BIGDATA TEAM

# HW #02: MapReduce

Deadline: 20.06.2019, 08:00



## 1. Описание задания.

В данном ДЗ нужно решить 1 задачу. Решение надо выполнить на Hadoop Streaming (для желающих, можно на Java, для этого см. документацию по Hadoop Java API по адресу - http://hadoop.apache.org/docs/r2.6.1/api/).

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 в этой базе могут быть, 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.

**Входные данные**

Список идентификаторов:

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

**Выходные данные**

Формат вывода (HDFS):

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

Вывод на печать (STDOUT): первые 50 строк.

Пример вывода:

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

## 2. Критерии оценивания.

Балл за задачу складывается из:

- **60%** - правильное решение задачи
- **20%** - поддерживаемость и читаемость кода (Clean Code, см. например Google Python Style Guide)
- **20%** - эффективность решения

Штрафы:

- **10%** за несоответствие правилам оформления задания
- **30%** за просрочку дедлайн

## 3. Сроки сдачи и правила оформления задания.

## Deadline: 20.06.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW2:MapReduce**.
- Решение задания должно содержаться в одной папке.
- Скрипт для запуска решения называется **run.sh**:
  - скрипт читает данные из HDFS-папки /data/ids
  - скрипт сохраняет данные в HDFS папку hw2_mr_data_ids
  - скрипт предварительно очищает данные из HDFS, чтобы туда можно было записать результаты вычислений

    - скрипт выводит на экран (STDOUT) указанное в задание число строк в нужном формате[1]
    - Вывод STDOUT сохраните в файл hw2_mr_data_ids.out
- Выполненное ДЗ запакуйте в архив MF2019Q2_<фамилия>_HW#.zip , к примеру -- MF2019Q2_Ivanov_HW2.zip. Например, ваше решение лежит в папке my_solution_folder, тогда чтобы на Linux и Mac OS создать архив под названием hw2.zip и пожать его с помощью zip выполните команду[2]:
    - zip -r hw2.zip my_solution_folder/

  На Windows 7/8/10: необходимо нажать правую кнопку мыши на директорию my_solution_folder/, выбрать в открывшемся меню "Отправить >", затем "Сжатая ZIP-папка". Теперь можно переименовать архив.
- Присылайте выполненное задание на почту **bigdata_mf2019q2@bigdatateam.org** с темой письма “Short name. ФИО.”. Например: “HW2:MapReduce. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по сссылке: http://rebrand.ly/mf2019q2_feedback_hw02. 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.

Любые вопросы / комментарии / предложения можно писать:

- в телеграм-канале: http://rebrand.ly/mf2019q2_telegram_join
- На почту: bigdata_mf2019q2@bigdatateam.org

## 4. Дорешка.

Решения после получения фидбека на решение ДЗ можно улучшить. Разрешается одна досылка в течение 1й недели после окончания дедлайна за ДЗ. Соответственно, фидбек за дорешенные ДЗ вы получите в течение 24 часов после окончания deadline следующего ДЗ.

Дорешивать неработающие задания - нельзя. Это позволит исправить дисбаланс между

присланными **НЕработающими** заданиями **ДО** deadline

VS

присланными **работающими** заданиями **ПОСЛЕ** deadline

Всем удачи!

[1] См. hdfs dfs -cat

[2] Флаг -r значит, что будет совершен рекурсивный обход по структуре директории

## Appendix. Подсказки (если не получается решить ДЗ).

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.