ТЕХНОСФЕРА

Практическое задание лекции№4

MapReduce в Hadoop WordCount и InputFormat-ы

Кисель Ян

Код для семинара

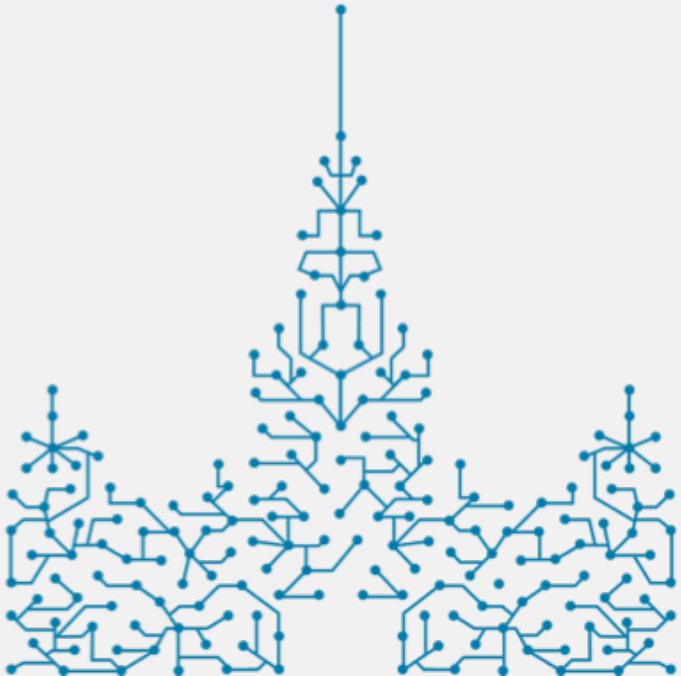
```
$ git clone \
https://github.com/dkrotx/hadoop_sem1
```

```
$ cd hadoop_sem1
$ ./gradlew jar
```

Выходной файл:

build/libs/hadoop_sem1.jar

WordCount



Запустим тестовый wordcount

`hdfs:///data/seminar1/word_count/lenta_ru.txt`

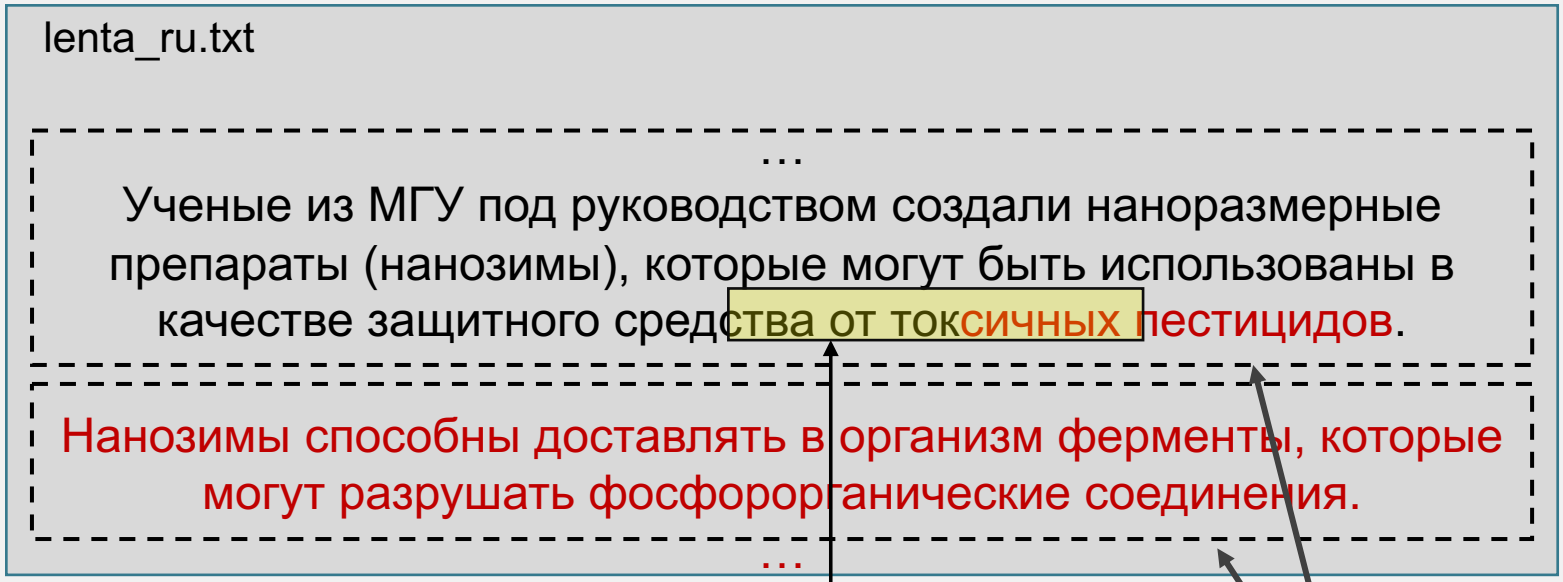
- Собрать hadoop_sem1 (WordCountJob.java)
- Запустить WordCount
- Посмотреть top по словам

```
$ hadoop jar build/libs/hadoop_sem1.jar WordCountJob \
  /data/seminar1/word_count/lenta_ru.txt out/sem1/1
$ hadoop fs -cat out/sem1/1/part-r-* | sort -k 2,2 -nr | head
```

Как работает чтение файла?

- В mapper-е вы получаете строку за строкой
- Блоки HDFS != строкам
- Почему не рвется строка?
- Как вообще происходит чтение?

Чтение на стыке строк



Входные данные

```
lenta_ru.txt        [1.9G]
lenta_ru.txt.gz     [547M]
lenta_ru.txt.bz2    [346M]
```

Запустить WordCount на каждом файле **отдельно**

Узнать размер блока:
$ hdfs getconf -confKey dfs.blocksize

Посмотреть блоки файлов:
$ hadoop fsck /data/seminar1/word_count/ -files -blocks

Работа со сжатыми данными

Обратите внимание:

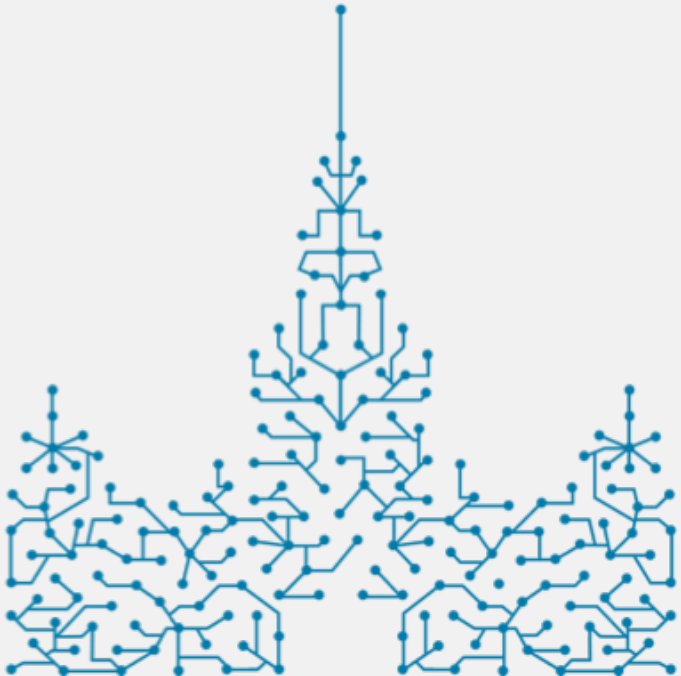
mapreduce.JobSubmitter: **number of splits:** N

- TXT: 15
- GZ: 1
- BZ2: 3

Можем сделать мелким (например, 32Мб):

```
-Dmapreduce.input.fileinputformat.split.maxsize=$[ 32 * 2**20 ]
```

InputFormat-ы



Чем регулируется как читать?

TextInputFormat.*addInputPath*(…);

Задает текстовый формат входа.

Разбиение: по HDFS-блокам, чтение – построчное

Но есть другие InputFormat-ы:

- DBInputFormat
- NLineInputFormat
- CombineFileInputFormat
- …

Что если нужен свой InputFormat?

Можно определить собственный.

Что необходимо:

- При запуске задачи:
  - Знать как разбить входные данные
  - Запуск mapper-ов согласно разбиению
- При запуске mapper-ов:
  - Выдать поставщик записей <key, value>
  - Будет вызываться дя функции map()

Как это ложиться на API?

FileInputFormat<K, V>

- List<InputSplit> getSplits
- RecordReader<K,V> createRecordReader()

RecordReader<K, V>

- void initialize()
- boolean nextKeyValue()
- K getCurrentKey()
- V getCurrentValue()
- float getProgress()

см. src/main/java/BMPCollectionInputFormat.java

А в самой Job-е?

Используем этот InputFormat:

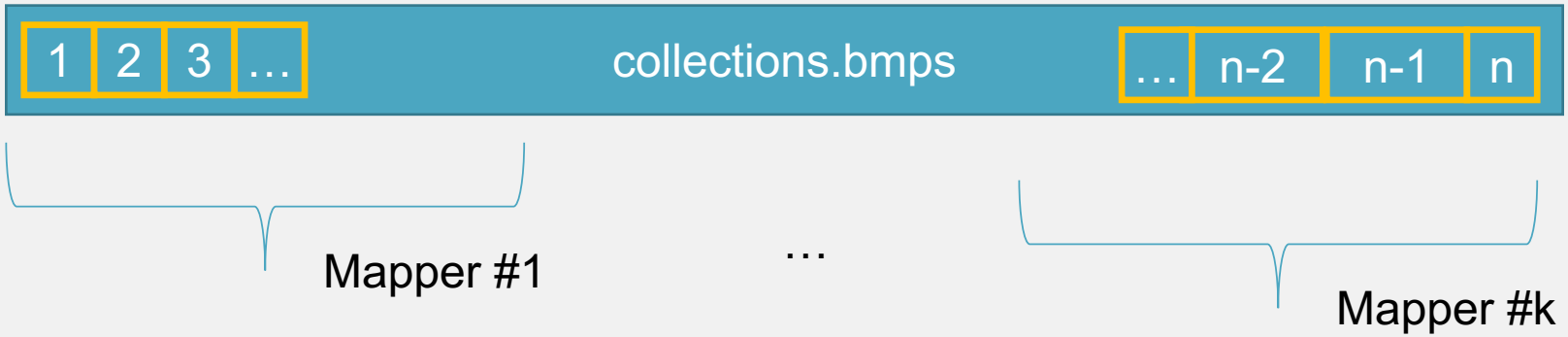
```
job.setInputFormatClass(BMPCollectionInputFormat.class);
FileInputFormat.addInputPath(job, new Path(input));
```

Используем данные в Mapper-е:

```
class ImgConverterMapper extends Mapper<K, V, …>
```

(см. ImgConverterJob.java)

Что должна делать ImgConverterJob?



- Все bmp изображения строго одинакового размера (getNumBytesPerImage)
- Необзодимо нарезать файл по границам изображений
  - Глупо создавать mapper на каждое изображение (getNumBytesPerSplit)
- Каждый mapper конвертирует свои изображения в jpeg
- Выводит в base64

Задание

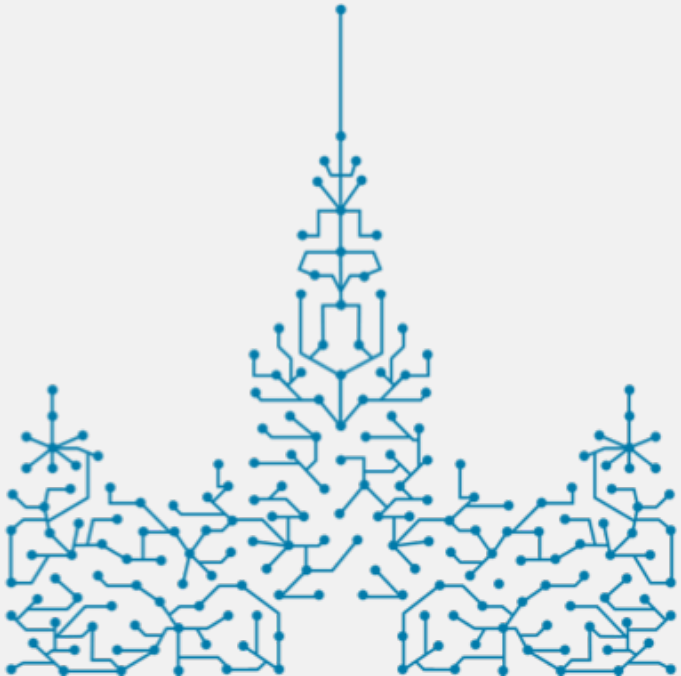
- Заправить BMPCollectionInputFormat.java
- Собрать hadoop_sem1.jar (./gradlew jar)
- Запустить в hadoop
- Собрать итоговый ролик

```
$ hadoop jar build/libs/hadoop_sem1.jar ImgConverterJob \
  /data/seminar1/collection.bmps jpegs

$ hadoop fs -get jpegs . # и скопируйте себе локально
$ ./make_movie.py jpegs/part-*

# посмотрите out.mp4 на локальной машине
```

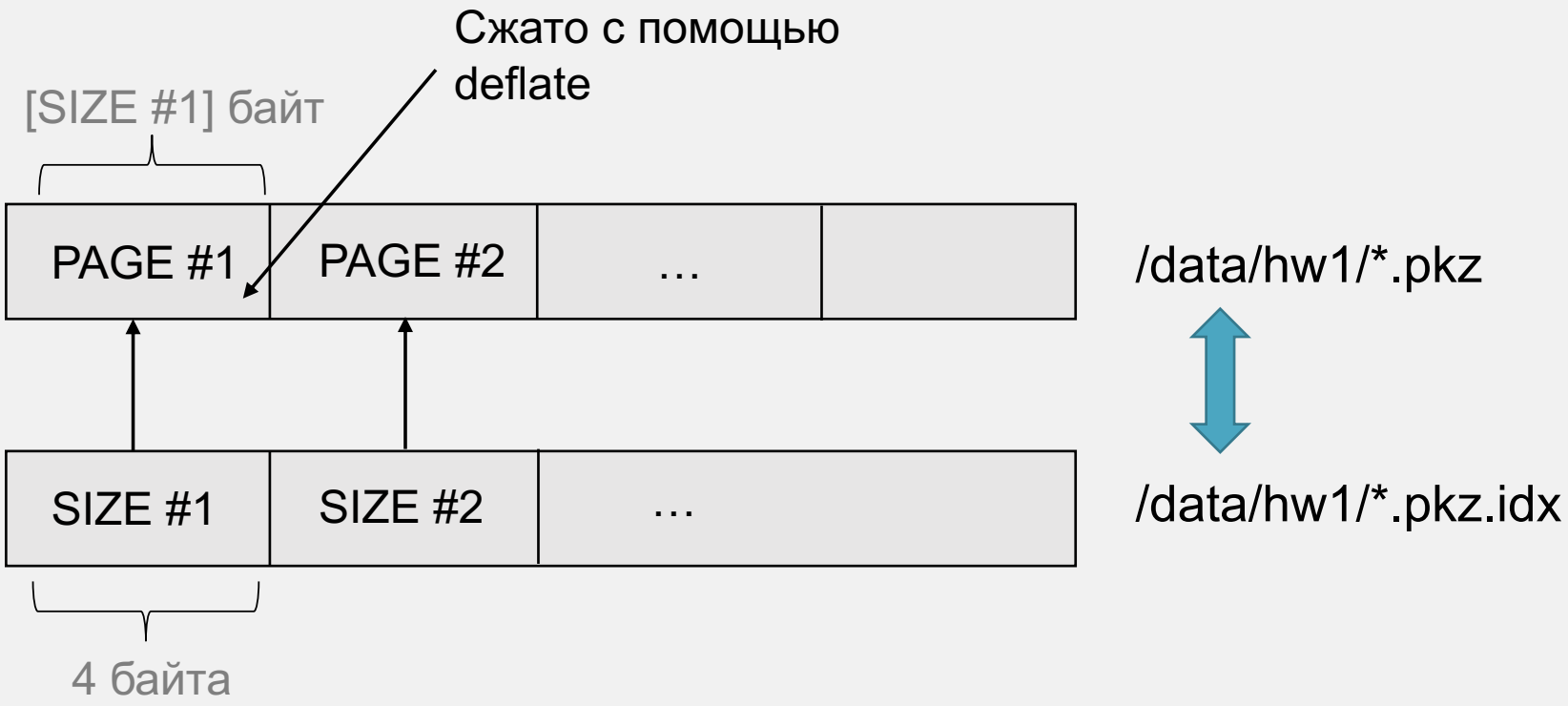
Домашнее задание



ДЗ

- Имеется дамп страниц новостного сайта
- Необходимо посчитать в скольких документах встречается каждое слово
- … реализовав свой InputFormat

ДЗ – формат входных данных



Куда присылать?

gz с исходным кодом и jar:

ts2018-hadoop@mail.ru

Тема: "[hadoop-ts] inputformat: Иван Иванов"