

## HW #04: Hive

---

1. Описание задания	<b>2</b>
2. Критерии оценивания	<b>2</b>
3. Описание данных	<b>2</b>
4. Задача #1: создание таблиц в Hive (Task ID: hive.ddl)	<b>4</b>
5. Задача #2: горячий денек (Task ID: hive.hot_day)	<b>5</b>
6. Задача #3: identify browser sex (Task ID: hive.sex_browser)	<b>6</b>
7. Правила оформления задания	<b>7</b>

---

автор задания: BigData Team, коллективная работа.



## 1. Описание задания

В данном ДЗ нужно решить **3 задачи**. Решение надо выполнить с помощью Hive.

## 2. Критерии оценивания

Веса задач:

1. 33.3%
2. 33.3%
3. 33.3%

Балл за задачу складывается из:

- **70%** - правильное решение задачи
- **0%** - поддерживаемость и читаемость кода
  - в общем случае см. Clean Code и [Google Python Style Guide](#)
- **30%** - 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 deadline
- **5%** за каждую дополнительную посылку в тестирующую систему (одна дополнительная посылка бесплатно)

Формула подсчета финальной оценки<sup>1</sup>:

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

## 3. Описание данных

### 3.1. Логи запросов пользователей новостных сайтов

logs\_raw:

- 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):

---

<sup>1</sup> результат умножается на 10 (максимальная оценка) и округляется до первой цифры после точки



1. ip STRING - ip-адрес, с которого пришел запрос,
2. date STRING - время запроса,
3. request STRING - пришедший с ip-адреса http-запрос,
4. page\_size INT - размер переданной клиенту страницы в байтах,
5. http\_status INT - http-статус запроса.
6. user\_agent 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 информация о 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

users:

- 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. ip STRING - IP-адрес, с которого пользователь выходит в интернет;
  2. browser STRING - браузер пользователя;
  3. sex STRING - пол (male / female);
  4. age INT - возраст.

Пример:



197.72.248.141    Opera/12.0   male   30

### 3.3. Геобазы - информация о соответствии ip-адресов регионам

ip\_regions:

- 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. ip STRING - IP-адрес;
  2. region STRING - регион.

Пример:

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

## 4. Задача #1: создание таблиц в Hive (Task ID: hive.ddl)

Создайте внешние (EXTERNAL) таблицы по исходным данным:

1. **logs\_raw** - логи пользователей;
2. **users** - таблица с информацией о пользователях;
3. **ip\_regions** - таблица с IP и регионами;

Из таблицы логов перенесите данные в другую таблицу, партиционированную по датам - одна партиция на каждый день:

4. **logs** - партиционированная таблица с логами.

Условия:

1. Таблицы и поля должны называться ровно так, как указано в описании задачи. Например поле для хранения даты (дня) в таблице **logs** оставить таким же, как и в **logs\_raw**:
  - ``date` STRING`;<sup>2</sup>
2. Сериализация и десериализация данных для таблицы **logs\_raw** должна осуществляться с использованием регулярных выражений, см.:
  - `org.apache.hadoop.hive.serde2.RegexSerDe`

---

<sup>2</sup> Обратите внимание на экранирование ключевых слов Hive

Проверить правильность создания таблиц можно с помощью простых SELECT-запросов:

```
SELECT * FROM <table> LIMIT 10
```

Рекомендации:

- предлагается начать с простых таблиц, а потом двигаться к сложным, например: `ip_regions` → `users` → `logs_raw` → `logs`;
- для создания таблиц `ip_regions` и `users` рекомендуется воспользоваться следующей конструкцией:
  - ROW FORMAT delimited
  - Документация по полям, разделяющим колонки, доступна по [адресу](#). Вам необходимо найти способ указать разделить `<tab>` вместо стандартного разделителя `^A`.

Подсказки по созданию партиционированной таблицы `logs`:

1. Чтобы выделить день в формате "YYYYMMDD", достаточно воспользоваться функцией для работы со строками SUBSTR.
2. Посчитайте, сколько уникальных (DISTINCT) дней в "сырых" логах (`logs_raw`). Это число должно получиться более 100 на датасете размера "\_M".
3. Используйте это число, чтобы задать переменную окружения Hive, которая позволит запустить динамическое создание партиций<sup>3</sup>:
  - `set hive.exec.max.dynamic.partitions.pernode=***;`
4. После этого можно написать запрос:
  - `INSERT OVERWRITE TABLE logs PARTITION(date) SELECT ... FROM logs_raw`

**На партиционированной таблице ``logs`` и нужно будет выполнять запросы в следующих задачах.**

## 5. Задача #2: горячий денек (Task ID: `hive.hot_day`)

Напишите запрос, который считает какое количество посещений новостных сайтов было в разрезе дней. Полученные результаты отсортируйте (`ORDER BY`) по убыванию популярности. На экран выведите TOP-10 самых "горячих" дней с точки зрения нагрузки на инфраструктуру новостных сервисов в формате:

---

<sup>3</sup> Подробную документацию по dynamic partitioning см. здесь:

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-DynamicPartitionInserts>



- день <tab> число посещений

*Пример вывода:*

```
20140308 96
20140409 94
20140318 89
...
```

Для этого задания таблица `logs` будет предоставлена, поэтому если вы используете названия колонок, которые не соответствуют схеме из раздела 3. Описание данных, то Grader не пропустит решение.

## 6. Задача #3: identify browser sex (Task ID: hive.sex\_browser)

Напишите запрос, который считает число употреблений браузера мужчинами и женщинами. Считаем статистику по таблице `logs`, но информацию о браузере берем из таблицы `users`<sup>4</sup>. Выведите **произвольные** 10 записей (LIMIT 10) в формате:

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

*Пример вывода:*

```
MsExplorer/11.0 1419872 621124
Chrome 1426114 623333
...
```

Подсказки:

- для решения задачи рекомендуется воспользоваться оператором IF, примеры его использования см. в официальной документации Hive (см. [здесь](#)) или в слайдах занятия.
- для решения этой задачи нужно сделать join двух таблиц. Сложность заключается в том, что: по умолчанию, из-за небольшого объема данных Hive преобразует этот запрос в Map-Side Join, НО у него **может** не хватить оперативной памяти, чтобы выполнить эту задачу, поэтому:
  1. Нужно отключить авто-конвертацию join в оптимизированный вид join. см. опцию:
    - `set hive.auto.convert.join`

---

<sup>4</sup> Да, мы согласны, что это глупое предположение и в реальной жизни информацию о браузере нужно брать из User Agent логов. Но мы еще придумаем задачи получше ;)



2. Из-за небольшого объема данных, Hive может запустить все вычисления в рамках Reduce-Side Join на одном редьюсере. Чтобы этого избежать, необходимо изменить число редьюсеров с помощью флага:
  - `set mapreduce.job.reduces`

**Для этого задания таблица logs будет предоставлена, поэтому если вы используете названия колонок, которые не соответствуют схеме из раздела 3. Описание данных, то Grader не пропустит решение.**

## 7. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW4:Hive**.
- Выполненное ДЗ запакуйте в архив `MADEBD2021Q1_<Surname>_<Name>_HW#.zip`, пример -- `MADEBD2021Q1_Dral_Alexey_HW4.zip`. (Проверяйте отсутствие пробелов и невидимых символов после копирования имени отсюда.<sup>5</sup>) Если ваше решение лежит в папке `my_solution_folder`, то для создания архива `hw.zip` на Linux и Mac OS выполните команду<sup>6</sup>:
  - `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.
- Название базы данных будет передаваться через CLI с помощью аргумента `--database=$(db_name)`, для локальных экспериментов рекомендуется использовать `<username>`, например `dral`; ваши скрипты **не должны** содержать использование директивы `"use"`.
- HQL-скрипты для запуска решений следует называть по суффиксу Task ID задачи `task_<Surname>_<Name>_<#task_ID_suffix>.hql`:
  - например решение задачи `"hive.hot_day"` должно называться `task_<Surname>_<Name>_hot_day.hql` и его можно запустить с помощью команды:
    - `hive --database=${DB_NAME}7 -f task_*_hot_day.hql`
  - скрипт выводит на экран (STDOUT) указанное в задании число строк в нужном формате
- Вывод STDOUT задач просьба сохранить в соответствующих файлах в архиве посылке домашнего задания (например, `task_<Surname>_<Name>_<#task_ID>.out`).

<sup>5</sup> Онлайн инструмент для проверки: <https://www.soscisurvey.de/tools/view-chars.php>

<sup>6</sup> Флаг -r значит, что будет совершен рекурсивный обход по структуре директории

<sup>7</sup> Это означает, что Вы не должны использовать `"use <database_name>"` внутри скриптов



- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
  - | MADEBD2021Q1\_<Surname>\_<Name>\_HW4.zip
  - | ---- task\_<Surname>\_<Name>\_ddl.hql
  - | ---- task\_<Surname>\_<Name>\_hot\_day.hql
  - | ---- task\_<Surname>\_<Name>\_hot\_day.out
  - | ---- task\_<Surname>\_<Name>\_sex\_browser.hql
  - | ---- task\_<Surname>\_<Name>\_sex\_browser.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:
  - 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: **HW4:Hive<sup>8</sup>**
  - 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: **"HW4:Hive. 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](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.

---

<sup>8</sup> Сервисный ID: hive.onsite\_hw





# BIGDATA TEAM

Любые вопросы / комментарии / предложения можно писать в [Discord-канал курса](#) или на почту [bigdata\\_made2021q1@bigdatateam.org](mailto:bigdata_made2021q1@bigdatateam.org).

Всем удачи!