# Hive’16:

1 - aggregate (2)

2 - aggregate + limit / sort / having (3)

3 - distribution / group by (4, 2)

4 - subquery / join (4)

(4) - join

(5) - udf

(6) - streaming

# 

# Basic Level

**Задача 1. (Одинаковая для всех вариантов)** Создайте внешние (EXTERNAL) таблицы по исходным данным. В результате будет три таблицы: логи пользователей, данные ip адресов, данные пользователей. 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.** Напишите запрос, выбирающий количество хитов для каждого дня. Полученные результаты отсортировать.

**Задача 2.2.** Напишите запрос, выбирающий количество различных HTTP-кодов для каждого дня. Полученные результаты отсортировать.

**Задача 2.3.** Напишите запрос, выбирающий количество хитов для каждого типа браузера. Браузеры берём из таблицы логов[[1]](#footnote-2). Если 2 браузера отличаются версиями, считаем их различными.Полученные результаты отсортировать.

**Задача 3.1.** Выберите регионы, в которых количество хитов *выше* среднего количества хитов по всем регионам.

**Задача 3.2.** Выведите возраст пользователей, у которых кол-во хитов *ниже*, чем среднее по всем значениям возраста.

**Задача 3.3.** Выберите возраст пользователей, у которых кол-во хитов *выше* среднего количества хитов по всем регионам.

**Задача 4.1.** Напишите запрос, выбирающий количество хитов от мужчин и от женщин по регионам.

**Задача 4.2.** Напишите запрос, выбирающий количество хитов от мужчин и от женщин по типам браузера (информацию о браузерах берём из таблицы логов).

**Задача 4.3.** Напишите запрос, выбирающий количество хитов от мужчин и от женщин по кодам HTTP ответов.

**Задача 5.1.** 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.

**Задача 5.2.** 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,...).

**Задача 5.3.** Создать UDF “мотиватор”. Функция принимает на вход число в виде строки и возвращает 100 - <это число>. 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,...).

**Задача 6.1.** Представьте ситуацию, что все новостные сайты переехали в домен .com. Вас попросили создать представление, в котором логи пользователей будут указывать 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. Для создания такого представления используйте стриминг в hive-sql запросе. (Рекомендуется обратить внимание на команды в awk и sed).

**Задача 6.2.** Аналогично, только заменяем http на ftp.

**Задача 6.3.** Аналогично, только заменяем Safari на Chrome в столбце браузеров.

# Исходные данные

hadoop fs -ls /data/user\_logs – описать, что там.

# Литература

1. Глава 12 из 3-го издания Definitive guide.

2. Книга по Hive.

1. Это важно, т.к. у нас в логах и пользователях разные браузеры. [↑](#footnote-ref-2)