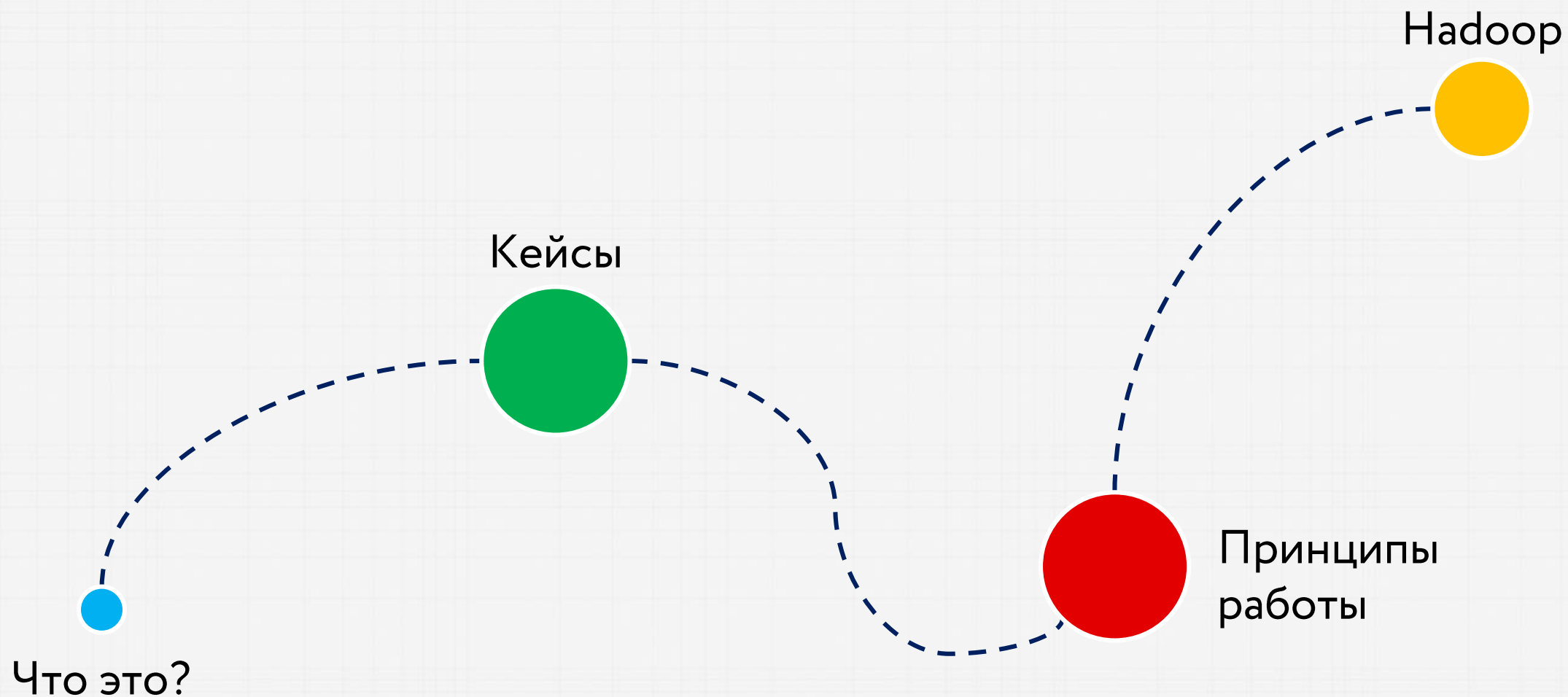




Введение в большие данные

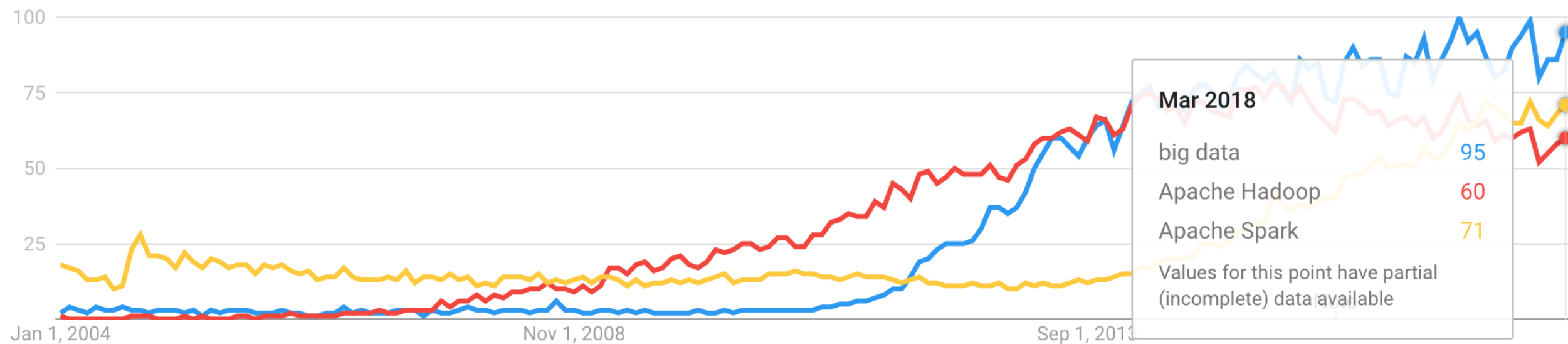
АРТЕМ ПИЧУГИН

План



Что это такое

Google Trends



Big Data

Данные

Большие массивы данных,
в том числе
неструктурированных

Алгоритмы

Аналитика и машинное
обучение, дающие новые
знания

Технологии

Распределенные
хранение и обработка
этих данных

3Vs of Big Data

Volume

Они просто физически
большие: терабайты и
выше

Variety

Они из разных
источников: CRM,
соцсети и т.д.

Velocity

Они обрабатываются в
режиме реального
времени

Кейсы



Ten years ago, we struggled to find 10 machine learning-based business applications. Now we struggle to find 10 that don't use it.

Alexander Linden, Research VP @ Gartner

1. Риск-менеджмент
2. Антифрод
3. Комплаенс
4. Сегментация
5. Персонализация
6. Банкоматы

HOME
CREDIT
BANK

Платформа позволяет обрабатывать до 80 000 запросов сутки и, в результате, значительно снижает риски при выдаче потребительских кредитов.

Данные о денежных переводах, данные социальных сетей. Ценные данные для кредитного скоринга предоставляют банкам операторы мобильной связи.

Для кредитного скоринга компаний используются тексты новостей с их упоминанием, положительная или отрицательная тональность которых определяется автоматически.

**СБЕРБАНК***Всегда рядом*



Тинькофф
Банк

Внедрена платформа VisionLabs LUNA, с помощью которой проводятся оффлайн-расследования: анализ клиентской базы с целью выявления признаков мошенничества и верификация клиентов, подавших заявку на получение кредита, с помощью фотографии.

В Сбербанке была разработана и внедрена система идентификации клиентов, которая сравнивает фотографий из базы с изображениями, получаемыми веб-камерами на стойках.



СБЕРБАНК

Всегда рядом

А Альфа·Банк

Клиентам, ведущим активный образ жизни, банк предлагает программу “Activity” - накопительный счет с повышенной ставкой, на которые будет начисляться сумма денег, пропорциональная количеству пройденных шагов.

Тем, кто часто делает переводы в благотворительные фонды, в Сбербанке предлагают карту “Подари жизнь”, а тем, кто часто бывает за границей - страховку для выезжающих за рубеж.

**СБЕРБАНК***Всегда рядом*

Банкоматы



**Райффайзен
БАНК**

Разработана для банка модель прогнозирования спроса на наличные в банкоматах. Внедрение данной системы позволит в перспективе уменьшить отклонение прогноза от реального спроса на 30%

На основе данных о работе пользователя с приложениями и сайтом банка банкомат автоматически определяет предпочитаемый клиентом язык и предлагает ему наиболее часто используемые им и рекомендуемые ему услуги.



HDFC BANK

We understand your world



Виртуальный оператор колл-центра Елена. Елена распознает вопросы и переадресует либо на один из вариантов IVR, либо на оператора КЦ, ускоряя получение необходимой информации.

Формирование полной картины о состоянии сети и качестве сервисов в масштабах всей страны. Предсказание инцидентов и превентивное тех. обслуживание.

The logo for Detectum, featuring the word "Detectum" in a dark grey sans-serif font, followed by a blue magnifying glass icon.

Поисковик для онлайн-магазинов. Использует всю мощь текущих достижений поисковиков: автодополнение, учет ошибок, понимание естественного языка. [ЧЕРНЫЕ БОСОНОЖКИ НА ВЫСОКОМ КАБЛУКЕ].

Построение рекомендательных систем. Увеличение конверсии блока на 7% в AB-тесте, хороший прирост в деньгах. Лучше всего работает комбинация: 40% Apache Spark (Python) + 50% Hive on TEZ + 10% Hive UDF (Java).

The logo for Ozon.ru, featuring the word "OZON.ru" in blue, with "OZON" in a bold sans-serif font and ".ru" in a lighter font. Below it are five colored dots (blue, green, yellow, red, purple) followed by the word "выбирайте" in blue.

The logo for easyJet, featuring the word "easyJet" in a bold, orange, sans-serif font.

Изменение цены билетов в зависимости от спроса по различным направлениям и сегментам в реальном времени.

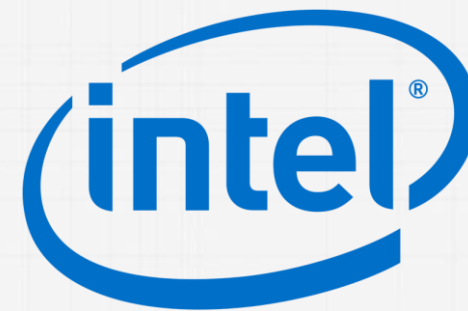
Прогнозирование поломок оборудования и детектирование аномалий, в том числе в реальном времени.

The logo for Southwest Airlines, featuring the word "Southwest" in a bold, blue, sans-serif font, followed by a small red and yellow heart-shaped icon.



Компания использует данные по продажам за предыдущие периоды и оптимизационные алгоритмы, автоматически определяет спрос на материалы и формирует логистические цепочки поставок.

Перед тем как выйти на рынок, каждый микропроцессор должен пройти около 19000 тестов. Анализируя данные по всему производственному процессу, компания выявляет, какие тесты проводить не потребуется, оставляя лишь часть необходимых проверок.



The Facebook logo, consisting of the word "facebook" in a blue, lowercase, sans-serif font, followed by a registered trademark symbol (®).

link prediction

The Tinder logo, featuring the word "tinder" in a red, lowercase, sans-serif font. The letter "i" has a small red flame above it. A trademark symbol (TM) is at the end.

matching

The logo for Donald Trump's 2016 presidential campaign. It features the word "TRUMP" in large, bold, blue, uppercase letters. Below it, in smaller blue uppercase letters, is the slogan "MAKE AMERICA GREAT AGAIN!". The entire text is enclosed in a thin red rectangular border. Above and below the text are five small red stars.

voting

Принципы работы с большими данными

Принципы

Независимое обрабатываем независимо

Позволяет обрабатывать
независимые данные
параллельно

Принцип локальности данных

Обрабатываем там же, где
и храним данные

Пошаговая обработка этапов

Сложный процесс
обработки можно разбить
на несколько простых

Игра

1. Делимся на 4 команды.
2. Каждая команда получает пакет с набором из стикеров 4 разных цветов.
3. Нужно подсчитать количество каждого цвета.

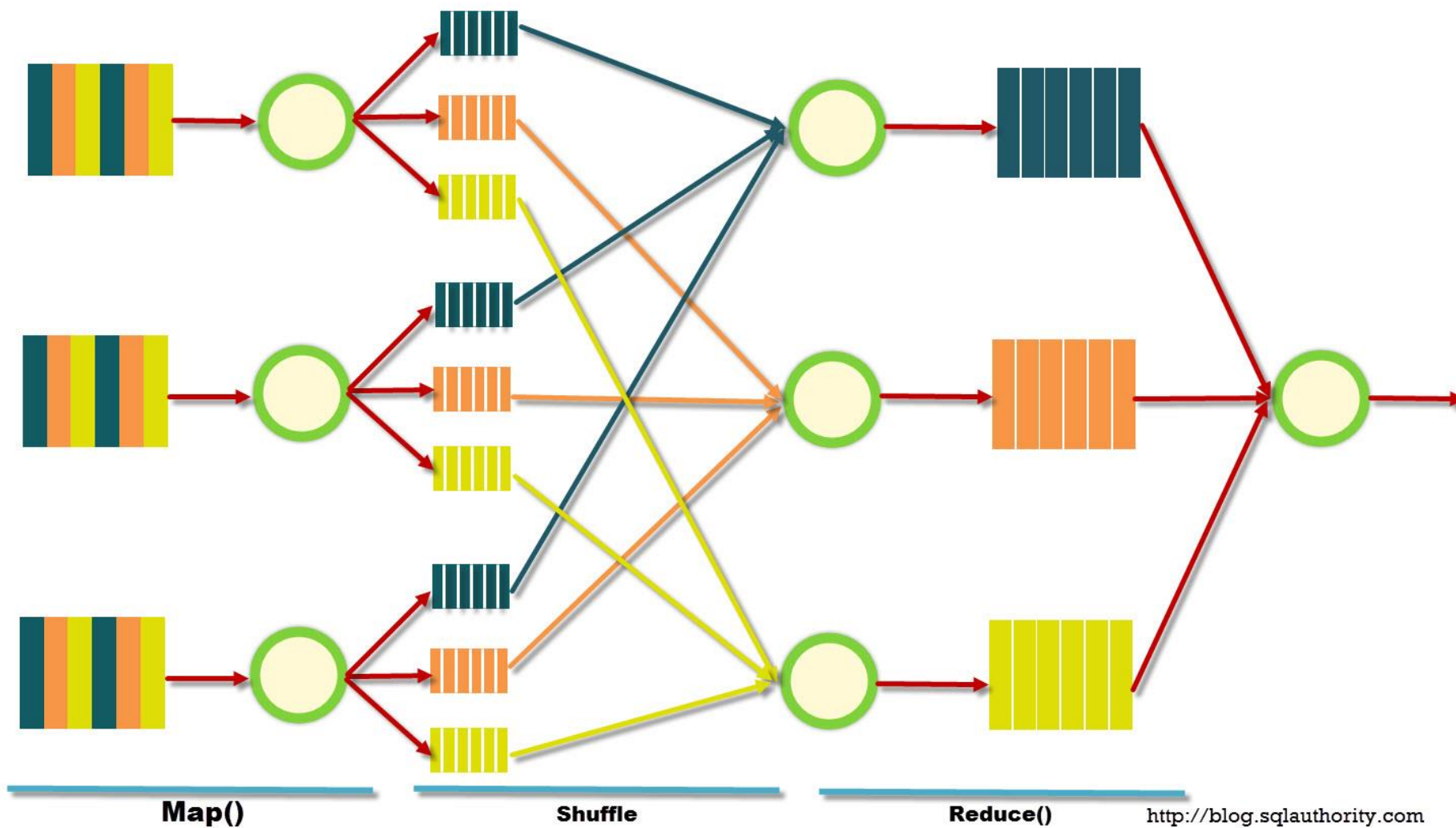
На подготовку к игре: 5 минут.

MapReduce

1. Самая известная парадигма обработки больших данных.
2. Компания Google предложила ее в 2004 году.
3. Много решений построено с этим алгоритмом под капотом (тот же Spark).

MapReduce

How MapReduce Works?



1. Стадия Map:

- вход: исходный объект
- выход: множество пар ключ-значение

2. Стадия Shuffle: данные сортируются по ключу и распределяются по редьюсерам

3. Стадия Reduce:

- вход: отсортированные ключи и список значений
- выход: ключ-значение

Не путать!

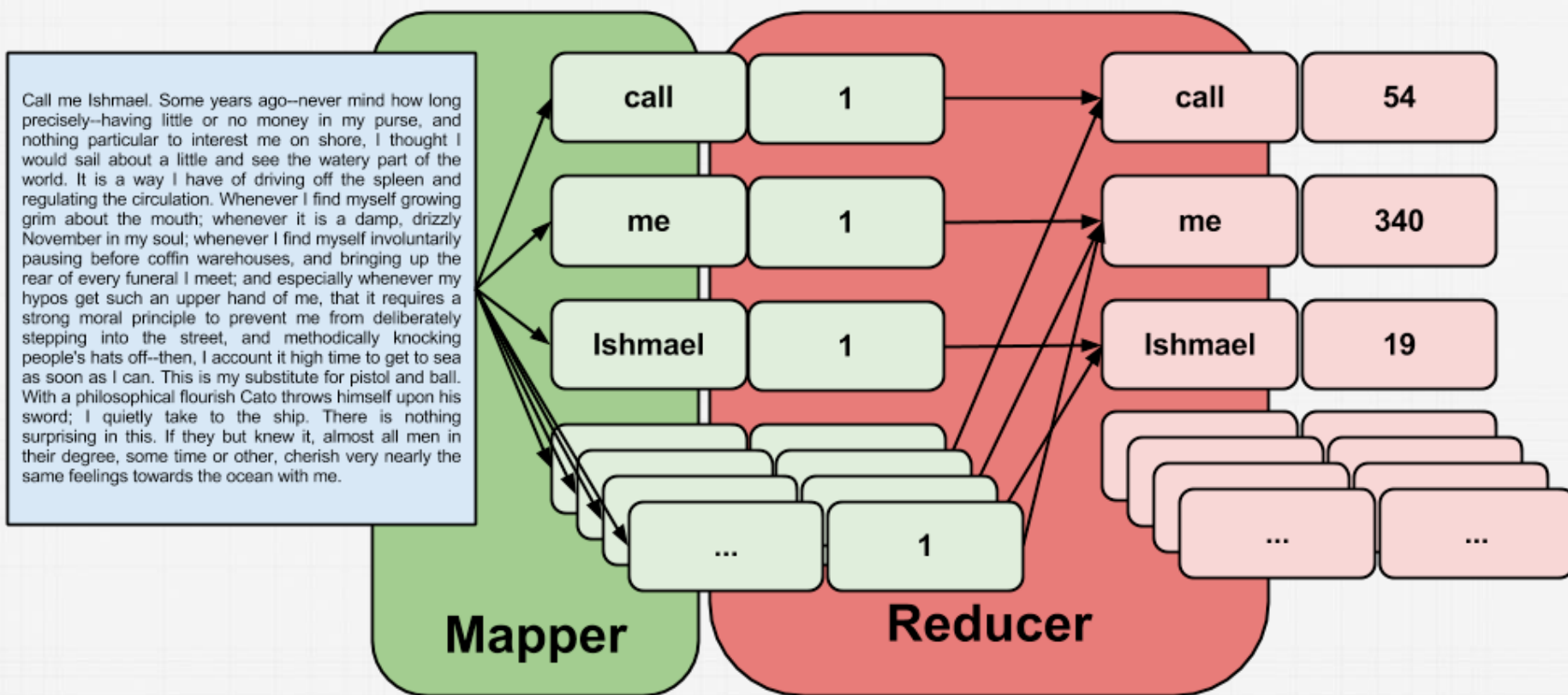
- Стадия Map, стадия Reduce
- Mapper, Reducer
- Функция Map, функция Reduce

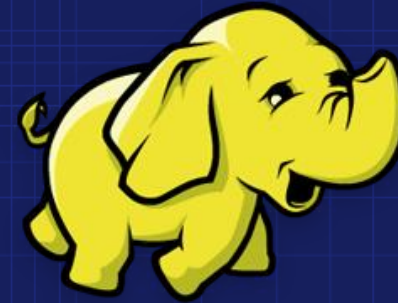
Word Count

- Дан файл со строками
 - считаем, что 1 строка – это 1 документ
- Посчитать, сколько раз встречается слово в исходном файле

```
def map(string):  
    for token in string.split():  
        print token , 1  
  
def reduce(key, values):  
    print key, sum(values)
```

Решение

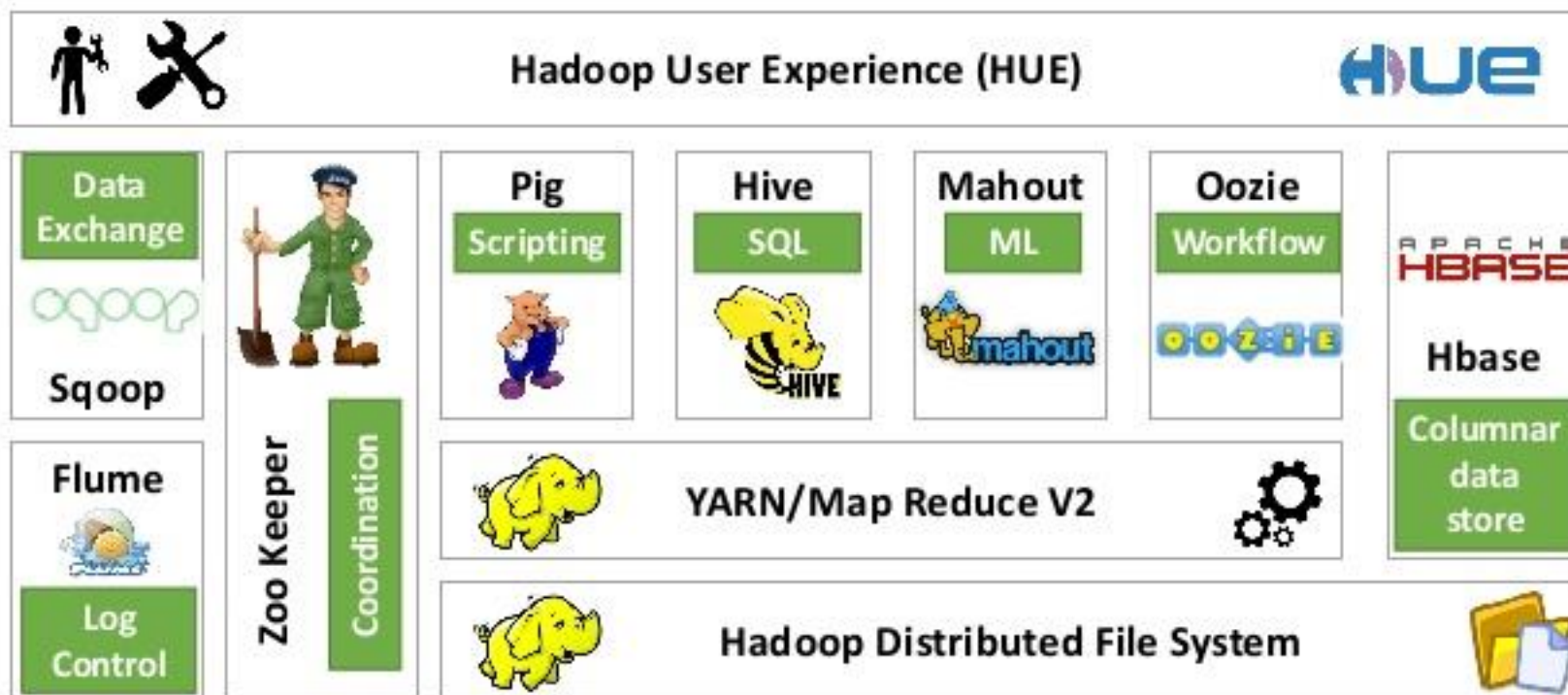




Hadoop

Hadoop

The Apache Hadoop Stack



Hadoop Streaming

- Реализация MapReduce в Hadoop
- Mapper и Reducer реализуются в виде отдельных программ
- И mapper и reducer читают входные данные с `sys.stdin` и пишут выходные на `sys.stdout`
- Ключ и значение отделяются друг от друга знаком табуляции

Word Count

#mapper.py:

```
for line in sys.stdin:
    for token in line.strip().split():
        print(token + "\t1")
```

#reducer.py

```
prev_key = None
sum = 0
for line in sys.stdin:
    if key != prev_key and prev_key is not None:
        print prev_key, sum
        sum = 0
    sum + 1
    prev_key = key
if key != prev_key and prev_key is not None:
    print prev_key, sum
```

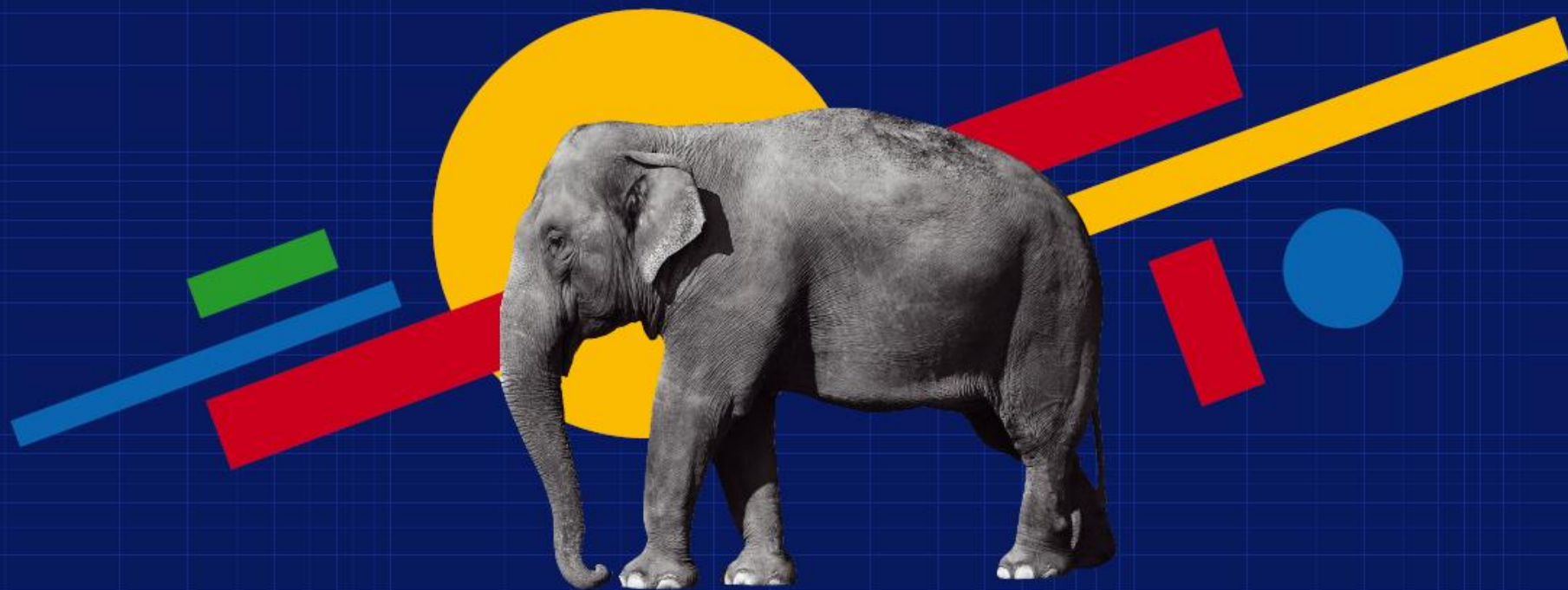

Статьи

- нашего преподавателя Александра Петрова - <https://habrahabr.ru/company/dca/blog/267361/>

Книги

- Hadoop: The Definitive Guide, Tom White
- Hadoop in Action, Chuk Lam

Официальная документация



BIG DATA IS LOVE

NEWPROLAB.COM