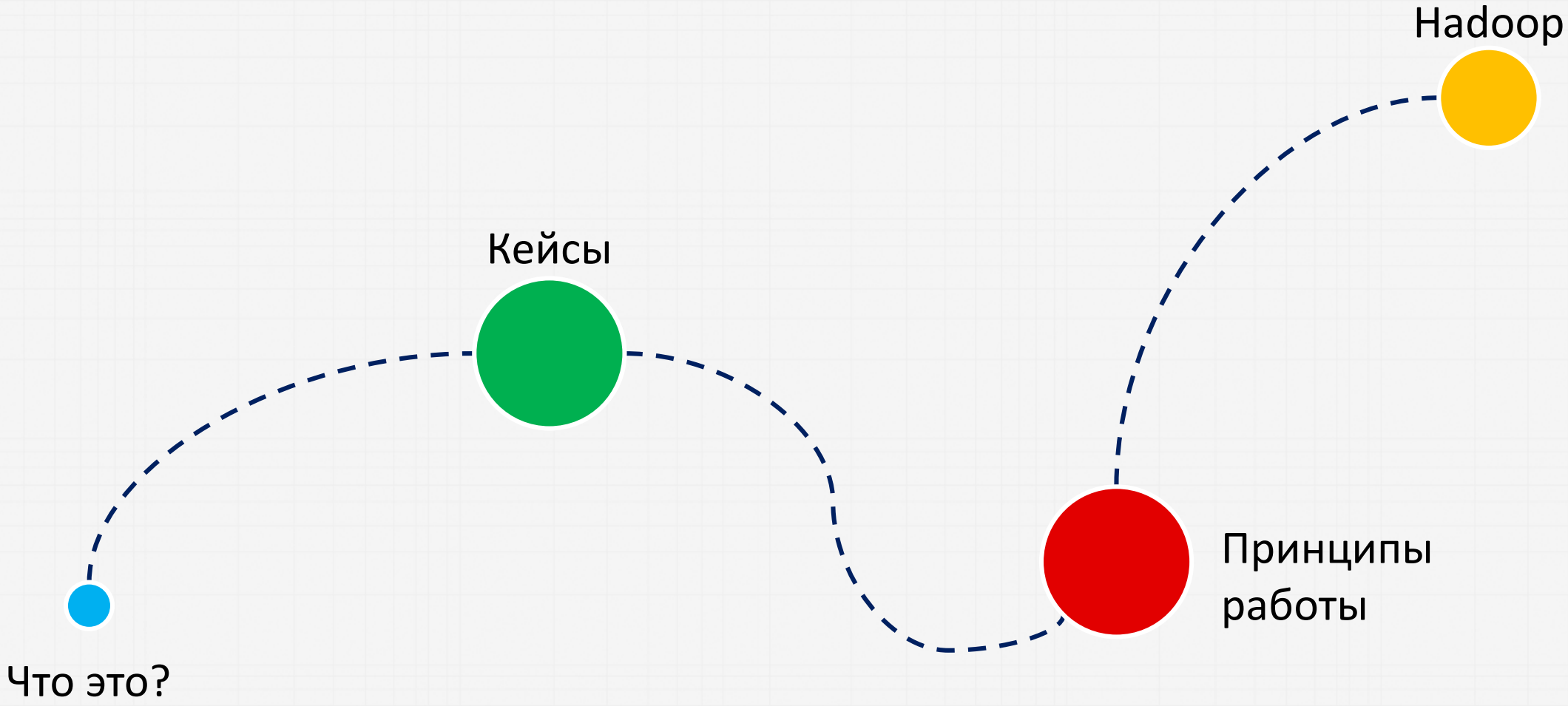
ВВЕДЕНИЕ В БОЛЬШИЕ ДАННЫЕ

АРТЕМ ПИЧУГИН

План



Что это такое

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

Банки

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

Скоринг

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.

Антифрод

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

Сегментация

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

Телекомы

МегаФон

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.

E-commerce

Detectum

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).

ozon.ru выбирайте

Авиакомпании

easyJet

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

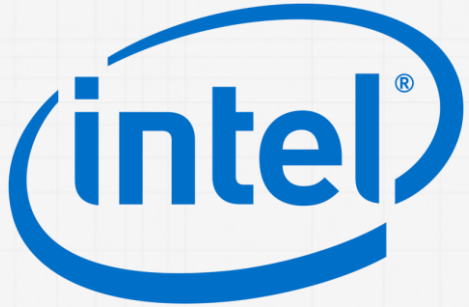
Прогнозирование поломок оборудования и детектирование аномалий, в том числе в реальном времени.

Southwest

Промышленность

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.



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

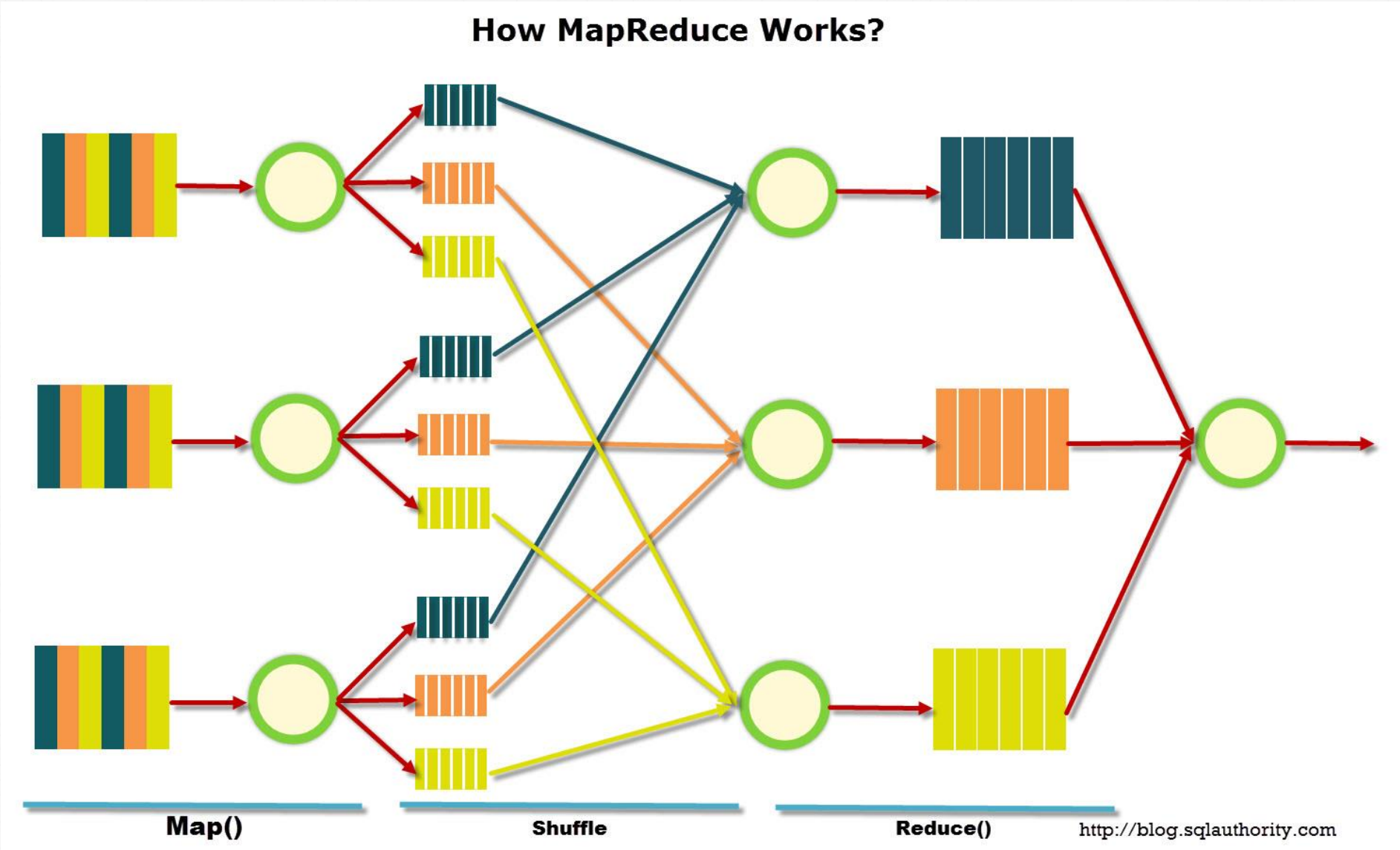
1. Делимся на 4 команды.
2. Каждая команда получает пакет с набором из стикеров 4 разных цветов.
3. Нужно подсчитать количество каждого цвета.

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?
Map()
Shuffle
Reduce()
http://blog.sqlauthority.com

MapReduce

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

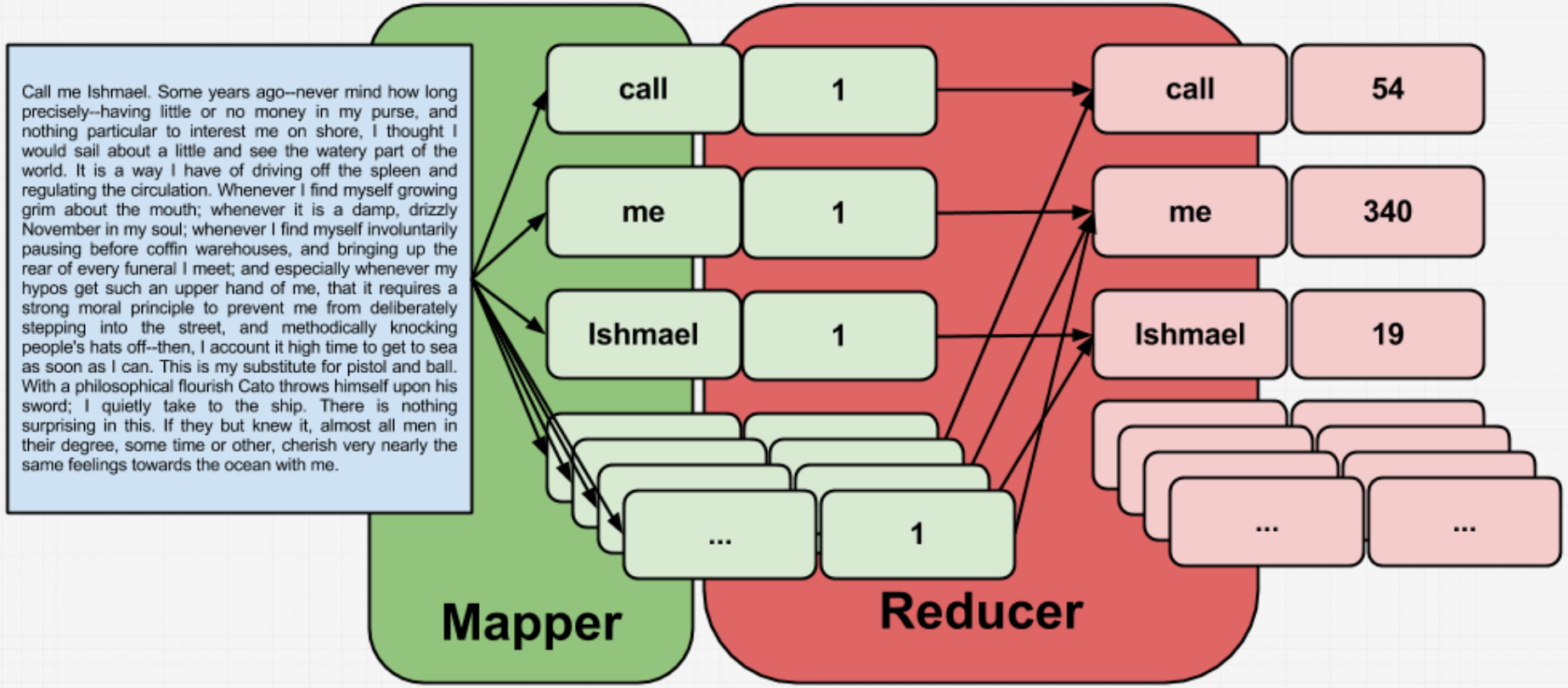
- Дан файл со строками
  - считаем, что 1 строка – это 1 документ
- Посчитать, сколько раз встречается слово в исходном файле

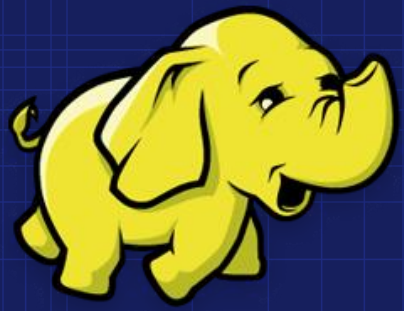
Решение

```
def map(string):
    for token in string.split():
        print(token, 1)

def reduce(key, values):
    print(key, sum(values))
```

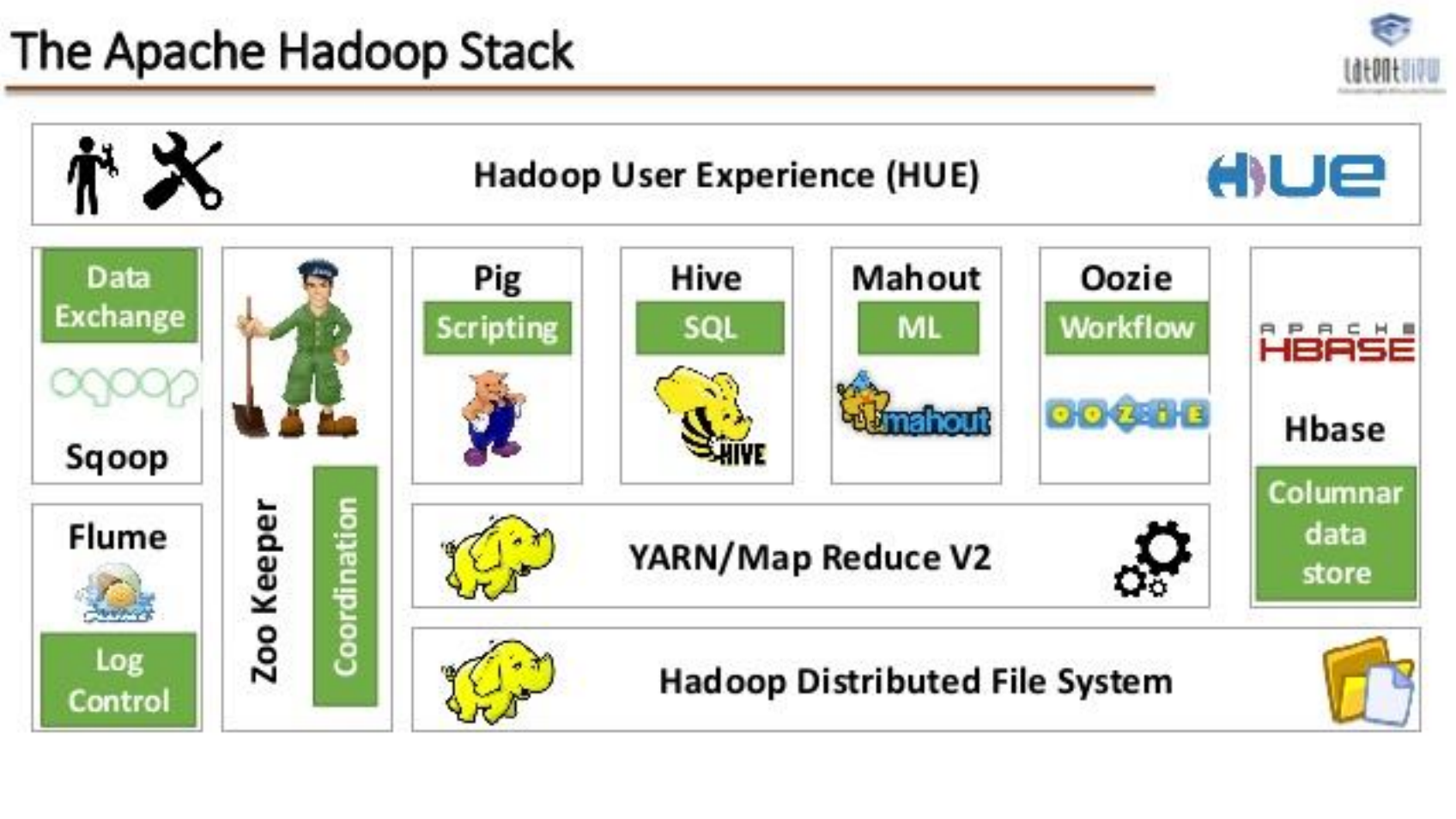
Решение


Call me Ishmael. Some years ago--never mind how long precisely--having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off--then, I account it high time to get to sea as soon as I can. This is my substitute for pistol and ball. With a philosophical flourish Cato throws himself upon his sword; I quietly take to the ship. There is nothing surprising in this. If they but knew it, almost all men in their degree, some time or other, cherish very nearly the same feelings towards the ocean with me.
call
1
me
1
Ishmael
1
...
1
Mapper
Reducer
call
54
me
340
Ishmael
19
...
...



Hadoop

Hadoop


The Apache Hadoop Stack
Hadoop User Experience (HUE)
HUE
Data Exchange
Sqoop
Flume
Log Control
Zoo Keeper
Coordination
Pig
Scripting
Hive
SQL
HIVE
Mahout
ML
mahout
Oozie
Workflow
OOZIE
APACHE HBASE
Hbase
Columnar data store
YARN/Map Reduce V2
Hadoop Distributed File System

Hadoop Streaming

- Реализация MapReduce в Hadoop
- Mapper и Reducer реализуются в виде отдельных скриптов
- 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

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

```python
#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 prev_key is not None:
        print(prev_key, sum)
```

https://goo.gl/49YVsD

Что почитать

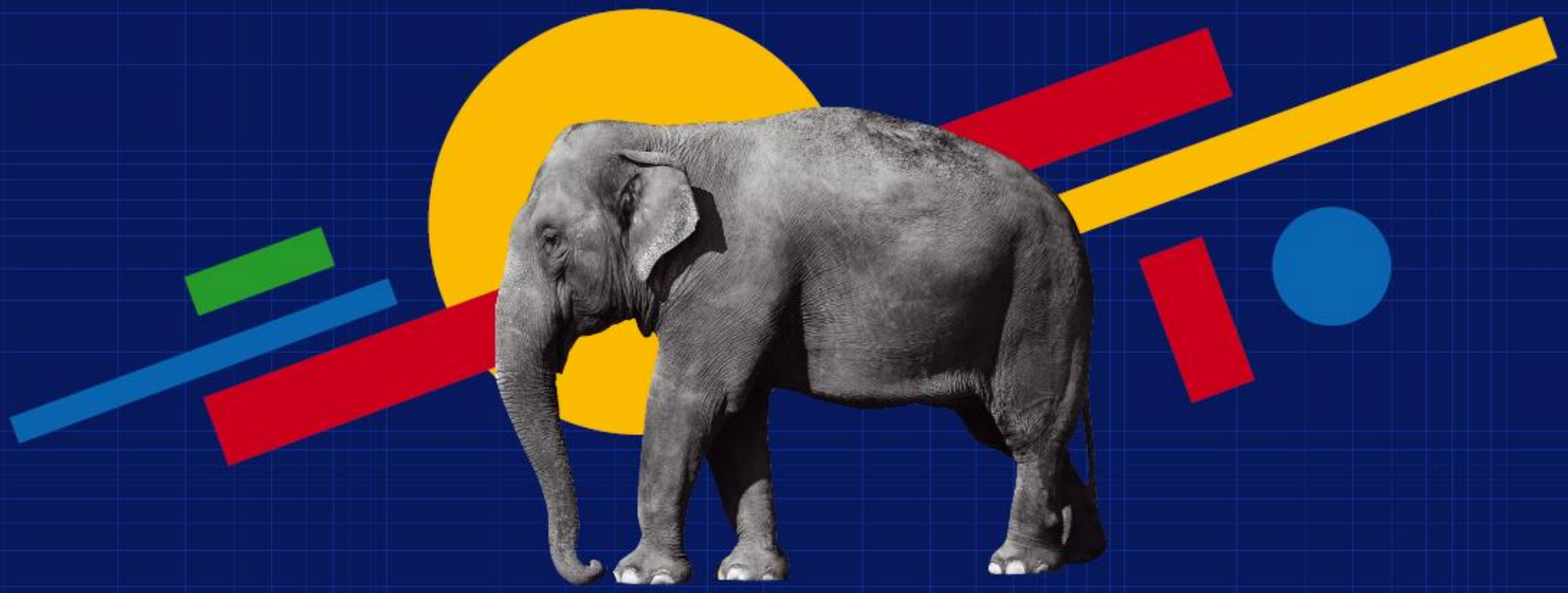
Статьи

- нашего бывшего 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