Streaming Word Count

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.

WIKIPEDIA
The Free Encyclopedia

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...

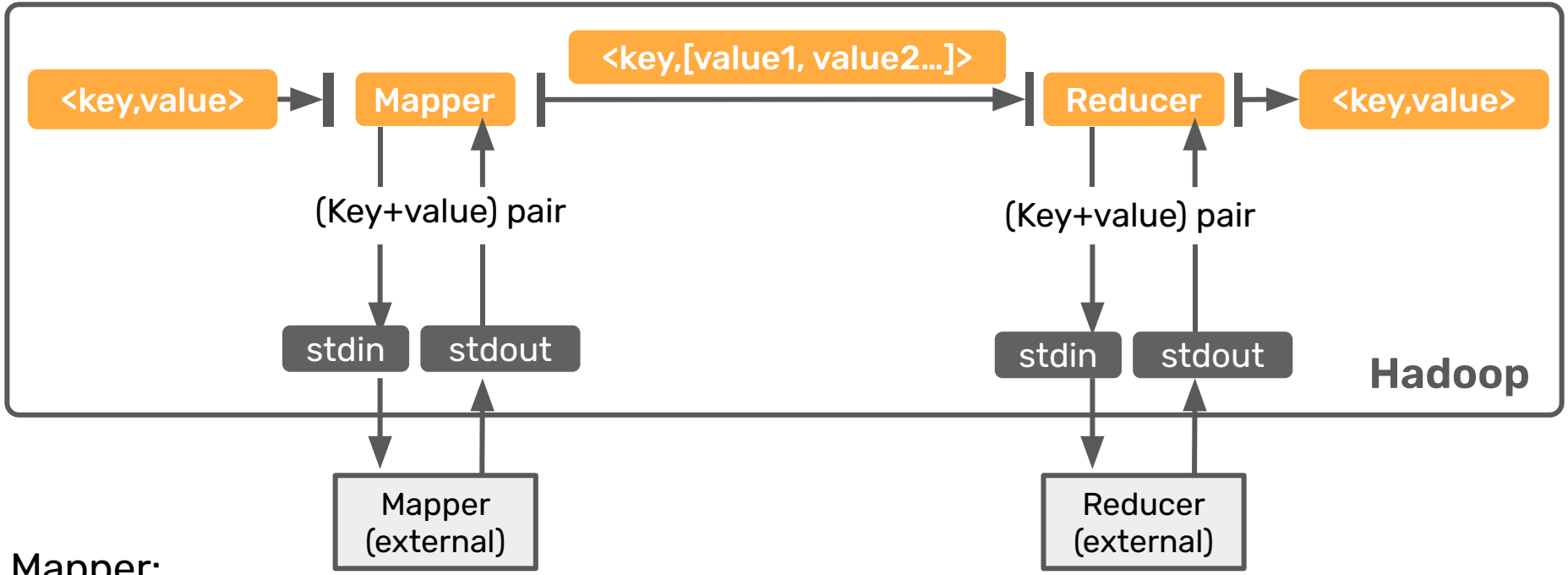
hadoop MapReduce

↓

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

BIGDATA TEAM

MapReduce Streaming



Mapper:

- Как данные читаем (input format)
- Как данные обрабатываем
- Как данные выводим (output format)

Тоже, что и Mapper, плюс:

- Как агрегируем по ключам отсортированные данные

Word Count Mapper

<article_id> <tab> <article_content>

key value

mapper.py

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

for line in sys.stdin:
    article_id, content = line.split("\t", 1)
    # ...
```

Word Count Mapper

mapper.py

```python
#!/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:
            key_interim, value_interim = word, 1
```

Word Count Mapper

mapper.py

```python
#!/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")
```

BIGDATA TEAM

Word Count Mapper

```
yarn jar $HADOOP_STREAMING_JAR \
    -files mapper.py \
    -mapper "python3 mapper.py" \
    -numReduceTasks 0 \
    -input /data/wiki/en_articles_part \
    -output word_count
```

```
$ hdfs dfs -ls -h word_count
```

```
Found 3 items
-rw-r--r--   3 aadral hdfs        0 2021-03-02 10:24 word_count/_SUCCESS
-rw-r--r--   3 aadral hdfs   83.6 M 2021-03-02 10:24 word_count/part-00000
-rw-r--r--   3 aadral hdfs   12.1 M 2021-03-02 10:24 word_count/part-00001
```

Word Count Mapper

```
$ hdfs dfs -tail word_count/part-00001 | head -5
```

```
ia,       1
Paris,    1
2005.     1
John      1
Newton,   1
```

Word Count Mapper

mapper.py

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

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

Word Count Reducer

```
yarn jar $HADOOP_STREAMING_JAR \
    -files mapper.py \
    -mapper "python3 mapper.py" \
    -numReduceTasks 0 \
    -input /data/wiki/en_articles_part \
    -output word_count
```

```
$ hdfs dfs -ls -h word_count
```

```
Found 3 items
-rw-r--r--   3 aadral hdfs        0 2021-03-02 10:46 word_count/_SUCCESS
-rw-r--r--   3 aadral hdfs   94.1 M 2021-03-02 10:46 word_count/part-00000
```

Word Count Reducer

```
$ hdfs dfs -text word_count/part-00000 | head -5
```

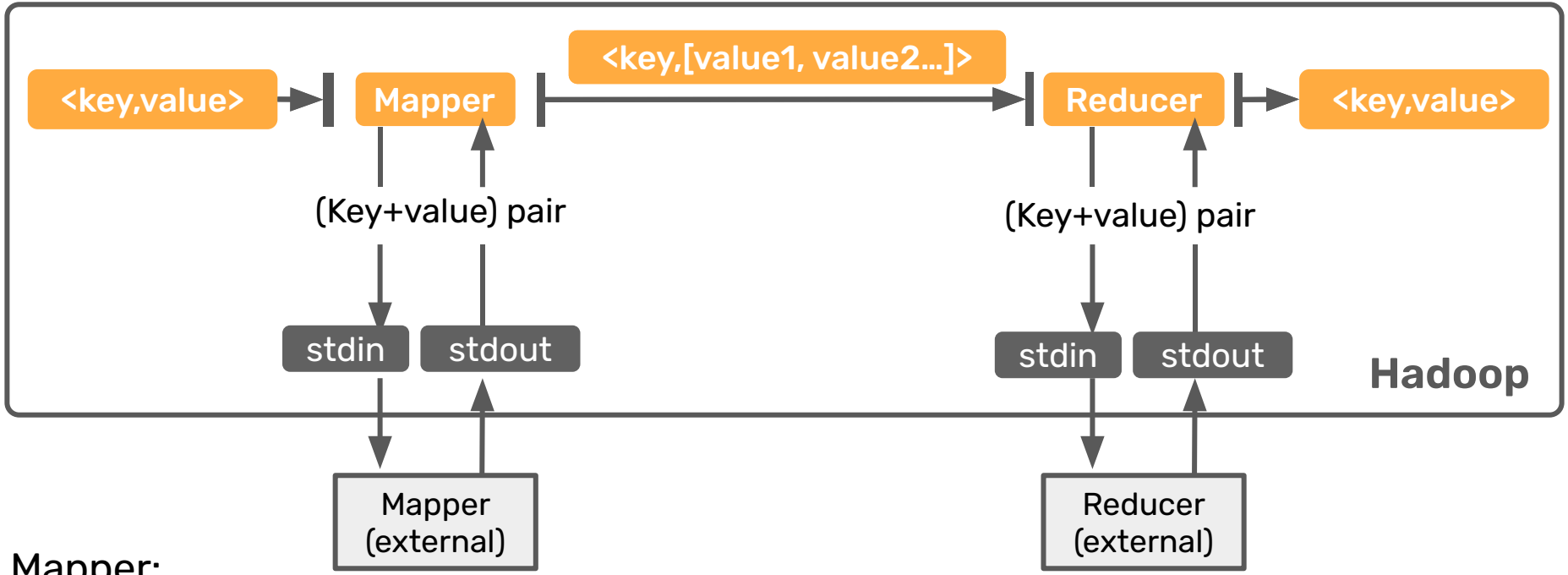
```
0   1
0   1
0   1
0   1
0   1
text: Unable to write to output stream.
```

```
$ hdfs dfs -tail word_count/part-00000 | tail -5
```

```
□   1
□   1
□   1
□   1
□   1
```

BIGDATA TEAM

MapReduce Streaming



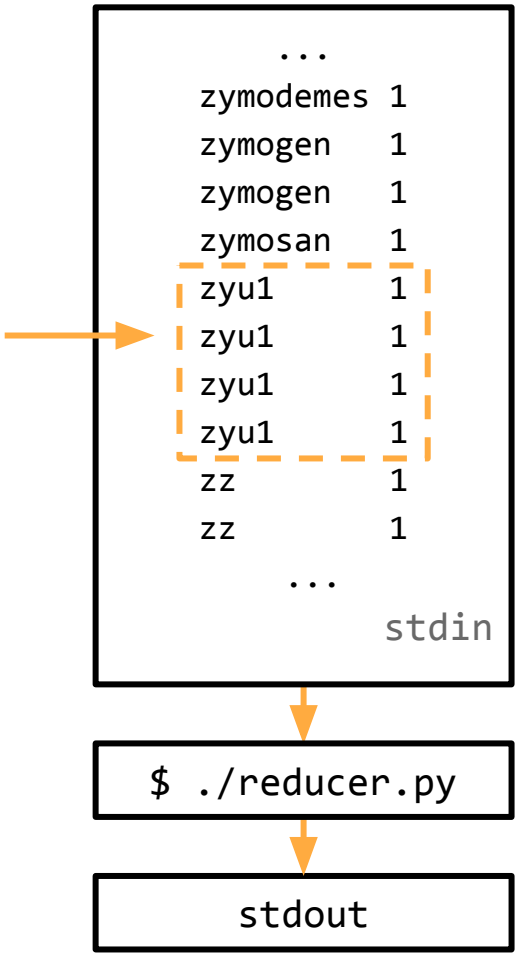
Mapper:

- Как данные читаем (input format)
- Как данные обрабатываем
- Как данные выводим (output format)

Тоже, что и Mapper, плюс:

- Как агрегируем по ключам отсортированные данные

Word Count Reducer



Word Count Reducer

reducer.py

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

current_word = None
word_count = 0

for line in sys.stdin:
    word, counts = line.split("\t", 1)
    counts = int(counts)
# ...
```

Word Count Reducer

reducer.py

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

current_word = None
word_count = 0

for line in sys.stdin:
    word, counts = line.split("\t", 1)
    counts = int(counts)
# ...
```

Word Count Reducer

reducer.py

```
# ...

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

```
...
zymosan 1
zyu1    1
zyu1    1
zyu1    1
zyu1    1
zz      1
...
```

stdin

Word Count Reducer

reducer.py

```
# ...
    if word == current_word:
        word_count += counts
    else:
        if current_word:
            print(current_word, word_count, sep="\t")
        current_word = word
        word_count = counts
# ...
```

```
...
zymosan 1
zyu1    1
zyu1    1
zyu1    1
zyu1    1
zz      1
...
```

stdin

Word Count Reducer

reducer.py

```
# ...
    if word == current_word:
        word_count += counts
    else:
        if current_word:
            print(current_word, word_count, sep="\t")
        current_word = word
        word_count = counts
# ...
```

```
...
zymosan 1
zyu1    1
zyu1    1
zyu1    1
zyu1    1
zz      1
...
```

stdin

Word Count Reducer

reducer.py

```
# ...
    if word == current_word:
        word_count += counts
    else:
        if current_word:
            print(current_word, word_count, sep="\t")
        current_word = word
        word_count = counts
# ...
```

```
...
zymosan 1
zyu1    1
zyu1    1
zyu1    1
zyu1    1
zz      1
...
```

stdin

Word Count Reducer

reducer.py

```python
# ...

for line in sys.stdin:
    # ...

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

```
...
zymosan 1
zyu1    1
zyu1    1
zyu1    1
zyu1    1
zz      1
...
```

stdin

Word Count Reducer

reducer.py

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

current_word = None
word_count = 0

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

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

Word Count Reducer

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

```
$ hdfs dfs -ls -h word_count
```

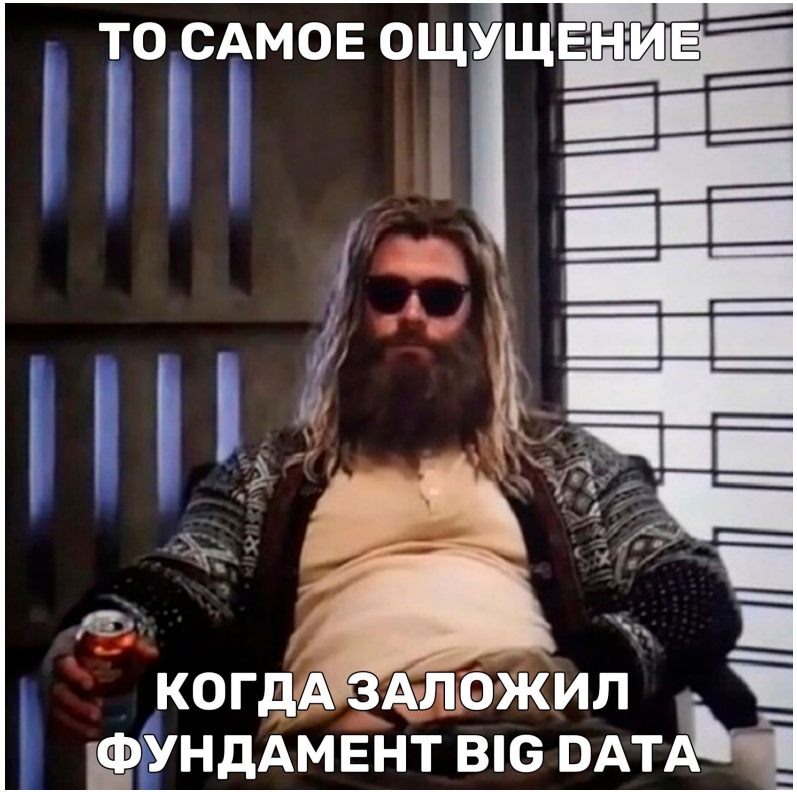
```
Found 3 items
-rw-r--r--   3 aadral hdfs       0 2021-03-02 11:21 word_count/_SUCCESS
-rw-r--r--   3 aadral hdfs   3.4 M 2021-03-02 11:21 word_count/part-00000
```

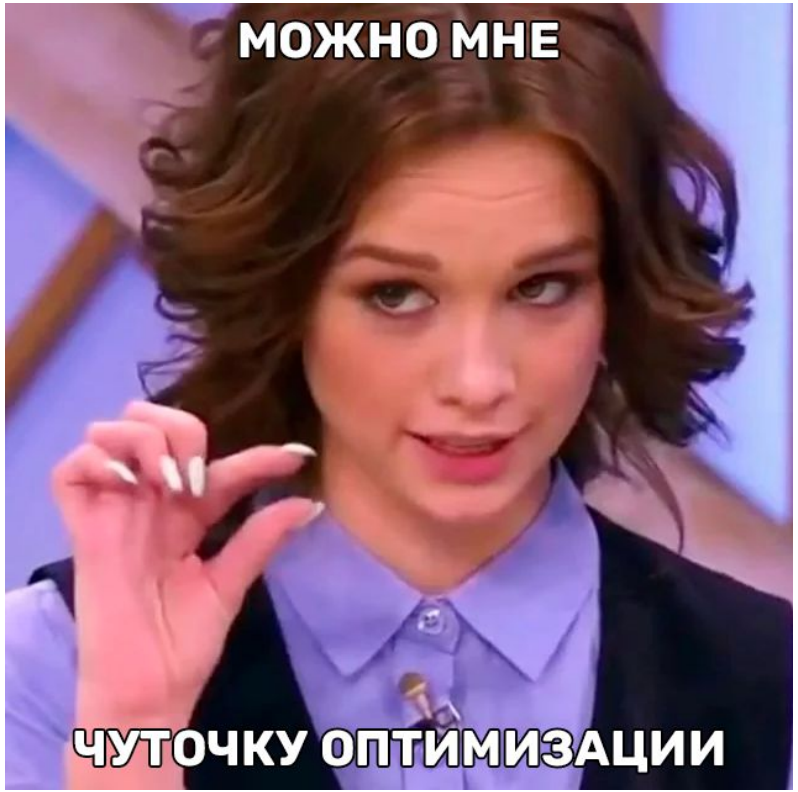
Word Count Reducer

```
$ hdfs dfs -text word_count/part-00000 | head -5
```

```
0       14891
00      844
000     8186
0000    55
00000   5
text: Unable to write to output stream.
```

Резюме


ТО САМОЕ ОЩУЩЕНИЕ
КОГДА ЗАЛОЖИЛ ФУНДАМЕНТ BIG DATA


МОЖНО МНЕ
ЧУТОЧКУ ОПТИМИЗАЦИИ