

# #02: Hadoop MapReduce. Workshop.

<ol> <li>Цель занятия.</li> <li>Проброс портов к ResourceManager (RM).</li> </ol>							
							3. Запуск первого MapReduce Streaming приложения. (см. слайд 50)
3. Задания.							
3.1. Задания для beginner.	3						
4. Обратная связь							

### 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 <u>User Guides</u>, 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:



							loca	alhost			(ئ				0	d l	
мы сталиигра	ı / Хабр — #9 – Подключк JavaSci	ript Using V	irtual Hortonworks Apple iClo	ud Facebook Bu	кипедия Яндекі	: Yahoo I	Новости 🗸 П	опулярные ч	Apple Yaho	o Bing Google	Wikipedia Facebook	Twitter LinkedIn 1	he Weather Channel Y	elp TripAdvisor App	le Yandex G	Roogle :	
hed	اووم						All	Appl	ication	s					Lo	iggad in as: m	
r	Cluster Metrics																
ations	Apps Submitted	Apps Pen	ding Apps Bunning	Apps Cor		Contains	ers Running		Memory Used	Memory Tota		Reserved	VCores Used	VCores Total	VCores F	Reserved	
ions	1534 0		0	1534	.0			0.8		174 GB	0 B	0	16	08	0		
NEW SAVING SUBMITTED ACCEPTED	Cluster Nodes Metrics																
	Active Nodes		Decommissionin	ig Nodes			Decommis	ssioned Nodes			Lost Nodes	Unhe	althy Nodes		Rebooted Nodes	sooted Nodes	
	User Metrics for mapred	0			9					0		0		0			
		Pending	Apps Running Apps Com	pleted Cont	tainers Running	Contain	ners Pending	Conta	iners Reserved	Memory Used 0 B	Memory Pending 0 B	Memory Rese	rved VCores Use	d VCores Pendir	g VCor	res Reserved	
	Show 20 ; entries														Search:		
	ID *	User 0	Name 0	Application Type ©	Queue 0	StartTime 0	FinishTime 0	State 0	FinalStatus 0	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress C	Tracking	
	application 1530537984253 2102	pakhtyamov	Sorter WordCount	MAPREDUCE	root pakhtyamov	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	pakhtyamov	Counter WordCount	MAPREDUCE	root.pakhtyamov	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	pakhtyamov	Sorter WordCount	MAPREDUCE	root.pakhtyamov	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	pakhtyamov	Counter WordCount	MAPREDUCE	root pakhtyamov	Wed Sep 19 14:41:36 +0300 2018	Wed Sep 19 15:02:26 +0300 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History	
			streamiob3937981754973372365.jar	MARRENINE	root.pakhtyamov	Wed Sep 19	Wed Sep 19	EINIGHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	_	History	

## 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<sup>1</sup>:

```
#/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 -1" \
    -numReduceTasks 0 \
    -input /data/wiki/en_articles_part \
    -output $OUT_DIR

echo $?
```

# 3.1. Задание #1 + FAQ.

Запустите приложение и посмотрите вывод в HDFS:

<sup>&</sup>lt;sup>1</sup> Для подсветки синтаксиса удобно пользоваться <a href="https://tohtml.com/">https://tohtml.com/</a>



- 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 (<a href="http://jupyter.org/">http://jupyter.org/</a>). [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: <a href="http://rebrand.ly/mf2019q2\_feedback\_02\_mr">http://rebrand.ly/mf2019q2\_feedback\_02\_mr</a>

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.