

Большие Данные с точки зрения MDM

Сергей Кузнецов

Генеральный Директор Informatica Россия

Руководитель Центра Разработки

Большие Данные с точки зрения MDM

❖ Подходы к работе с Большими Данными, Hadoop

❖ Управление Большими Данными –

Master Data Management (MDM)

❖ Informatica MDM для Больших Данных

Традиционный подход

Больше Памяти

Быстрее Процессор



Мир Данных

Корпорации - терабайты/день

Facebook = 15Pb, eBay = 5Pb



Доставка данных для процессоров



Новый Подход

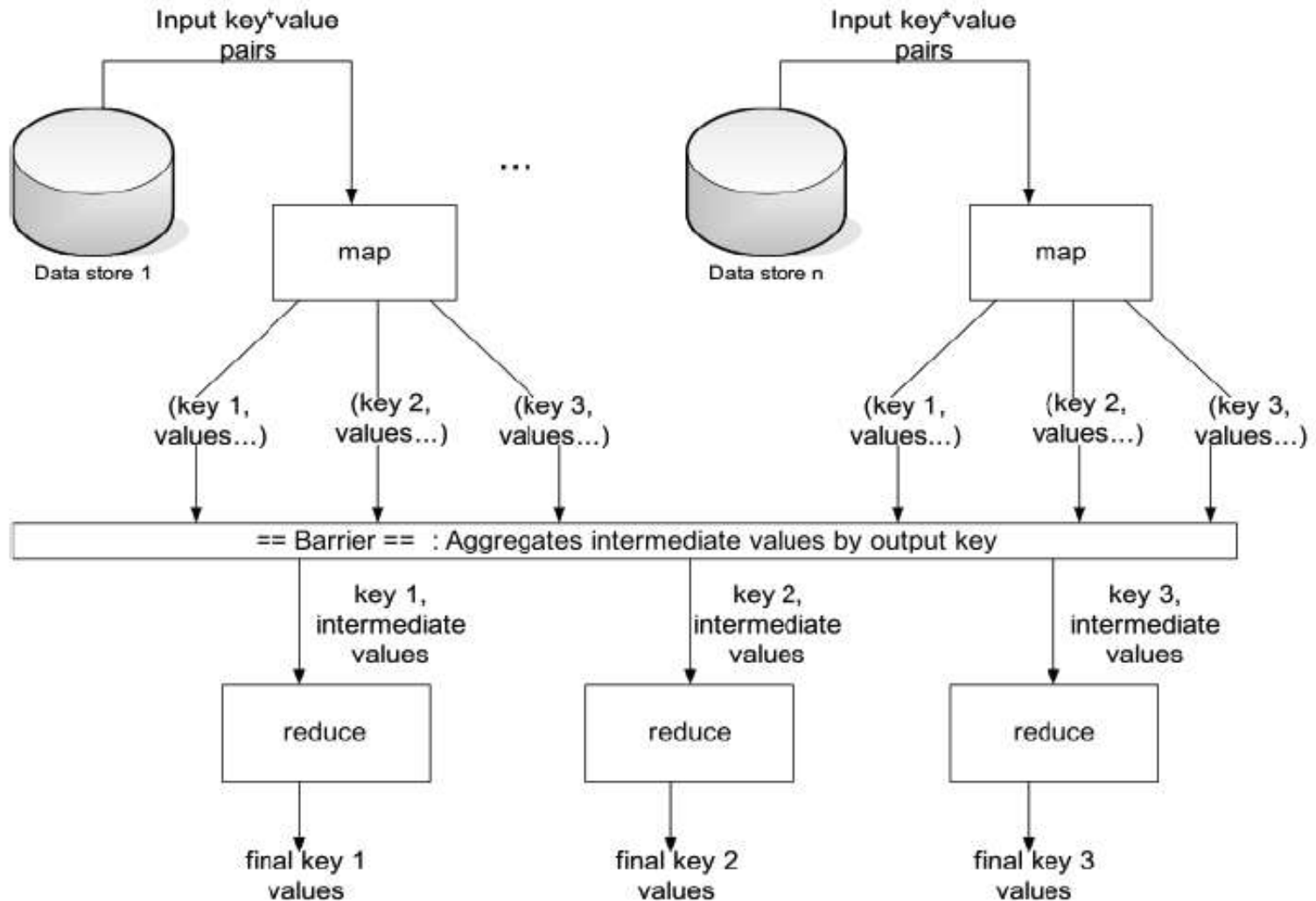




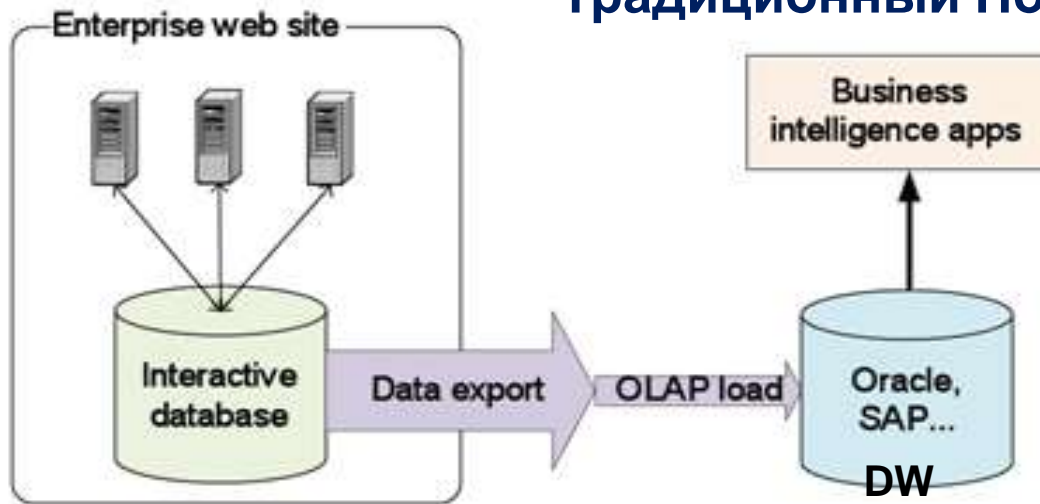
Проект Hadoop

- **Общедоступный проект (Apache), свой вклад внесли компании Yahoo!, Facebook, Cloudera**
- **Состоит из двух основных компонент –**
 - HDFS (The Hadoop Distributed File System) – хранение данных на кластере дублирующими блоками 64/128 Mb
 - MapReduce – распределенные вычисления среди узлов кластера
- **Экосистема Hadoop**
 - Pig, Hive – оболочки для использования традиционных SQL запросов
 - HBase - База Данных для хранения больших данных и широких таблиц, имеет ограниченную модель доступа
 - Oozie, Sqoop, HUE, Flume и т.д.

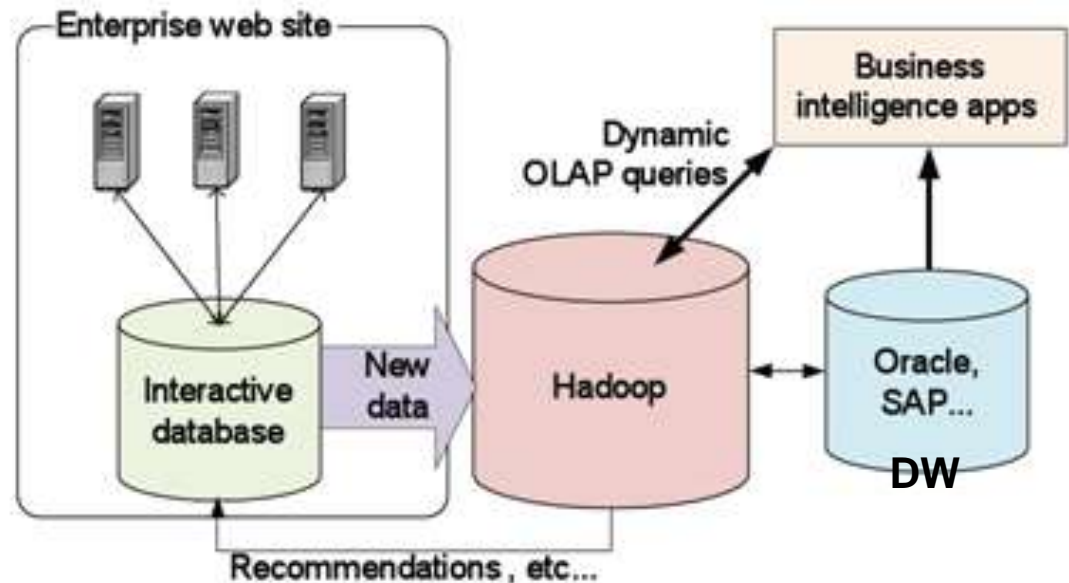
Hadoop MapReduce –



Традиционный Подход



Использование Hadoop для Интерактивных Данных



Большие Данные с точки зрения MDM

❖ Подходы к работе с Большими Данными, Hadoop

❖ Управление Большими Данными –

Master Data Management (MDM)

❖ Informatica MDM для Больших Данных

Получения Достоверного Источника Данных



Управление мастер-данными

- Целостность, Расширяемость,
- Консолидация данных

Качество Данных

- Точность, Очищение данных

Интеграция Данных

- Доступ к данным из любых источников
- Репликация, защита и маскирование данных

Управление Данными
Полная интеграция в существующую инфраструктуру приложений, процессов и пр

Продукты компании Informatica



**Data
Integration**



**Data
Quality**



**Master Data
Management**



INFORMATICA
The Data Integration Company™
PowerCenter



INFORMATICA
The Data Integration Company™
Data Quality

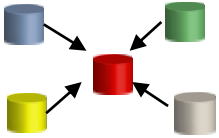
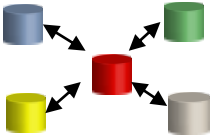


INFORMATICA
The Data Integration Company™
MDM

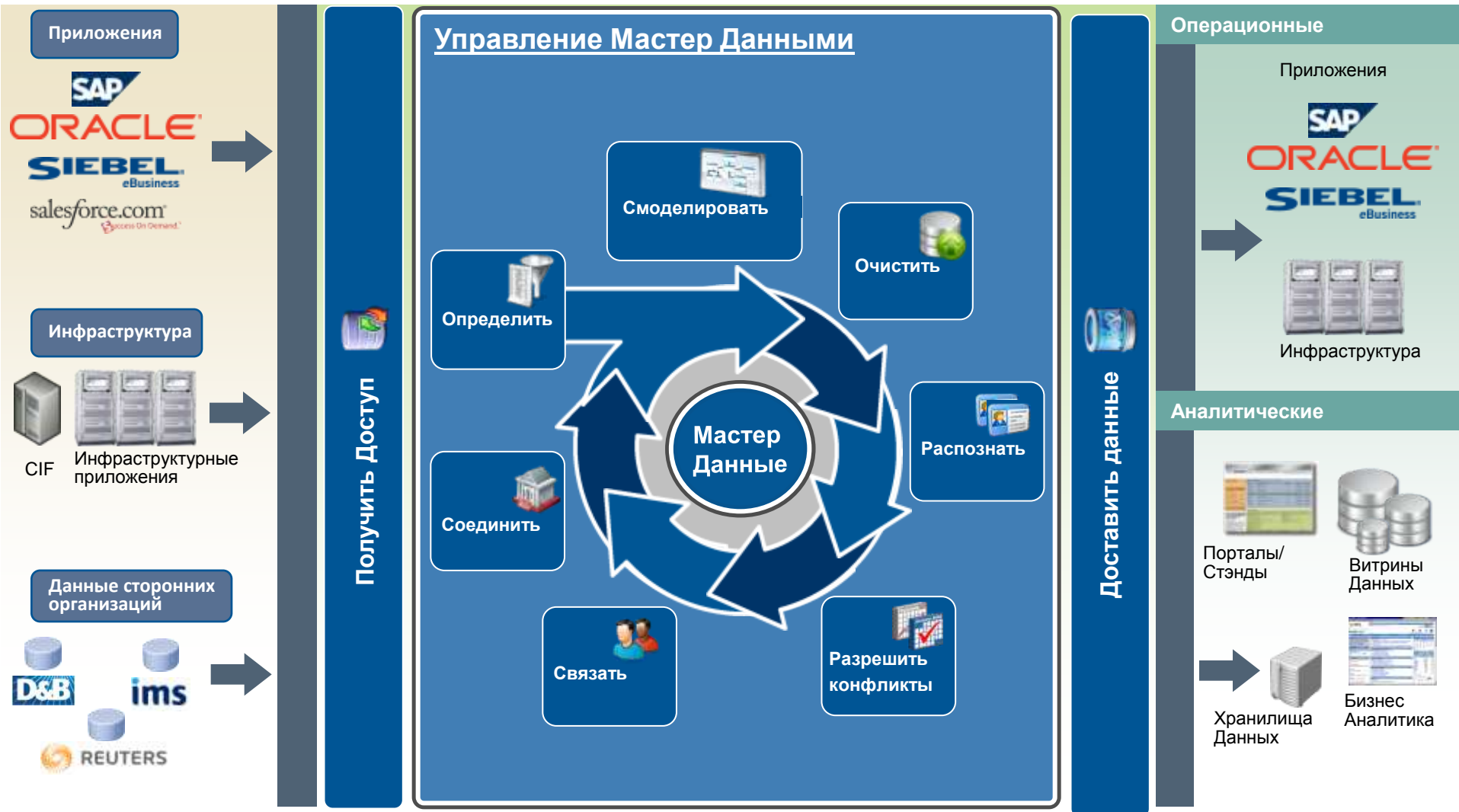


Единый репозиторий метаданных (правила, библиотеки и пр.)

























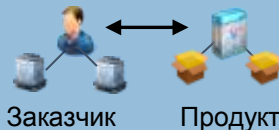
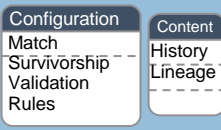








MDM – различные подходы


	Единый <u>Образ</u> <u>Данных</u>	Единый <u>Источник</u> <u>Данных</u>		
	Реестр	Консолидация	Синхронизация	Централизация
Стиль				
Процесс	Реестр идентификаторов-указателей на исходные источники данных	Мастер-данные собраны в Хабе, далее - синхронизация с Хранилищем Данных	Мастер-данные собраны в Хабе, синхронизация с исходными приложениями – источниками данных	Мастер-данные распределяются в приложения из Хаба
Обработка Данных	Реальное время	Пакетное	Пакетное <u>и</u> Реальное время	Пакетное <u>и</u> Реальное время
Направление потока данных	Одностороннее	Одностороннее	Двустороннее	Одностороннее


Как Informatica решает задачу получения мастер-данных?





Informatica MDM – основные возможности

Доставка Данных 	Синхронизация 	Вывод данных через API 	Бизнес Транзакции 
Управление данными 	Мониторинг KPI 	Аналитика Данных 	Бизнес-процессы 
Построение Связей 	Контр- агент 	Продукт 	Контр-агент & Продукт 
Разрешение Конфликтов 	Соединение 	Функции Доверия 	Разъединение 
Распознавание 	Deterministic & Fuzzy Logic 	Интернационализация 	
Очистка 	Очистка Данных 	Стандартизация адресов 	Открытая Архитектура 
Модель Данных 	Один или Несколько Доменов 	Метаданные 	
Определение 	Профилирование 	Анализ 	Распознавание 
Получение Доступа 	Пакетное Реальное время 	Любые Источники Данных 	Любые Форматы 

Сервисы
Данных 

Управление
Мастер
Данными 

Качество
Данных 

Интеграция
Данных 

Большие Данные с точки зрения MDM

- ❖ Подходы к работе с Большими Данными, Hadoop
- ❖ Управление Большими Данными –

Master Data Management (MDM)

❖ Informatica MDM для Больших Данных

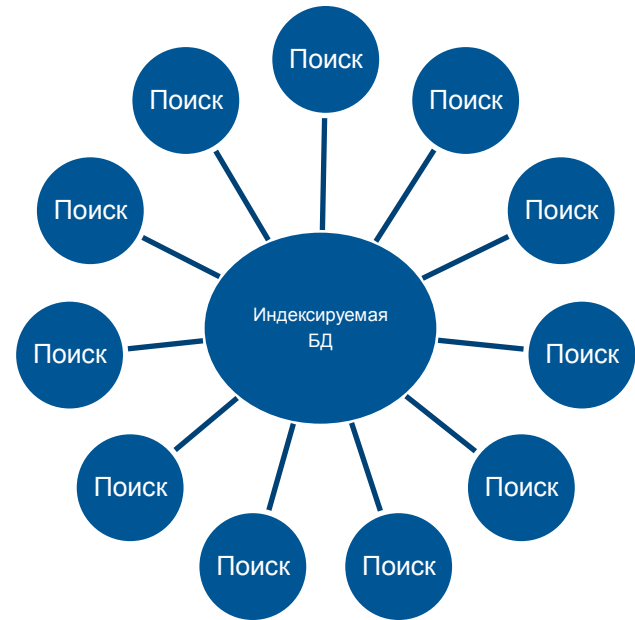
Продукты Informatica и Hadoop

- **Общие цели, дополняют друг друга**
 - Большие Данные
 - Расширяемость, Надежность, Доступность, Переносимость
 - Снижение стоимости хранения информации
- **Управление Большими Данными и MDM**
 - Большие данные везде (Volume, Velocity, Variety)
 - Сбор транзакционных данных – традиционных, соц сетей и пр
 - Группирование данных больших объемов – в силу множественности является задачей Больших Данных

Что меняется?

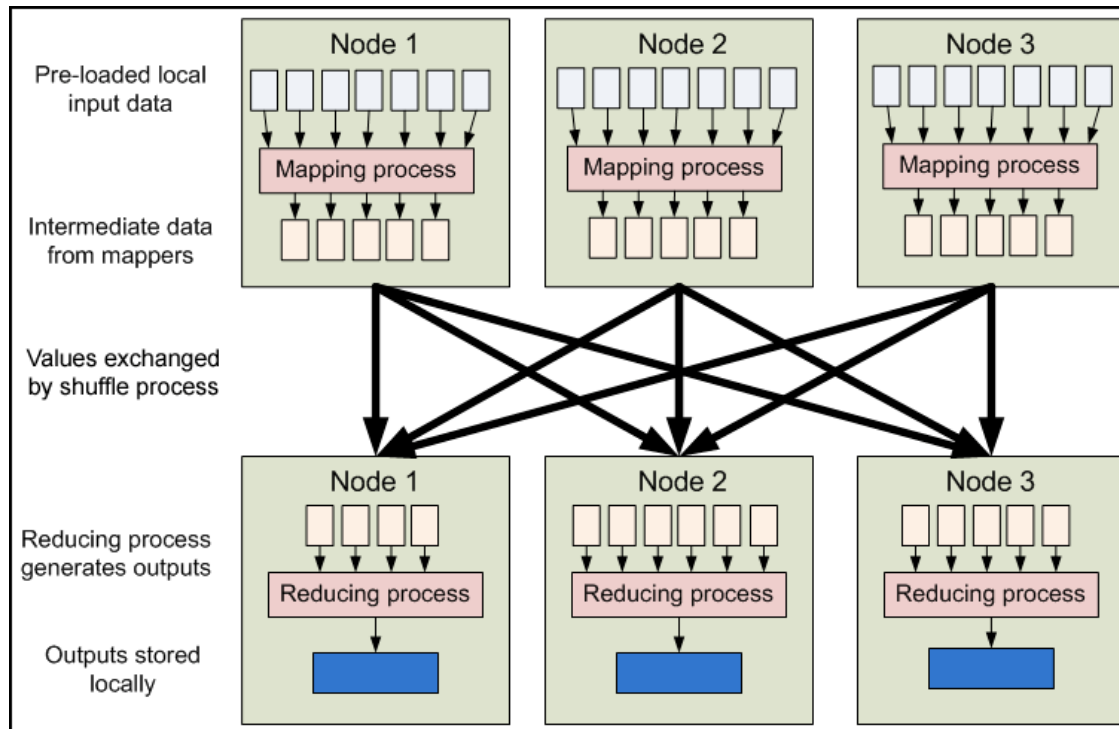
Традиционный Подход

- Основан на БД
- Несколько потоков используют поиск в индексируемой БД
- Издержки – Индексы БД в случае чтения и записи одного фрагмента данных

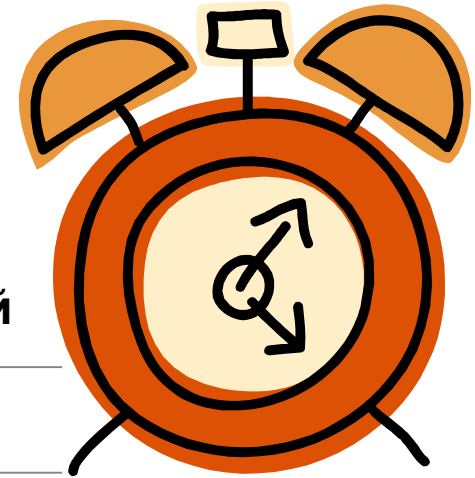


Что меняется?

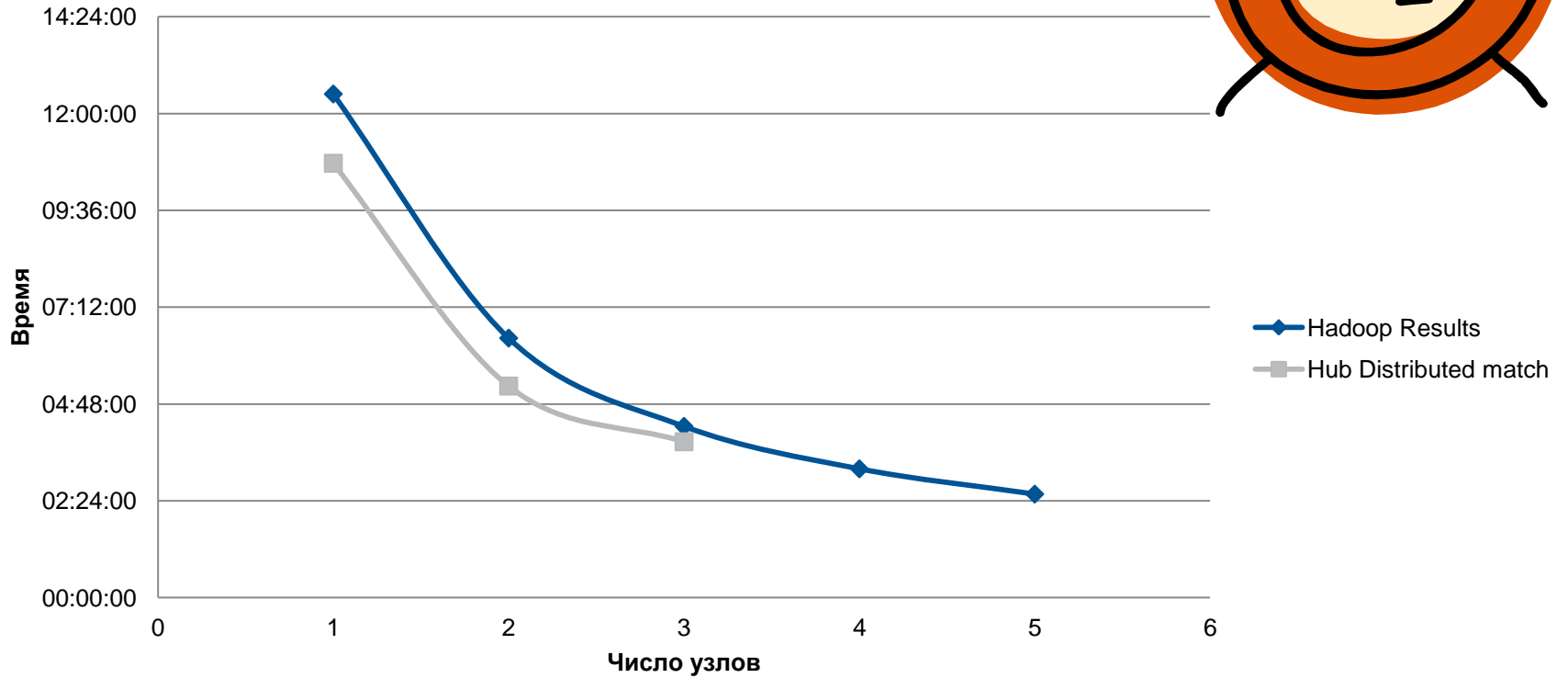
- Hadoop решает эту задачу на уровне данных
- Издержки – обмен данными между узлами кластера



Результаты

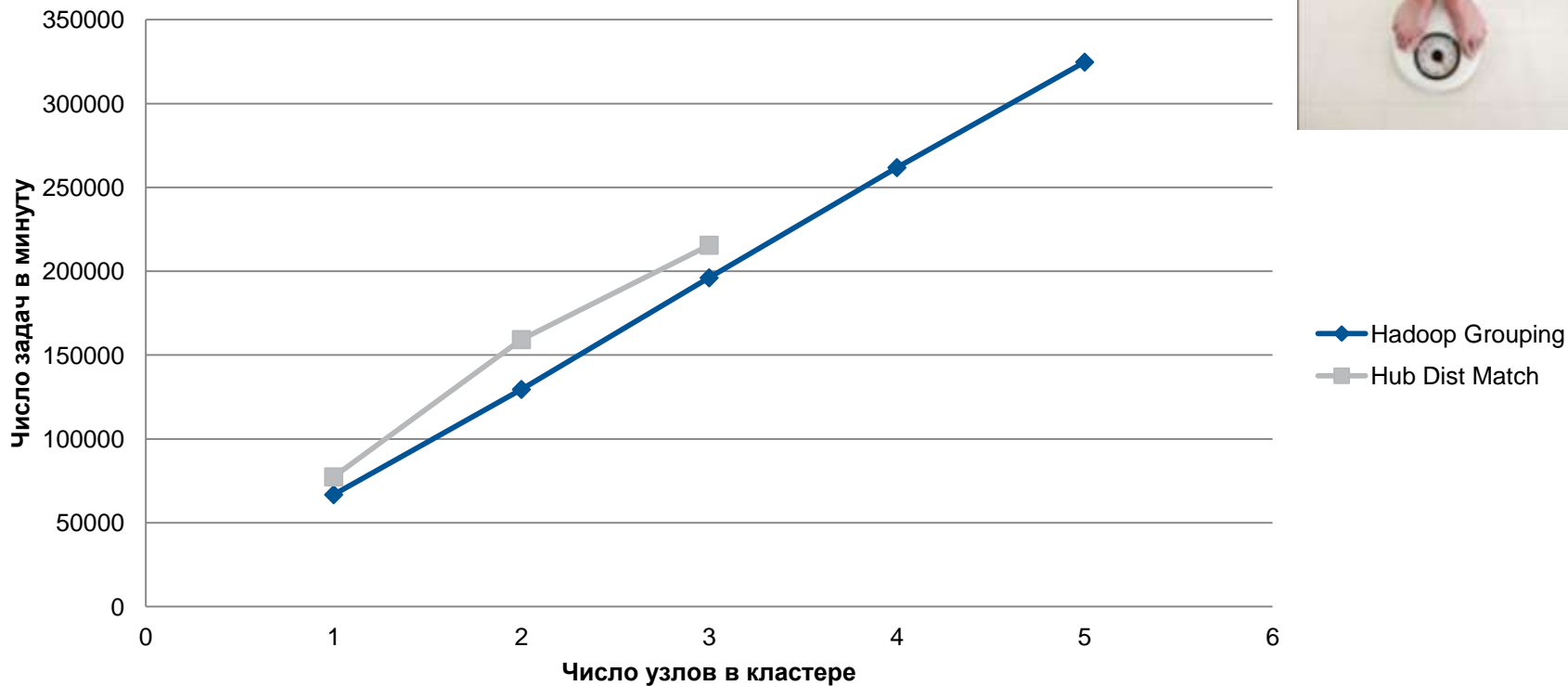


Результаты Hadoop на 50М записей



Тесты на расширяемость

Расширяемость 50М



Informatica в России

- **Informatica Россия & СНГ (офис продаж)**
 - Смоленский Пассаж, 6й этаж
 - Смоленская пл. д.3
 - 121099 Москва, Россия
 - Тел +7(495) 771-7150
 - Email: info-ru@informatica.com
- **Informatica R&D Центр**
 - Средний пр 88А, 7й этаж
 - 199106 С-Петербург, Россия
 - Тел +7(812) 320-9143

Вопросы?

