EMC Greenplum

Унифицированная платформа для аналитики Больших Данных

Денис Серов
Руководитель направления
технического консультирования
EMC Россия и СНГ
Denis.serov@emc.com

GREENPLUM.

EMC²

Greenplum стал фундаментом Аналитики Больших Данных EMC (Июль 2010)

EMC ПРИОБРЕТАЕТ GREENPLUM

"В течение трех лет, Gartner именовал Greenplum как **самого продвинутого вендора среди Визионеров** в своем ежегодном DBMS Magic Quadrant…."

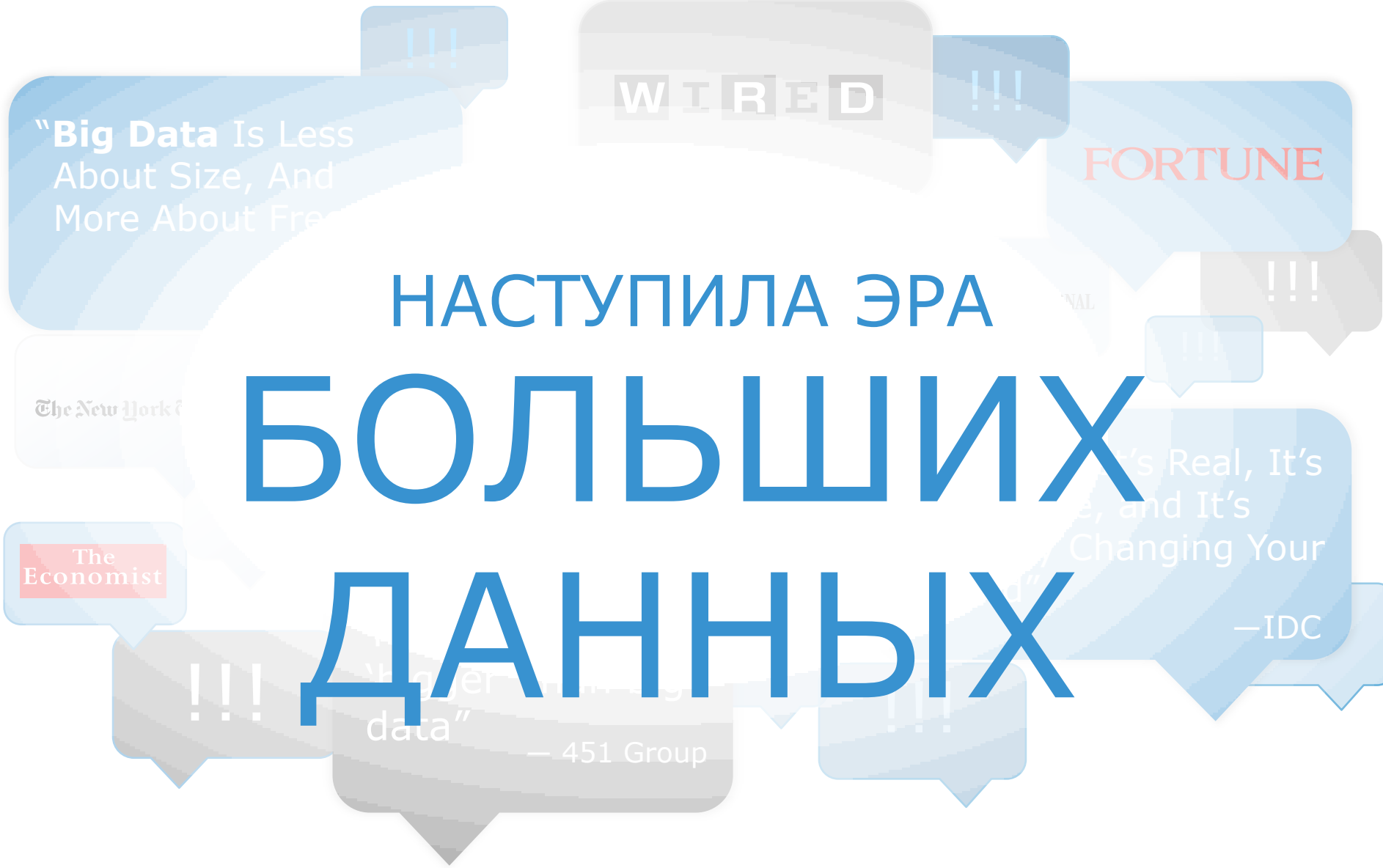
– Gartner

GREENPLUM.

EMC²

1,000,000,000 запросов в день

900ms среднее время выполнения



НАСТУПИЛА ЭРА БОЛЬШИХ ДАННЫХ

GREENPLUM. EMC²

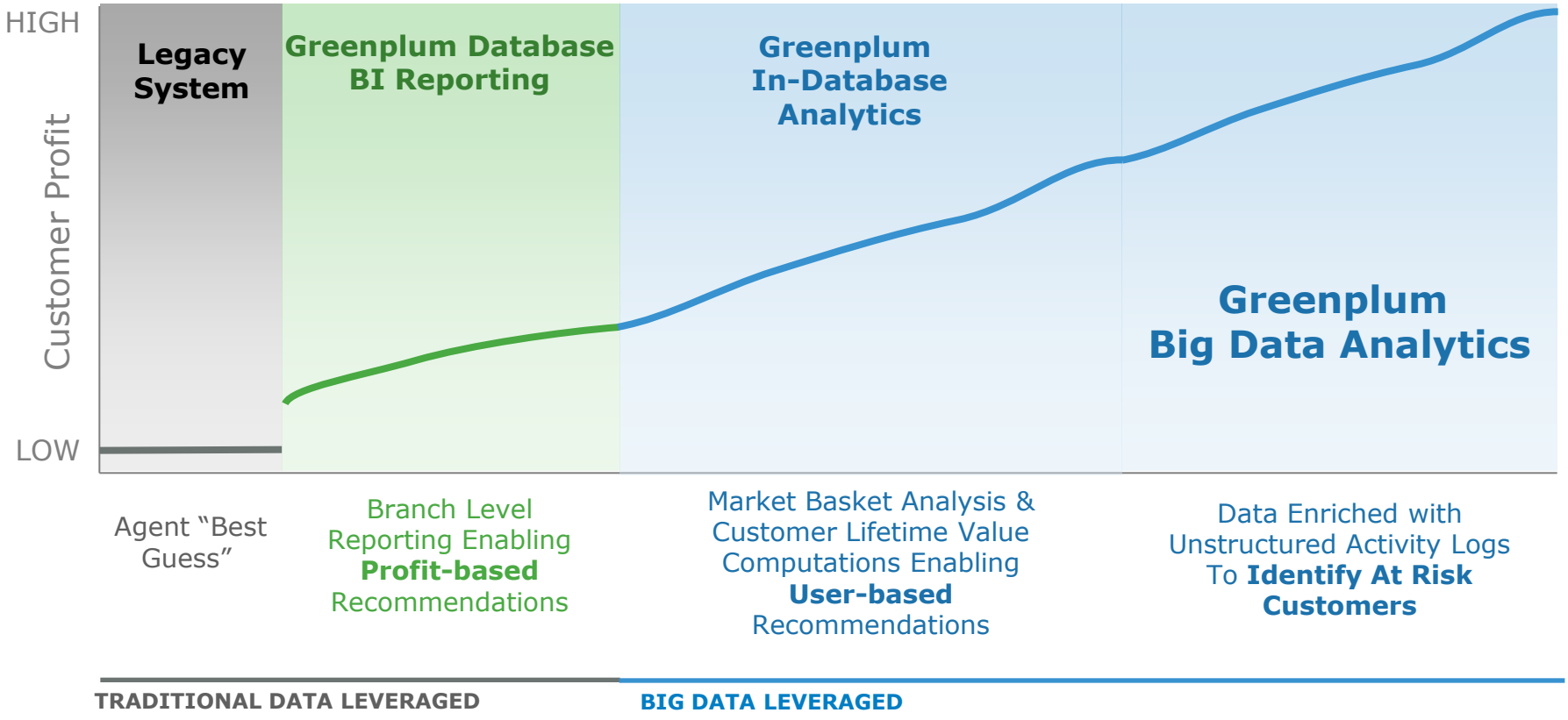
В ЭРУ БОЛЬШИХ ДАННЫХ АНАЛИТИКА ЯВЛЯЕТСЯ КЛЮЧОМ К УСПЕХУ

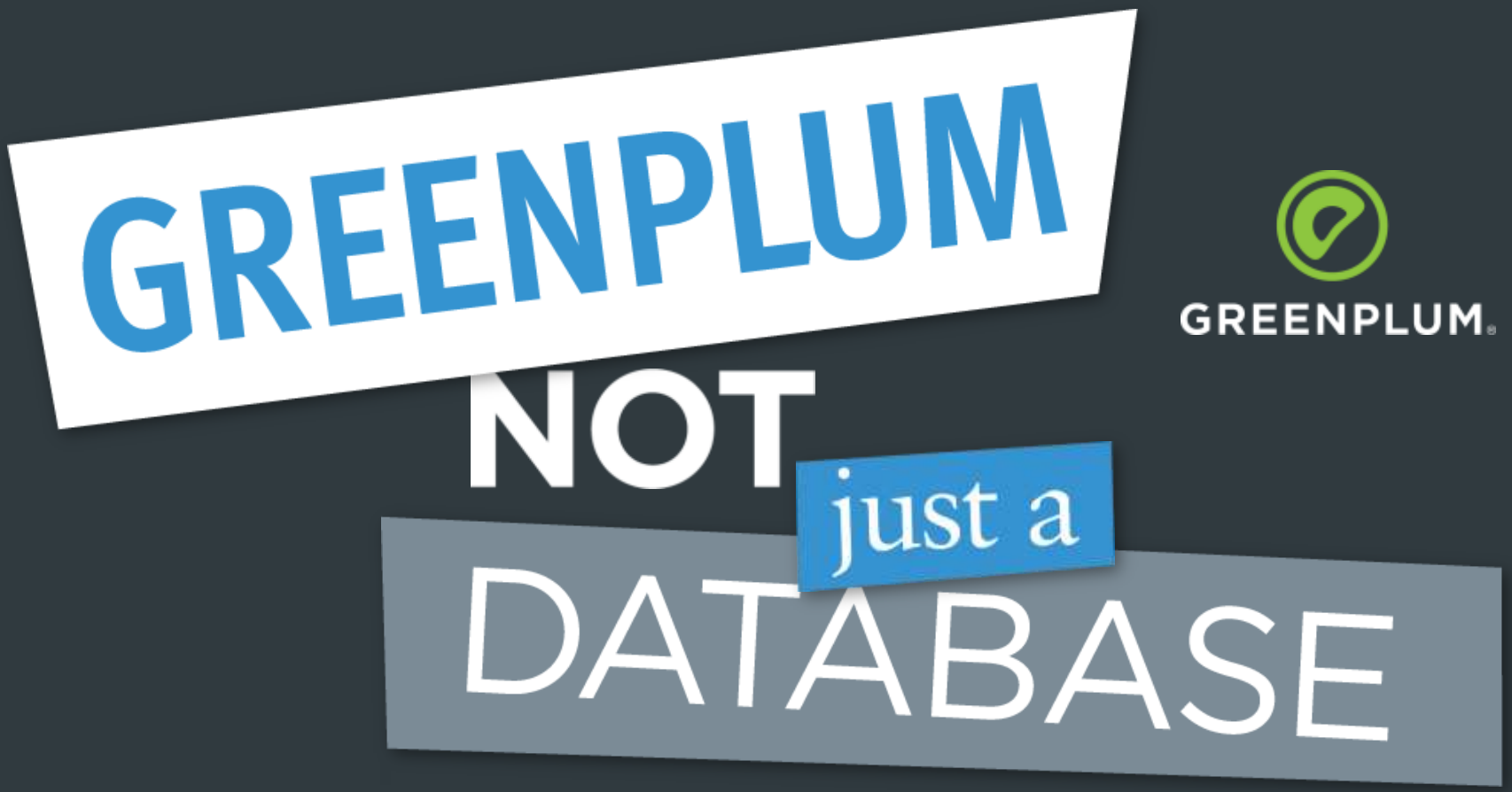
Аналитика Больших Данных: Путь к Выгоде для Бизнеса

Пример использования

Предсказание поведения покупателей для повышения дохода

Аналитика Больших Данных делает возможным увеличение прибыли на покупателя





GREENPLUM

UNIFIED ANALYTICS PLATFORM FOR BIG DATA.

GREENPLUM.

EMC²

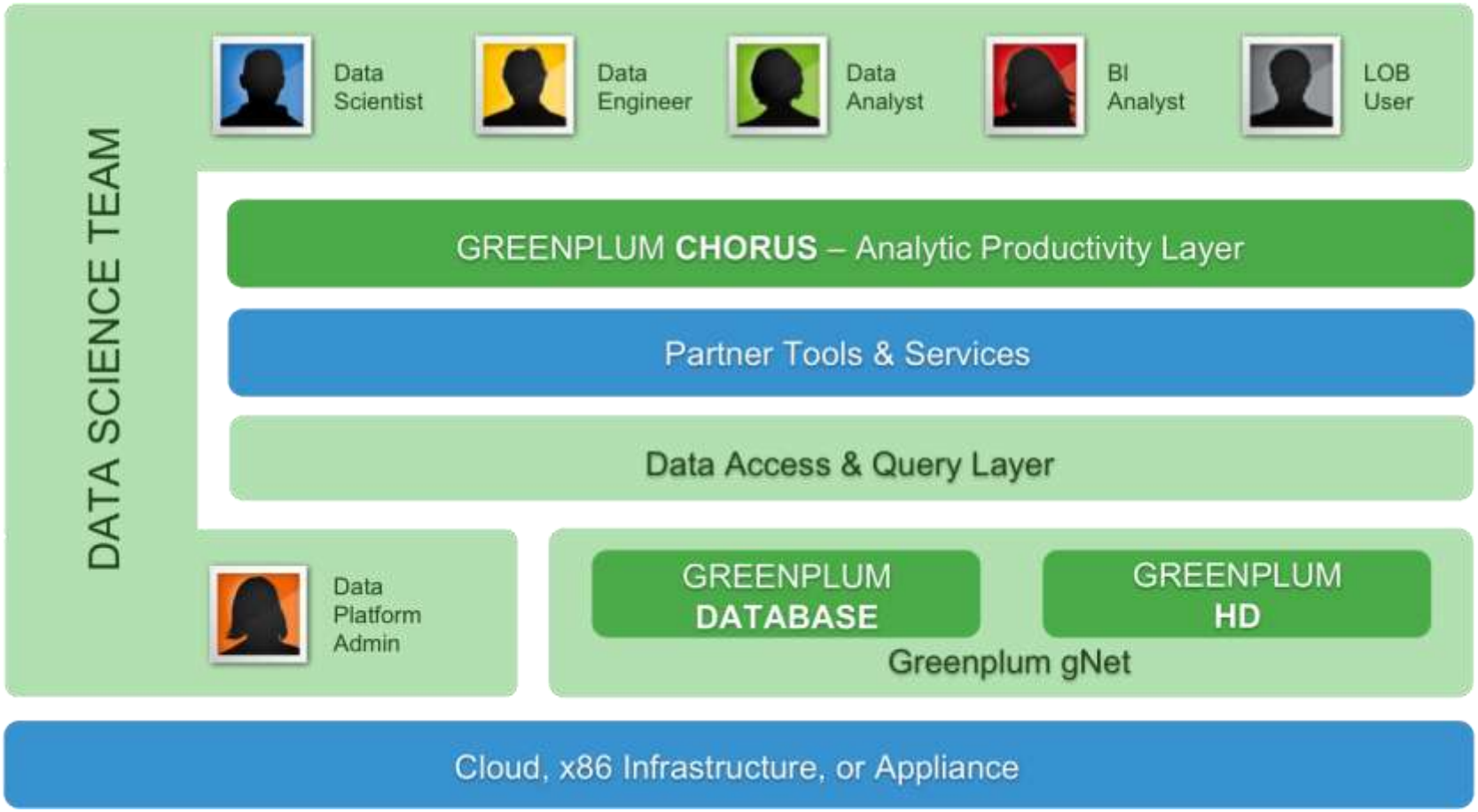
GREENPLUM

UNIFIED ANALYTICS PLATFORM FOR BIG DATA.

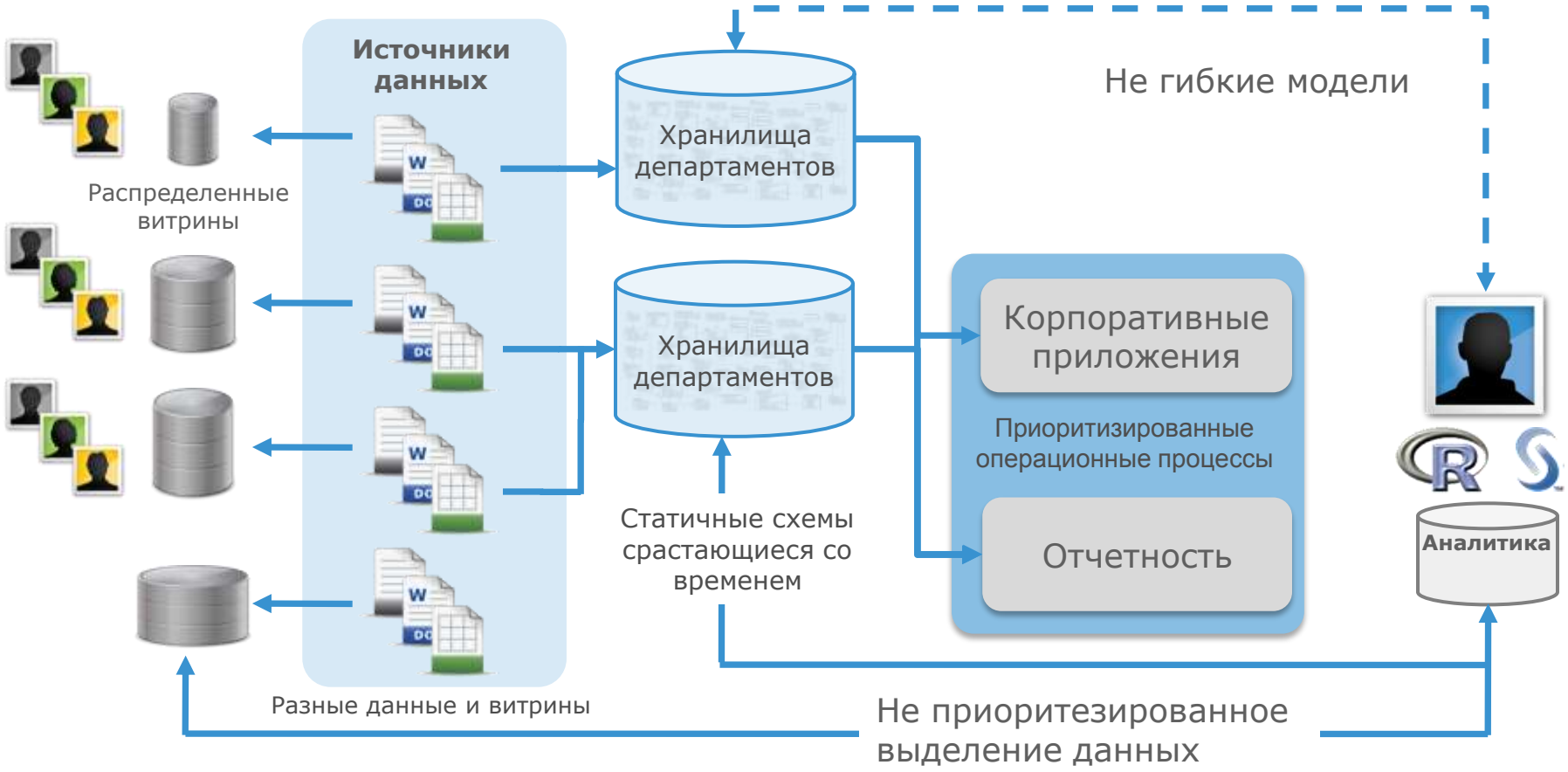
GREENPLUM.

EMC²

Greenplum: унифицированная платформа для аналитики



Типичная архитектура хранилища данных для аналитики



Архитектура Greenplum для аналитики

GPDB + Hadoop + Chorus + Labs

Аналитика вместе с Greenplum

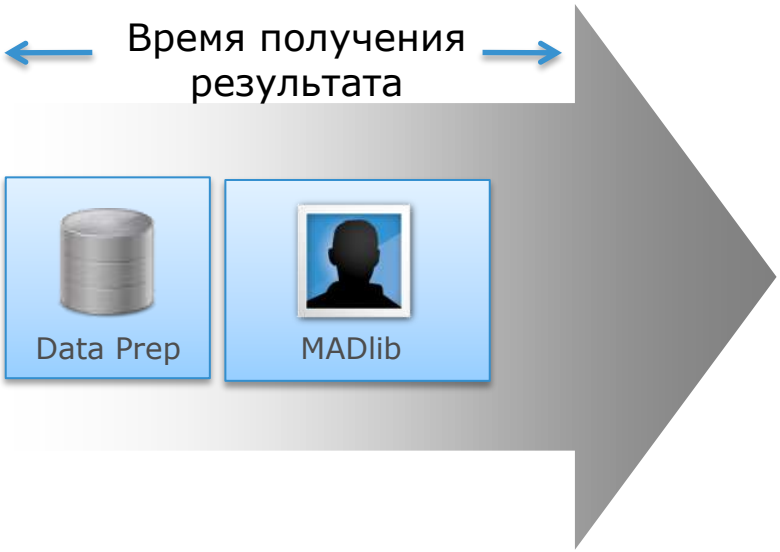
Массивная масштабируемость скоростной аналитики SAS-Greenplum


```
libname GPlib &dbms &CONNOPT
   connection=unique;
data work.combined;
  merge GPlib.STAFF GPlib.SUPERV(in=super
    rename=(SUPID=IDNUM));
  by IDNUM;
  if super;
run;

proc hplogistic data=GPlib.sgf_binary;
   class A B C;
   model y = a b c x1 x2 x3;
   performance details host="gpprod";
run;
```

Аналитика вместе с Greenplum

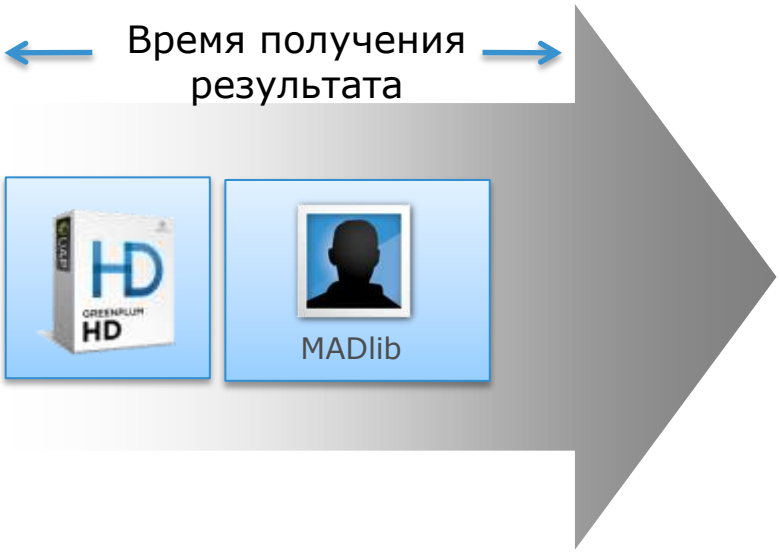
Маркетинговая оптимизация при помощи MADlib




```
> SELECT householdID, variables
    FROM households
    ORDER BY RANDOM()
    LIMIT 100000;
> SELECT run_univariate_analysis (
        'households_training',
        'variables');
    WHERE pvalue<.01 AND r2>.01;
> SELECT run_regression(
        'univariate_results',
        'households_training');
> SELECT householdID,
madlib.array_dot(
        coef::REAL[],
        xmatrix::REAL[])
    FROM coefficients, households;
```

Аналитика с Greenplum

Текстовая аналитика с Hadoop и MADlib




```
-MAP:
    NAME:       extract_terms
    PARAMETERS: [id integer, body text]
    RETURNS:    [id int, title text, doc _text]
    FUNCTION: |

                        if 'parser' not in SD:
                            import ...
                            class MyHTMLParser(HTMLParser):
                              ...
                            SD['parser'] = MyHTMLParser()

                        parser = SD['parser']
                        parser.reset()
                        parser.feed(body)
```

```
> CREATE EXTERNAL TABLE document_feature (
    docid integer,
    term text,
    freq integer)
  LOCATION ('gphdfs://localhost:9000/user/schopf/docs.txt')
  FORMAT 'text' (delimiter '|');
```

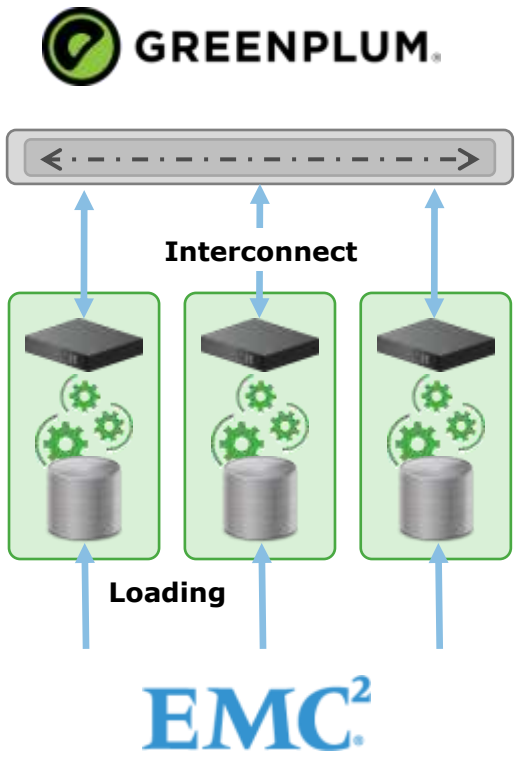
```
id  |                      tfxidf
-----+-------------------------------------------------------------------
2482 | {3,1,37,1,18,1,29,1,45,1,...}:{0,8.25206814635817,0,0.34311110...}
   1 | {41,1,34,1,22,1,125,1,387,...}:{0,0.771999985977529,0,1.999427...}
  10 | {3,1,4,1,30,1,18,1,13,1,4,...}:{0,2.95439664949608,0,3.2006935...}
...
```

СУБД Greenplum: экстремальная производительность для аналитики

- Оптимизирована для BI и аналитики
  - Глубокая интеграция со статистическими инструментами
  - Высокая производительность параллельной обработки
- Простая и автоматизированная
  - Загрузка и запросы как в обычной СУБД
  - Таблицы автоматически распределяются по узлам
- Экстремально масштабируемая
  - MPP архитектура без разделения ресурсов
  - Все узлы обрабатывают данные параллельно
  - Линейная масштабируемость

Производительность достигаемая через параллелизм обработки данных

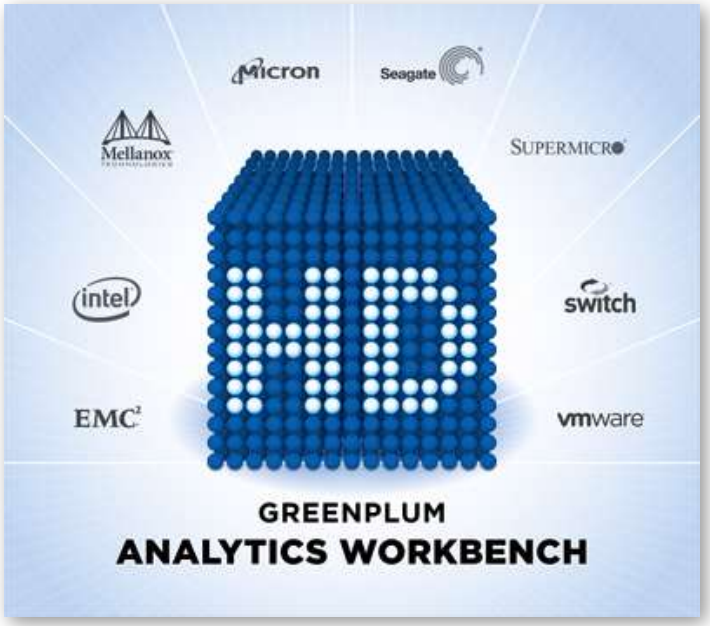
- Горизонтально масштабируемая архитектура на стандартном оборудовании
- Автоматический параллелизм
  - Загрузка и запросы как в обычной СУБД
  - Автоматическое распределение таблиц
  - Не требует ручной настройки
- Экстремально масштабируемая MPP архитектура без разделения ресурсов
  - Все узлы обрабатывают параллельно
  - Линейная масштабируемость
  - Расширение без остановки работы



Greenplum HD
решение Hadoop для предприятий

- Apache Hadoop
  - На базе наиболее стабильной версии
- Корпоративная поддержка
  - 24x7 поддержка от EMC
- Проверенная
  - Сертифицировано EMC
- Скоростное хранилище Isilon в виде опции
  - Лучшая в своем классе СХД без необходимости менять Hadoop приложения

Проверенное решение с поддержкой мирового класса



Bringing Rapid Innovation to Hadoop

- Крупнейшая в индустрии команда поддержки Hadoop
  - Ведущие специалисты по Hadoop (из Yahoo!, LinkedIn, Talend, и т.д.)
- Проверено в масштабе
  - 1,000-узлов, 24-Петабайт
  - Многомиллионные инвестиции
  - Сниженный риск для заказчиков EMC
  - Сертификация с партнерами

Greenplum Chorus делает возможной гибкость для Больших Данных

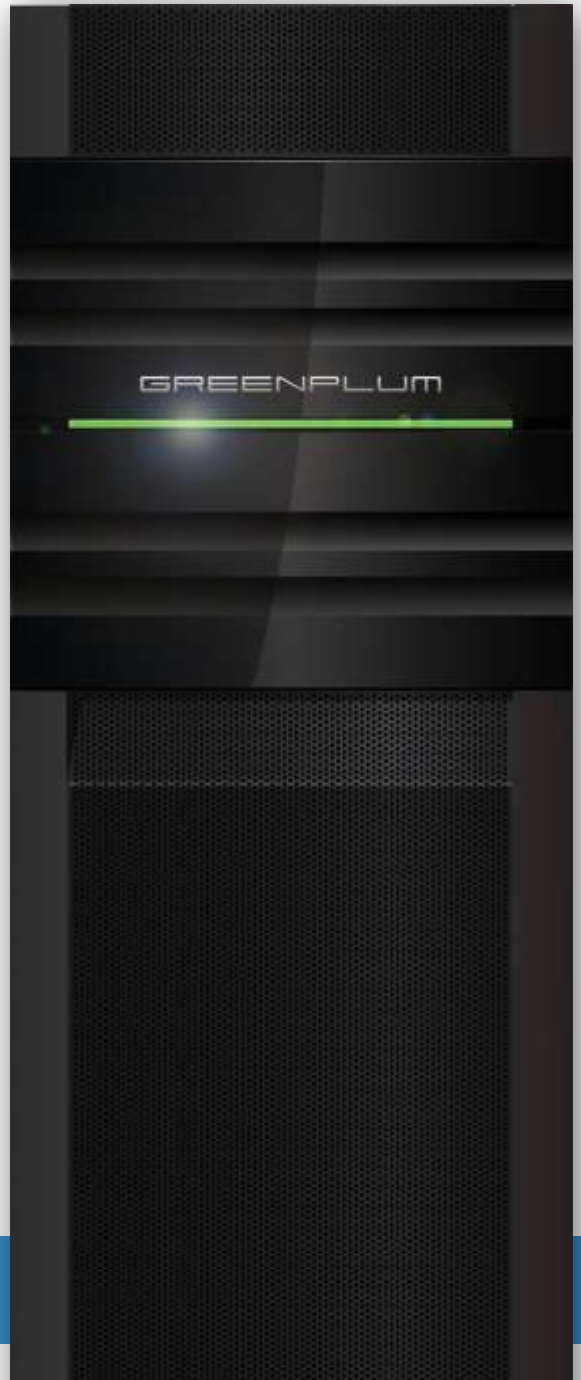
- Первая в мире платформа аналитической продуктивности
  - Поиск, изучение, визуализация и импорт данных со всего предприятия
  - Самообслуживание и выделение аналитических рабочих сред
  - Создание, совместное использование и публикация инсайтов для гибкой аналитики

Коллаборативная среда для аналитики

- Рабочие окружения для аналитики больших данных
- Более прозрачные проекты
- Коллаборация внутри проектов, обмен информацией между командами

Модульный Greenplum

- Представляем первое в мире:
  - Высокоскоростное
  - Специально созданное
  - Устройство для обработки данных
- Комбинация СУБД Greenplum и Greenplum Hadoop в одном устройстве



Greenplum это выбор и гибкость

Greenplum Data Computing Appliance

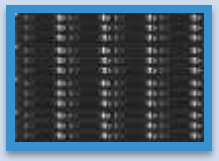
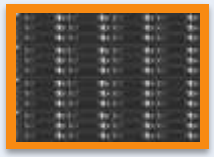
- Выбор модулей Greenplum СУБД и/или Hadoop инкрементами по ¼ шкафа
- Масштабирование добавлением новых узлов по Вашему выбору
- Минимальное время развертывания

Greenplum Software Solutions

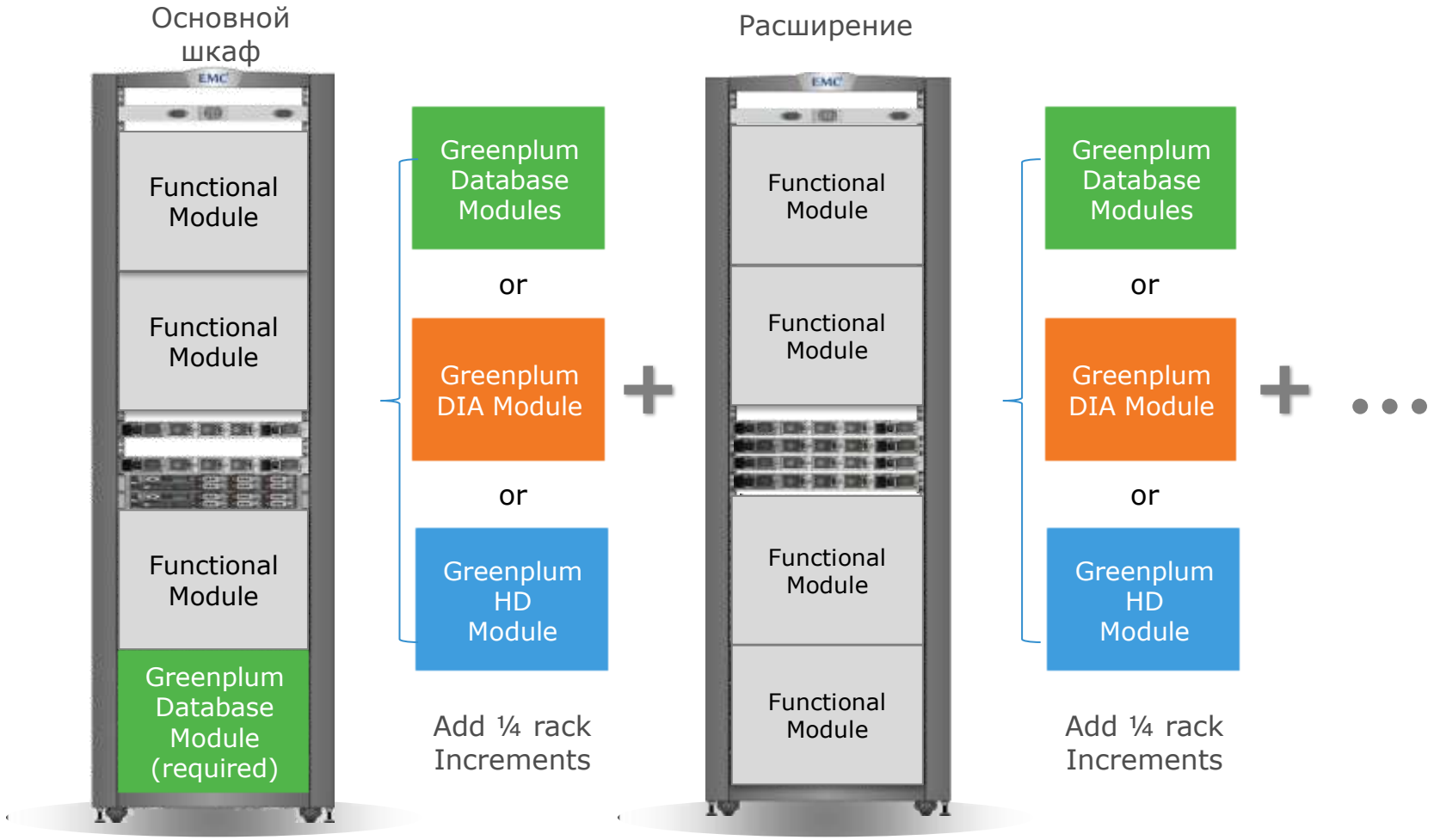
- Greenplum СУБД, Hadoop, & Chorus на Вашем x86 оборудовании
- Гибкость для любой нагрузки и окружения
- Постоянные лицензии, либо годовая подписка

Greenplum Data Computing Appliance

Революционная модульная архитектура

Стандартный модуль Greenplum СУБД		9TB (без сжатия)	Состав каждого сервера: • 2 sockets/12 ядер - 48GB памяти • 12x 600GB 10k rpm
Высокоемкий модуль Greenplum СУБД		31TB (без сжатия)	Состав каждого сервера: • 2 sockets/12 ядер - 48GB памяти • 12x 2TB 7.2k rpm
Модуль Greenplum HD (Hadoop)		28TB (3 копии, без сжатия)	Состав каждого сервера: • 2 sockets/12 ядер - 48GB памяти • 12x 2TB 7.2k rpm
Модуль ускорения интеграции данных (DIA)		70TB	Состав каждого сервера: • 2 sockets/12 ядер - 48GB памяти • 12x 2TB 7.2k rpm

Масштабируемость до нескольких шкафов инкрементами по ¼ шкафа



Примеры кластерных конфигураций DCA Greenplum

Модули СУБД Greenplum

Module Type	Стандартный модуль Greenplum Database		Модуль высокой емкости Greenplum Database	
К-во модулей	4	48	4	48
К-во шкафов	1	12	1	12
Полезная емкость (без сжатия)	36 TB	432 TB	124 TB	1,488 TB
Полезная емкость (со сжатием)	144 TB	1,728 TB	496 TB	5,952 TB
Скорость сканирования	24 GB/Sec	288 GB/Sec	14 GB/Sec	168 GB/Sec
Скорость загрузки	10 TB/Hour	120 TB/Hour	10 TB/Hour	120 TB/Hour

Бесшовная интеграция с инфраструктурой

EMC Data Domain
для резервного
копирования и
восстановления

Isilon Scale Out
для Больших Данных

EMC VMAX/VNX SAN Mirror
для повышенной
катастрофоустойчивости

EMC VMAX SRDF
EMC Data Domain
Replication
For Disaster Recovery

GREENPLUM

EMC²

Инновационные компании использующие Greenplum

Мощная партнерская экосистема

GREENPLUM®

EMC²

Преимущества модульной архитектуры

ПРОСТО

- All software infrastructure optimally pre-configured
- All hardware automatically monitored by EMC

ГИБКО

- Big Data Analytics need more than just database; it needs an agile platform
- Able to scale as dictated by requirements; In-place expansion; no forklifting needed

ЭФФЕКТИВНО

- Minimized time to value
- Built on best-of-breed commodity hardware
- Automated health monitoring and EMC Dial Home

Greenplum: это не только технология

- Data Science teams will become the driving force for success with big data analytics
- Greenplum is committed to the future of data science
  - University data science program collaboration with Stanford and UC Berkeley
  - Community investment including the Greenplum Analytic Workbench, Community edition software, and Data Science Summits
- Greenplum built its own Data Science practice
  - Leading PhDs with analytic tools expertise

Научный подход к работе с данными

Broad and Deep Experience to Assist Your Teams

Academic Affiliations:
Cal Berkeley
Princeton
Stanford
Australian National Univ.
Oxford
Courant Institute

Fields of Specialization:
Statistics
Applied Mathematics
Bayesian Analysis
Analytics Software
Quantitative Methods
Risk Analytics
Marketing Optimization
Stochastic Machine Learning
Healthcare Data Mining
Engineering

Experience:
Yahoo
DemandTec
Fox Interactive
eHarmony
Amazon

Academic Level:
Masters
PhD

С чего начать:
Greenplum Analytics Labs

- Packaged solutions that produce business value and actionable results
- Accelerate analytics capabilities on your data with your analysts
- Leverage the expertise of Greenplum's Data Scientists
- Establish a strategic vision for analytics development

Большое спасибо и ждем ваших заявок!