Домашнее задание по Hive

Исходные данные

Исходные данные находятся в HDFS по адресу `/data/user_logs/`. 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.

1. Логи запросов пользователей к новостным сообщениям (user_logs).
    1. Ip-адрес, с которого пришел запрос (INT),
    2. Время запроса (TIMESTAMP),
    3. Пришедший с ip-адреса http-запрос (INT),
    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.

2. Информация о пользователях (user_data).
    1. IP-адрес (STRING),
    2. Браузер пользователя (STRING),
    3. Пол (STRING) //male, female,
    4. Возраст (INT).
3. Информация о местонахождении IP адресов пользователей (ip_data).
    1. IP-адрес (STRING),
    2. Регион (STRING).

Для каждой поддиректории созданы соответствующие директории с семплами, они имеют суффикс “_S”.

Создание таблиц

Как было рассказано на лекциях в hive существуют два типа таблиц managed и external. По умолчанию создается managed таблица. Если вы создаете управляемую таблицу, значит вы можете удалять и модифицировать прямо те данные, по которым была создана таблица, и это означает, что вам нужны права на запись в директорию с данными. Прав на изменение исходных данных у вас нет, поэтому вам необходимо создавать external таблицы.

Литература

Tom White. Hadoop: The Definitive Guide, 3rd edition. O’Reilly, 2012, глава 12.

Edward Capriolo, Dean Wampler, Jason Rutherglen. Programming Hive. O’Railly, 2012.

(обе книги можно найти в разделе Resources данного курса).