BIGDATA TEAM

HW #02: MapReduce

автор задания:

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

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
  - 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
- **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)).

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[1]:

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

[1] результат умножается на 10 (максимальная оценка) и округляется до первой цифры после точки

3. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW2:MapReduce(IDs)**.
- Выполненное ДЗ запакуйте в архив **MADEBD2021Q1**_<Surname>_<Name>_HW#.zip, пример -- **MADEBD2021Q1**_Dral_Alexey_HW2.zip. (Проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда.[2]) Если ваше решение лежит в папке my_solution_folder, то для создания архива hw.zip на Linux и Mac OS выполните команду[3]:
  - `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), будет использоваться - /data/ids
  - скрипт сохраняет данные в HDFS папку $2 (можете использовать hw2_mr_data_ids для тестирования)
  - скрипт выводит на экран (STDOUT) указанное в задание число строк в нужном формате[4]
  - вывод STDOUT сохраните в файл **hw2_mr_data_ids.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)
- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
  - **| MADEBD2021Q1_<Surname>_<Name>_HW2.zip**
  - **| ---- run.sh**
  - **| ---- mapper.py**
  - **| ---- reducer.py**
  - **| ---- hw2_mr_data_ids.out**

[2] Онлайн инструмент для проверки: https://www.soscisurvey.de/tools/view-chars.php
[3] Флаг -r значит, что будет совершен рекурсивный обход по структуре директории
[4] См. hdfs dfs -cat

    - 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: BDT-grader-MADE-BD
    - Выбрать вкладку Submit Job (если отображается иная).
    - Выбрать в качестве “Task” значение: **HW2:MapReduce(IDs)** [5]
    - 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: “**HW2:MapReduce(IDs)**. 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. Это позволит 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 Discord-канал курса или на почту bigdata_made2021q1@bigdatateam.org .

Всем удачи!

[5] Сервисный ID: map_reduce.ids

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.