Работа с большими данными

Содержание курса

01 Введение. Распределенные вычисления. MapReduce.

02 HDFS. Apache Spark. RDD.

03 Базы данных. Spark SQL. Хранение больших данных.

04 Подробнее о модели вычислений Spark. Знакомство со Scala.

05 Алгоритмы машинного обучения на больших данных. spark.ml.

06 Рекомендательные системы. Виды. Метрики.

07 Обработка потоковых данных. Structured streaming и интеграция с spark.ml.

08 Модели в продакшен. Управление кластером.

1. Введение. Распределенные вычисления. MapReduce.

План:

1. Содержание курса. Организационные вопросы.
2. Сферы производящие большие данные. Data explosion.
3. Большие данные - где начало. 3 основных принципа.
4. Как компании справляются с большими данными IaaS/PaaS/SaaS.
5. Оперативная память, жесткий дик. Сортировка во внешней памяти.
6. Плюсы и минусы распределенных систем. Предпосылки к созданию MapReduce.
7. Задача подсчета слов. Map. Shuffle. Reduce.

Организационные вопросы

На первое время всем рекомендую зарегистрироваться на https://databricks.com/ выбрать community edition. Там будут доступна работа с ноутбуком и с установленным Spark.

Здесь есть ссылка на материалы, чтобы самостоятельно вне зависимости от используемой системы установить себе виртуальную машину с убунту, а также последовательно установить hadoop, spark.

Результаты опросника:

Вопросы можно писать в чат, когда будет кончаться секция - буду отвечать на вопросы по этой секции.

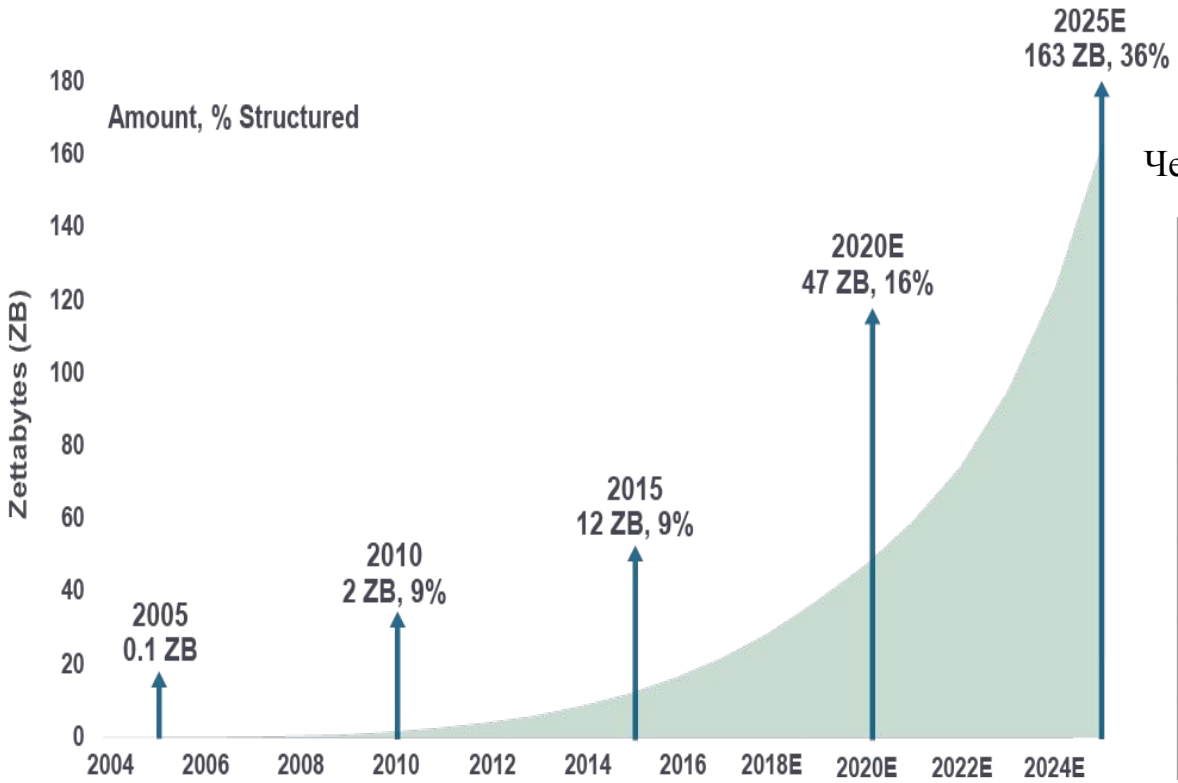
Сферы производящие большие данные.
Data explosion.

Откуда пришла Big Data

Сферы:

- Телеком
- Банки
- Социальные сети
- Медиа
- Промышленность
- Биоинформатика
- Интернет вещей

Data explosion.



Чему равен зеттабайт? - триллиону гигабайт

1 kilobyte	**1 000**
1 megabyte	**1 000 000**
1 gigabyte	**1 000 000 000**
1 terabyte	**1 000 000 000 000**
1 petabyte	**1 000 000 000 000 000**
1 exabyte	**1 000 000 000 000 000 000**
1 zettabyte	**1 000 000 000 000 000 000 000**

Большие данные, где начало. 3 основных принципа.

?
Veracity

Volume

Зеттабайты данных
6 миллиардов людей имеют телефон

- Терабайты
- Записи
- Транзакции
- Таблицы, файлы

‘десятки миллионов номеров телефонов’

‘миллионы транзакций’

‘youtube - 3 миллиарда просмотров в день’

Velocity

Твиты
Посты в facebook
Датчики устройств

‘Каждую минуту производится 511,200 твитов’

- Скорость генерируемых данных
- генерируемые в реал-тайм
- онлайн и оффлайн данные
- По стримам, батчам ли битам

Variety

Видео
Аудио
Текстовые данные
Временные ряды

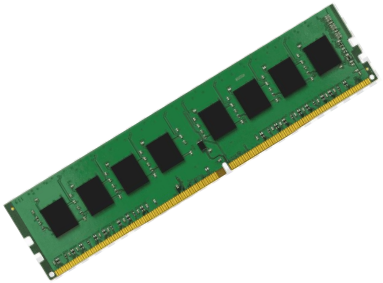
- Структурированые
- Неструктурированые
- Полуструктурированные

Как компании справляются с большими данными Iaas/Paas/Saas.

IaaS/PaaS/SaaS.

- **IaaS — это Infrastructure as a Service.** Инфраструктура как услуга. К инфраструктуре относят вычислительные ресурсы: виртуальные серверы, хранилища, сети.
  - Перенос IT-систем в облако.
  - Экономия на инфраструктуре.
  - Быстрый запуск бизнеса.
  - Расширение инфраструктуры.
  - Инфраструктура для компаний со скачками спроса.
  - Разработка и тестирование.
- PaaS - **это Platform as a Service**, платформа как услуга.
  - Базы данных.
  - Разработка приложений в контейнерах.
  - Аналитика больших данных.
  - Машинное обучение.
- SaaS — **это Software as a Service**, программное обеспечение как сервис
  - электронная почта
  - CRM-системы
  - планировщики задач
  - веб-конструкторы для создания сайтов

Оперативная память, жесткий диск. Сортировка во внешней памяти.



Оперативная память (англ. *Random Access Memory, RAM* — память с произвольным доступом) — энергозависимая часть системы компьютерной памяти, в которой во время работы компьютера хранится выполняемый машинный код (программы), а также входные, выходные и промежуточные данные, обрабатываемые процессором.

Жёсткий диск (англ. *hard drive, HDD*)— запоминающее устройство (устройство хранения информации, накопитель) произвольного доступа, основанное на принципе магнитной записи. Является основным накопителем данных в большинстве компьютеров.

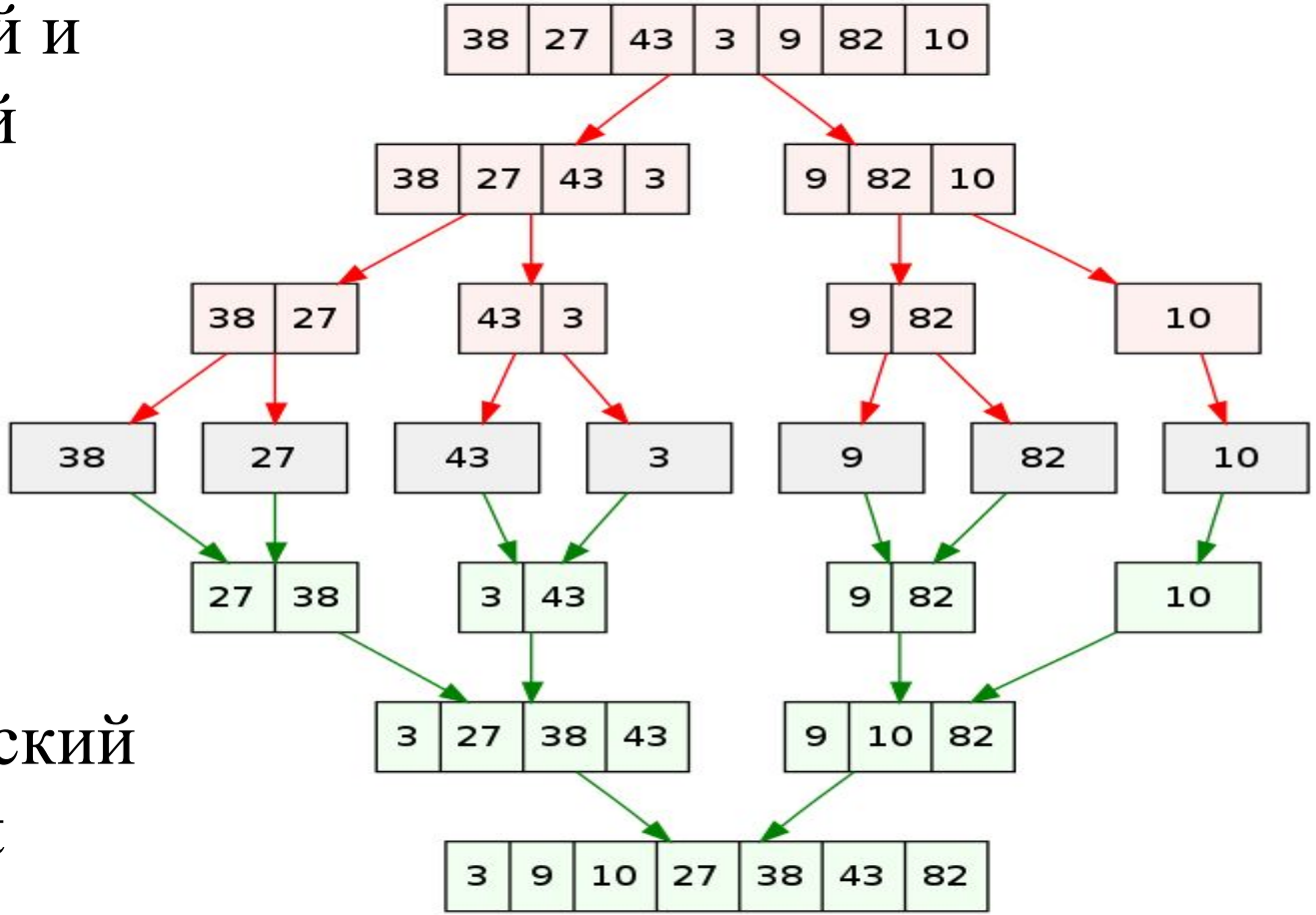
Центральный процессор (англ. *central processing unit*, *CPU*) — электронный блок либо интегральная схема, исполняющая (код программ).

Задача:

Отсортировать массив

38	27	43	3	9	82	10

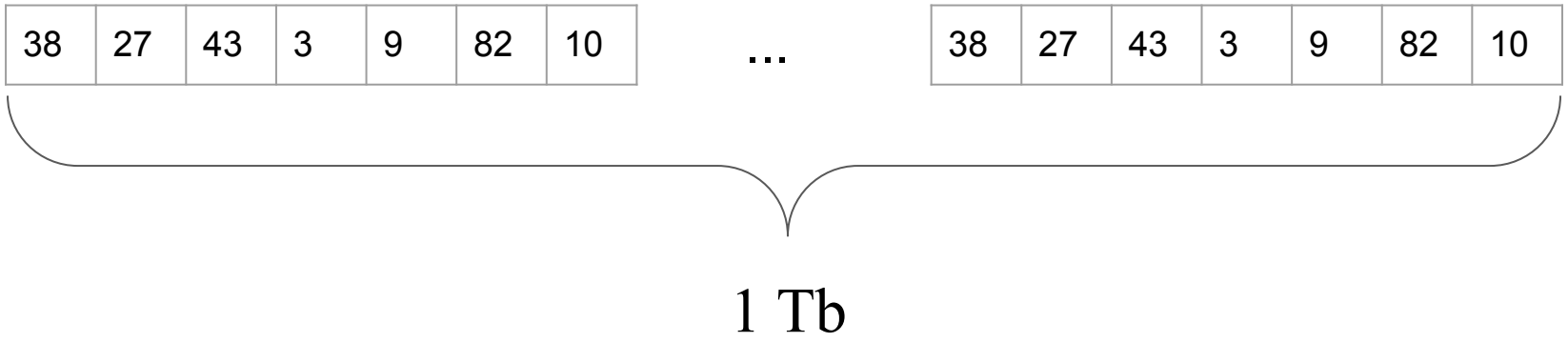
Разделяй и властвуй

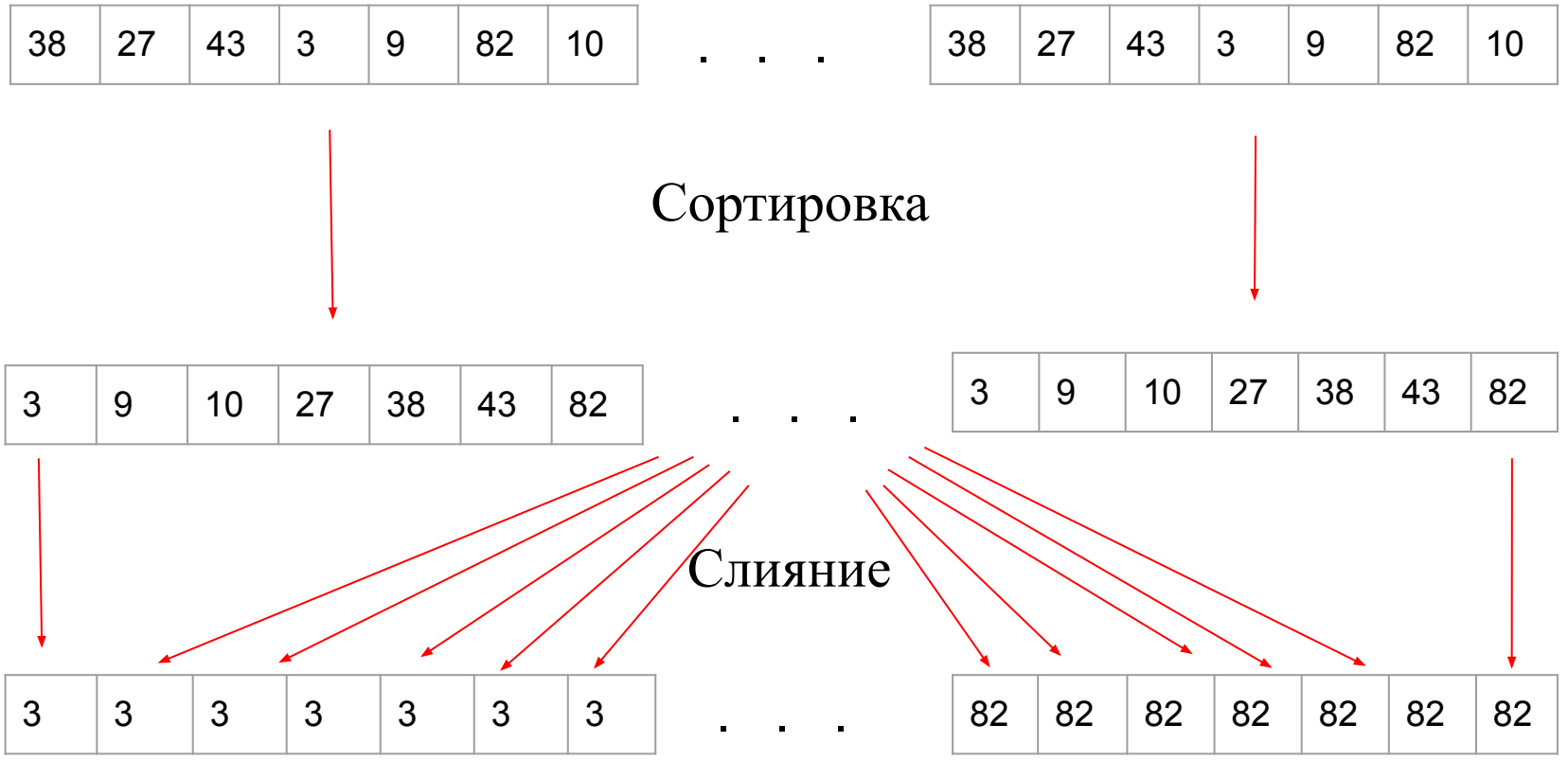


Классический merge sort

Задача:

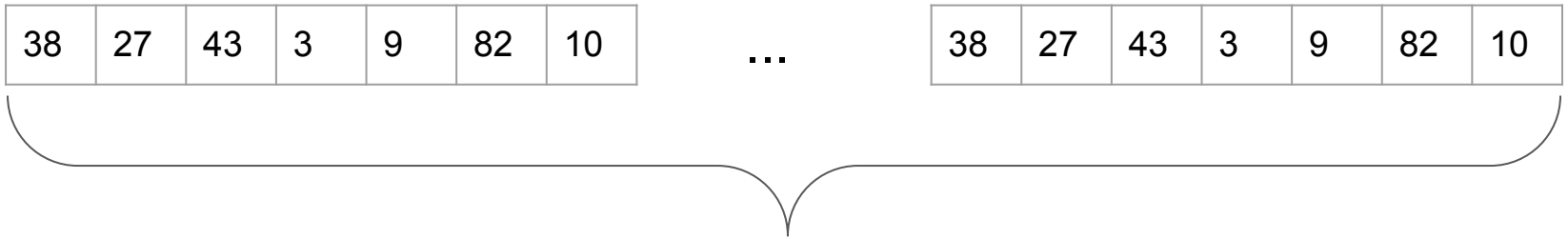
Отсортировать массив, который не помещается в оперативную память.



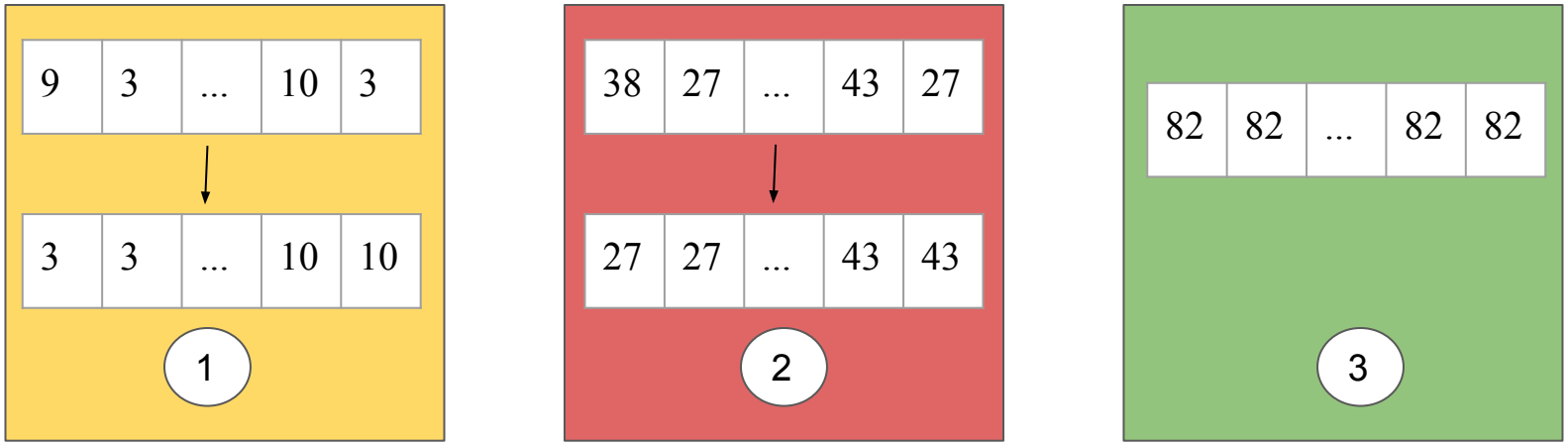

38 27 43 3 9 82 10
. . .
38 27 43 3 9 82 10
Сортировка
3 9 10 27 38 43 82
. . .
3 9 10 27 38 43 82
Слияние
3 3 3 3 3 3 3
. . .
82 82 82 82 82 82 82

Задача:

Отсортировать массив, который не помещается на доступный жесткий диск.




38 27 43 3 9 82 10
...
38 27 43 3 9 82 10


9 3 ... 10 3
3 3 ... 10 10
1
38 27 ... 43 27
27 27 ... 43 43
2
82 82 ... 82 82
3

Плюсы и минусы распределенных систем. Предпосылки к созданию MapReduce.

Плюсы:

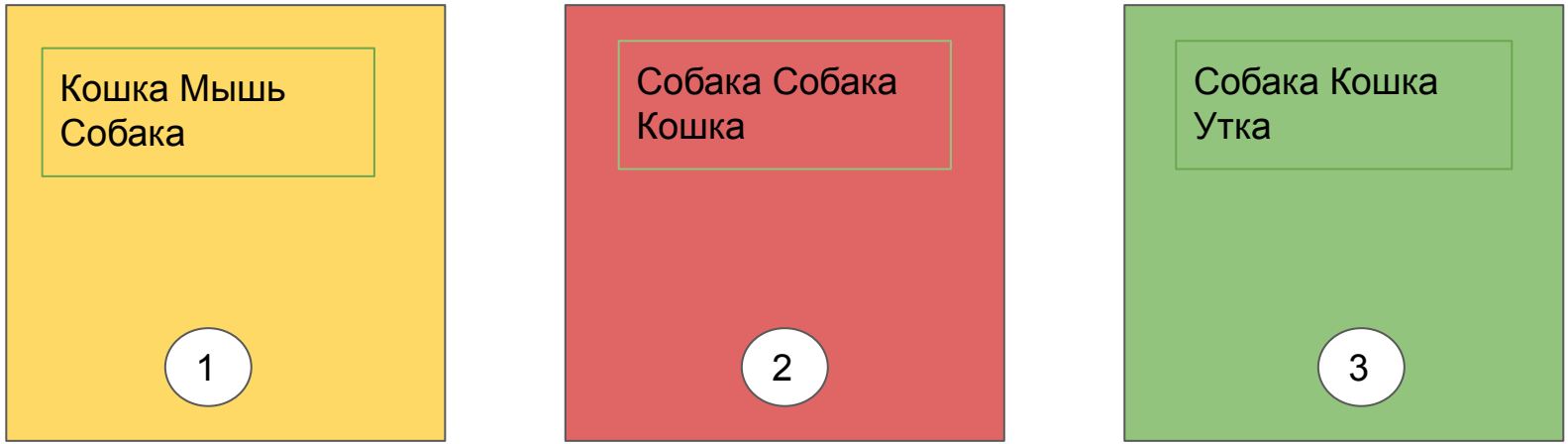
1. Высокая производительность
2. Отказоустойчивость
3. Поддержка физической удаленности ресурсов

Минусы:

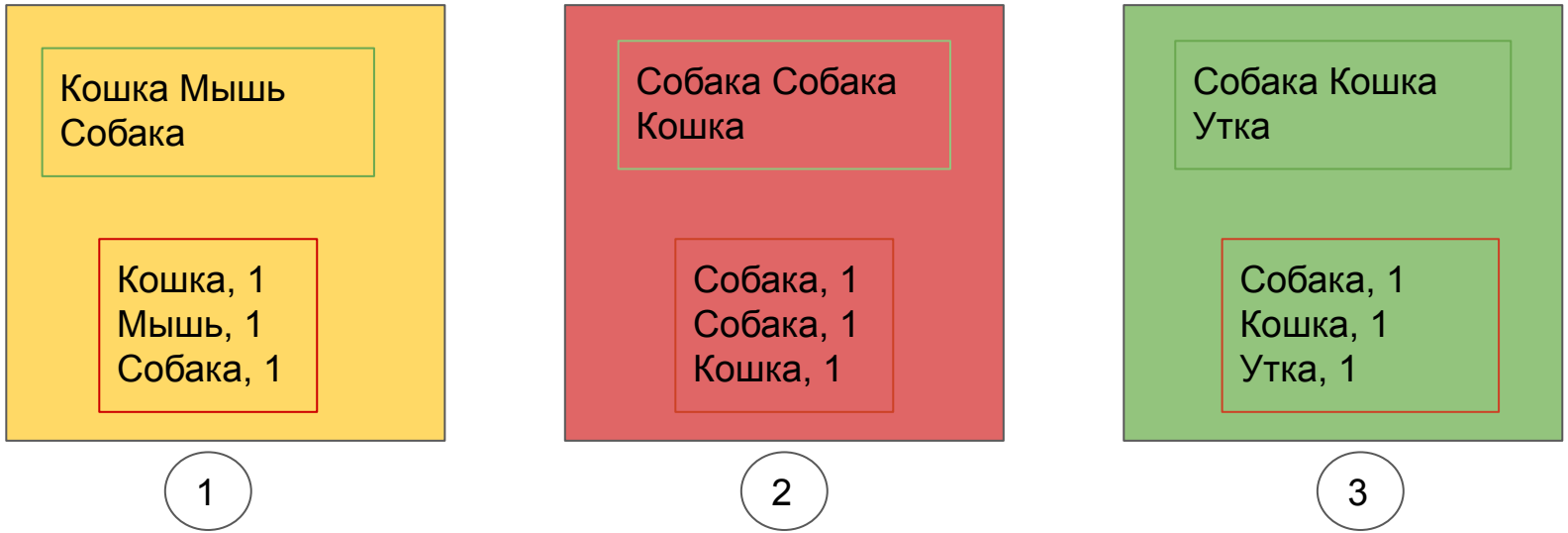
1. Большое количество задач
2. Конкуренция за ресурсы
3. Частичные падения
4. Неочевидные схемы падения
5. Передача данных по сети
6. Выигрыш вычислениях - тонкая настройка ресурсов и передачи данных

И всегда стоит помнить, что если вы можете решить задачу без использования распределенной системы, скажем с использованием только одного компьютера, то стоит отойти от ее создания.

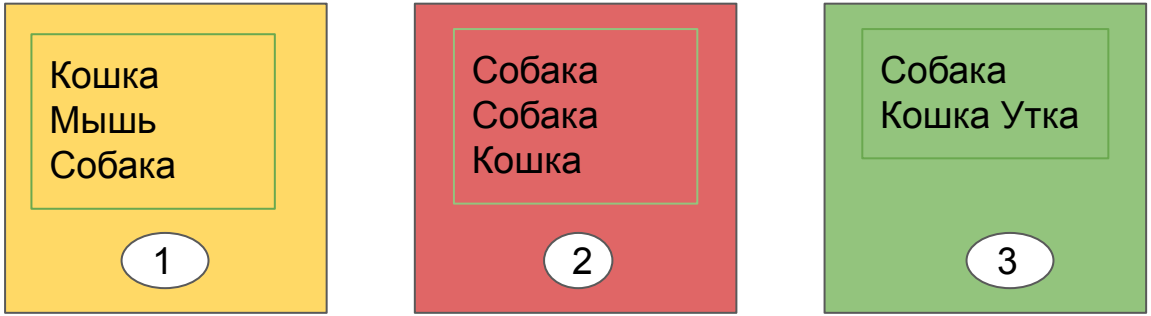
Задача подсчета слов. Map. Shuffle. Reduce.


Кошка Мышь
Собака
1
Собака Собака
Кошка
2
Собака Кошка
Утка
3

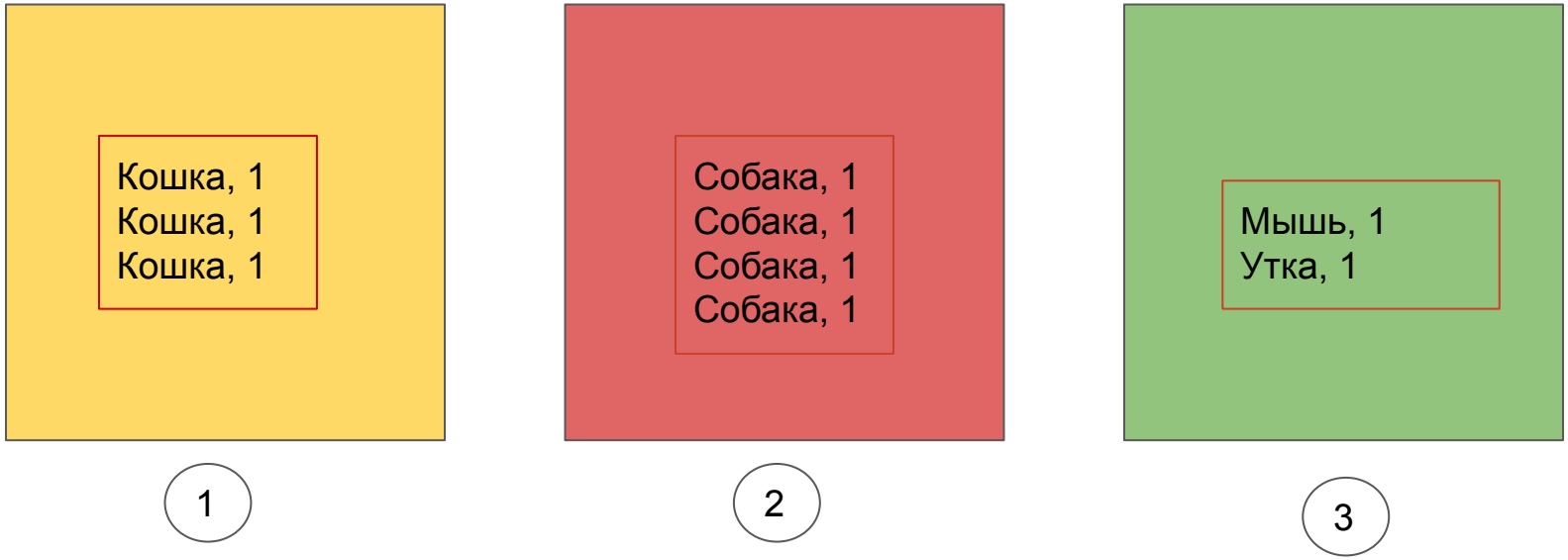
Map:


Кошка Мышь
Собака
Кошка, 1
Мышь, 1
Собака, 1
1
Собака Собака
Кошка
Собака, 1
Собака, 1
Кошка, 1
2
Собака Кошка
Утка
Собака, 1
Кошка, 1
Утка, 1
3

Shuffle:



По сути - сортировка:



Reduce: