BIGDATA TEAM

#05: Hive Optimization. Workshop.

1. Цель занятия. 1

2. Общие рекомендации. 1

3. Оптимизация аналитический запросов. 1

4. Обратная связь 3

1. Цель занятия.

1. Научиться оптимизироваться аналитические запросы с помощью правильного выбора Data Layout.

DISCLAIMER: Не надо беспокоиться, если у Вы что-либо не успели. Всегда остается возможность продолжить погружение дома и иметь возможность спрашивать вопросы в Telegram-канале.

2. Общие рекомендации.

Чтобы пробросить порты 8088, 50070 воспользуйтесь инструкцией из User Guides.

Для удобства необходимые материалы для копирования лежат в папке:

`/home/aadral/public_examples/hive`

3. Оптимизация аналитический запросов.

В рамках решения ДЗ по Hive, у вас появилась партиционированная таблица с логами пользователей новостных сайтов `user_logs`, а также персональные данные по пользователям `user_data`.

Вам предлагается решить следующие задачи (отработатать задачи на семплах _S, _M и получить решение или оценки роста производительности для полного датасета).

ДЗ по Hive. Задача #3 (вариант 2): identify browser sex.

Напишите запрос, который считает число употреблений браузера мужчинами и женщинами. Группируем браузеры из таблицы **`logs`**. Выведите **произвольные** 10 записей (LIMIT 10) в формате:

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

Пример вывода:

```
Firefox/5.0 1419872 621124
Opera/5.0 1426114 623333
...
```

Задачи для практического занятия.

Задача 1. Переложить данные logs в таблицу logs_optimized, где будет использоваться формат хранения данных ORC. С помощью параметров TBLPROPERTIES найдите оптимальный набор параметров, чтобы получить максимальное сжатие данных. Какой оптимизации пространства удалось добиться?

Задача 2. Проверьте скорость выполнения простых аналитических запросов на основе таблицы logs и logs_optimized. Какая оптимизация по скорости выполнения получена? Сделайте релевантные таблицы для датасетов _S, _M и _full и сравните наблюдения.

Задача 3. Для того, чтобы оптимизировать скорость выполнения запроса для решения задачи “identify browser sex” предлагается поиграться с параметрами TBLPROPERTIES, чтобы оптимизировать не только сжатие, но и скорость выполнения самих запросов. Произведите релевантные исследования (рекомендуется использоваться Managed таблицы и перезаписывать logs_ с помощью INSERT OVERWRITE запроса). Какая оптимизация по скорости получена? Какие параметры были выбраны?

***Задача 4**. Попробуйте добавить бакетирование и сортировку данных. Какая оптимизация по скорости выполнения запроса получена? Для сравнения: сколько времени тратится на переукладку данных?

***Задача 5**. Попробуйте заменить в логах информацию про браузер таким образом, чтобы 90% данных содержало одинаковый браузер (или браузер “unknown”), запишите результат в таблицу logs_broken. Попробуйте посчитать запрос в задаче “identify browser sex”. Оцените время на выполнение запроса. Для того, чтобы пофиксить проблему:

1. В реальной жизни рекомендуется сделать запрос в формате TABLESAMPLE (и увидеть по каким параметрам происходит перекос)
2. По результатам 5.1 вы знаете по каким данным происходит перекос, дайте эту информацию в формате SKEWED TABLE для Hive

Оцените скорость выполнения запроса в этом случае (на заметку: не забывайте отслеживать параметр числа редьюсеров, если их недостаточно для выполнения запроса).

4. Обратная связь

Обратная связь: http://rebrand.ly/mf2019q2_feedback_05_datalayout

Просьба потратить 1-2 минут Вашего времени, чтобы поделиться впечатлением, описать что было понятно, а что непонятно. Мы учитываем рекомендации и имеем возможность переформатируем учебную программу под Ваши запросы.