Домашнее задание по Hive

Исходные данные:

I. 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 или пробелами.

A. Логи запросов пользователей к новостным сообщениям (user_logs).

1. Ip-адрес, с которого пришел запрос (STRING),
2. Время запроса (TIMESTAMP или INT),
3. Пришедший с ip-адреса http-запрос (STRING),
4. Размер переданной клиенту страницы (SMALLINT),
5. Http-статус код (SMALLINT).
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 (TINYINT).

C. Информация о местонахождении IP адресов пользователей (ip_data).

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

II. Подсети

Данные находятся по адресу /data/subnets. В директории 3 датасета (/data/subnets/variant[1-3], выберите соответствующий для своего варианта), но все они имеют одинаковый формат.

1. IP-адрес (STRING),
2. Маска подсети (STRING).

Датасеты в каждой директории отличаются 1-м полем. * 1-й вариант. В качестве 1-го поля дан адрес сети. * 2-й вариант. Адрес произвольного хоста в сети. * 3-й вариант. Широковещательный адрес (broadcast).

Семплы находятся в `/data/subnets_S`. Если в полных данных 5000 записей, то в семплах всего 20.

Задачи

Задача 1 (411). Создайте внешние (EXTERNAL) таблицы по исходным данным. В результате будет 4 таблицы: логи пользователей, данные ip адресов, данные пользователей и подсети. Из таблицы логов перенесите данные в другую таблицу, партицированную 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`). Эти Select запросы нужно также добавлять в скрипт задачи.

Пример результата:

```
33.49.147.163   http://lenta.ru/4303000 1189    451 Chrome/5.0 (compatible; MSIE 9.0; Windows
75.208.40.166   http://newsru.com/3330815   60  306 Safari/5.0 (Windows; U; MSIE 9.0; Windows
```

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

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 количества.

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 количества.

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.

Пример результата:

```
511 90675090    39459549
412 87782696    38146030
```

Задача 4.1 (461). Создать UDF «перевертыш». Функция принимает строку и возвращает данную строку записанную в обратном порядке. Подключите данную функцию к hive и выполните выборку «перевертышей» для ip адресов. Вывести TOP-10 записей.

Пример результата:

```
661.04.802.57
791.39.552.861
661.04.802.57
```

Задача 4.2 (462). Создать UDF “мегабайт”. Функция принимает на вход число в виде строки и возвращает его же, деленное на 1024. Подключите данную функцию к hive и переведите размер HTML страницы в таблице логов в мегабайты (он дан в килобайтах). Поведение функции в аварийной ситуации (если на вход подано не число) - на ваше усмотрение (NULL, 0, пустая строка,. . .). Вывести TOP-10 записей.

Пример результата:

```
0
0
1
```

Задача 4.3 (463). Создать UDF “мотиватор”. Функция принимает на вход число в виде строки и возвращает 100 - . Подключите данную функцию к hive и посчитайте, сколько лет осталось до 100 участникам логов. Поведение функции в аварийной ситуации (если на вход подано не число) - на ваше усмотрение (NULL, 0, пустая строка,. . .). Вывести TOP-10 записей.

Пример результата:

```
35
78
77
```

Задача 5.1. (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
```

Задача 5.2. (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
```

Задача 5.3. (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
```

Задачи 5.x должны быть решены с использованием Hive Streaming.