Combiner

Драль Алексей, study@bigdatateam.org

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

https://www.facebook.com/bigdatateam

Combiner

Постановка задачи

```
input: word word a word b c d word d e ...
```

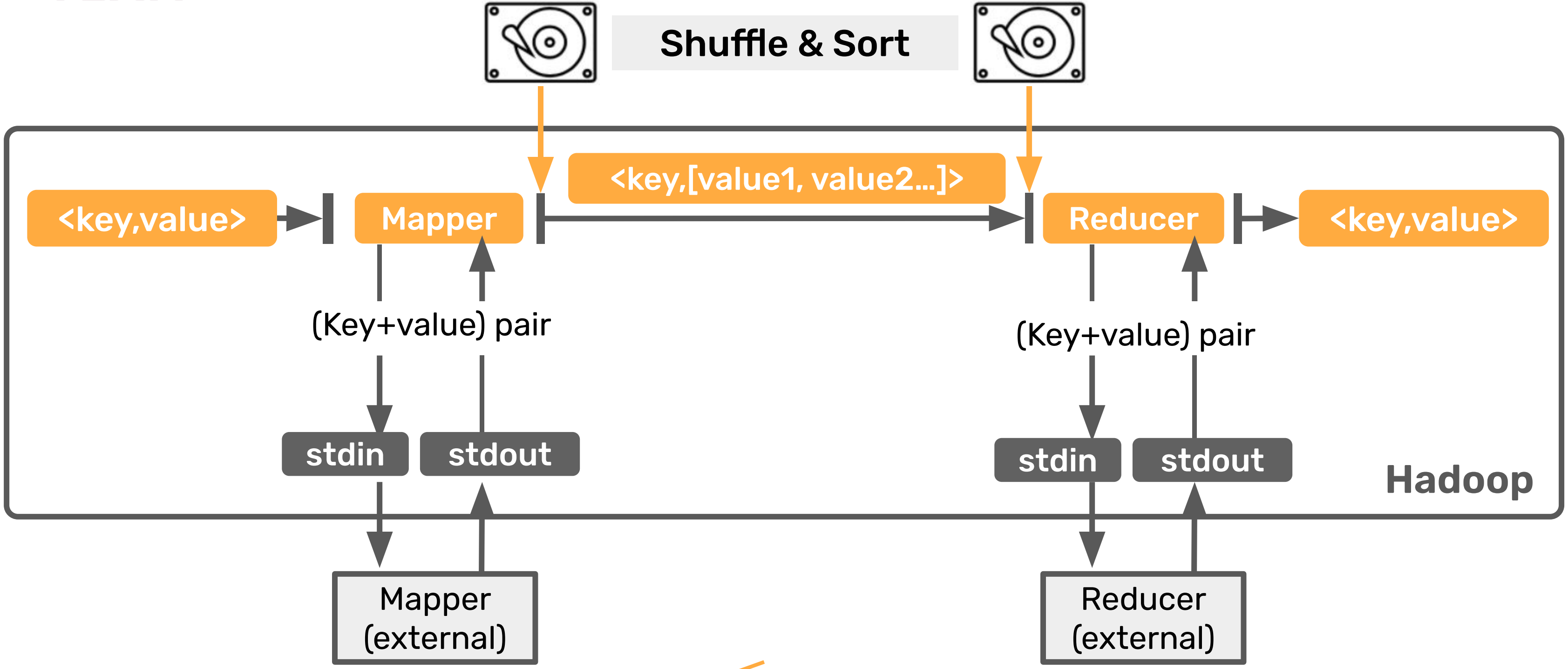
mapper.py

```
#!/usr/bin/env python3
import sys

for line in sys.stdin:
    article_id, content = line.split("\t", 1)
    words = content.split()
    for word in words:
        if word:
            print(word, 1, sep="\t")
```

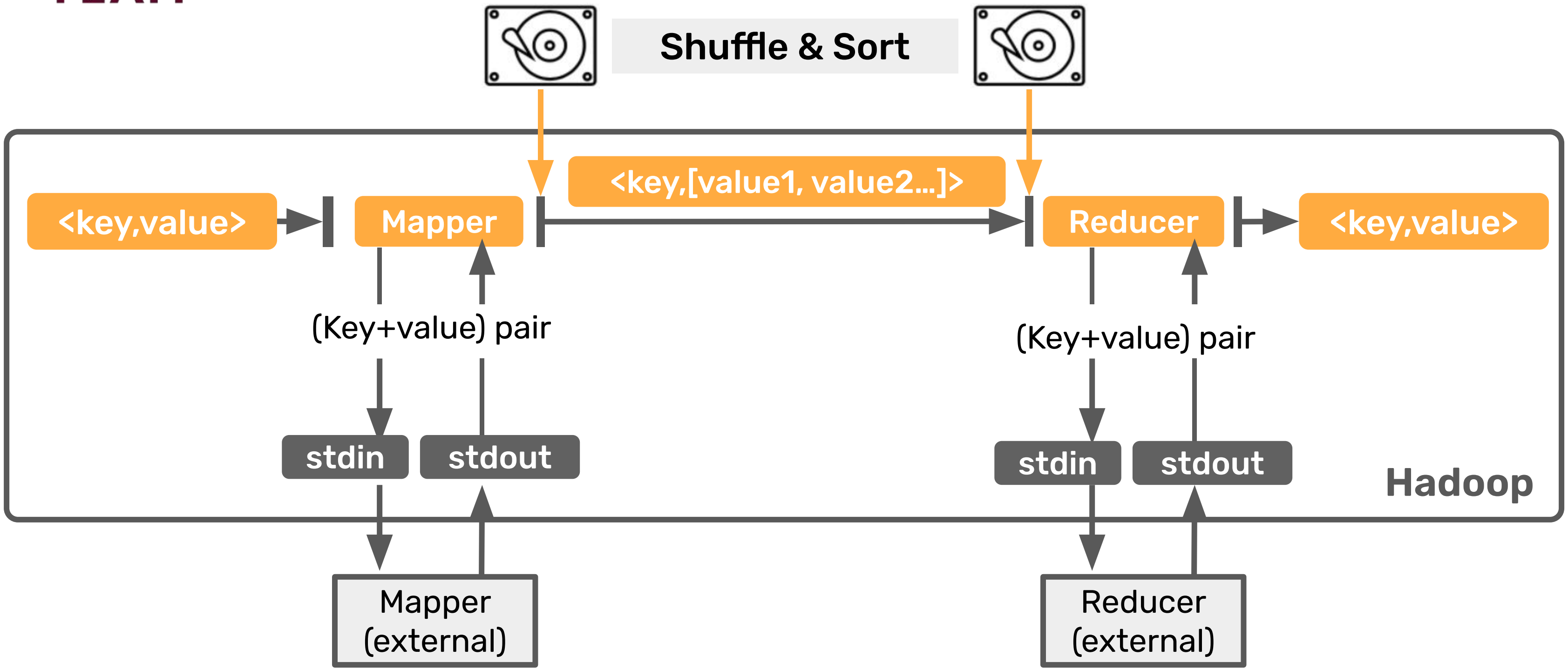
```
output: (word,1) (word,1) (a,1) ...
```

Постановка задачи



output: (word, 1), (word, 1), (a, 1), ...

Решение



output: ~~(word, 1), (word, 1), (a, 1), ...~~
(word, 2), (a, 1), ...

Решение

```
input: word word a word b c d word d e ...
```

mapper.py

```
#!/usr/bin/env python3
import sys
from collections import Counter

for line in sys.stdin:
    article_id, content = line.split("\t", 1)
    words = content.split()
    counts = Counter(words)
    for word, word_count in counts.items():
        print(word, word_count, sep="\t")
```

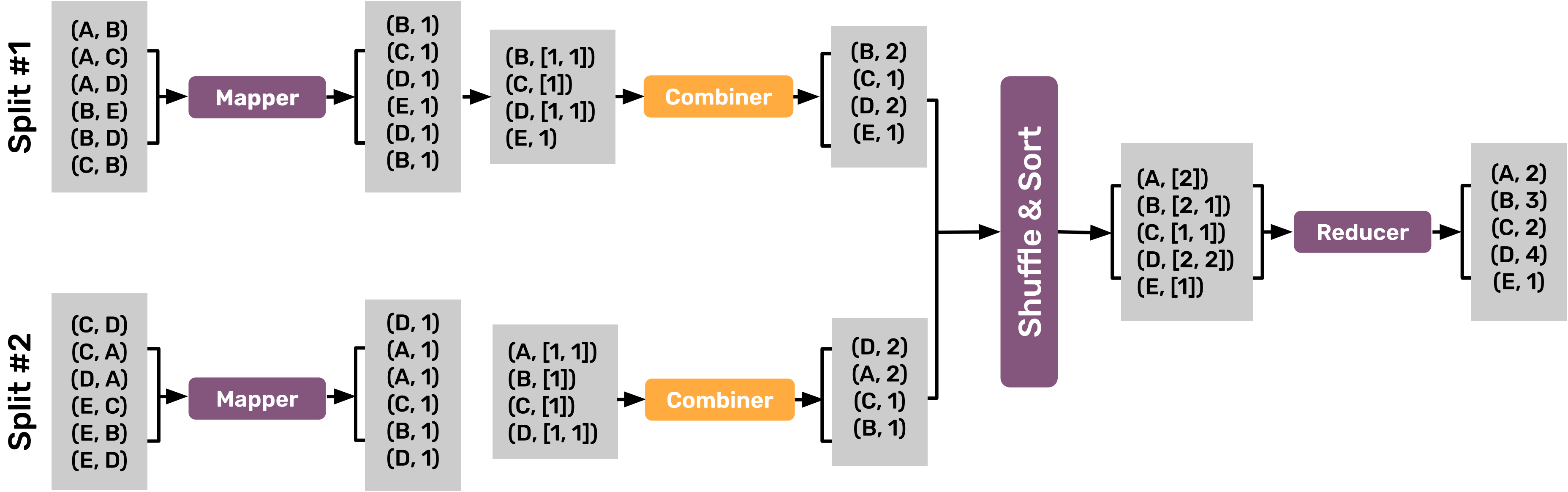
```
output: (word,28) (a,1) ...
```

Сравнение

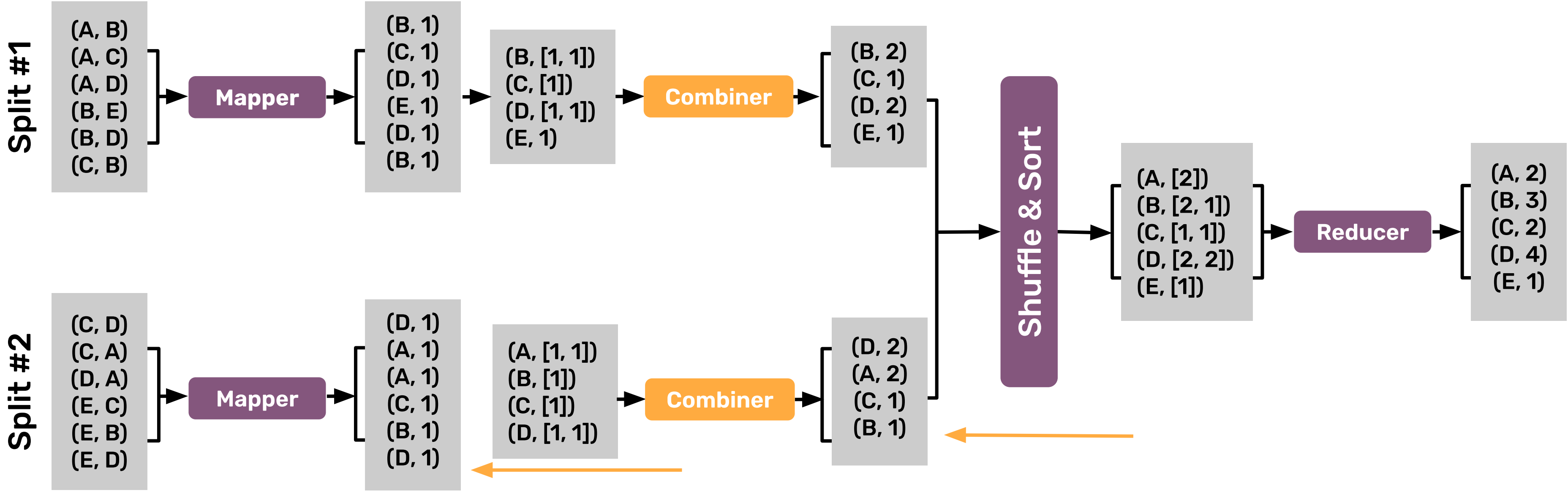
	without Combiner	with Combiner
Wall time (sec)	**935**	**528**
CPU time (sec)	**9790**	**6584**
Local FS Read (MB)	**3006**	**1324**
Local FS Write (MB)	**4527**	**1963**
Peer Map phys. memory (MB)	**526**	**606**
Peek Map virt. memory (MB)	**2131**	**2144**
Peek Reduce phys. memory (MB)	**2744**	**631**
Peer Reduce virt. memory (MB)	**3196**	**3194**

BIGDATA TEAM

Уточнение MapReduce: Combiner

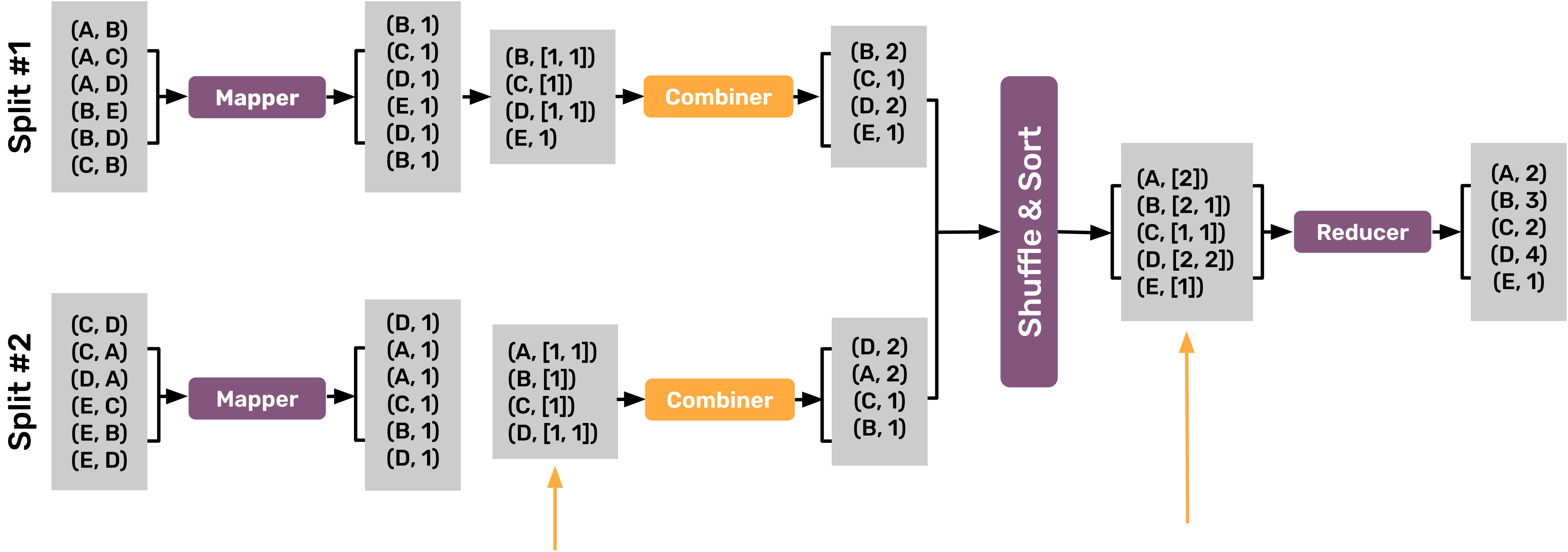


Формальная модель



- read: [(k_in, v_in), …]
- map: (k_in, v_in) -> [(k_interm, v_interm), …]
- combine: (k_interm, [(v_interm, …)]) -> [(k_interm, v_interm), …]
- Shuffle & Sort: sort and group by k_interm
- reduce: (k_interm, [(v_interm, …)]) -> [(k_out, v_out), …]

Формальная модель



- read: [(k_in, v_in), …]
- map: (k_in, v_in) -> [(k_interm, v_interm), …]
- combine: (k_interm, [(v_interm, …)]) -> [(k_interm, v_interm), …]
- Shuffle & Sort: sort and group by k_interm
- reduce: (k_interm, [(v_interm, …)]) -> [(k_out, v_out), …]

Использование

```
$ yarn jar $HADOOP_STREAMING_JAR \
    -files mapper.py,reducer.py \
    -mapper "python3 mapper.py" \
    -combiner "python3 reducer.py" \
    -reducer "python3 reducer.py" \
    -input /data/wiki/en_articles_part \
    -output word_count
```

Использование

```
Map-Reduce Framework
    Map input records=4100
    Map output records=12047715
    Map output bytes=100345949
    Map output materialized bytes=12258223
    Input split bytes=266
    Combine input records=13028233
    Combine output records=1858345
    Reduce input groups=773558
    Reduce shuffle bytes=12258223
    Reduce input records=877827
    Reduce output records=773558
```

Combiner: mean

```
input: word word a word b c d word d e ...
```

?

```
output: (word,#mean) (a,#mean) ...
```

Если Mapper

input: **word word a word b c d word d e ...**

mapper.py

```
#!/usr/bin/env python3
import sys
from collections import Counter

for line in sys.stdin:
    article_id, content = line.split("\t", 1)
    words = content.split()
    counts = Counter(words)
    for word, word_count in counts.items():
        print(word, word_count, sep="\t")
```

output: **(word,28) (a,1)** ...

И если Reducer

reducer.py

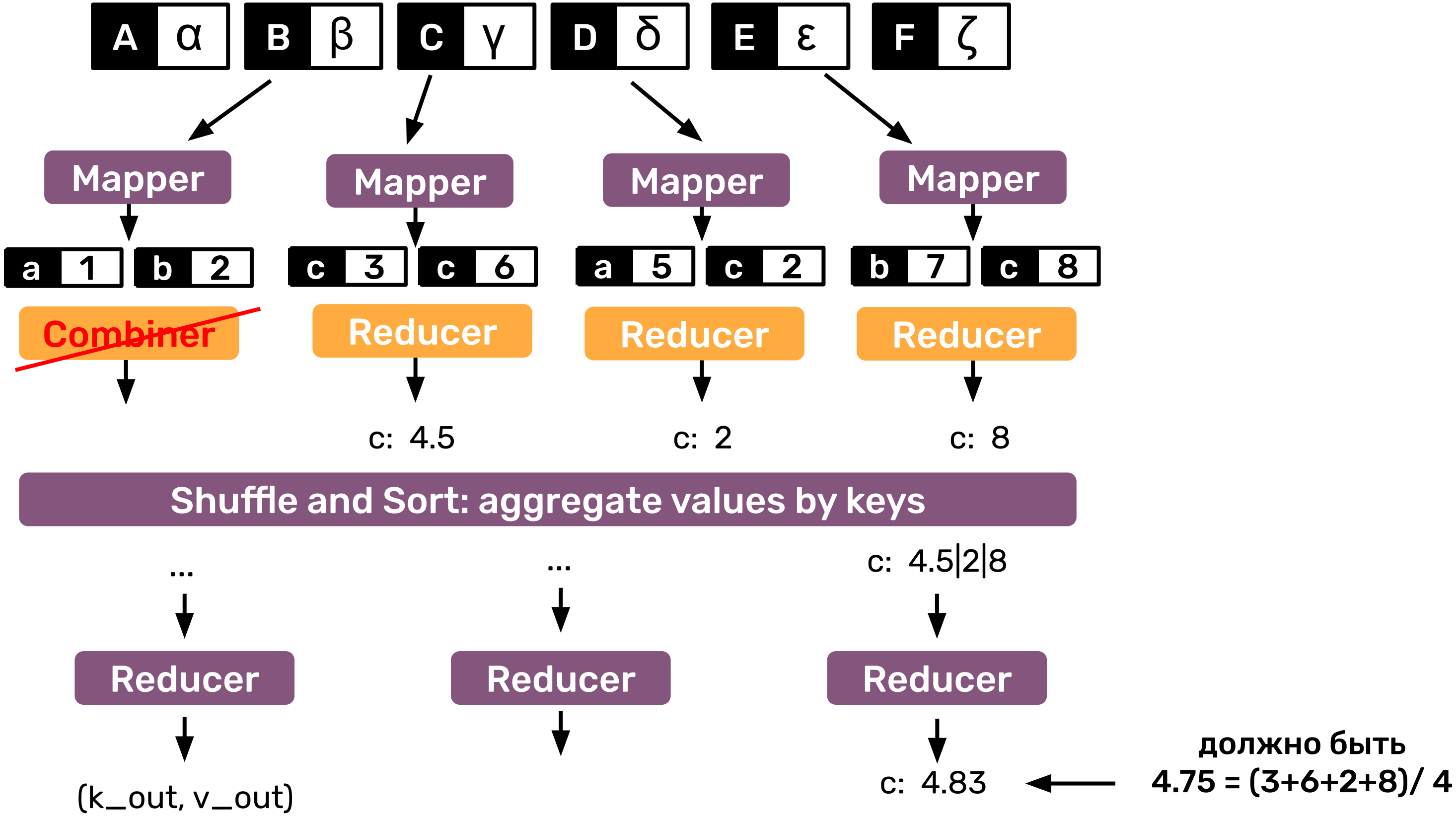
```
#!/usr/bin/env python3
import sys

current_word, word_count, article_count = None, 0, 0

for line in sys.stdin:
    word, counts = line.split("\t", 1)
    counts = int(counts)
    if word == current_word:
        word_count += counts
        article_count += 1
    else:
        if current_word:
            print(current_word, word_count / article_count, sep="\t")
        current_word, word_count, article_count = word, counts, 1

if current_word:
    print(current_word, word_count / article_count, sep="\t")
```

Постановка задачи: mean



Правильный Mapper

input: **word word a word b c d word d e ...**

mapper.py

```python
#!/usr/bin/env python3
import sys
from collections import Counter

for line in sys.stdin:
    article_id, content = line.split("\t", 1)
    words = content.split()
    counts = Counter(words)
    for word, word_count in counts.items():
        print(word, 1, word_count, sep="\t")
```

output: (**word,(1,28)**) (**a,(1,1)**) ...

Правильный Reducer

reducer.py

```
#!/usr/bin/env python3
import sys

current_word, word_count, article_count = None, 0, 0

for line in sys.stdin:
    word, articles, counts = line.split("\t", 2)
    articles, counts = int(articles), int(counts)
    if word == current_word:
        word_count += counts
        article_count += articles
    else:
        if current_word:
            print(current_word, word_count / article_count, sep="\t")
        current_word, word_count, article_count = word, counts, articles

if current_word:
    print(current_word, word_count / article_count, sep="\t")
```

Правильный ли Combiner?

combiner.py

```
#!/usr/bin/env python3
import sys

current_word, word_count, article_count = None, 0, 0

for line in sys.stdin:
    word, articles, counts = line.split("\t", 2)
    articles, counts = int(articles), int(counts)
    if word == current_word:
        word_count += counts
        article_count += articles
    else:
        if current_word:
            print(current_word, word_count / article_count, sep="\t")
        current_word, word_count, article_count = word, counts, articles

if current_word:
    print(current_word, word_count / article_count, sep="\t")
```

Правильный Combiner

combiner.py

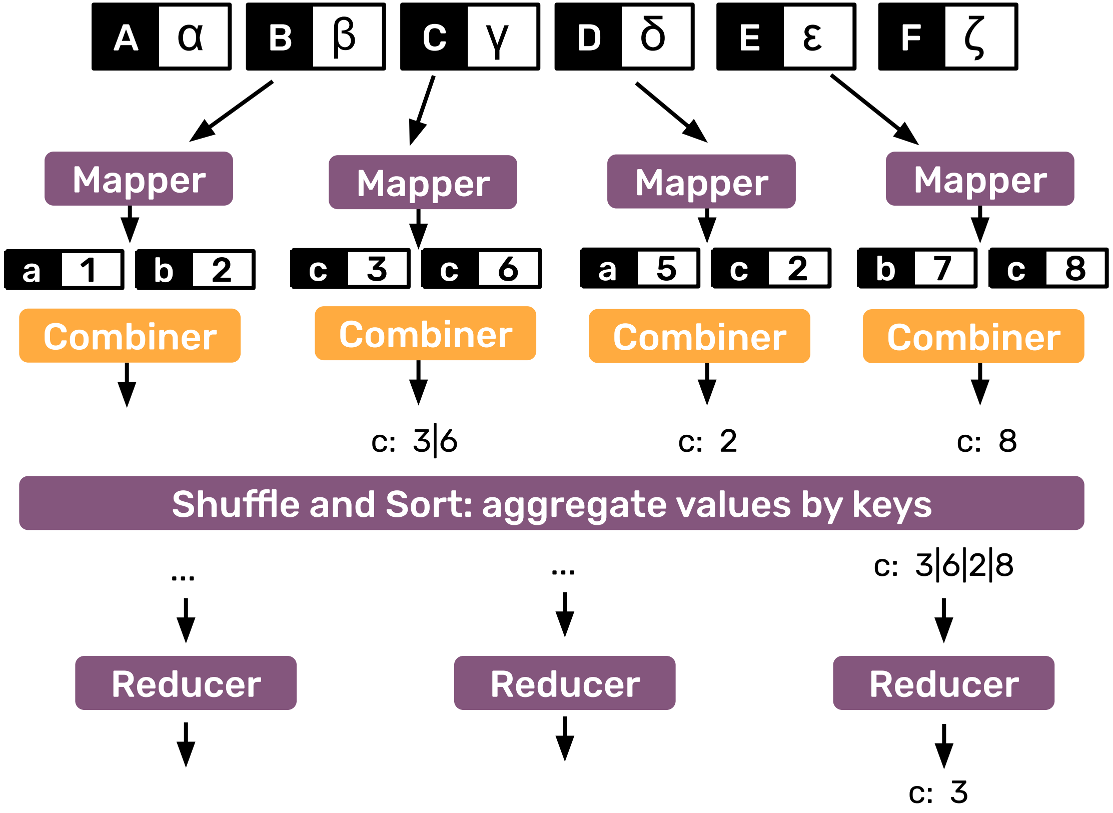
```
#!/usr/bin/env python3
import sys

current_word, word_count, article_count = None, 0, 0

for line in sys.stdin:
    word, articles, counts = line.split("\t", 2)
    articles, counts = int(articles), int(counts)
    if word == current_word:
        word_count += counts
        article_count += articles
    else:
        if current_word:
            print(current_word, article_count, word_count, sep="\t")
        current_word, word_count, article_count = word, counts, articles

if current_word:
    print(current_word, article_count, word_count, sep="\t")
```

Combiner: Median



Резюме

Резюме

Теперь вы:

- Знаете что такое Combiner

Резюме

Теперь вы:

- Знаете что такое Combiner
- Умееете вычислять сигнатуру функции combine

Резюме

Теперь вы:

- Знаете что такое Combiner
- Умеете вычислять сигнатуру функции combine
- Можете объяснить где надо и каким образом использовать Combiner, а где - не стоит