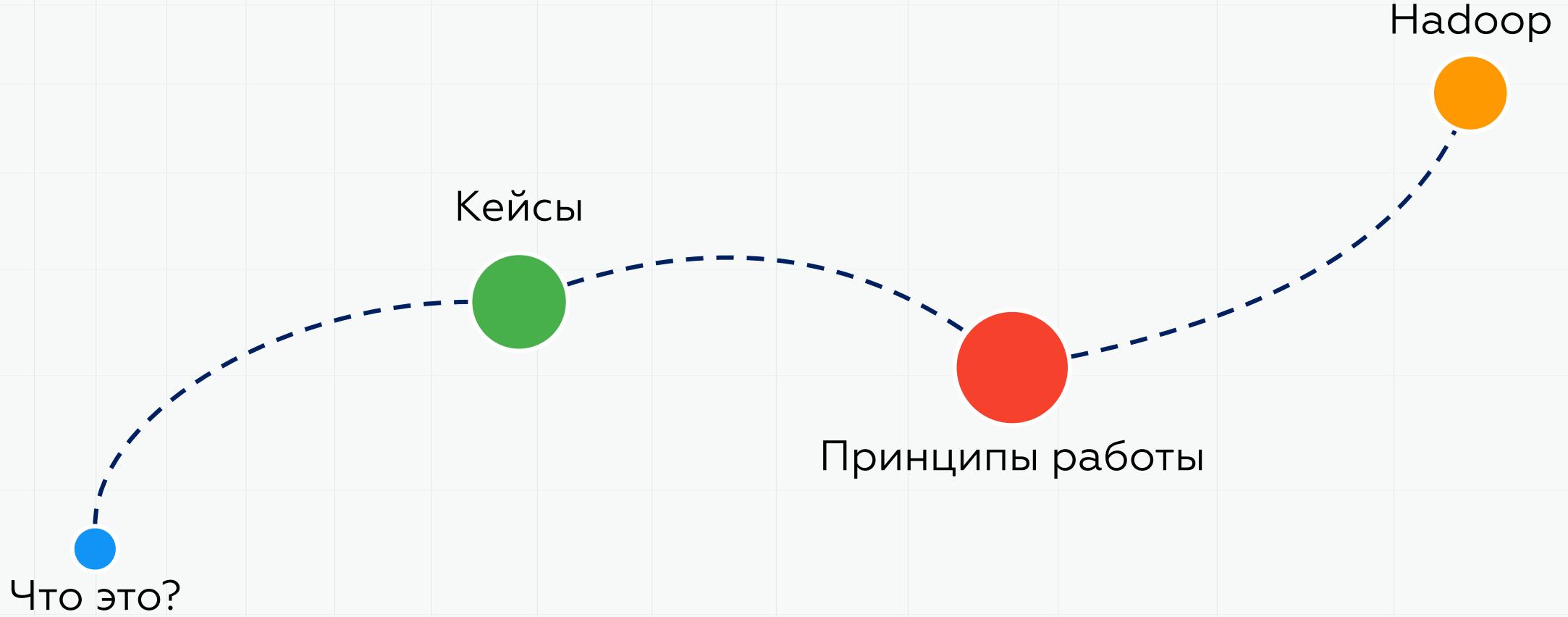




ВВЕДЕНИЕ В БОЛЬШИЕ ДАННЫЕ

Весна 2021

План



ЧТО ЭТО ТАКОЕ?

Это Big Data?

Пример 1

Коллега из отдела маркетинга попросил обработать CSV файл с "большими данными". В файле немногим больше 1 млн. записей.

Пример 2

IT отдел обрабатывает данные от CDN сетей и считает кол-во переданных байт пользователям. Расчет занимает ~1 неделю.

Big Data

Данные

Большие массивы
данных, в том числе
неструктурированных

Алгоритмы

Аналитика и машинное
обучение, дающие
новые знания

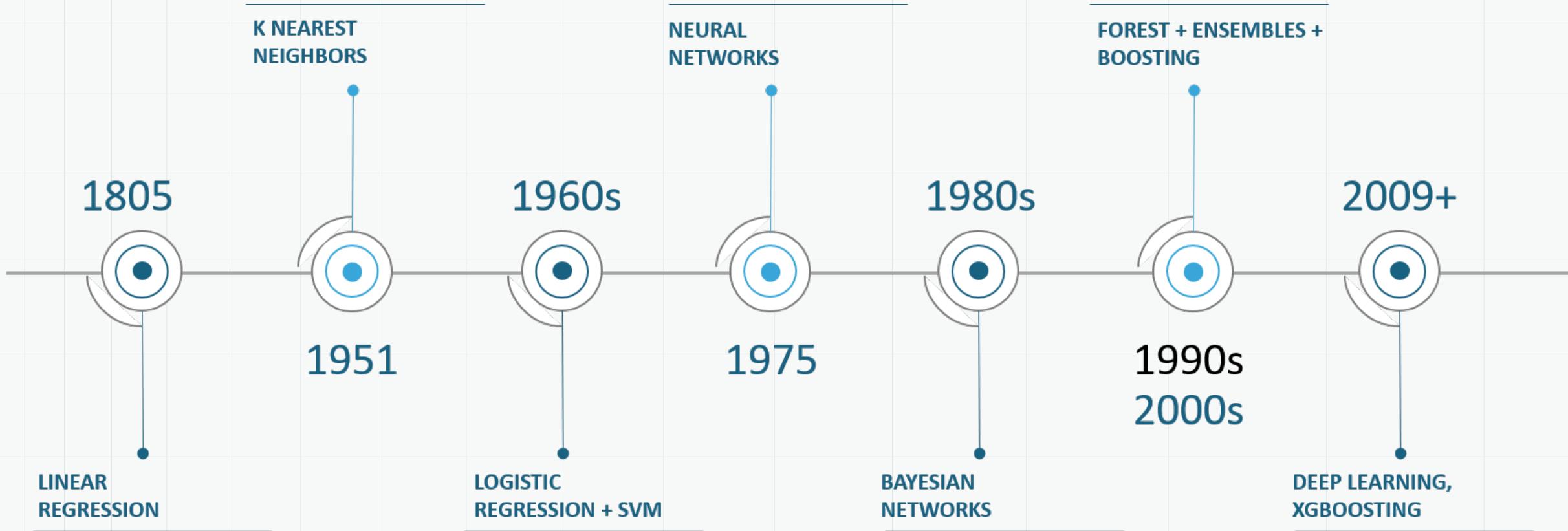
Технологии

Распределенные
хранение и обработка
этих данных

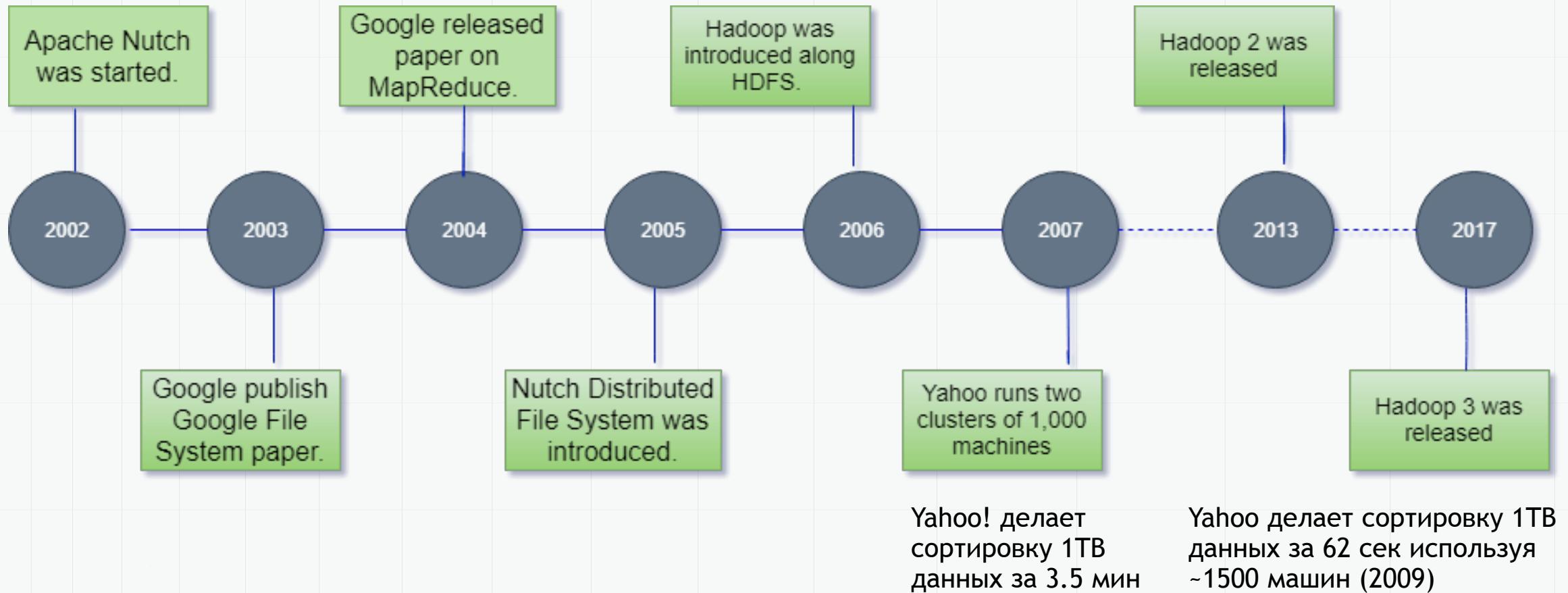
Данные



Алгоритмы



Технологии



3Vs of Big Data

Volume

Они просто физически большие: терабайты, петабайты и выше.

Variety

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

Velocity

Они быстро поступают и обрабатываются в режиме реального времени.

4-5-6-7Vs ?

Volume

Variety

Velocity

Veracity

Viability

Value

Достоверность (качество)

Жизнеспособность

Ценность

Variability

Переменчивость

NEW
PRO
LAB

КЕЙСЫ

Цитата

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

IT компаний

Пионеры применения ML + Big Data в продакшне

YAHOO!

Решали проблему поиска и придумали Hadoop, а также реализовали подходы к распределенным вычислениям

Google

Google Translate / Google Lens
Google Voice Search (“OK Google”)

Яндекс

Контекстная реклама (Директ)
Беспилотники



IT сервисы

Компании-сервисы на стыке онлайн и офлайна

Uber

Поиск оптимального водителя из пула доступных
Оптимизация цен

Booking.com

Поиск оптимального отеля на указанные даты
Эластичные цены

Банки

1. Риск-менеджмент (скоринг)
2. Антифрод
3. Сегментация и персонализация
4. Банкоматы

Скоринг



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

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



Всегда рядом

Антифрод



Тинькофф
Банк

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

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.

 **СБЕРБАНК**
Всегда рядом

Сегментация



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.



Банкоматы



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 предпочтаемый клиентом язык и предлагает ему наиболее часто используемые им и рекомендуемые ему услуги.



Телекомы



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

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

E-commerce

Detectum 

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

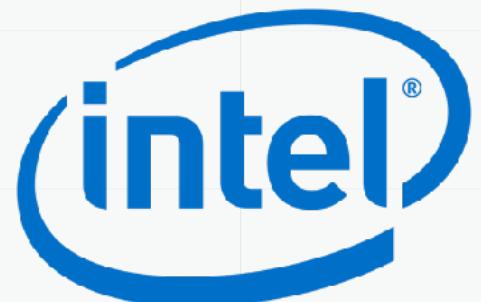
ozon.ru
••••• выбирайте

Промышленность



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.

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



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

Принципы

Пример: международный интернет-магазин, гео-распределённые данные

Задача: посчитать средний чек по всем покупателям

Независимое
обрабатываем
независимо

Считать метрики по
странам можно
параллельно т.к. они не
зависят друг от друга

Принцип
локальности
данных

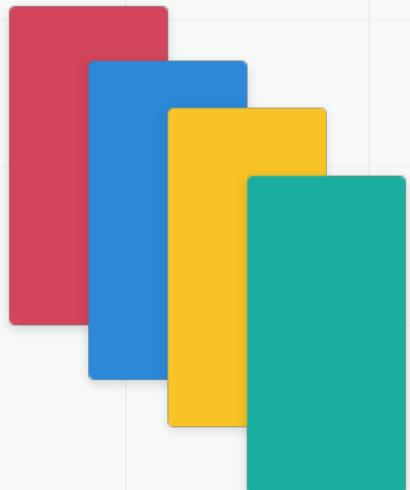
Считаем метрики там
же, где храним данные,
не нужно передавать их
в одно место

Пошаговая
обработка
этапов

Считаем сумму продаж
и количество
покупателей по
странам, затем делаем
метрику среднего чека

Задачка

1. Представьте, что вы в команде из 5 человек.
2. Команда получила пакет с набором из стикеров 4 разных цветов.
3. Нужно подсчитать количество каждого цвета.

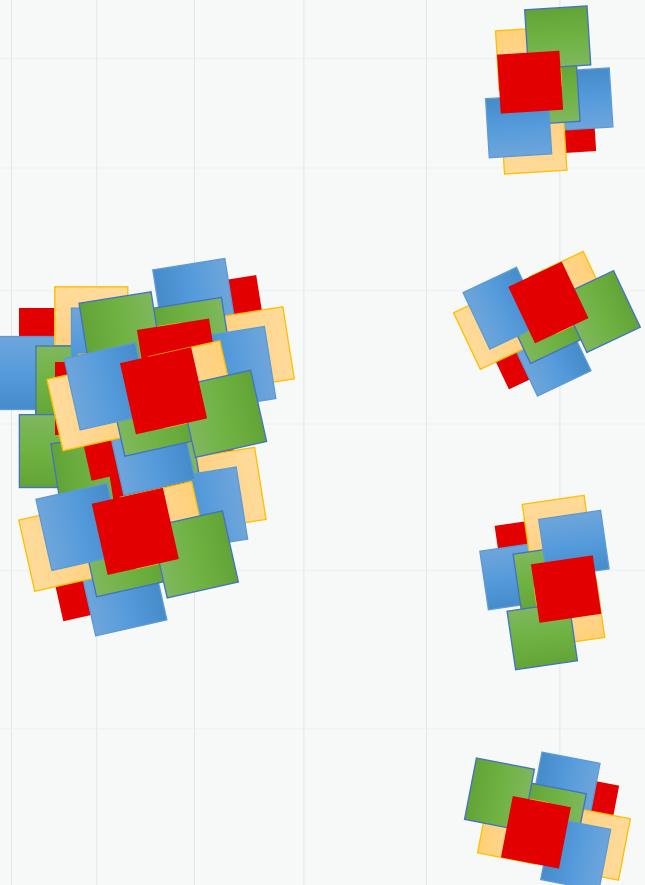


На обдумывание алгоритма 7-8 минут.

MapReduce

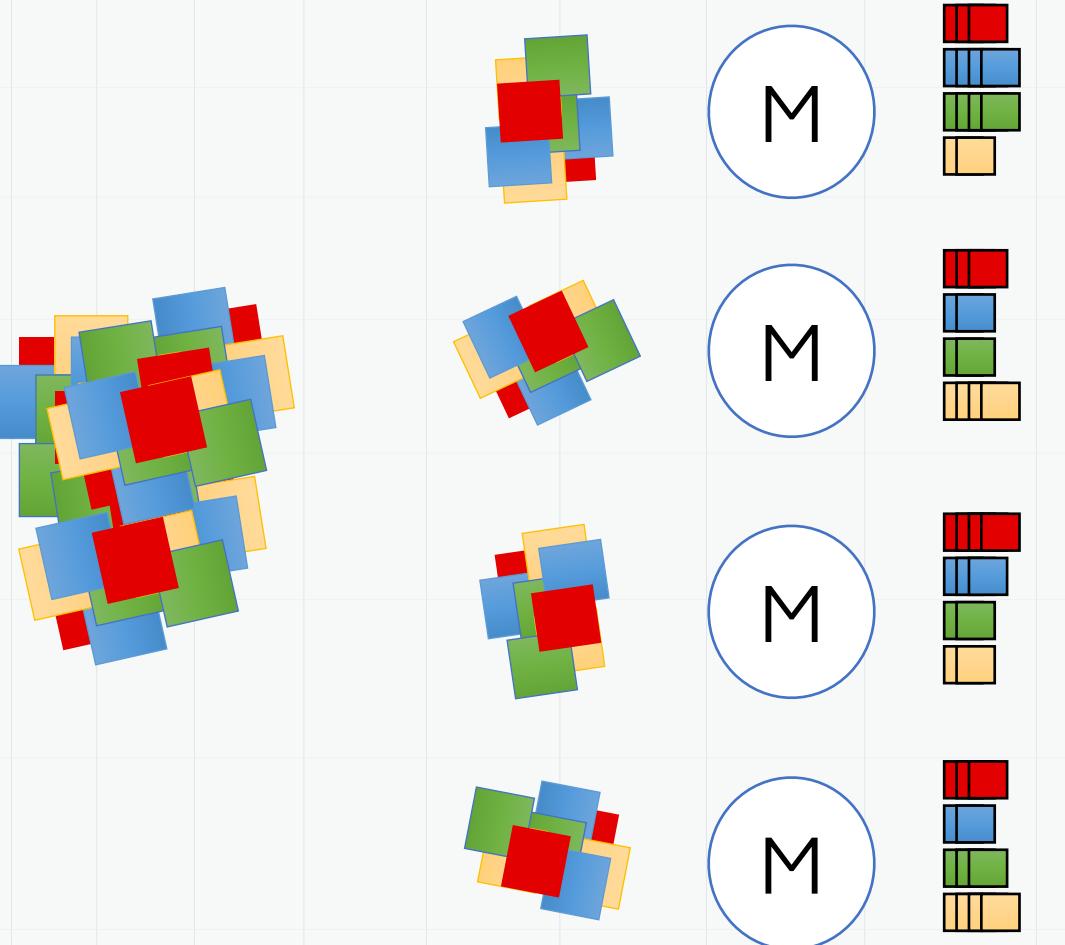
1. Самая известная парадигма обработки больших данных
2. Компания Google предложила ее в 2004 году
3. Много решений построено с этим алгоритмом под капотом

MapReduce



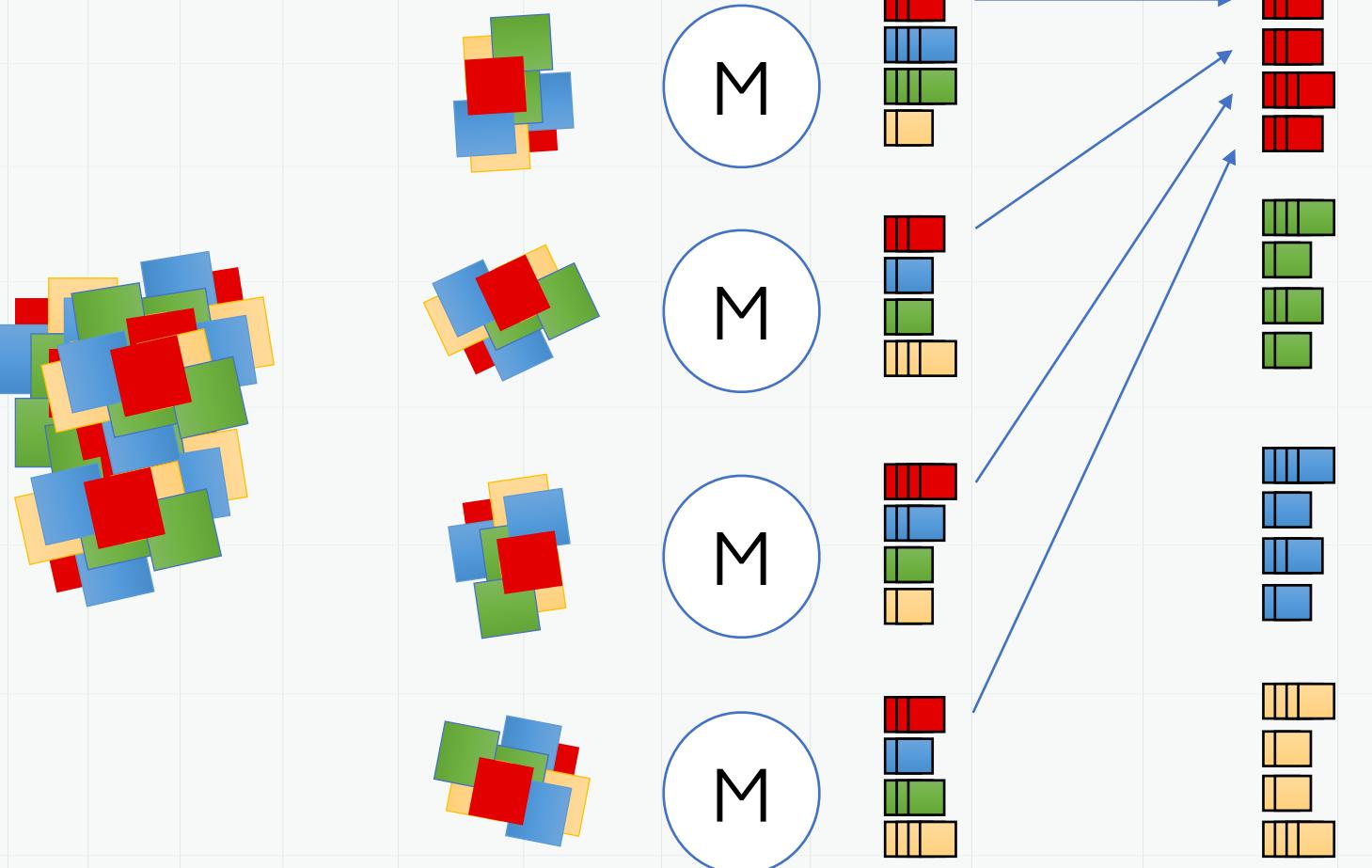
Шаг 0 - делим на кучки

MapReduce



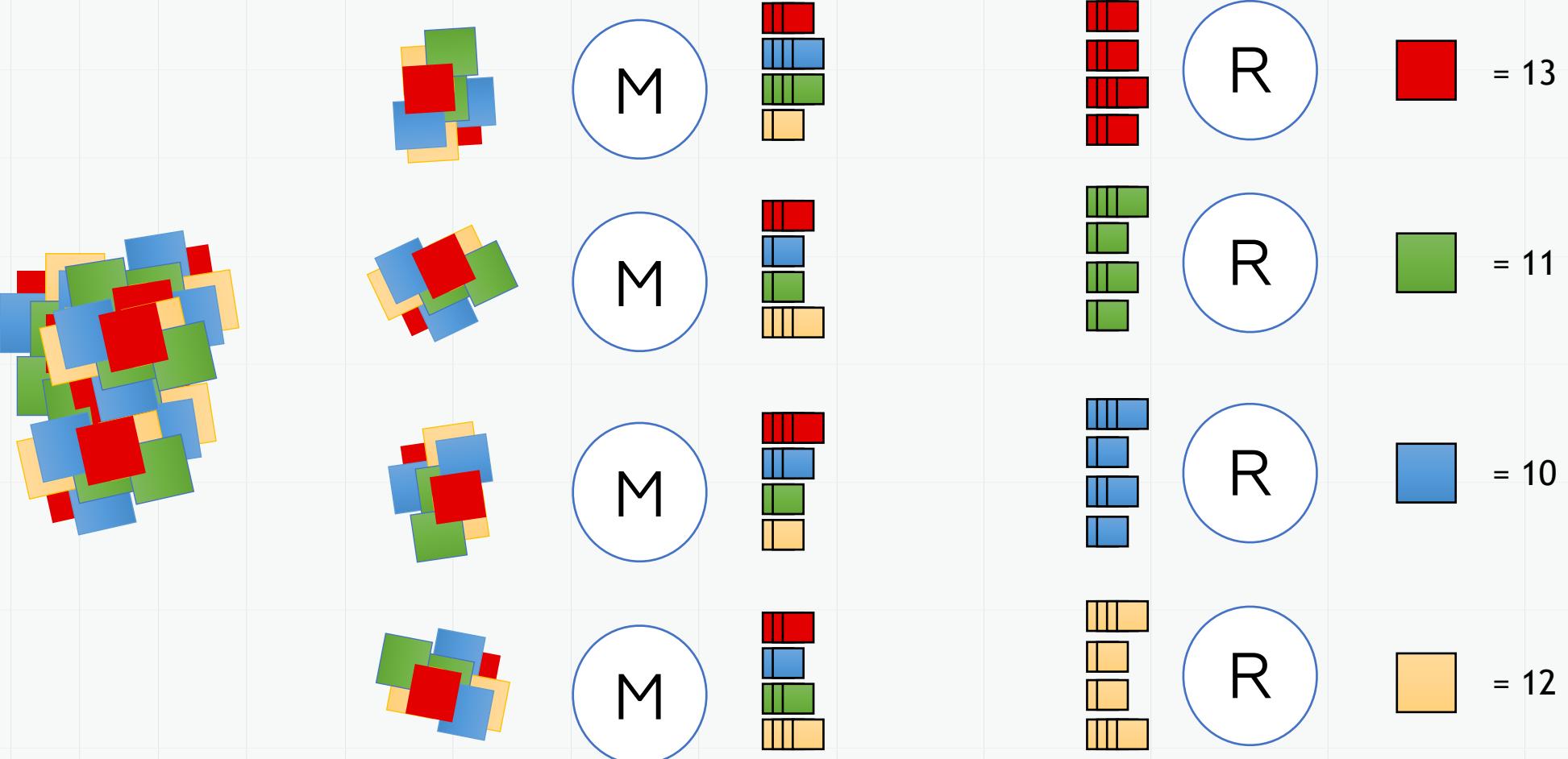
Шаг 1 - стадия Мар: на входе кучка, на выходе отсортированные по ключу (цвету) кучки

MapReduce



Шаг 2 - стадия **Shuffle**: сортирует все выходы мапперов по ключу (цвету)

MapReduce



Шаг 3 - стадия Reduce: считаем количество ключей и выдаём ответ

MapReduce

1. Стадия **Мар**:

- 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



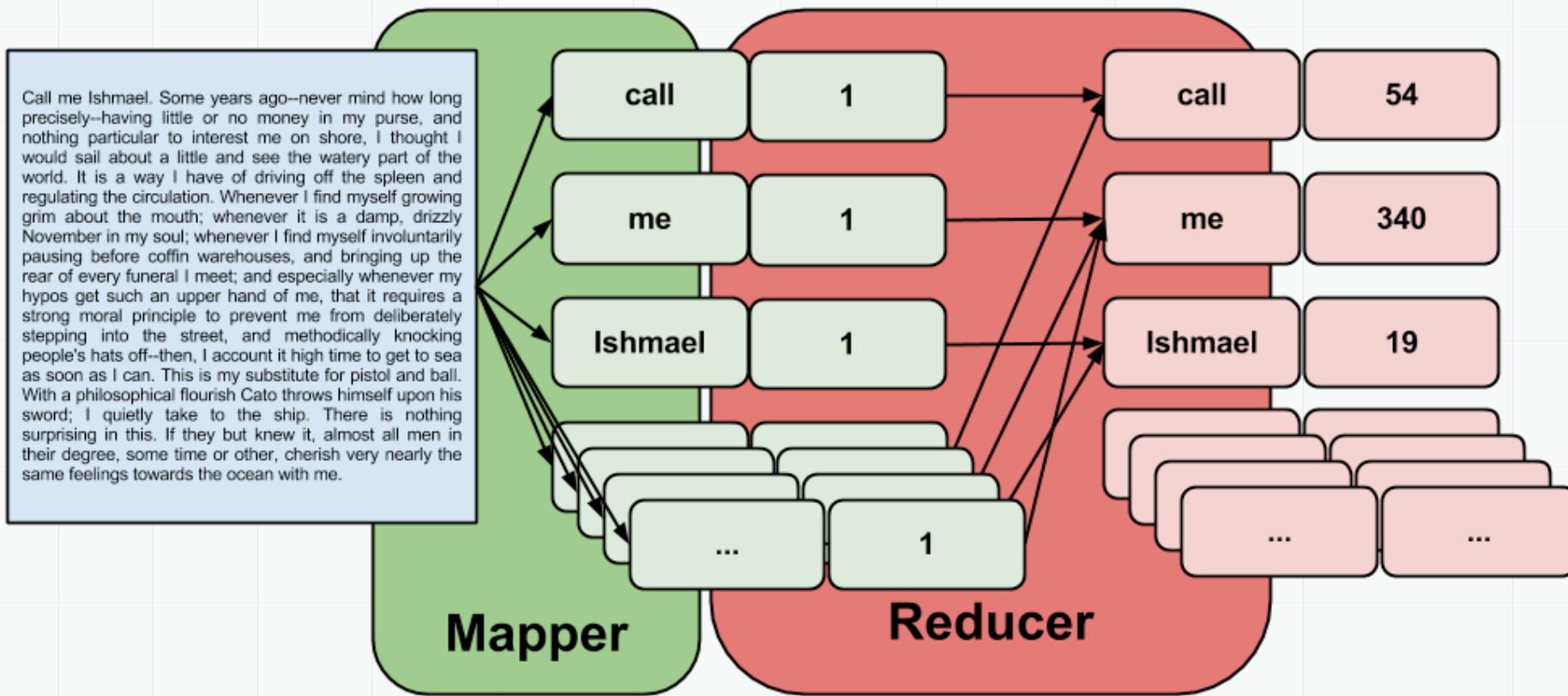
Не путать!

1. Стадия Map, стадия Reduce
2. Mapper, Reducer
3. Функция Map, функция Reduce

WordCount

1. Дан файл со строками.
2. Считаем, что 1 строка – это 1 документ.
3. Посчитать, сколько раз встречается слово в исходном файле.

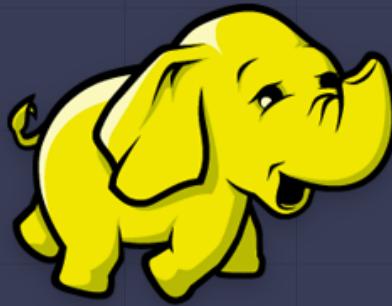
Решение



Решение

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

HADOOP



Экосистема

The Apache Hadoop Stack



HUE



Hadoop User Experience (HUE)

Data Exchange



Sqoop

Zoo Keeper

Coordination

Pig Scripting



Hive SQL



HIVE

Mahout ML



Mahout

Oozie Workflow



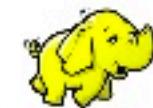
Oozie

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

WordCount

```
#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:  
    key, value = line.split("\t")  
    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)
```

Reducer

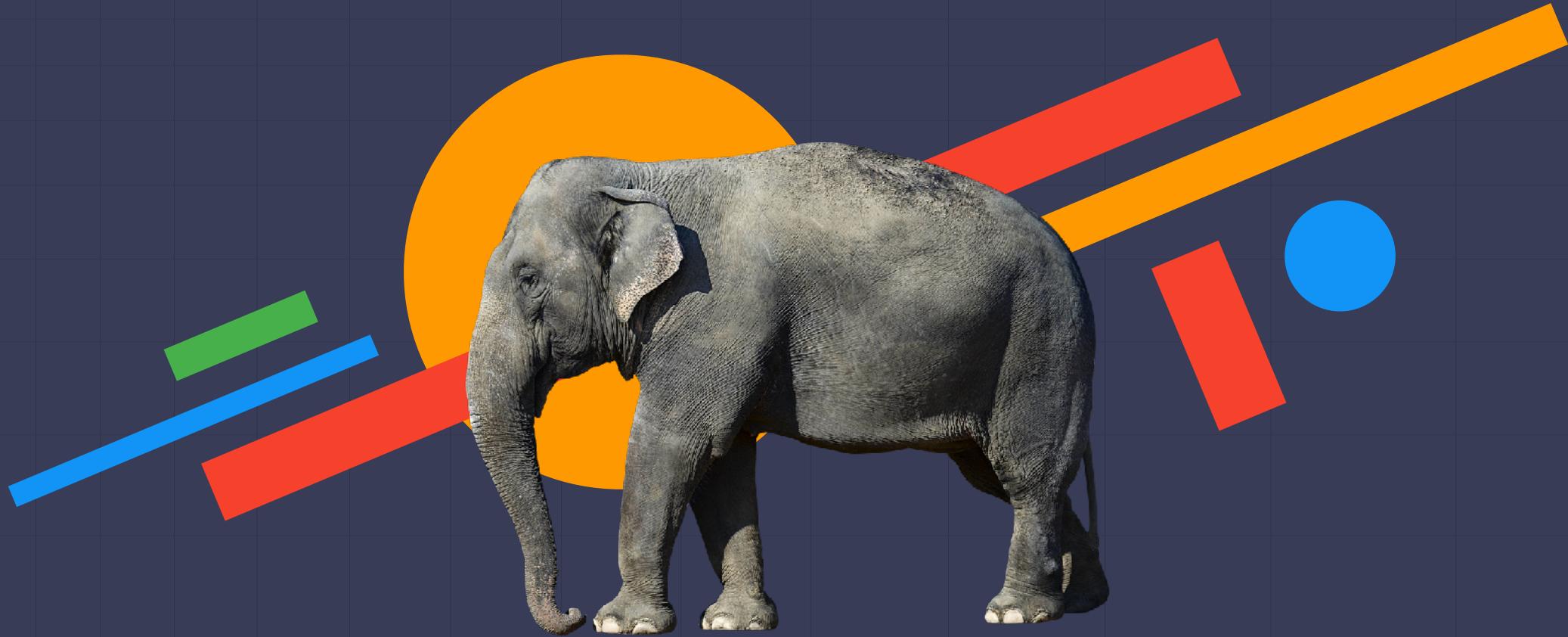
```
never1
never1
never1
ever1
forever1
forever1
```

```
#reducer.py
prev_key = None
sum = 0
for line in sys.stdin:
    key, value = line.split("\t")
    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)
```

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, Chuck Lam
- Официальная документация



Big Data is Love

NEWPROLAB.COM