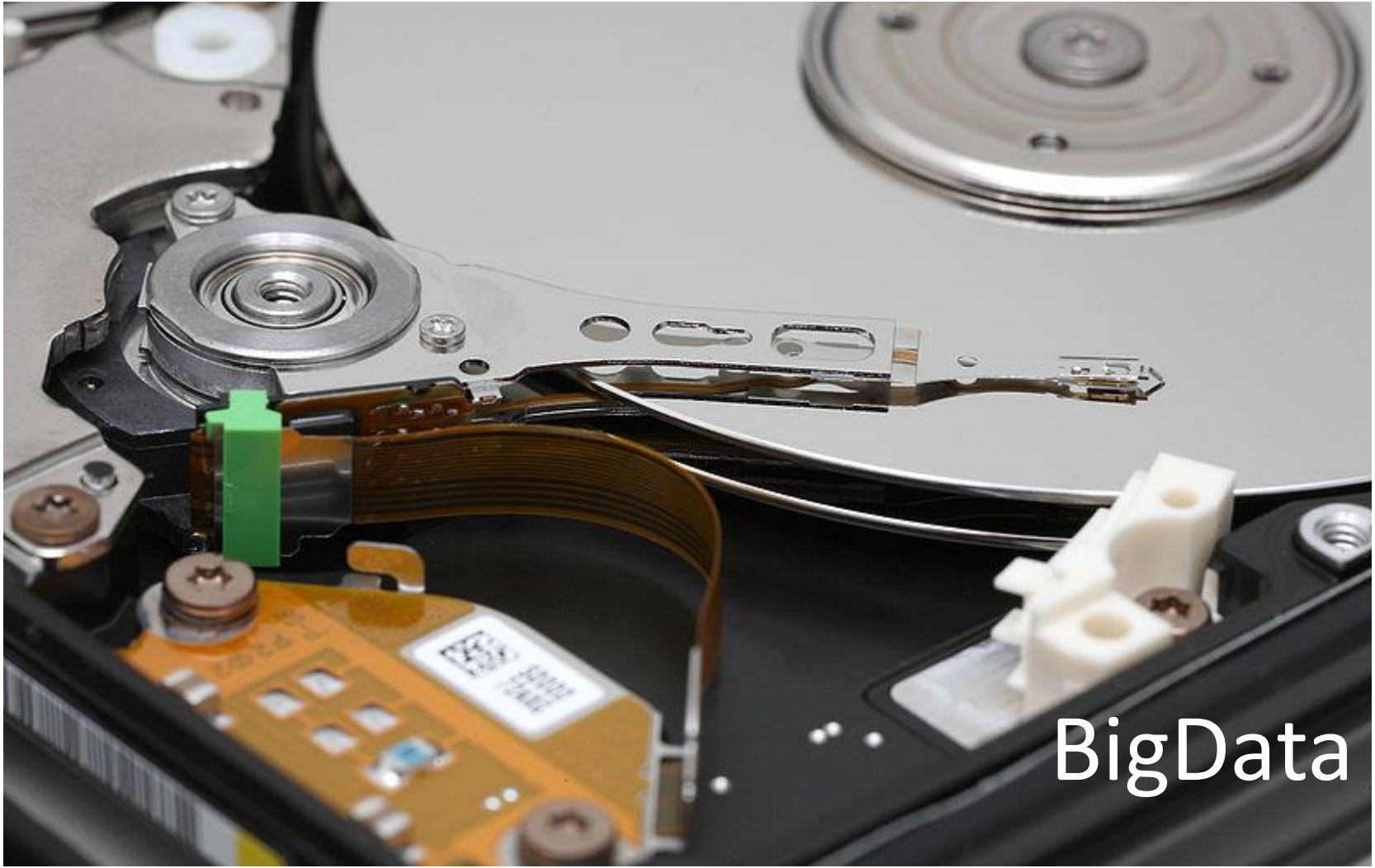
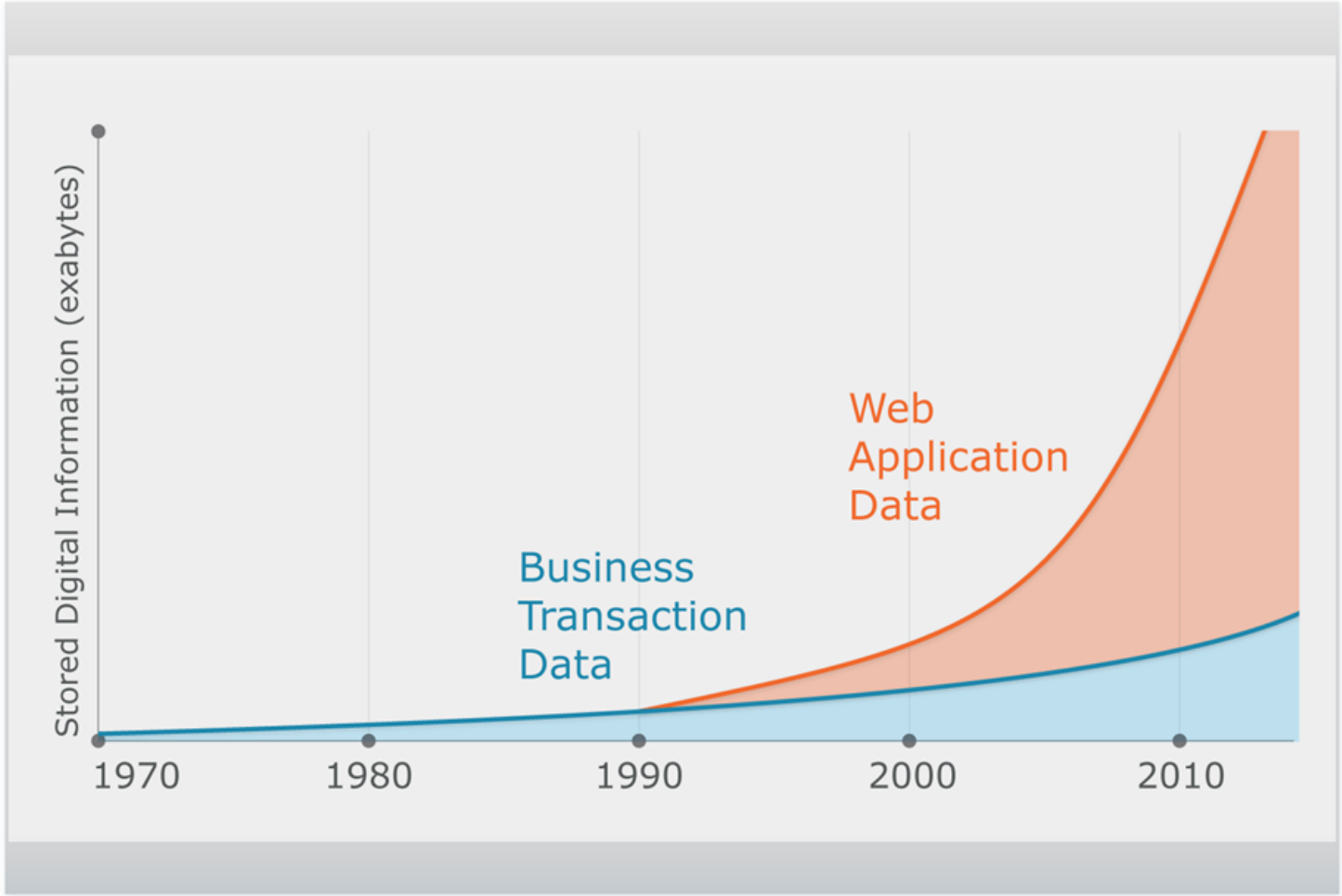

BigData


Stored Digital Information (exabytes)
Web
Application
Data
Business
Transaction
Data
1970
1980
1990
2000
2010

Общий размер данных (2010г)*:

1.2 Зеттабайт (1.2 Триллиона Гигабайт)

* По данным Information Data Corporation (IDC)

Обрабатывает 20 PB в день (2008)
Скачивает 20B веб-страниц в день (2012)

>10 PB данных,
75B DB (6/2012)

>100 PB польз. данных +
500 TB/день (8/2012)

S3: 449B объектов (7/2011)

JPMorganChase

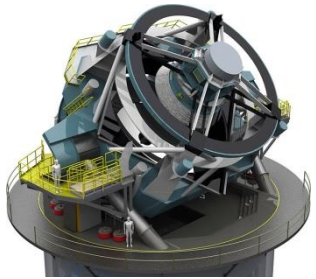
150 PB на 50k+ серверов
работает 15k apps (6/2011)

INTERNET ARCHIVE

Wayback Machine: 240B веб-страниц в архиве, 5 PB (1/2013)

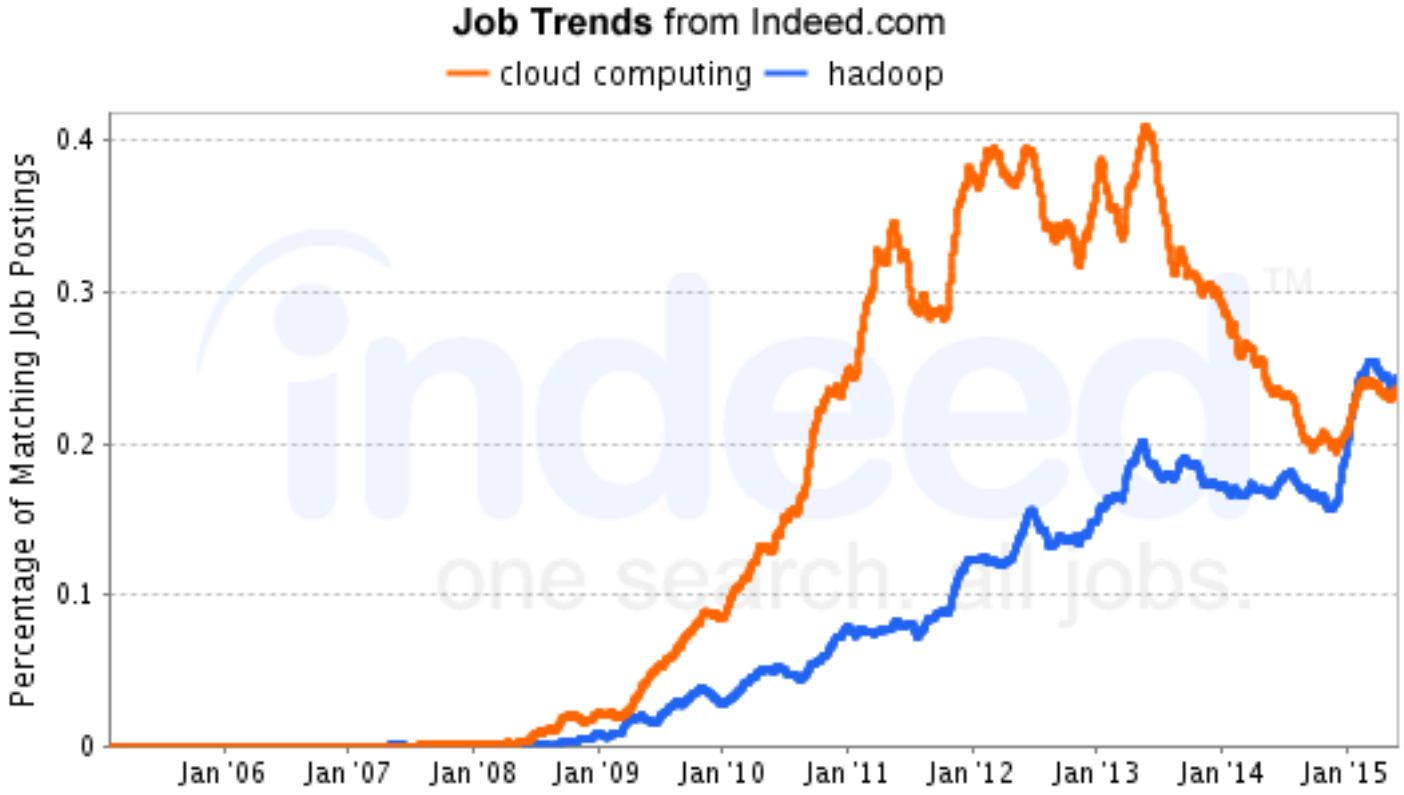
LHC: ~15 PB в год

CERN



LSST: 6-10 PB в год
(~2015)

Много данных – это сколько?


Job Trends from Indeed.com
cloud computing
hadoop
Percentage of Matching Job Postings
0.4
0.3
0.2
0.1
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
Jan '14
Jan '15

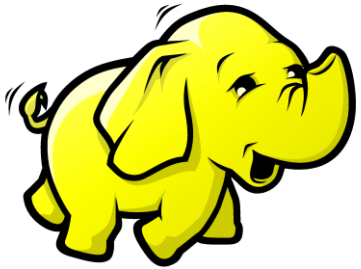
Top Job Trends:

1. HTML5
2. MongoDB
3. iOS
4. Android
5. Mobile app
6. Puppet
7. Hadoop
8. jQuery
9. PaaS
10. Social Media


hadoop

- Проект "The Apache™ Hadoop™“ разрабатывает open-source ПО для отказоустойчивых, масштабируемых и распределенных вычислений
- Особенности:
  - Работает с BigData на обычных серверах
  - Сильное open-source комьюнити
  - Много различных продуктов и средств используют Hadoop

История Hadoop



- Начинался как подпроект в Apache **Nutch**
  - **Nutch** – это открытый Web Search Engine
  - OpenSource альтернатива Google
  - Начинал его **Doug Cutting**
- В **2004** году Google публикует статьи про GFS и MapReduce
- **Doug Cutting** и команда Nutch реализовала свой фреймворк на основе этих статей
- В **2006** Yahoo! Нанимает **Doug Cutting** для работы над Hadoop в своей команде
- В **2008** Hadoop становится Apache Top Level Project
  - http://hadoop.apache.org

Кто использует Hadoop

Хранение данных

- Емкость дисков выросла экспоненциально, в отличии от скорости чтения
  - 1990
    - Емкость 1400 Мб
    - Скорость чтения 4.5 Мб/сек
    - Чтение всего диска за ~5 мин
  - 2010
    - Емкость 1Тб
    - Скорость чтения 100 Мб/сек
    - Чтение всего диска за ~3 часа

- Hadoop:
  - 100 HDD одновременно могут прочитать 1Тб данных за 2 мин

Кластер Hadoop

- “Дешевое” обычное железо:
  - Не суперкомпьютеры
  - Не десктопы
- Соединенное по сети
- Расположено в одном месте
  - Сервера в стойках в датацентре

Системные принципы Hadoop

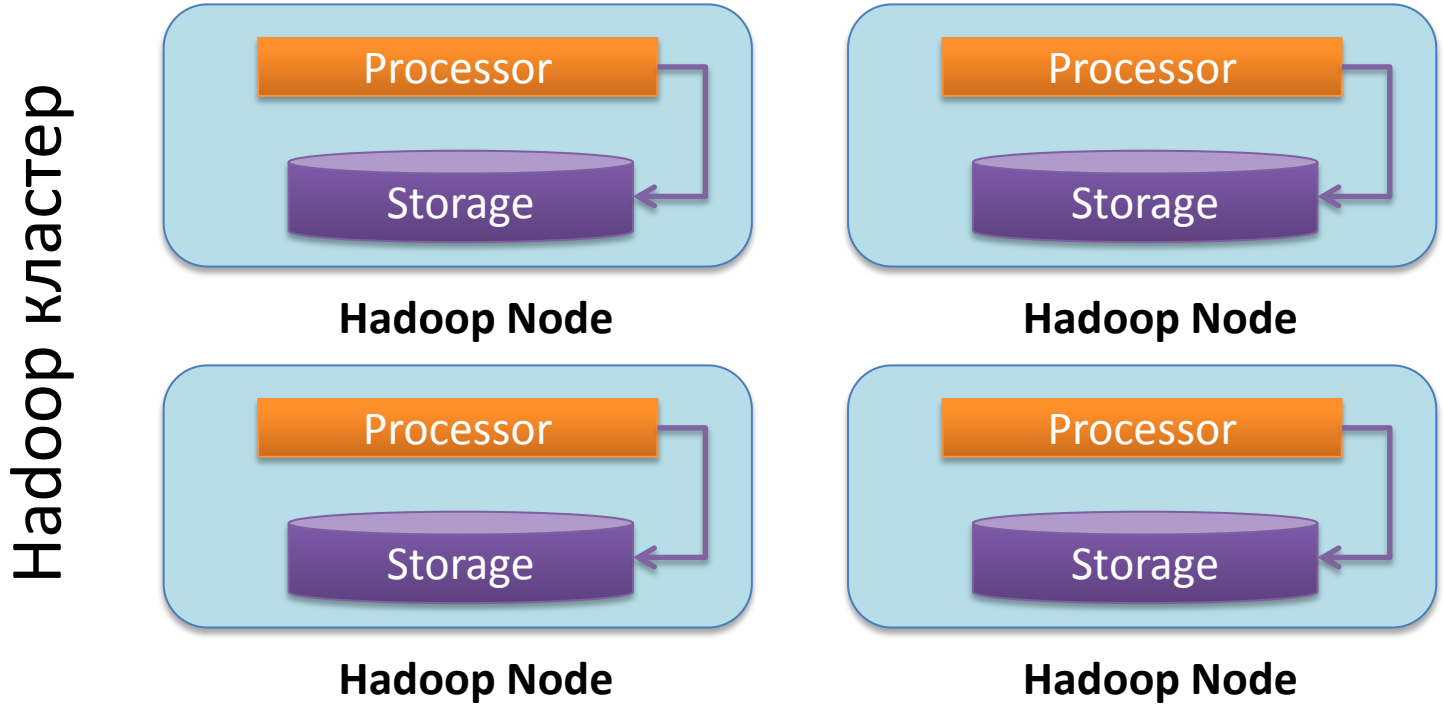
- Горизонтальное (Scale-Out) масштабирование вместо вертикального (Scale-Up)
- Отправляем код к данным
- Уметь обрабатывать падения нод и отказы оборудования
- Инкапсуляция сложности работы распределенных и многопоточных приложений

Масштабирование

- Вертикальное:
  - Добавить дополнительные ресурсы к существующему железу (CPU, RAM)
  - Если нельзя улучшить железо, то надо покупать более мощное новое
  - Закон Мура не успевает за ростом объема данных
- Горизонтальное:
  - Добавить больше машин к существующему кластеру
  - Приложение поддерживает добавление/удаление серверов
  - Просто масштабироваться “вниз”

Данные к коду

Код к данным


Hadoop кластер
Processor
Storage
Hadoop Node
Processor
Storage
Hadoop Node
Processor
Storage
Hadoop Node
Processor
Storage
Hadoop Node

Отказы оборудования

- Чем больше количество машин, тем чаще будут отказы железа
- Hadoop разрабатывался с учетом отказов железа
  - Репликация данных
  - Перезапуск тасков

Инкапсуляция сложности реализации

- Hadoop скрывает многие сложности распределенных и многопоточных систем
- Освобождает разработчика от заботы о проблемах системного уровня
  - Race conditions, ожидание данных
  - Организация передачи данных, распределение данных, доставка кода и т.д.
- Позволяет разработчику фокусироваться на разработке приложения и реализации бизнес-логики


Экосистема Hadoop

Экосистема Hadoop

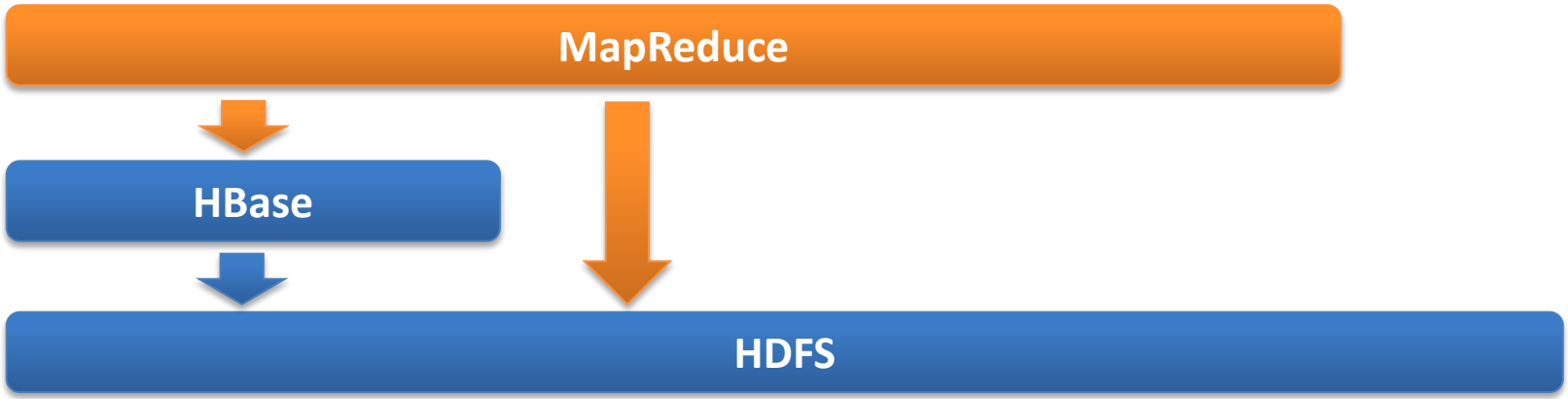
- Главные компоненты Hadoop:
  - HDFS: Hadoop Distributed FileSystem
  - MapReduce: Фреймворк распределенной обработки данных
- Другие компоненты:
  - Hbase: Column-oriented DB, поддержка последовательного и произвольного чтения, поддержка простых запросов
  - Zookeeper: Highly-Available Coordination Service
  - Oozie: Диспетчер задача для Hadoop
  - Pig: Язык обработки данных и среда выполнения
  - Hive: Data warehouse с SQL интерфейсом

Экосистема Hadoop

- Для разработки приложения необходима файловая система
  - В Linux: ext3 и ext4
  - В мире Hadoop обычно Hadoop Distributed File System (HDFS)
- Также нужен удобный интерфейс для работы с данными
  - Реляционная СУБД поверх локальной файловой системы
  - Hbase: Это key/value хранилище, реализованное поверх HDFS

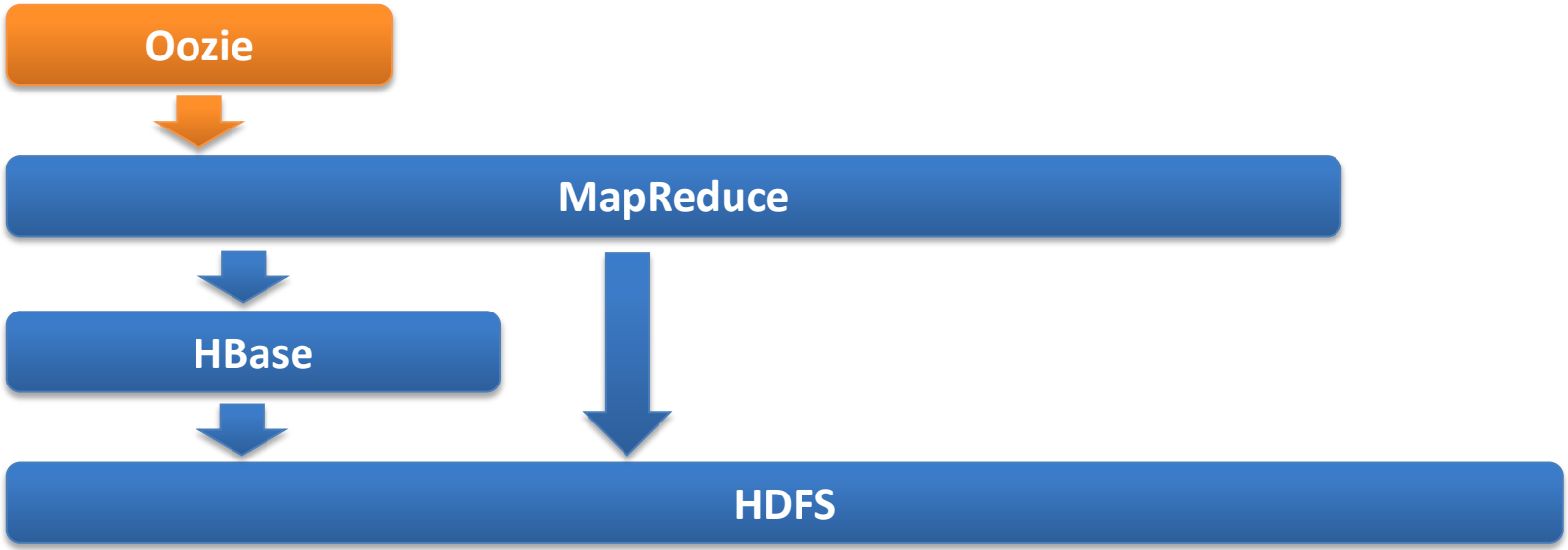
Экосистема Hadoop

- Фреймворк для запуска MapReduce задач



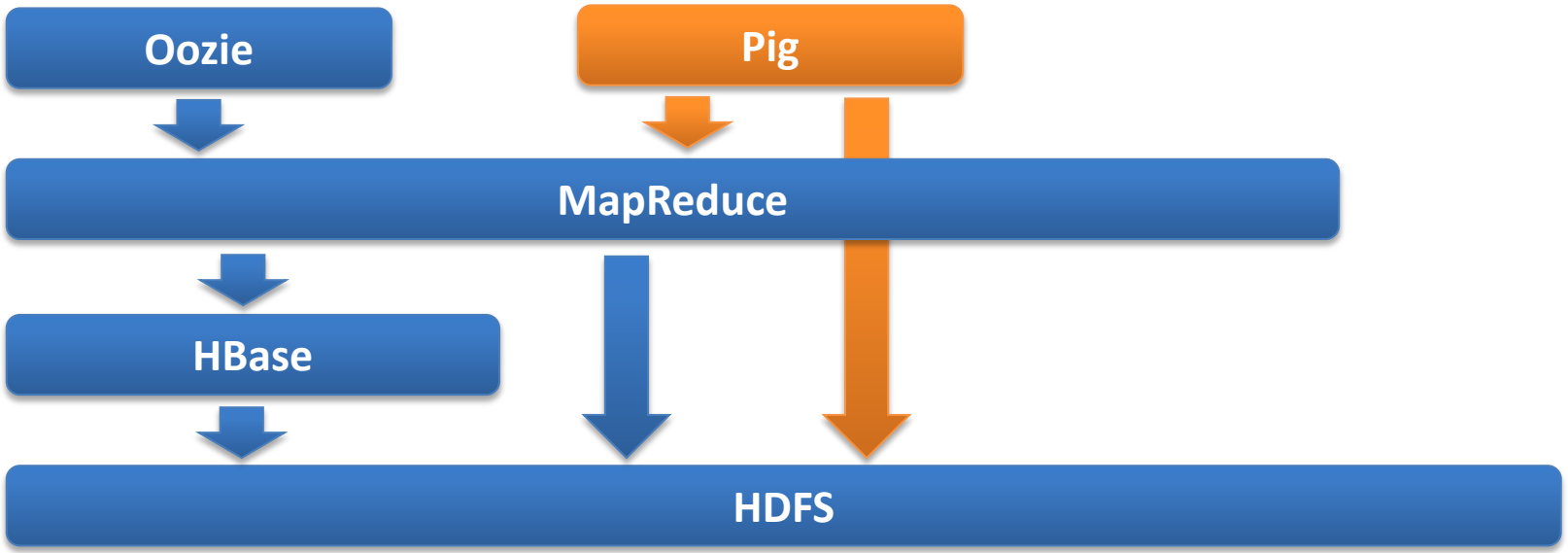
Экосистема Hadoop

Apache Oozie: популярный продукт для координации рабочего процесса MR задач



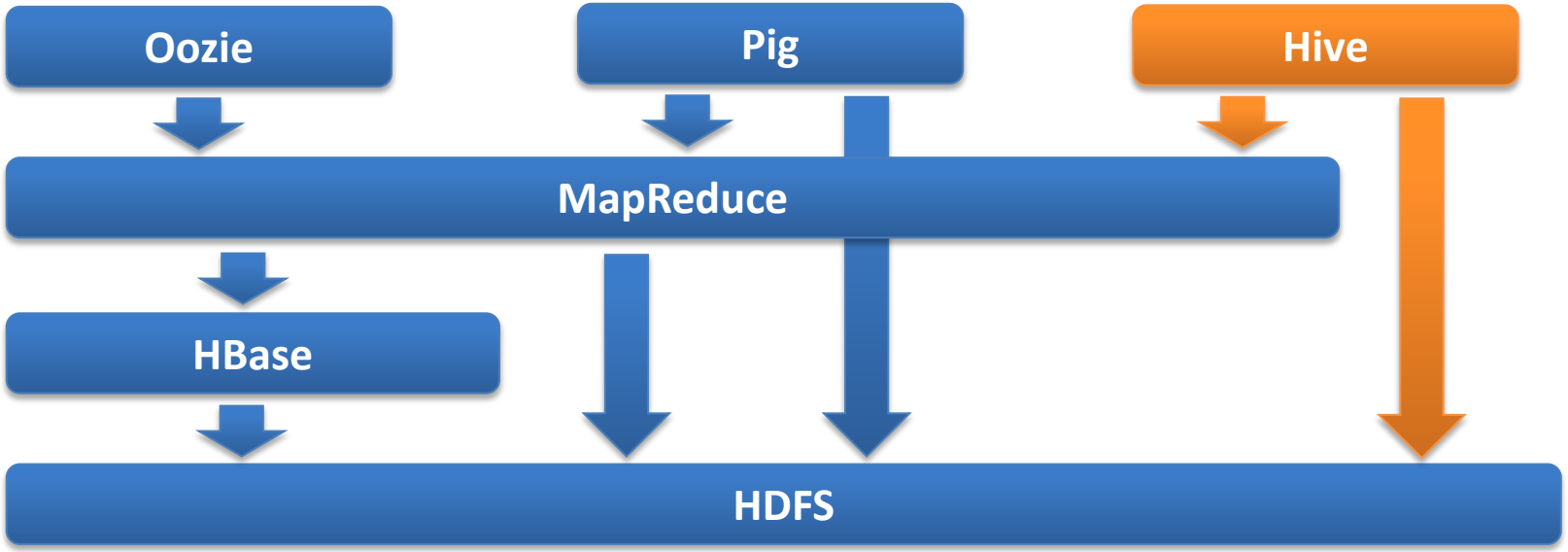
Экосистема Hadoop

Apache Pig инструмент для обработки данных с помощью высокоуровневых команд



Экосистема Hadoop

Apache Hive обработка данных с помощью SQL-подобных запросов



Дистрибутивы Hadoop

Установка Hadoop

- Скачиваем и устанавливаем HDFS и MapReduce с http://hadoop.apache.org/
- Потом скачиваем HBase, не работает с текущим HDFS -> переустанавливаем HDFS
- Устанавливаем Pig -> не работает с нашим HDFS
- Меняем HDFS – перестает работать HBase

Дистрибутивы Hadoop

- Решают проблему несовместимости версий
- Вендоры дистрибутивов обеспечивают:
  - Интеграционные тесты компонентов Hadoop
  - Инсталляционные пакеты в различных форматах
  - Некоторые вендоры делают дополнительные фичи и исправляют баги в стандартной версии Hadoop

Вендоры дистрибутивов

- Cloudera Distribution for Hadoop (CDH)
- MapR Distribution
- Hortonworks Data Platform (HDP)
- Apache BigTop Distribution
- Greenplum HD Data Computing Appliance

Cloudera Distribution for Hadoop (CDH)

- Cloudera является лидером в распространении Hadoop
- Самый популярный дистрибутив
  - http://www.cloudera.com/hadoop
  - 100% open-source
- В Cloudera работает большой процент коммитеров Hadoop
- CDH распространяется в различных форматах
  - RPM, Virtual Machine Images и tarballs

Cloudera Distribution for Hadoop (CDH)

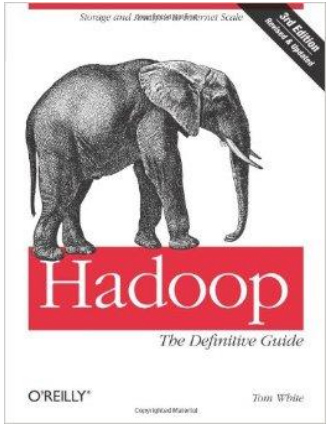
- Включает большинство популярных продуктов Hadoop
  - HDFS, MapReduce, Hbase, Hive, Pig, Oozie, Mahout, Sqoop, Zookeeper, Flume

Поддерживаемые операционные системы

- Каждый дистрибутив поддерживает свой собственный набор операционных систем
- Обычно поддерживаются
  - Red Hat Enterprise
  - CentOS
  - Oracle Linux
  - Ubuntu
  - SUSE Linux Enterprise Server

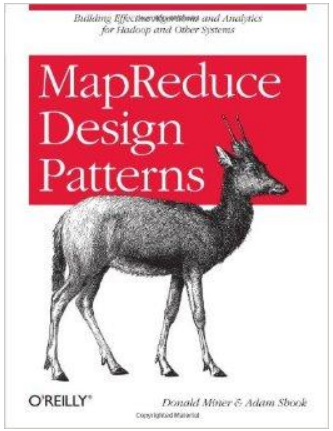
Ресурсы

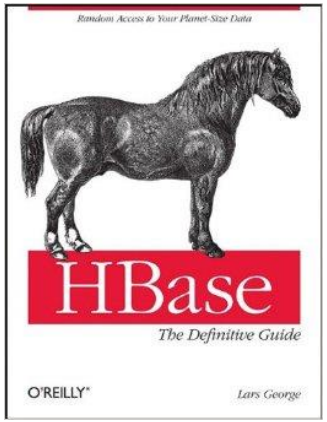
- Apache Hadoop Documentation
  - http://hadoop.apache.org
- Каждый отдельный продукт имеют свою собственную документацию
- Каждый вендор Hadoop предоставляет свою документацию
  - https://ccp.cloudera.com/display/DOC/Documentation



Hadoop: The Definitive Guide
Tom White (Author)
O'Reilly Media; 3rd Edition
Hadoop. Подробное руководство

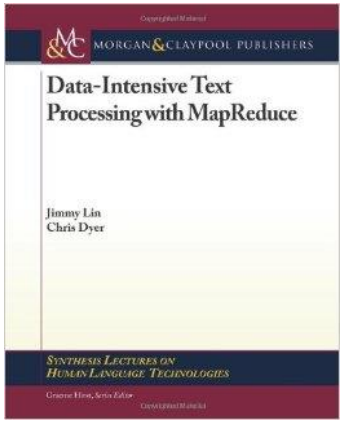
MapReduce Design Patterns
Donald Miner, Adam Shook (Authors)
O'Reilly Media

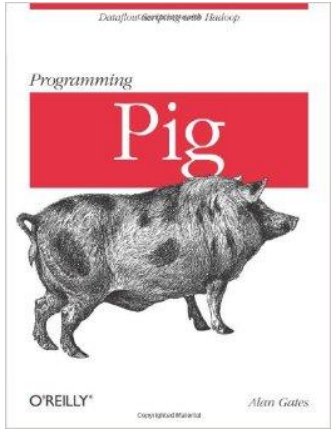




HBase: The Definitive Guide
Lars George (Author)
O'Reilly Media; 1 edition

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer (Authors)
(April, 2010)





Programming Pig
Alan Gates (Author)
O'Reilly Media; 1st Edition

Programming Hive
Edward Capriolo, Dean Wampler,
Jason Rutherglen (Authors)
O'Reilly Media; 1 edition