BI в "облаках"

Кудрявцев Юрий, mail@ykud.com

24 апреля 2009

- Год назад "облака" были еще не столь раскрученным термином
- Было много "сильных" утверждений о том, как использование этой технологии изменит BI
- Интересны как исходные тенденции, так и то, что произошло за год

Ключевые моменты

1. BI как услуга, пример Panorama + Google
2. Использование "облака" как платформы для MPP Хранилищ
3. Map/Reduce и (или) MPP ХД?
4. Использование MapReduce для расчета OLAP-кубов

BI как услуга, пример Panorama + Google

О компании Panorama

- Израильская компания – один из старейших BI вендоров
- Команда разработчиков Microsoft Analysis Services была куплена из Panorama
- После объявления о создании Microsoft PerformancePoint, Panorama оказалась в сложном положении

Panorama Analytics for Google Applications

- Добавление аналитических возможностей в Google Spreadsheets
- Аналог Pivot Tables в Microsoft Excel
- Нужно строить кубы, считать агрегаты, рисовать графики
- Для быстрого расчета кубов используется Google App Engine
- Есть возможность использовать Google Docs как клиента к Ms Analytical Services

http://www.panorama.com/google/

Panorama PowerApps

- Сервис, позволяющий использовать Panorama Analytics во внешних приложениях/сайтах
- Например, графический анализ на сайте учета личных финансов
- Аналитические кубы в таком случае собираются на "облаке" Google
- OLAP как сервис или OLAP 2.0

Что изменилось за год?

- Интеграция в Google Docs усилилась, увеличивается количество пользователей
- Не видно, чтобы PowerApps широко использовались

Использование "облака" как платформы для MPP Хранилищ

Что предоставляют вычислительные "облака"?

- Возможность взять в аренду любое количество серверов заданной конфигурации
- Забота о поддержке серверов ложится на провайдера "облака"
- Вы получаете ip сервера и root доступ. Можно заливать готовые образы ОС
- Дешево)

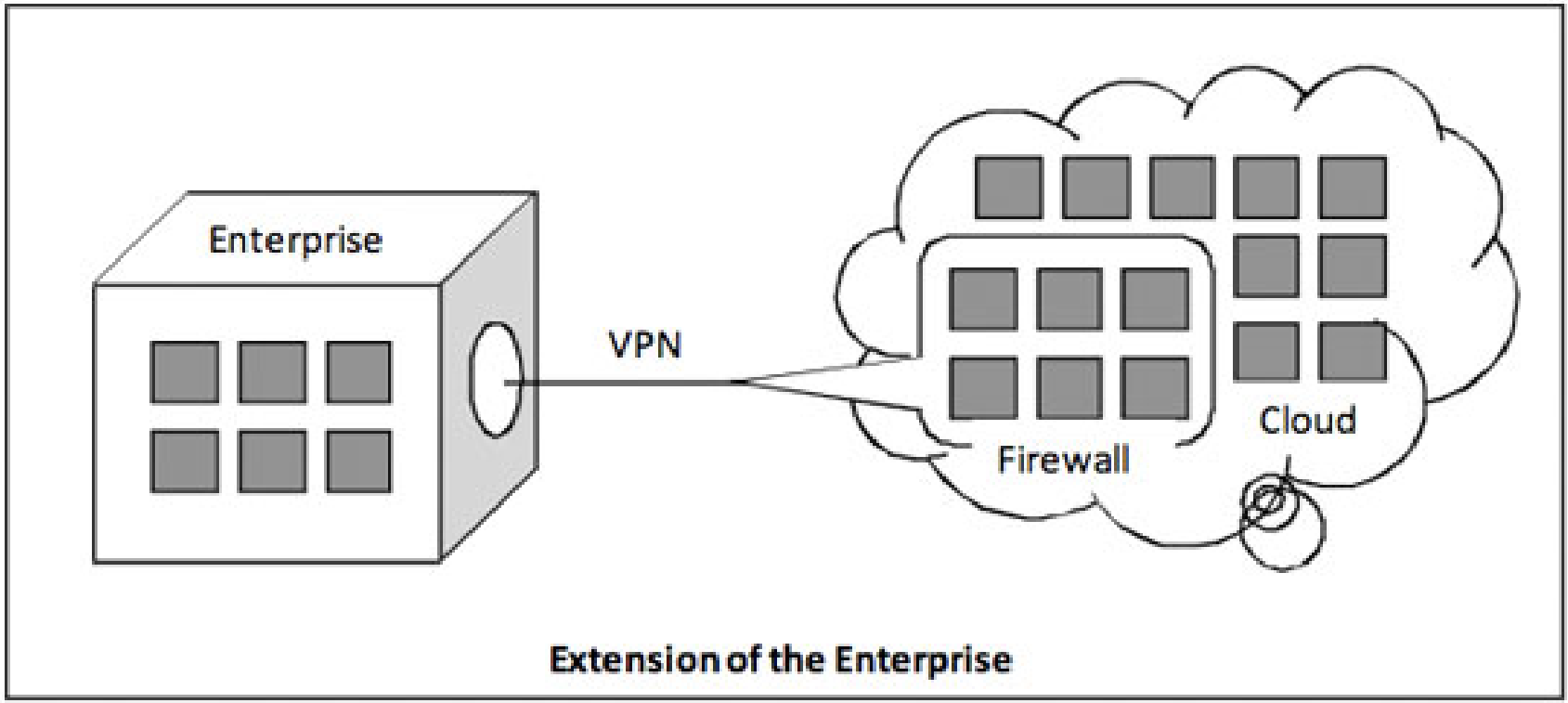
Зачем это нужно для MPP ХД?

- Удобство тестирования технологии
- Легкое масштабирование
- Отсутствие проблем с конфигурацией серверов
- Основные апологеты: Vertica, AsterData

Проблемы

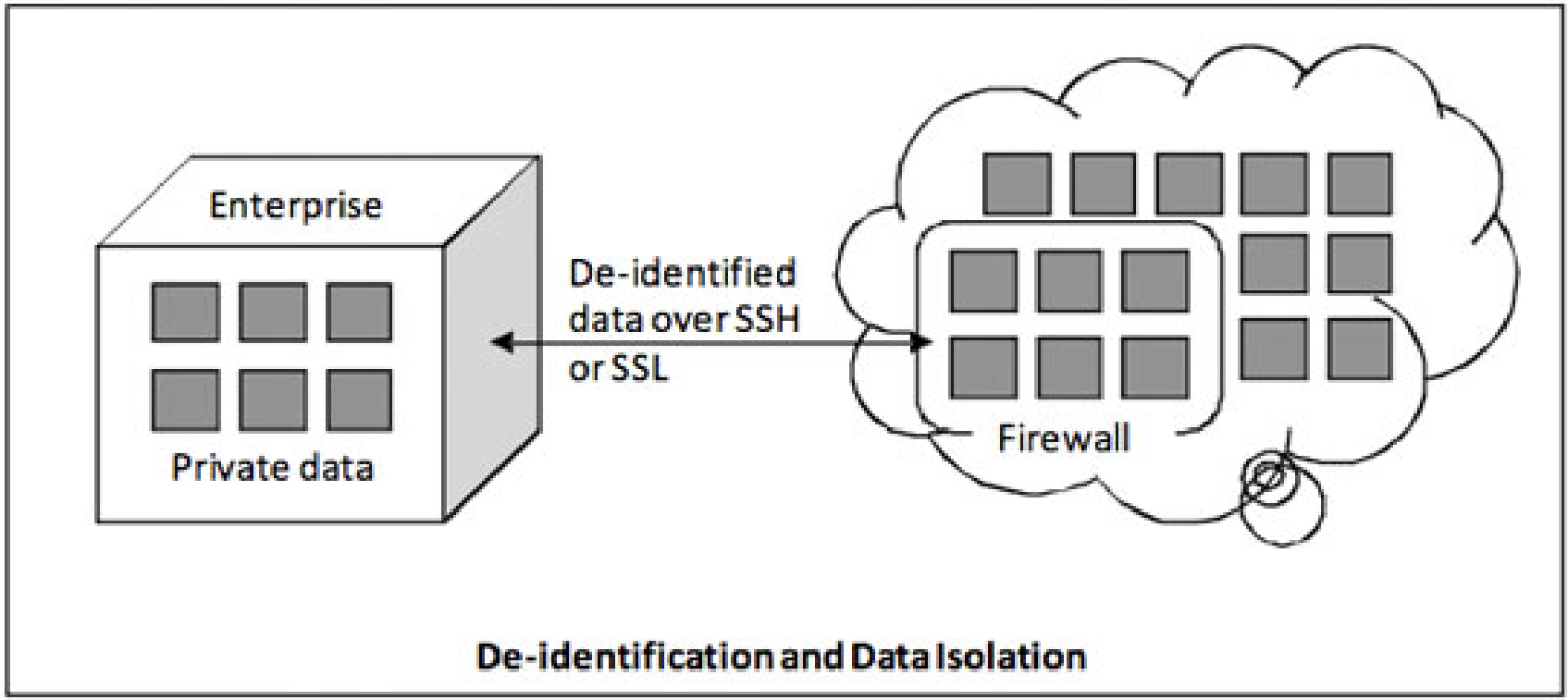
- Безопасность данных
- Надежность провайдера
- Перенос данных в "облако"

За этот год придуманы варианты решения: Добавление части облака в VPN



http://www.databasecolumn.com/2009/03/securing-your-data-in-the-cloud.html

Шифрование части "важных" данных перед joinom



Перенос данных в облако

Сервисы хранения данных – оплата за гигабайты данных, хранящихся на облаке.
Скорость импорта = скорости сетевого канала до провайдера.

Что изменилось за год?

- Увеличилось количество провайдеров "облаков": Amazon EC2, AppNexus, GoGrid, Rackspace Cloud
- У Vertica около 5ти клиентов в "облачном" варианте

Map/Reduce и (или) MPP ХД?

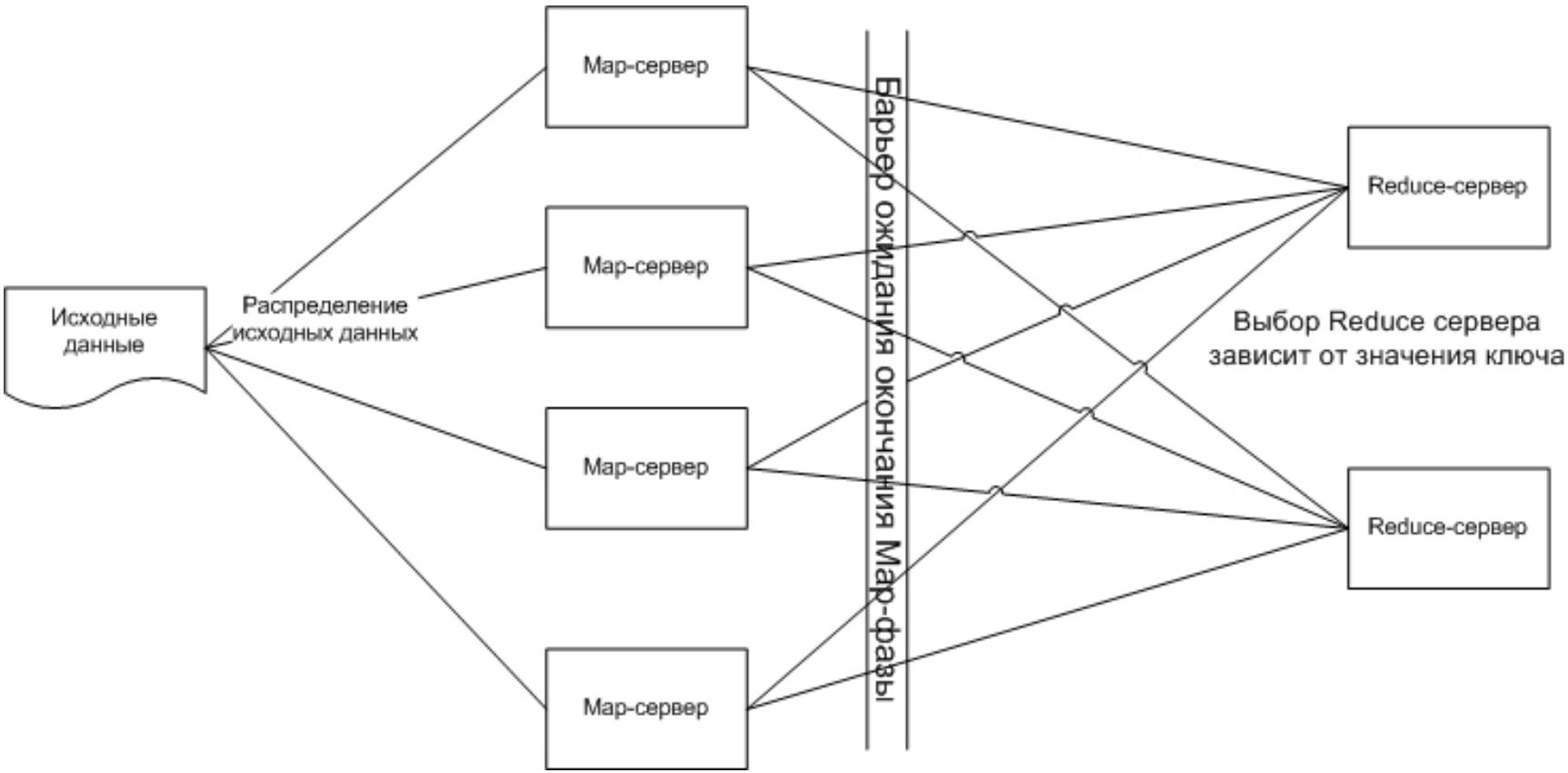
Что такое Map Reduce

Map/Reduce – вычислительная парадигма + программная платформа исполнения задач на кластере серверов, созданная в компании Google
Задача декомпозируется на 2х фазы – map и reduce (эквивалентно map и fold в функциональных ЯП)

- Map-процессы запускаются над подмножествами исходных данных и выполняются абсолютно независимо друг от друга.
- Reduce-процессы обрабатывают результаты map-фазы, разбивая их по значениям ключей на непересекающиеся блоки, что также позволяет выполнять их независимо.
- Таким образом, каждая из фаз может обрабатываться на любом количестве серверов параллельно.

Open-Source проект Map/Reduce платформы – Apache Hadoop

Схема Map Reduce



http://labs.google.com/papers/mapreduce.html

Пример Map-Reduce, Алгоритм подсчета слов в тексте

map – разбивает блок строк на слова, для каждого слова генерирует пару (слово, 1)
reduce – для каждого слова, складывает полученные 1, вычисляя таким образом количество повторений слова

Почему возникло противостояние M/R и MPP ХД

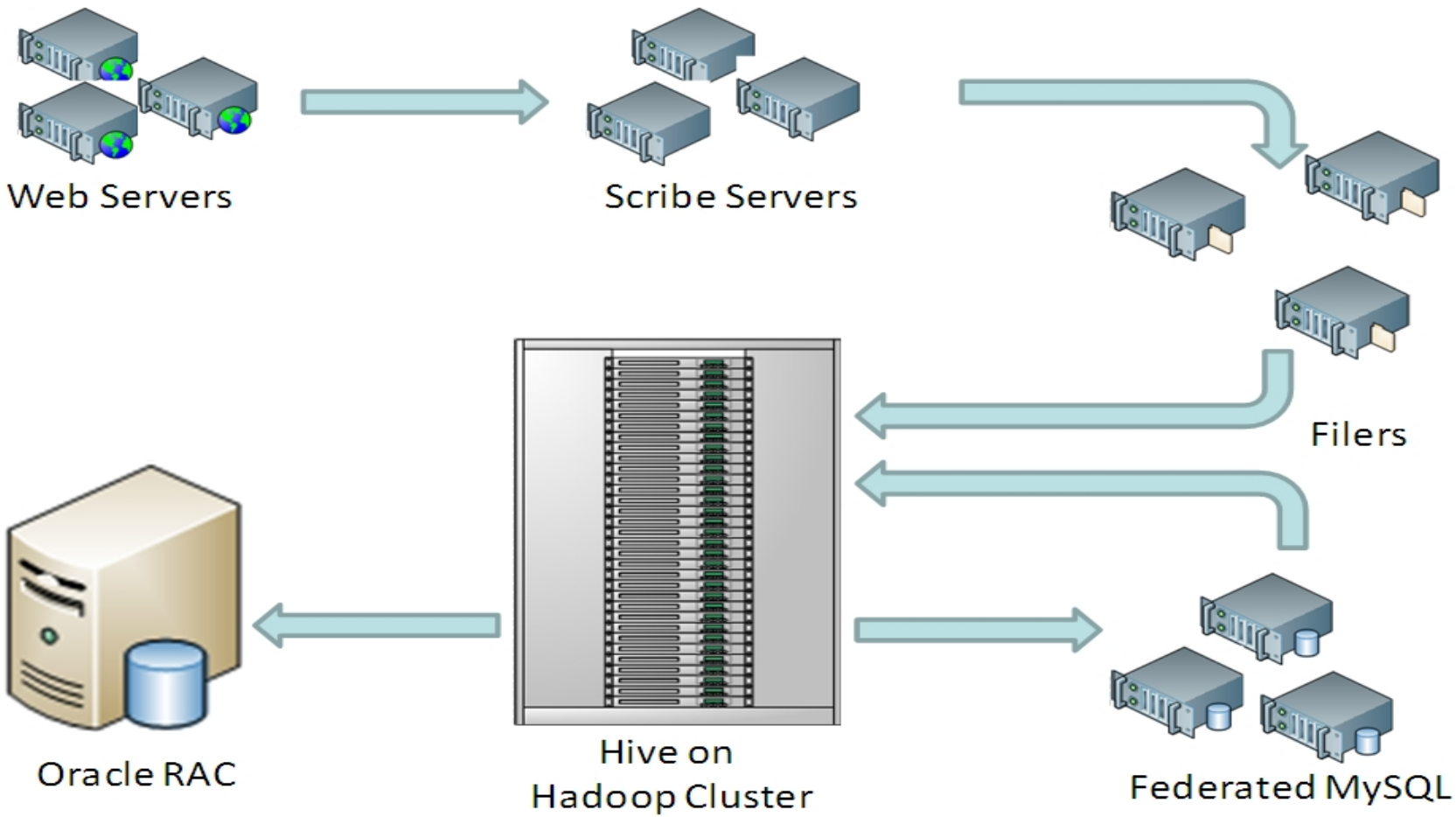
M/R используется для решения аналитических задач на сверхбольших объемах данных – та же область, что и MPP DWH.
2 замечательных позиции в дискуссии

- Map Reduce – платформа, чье существование обусловлено недостаточной грамотностью разработчиков, не знающих истории развития параллельных СУБД
- Map Reduce – средство работы с большими объемами, бесконечно масштабируемое и не имеющее ограничения в виде схем и типизации. Зачем нужны РСУБД?

Самый известный пример Map/Reduce аналитики - Hive

- Социальная сеть Facebook – десятки миллионов пользователей
- Использует Hadoop для анализа данных о пользователях
  - 2,5 Петабайта данных
  - В день добавляется 15 Тб данных
  - Таргетирование рекламы
  - Аналитические запросы
  - Инструментальные панели
  - Text Mining
- Написана специальная компонента – Hive, трансформирующая SQL запрос в Map Reduce задачу
- Hive теперь является одним из подпроектов Hadoop

Схема ХД Facebook



http://www.slideshare.net/jsensarma/hadoop-hive-talk-at-iitdelhi

Почему Facebook использует Hadoop:

- Цена
- Использование уже написаного кода(обработка текста на Python) для ETL
- Масштабируемость
- Гибкость схемы хранения

Все равно нужен Oracle RAC для онлайн запросов
http://www.dbms2.com/2009/04/15/
cloudera-presents-the-mapreduce-bull-case/

С другой стороны: MPP ХД в Ebay

- 5 Птб данных в Terradata
- 10ки тысяч серверов
- После тестирования Hadoop: меньше загружает процессора и требует больше серверов – дороже

Когда стоит использовать MPP ХД или Map/Reduce

Статья Стоунбрейкера сотоварищи:
A comparison of approaches to large-scale data analysis: MapReduce vs DBMS Benchmarks
http://database.cs.brown.edu/sigmod09/
Сравнение Hadoop, DBMS X (построчное MPP ХД) и Vertica (поколоночное MPP ХД) на ряде типичных задач:

- Поиск значения по текстовой маске
- Расчет агрегата
- Сложная UDF, подсчет количества внутренних ссылок в наборе HTML страниц

Выводы тестирования:

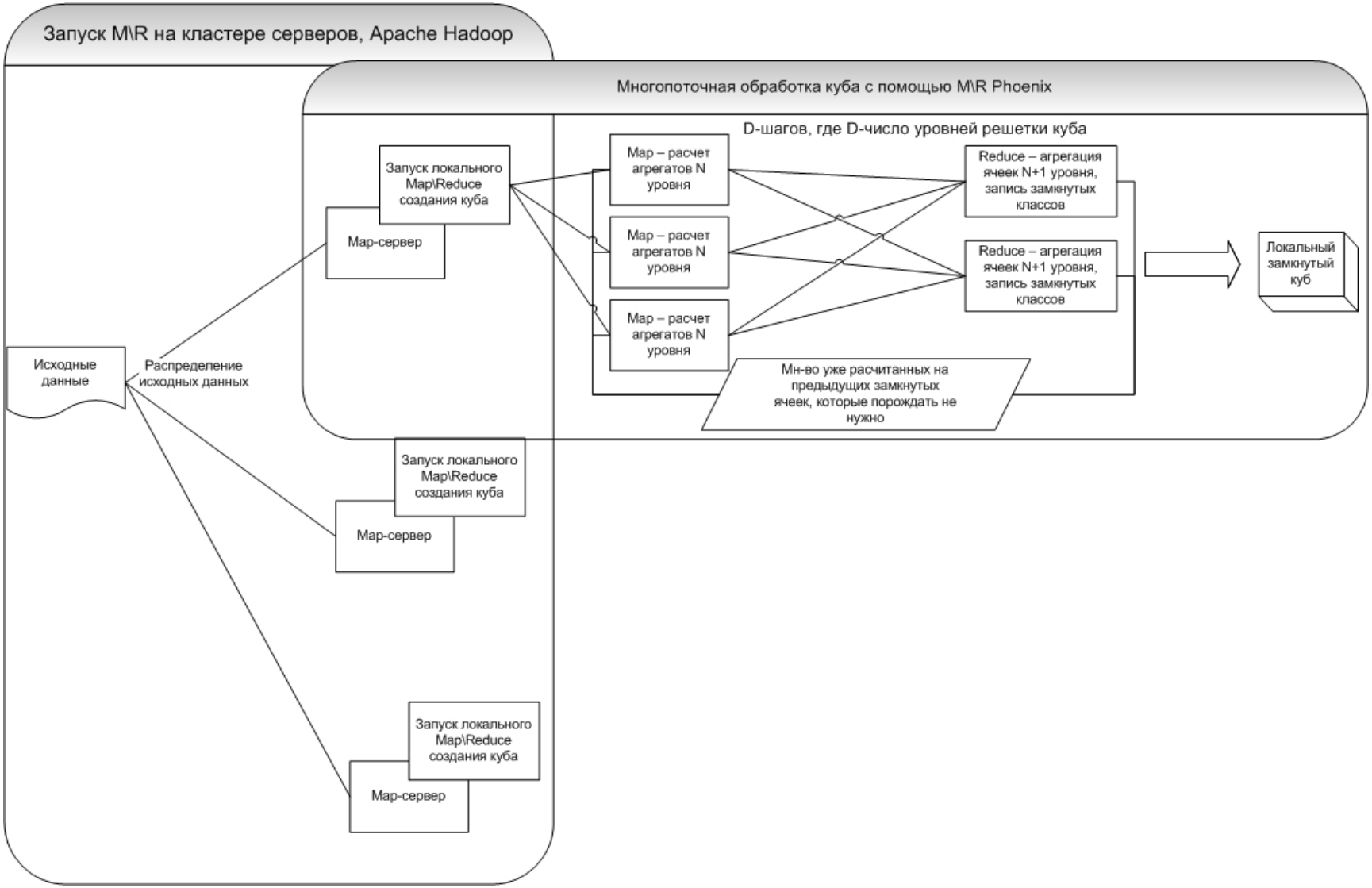
- Загрузка данных в Hadoop быстрее всего
- Поиск и агрегация – быстрее в DBMS X и кластере
- UDF быстрее в Hadoop
- Vertica быстрее всех на 100 узловом кластере)

Map/Reduce как средство расширения MPP ХД

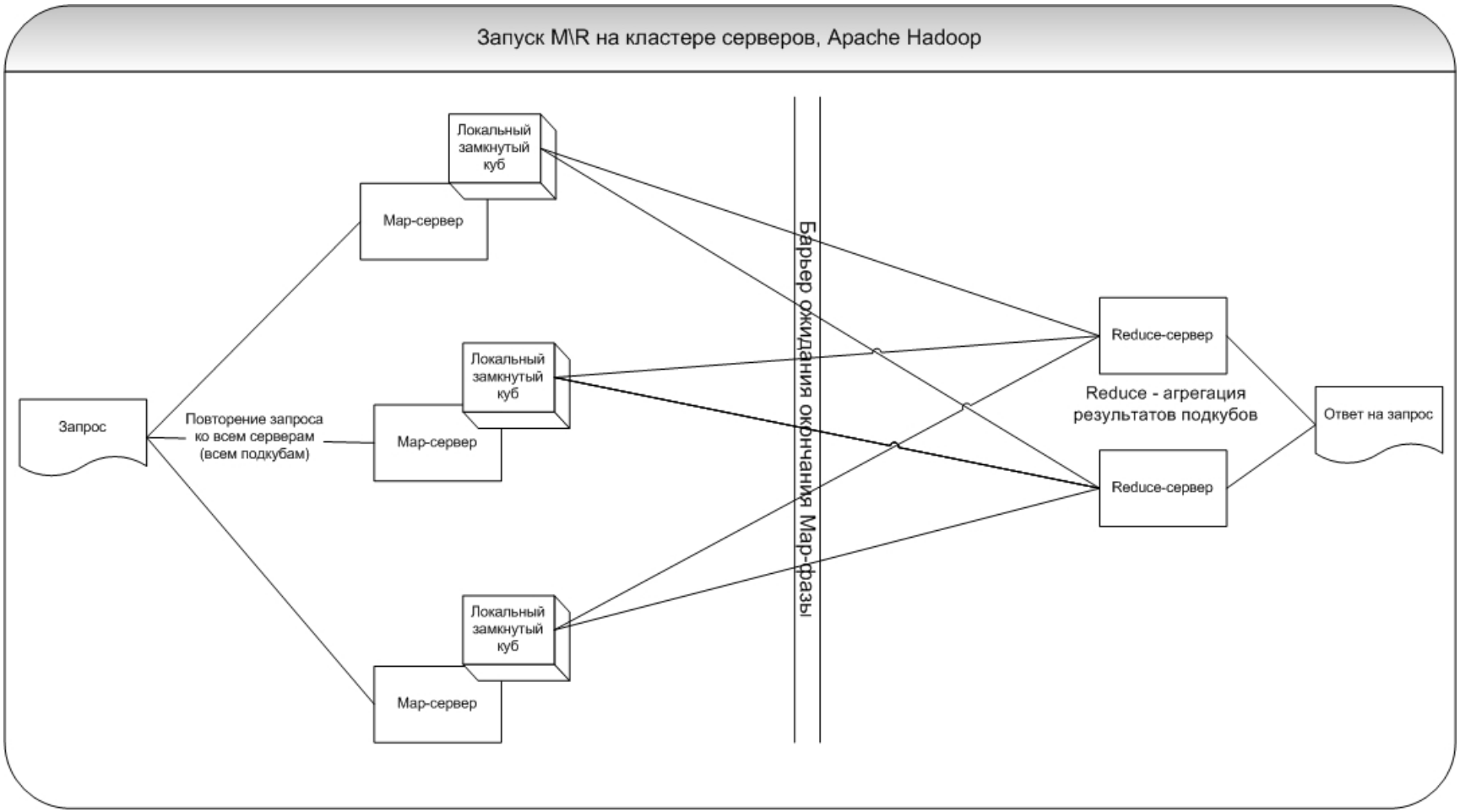
SQL/MR в ХД Aster Data – совмещение SQL и Map/Reduce в запросах к хранилищу
Часть задач в SQL, часть на удобном языке программирования в MR
http://www.asterdata.com/blog/index.php/2009/04/02/enterprise-class-mapreduce/

Пример использования M/R для аналитики. Создание OLAP кубов

Пример использования M/R для аналитики. Создание OLAP кубов



Пример использования M/R для аналитики.Схема ответов на запросы



- Не важно, какой OLAP-сервер использовать
- Скорость построения кубов растет линейно с уменьшением объема данных на узле
- Сложная задача создания OLAP-куба легко переписывается в MR

Спасибо! Вопросы?