

#02: Hadoop MapReduce. Workshop.

1. Цель занятия	2
2. Проброс портов к ResourceManager (RM)	2
3. Запуск первого MapReduce Streaming приложения	3
3.1. Задание #1 + FAQ	4
3.2. Задание #2	6
4. Обратная связь	6

1. Цель занятия

Научиться запускать MapReduce Streaming приложения.

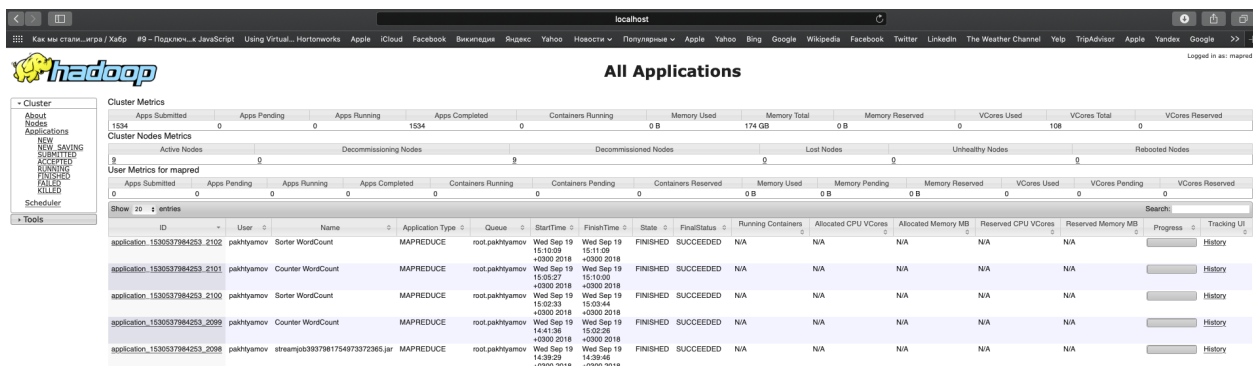
Не надо беспокоиться, если у Вас что-либо не успели. Всегда остается возможность продолжить погружение дома и иметь возможность спрашивать вопросы в Telegram-канале.

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 (ResourceManager), 19888 (Job History Server)

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



ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	Progress	Tracking UI
application_1530537984253_2102	pakhtyanov	Sorter WordCount	MAPREDUCE	root.pakhtyanov	Wed Sep 19 15:10:09 +0300 2018	Wed Sep 19 15:11:09 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1530537984253_2101	pakhtyanov	Sorter WordCount	MAPREDUCE	root.pakhtyanov	Wed Sep 19 15:05:27 +0300 2018	Wed Sep 19 15:10:00 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1530537984253_2100	pakhtyanov	Sorter WordCount	MAPREDUCE	root.pakhtyanov	Wed Sep 19 15:02:33 +0300 2018	Wed Sep 19 15:03:44 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1530537984253_2099	pakhtyanov	Sorter WordCount	MAPREDUCE	root.pakhtyanov	Wed Sep 19 14:41:38 +0300 2018	Wed Sep 19 15:02:28 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1530537984253_2098	pakhtyanov	streamjob3937981754973372365.jar	MAPREDUCE	root.pakhtyanov	Wed Sep 19 14:39:29 +0300 2018	Wed Sep 19 14:39:46 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History

3. Запуск первого MapReduce Streaming приложения

Для удобства копирования, исходные файлы лежат в папке:

`/usr/local/share/big_data_course/public_examples/map_reduce`

run.sh выглядит следующим образом¹:

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

HADOOP_STREAMING_JAR=/usr/local/hadoop/share/hadoop/tools/lib/hadoop-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 $?
```

¹ Просьба не копировать из PDF, поскольку могут скопировать невидимые человеческому глазу юникодные символы, которые будут мешать запустить скрипт. Копируйте файлы в свою домашнюю папку на edge node из указанных локаций.



3.1. Задание #1 + FAQ

Запустите приложение и посмотрите вывод в HDFS:

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

FAQ

Как отслеживать статус выполнения задачи?

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 (**активного приложения**) 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. В случае если приложение уже неактивно, то оно будет доступно только в JobHistory Server и нужно будет повторить операцию из п.2 несколько раз (рекомендация для ускорения - воспользоваться curl из консоли)

Как удобно отличать свои задачи от других?

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"
```

Как убить задачу?



Убить задачу с помощью CTRL+C не выйдет, поскольку задание запущено на кластере. Находите интересующее Вас приложение (**application_ID**), можно даже не свое (но просьба не злоупотреблять ;)). После этого в консоли пишем:

```
yarn application -kill <application_ID>
```



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 edge node (обычно в hostname можно увидеть "client");
2. Использовать любимый редактор на ноутбуке и копировать файлы на кластер с помощью SCP (или PSCP на Windows);
3. Воспользоваться "jupyter notebook --port **port_1**" и пробросом портов, чтобы загружать и обновлять файлы в привычном интерфейсе [Jupyter](#).
Для доступа (может быть³) нужно добавить порт, но если раньше было "brain-master:50070" и "brain-master:8088", то теперь надо пробросить порт **port_1** на "localhost:**port_1**" 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/x5bd2021q1_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.

² См. раздел "Соответствие логина и портов" в User Guides

³ Зависит от строгости Hadoop администратора кластера в текущий момент времени ;)