

HW #10: Sqoop

Deadline: 26.08.2019, 08:00

1. Описание задания.	1
2. Критерии оценивания.	1
3. Описание данных.	2
4. Условие первой задачи: загрузка таблицы person_details из MySQL в HIVE	2
5. Условие второй задачи: загрузка подмножества строк на основе уникального SQL запроса из MySQL в HIVE	2
5. Сроки сдачи и правила оформления задания.	3
6. Дорешка.	4

1. Описание задания.

В данном ДЗ нужно решить **2 задачи**. Решение надо выполнить с помощью Sqoop (import/export), Hive, MySQL.

2. Критерии оценивания.

Балл за задачу складывается из:

- **30%** - правильное решение первой задачи
- **70%** - правильное решение второй задачи

Штрафы:

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

3. Описание данных.

- Описание: имеется MySQL база moon_db
- В базе созданы две таблицы: person и person_details
- Логин и пароль от базы sqoop

4. Условие первой задачи: загрузка таблицы person_details из MySQL в HIVE

В этом домашнем задании вам предстоит загрузить таблицу person_details из MySQL в HIVE.

Условия:

- Исходная таблица находится в базе moon_db (moon_db.person_details).
- Решение должно быть в виде sqoop import скрипта.
- Результат — это таблица в базе HIVE (Название базы sqoop_db). Название таблицы person_details_<фio пользователя>. Например, person_details_gsi.

5. Условие второй задачи: загрузка подмножества строк на основе уникального SQL запроса из MySQL в HIVE

В этом домашнем задании вам предстоит загрузить результат запроса к таблицам: person и person_details из MySQL в HIVE.

Условия:

- Исходные таблицы находятся в базе moon_db (moon_db.person и moon_db.person_details).
- Sql запрос:

```
select p.names, count(*) as cards_count from person p  
left join person_details pd on pd.person_id = p.id
```



```
group by p.id, p.names  
order by cards_count DESC
```

- Решение должно быть в виде sqoop import скрипта.
- Результат — это таблица в базе HIVE (Название базы sqoop_db). Название таблицы person_sql_query_<фio пользователя>. Например, person_sql_query_gsi.

5. Сроки сдачи и правила оформления задания.

Deadline: 26.08.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW10:Sqoop**.
- Решение задачи (для каждой задачи) – скрипта sqoop import (сохранить в файл **sqoop_import_<task_#>.sh**) и результат вывода команды Hive “show create table sqoop_db.<название вашей таблицы>” (сохранить в файл **show_create_table_<task_#>.out**) .
- Выполненное ДЗ запакуйте в архив MF2019Q2_<фамилия>_HW#.zip , например -- MF2019Q2_Ivanov_HW10.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 команду¹:
 - 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: “**HW10:Sqoop**. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по ссылке: http://rebrand.ly/mf2019q2_feedback_10_sqoop. 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

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



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!