BIGDATA TEAM

#02: Hadoop MapReduce. Workshop.

1. Цель занятия.

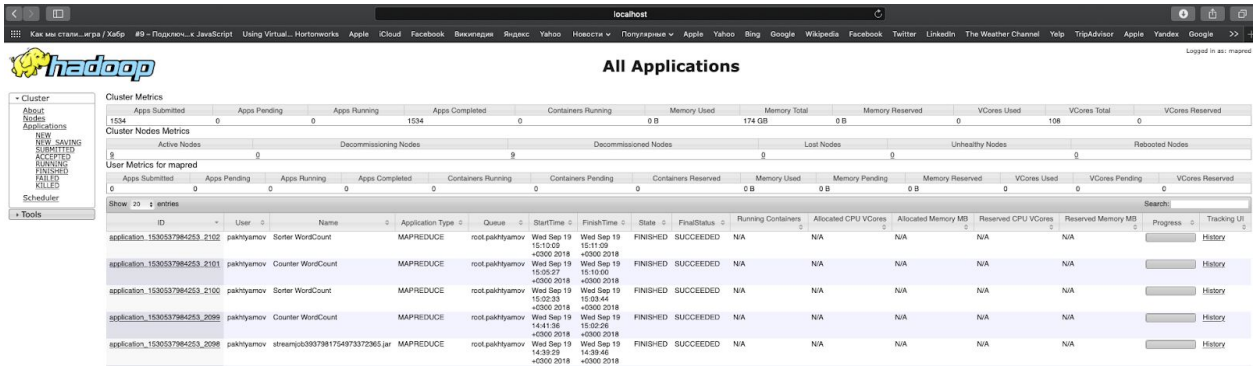
Научиться запускать MapReduce Streaming приложения.

Не надо беспокоиться, если у Вы что-либо не успели. Всегда остается возможность продолжить погружение дома и иметь возможность спрашивать вопросы в Telegram-канале. **Для отслеживания прогресса в сравнении с остальными членами группы, мы будем пользоваться “Poll” в Telegram.**

2. Проброс портов к ResourceManager (RM).

Следуйте инструкции из User Guides, только в дополнение к порту 50070 аналогичным образом добавьте порт 8088.

После этого в браузере введите localhost:8088. Если у вас появилось следующее изображение, то все хорошо:



3. Запуск первого MapReduce Streaming приложения. (см. слайд 50)

Для удобства копирования, исходные файлы лежат в папке:

/home/aadral/public_examples/map_reduce

run.sh выглядит следующим образом[1]:

```
#/usr/bin/env bash
set -x

HADOOP_STREAMING_JAR=/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/had
oop-streaming.jar
OUT_DIR=my_hdfs_output
hdfs dfs -rm -r $OUT_DIR

yarn jar $HADOOP_STREAMING_JAR \
    -mapper "wc -l" \
    -numReduceTasks 0 \
    -input /data/wiki/en_articles_part \
    -output $OUT_DIR

echo $?
```

3.1. Задание #1 + FAQ.

Запустите приложение и посмотрите вывод в HDFS:

[1] Для подсветки синтаксиса удобно пользоваться https://tohtml.com/

1. Сколько файлов на выходе?
2. Что находится внутри файлов?

FAQ

Q1: как отслеживать статус выполнения задачи?
A1:
Статус выполнения задачи и логи можно отслеживать через Resource Manager:
http://localhost:8088/cluster

Если зашли без проброса порта 8088 - перезайдите ещё раз (см. Раздел 2).

1. Если задача еще не завершила выполнение, то будет доступна ссылка на ApplicationMaster.
2. Если задача уже завершилась, то будет доступна ссылка History.

Для того чтобы перейти по ссылке ApplicationMaster или History, необходимо:

1. Скопировать ссылку, пример:
   - http://virtual-master.bigdatateam.ru:8088/proxy/application_1538682956624_0865/
2. Заменить virtual-master.bigdatateam.ru на localhost, пример:
   - http://localhost:8088/proxy/application_1538682956624_0865/
3. Перейти по полученной ссылке из п. 2

Q2: как удобно отличать свои задачи от других?
A2:
Чтоб отличать свою задачу от других, удобно присвоить ей имя. Для этого в нужно добавить еще один флаг (**одним из самых первых флагов**):

```
-D mapreduce.job.name="surname: my first line_count"
```

3.2. Задание #2.

Нужно обновить приложение (run.sh), чтобы оно считало число строк во всем датасете. Для этого необходимо воспользоваться reducer.sh (находится в этой же папке).

reducer.sh

```
#!/usr/bin/env bash
awk '{line_count += $1} END { print line_count }'
```

Для редактирования файлов можете использовать любимую опцию:

1. Использовать редактор vim или nano в терминале на virtual-client;
2. Использовать любимый редактор на ноутбуке и копировать файлы на кластер с помощью SCP (или PSCP на Windows);
3. Воспользоваться "jupyter notebook --port 30XX" (XX - ваш ID, дополненный лидирующим нулем в случае необходимости) и пробросом портов, чтобы загружать и обновлять файлы в привычном интерфейсе Jupyter (http://jupyter.org/). [Ваш ID указан рядом с вашей фамилией в столбце A в таблице с оценками.]
   Для доступа нужно добавить порт, но если раньше было “virtual-master:50070” и “virtual-master:8088”, то теперь надо пробросить порт 30XX на “localhost:30XX” аналогичным образом.
4. Если придумали свою опцию - поделитесь этим знанием с другими.

Вопрос: Чтобы использовать reducer.sh в run.sh, какие 3 поля нужно было обновить?

4. Обратная связь

Обратная связь: http://rebrand.ly/mf2019q2_feedback_02_mr

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.