

## Д3-2. Hive

Номер варианта определяется по формуле: `int(logname[3:]) % 3 + 1`, где `logname` - это аккаунт на кластере.

Срок сдачи задания - **10.12.17** (включительно).

**Исходные данные:** Логи пользователей.

Данные находятся в HDFS по адресу `/data/user_logs/*_M`. Они состоят из трёх частей, каждая из которых находится в своей поддиректории. Данные в каждой части отличаются количеством и типом колонок, разделенных знаками табуляции (`\t`) или пробелами.

*A. Логи запросов пользователей к новостным сообщениям (user\_logs).*

1. IP-адрес, с которого пришел запрос (STRING),
2. Время запроса (TIMESTAMP),
3. Пришедший с IP-адреса HTTP-запрос (STRING),
4. Размер переданной клиенту страницы (INT),
5. HTTP-статус код (INT).
6. Информация о клиентском приложении, с которого осуществлялся запрос на сервер, в том числе, информация о браузере (STRING).

**Важно:** информация о браузере содержится в начале 6-ого поля лога (символы с нулевой позиции до позиции первого пробельного символа), содержание оставшейся части строки не определяет браузер пользователя. Разделитель между IP и временем запроса имеет 3 табуляции.

*B. Информация о пользователях (user\_data).*

1. IP-адрес (STRING),
2. Браузер пользователя (STRING),
3. Пол (STRING) //male, female,
4. Возраст (INT).

*C. Информация о местонахождении IP адресов пользователей (ip\_data).*

1. IP-адрес (STRING),
2. Регион (STRING).

**Задача 1 (411).** Создайте внешние (EXTERNAL) таблицы по исходным данным. В результате будет 3 таблицы: логи пользователей, данные IP-адресов, данные пользователей. Из таблицы логов перенесите данные в другую таблицу, партиционированную по датам — одна партиция на каждый день. На партиционированных таблицах и нужно будет выполнять запросы в следующих задачах. Требуется, чтобы сериализация и десериализация данных осуществлялась с использованием регулярных выражений

(см. `org.apache.hadoop.hive.contrib.serde2.RegexSerDe`, `org.apache.hadoop.hive.serde2.RegexSerDe`).

Проверить правильность создания таблиц с помощью простейших запросов (`SELECT * FROM <table> LIMIT 10`).

**Задача 2.1. (421).** Напишите запрос, выбирающий количество посещений для каждого дня. Полученные результаты отсортируйте по убыванию.

*Пример результата:*

20140308	96
20140409	96
20140318	96

**Задача 2.2. (422).** Напишите запрос, выбирающий количество различных HTTP-кодов для каждого дня. Полученные результаты отсортируйте по убыванию.

*Пример результата:*

20140207	46
20140126	46

20140112	46
----------	----

**Задача 2.3. (423).** Напишите запрос, выбирающий количество посещений для каждого типа браузера. Браузеры берём из таблицы логов. Если 2 браузера отличаются версиями, считаем их различными. Полученные результаты отсортируйте по убыванию.

*Пример результата:*

Firefox/5.0	25
Opera/5.0	21

**Задача 3.1. (441).** Напишите запрос, выбирающий количество посещений от мужчин и от женщин по регионам.

*Пример результата:*

Tver	66968157 29097223
Voronezh	60445347 26333509

**Задача 3.2. (442).** Напишите запрос, выбирающий количество посещений от мужчин и от женщин по типам браузера (информацию о браузерах берём из таблицы логов).

*Пример результата:*

Firefox/5.0	1419872	621124
Opera/5.0	1426114	623333

**Задача 3.3. (443).** Напишите запрос, выбирающий количество посещений от мужчин и от женщин по кодам HTTP ответов.

Подсказка: используйте конструкцию `IF`.

*Пример результата:*

511	90675090 39459549
412	87782696 38146030

*Задачи 4 и 5 рекомендуется выполнять с помощью Hive Streaming.*

**Задача 4.1. (471).** Представьте ситуацию, что все новостные сайты переехали в домен `.com`. Вас попросили обновить базу логов, чтоб логи пользователей указывали не на старые домены, а на новые. Например, новостная ссылка `http://news.rambler.ru/8744806` теперь должна выглядеть в ваших запросах как `http://news.rambler.com/8744806`. Используйте стриминг в `hive-sql` запросе. (Рекомендуется обратить внимание на команды `awk` и `sed`). Выведите TOP-10 записей логов без сортировки.

*Пример результата:*

49.203.96.67	20140102	http://lenta.com/2296722	716	499	Safari/5.0
33.49.147.163	20140102	http://news.yandex.com/5605690	850	300	Safari/5.0
0					

**Задача 4.2. (472).** Аналогично, только заменяем `http` на `ftp`.

*Пример результата:*

222.131.187.37	20140101	ftp://news.mail.ru/8805842	1017	416	Opera/5.0
197.72.248.141	20140101	ftp://news.rambler.ru/2816512	2042	428	Safari/5.0
0					

**Задача 4.3. (473).** Аналогично, только заменяем Safari на Chrome в столбце браузеров.

Пример результата:

222.131.187.37	20140101	http://news.mail.ru/8805842	1017	416	Opera/5.0
197.72.248.141	20140101	http://news.rambler.ru/2816512	2042	428	Chrome/5.0

**Задача 5.1 (461).** Выведите TOP-10 IP-адресов с октетами в обратном порядке. (Например, 192.168.0.125 должно превратиться в 125.0.168.192 ).

Пример результата:

166.40.208.75  
197.93.255.168

**Задача 5.2. (462).** Переведите размер HTML страницы в таблице логов в мегабайты (он дан в килобайтах). Поведение запроса в аварийной ситуации (если на вход подано не число) - на ваше усмотрение (NULL, 0, пустая строка,...). Выведите TOP-10 записей.

Пример результата:

1  
0

**Задача 5.3. (463).** Посчитайте, сколько лет осталось до 100 каждому пользователю из таблицы пользователей. Поведение запроса в аварийной ситуации (если на вход подано не число) - на ваше усмотрение (NULL, 0, пустая строка,...). Выведите TOP-10 записей.

Пример результата:

73  
35

**Комментарий.** В некоторых задачах требуется вывести TOP-N записей. Это означает просто использование `LIMIT N`. Сортировать результат запроса при этом не обязательно. Там, где сортировка необходима (задача 2), это указано отдельно.

#### Литература

1. [План семинара](#) по Hive.
2. Tom White. Hadoop: The Definitive Guide, 4rd edition. O'Reilly, 2015, глава 17.
3. Edward Capriolo, Dean Wampler, Jason Rutherglen. Programming Hive. O'Reilly, 2012.