BIGDATA TEAM

HW #05: Data Layout

Deadline: 22.07.2019, 08:00

---

1. Описание задания.

В данном ДЗ нужно решить **5 задач**. Решение надо выполнить с помощью Hive. Задачи общие для всех.

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

Веса задач:

1. 20%
2. 20%
3. 20%
4. 20%
5. 20%

3. Описание данных

3.1. Логи запросов пользователей новостных сайтов.

user_logs:

- Путь на кластере: полный датасет - **`/data/user_logs/user_logs_M`**
- Семпл (для тестирования): **`/data/user_logs/user_logs_S`**
- Формат: текст
- В каждой строке находятся следующие поля, разделенные знаком табуляции (**иногда не одним**):
    1. STRING - ip-адрес, с которого пришел запрос,
    2. STRING (TIMESTAMP) - время запроса,
    3. STRING - пришедший с ip-адреса http-запрос,
    4. INT - размер переданной клиенту страницы в байтах,
    5. INT - http-статус запроса.

6. STRING - User Agent, информация о клиентском приложении, с которого осуществлялся запрос на сервер, в том числе, информация о браузере.

Пример:

```
135.124.143.193              20150601013300
http://newsru.com/4712386    235   412   Firefox/5.0 (compatible; MSIE
9.0; Windows NT 6.1; Win64; x64; Trident/5.0)n
```

Важно:

- разделитель между IP и временем запроса состоит из 3 символов табуляции;
- Будем считать, что браузере содержится в начале 6-ого поля лога - символы с нулевой позиции до позиции первого пробельного символа.
  - пример User Agent:
  - `Chrome/5.0 (compatible; MSIE 9.0; Windows NT 8.0; WOW64; Trident/5.0; .NET CLR 2.7.40781; .NET4.0E; en-SG)`
  - тогда браузером будет: `Chrome/5.0`

Подсказка:

- поскольку нас не интересует оставшаяся часть User Agent, то получить тип браузера пользователя можно с помощью правильного регулярного выражения в период чтения logs_raw.

3.2. Информация о пользователях.

user_data:

- Путь на кластере: полный датасет - **/data/user_logs/user_data_M**
- Семпл (для тестирования): **/data/user_logs/user_data_S**
- Формат: текст
- В каждой строке находятся следующие поля, разделенные знаком табуляции:
  1. STRING - IP-адрес, с которого пользователь выходит в интернет;
  2. STRING - браузер пользователя;
  3. STRING - пол (male / female);
  4. INT - возраст.

Пример:

```
197.72.248.141    Opera/12.0  male  30
```

3.3. Геобаза - информация о соответствии ip-адресов регионам.

ip_data:

- Путь на кластере: полный датасет - **/data/user_logs/ip_data_M**
- Семпл (для тестирования): **/data/user_logs/ip_data_S**

- Формат: текст
- В каждой строке находятся следующие поля, разделенные знаком табуляции:
    1. STRING - IP-адрес;
    2. STRING - регион.

Пример:

```
33.49.147.163     Kemerovo Oblast
197.72.248.141    Belgorod Oblast
135.124.143.193   Krasnoyarsk Krai
...
```

4. Задачи.

В рамках решения ДЗ по Hive, у вас появилась партиционированная таблица с логами пользователей новостных сайтов `user_logs`, а также персональные данные по пользователям `user_data`.

Вам предлагается решить следующие задачи (отработатать задачи на семплах _S, _M и получить решение или оценки роста производительности для полного датасета).

ДЗ по Hive. Задача #3 (вариант 2): identify browser sex.

Напишите запрос, который считает число употреблений браузера мужчинами и женщинами. Группируем браузеры из таблицы **logs**. Выведите **произвольные** 10 записей (LIMIT 10) в формате:

- браузер <tab> посещаемость мужчинами <tab> посещаемость женщинами

Пример вывода:

```
Firefox/5.0 1419872 621124
Opera/5.0 1426114 623333
...
```

Задачи.

Задача 1. Переложить данные logs в таблицу logs_optimized, где будет использоваться формат хранения данных ORC. С помощью параметров TBLPROPERTIES найдите оптимальный набор параметров, чтобы получить максимальное сжатие данных. Какой оптимизации пространства удалось добиться?

Задача 2. Проверьте скорость выполнения простых аналитических запросов на основе таблицы logs и logs_optimized. Какая оптимизация по скорости выполнения получена? Сделайте релевантные таблицы для датасетов _S, _M и _full и сравните наблюдения.

Задача 3. Для того, чтобы оптимизировать скорость выполнения запроса для решения задачи “identify browser sex” предлагается поиграться с параметрами TBLPROPERTIES, чтобы оптимизировать не только сжатие, но и скорость выполнения самих запросов. Произведите релевантные исследования (рекомендуется использоваться Managed таблицы и перезаписывать logs_ с помощью INSERT OVERWRITE запроса). Какая оптимизация по скорости получена? Какие параметры были выбраны?

Задача 4. Попробуйте добавить бакетирование и сортировку данных. Какая оптимизация по скорости выполнения запроса получена? Для сравнения: сколько времени тратится на переукладку данных?

Задача 5. Попробуйте заменить в логах информацию про браузер таким образом, чтобы 90% данных содержало одинаковый браузер (или браузер “unknown”), запишите результат в таблицу logs_broken. Попробуйте посчитать запрос в задаче “identify browser sex”. Оцените время на выполнение запроса. Для того, чтобы пофиксить проблему:

1. В реальной жизни рекомендуется сделать запрос в формате TABLESAMPLE (и увидеть по каким параметрам происходит перекос)
2. По результатам 5.1 вы знаете по каким данным происходит перекос, дайте эту информацию в формате SKEWED TABLE для Hive

Оцените скорость выполнения запроса в этом случае (на заметку: не забывайте отслеживать параметр числа редьюсеров, если их недостаточно для выполнения запроса).

5. Disclaimer.

Это экспериментальное домашнее задание. Просьба первым экспериментатором присылать вопросы и рекомендации в телеграм-чат, чтобы преподаватели оперативно реагировали, корректировали формулировки заданий и содержание задач. Это в свою очередь позволит скорректировать нагрузку на выполнение ДЗ, чтобы было реалистично выполнить в указанные сроки на решение ДЗ.

6. Сроки сдачи и правила оформления задания.

Deadline: 22.07.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW5:DataLayout**.
- Решения задач должны содержаться в одной папке.
- По результатам решения ожидается отчет в формате PDF с описанием результатов оптимизации (ответов на поставленные исследовательские вопросы).
- HQL-скрипты для запуска решений следует называть по номеру задачи и варианта **task_<#task>_<#variant>.hql**:
  - например решение задачи #2 для 3го варианта должно называться task_2_3.hql и его можно запустить с помощью команды:
    - `hive -f task_2_3.hql`
  - скрипт выводит на экран (STDOUT) указанное в задание число строк в нужном формате
  - Вывод STDOUT нужно писать в соответствующие файлы, например task_2_3.out.
- Выполненное ДЗ запакуйте в архив **MF2019Q2_<фамилия>_HW#.zip**, к примеру -- **MF2019Q2_Ivanov_HW3.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: “**HW5:DataLayout**. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по сссылке: http://rebrand.ly/mf2019q2_feedback_hw05. 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

Всем удачи!

7. Дорешка.

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

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