Урок 3

1. Может ли стадия Reduce начаться до завершения стадии Map? Почему?
2. Приведите пример Map only и Reduce задачи.
3. Разверните кластер hadoop, соберите WordCount приложение, запустите на датасете ppkm_sentiment и выведите 10 самых редких слов*
4. Измените маппер в WordCount так, чтобы он удалял знаки препинания и приводил все слова к единому регистру
5. *У вас есть два датасета с одинаковыми ключами. Вам нужно их объединить, суммировав значения с одинаковыми ключами. Как это сделать в MapReduce?
6. *На кластере лежит датасет, в котором ключами является id сотрудника и дата, а значением размер выплаты. Руководитель поставил задачу рассчитать среднюю сумму выплат по каждому сотруднику за последний месяц. В маппере вы отфильтровали старые записи и отдали ключ-значение вида: id-money. А в редьюсере суммировали все входящие числа и поделили результат на их количество. Но вам в голову пришла идея оптимизировать расчет, поставив этот же редьюсер и в качестве комбинатора, тем самым уменьшив шафл данных. Можете ли вы так сделать? Если да, то где можно было допустить ошибку? Если нет, то что должно быть на выходе комбинатора?

Задание 1

Не может так как стадии сортировки и шафла должны производиться над всеми данными. Иначе редьюсеры могут отработать неправильно.

Задание 2

1. Предположим в hdfs у нас хронятся логи воркеров каких-то приложений. В каждой строке лога указано имя воркера из которого она. Можно написать мапер, который будет выводить только логи какого-то конкретного воркера.

2. А в качестве Reduce задачи мы можем считать сколько логов произведено каждым воркером(мапер будет уже не из пункта 1, а такой, что выводит: worker-name 1, в stdout)

Задание 3

Запускаем WC

```
hduser@localhost:~$ hadoop fs -ls
Found 1 items
drwxrwxrwx   - hduser supergroup          0 2022-05-15 14:07 ppkm
hduser@localhost:~$ hadoop jar ./hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.1.jar \
> -file /home/hduser/word-count/mapper.py -mapper /home/hduser/word-count/mapper.py \
> -file /home/hduser/word-count/reducer.py -reducer /home/hduser/word-count/reducer.py \
> -input /user/hduser/ppkm -output /user/hduser/ppkm_wc
22/05/31 14:59:24 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hduser/word-count/mapper.py, /home/hduser/word-count/reducer.py] [] /tmp/streamjob4476787854601311644.jar tmpDir=null
22/05/31 14:59:25 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
22/05/31 14:59:25 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
22/05/31 14:59:25 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
22/05/31 14:59:25 INFO mapred.FileInputFormat: Total input files to process : 3
22/05/31 14:59:25 INFO mapreduce.JobSubmitter: number of splits:3
```

Выводим топ-10 результатов из файла с расчетами

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/ppkm_wc/* | sort -nk2 | tail -n 10
perpanjangan    32
#PPKMMikro      34
Pembatasan      36
Masyarakat      38
Kegiatan        39
Mikro   56
yg      56
di      76
PPKM    85
dan     88
hduser@localhost:~$
```

Задание 4

Так если просто заменять знаки препинания на пробелы возникает много багов: считаются некоторые пробелы, ссылки рассыпаются и тд. Решил использовать регулярки(но я так понимаю лучше обходится без них так как это тяжелые операции). Использовал две регулярки: чтобы найти все url'ы и выкинуть их и для поиска слов.

```python
#!/usr/bin/env python

import sys
import re


for line in sys.stdin:
    line = line.strip()
    urls = re.findall(r"http\S*\b", line)
    if urls:
        for url in urls:
            print(f"{url}\t{1}")
            line = line.replace(url, "")
    words = re.findall("[\w]+", line)

    for word in words:
        print(f"{word.lower()}\t{1}")
```

Топ-40 результат(чтобы попал url)

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/ppkm_wc_2/* | sort -nk2 | tail -n 40
desa    20
diperpanjang    20
gara    20
masker  20
mulai   20
pemberlakuan    20
https://t.co/ROoMoqoBGC 21
itu     21
jogjaelinglanwaspada    21
pemberlakukan   21
pengumuman      21
psbb    22
22      23
atau    23
pak     24
humas_jogja     25
ini     25
jogjaistimewa   25
2021    27
maret   28
untuk   28
berbasis        31
yang    33
rt      34
ada     36
pembatasan      43
perpanjangan    43
kegiatan        44
yg      61
masyarakat      67
19      68
ppkmmikro       74
di      81
covid   93
dan     93
negatif 101
netral  101
positif 106
mikro   111
ppkm    141
hduser@localhost:~$
```

Задание 5

Мапер и редьюсер похожи на WC. Мапер выводит ключ-значение, редьюсер сумирует по этим значениям. Если при запуске джобы мы укажим параметр `-D mapred.reducer.tasks=1`, то на выходе получим один файл который по сути будет объединением этих таблиц.

Задание 6

Так нельзя делать так как, среднее от суммы средних не будет равняться просто среднему. (1+2+3+4)/4 != ((1+2+3)/3 + 4/1)/2. Но можно написать комбайнер который будет писать в вывоод сумму и количество выплат.