Распределенный кеш (Distributed Cache)

Драль Алексей, study@bigdatateam.org
CEO at BigData Team, https://bigdatateam.org
https://www.facebook.com/bigdatateam

Фильтрация по словарю

Фильтрация по словарю

mapper.py

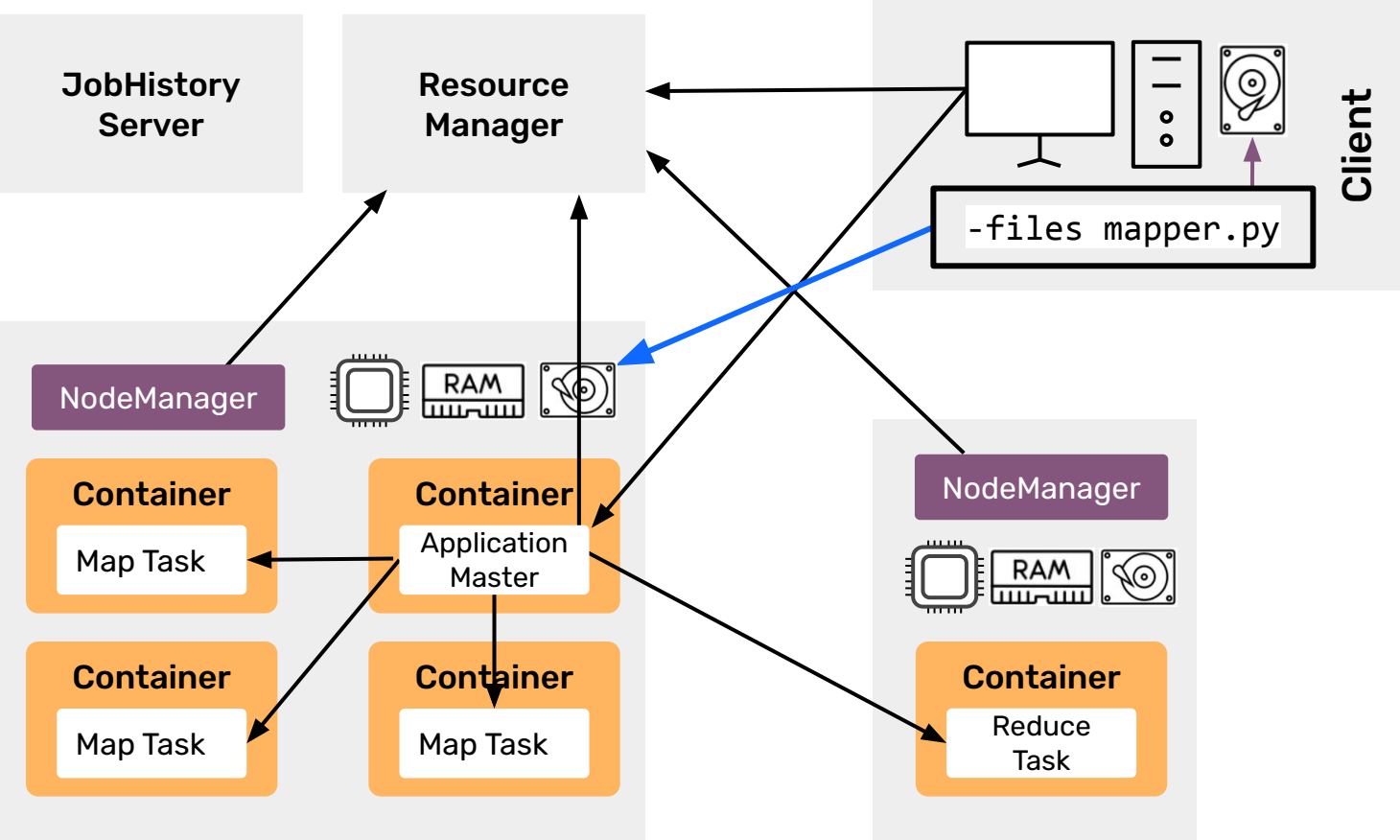
```
import re
import sys

def read_vocabulary(file_path):
    return set(line.strip() for line in open(file_path))

vocabulary = read_vocabulary("vocabulary.txt")

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

BIGDATA TEAM

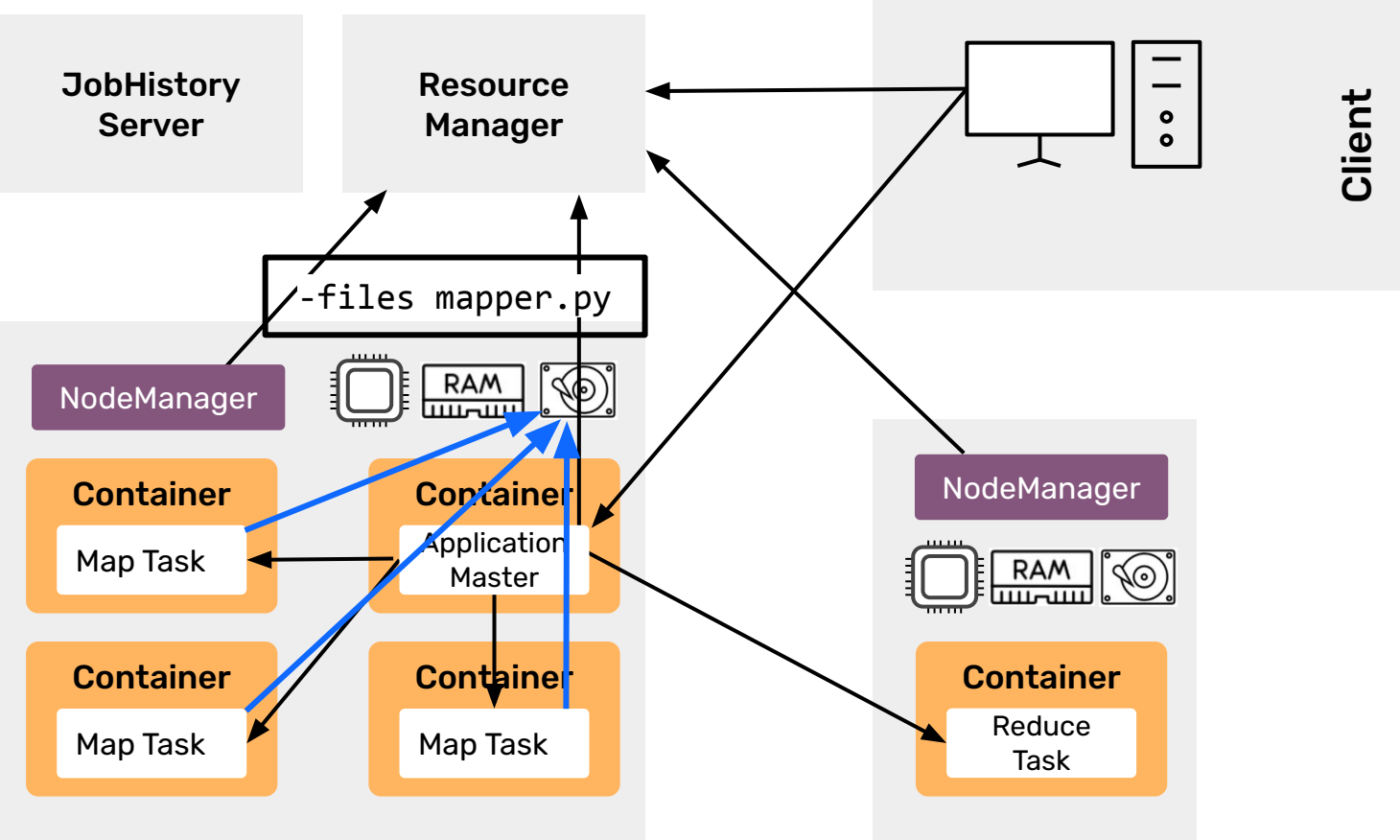
Фильтрация по словарю

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

Распространение данных



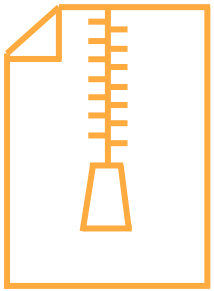
Использование данных



BIGDATA TEAM

Вараинаты Distributed Cache

-files



-archives

-libjars

Фильтрация по именам

female.txt

male.txt

names.tar

```
$ tar -cf names.tar male.txt female.txt
```

Передача архива

```
$ tar -cf names.tar male.txt female.txt
```

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

Фильтрация по словарям

mapper.py

```
import re
import sys

def read_vocabulary(file_path):
    return set(line.strip() for line in open(file_path))

male_names = read_vocabulary("names.tar/male.txt")
female_names = read_vocabulary("names.tar/female.txt")

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

BIGDATA TEAM

Ожидание vs Реальность?


```
$ hdfs dfs -text word_count/* | sort -nrk2,2
```

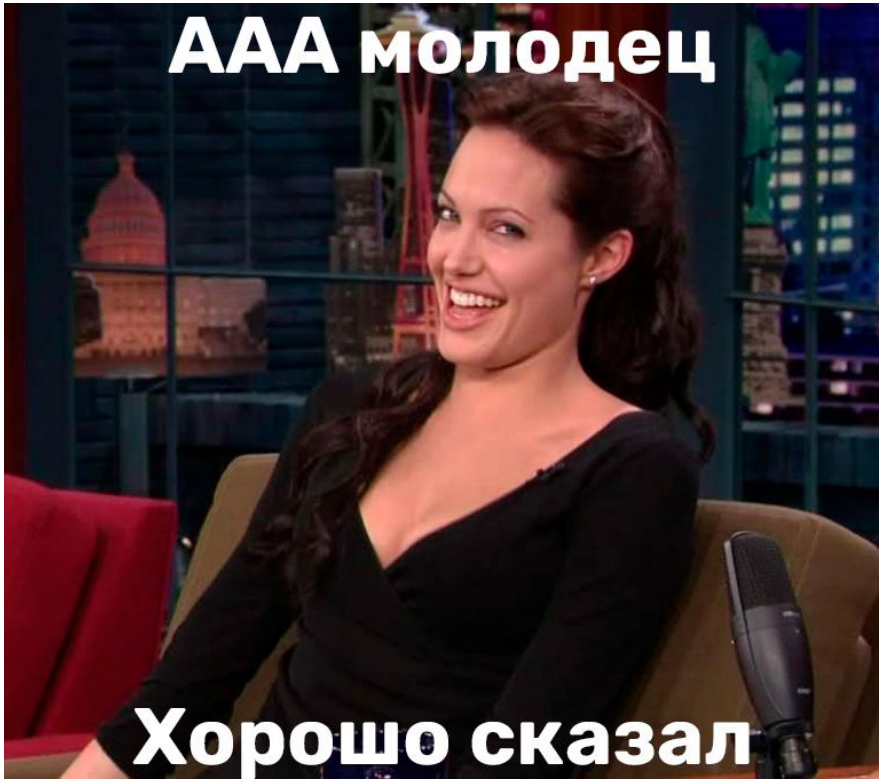
James	2284
Thomas	1941
Jack	786
Harry	504
Muhammad	444
Oliver	250
Charlie	250
Jacob	234
Emily	128
Isabella	99

Sophia	92
Noah	80
Sophie	64
Lily	31
Olivia	27
Ethan	27
Ella	25
Amelia	25
Isabelle	18
Chloe	9

Резюме

Теперь вы:

- Знаете форматы использования Distributed Cache и как он работает под капотом



BIGDATA TEAM

Резюме