Word Count и формальная модель MapReduce

Драль Алексей, study@bigdatateam.org

CEO at BigData Team, https://bigdatateam.org

https://www.facebook.com/bigdatateam

Word Count

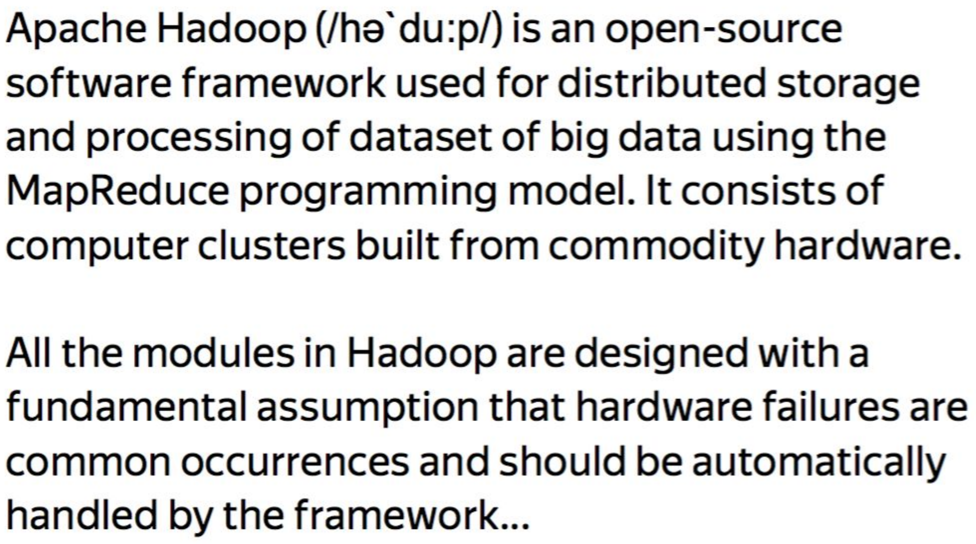
Apache Hadoop (/hə`du:p/) is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware.

All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework...

Wikipedia
The Free Encyclopedia

hadoop
Map Reduce

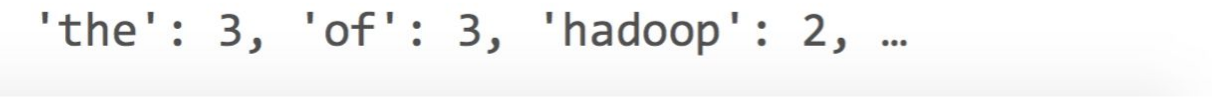
Word Count



Apache Hadoop (/hə`du:p/) is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware.

All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework...

↓



```
'the': 3, 'of': 3, 'hadoop': 2, …
```

Word Count на одном компьютере

```
$ cat dataset.txt
```

```
   Apache Hadoop is a collection of open-source
software utilities that facilitates using a
network of many computers to solve problems
involving massive amounts of data and computation.
It provides a software framework for distributed
storage and processing of big data using the
MapReduce programming model...
```

Word Count на одном компьютере

```
$ cat dataset.txt | tr ' ' '\n'
```

```
   Apache
   Hadoop
   is
   a
   collection
   of
   ...
```

Word Count на одном компьютере

```
$ cat dataset.txt | tr ' ' '\n' | sort
```

```
   All
   Apache
   Hadoop
   Hadoop
   Hadoop
   It
   ...
```

Word Count на одном компьютере

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
```

```
   1 All
   1 Apache
   3 Hadoop
   2 It
   1 MapReduce
   4 a
   ...
```

Распределенный Word Count

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
```

```
    1 All
    1 Apache
    3 Hadoop
    2 It
    1 MapReduce
    4 a
    ...
```

Распределенный Word Count

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
   1 All
   1 Apache
   3 Hadoop
   2 It
   1 MapReduce
   4 a
   ...
```

Распределенный Word Count

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
```

```
      1 All
      1 Apache
      3 Hadoop
      2 It
      1 MapReduce
      4 a
      ...
```

Что делаем с “sort”?

Распределенный Word Count

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
```

❌ Фаза Map (нужна агрегация)

Распределенный Word Count

```
$ cat dataset.txt | tr ' ' '\n' | sort | uniq -c
```

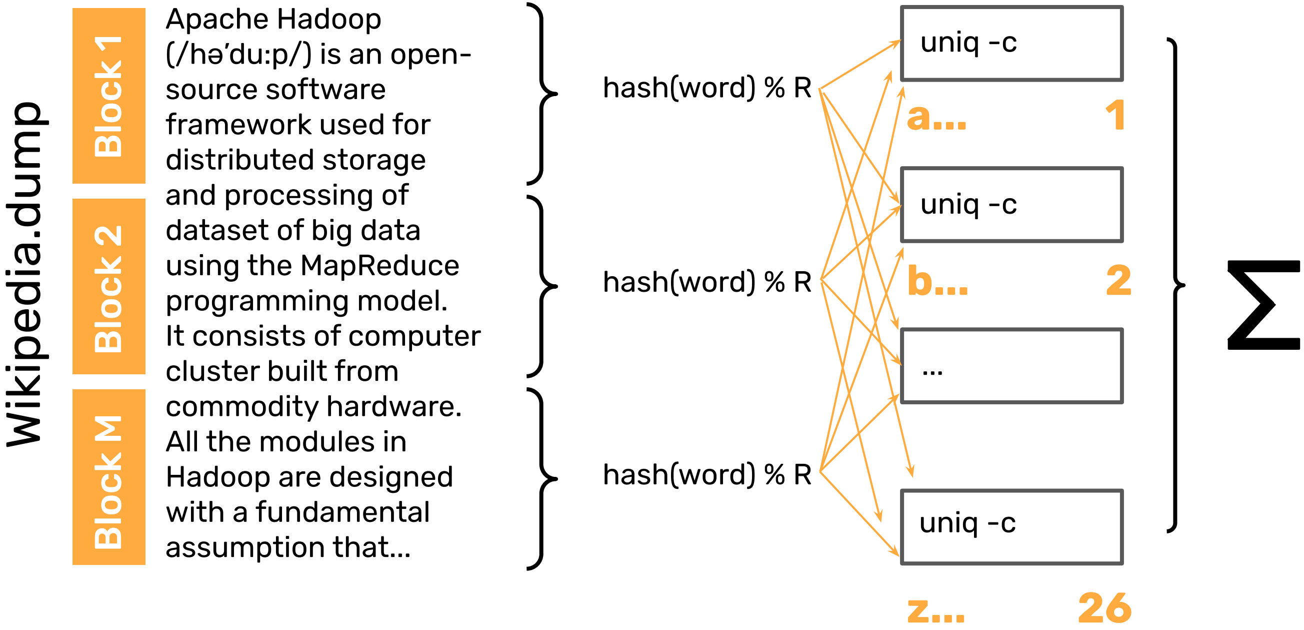
❌ Фаза Map (нужна агрегация)
❌ Фаза Reduce (не хватит RAM / HDD)

MapReduce =

Map → Shuffle & Sort → Reduce

Word Count детализация

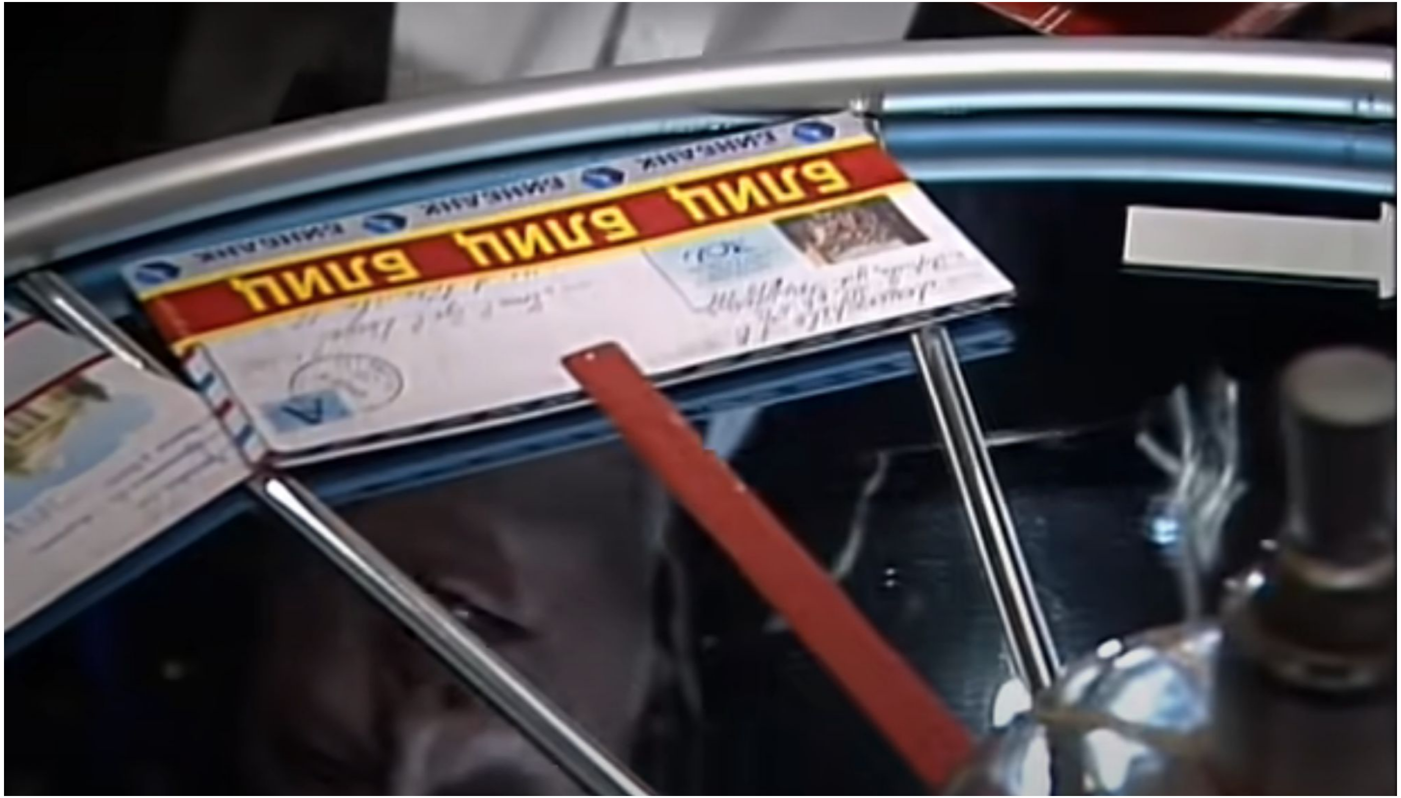
wikipedia.dump | tr ' ' '\n' | sort | uniq -c



Apache Hadoop (/hə'du:p/) is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer cluster built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that...

wikipedia.dump -> map () -> word shuffle & sort reduce()

Блиц-турнир



Формальная модель MapReduce

Формальная модель MapReduce

Фазы:

Формальная модель MapReduce

Фазы:

1. Map

BIGDATA TEAM

Формальная модель MapReduce

Фазы:

1. Map
2. Shuffle & Sort

Формальная модель MapReduce

Фазы:

1. Map
2. Shuffle & Sort
3. Reduce

Формальная модель MapReduce

Фазы:

1. Map
2. Shuffle & Sort
3. Reduce

Worker’ы (контейнеры):

Формальная модель MapReduce

Фазы:

1. Map
2. Shuffle & Sort
3. Reduce

Worker’ы (контейнеры):

- Фаза Map → Mapper (использует функцию map)

Формальная модель MapReduce

Фазы:

1. Map
2. Shuffle & Sort
3. Reduce

Worker’ы (контейнеры):

- Фаза Map → Mapper (использует функцию map)
- Фаза Reduce → Reducer (использует функцию reduce)

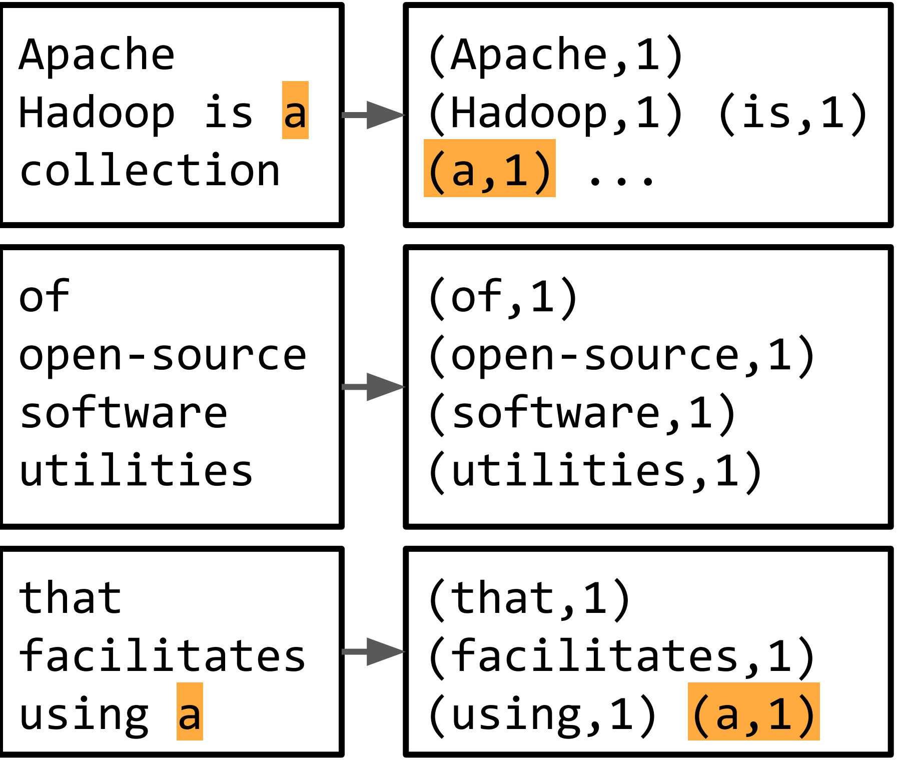
Word Count v.2

```
Apache
Hadoop is a
collection
```

```
of
open-source
software
utilities
```

```
that
facilitates
using a
```

Word Count v.2



Apache
Hadoop is a
collection

(Apache,1)
(Hadoop,1) (is,1)
(a,1) ...

of
open-source
software
utilities

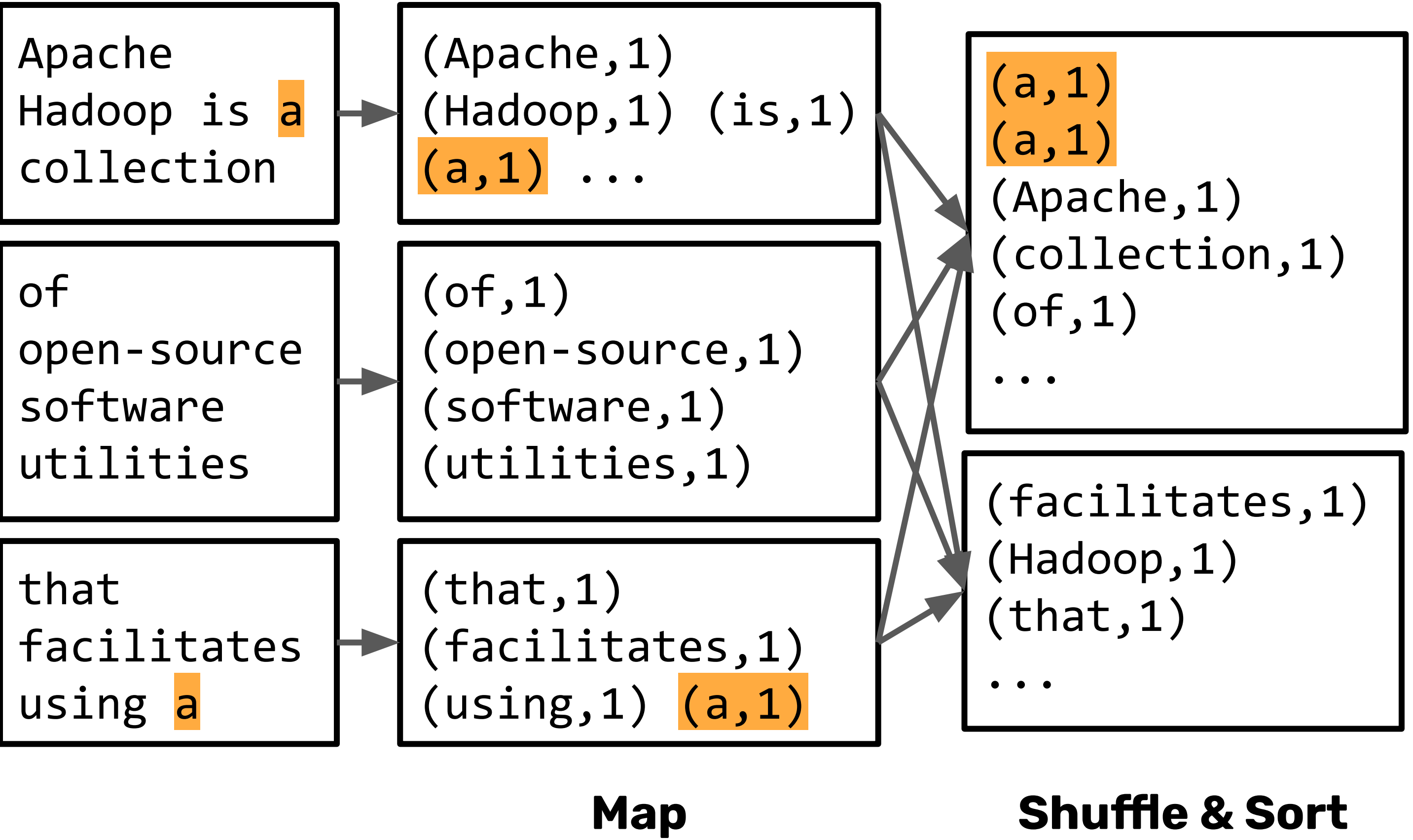
(of,1)
(open-source,1)
(software,1)
(utilities,1)

that
facilitates
using a

(that,1)
(facilitates,1)
(using,1) (a,1)

Map

Word Count v.2



Apache Hadoop is a collection

of open-source software utilities

that facilitates using a

(Apache,1) (Hadoop,1) (is,1) (a,1) ...

(of,1) (open-source,1) (software,1) (utilities,1)

(that,1) (facilitates,1) (using,1) (a,1)

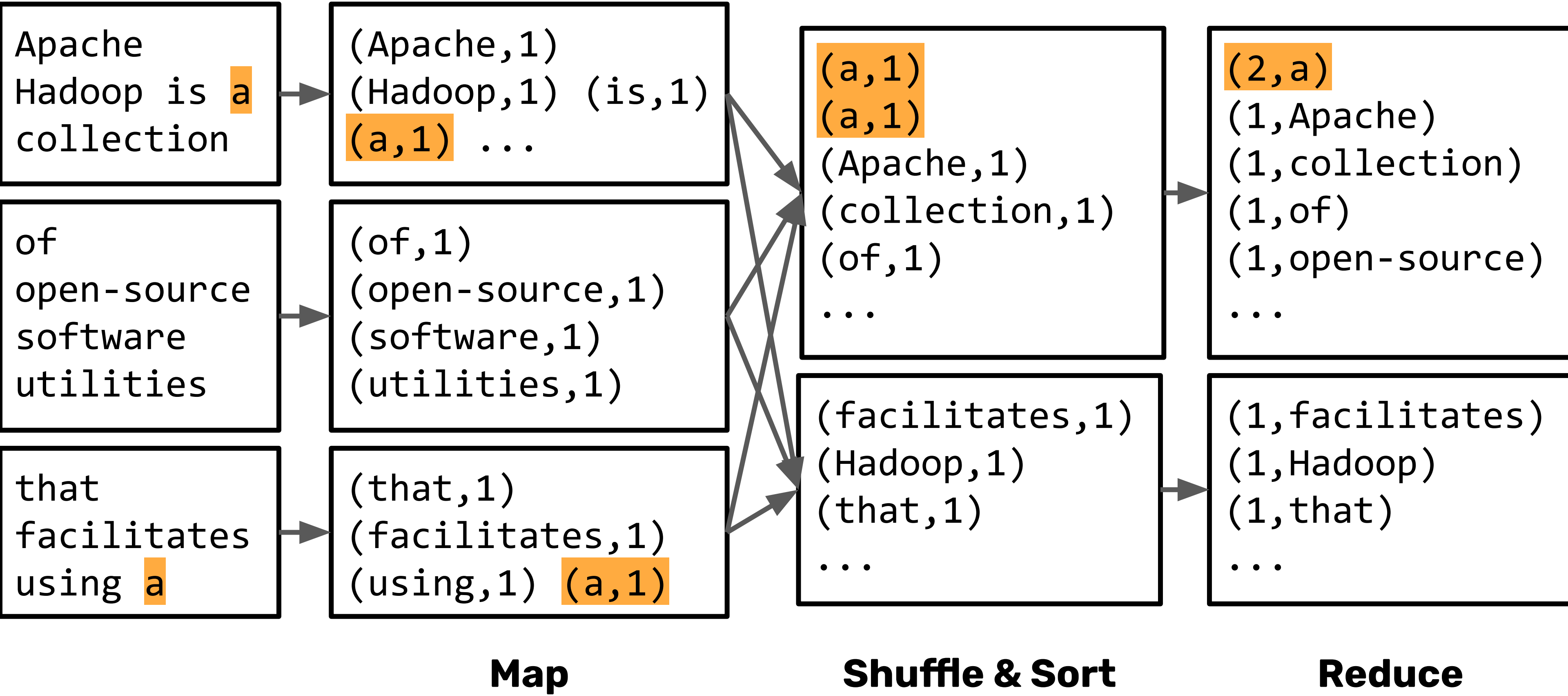
(a,1) (a,1) (Apache,1) (collection,1) (of,1) ...

(facilitates,1) (Hadoop,1) (that,1) ...

Map

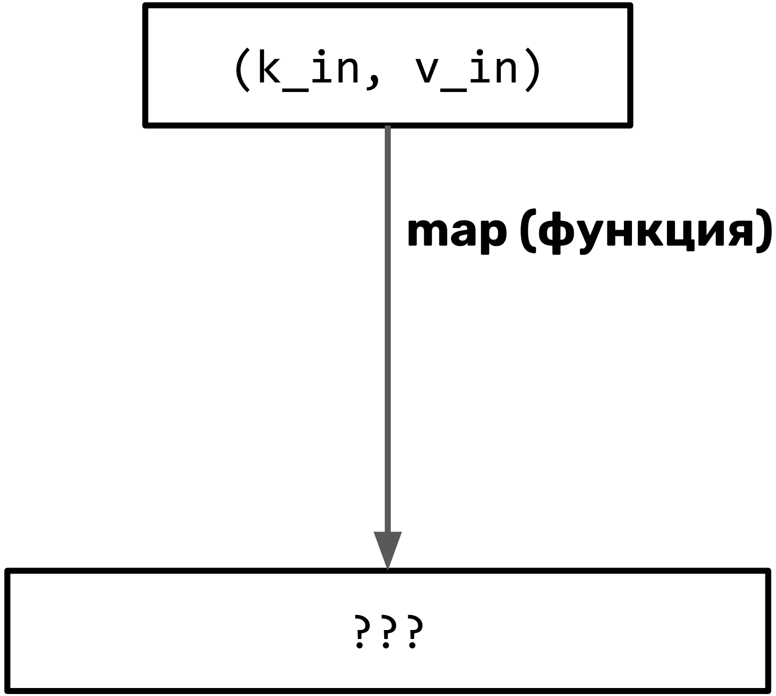
Shuffle & Sort

Word Count v.2


Apache Hadoop is a collection
of open-source software utilities
that facilitates using a
(Apache,1) (Hadoop,1) (is,1) (a,1) ...
(of,1) (open-source,1) (software,1) (utilities,1)
(that,1) (facilitates,1) (using,1) (a,1)
(a,1) (a,1) (Apache,1) (collection,1) (of,1) ...
(facilitates,1) (Hadoop,1) (that,1) ...
(2,a) (1,Apache) (1,collection) (1,of) (1,open-source) ...
(1,facilitates) (1,Hadoop) (1,that) ...
Map
Shuffle & Sort
Reduce

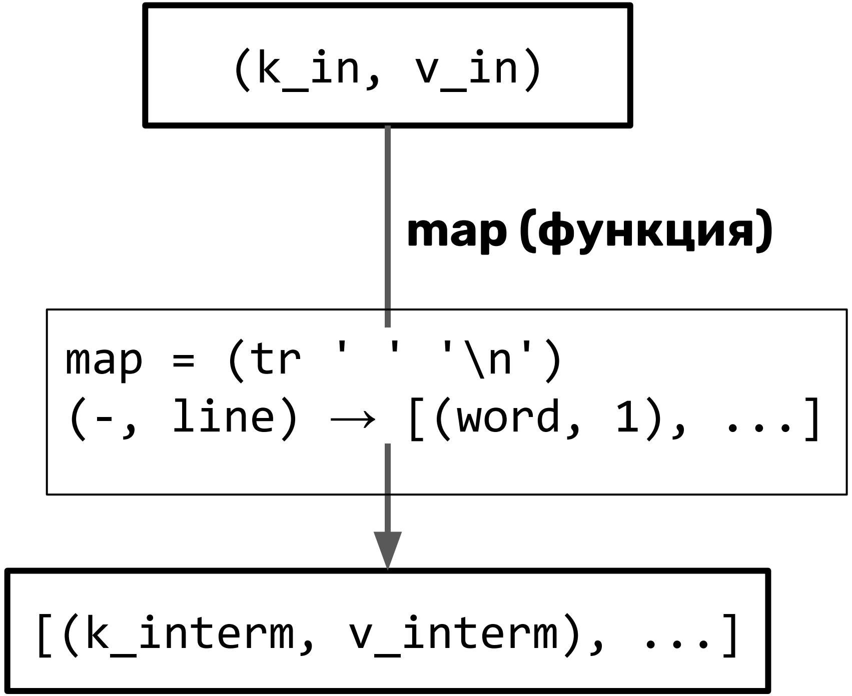
Формальная модель Word Count

Формальная модель Word Count



Map (фаза)

Формальная модель Word Count



```
(k_in, v_in)
```

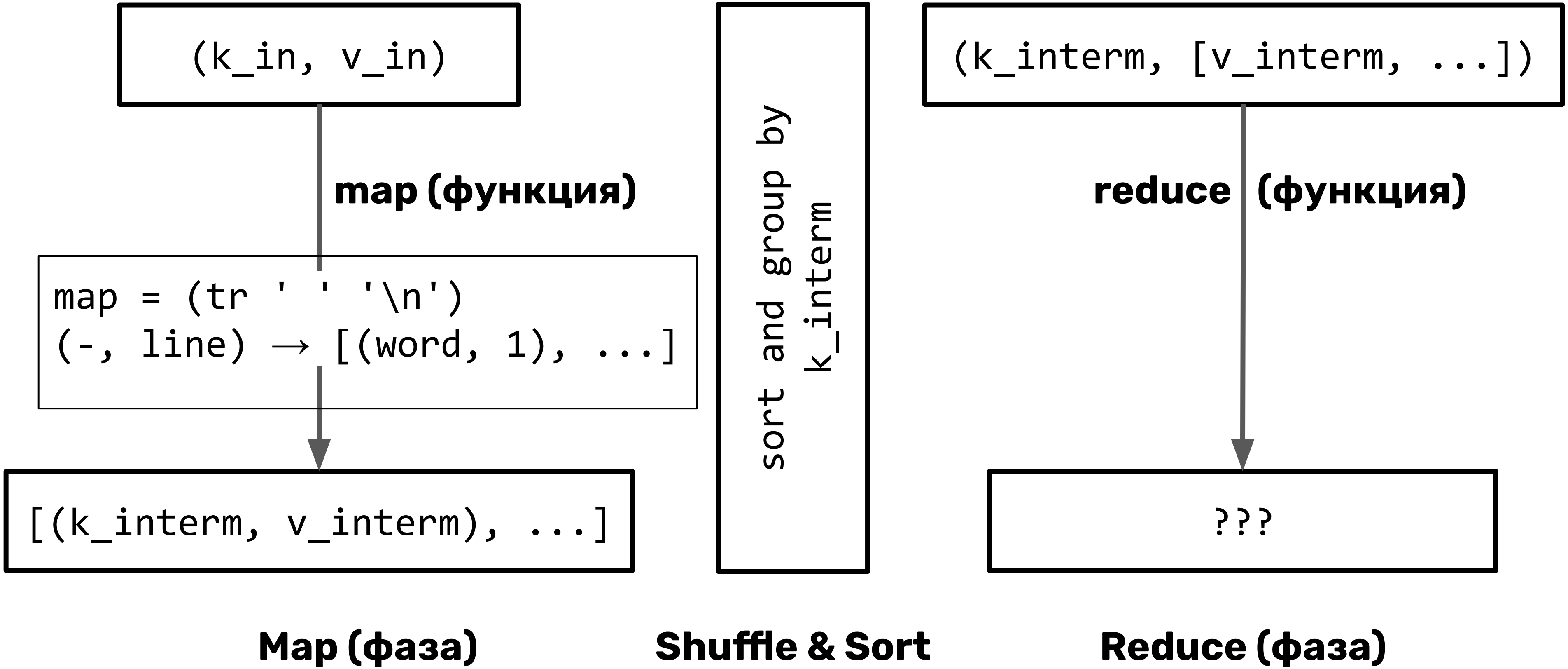
map (функция)

```
map = (tr ' ' '\n')
(-, line) → [(word, 1), ...]
```

```
[(k_interm, v_interm), ...]
```

Map (фаза)

Формальная модель Word Count



(k_in, v_in)

map (функция)

```
map = (tr ' ' '\n')
(-, line) → [(word, 1), ...]
```

[(k_interm, v_interm), ...]

sort and group by k_interm

(k_interm, [v_interm, ...])

reduce (функция)

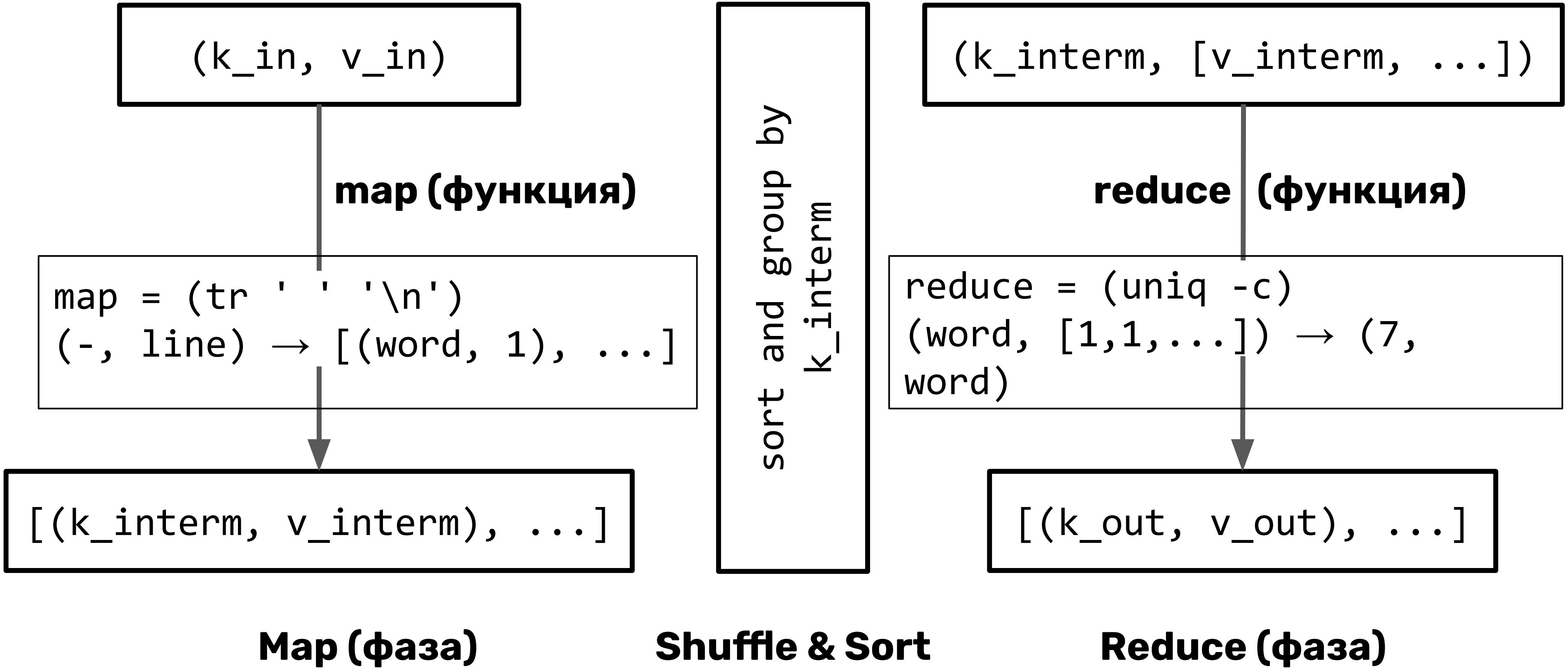
???

Map (фаза)

Shuffle & Sort

Reduce (фаза)

Формальная модель Word Count



(k_in, v_in)

map (функция)

```
map = (tr ' ' '\n')
(-, line) → [(word, 1), ...]
```

[(k_interm, v_interm), ...]

sort and group by k_interm

(k_interm, [v_interm, ...])

reduce (функция)

```
reduce = (uniq -c)
(word, [1,1,...]) → (7, word)
```

[(k_out, v_out), ...]

Map (фаза)

Shuffle & Sort

Reduce (фаза)

Резюме

Теперь вы:

Резюме

Теперь вы:

- Можете насчитать больше 2х фаз в MapReduce

Резюме

Теперь вы:

- Можете насчитать больше 2х фаз в MapReduce
- Понимаете как решать Word Count при помощи MapReduce

Резюме

Теперь вы:

- Можете насчитать больше 2х фаз в MapReduce
- Понимаете как решать Word Count при помощи MapReduce
- Знаете какие 3 типа пар (key, value) указываются при запуске (например) Java MapReduce приложения