BIGDATA TEAM

HW #09: Real Time

Deadline: 19.08.19, 23:59

1. Описание задания. 1

2. Критерии оценивания. 1

3. Описание данных. 2

4. Условие задачи: поиск наиболее популярных user-agent 3

5. Сроки сдачи и правила оформления задания. 3

6. Дорешка. 4

1. Описание задания.

В данном ДЗ нужно решить **1 задачу**. Решение надо выполнить с помощью Spark Streaming.

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, для этого поможет использование линтеров (к примеру pylint, flake8) и/или форматеров (к примеру black))
- **20%** - эффективность решения

Штрафы:

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

3. Описание данных.

- Путь на кластере: **/data/course4/uid_ua_100k_splitted_by_5k**
- Формат: tsv
- В каждой строке находятся следующие поля, разделенные знаком табуляции:
  - STRING - UID пользователя,
  - STRING - User-Agent пользователя

Пример:

```
f78366c2cbed009e1febc060b832dbe4	Mozilla/5.0 (Linux; Android 4.4.2; T1-701u Build/HuaweiMediaPad) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3202.73 Safari/537.36
62af689829bd5def3d4ca35b10127bc5	Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36
a84ad52d5846a231102d867c28a1c008	Mozilla/5.0 (iPad; CPU OS 7_0_4 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like Gecko) Version/7.0 Mobile/11B554a Safari/9537.53
2d43b507519f7ec18ee21534ad9069a5	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3202.89 Safari/537.36
2a30a05b6bf678e4a5157419c16b9da3	Mozilla/5.0 (Linux; Android 4.3; ru-ru; SAMSUNG GT-I9500 Build/JSS15J) AppleWebKit/537.36 (KHTML, like Gecko) Version/1.5 Chrome/28.0.1500.94 Mobile Safari/537.36
1c0db4f8abc7296800ad269e6752a5c8	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/57.0.2987.137 YaBrowser/17.4.1.1026 Yowser/2.5 Safari/537.36
2d43b507519f7ec18ee21534ad9069a5	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3202.89 Safari/537.36
d310e573112dd85839afc107971fb057	Mozilla/5.0 (Windows NT 6.1; rv:56.0) Gecko/20100101 Firefox/56.0
...
```

4. Условие задачи: поиск наиболее популярных user-agent

В этом домашнем задании вам предстоит реализовать поиск наиболее популярных user_agent из предложенного лога.

Условия:

- Решение должно быть написано на Spark Streaming.
- Входные данные читаются с помощью queueStream (аналогично примеру с семинара). Один файл из папки - один батч в RT обработке.
- Ваше решение должно печатать в STDOUT топ10 самых популярных user-agent вместе с числом строк (count), в которых они встретились, в порядке убывания count. Результат это 10 пар вида:

  `count <tab> user_agent`
- Результат это кумулятивная статистика за всё время работы Streaming (для этого стоит использовать функцию updateStateByKey)
- Результат печатается по окончании обработки каждого батча

Пример результата:

```
41    Mozilla/5.0  (Windows  NT  6.1;  Win64;  x64)  AppleWebKit/537.36
(KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36
23    Mozilla/5.0    (iPad;    CPU    OS    7_0_4    like    Mac    OS    X)
AppleWebKit/537.51.1  (KHTML,  like  Gecko)  Version/7.0  Mobile/11B554a
Safari/9537.53
18    Mozilla/5.0  (Windows  NT  10.0;  Win64;  x64)  AppleWebKit/537.36
(KHTML, like Gecko) Chrome/62.0.3202.89 Safari/537.36
...
```

5. Сроки сдачи и правила оформления задания.

Deadline: 19.08.19, 23:59

Оформление задания:

- Код задания (Short name): **HW9:RealTime**.
- Решение задачи - ipython notebook или скрипт, который можно запустить через spark-submit

- Если выполняете задание в ipython notebook, то в качестве решения требуется приложить нотебук с выполненными ячейками, если выполняете задание в виде скрипта на python/java/scala, то перенаправьте stdout приложения в файл result.out и в качестве результата помимо скрипта приложите также его.
- Выполненное ДЗ запакуйте в архив MF2019Q2_<фамилия>_HW#.zip , например -- MF2019Q2_Ivanov_**HW9**.zip. Например, ваше решение лежит в папке my_solution_folder, тогда чтобы на Linux и Mac OS создать архив под названием hw.zip и пожать его с помощью zip выполните команду[1]:
    - zip -r hw.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: “**HW9:RealTime**. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по сссылке: http://rebrand.ly/mf2019q2_feedback_hw09. 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

6. 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 **работающими** заданиями **после** deadline

VS

присланными **НЕработающими** заданиями **до** deadline

Всем удачи!

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