MapReduce (2)

План семинара

1. Элементы Hadoop Main API. Элементы классов IO
2. Wordcount. Две MapReduce job'ы (screencast)
3. Reduce-side Join

Конспект

https://gitlab.com/pd2020-supplementary/8xx-GLOBAL/-/blob/master/practice/09-mapreduce_part2.md

Job, Mapper, Reducer

Классы должны быть помещены в один JAR.

Определение **main-класса**

```
class … extends Configured implements Tool
public int run(String[]) throws Exception;
```

Определение **Mapper**:

```
class … extends Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>
void map(KEYIN, VALUEIN, MapContext);
```

Определение **Reducer**:

```
class … extends Reducer<KEYIN,VALUEIN,KEYOUT,VALUEOUT>
void reduce(KEYIN, Iterable<VALUEIN>, ReduceContext);
```

Определение task производится классом **Job**:

```
Job job = Job.getInstance(); job.setMapperClass(M.class);
```

Классы ключей и значений

Кортежи **промежуточных стадий** должны реализовывать `interface WritableComparable<T>`

Кортежи **на входе и выходе** должны реализовывать `interface Writable`

Входные и выходные данные из файлов могут быть обработаны с помощью различных **классов-formatter'ов**. Перечень встроенных:
https://timepasstechies.com/input-formats-output-formats-hadoop-mapreduce/

Установка форматов ввода-вывода выполняется в **`Job`**

Есть дополнительные возможности (компрессия, …)

Wordcount

(screencast)
/home/velkerr/seminars/pd2020/05_wordcount_java

Wordcount с (глобальной) сортировкой

(screencast)

/home/velkerr/seminars/pd2020/10-globalsort

http://blog.ditullio.fr/2016/01/04/hadoop-basics-total-order-sorting-mapreduce/

Join (reduce-side)

(screencast)

/home/mobod2020/mob202007/pd2020/11-joins

https://gitlab.com/pd2020-supplementary/8xx-GLOBAL/-/blob/master/practice/09-mapreduce_part2.md#схема-reduce-side-join

Полезные материалы

Документация Hadoop 2 Main API
https://hadoop.apache.org/docs/r2.4.1/api/overview-summary.html