FUNBOX

Квалификационные задания для разработчиков BigData

О заданиях

Результат присылайте нам вместе с рассказом о себе на почтовый ящик wanted@fun-box.ru с пометкой «Вакансия разработчика BigData». После проверки заданий, мы обязательно сообщим вам о нашем решении.

В общем, это всё. Перейдя на следующую страницу, вы увидите наши вопросы и задания.

Имеется

- набор **tsv** файлов в **HDFS**, сгруппированных по датам, в которых лежат факты;
- посещения некоторыми пользователями различных интернет-ресурсов с использованием мобильных устройств:
    - шаблон пути к файлам: `/files/yyyyMMdd/part*.tsv`, где `yyyyMMdd` - собственно формат даты
    - в файлах набор записей из двух полей:
        - `user_key` - строка фиксированной длины - идентификатор пользователя;
        - `resource` - адрес интернет-ресурса, например: www.google.com, woman.ru, www.bbc.co.uk и т.п.
    - пары ключ-значение в пределах одного файла могут быть неуникальны.
- таблица в **Cassandra**, содержащая информацию об операционной системе девайса конкретного пользователя:
    - схема: `CREATE TABLE my_keyspace.table1 (user_key TEXT PRIMARY KEY, device_os TEXT);`
    - `user_key` - идентификатор пользователя;
    - `device_os` - операционная система мобильного устройства, например: `android, ios, windows phone` и т.п.;
- объем tsv файлов - 100+ Гб;
- количество записей в таблице в **Cassandra** - до 100 млн.

Задача

Разработать приложение, которое будет составлять перечень пользователей, посетивших заданный интернет-ресурс с телефона указанной операционной системы в заданный диапазон дат с подсчетом этих посещений, причем:

- конкретный интернет-ресурс может быть задан в произвольном виде, например: https://www.google.com, Google или google.com;
- операционная система устройства также может быть задана в произвольном виде: iOs, ios, Android и т.п.;

Окружение

- Hadoop 2.7.x
- Cassandra 3.x

Требования

- разработать собственно MapReduce задачу фильтрации и подсчета количества;
- покрыть код unit тестами;
- разработать простейший набор интеграционных тестов, выполняемых в Docker, и подготовить небольшой набор данных для него для демонстрации работоспособности приложения.

Спасибо за время, потраченное на выполнение заданий!