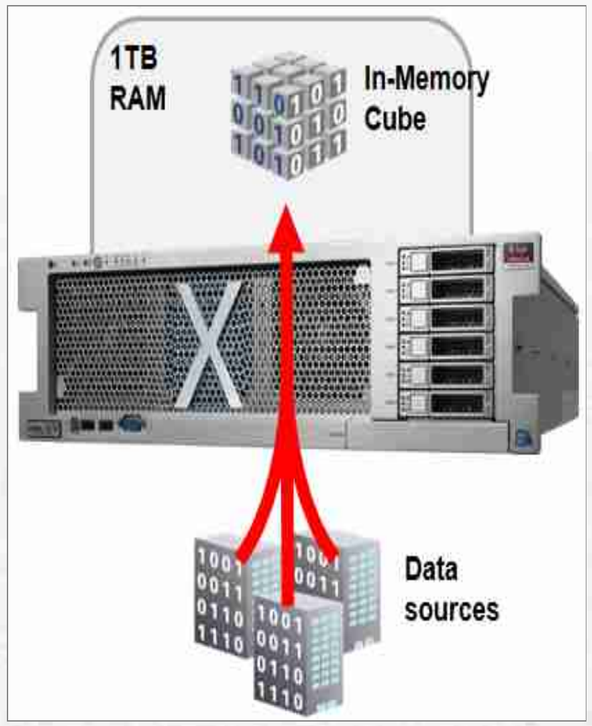
Адаптивные in-memory акселераторы

Кэширование данных в оперативной памяти



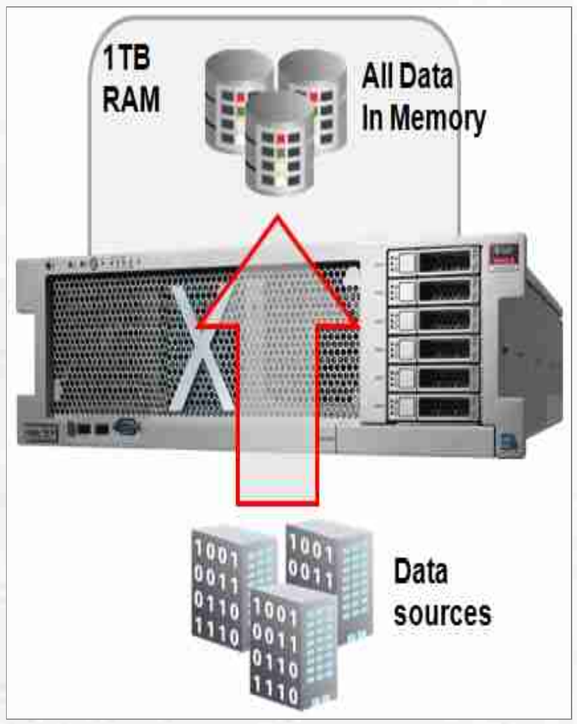
In-Memory Intelligent Result Cache

Хранение в оперативной памяти результатов выполненных запросов



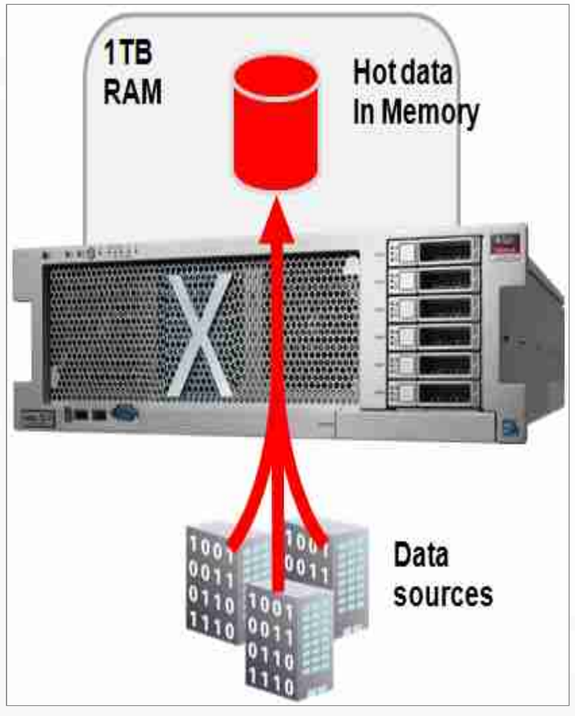
In-Memory Essbase Cubes

Кэширование в оперативной памяти Essbase-кубов



In-Memory Data Warehouse

Хранение в оперативной памяти всего хранилища данных



In-Memory Adaptive Data Marts

Кэширование в TimesTen
Автоматическое формирование витрины на основе статистики запросов

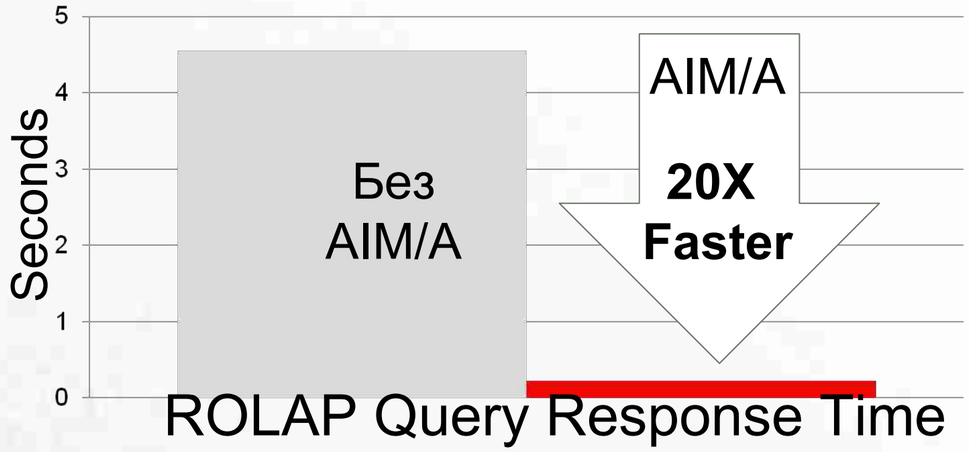
Что дают in-memory акселераторы

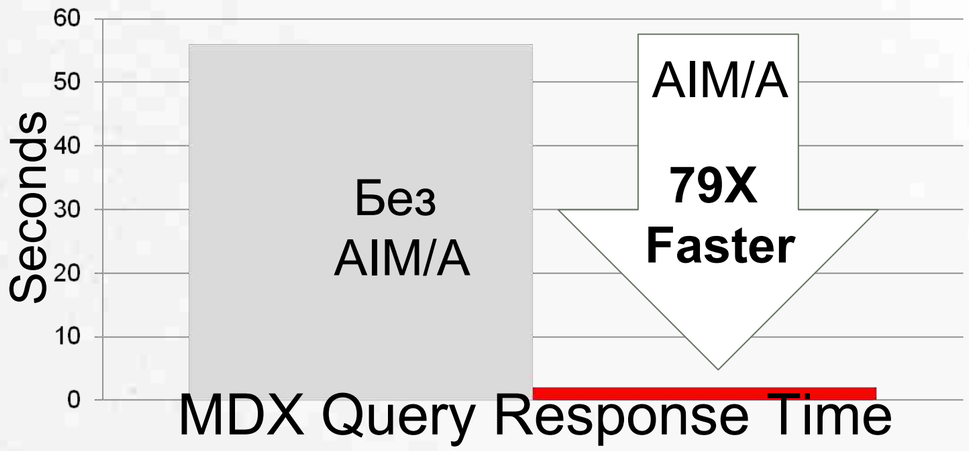
Экстремальная производительность

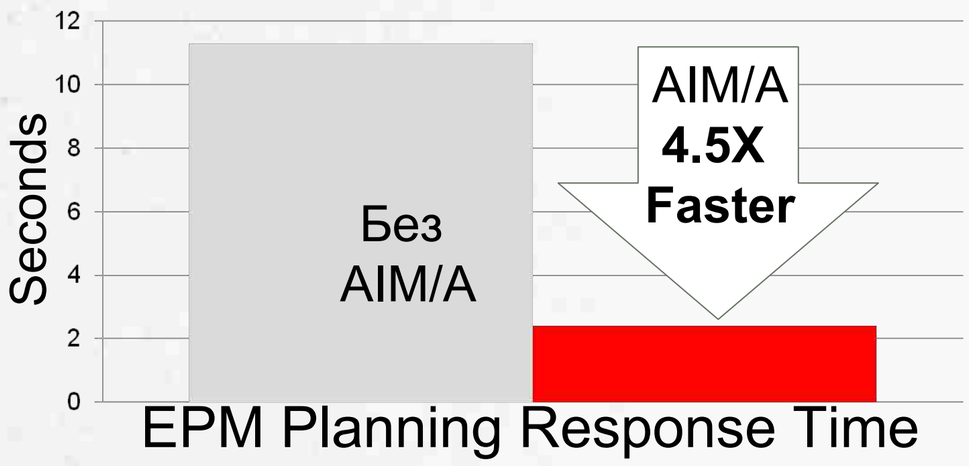
- **In-Memory Analytics: ROLAP**
  - **20X** снижение времени отклика
  - **50,000** пользователей на одном комплексе

- **In-Memory Analytics: MOLAP**
  - **79X** снижение времени отклика при чтении
  - **16X** снижение времени отклика при записи

- **In-Memory Analytics: EPM Planning**
  - **4.5X** снижение времени отклика
  - **10,000** пользователей на одном комплексе







Бизнес-анализ и источники данных

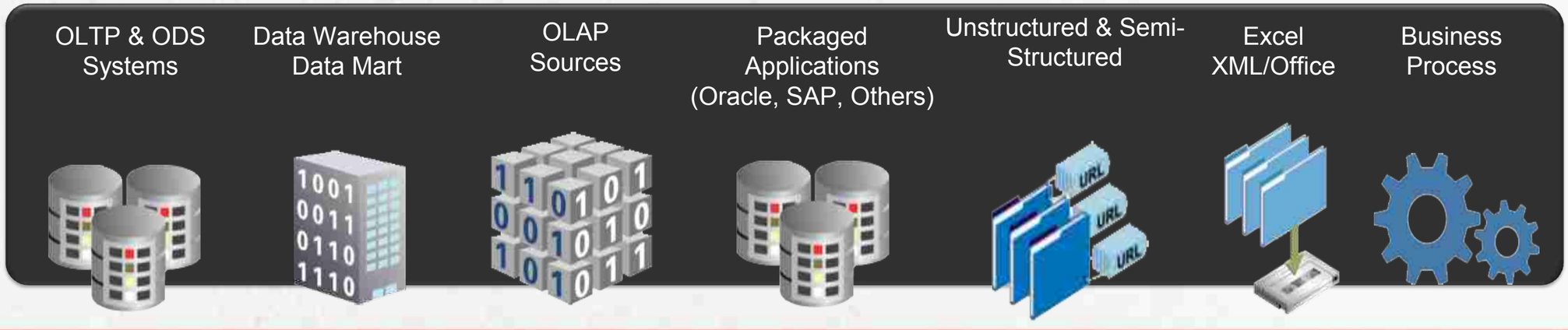
Работает с любыми источниками, оптимизирована для Exadata

Бизнес-анализ «со скоростью мысли»

InfiniBand

Экстремальная производительность хранилища данных

Любые источники данных



Результаты практического использования

Первые заказчики

Nykredit

- Крупнейший поставщик ипотечных услуг Дании
- Цель -- добиться экстремальной производительности для агрегированных и транзакционных данных
- **35x to 70x**, использовались Exadata + Exalytics

Polk

- Поставщик маркетинговой аналитики для автомобильной индустрии
- Цель – высокая интерактивность панелей и визуализаций для глобальной компании
- от **10x** и до **100x** в отдельных случаях

Key Energy Services

- Large oilfield services company with about ~860 rigs deployed around the world
- Цель – внедрение BI Apps на предприятии
- **5x** – сокращение времени внедрения; **50x** повышение скорости выполнения отчетов

SAVVIS

- Услуги в области облачных инфраструктур
- Цель – высокие требования к визуализации для разнообразных наборов данных
- Интерактивность на уровне **долей секунд**, в текущей системе было 30 сек

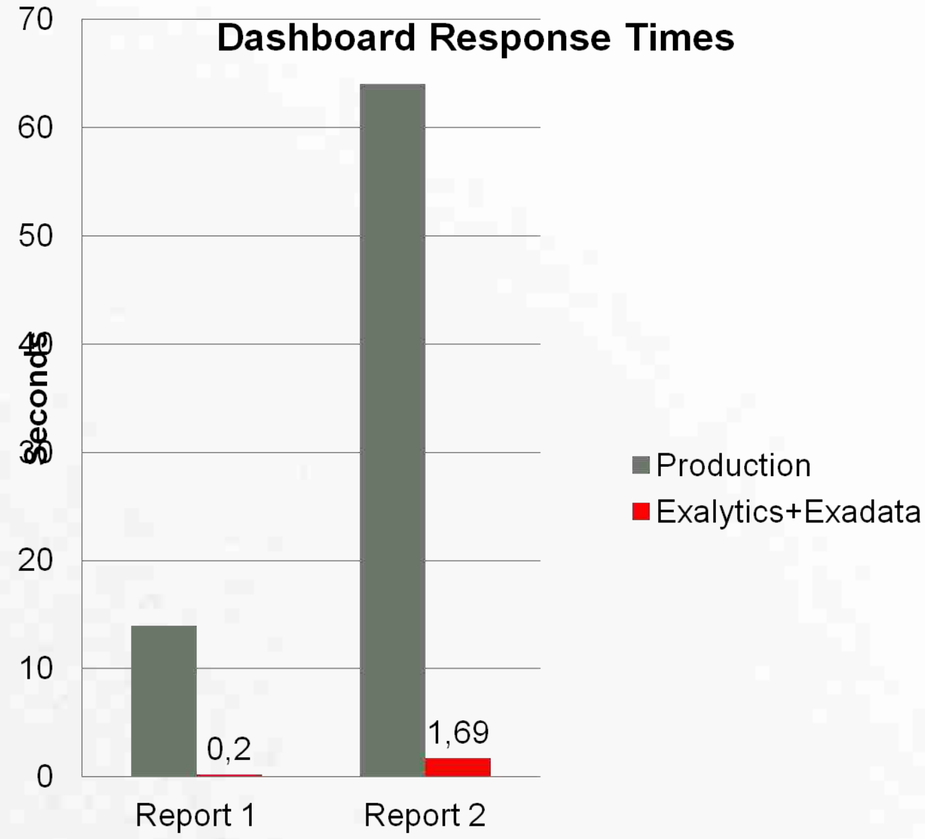
A Global CPG Company

- Global consumer pre-packaged foods company
- Цель – сократить цикл бюджетирования для 2000+ пользователей
- **6x** – ускорение цикла; **4 часа** по сравнению с **24 часами** для существующей системы

Первые проекты и результаты

Nykredit

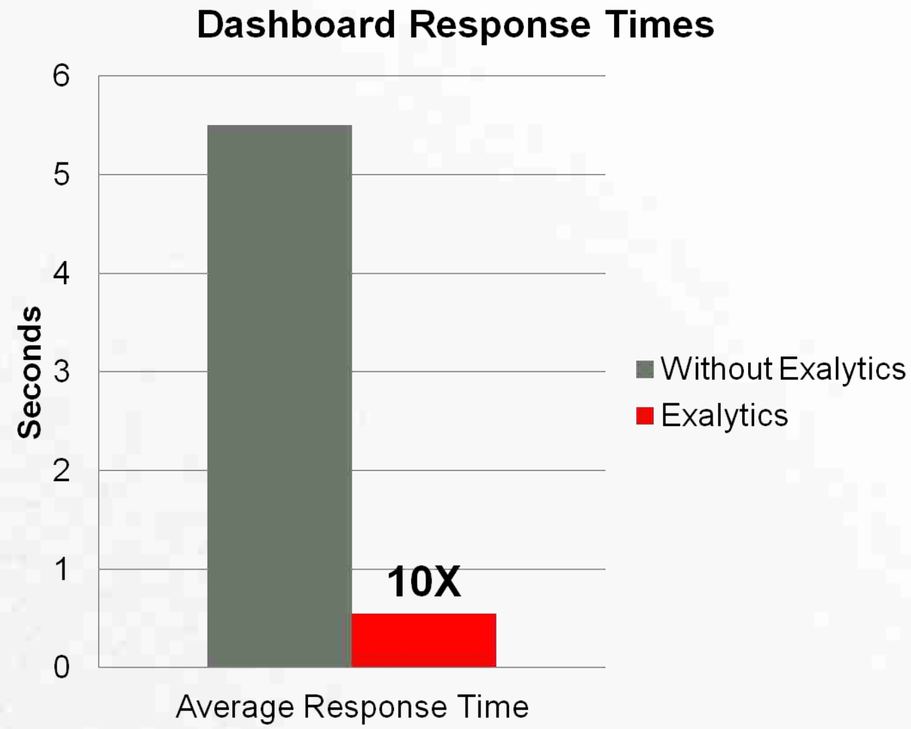
- Одна из крупнейших финансовых корпораций Дании, поставщик ипотечных услуг, 4 000 сотрудников
- BI система:
  - 1 700 пользователей
  - BI Aplications с быстро растущими объемами данных (до 50 Тб)
  - Необходима высокая производительность как на агрегированных, так и для транзакционных данных
- **Exadata + Exalytics:** От **35** до **70** раз быстрее



"Используя Exalytics и ее технологии аналитической обработки данных в оперативной памяти (in-memory), мы видели сокращение времени отклика аналитической системы в 35 - 70 раз по сравнению с тем, что есть сейчас!"

Первые проекты и результаты

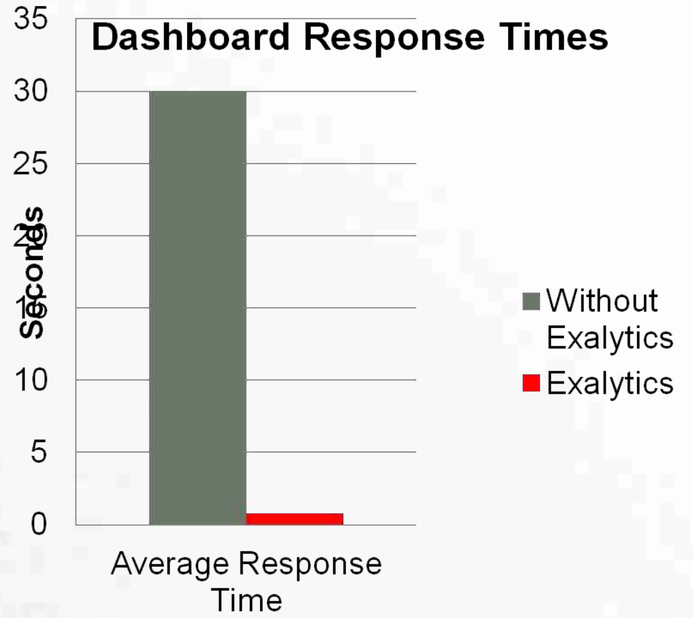
- Поставщик маркетинговой аналитики и решений для автомобильной индустрии, 500M vehicles, 195M people , 17M businesses
- BI система
  - 6 000 пользователей, ориентация на сложный анализ и исследования
  - Быстрое рост числа пользователей
  - Высокие требования к интерактивности , скорости, визуализации (прямое влияние на конкурентноспособость)
- **Exalytics:** В среднем ускорение более чем в **10** раз и в отдельных случаях до 100 раз



"Аналитическая машина Exalytics продемонстрировала снижение времени отклика информационных панелей в среднем в 10 раз, а в отдельных случаях более чем в 100 раз по сравнению системой, которую мы используем сейчас ."

Первые проекты и результаты

- Крупный поставщик услуг в области облачных инфраструктур, 45 000 сотрудников
- Текущая корпоративкая BI система
  - Более 1500 BI- пользователей
  - Проблемы -- недостаточный уровень интерактивности корпоративной BI системы, выгрузка данных и использование desktop инструментов, отклик --до 30 сек
- **Exalytics**: согласованная интерактивность на уровне долей секунд



"Exalytics продемонстрировала интерактивность «со скоростью мысли», которая до этого была возможна только для инструментов desktop-уровня. Теперь мы готовы отказаться от настольных BI и перейти с использованию Exalytics как только эта машина будет готова."

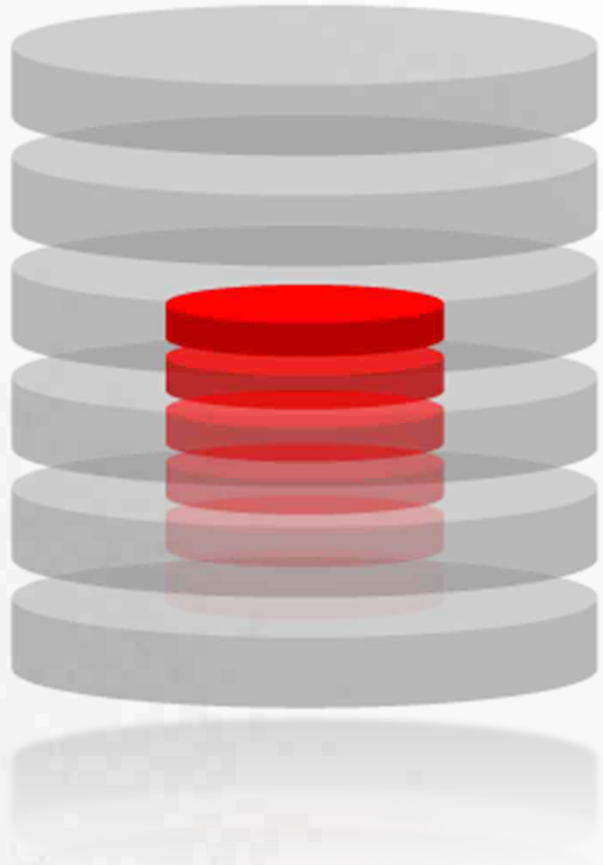
План

- Oracle Exalytics -- аналитика в оперативной памяти
- **Oracle R Enterprise – статистический анализ и визуализация для Больших Данных**
- Endeca – платформа для систем исследования неструктурированной информации

Oracle Advanced Analytics

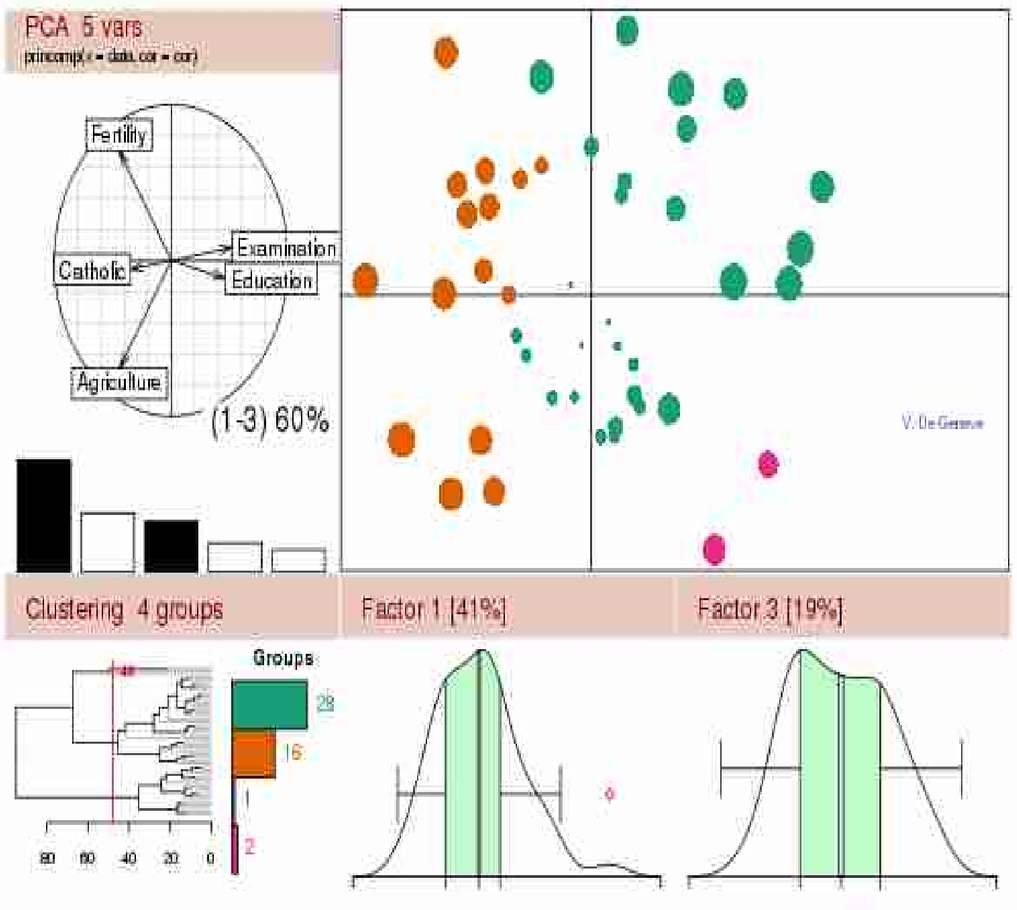
Встроенная аналитика для Больших Данных

- Новая опция для СУБД Oracle Database 11g
- Среда статистических исследований Oracle R Enterprise + Oracle Data Mining
- Углубленная аналитика в базе данных



ORACLE R Enterprise
Open Source

Проект R для статистических исследований



- Язык для статистических исследований и работы с графикой (Росс Айхэк, Роберт Джентельмен, Оклендский ун-т, 1997)

- **Open source проект, R Foundation**

- Широкий спектр различных функций (временные ряды, прогнозирование, классификация, кластеризация и др)

- Важное отличительное преимущество – простые средства построения самых сложных графиков и диаграмм

- Возможность расширения, технология разработки дополнительных пакетов участниками проекта

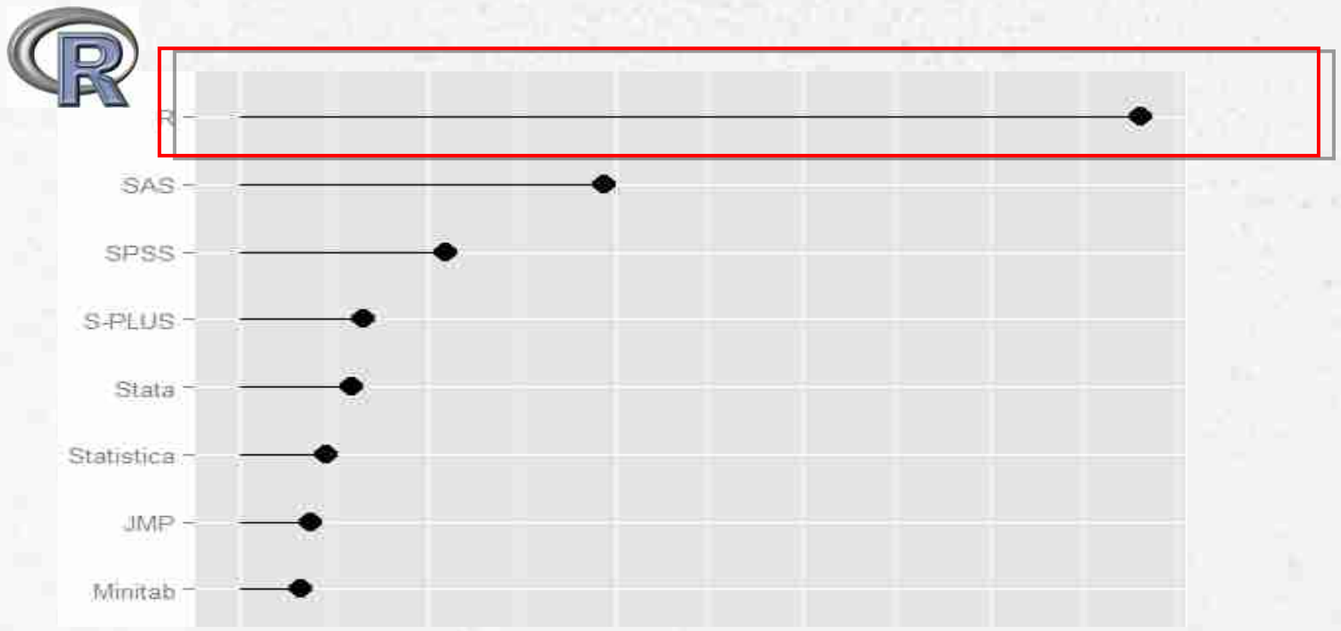
Open Source

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

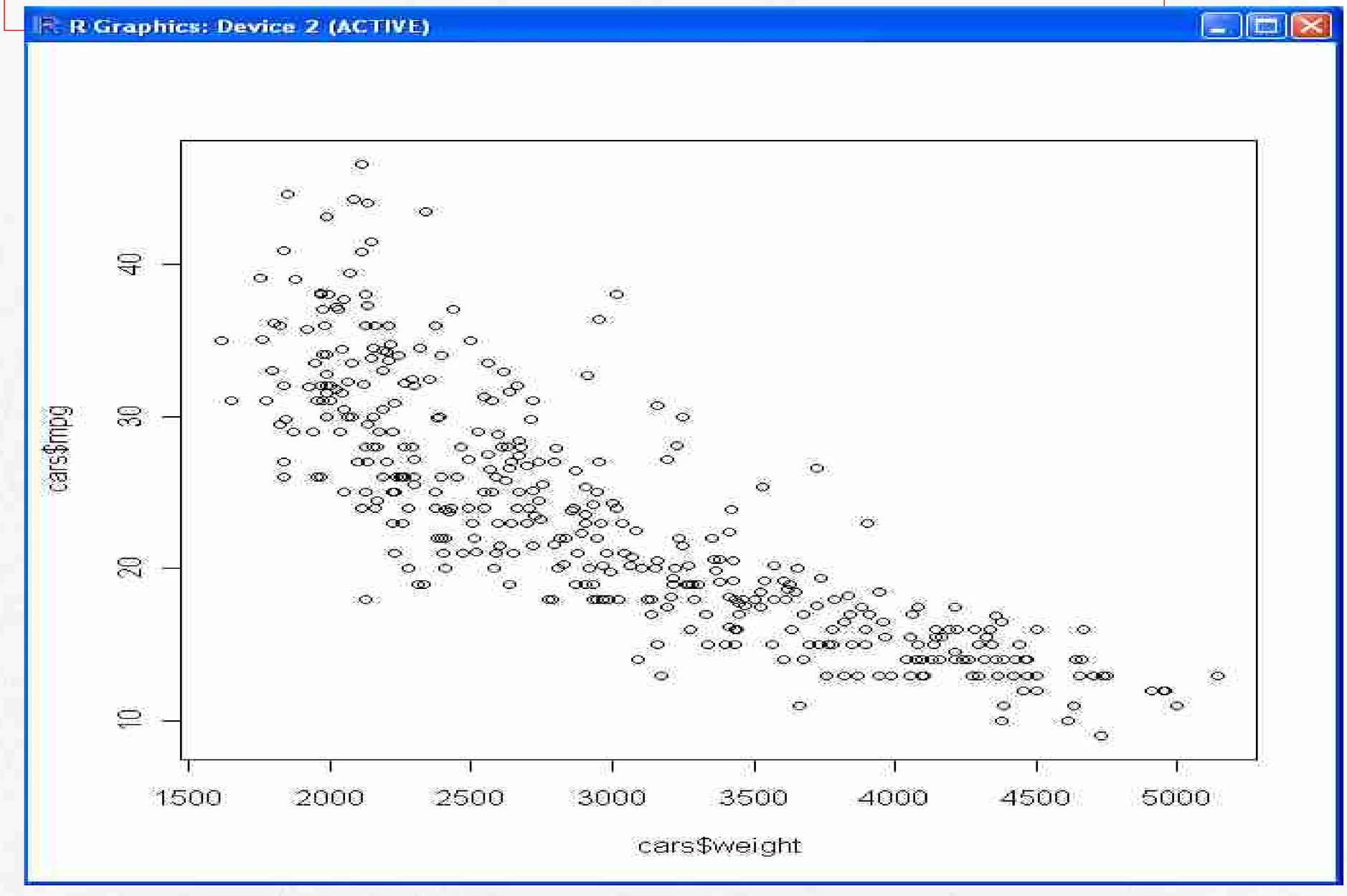


Кол-во f web site линков, которые указывают на основной сайт инструментальной среды March 19, 2011.

http://www.r4stats.com/popularity

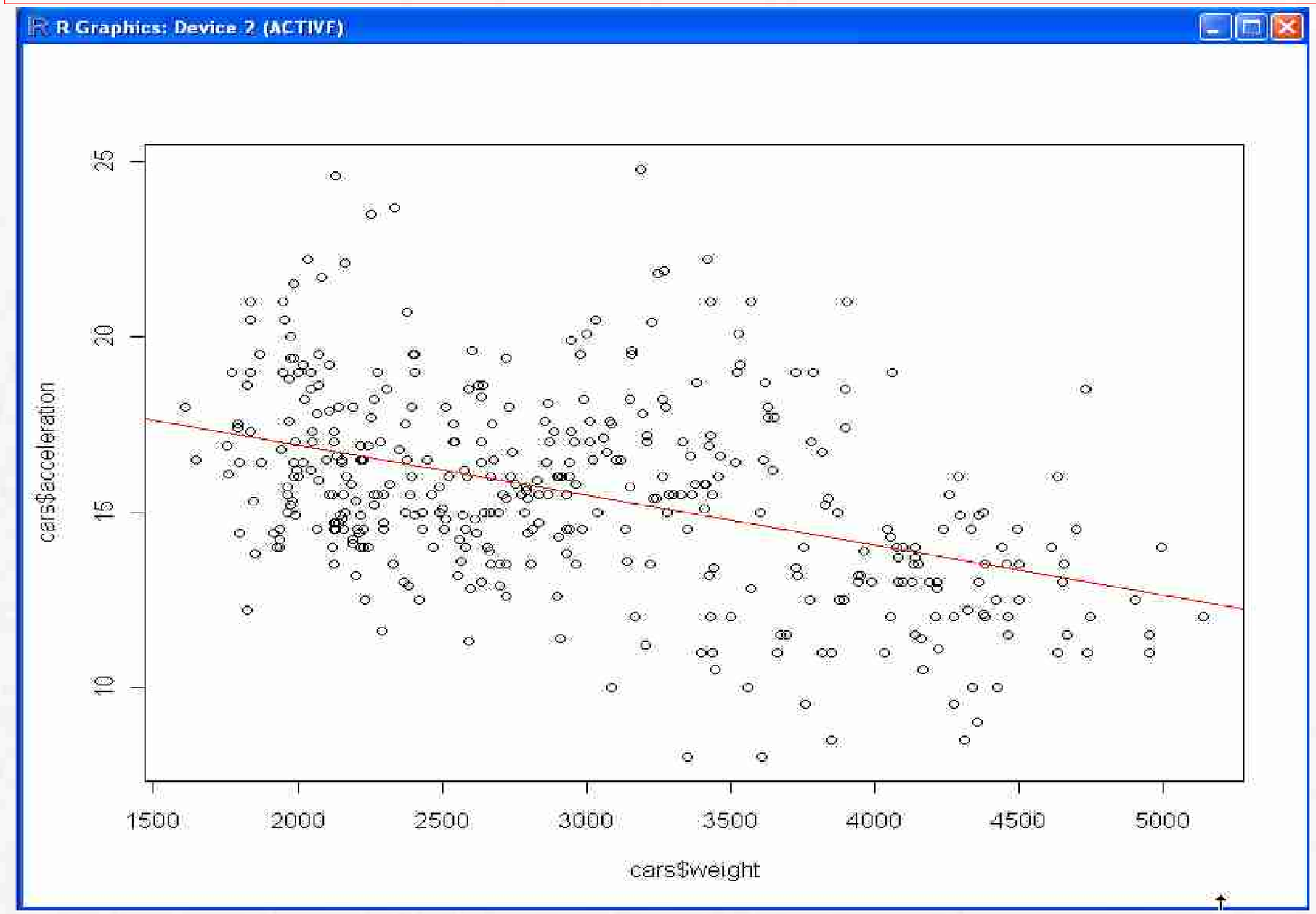
R Graphics

```
R> plot(cars$weight, cars$mpg)
```



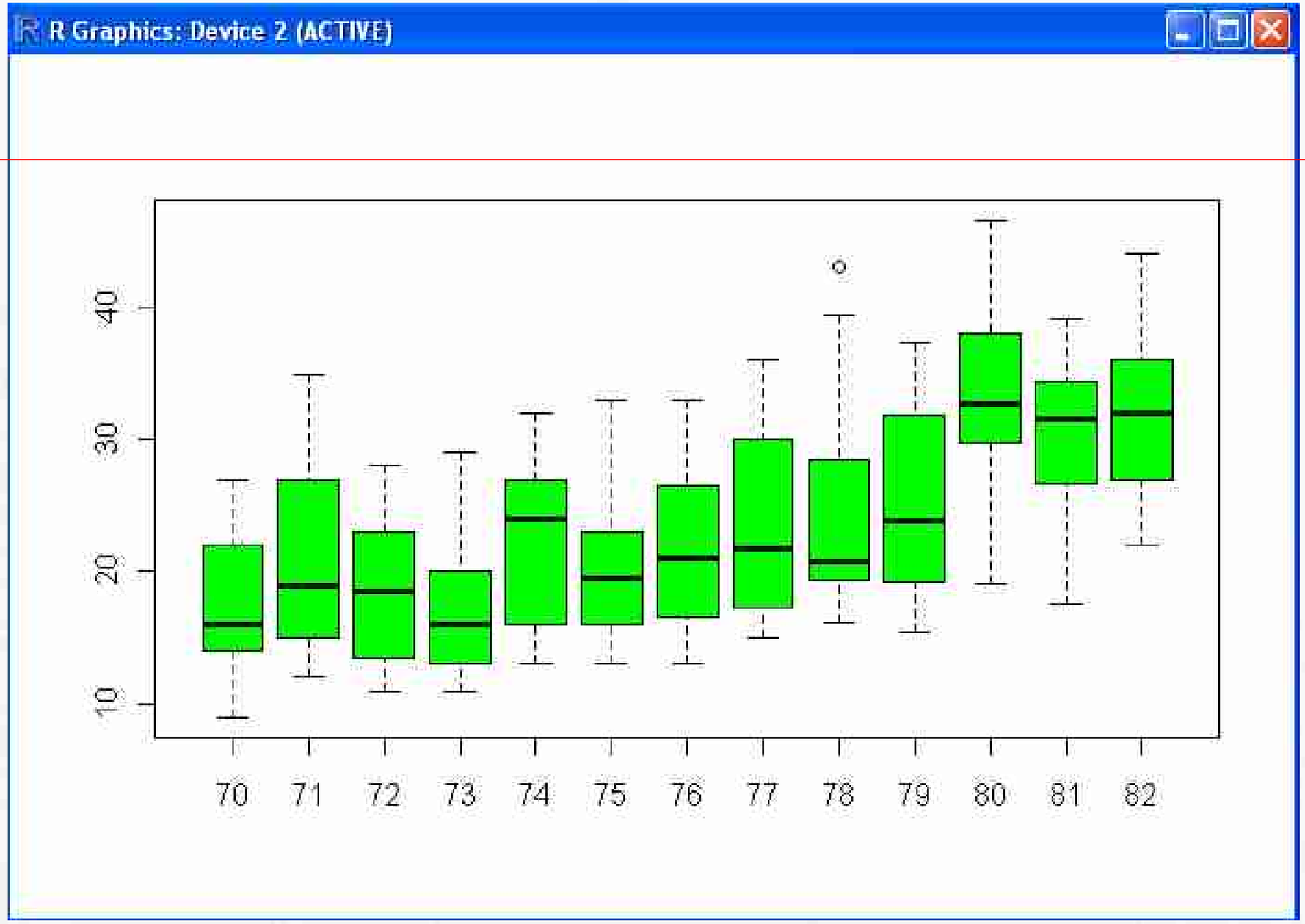
R Graphics

```
R> abline(coef(lm(acceleration ~ weight, cars)), col = "red")
```



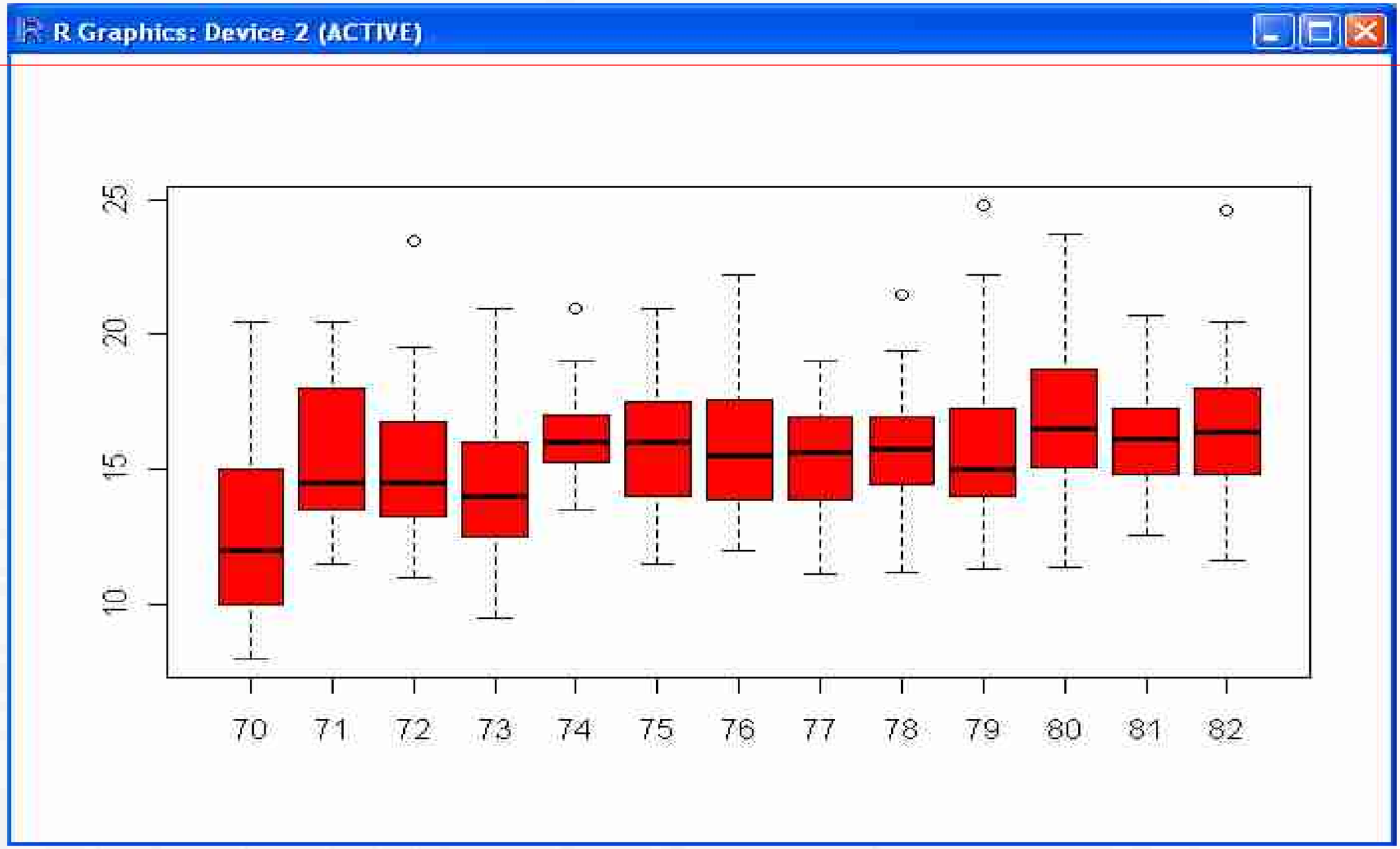
R Graphics

```
R> boxplot(split(cars$mpg,
  cars$model.year), col = "green")
```



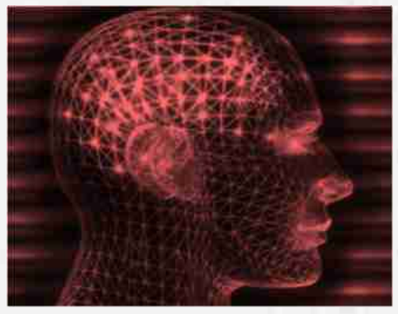
R Graphics

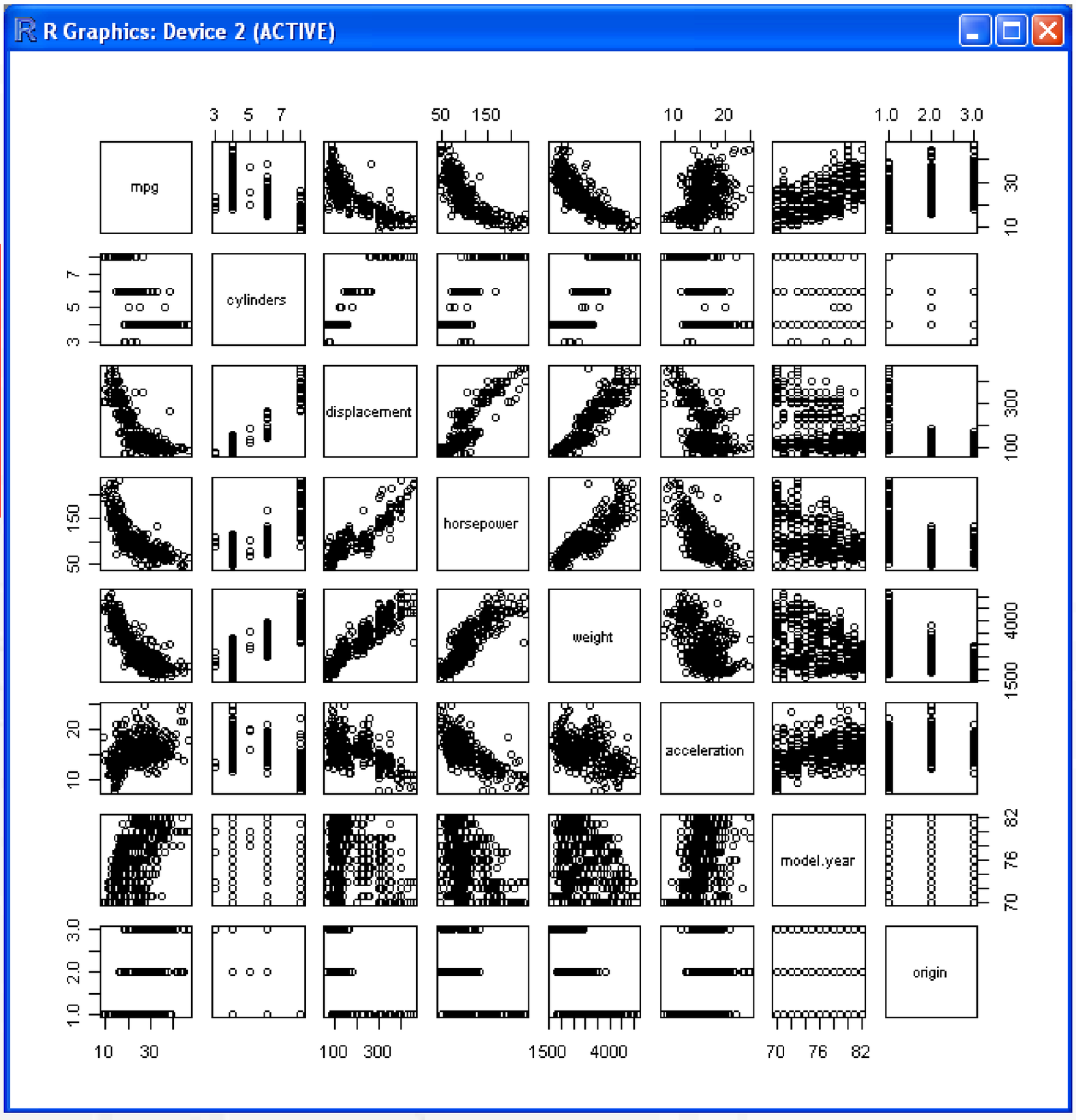
```
R> boxplot(split(cars$acceleration, cars$model.year),
  col = "red")
```



R графики

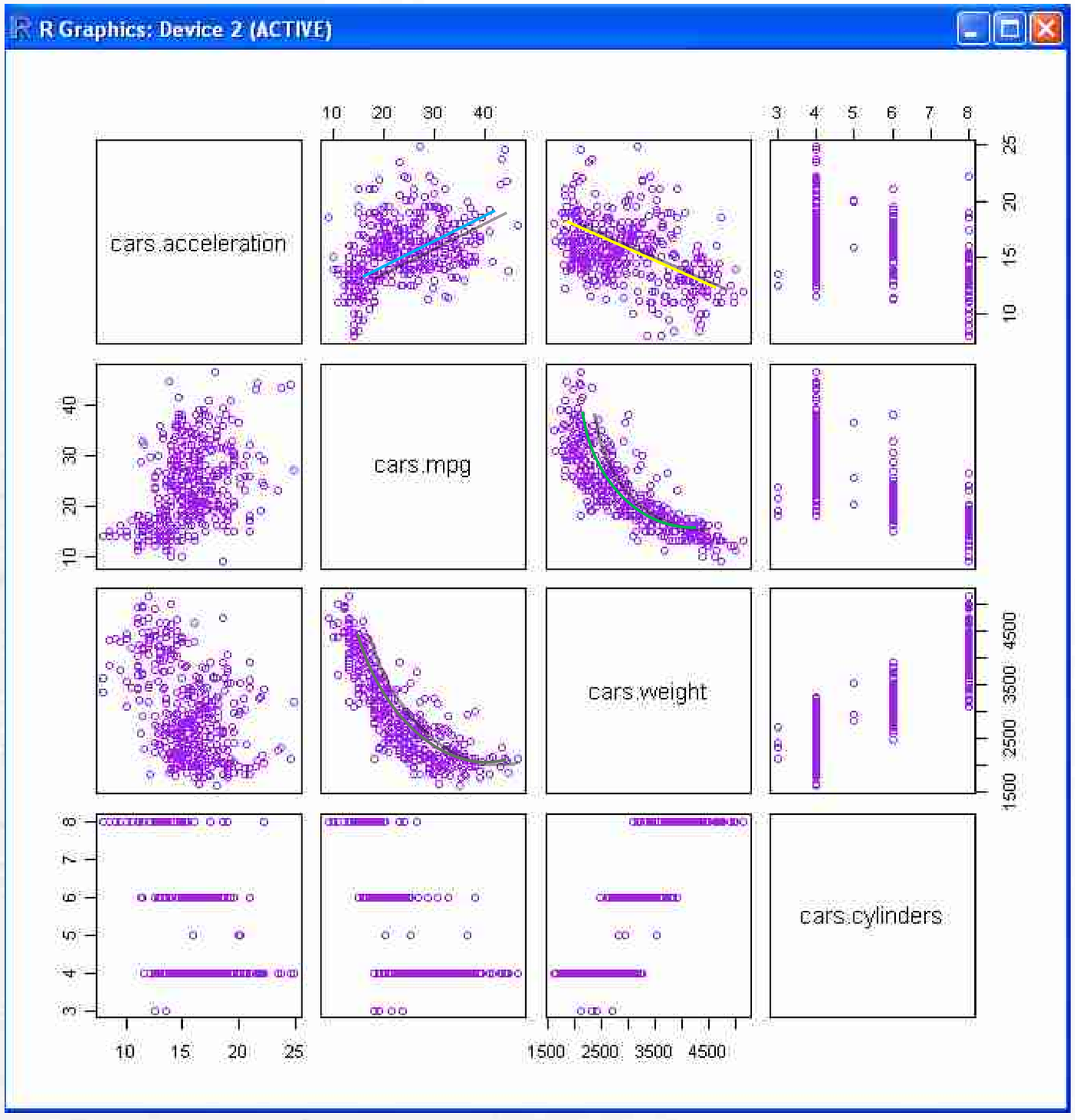
```
R> plot(cars)
```

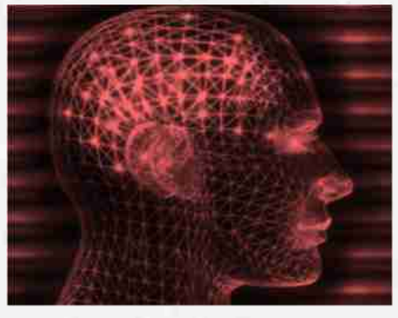




R графики

```
R>
   plot(data.frame(cars$accele
   ration,cars$mpg,
   cars$weight,
   cars$cylinders), col =
   "purple")
```

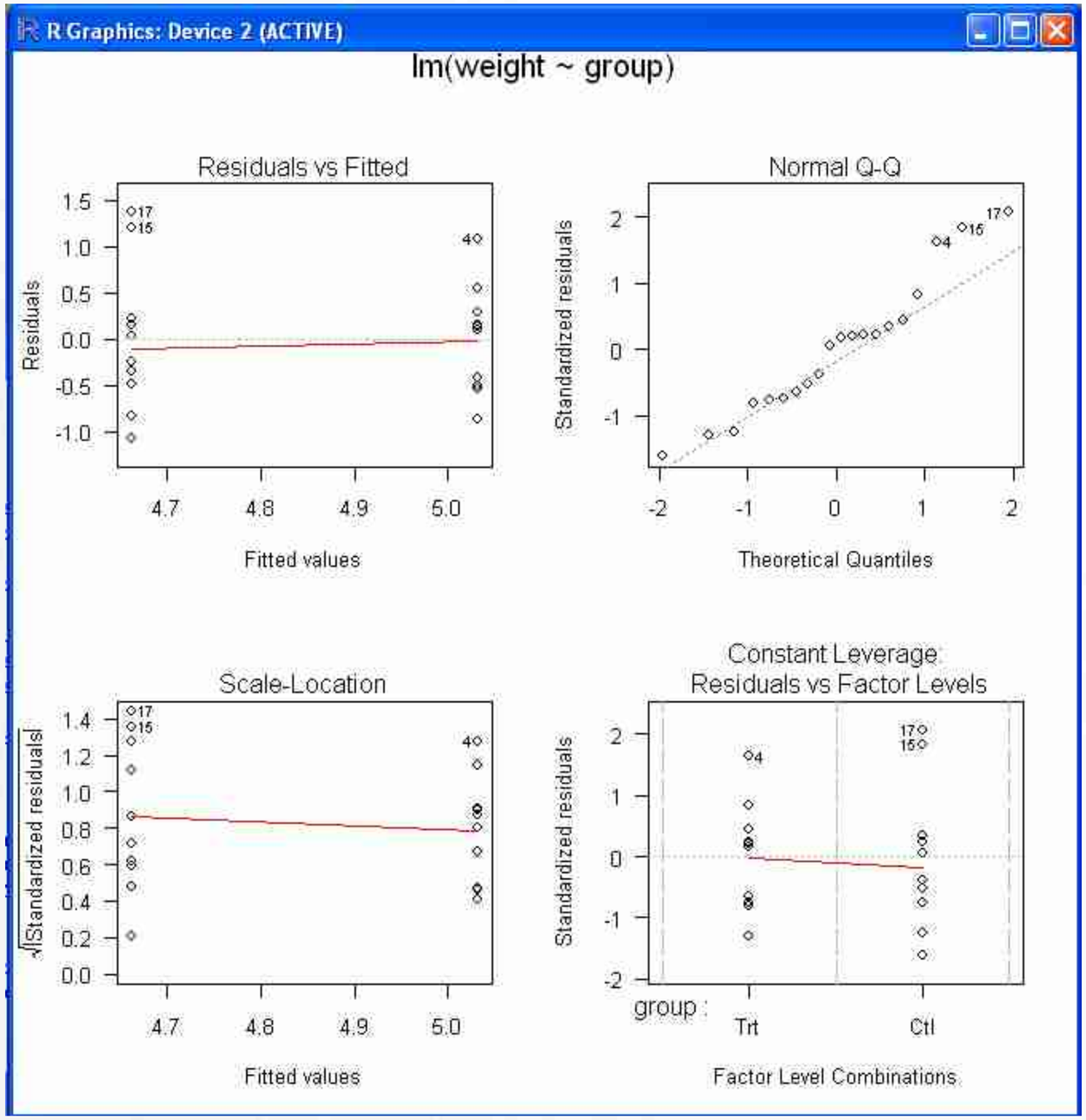




Линейное моделирование

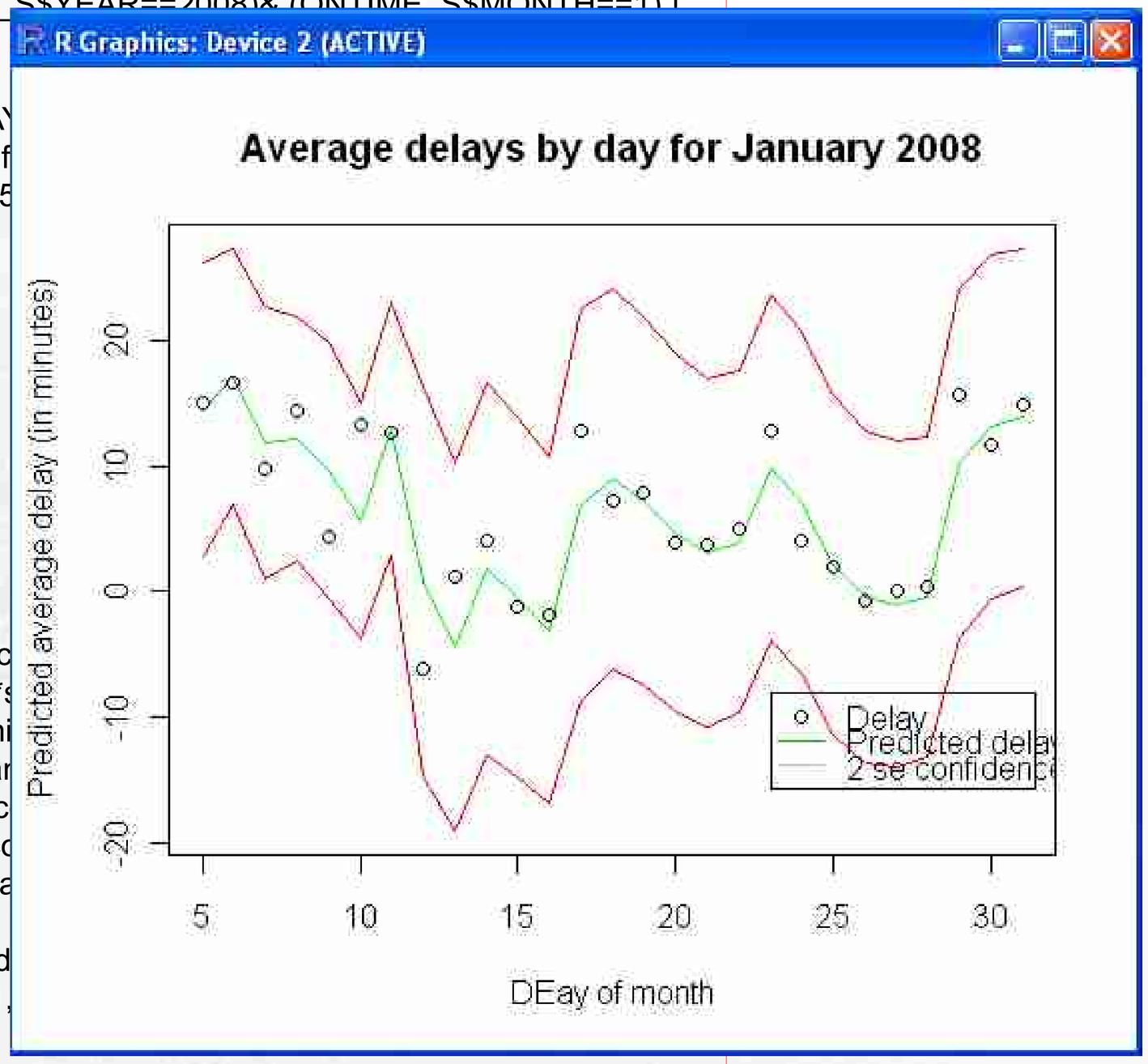
R Console

```
R> anova(lm.D9 <- lm(weight ~ group))
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value Pr(>F)
group      1 0.6882 0.68820  1.4191  0.249
Residuals 18 8.7293 0.48496
R> summary(lm.D90 <- lm(weight ~ group - 1))# omitting intercept

Call:
lm(formula = weight ~ group - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4938  0.0685  0.2462  1.3690

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
groupCtl   5.0320     0.2202   22.85 9.55e-15 ***
groupTrt   4.6610     0.2202   21.16 3.62e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared: 0.9818,     Adjusted R-squared: 0.9798
F-statistic: 485.1 on 2 and 18 DF,  p-value: < 2.2e-16
```

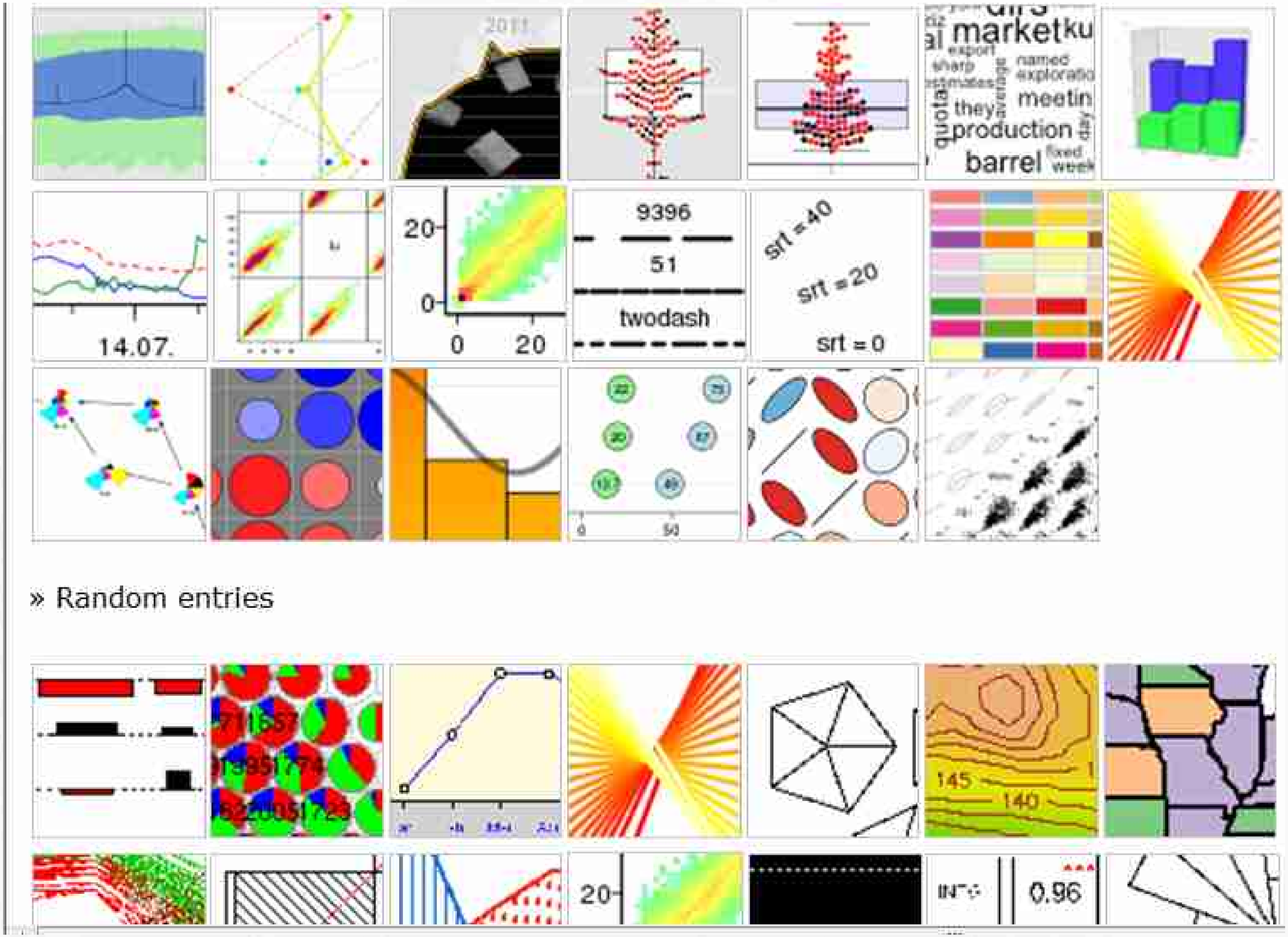


Oracle R Enterprise ARIMA --прогнозирование

```
year200801 <- ONTIME_S[(ONTIME_S$YEAR==2008)& (ONTIME_S$MONTH==1),]
y <- ore.pull(year200801)
gc()
delays <- tapply(y$ARRDELAY, y$DAY
delays <- ts(delays, start=1, end=31, f
# Create a Kalman filter with the first 5
preds <- c()
ses <- c()
# 1 step predictions
for (i in 5:length(delays))
{
  fit <- arima(delays[1:i], c(1,2,1))
  # predict 1 step into the future.
  pred <- predict(fit)
  preds <- c(preds, pred$pred)
  ses <- c(ses, pred$se)
}
plot(5:length(delays), preds, type='l', c
  ylim=range(c(preds+2*ses, preds-2*s
  ylab="Predicted average delay (in mi
  main="Average delays by day for Jar
lines(5:length(delays), preds+2*ses, c
lines(5:length(delays), preds-2*ses, cc
points(5:length(delays), as.vector(dela

legend( 23, -8, c("Delay", "Predicted d
  col=c(1, 3, 8), lty=c(0, 1, 1), pch=c(1,
```

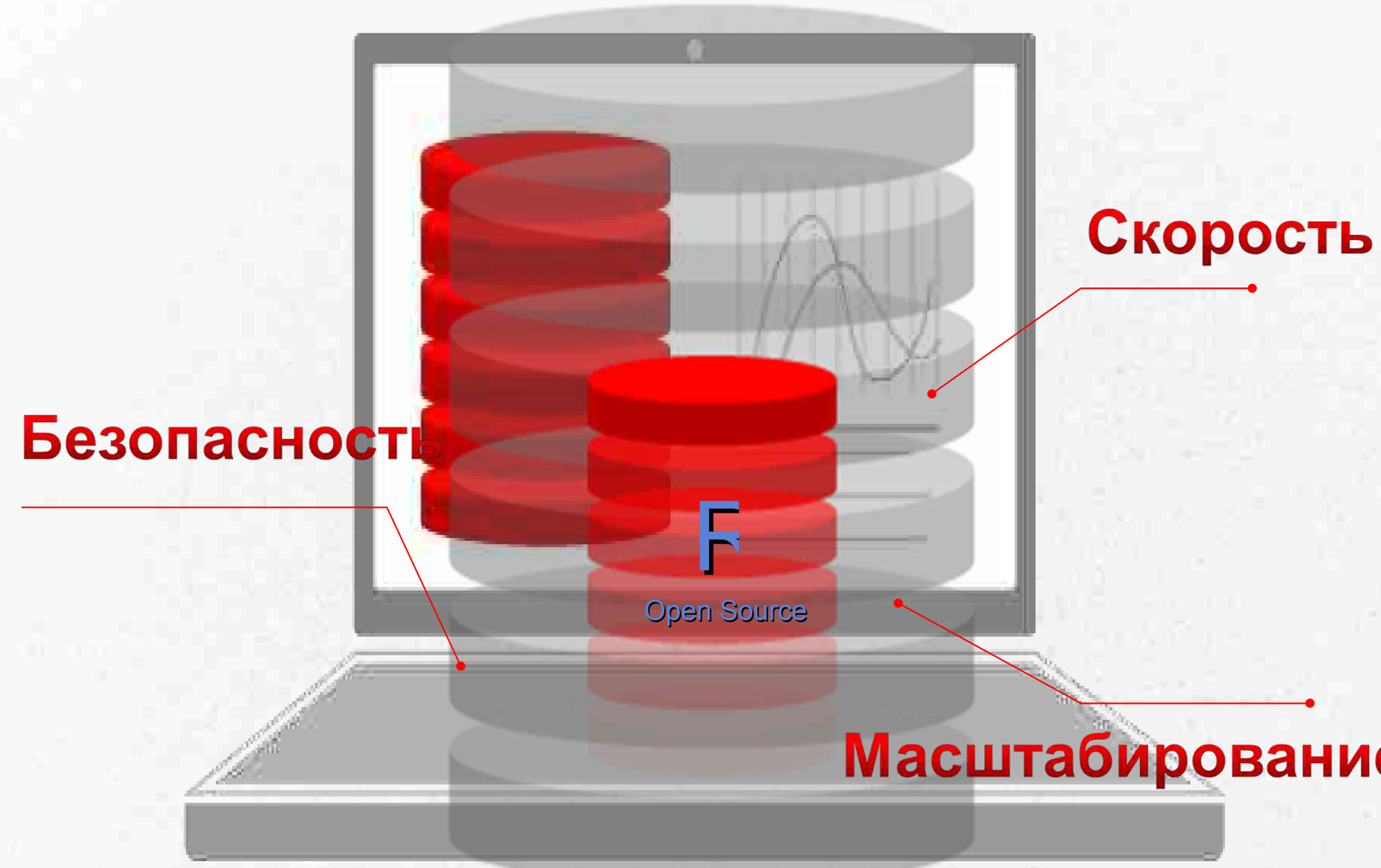


Примеры графиков, генерируемых R



» Random entries

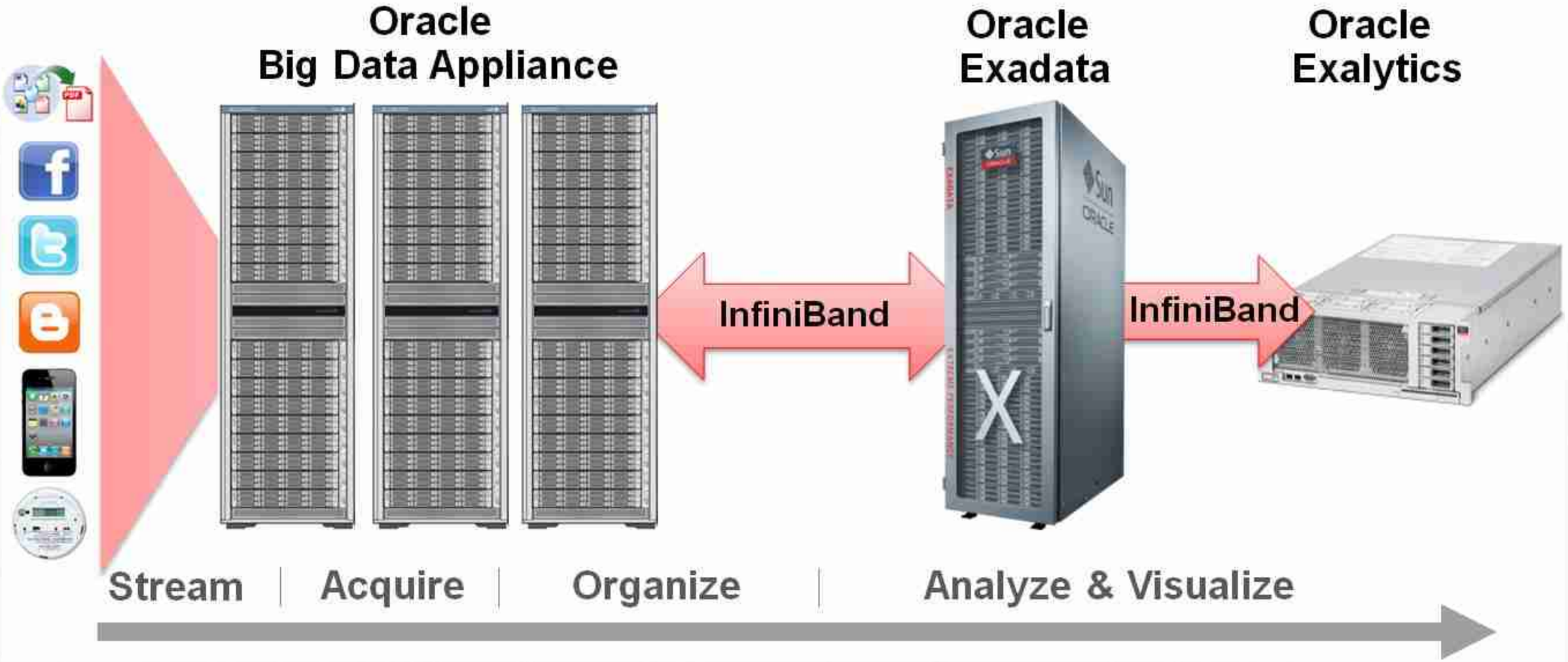
Oracle R Enterprise



- R «встраивается» в Oracle database
- Данные сохраняются и статистические вычисления выполняются в базе данных
- 100% совместимость с R интерфейсом и клиентскими приложениями
- Дополняет Oracle Data Mining

ORACLE R Enterprise
Open Source

Анализ любых данных



План

- Oracle Exalytics -- аналитика в оперативной памяти
- Oracle R Enterprise – статистический анализ и визуализация для Больших Данных
- **Endeca – платформа для систем исследования неструктурированной информации**

Endeca

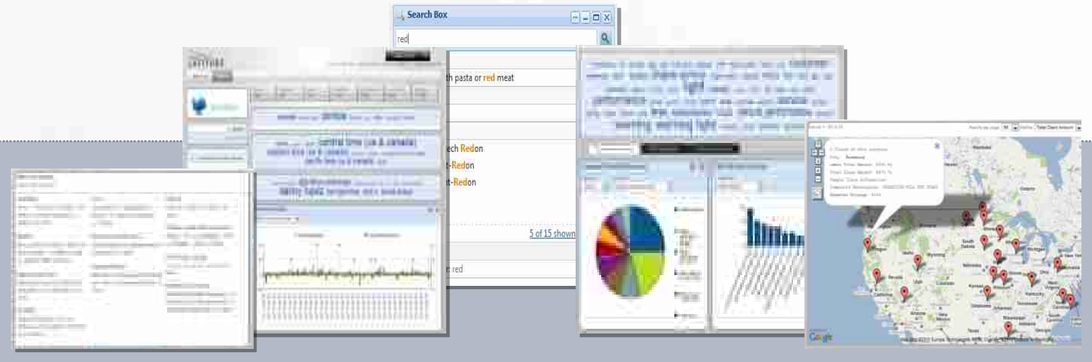
Data Exploration and Discovery

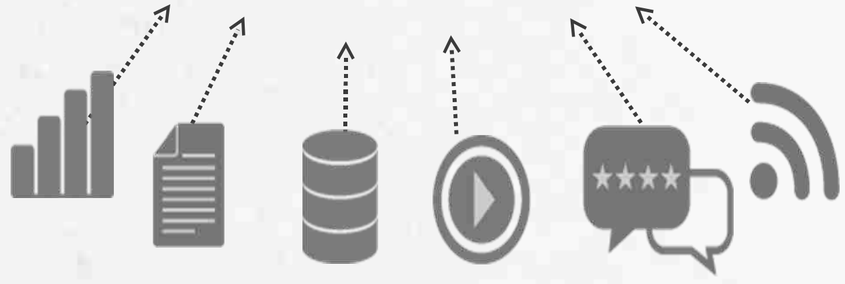
- Основана в Кембридже, MA в 1999
- Более 600 клиентов
- 33% of the Fortune 100
- Анилиз неструктурированной информации, Большие данные
- От систем поиска для электронной коммерции к «BI beyond the data warehouse»
- *Entdecken* (немецкий) = to discover.

Oracle Endeca Information Discovery

Платформа для систем исследования данных

- Структурированные и неструктурированные данные из различных источников
- Поиск по «нечетким» критериями
- Быстрая разработка систем класса Data Discovery
- Основной компонент – Endeca сервер
- Фасетный поиск





Фасетный поиск, фасетная навигация

- Поиск путем уточнений
- Модель информационного поиска
- Набор независимых параметров-фильтров, применяемых в объектам

Data Discovery – основные возможности

- Удобство и простота использования
  - На основе 10-летнего опыта работы в области разработки поисковых систем для электронной коммерции
- Поиск + Фасетная навигация + Визуальный анализ
  - Поиск и выбор атрибутов в стиле вэб сайтов
- Интерактивные исследования
  - Поддерживаются сервером Endeca

Разработка системы исследования данных

Источники данных

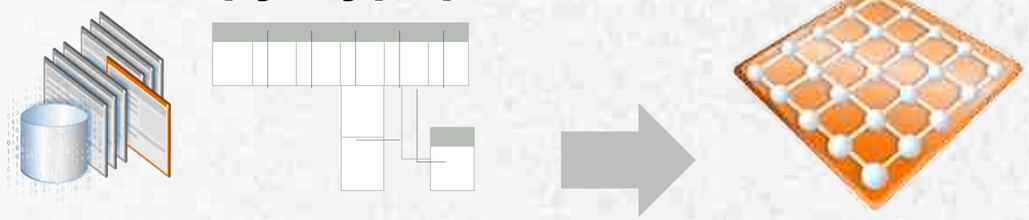
Автоматическая загрузка в Endeca Server (без модели)

Drag-and-drop инструменты создания приложения

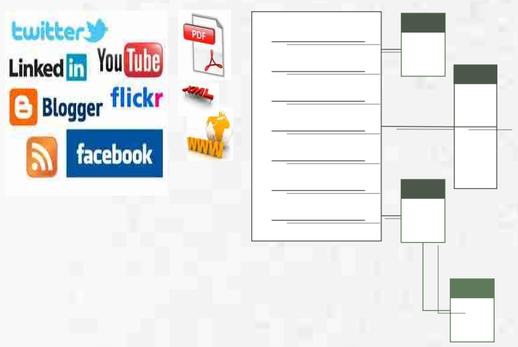
Интерактивный поиск, навигация и анализ

Структурированные

Слабоструктурированные



Неструктурированныеd



Более 600 клиентов

Полная аналитическая платформа Oracle

Спасибо за внимание!