Промышленное машинное обучение на Spark

Содержание курса

01 Как работают и где живут большие данные.

02 Погружение среду Spark. Spark RDD / Spark SQL. Part1.

03 Погружение среду Spark. Spark RDD / Spark SQL. Part2.

04 Spark ML Part 1

05 Spark ML Part 2 & Model validation

06 Spark Streaming

07 Spark Ecosystem (MLFlow, AirFlow, H2O AutoML)

08 Spark в архитектуре проекта / Spark CI/CD

1. Как работают и где живут большие данные.

План:

1. Сферы производящие большие данные. Data explosion.
2. Большие данные - где начало. 3 основных принципа.
3. Отказоустойчивость вычислений.
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

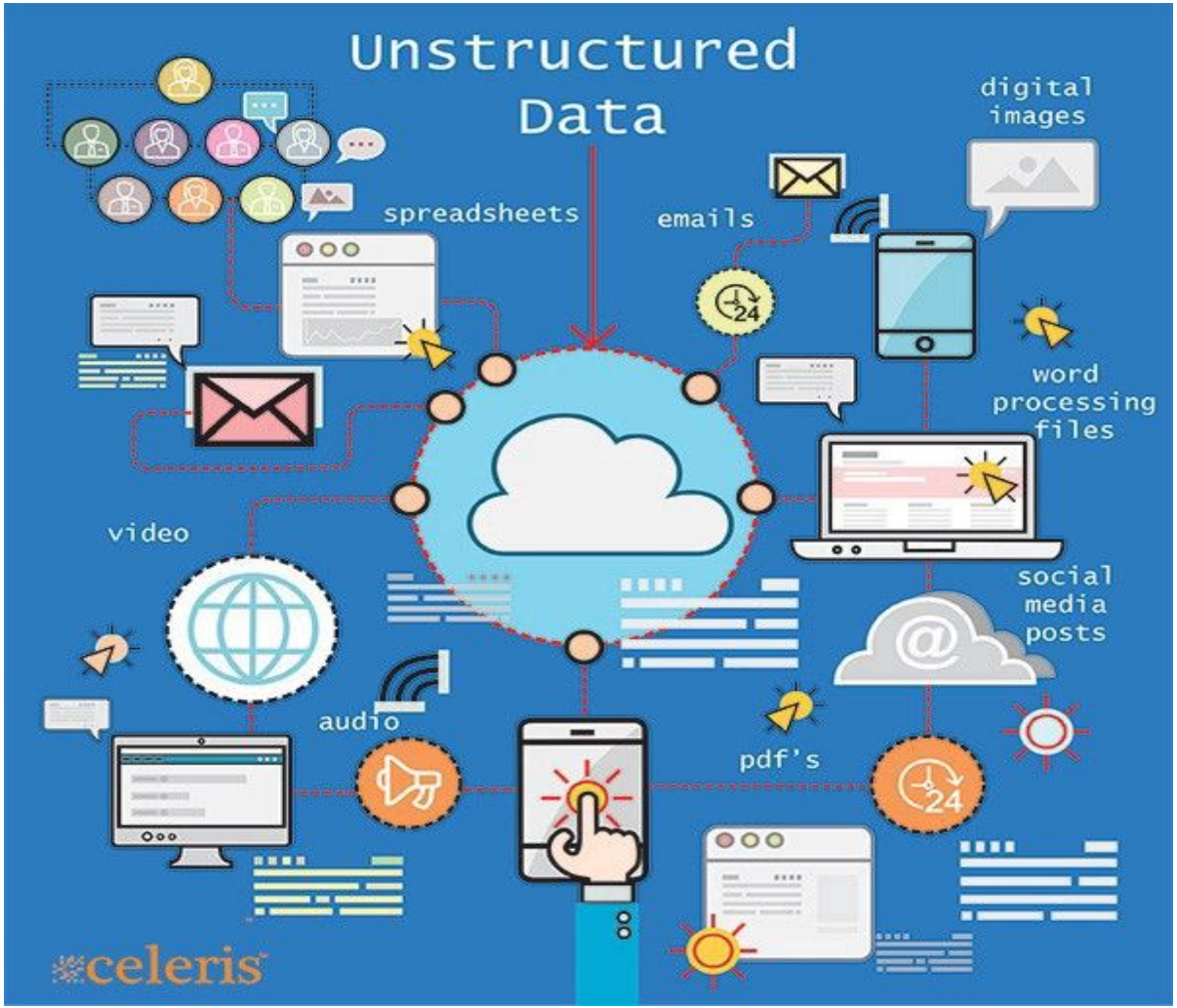
На первое время всем рекомендую зарегистрироваться на https://databricks.com/ выбрать community edition. Там будут доступна работа с ноутбуком и с установленным Spark.

Сферы производящие большие данные.
Data explosion.

Откуда пришла Big Data

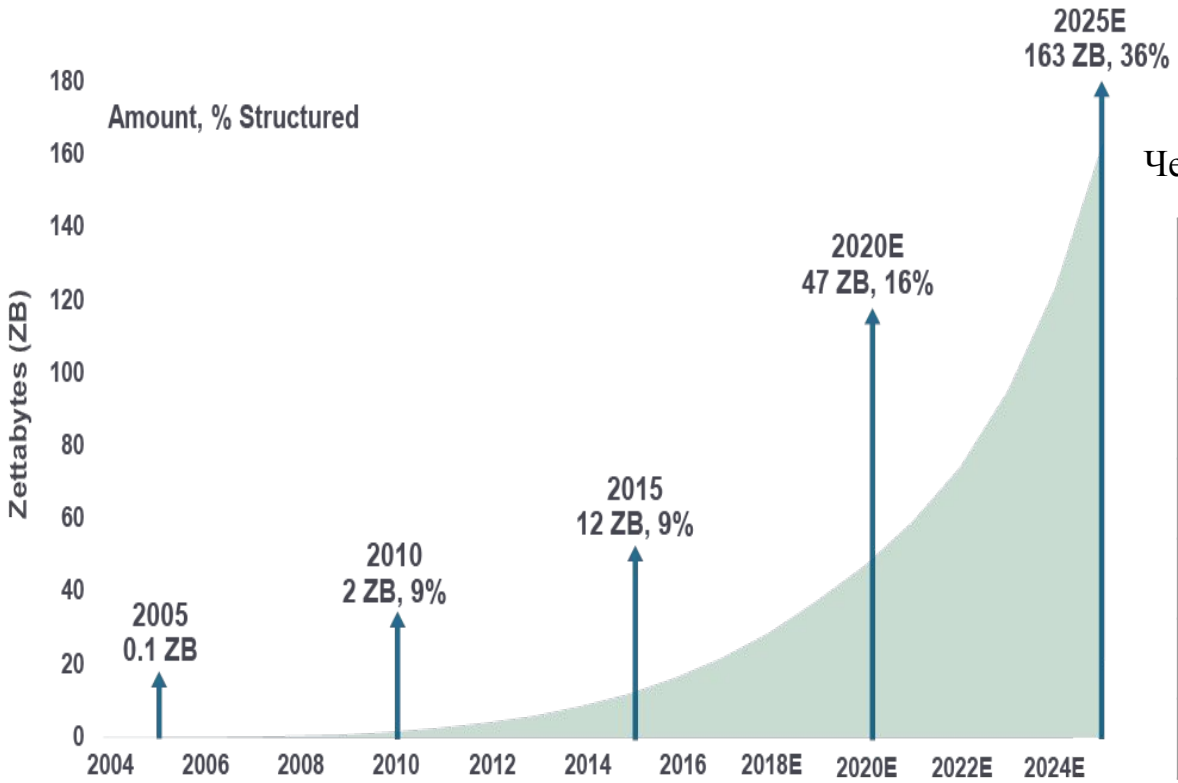
Сферы:

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



Data Explosion

Data explosion.

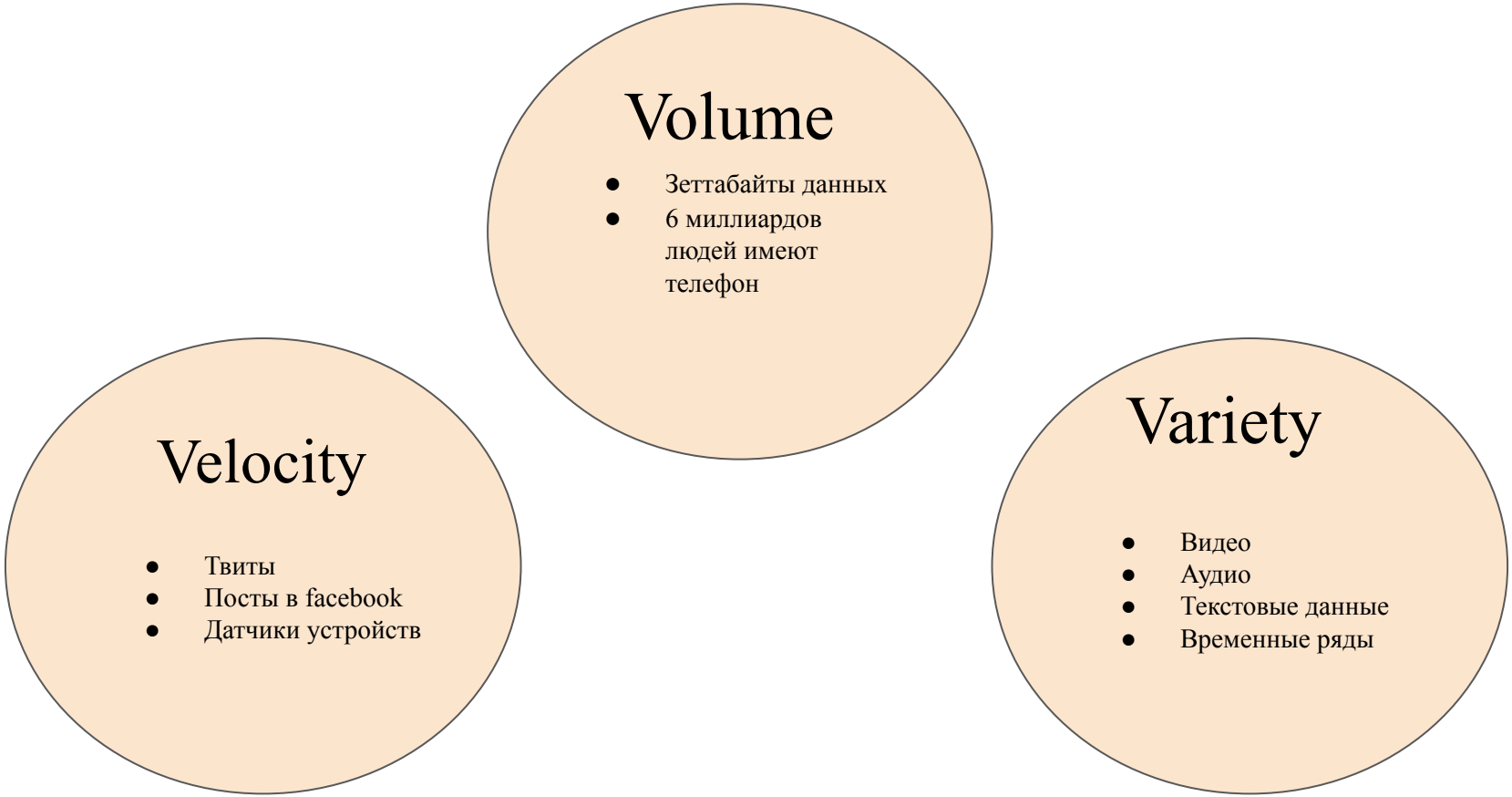


Чему равен зеттабайт? - триллиону гигабайт

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.

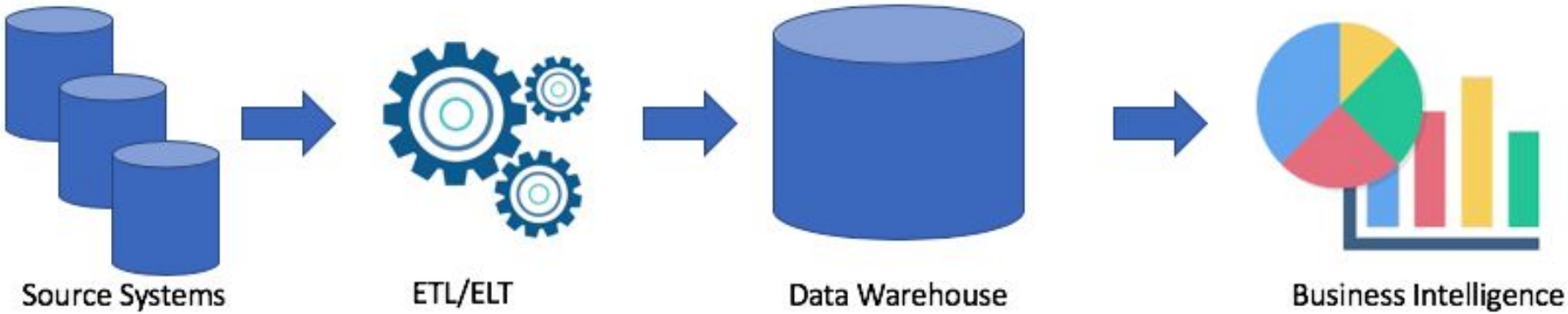
Большие данные, где начало. 3 основных принципа.



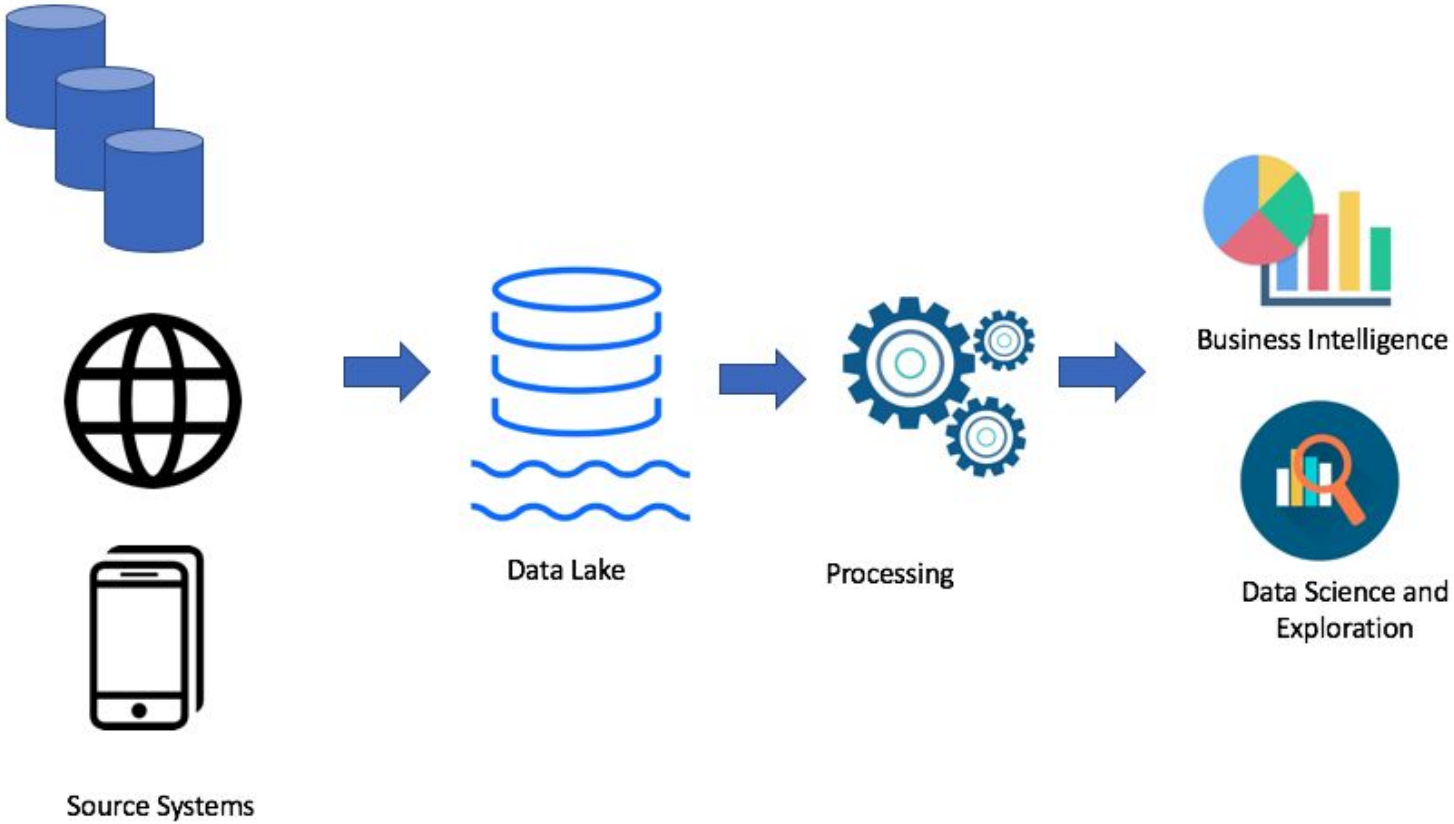
Данные - их нужно где-то надежно хранить, да еще и быстро и эффективно обрабатывать обрабатывать

Как компании справляются с большими данными. Data Lake & Data Warehouse

Data Warehouse



Data Lake



https://habr.com/ru/post/485180/

А что предлагает рынок?

- **IaaS — это Infrastructure as a Service.** Инфраструктура как услуга. К инфраструктуре относят вычислительные ресурсы: виртуальные серверы, хранилища, сети. (Google Compute Engine, DigitalOcean, Amazon Web Services (AWS), and Cisco Metacloud).
    - 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. (Windows Azure, OpenShift, Heroku, and Google App Engine).
    - Базы данных.
    - Разработка приложений в контейнерах.
    - Аналитика больших данных.
    - Машинное обучение.
- SaaS — **это Software as a Service**, программное обеспечение как сервис (Google App Engine, Dropbox, JIRA, and others).
    - электронная почта
    - CRM-системы
    - планировщики задач
    - веб-конструкторы для создания сайтов

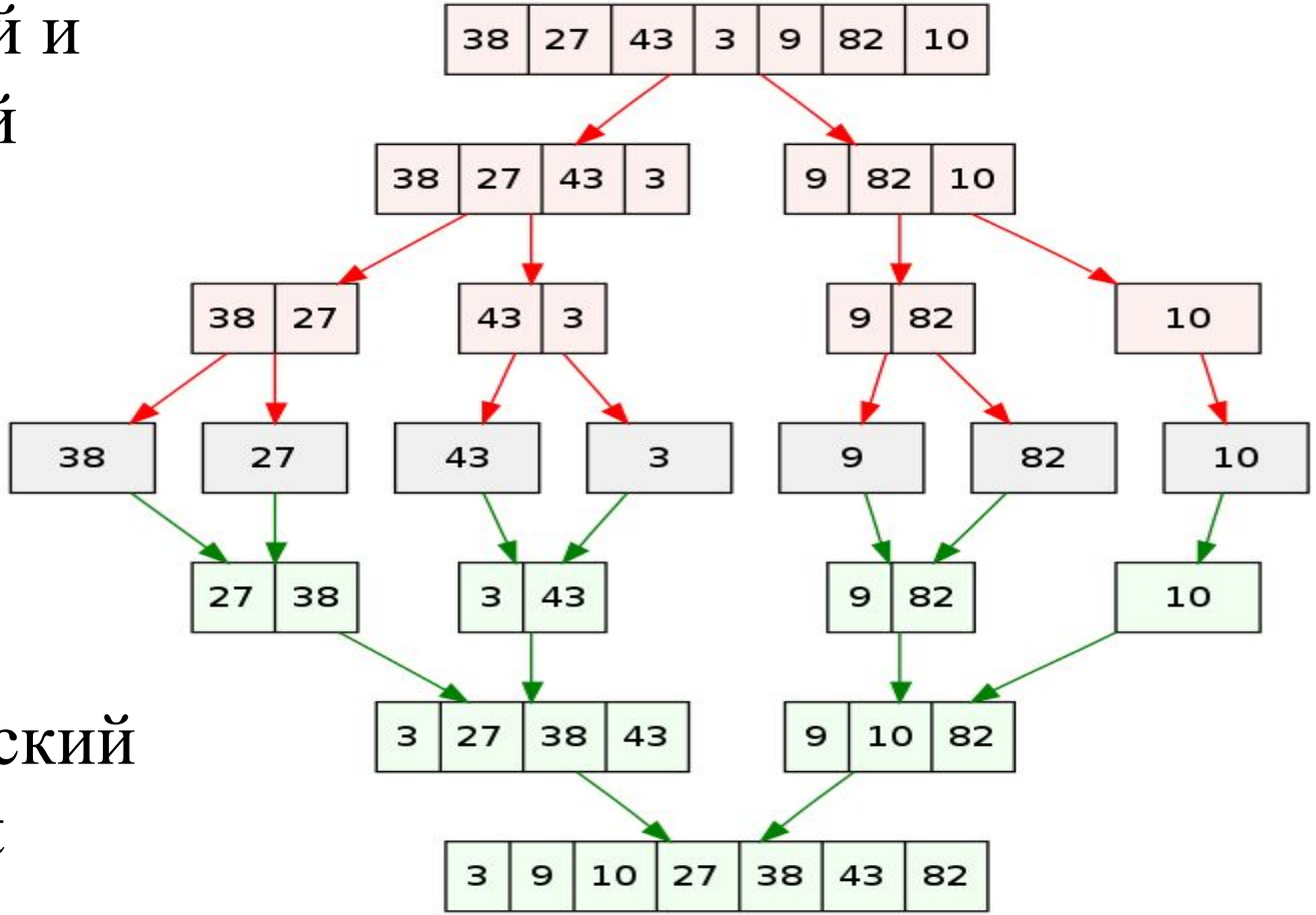
Теперь про вычисления!

Задача:

Отсортировать массив

38	27	43	3	9	82	10

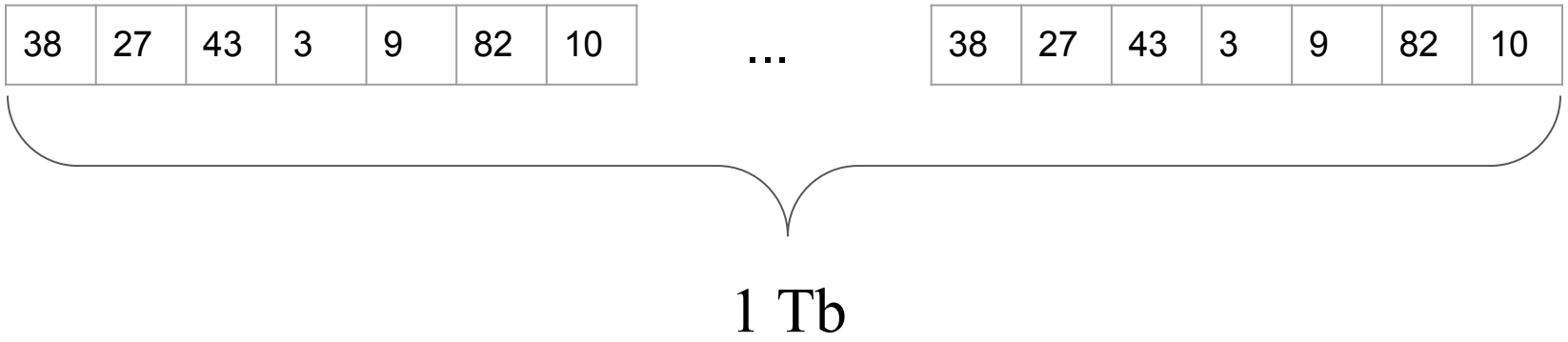
Разделяй и властвуй

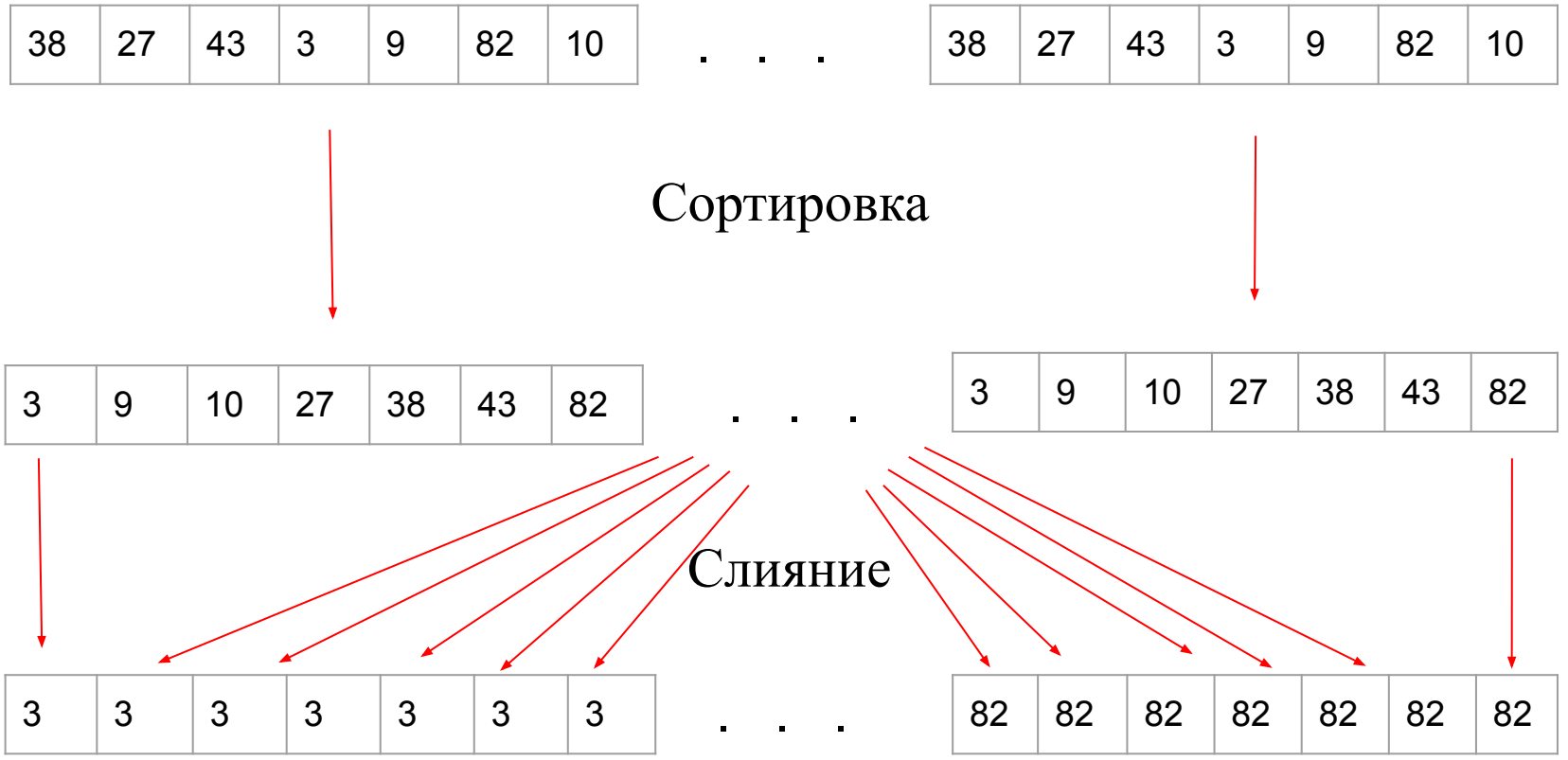


Классический merge sort

Задача:

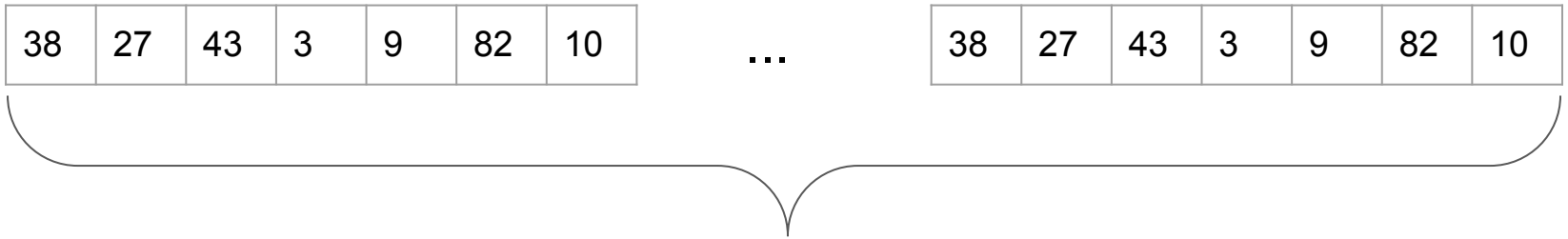
Отсортировать массив, который не помещается в оперативную память.



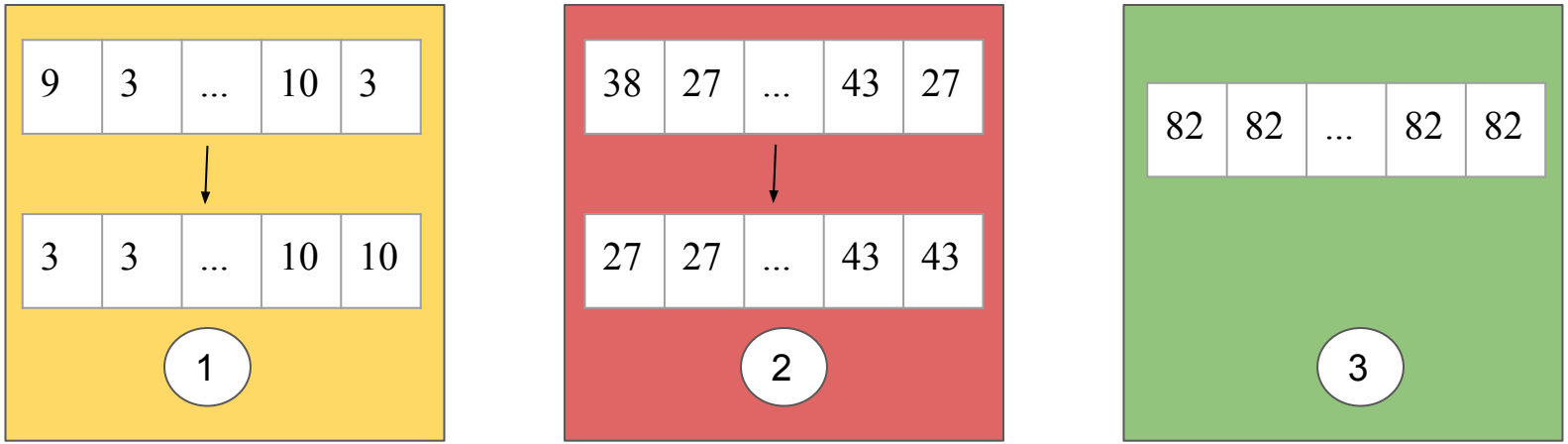

38 27 43 3 9 82 10
. . .
38 27 43 3 9 82 10
Сортировка
3 9 10 27 38 43 82
. . .
3 9 10 27 38 43 82
Слияние
3 3 3 3 3 3 3
. . .
82 82 82 82 82 82 82

Задача:

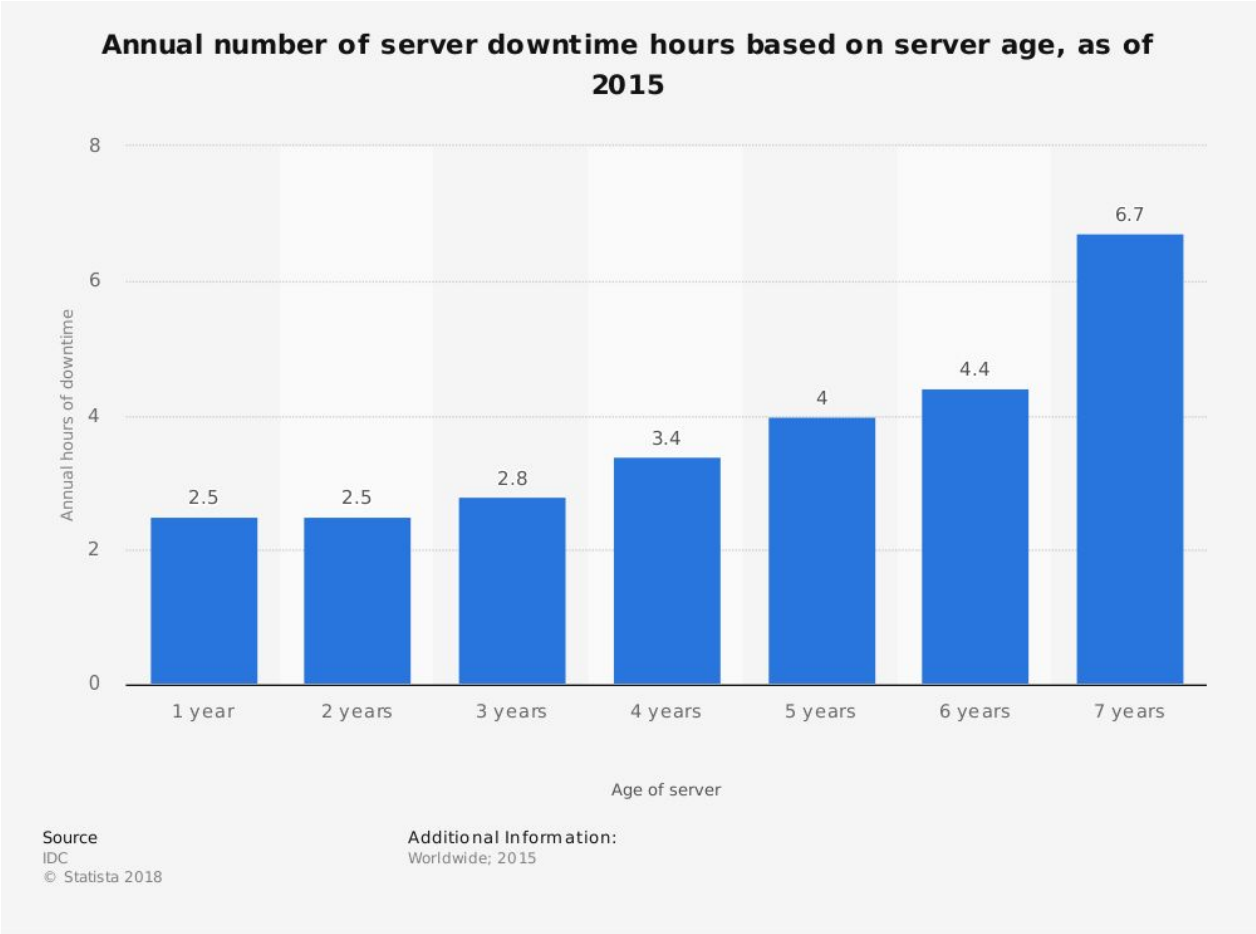
Отсортировать массив, который не помещается на доступный жесткий диск.




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

Отказоустойчивость вычислений.



Вероятность, что в следующий час случится поломка

$P = 2.5 / (24 * 365) = 0.00028$

$P(\text{не выйдет из строя}) = (1 - P) = 0.9997$

1000 машин в кластере

Вероятность, что какой-нибудь сломается ближайший час

$1 - (0.9997)^{1000} = 0.25$

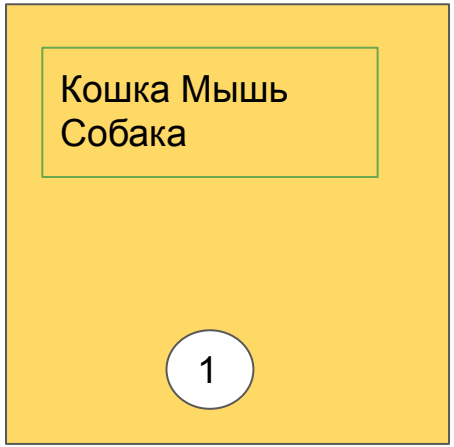
Задача подсчета слов. Map. Shuffle. Reduce.

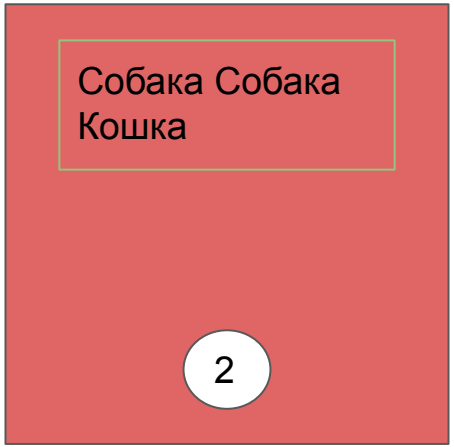
Задача подсчета слов

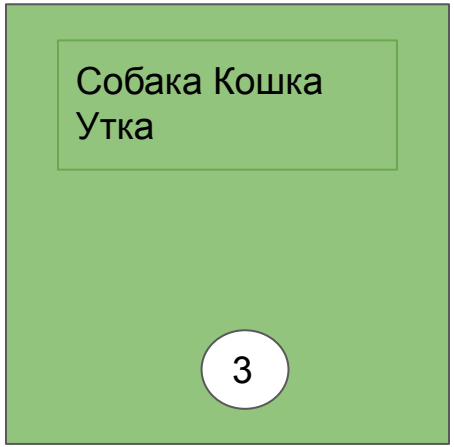
Кошка Мышь Собака

Собака Собака Кошка

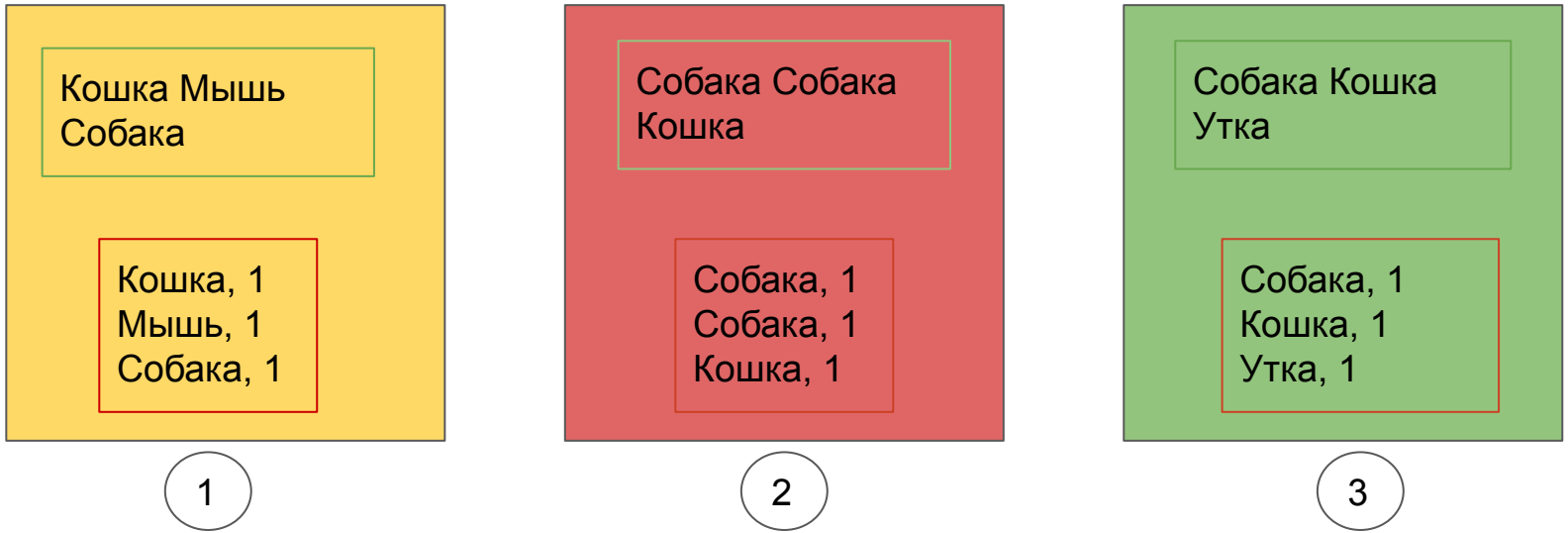
Собака Кошка Утка



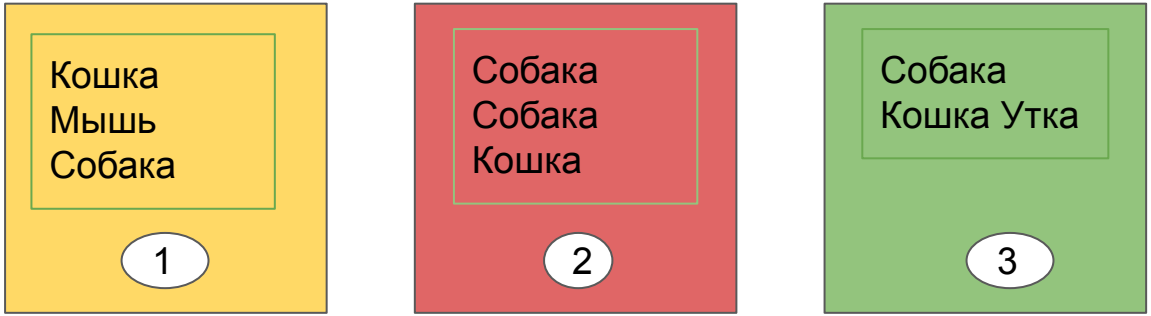




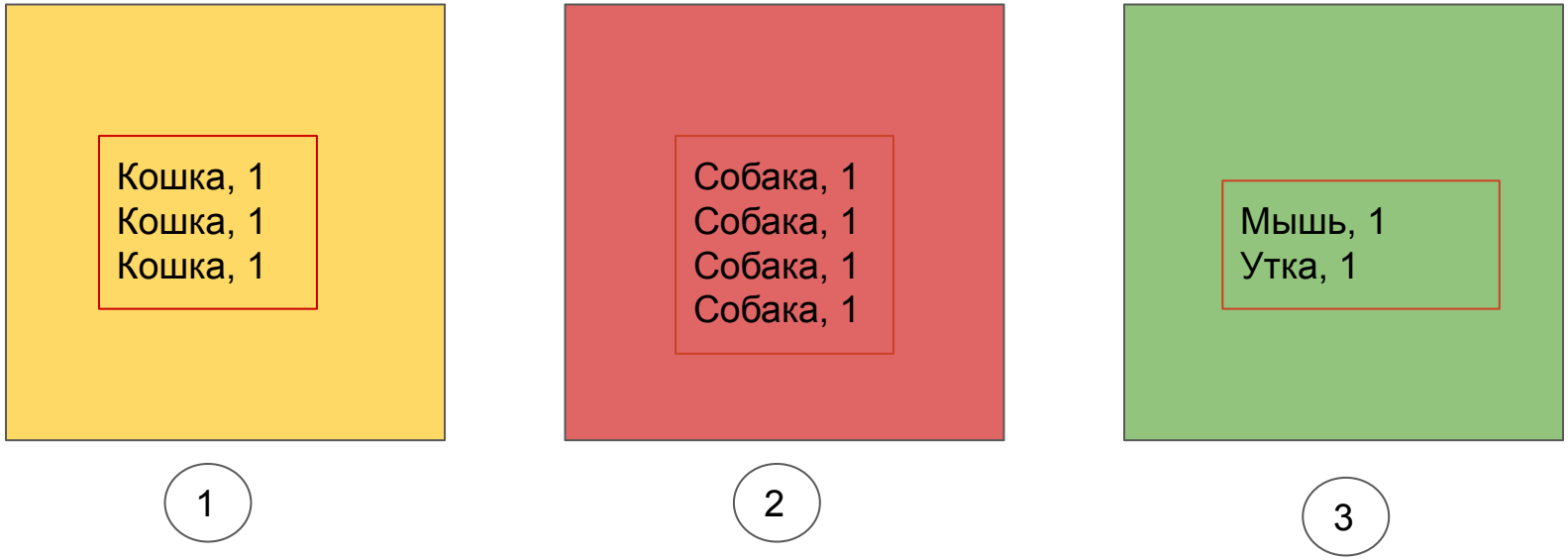
Map:


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

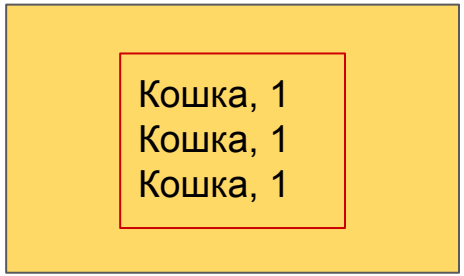
Shuffle:



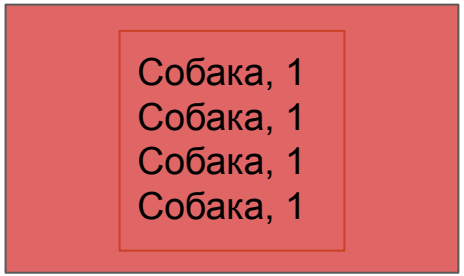
По сути - сортировка:



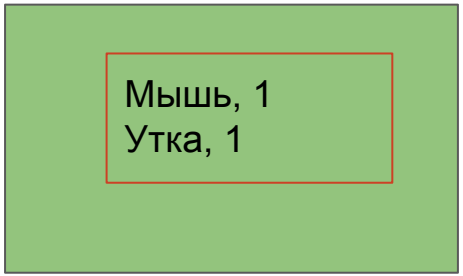
Reduce:



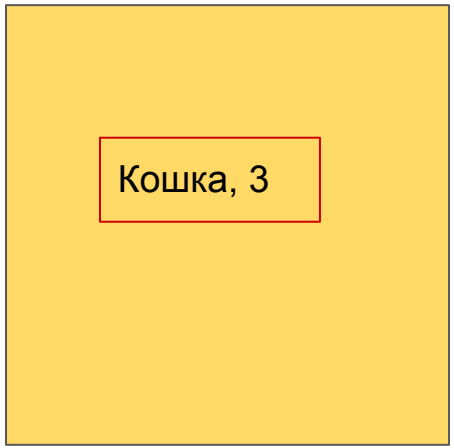
1



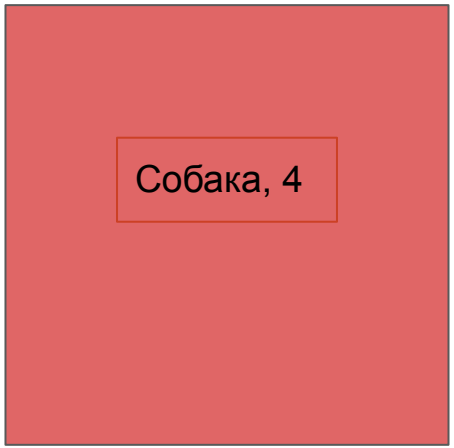
2



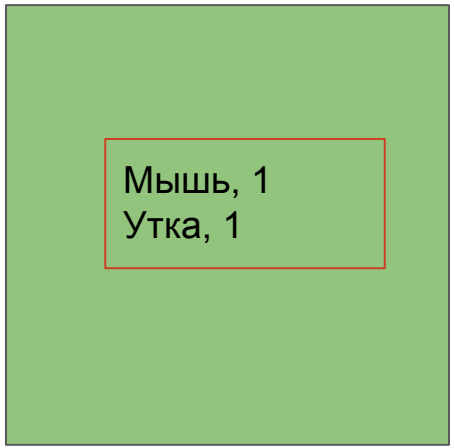
3



1



2



3

И где тут Apache Spark?

Вызовы, с которыми столкнулись, пытаясь обработать данные

- Дорогое оборудование - если хотим вычисления на одном компьютере
- Поломки дешевого оборудования - если используем кластер
- Типичный файл в распределенной файловой системе имеет размер от гигабайтов до терабайт.
- Вычисления занимают множество времени на одном компьютере
- Если процессов много - как корректно организовать пересылку

Все вызовы, брошенные выше, решаются красиво и userfrendly благодяря!

- Открытый исходный код
- Целый набор библиотек для обработки данных на кластерах компьютеров
- Самый активно разрабатываем фреймворк с открытым исходным кодом!
- Поддержка Python, Scala, R
- Поддержка приятного и привычного SQL
- Библиотеки - для ML, анализа графов, обработки стриминговых данных,

Плюсы и минусы распределенных систем.

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.

Полезные определения

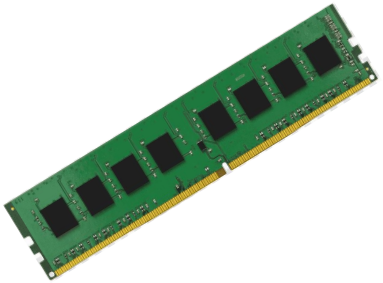
Кластер - совокупность компьютеров объединенных сетью и выполняющих задачи, посылаемые клиентами.

Нода (Node) - один из компьютеров подключенных к кластеру.

Стойка (Rack) - совокупность нескольких нод, объединенных сетью

Демон - компьютерная программа в системах класса UNIX, запускаемая самой системой и работающая в фоновом режиме без прямого взаимодействия с пользователем

Клиент - это аппаратный или программный компонент вычислительной системы, посылающий запросы серверу.



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).

Литература\ссылки

https://www.ozon.ru/context/detail/id/34973964/ - **Hadoop: The Definitive Guide | Уайт Том**

Статья на Medium с полезными книгами