BIGDATA TEAM

#03: MapReduce Optimization. Workshop.

1. Цель занятия.

Научиться оптимизировать MapReduce Streaming приложения с помощью Combiner, а также пользоваться Partitioner и Comparator.

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 копирования, исходные файлы лежат в папках:

```
/home/aadral/public_examples/map_reduce
/home/aadral/public_examples/map_reduce_word_count
```

3. MapReduce Python Streaming, `-files`, тестирование.

См. слайды 8-14 для написания задачи “Line Count” полностью на Python.

Предлагается действовать в 2 этапа:

1. Написать Python-скрипты и протестировать их локально.
2. Обновить run.sh, чтобы запустить полностью Python Streaming MR-приложение.

Эмуляция работы Hadoop локально достигается следующим образом:

- Скачиваем часть данных себе локально из HDFS (см. **`/data/wiki/en_articles_part`**) под названием **`sample.txt`**:
  ```
  hdfs dfs -cat /data/wiki/en_articles_part/* | head -n 50 > sample.txt
  ```
- Тестируем MapReduce-скрипты:
  ```
  cat sample.txt | python3 mapper.py | sort | python3 reducer.py > out.txt
  ```

Таким образом, мы можем оперативно проверить работоспособность скриптов до запуска в распределенном режиме.

4. Задача Word Count на Python. (см. слайды 17-33)

Исходный код базового WordCount возьмите по адресу:

`/home/aadral/public_examples/map_reduce_word_count`

Задача 4.1. Написать приложение WordCount, не учитывающее мусор (знаки пунктуации).

Задача 4.2. Написать приложение WordCount, приводящее все слова к нижнему регистру.

5. Distributed Cache. (см. слайды 35-46)

1. Найдите какую-нибудь статистику употребления имен / слов в интернете.
2. Сохраните статистику как минимум в два файла (например: male.txt и female.txt).
3. Сделайте из них tar-архив.
4. Посчитайте WordCount на основе (семпла) Википедии (**`/data/wiki/en_articles_part`**) с помощью Distributed Cache.
5. Сравните полученные статистики.
6. Напишите в телеграм-чат что нашли интересного.

6. Обратная связь

Обратная связь: http://rebrand.ly/mf2019q2_feedback_03_mro

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.