BIGDATA TEAM

HW #08: Real Time

автор задания:
- Vybornov Artyom, avybornov@bigdatateam.org
- Big Data Instructor @ BigData Team
- Head of Data Platform @ Rambler Group

редактор задания[1]:
- Дмитрий Зверев
- Big Data Mentor @ BigData Team

[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 автору задания или менеджеру учебного курса.

1. Описание задания

В данном ДЗ нужно решить 2 задачи. Решение надо выполнить с помощью Spark Structured Streaming.

WARNING: маловероятно, но при условии перезагрузки (или прочих проблем) на сервере типа client, поток данных в Kafka может быть прерван. Для возобновления потока данных обратитесь в чатике курса к преподавателям и/или поддержке курса. При отсутствии стрима свежих данных попробуйте установить отступы на чтение данных из Kafka вручную (подробнее - Structured Streaming Kafka Integration).

1.1. Задача #1 (Task ID: realtime.domain_stat): статистика посещения доменов

В этом домашнем задании вам предстоит определить наиболее популярные домены по посещаемости и подсчитать число уников (то есть уникальных пользователей), которые зашли на этот домен.

В этом и последующем заданиях для парсинга домена используйте функцию из spark.sql:

```
select parse_url(`url_col`, 'HOST') as domain;
```

1.2. Входные данные

- Входные данные - поток событий просмотра страниц в Kafka
- Брокеры кафка:
  `brain-node1.bigdatateam.org:9092,brain-node2.bigdatateam.org:9092,brain-node3.bigdatateam.org:9092`
- Топик кафка:
  `page_views`
- Формат строчки: `tsv`
- В каждой строке находятся следующие поля, разделенные знаком табуляции:
  - `DOUBLE` - TS (unixtime) события,
  - `STRING` - UID пользователя,
  - `STRING` - URL,

- STRING - Title страницы,
- STRING - User-Agent пользователя,

Пример:

```
1522588842.557  1129fa876d6a79497387723a77d3f24c
https://www.adamas.ru/catalog/kolca/?utm_medium=cpc&utm_source=yandex.d
irect&utm_campaign=Koltsa_Msk_RSYA%7c15392911&utm_term=%25D0%25BA%25D0%
25BE%25D0%25BB%25D1%258C%25D1%2586%25D0%25BE&utm_content=k50id%7c010000
004614872683_%7ccid%7c15392911%7cgid%7c1053311384%7caid%7c5569400968%7c
adp%7cno%7cpos%7cnone0%7csrc%7ccontext_com.yandex.browser%7cdvc%7cmobil
e%7cmain&k50id=010000004614872683_&_openstat=ZGlyZWN0LnlhbmRleC5ydTsxNT
M5MjkxMTs1NTY5NDAwOTY4O2NvbS55YW5kZXguYnJvd3NlcjpndWFyYW50ZWU&yclid=162
0688752103923060
%D0%97%D0%BE%D0%BB%D0%BE%D1%82%D1%8B%D0%B5%20%D0%BA%D0%BE%D0%BB%D1%8C%D
1%86%D0%B0%20-%20%D0%BA%D1%83%D0%BF%D0%B8%D1%82%D1%8C%20%D0%BA%D0%BE%D0
%BB%D1%8C%D1%86%D0%BE%20%D0%B8%D0%B7%20%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D
0%B0%20%D0%B2%20%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%82-%D0%BC
%D0%B0%D0%B3%D0%B0%D0%B7%D0%B8%D0%BD%D0%B5%20Adamas.ru      Mozilla/5.0
(Linux; Android 7.1.2; Redmi 5 Plus Build/N2G47H) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/63.0.3239.132 YaBrowser/18.1.1.645.00 Mobile
Safari/537.36
1522588842.564  fe2042e800cbb63cff03f1152ebf74b6
https://www.gtavicecity.ru/gta-4/mods/
%D0%9C%D0%BE%D0%B4%D1%8B%20%D0%B4%D0%BB%D1%8F%20GTA%204%20%D1%81%20%D0%
B0%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0
%BE%D0%B9%20%D1%83%D1%81%D1%82%D0%B0%D0%BD%D0%BE%D0%B2%D0%BA%D0%BE%D0%B
9%3A%20%D1%81%D0%BA%D0%B0%D1%87%D0%B0%D1%82%D1%8C%20%D0%B1%D0%B5%D1%81%
D0%BF%D0%BB%D0%B0%D1%82%D0%BD%D0%BE%20%D0%BC%D0%BE%D0%B4%D1%8B%20%D0%B4
%D0%BB%D1%8F%20GTA%20IV      Mozilla/5.0 (Windows NT 6.2; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/55.0.2883.87
UBrowser/7.0.185.1002 Safari/537.36
1522588842.564  dc215986678c3b4190a102db669cf86d
https://utro.ru/politics/2018/03/29/1355676.shtml?utm_campaign=utro&utm
_medium=referral&utm_source=push
%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B0%20%D0%B6%D0%B5%D1%81%D1%82%D0%BA%D
0%BE%20%D0%BE%D1%82%D0%BF%D0%BB%D0%B0%D1%82%D0%B8%D0%BB%D0%B0%20%D0%A1%
D0%A8%D0%90%20%D0%B7%D0%B0%20%D1%81%D0%B2%D0%BE%D0%B8%D1%85%20%D0%B4%D0
%B8%D0%BF%D0%BB%D0%BE%D0%BC%D0%B0%D1%82%D0%BE%D0%B2%20%3A%3A%20%D0%9E%D
1%82%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5%20%D0%A1%D0%BA%D1%
80%D0%B8%D0%BF%D0%B0%D0%BB%D1%8F     Mozilla/5.0 (Linux; Android 7.0;
```

```
MI 5 Build/NRD90M) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/65.0.3325.109 Mobile Safari/537.36
```

1.3. Требования к реализации

Условия:

- Решение должно быть реализовано на Spark Structured Streaming.
- Ваше решение должно печатать в STDOUT топ-10 самых популярных (по просмотрам) доменов с информацией об общем числе просмотров этого домена и числа уников, которые на него зашли.
- Результат - это кумулятивная статистика за всё время работы Streaming, отсортированная по убыванию числа просмотров.
- Результат должен выводиться в консоль каждые 5 секунд
    - Если ваш код не успевает уложиться в этот интервал - возможно, проблема в избыточном числе партиций
    - Важно выводить таблицу целиком и не обрезать длину столбцов (опция truncate должна быть выключена)

Пример результата:

```
-------------------------------------------
Batch: 10
-------------------------------------------
+---------------------------+----+------+
|domain                     |view|unique|
+---------------------------+----+------+
|news.rambler.ru            |18  |15    |
|m.lenta.ru                 |9   |7     |
|yandex.ru                  |9   |8     |
|www.championat.com         |7   |7     |
|www.yaplakal.com           |7   |7     |
|www.mk.ru                  |7   |7     |
|www.gazeta.ru              |6   |6     |
|www.coins-spb.ru           |6   |1     |
|miss-tramell.livejournal.com|6   |6     |
|woman.rambler.ru           |6   |5     |
+---------------------------+----+------+
```

- Решение должно предоставлять CLI интерфейс со следующими параметрами:
    - Общие настройки (должны быть заполнены все)
        - --topic-name - имя топика
        - --starting-offsets - отступ, с которого скрипт начинает работать

        - --kafka-brokers - координаты брокеров Kafka
    - Настройка триггера (должен быть заполнен один из двух)
        - --processing-time - микробатчевый триггер по времени (запускает триггер с заданной настройкой)
        - --once - триггер, который запустит вычисление датасета только один раз
    - Пример запуска решения

```
...runet_stat.py --topic-name page_views --starting-offsets
latest --processing-time "5 second" --kafka-brokers
brain-node1.bigdatateam.org:9092,brain-node2.bigdatateam.org:9092
,brain-node3.bigdatateam.org:9092
```

- Пример кода решения для инициализации нужных параметров:

```
import argparse
parser = argparse.ArgumentParser()
parser.add_argument("--kafka-brokers", required=True)
parser.add_argument("--topic-name", required=True)
parser.add_argument("--starting-offsets", default='latest')

group = parser.add_mutually_exclusive_group()
group.add_argument("--processing-time", default='0 seconds')
group.add_argument("--once", action='store_true')

args = parser.parse_args()
if args.once:
    args.processing_time = None
else:
    args.once=None
...
   .trigger(once=args.once, processingTime=args.processing_time)
\
...
```

- PySpark-скрипты для запуска решений следует называть task_<Surname>_<Name>_<#task_ID.**suffix**>.py:
    - решение задачи #1 должно называться "task_*_domain_stat.py" и запускаться с помощью команды:

```
PYSPARK_DRIVER_PYTHON=python3.6 PYSPARK_PYTHON=python3.6
spark-submit --packages
```

```
    org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.0
    task_*_domain_stat.py %cli_args%
```

- скрипты выводят на экран (STDOUT) указанное в задании число строк в нужном формате каждый батч. Если вы запускаете spark streaming с триггером по времени или по умолчанию, то первый батч будет пустым и это нормально. Запуск решения при проверке будет запускаться с тригером once и если вы все написали правильно - то первый батч пустым не будет;
- вывод STDOUT задач с результатом обработки 4 батчей нужно сохранить в соответствующих файлах в архиве посылки домашнего задания (например, `task_*_suffix.out`).[2]
    - формат файла произвольный
    - этот файл не влияет на успешность сдачи грейдеру

1.4. Задача #2 (Task ID: realtime.runet_stat): оконная статистика посещения рунета

В этом домашнем задании вам предстоит определить видимый трафик в зоне ru и в остальном интернете. Сравнение производится на окне размером в 2 секунды каждую секунду (нас в обоих случаях интересует время события (поле TS из лога), а не обработки). Для трафика требуется подсчитать характеристики: число просмотров и число уников.

1.5. Входные данные

Входные данные для задания описаны в пункте 1.2.

1.6. Требования к реализации

Условия:

- Решение должно быть реализовано на Spark Structured Streaming.
- Ваше решение должно печатать в STDOUT агрегированную статистику для сайтов зоны RU и остальных.

[2] Для подготовки архива с решением и выводом результатов запуска можно воспользоваться командой “tee”

- Статистика - это число просмотров и число уников, которые в определенный интервал зашли на искомую группу доменов.
- Статистика рассчитывается за две секунды лога каждую секунду (под временем здесь подразумевается именно время события)
- Результат - это кумулятивная статистика за всё время работы Streaming, отсортированная по времени окна и убыванию числа просмотров в каждом окне.
- Решение должно выводить в консоль только первые 20 результатов работы
- Результат выводится в консоль по мере готовности:
    - Важно выводить таблицу целиком и не обрезать длину столбцов (опция truncate должна быть выключена)

Пример результата:

```
-------------------------------------------
Batch: 6
-------------------------------------------
+------------------------------------------+------+----+------+
|window                                    |zone  |view|unique|
+------------------------------------------+------+----+------+
|[2018-04-01 16:20:43, 2018-04-01 16:20:45]|ru    |719 |683   |
|[2018-04-01 16:20:43, 2018-04-01 16:20:45]|not ru|242 |255   |
|[2018-04-01 16:20:44, 2018-04-01 16:20:46]|ru    |719 |702   |
|[2018-04-01 16:20:44, 2018-04-01 16:20:46]|not ru|259 |255   |
...
|[2018-04-01 16:20:49, 2018-04-01 16:20:51]|ru    |717 |668   |
|[2018-04-01 16:20:49, 2018-04-01 16:20:51]|not ru|265 |257   |
+------------------------------------------+------+----+------+
```

- Решение должно предоставлять CLI интерфейс аналогичный тому, который был описан в пункте 1.3.
- PySpark-скрипты для запуска решений следует называть `task_<Surname>_<Name>_<#task_ID>.suffix.py`:
    - решение задачи #2 должно называться "`task_*_runet_stat.py`" и запускаться с помощью команды:

```
PYSPARK_DRIVER_PYTHON=python3.6 PYSPARK_PYTHON=python3.6
spark-submit --packages
org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.0
task_*_runet_stat.py %cli_args%
```

2. Рекомендации

При решении задач старайтесь следовать следующим рекомендациям:

- следите за качеством кода и проверяйте “глупые” ошибки с помощью pylint, следите за поддерживаемостью и читаемостью кода;
- очень часто бывает полезно использовать spark.sql. Не забывайте что streaming dataframe можно преобразовать в view и обрабатывать с помощью sql, а результат всё так же будет являться streaming df (см. пример в лекции).

3. Критерии оценивания

Веса задач:

1. 50% - Задача #1 (Task ID: realtime.domain_stat): статистика посещения доменов
2. 50% - Задача #2 (Task ID: realtime.runet_stat): оконная статистика посещения рунета

Балл за задачу складывается из:

- **80%** - 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 invalid-name,missing-docstring --ignored-modules=pyspark.sql.functions`
        - качество кода должно оцениваться выше 8.0 / 10.0
        - проверяем код **Python версии 3** с помощью `pylint==2.5.3`
- **0%** - эффективность решения (такие как потребляемые CPU-ресурсы, скорость выполнения (в предположении свободного кластера)).

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 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.

4. Инструкция по отправке задания

Перед отправкой задания оставьте, пожалуйста, отзыв о домашнем задании по ссылке: https://rebrand.ly/bdmade2022q2_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): **HW08:RealTime**.
- Выполненное ДЗ запакуйте в архив **BD_MADE_2022_Q2**_<Surname>_<Name>_HW#**.zip**, пример -- **BD_MADE_2022_Q2**_Dral_Alexey_**HW08**.zip. (Проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда.[3]). Рекомендуем формирование архива для грейдера производить на машине кластера. Это исключить сразу несколько возможных ошибок[4]. Если вы файлы решения размещены в папке hw, то для формирование архива сначала перейдите в эту папку:
  `cd hw`
  а затем сформируйте архив следующей командой:
  `zip BD_MADE_2022_Q2_<Surname>_<Name>_HW08.zip *`
- (***для рисковых***[5]) 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.
- Решения заданий должны содержаться в одной папке.
- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
  - **| BD_MADE_2022_Q2_<Surname>_<Name>_HW08.zip**
  - **| ---- task_<Surname>_<Name>_domain_stat.py**
  - **| ---- task_<Surname>_<Name>_domain_stat.out**
  - **| ---- task_<Surname>_<Name>_runet_stat.py**
  - **| ---- task_<Surname>_<Name>_runet_stat.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:

[3] Онлайн инструмент для проверки: https://www.soscisurvey.de/tools/view-chars.php
[4] Студенты, сталкивались со следующими проблемами: изменение символа перевода строки при переносе файлов обратно на локальную машину, попадание в архив скрытых файлов при архивировании в Mac OS / Windows, несоответствие версий (на кластере ошибку поправили, а скачать на локальную машину последнюю версию забыли и в грейдер закачали предыдущую)
[5] Ответственность за сборку архива и его валидность берете на себя

- Зарегистрироваться и залогиниться в сервисе Everest
- Перейти на страницу приложения: MADE BigData Grader
- Выбрать вкладку Submit Job (если отображается иная).
- Выбрать в качестве “Task” значение: **HW08:RealTime**[6]
- Загрузить в качестве “Task solution” файл с решением
- В качестве Access Token указать тот, который был выслан по почте

Любые вопросы / комментарии / предложения можно писать в телеграм-канал курса или на почту bd_made2022q2@bigdatateam.org .

Всем удачи!

[6] Сервисный ID: realtime.onsite_hw

5. FAQ (часто задаваемые вопросы)

"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 помощи: в случае неправильно указанного Access Token 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: “**HW08:RealTime**. 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.

Что в отчете Grader означает проверка X ?
Как читать отчет: Для каждого теста • Raw_score - балл за конкретный тест. Может быть как бинарным (1\0), так и находиться в интервале от 0 до 1 • Score - Raw_score*weight (вес теста в общей оценке). Вес указан для каждого теста ниже

Итоговая оценка: смотрите строку Score (сумма Score всех индивидуальных тестов) внизу отчета.

Правильность решения задачи:
test_unzip_is_succesful (weight = 0) - ДЗ заархивировано в .zip архив и грейдер может его разархивировать.

test_pyspark_execution_successful (weight = 0.1) - успешность выполнения скрипта

test_exact_run_stdout_comparison (weight = 0.7) - корректность вывода (stdout)

Поддерживаемость и читаемость кода:
test_py_files_min_lint_score (weight = 0.2) - качество кода в .py файлах оценивается выше 8.0