Описание

Для выполнения лабораторная работы №1 и №2 необходимо установить https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html

Лабораторная работа №1 (Map/Reduce)

Задание

1. Напишите MR задачу:
   - Считает средние количество и общее количество байтов на запрос по IP

Пример строки входной строки:
ip13 - - [24/Apr/2011:04:41:53 -0400] "GET /logs/access_log.3 HTTP/1.1" 200 4846545 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

   - Необходимо использовать комбайнер (Combiner).
   - Вывод должен быть файл CSV со строками следующим образом:

   IP, 175.5 (***среднее количество байт***),109854 (***сумма***)

2. Добавьте юнит-тесты MR для вашей Map/Reduce задачи
3. Сохраните вывод в виде файла последовательности, сжатого с помощью Snappy (ключ - IP, а значение - пользовательский объект для avg и общего размера)
4. Используйте счетчики, чтобы получить статистику, сколько пользователей IE, Mozzila или других было обнаружено и распечатайте их в STDOUT Driver.

Используйте стороннюю библиотеку для анализа UserAgent, например:
   - https://github.com/HaraldWalker/user-agent-utils

Исходные данные

https://drive.google.com/file/d/1911EoiA2nqJDefJB97vVng4cG4T18mbJ/view?usp=sharing

Ссылки

- https://habr.com/ru/post/103467/
- https://habr.com/ru/post/103490/
- https://habr.com/ru/company/intersystems/blog/310180/
- https://habr.com/ru/company/dca/blog/268277/
- http://shop.oreilly.com/product/0636920025122.do

Отчет

- Ссылку на Github репозиторий (Никакого кода в отчете не должно быть !)
- Скриншоты успешно завершенной джобы.
- Скриншоты счетчиков.
- А также скриншоты пройденных тестов.

Лабораторная работа № 2 (Hive)

Задание

Напишите запрос (запросы) на HQL:

- Создает таблицу ***Logs*** и загружает в нее данные из файла.
- Считает среднее и сумму байтов для каждого IP адреса.
- Результат предыдущего запроса записывает в новую таблицу (***Statistic***)

Исходные данные

https://drive.google.com/file/d/1911EoiA2nqJDefJB97vVng4cG4T18mbJ/view?usp=sharing

Ссылки

- https://habr.com/ru/post/283212/
- https://habr.com/ru/company/dca/blog/305838/
- https://habr.com/ru/post/223217/
- https://www.youtube.com/watch?v=qC_GbpPu1aU

Отчет

- Ссылку на GitHub с исходными HQL скриптами
- Скриншот успешно завершенного запроса
- Скриншот таблицы с исходными данными

Лабораторная работа № 3 (Spark)

Задание

Необходимо реализовать Spak джобу которая:

- Считает средние количество и общее количество байтов на запрос по IP
- Вывод должен быть файл CSV со строками следующим образом:
  IP, 175.5 (***среднее количество байт***),109854 (***сумма***)
- Добавить юнит тесты.
- Используйте счетчики, чтобы получить статистику, сколько пользователей IE, Mozzila или других было обнаружено и распечатайте их в STDOUT Driver.
- Используйте стороннюю библиотеку для анализа UserAgent, например:
- - https://github.com/HaraldWalker/user-agent-utils
- Весь процесс запуска должен быть автоматизирован, по запуску одной команды должны пройти тесты, запуститься джоба и открыть выходной файл.

Исходные данные

https://drive.google.com/file/d/1911EoiA2nqJDefJB97vVng4cG4T18mbJ/view?usp=sharing

Ссылки

- https://habr.com/ru/company/piter/blog/276675/
- https://habr.com/ru/company/mlclass/blog/250811/
- https://habr.com/ru/post/329838/
- https://habr.com/ru/company/epam_systems/blog/336090/
- https://habr.com/ru/company/jugru/blog/325070/
- https://habr.com/ru/post/330986/

Отчет

- Ссылка на Github с исходным кодом
- Скриншоты на пройденные юнит тесты
- Счетчики браузеров