

HW #03: MapReduce-advanced

1. Описание задания	2
2. Задача (Task ID: map_reduce.stackoverflow): популярные теги stackoverflow	2
3. Критерии оценивания	3
4. Job Chaining	4
5. Общие рекомендации	5
6. Правила оформления задания	5

автор задания: BigData Team, коллективная работа.



1. Описание задания

В данном ДЗ нужно решить 1 задачу. Решение надо выполнить на Hadoop Streaming¹.

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)

Перед решением задачи **обязательно** изучите разделы (4) Job Chaining и (5) Общие рекомендации.

2. Задача (Task ID: `map_reduce.stackoverflow`): популярные теги `stackoverflow`

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 строчки, 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"
```

¹ Java-разработчикам просьба обратить внимание на документацию по Hadoop Java API по адресу <http://hadoop.apache.org/docs/r2.6.1/api/>



```
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
```

3. 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](https://google.github.io/styleguide/python/)
 - 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 обрабатывать² 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).

² Оценка производится на основе счетчика "CPU time spent (ms)"



Discounts (скидки и другие акции):

- **100%** за плагиат в решениях (всем участникам процесса)
- **100%** за посылку решения после hard deadline
- **30%** за посылку решения в после soft deadline и до hard deadline
- **5%** за каждую новую посылку (одна дополнительная посылка бесплатно)

4. Job Chaining

Пример запуска связанных MapReduce задач (Job Chaining), представлен, ниже.

run.sh (обратите внимание на конструкцию "(... && ...) || echo 'smth' "):

```
#!/usr/bin/env bash
set -x
```

```
HADOOP_STREAMING_JAR=/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar
OUT_DIR="streaming_wc_result"
NUM_REDUCERS=8
```

```
hdfs dfs -rm -r -skipTrash ${OUT_DIR}* > /dev/null
```

```
# Wordcount
```

```
( yarn jar $HADOOP_STREAMING_JAR \
  -D mapreduce.job.name="Streaming WordCount" \
  -files count_mapper.py,sum_reducer.py \
  -mapper "python3 count_mapper.py" \
  -reducer "python3 sum_reducer.py" \
  -numReduceTasks $NUM_REDUCERS \
  -input /data/wiki/en_articles_part \
  -output ${OUT_DIR}_tmp &&
```

```
# Global sorting as we use only 1 reducer
```

```
yarn jar $HADOOP_STREAMING_JAR \
  -D
  mapreduce.job.output.key.comparator.class=org.apache.hadoop.mapreduce.lib.partition.KeyFieldBasedComparator \
  -D mapreduce.map.output.key.field.separator=. \3
  -D mapred.text.key.comparator.options="-k2,2nr" \
```

³ Если у вас разделитель <tab> и вы хотите использовать 2 колонки для ключа, то указанную строчку нужно заменить на:

```
-D stream.num.map.output.key.fields=2 \
```



```
-mapper cat \  
-reducer cat \  
-numReduceTasks 1 \  
-input ${OUT_DIR}_tmp \  
-output ${OUT_DIR}  
) || echo "Error happens"
```

```
hdfs dfs -rm -r -skipTrash ${OUT_DIR}_tmp
```

```
hdfs dfs -cat ${OUT_DIR}/part-00000 | head
```

Для удобства копирования run.sh, count_mapper.py и sum_reducer.py доступны по адресу:

`/usr/local/share/big_data_course/public_examples/job_chain`

5. Общие рекомендации

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.



6. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW3:MR++(SO)**.
- Выполненное ДЗ запакуйте в архив **X5BD2021Q1_<Surname>_<Name>_HW#.zip**, пример -- **X5BD2021Q1_Dral_Alexey_HW3.zip**. (Проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда.⁴) Если ваше решение лежит в папке `my_solution_folder`, то для создания архива `hw.zip` на Linux и Mac OS выполните команду⁵:
 - `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 формате⁶
- вывод `STDOUT` сохраните в файл `hw3_mr_advanced_output.out` и приложите к архиву с решением
- 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)

⁴ Онлайн инструмент для проверки: <https://www.soscisurvey.de/tools/view-chars.php>

⁵ Флаг `-r` значит, что будет совершен рекурсивный обход по структуре директории

⁶ См. `hdfs dfs -cat`



- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
 - | X5BD2021Q1_<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-X5-BD](#)
 - Выбрать вкладку Submit Job (если отображается иная).
 - Выбрать в качестве "Task" значение: **HW3:MR++(SO)**⁷
 - 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:MR++(SO). 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 домашнем задании по ссылке: http://rebrand.ly/x5bd2021q1_feedback_hw03. Это позволит нам скорректировать учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересные вопросы.

Любые вопросы / комментарии / предложения можно писать в телеграм-канал курса или на почту bigdata_x52021q1@bigdatateam.org . Всем удачи!

⁷ Сервисный ID: map_reduce.stackoverflow