

# HW #03: MapReduce Optimization

---

1. Описание задания: популярные теги stackoverflow	<b>2</b>
2. Критерии оценивания	<b>3</b>
3. Job Chaining	<b>4</b>
4. Рекомендации по выполнению ДЗ	<b>4</b>
5. Правила оформления задания	<b>5</b>

---

автор задания: BigData Team, коллективная работа.



## 1. Описание задания: популярные теги stackoverflow

В данном ДЗ нужно решить 1 задачу. Решение нужно выполнить с помощью Hadoop Streaming.

На основе выборки из постов stackoverflow необходимо найти TOP-10 самых популярных тегов, которые люди ставили в 2010 и в 2016 годах (соответственно).

Ограничения:

- из тегов удалить ненужные html-символы < и >. Например, если на входе Tags="<html><browser><timezone>", то тегами будут html, browser и timezone;
- Тройки (year, tag, counts) отсортировать сначала по году (по возрастанию), затем по counts (по убыванию).

### Входные данные

Stackoverflow:

- Путь на кластере: /data/stackexchange/posts
- Семпл (для тестирования): /data/stackexchange\_part/posts
- Формат: XML;
- Необходимо рассматривать только строки, начинающиеся на "<row" (в начале строки могут быть еще пробельные символы)

Пример:

```
<row Id="13" PostTypeId="1" AcceptedAnswerId="357"
CreationDate="2008-08-01T00:42:38.903" Score="440" ViewCount="128370"
Body="<p>Is there any standard way for a Web Server to be able to
determine a user's timezone within a web page? Perhaps from a HTTP
header or part of the user-agent string?</p>" OwnerUserId="9"
LastEditorUserId="3604745" LastEditorDisplayName="Rich B"
LastEditDate="2016-11-29T02:17:23.667"
LastActivityDate="2016-11-29T02:17:23.667" Title="Determine a User's
Timezone" Tags="<html><browser><timezone><timezoneoffset>"
AnswerCount="24" CommentCount="3" FavoriteCount="120" />
```

### Выходные данные

формат вывода (HDFS):

year <tab> tag <tab> число\_постов\_с\_указанным\_тегом\_в\_заданный\_год

Вывод на печать (STDOUT):

вывести TOP-10 тегов для каждого года, сначала для 2010, затем - для 2016.

Пример вывода (посчитан на подвыборке Stackoverflow):

```
2010 .net 2139
2010 asp.net 2041
2016 javascript 9263
2016 java 7435
2016 python 6183
```

## 2. Критерии оценивания

Балл за задачу складывается из:

- **60%** - правильное решение задачи
- **20%** - поддерживаемость и читаемость кода
  - в общем случае см. Clean Code и [Google Python Style Guide](#)
  - оценка качества будет проводиться автоматическим вызовом pylint:
    - `pylint *.py -d invalid-name,missing-docstring`
    - качество кода должно оцениваться выше 8.0 / 10.0
    - проверяем код **Python версии 3** с помощью `pylint==2.5.3`
- **20%** - эффективность решения (для сравнения: решение должно обрабатывать<sup>1</sup> в течение 5 минут на ресурсах 3х вычислительных узлов; в решении для закрепления навыков обязательно использование сложного ключа и должны использоваться как минимум 2 из 3х оптимизаций: combiner, partitioner, comparator).

Discounts (скидки и другие акции):

- **100%** за плагиат в решениях (всем участникам процесса)
- **100%** за посылку решения после deadline
- **5%** за каждую дополнительную посылку в тестирующую систему (одна дополнительная посылка бесплатно)

Формула подсчета финальной оценки<sup>2</sup>:

$\max(0, 0.95^{\max(0, \# \text{доп. посылки} - 1)} * (1 - \text{штраф за дедлайн и списывание})) * \text{оценка по тестам}$

<sup>1</sup> Оценка производится на основе счетчика "CPU time spent (ms)"

<sup>2</sup> результат умножается на 10 (максимальная оценка) и округляется до первой цифры после точки



## 3. Job Chaining

При решении задач старайтесь использовать оптимальный MapReduce-алгоритм:

- использовать как можно меньшее кол-во Hadoop Job;
- использовать combiner для ускорения вычислений;
- использовать больше, чем 1 reducer (1 reducer разрешается использовать только в финальной job'e, при сортировке результата)

Пример запуска связанных MapReduce задач (Job Chaining), представлен в run.sh доступному на github:

[github:big-data-team/big-data-course/.../map\\_reduce/job\\_chain/run\\_job\\_chain.sh](https://github.com/big-data-team/big-data-course/blob/master/map_reduce/job_chain/run_job_chain.sh)

Обратите внимание на конструкцию "`( ... && ... ) || echo 'smth'` ", которая позволяет отлавливать исключительные события и не запускать зависимую задачу, если первая не отработала.

Для удобства копирования run.sh, count\_mapper.py и sum\_reducer.py доступны по адресу:

[github:big-data-team/big-data-course/.../map\\_reduce/job\\_chain](https://github.com/big-data-team/big-data-course/blob/master/map_reduce/job_chain)

## 4. Рекомендации по выполнению ДЗ

Чтобы быть уверенным, что Grader (скрипт оценки решения) правильно обработает ваше решение, предлагаем следующие рекомендации:

- Для временных данных используйте HDFS-папку с суффиксом `_tmp` (например `my_hdfs_folder_tmp`);
- Убедитесь, что Вы удаляете все временные данные после завершения выполнения задачи;
- Отслеживайте код возврата MapReduce задач (Job'ов). В случае ошибки первой задачи в цепочке **нет** необходимости запускать следующие;
- Обращайте внимание на вывод в "STDOUT". Его форматирование является критически важным для прохождения тестов. Формат должен соответствовать выходному HDFS-формату. Вам нужно прочитать ровно столько строчек в STDOUT из HDFS, сколько указано в задании;
- Вы **НЕ** можете прочитать весь HDFS output в RAM для сортировки. Даже если получится с игрушечными примерами на нашем кластере, в бою это будет больно отстреливать в ногу;



- Вероятно Вы решите задачу в 2+ стадии MapReduce, **От Вас ожидается посчитать статистику по всем парам** (tag, year) на первой стадии, фильтрацию можно производить только на стадии 2.

## 5. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW3:MapReduce-advanced(Stackoverflow)**.
- Выполненное ДЗ запакуйте в архив **MADEBD2021Q1\_<Surname>\_<Name>\_HW#.zip**, например, для Алексея Драля -- **MADEBD2021Q1\_Dral\_Alexey\_HW3.zip** (проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда<sup>3</sup>). Если ваше решение лежит в папке my\_solution\_folder, то для создания архива hw.zip на Linux и Mac OS выполните команду<sup>4</sup>:
  - `zip -r hw.zip my_solution_folder/*`
- На Windows 7/8/10: необходимо выделить все содержимое директории my\_solution\_folder/ нажать правую кнопку мыши на одном из выделенных объектов, выбрать в открывшемся меню "Отправить >", затем "Сжатая ZIP-папка". Теперь можно переименовать архив.
- Решение задания должно содержаться в одной папке.
- Скрипт для запуска решения должен называться **run.sh**:
  - скрипт будет запускаться с помощью команды:

```
bash run.sh $(input_ids_hdfs_path) $(output_hdfs_path) $(job_name)
```

- скрипт читает данные из HDFS-папки, указанной первым аргументом (используйте \$1 в run.sh)
- скрипт сохраняет данные в HDFS папку \$2
- скрипт очищает все временные директории в HDFS до и после запуска вычислений, выходящая папка будет предварительно очищена фреймворком для проверки решения
- run.sh не должен содержать "echo \$?", поскольку эта информация будет содержаться в STDOUT и использоваться для оценки решения
- скрипт выводит на экран (STDOUT) указанное в задании число строк в нужном формате<sup>5</sup>
- вывод STDOUT сохраните в файл **hw3\_mr\_advanced\_output.out** и приложите к архиву с решением

<sup>3</sup> Онлайн инструмент для проверки: <https://www.soscisurvey.de/tools/view-chars.php>

<sup>4</sup> Флаг -r значит, что будет совершен рекурсивный обход по структуре директории

<sup>5</sup> См. `hdfs dfs -cat`

- скрипт использует следующий путь до `hadoop-streaming.jar` на кластере:  
`/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming.jar`
- в заголовке `bash`-скрипта указана опция `"set -x"`, вывод `STDERR` никуда не перенаправляется (он используется для анализа логов исполнения задачи)
- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
  - | `MADEBD2021Q1_<Surname>_<Name>_HW3.zip`
  - | `---- run.sh`
  - | `---- *.py`
  - | `---- hw3_mr_advanced_output.out`
  - При несовпадении дерева вашего архива с представленным деревом ваше решение не будет возможным автоматически проверить, а значит, и оценить его.
- Для того, чтобы сдать задание необходимо:
  - Зарегистрироваться и залогиниться в сервисе [Everest](#)
  - Перейти на страницу приложения: [BDT-grader-MADE-BD](#)
  - Выбрать вкладку `Submit Job` (если отображается иная).
  - Выбрать в качестве `"Task"` значение:  
**HW3:MapReduce-advanced(Stackoverflow)<sup>6</sup>**
  - Загрузить в качестве `"Task solution"` файл с решением
  - В качестве `Sender ID` указать тот, который был выслан по почте
- Если Вы видите надпись `"You are not allowed to run this application"` во вкладке `Submit Job` в Everest, то на данный момент сдача закрыта (нет доступных для сдачи домашних заданий, по техническим причинам или другое). Попробуйте, пожалуйста, еще раз через некоторое время. Если Вы еще ни разу не сдавали, у коллег сдача работает, но Вы видите такое сообщение, сообщите нам об этом.
- Ситуации:
  - \* система оценивания показывает оценку `(Grade) < 0`, а отчет `(Grading report)` не помогает решить проблему (пример помощи: в случае неправильно указанного `Sender ID` система вернет `-2` и информацию о том, что его нужно поправить);
  - \* показывает `0` и в отчете `(Grading report)` не указано, какие тесты не пройдены. Если Вы столкнулись с какой-то из них присылайте ссылку на выполненное задание (`Job`) на почту с темой письма `"Short name. ФИО."`. Например: **"HW3:MapReduce-advanced(Stackoverflow). Иванов Иван Иванович."**  
Пример ссылки: <https://everest.distcomp.org/jobs/67893456230000abc0123def>
- **Внимание:** Если до дедлайна остается меньше суток, и Вы знаете (сами проверили или коллеги сообщили), что сдача решений сломана, обязательно сдайте свое решение и напишите письмо, как написано выше, чтобы мы видели, какое решение Вы имели до дедлайна и смогли его оценить.
- Перед отправкой задания, оставьте, пожалуйста, отзыв о нём по ссылке: [http://rebrand.ly/mailbd2021q1\\_feedback\\_hw](http://rebrand.ly/mailbd2021q1_feedback_hw). Это позволит скорректировать

<sup>6</sup> Сервисный ID: `map_reduce.stackoverflow`



# BIGDATA TEAM

учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересующие вопросы.

Любые вопросы / комментарии / предложения можно писать в [Discord-канал курса](#) или на почту [bigdata\\_made2021q1@bigdatateam.org](mailto:bigdata_made2021q1@bigdatateam.org).

Всем удачи!