BIGDATA TEAM

HW #02: MapReduce

автор задания:
- Горохов Антон, anton.gorokhov@bigdatateam.org
- Big Data Instructor @ BigData Team
- Senior SDE @ Yandex

редакторы задания[1]:
- Александр Ким, Николай Попов*, Ксения Пеньевская**
- Big Data Mentor @ BigData Team
- *Data Engineer @ inDriver
- **Big Data Analyst

[1] Хочешь стать ментором и оставить след в истории Big Data? Тогда хорошо учись, помогай другим и дай нам знать о своем желании. Смело пиши преподавателям и менеджерам учебных курсов.

1. Описание задания

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

Представьте следующую ситуацию: вам нужно оценить поведение нового сервиса (например, базу данных) под нагрузкой. Для этого вы решаете “обстрелять” сервис и залогировать его поведение. На первом этапе вам нужно подготовить “патроны”, которые будут представлять запросы к этому сервису (БД). Вам известен список ключей, которые могут быть в этой базе, а также вам известно, что в одном запросе таких ключей до 5 штук (включительно).

Таким образом, ваша задача состоит в следующем: имея список идентификаторов, перемешать его в случайном порядке. Далее в каждой строке записать через запятую случайное число идентификаторов - от 1 до 5.

1.1. Входные данные

Список идентификаторов:

- Путь на кластере: полный датасет - **`/data/ids`**, семпл - **`/data/ids_part`**
- Формат: текст, один идентификатор в строке

1.2. Выходные данные

Формат вывода (HDFS и STDOUT):

```
id1,id2,...
...
```

Вывод MapReduce задачи в HDFS должен содержать все идентификаторы, которые были на входе.

Вывод на печать (STDOUT): первые 50 строк.

Пример вывода:

```
1cf54b530128257d72,4cdf3efa01036a9a48,8c3e7fb30261aaf9cf
4cfe6230016553c3ed,76e1b8690176f801bb,e7409c39013c9db7b4,a5f1519c02b22550e6
83a119ef02346d0879
...
```

1.3. Требования к реализации

Скрипт для запуска решения должен называться **run.sh** и запускаться с помощью команды:

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

Требования:

- скрипт читает данные из HDFS-папки, указанной первым аргументом (используйте `$1` в `run.sh`), будет использоваться - `/data/ids`
- скрипт сохраняет данные в HDFS папку `$2` (можете использовать `hw02_mr_data_ids` для тестирования)
- скрипт выводит на экран (STDOUT) указанное в задании число строк в нужном формате[2]
- вывод STDOUT должен быть сохранен в файле **`hw02_mr_data_ids.out`**[3] и приложен 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)

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 C0111,C0103`
    - 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 отрабатывать в течение 5 минут на ресурсах 3-х вычислительных узлов; не должно грузить все данные в RAM для обработки как на фазе Map, так и на фазе Reduce; работать в распределенном режиме (например, использовать минимум 2 редьюсера)).

[2] См. `hdfs dfs -cat`

[3] Содержимое файла `Grader’ом` не проверяется. Служит для визуальной оценки вывода написанного решения

Discounts (скидки и другие акции):

- **100%** за плагиат в решениях (всем участникам процесса)
- **100%** за посылку решения после hard deadline
- **30%** за посылку решения после soft deadline и до hard deadline
- **5%** за каждую дополнительную посылку в тестирующую систему (всего можно делать до 3-х посылок без штрафа):

Пример работы системы штрафов:

День	Посылка	Штраф
День 1	Посылка 1	Без штрафа
День 1	Посылка 2	Без штрафа
День 1	Посылка 3	Без штрафа
День 1	Посылка 4	-5%
День 2	Посылка 5	-5%
День 3	Посылка 6	-5%
Итоговый штраф: -15%		

Для подсчета финальной оценки **всегда** берется **последняя** оценка из Grader.

3. Инструкция по отправке задания

Перед отправкой задания оставьте, пожалуйста, отзыв о домашнем задании по ссылке: https://rebrand.ly/bdb2c2022q2_feedback_hw. Это позволит нам скорректировать учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересующие вопросы.

Оформление задания:

- Код задания (Short name): **HW02:MapReduce**.
- Выполненное ДЗ запакуйте в архив **BD-B2C-2022-Q2**<Surname>_<Name>_HW#**.zip**, пример -- **BD-B2C-2022-Q2**_Dral_Alexey_**HW02**.zip. (Проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда.[4]) Если ваше

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

решение лежит в папке `my_solution_folder`, то для создания архива hw.zip на Linux и Mac OS выполните команду[5]:
    - `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.
- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
    - **`| BD-B2C-2022-Q2_<Surname>_<Name>_HW02.zip`**
    - **`| ---- run.sh`**
    - **`| ---- mapper.py`**
    - **`| ---- reducer.py`**
    - **`| ---- hw02_mr_data_ids.out`**
    - При несовпадении дерева вашего архива с представленным деревом, ваше решение будет невозможно автоматически проверить, а значит, и оценить его.
- Для того, чтобы сдать задание, необходимо:
    - Зарегистрироваться и залогиниться в сервисе Everest
    - Перейти на страницу приложения: B2C Big Data Grader
    - Выбрать вкладку Submit Job (если отображается иная).
    - Выбрать в качестве “Task” значение: **HW02:MapReduce** [6]
    - Загрузить в качестве “Task solution” файл с решением
    - В качестве Access Token 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 Access Token система вернет -2 и информацию о том, что его нужно поправить);
  * система 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: “**HW02:MapReduce**. Иванов Иван Иванович.”
  Пример ссылки: https://everest.distcomp.org/jobs/67893456230000abc0123def

[5] Флаг -r значит, что будет совершен рекурсивный обход по структуре директории
[6] Сервисный ID: map_reduce.ids

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.

Любые вопросы / комментарии / предложения можно писать в телеграм-канал курса или на почту bd_b2c2022q2@bigdatateam.org.

Всем удачи!

4. FAQ (часто задаваемые вопросы)

Никогда не программировал на Python, где можно получить максимально быстрый ликбез?

Быстрое введение в основы работы с языком Python:

- https://pythonworld.ru/samouchitel-python
- https://learnxinyminutes.com/docs/python/

Необходимый минимум для данного ДЗ:
Операции над числами, строками, сравнения, присваивания; индексы и срезы, списки, циклы, условные операторы и умение импортировать библиотеки.

"You are not allowed to run this application", что делать?

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.

Grader показывает 0 или < 0, а отчет (Grading report) не помогает решить проблему

Ситуации:

- система оценивания показывает оценку (Grade) < 0, а отчет (Grading report) не помогает решить проблему. Пример: в случае неправильно указанного access token система вернет -401 и информацию о том, что его нужно поправить;
- система показывает 0 и в отчете (Grading report) не указано, какие тесты не пройдены. Пример: вы отправили невалидный архив (rar вместо zip), не приложили нужные файлы (или наоборот приложили лишние - временные файлы от Mac OS и т.п.), рекомендуется проверить содержимое архива в консоли:

```
unzip -l your_solution.zip
```

Если Вы столкнулись с какой-то из них, присылайте ссылку на выполненное задание (Job) в чат курса. Пример ссылки:
https://everest.distcomp.org/jobs/67893456230000abc0123def

Что в отчете Grader означает проверка X ?

Правильность решения задачи:
test_unzip_is_succesful - ДЗ заархивировано в .zip архив и грейдер может его разархивировать

test_map_reduce_execution_is_successful - map-reduce задача выполнилась без ошибок

test_run_output_contains_expected_line_count - run.sh выводит из результата map-reduce задачи, записанного в $(output_hdfs_path), ровно 50 строчек в STDOUT (см. команды hdfs dfs)

test_expected_reduce_output_records_count - количество строк в результате map-reduce задачи, записанном в $(output_hdfs_path), попадает в ожидаемый интервал (между минимальным возможным значением и максимальным возможным)

test_run_output_contains_expected_distribution_of_ids - в выводе в STDOUT в каждой строке количество идентификаторов - от 1 до 5 и отсутствуют пустые строки

test_run_output_is_not_globally_sorted
Глобальная сортировка означает следующее: если мы отсортируем идентификаторы внутри строчек, а затем все строчки склеим в один массив, то мы получим полностью отсортированный массив. Указанный тест проверяет, что глобальной сортировки по id - нет (на основе вывода в STDOUT).

test_run_output_contains_randomly_sorted_ids - если id внутри каждой строки в STDOUT отсортированы, значит вы делаете что-то не так. Тест проверяет, что найдется хотя бы одна строка, где идентификаторы не отсортированы.

test_all_ids_are_of_expected_size - в выводе в STDOUT каждый отдельно взятый идентификатор не видоизменялся и имеет ожидаемую длину (какую имел в датасете)

test_does_not_load_ids_in_memory - map-reduce задача работала с датасетом на стриминге (потоке), а не загружая все данные в оперативную память на Mapper'е или Reducer'е

Поддерживаемость и читаемость кода:
test_py_files_min_lint_score - качество кода в .py файлах оценивается выше 8.0

Эффективность решения:
test_execution_time_below_threshold - map-reduce задача выполняется не дольше 5 минут (на основе счетчика "CPU time spent (ms)")

test_expected_reduce_task_count - фаза reduce происходит в распределенном режиме (см. -numReduceTasks)

Appendix. Подсказки (если не получается решить ДЗ).

Если у Вас возникли трудности с пониманием, что нужно сделать:

1. Пройдите Workshop по MapReduce
2. Попробуйте самостоятельно запустить решение по подсчету слов на кластере (github word_count)

Для случайного выбора элементов последовательности можно использовать библиотеку random (random.randint).

Для вывода на печать (STDOUT) первых 50 строк после команды hdfs dfs -cat … применить команду head -n кол-во строк (hdfs dfs -cat ${2}/part-00000 | head -n 50)

При реализации перестановок можно воспользоваться следующей идеей:

1. Добавьте к каждому ID префикс в виде случайного числа.
2. Отсортируйте ID с помощью MapReduce.
3. Сгруппируйте ID по группам, длина группы от 1 до 5.
4. Удалите все префиксы перед выводом.