

Hadoop Fundamentals: Demo

Семинар курса «Управление разно-структурированными большими данными»

<http://synthesis.ipi.ac.ru/synthesis/student/BigData/seminar-hadoop/hadoop2014>

alexey.vovchenko@gmail.com

Начало работы

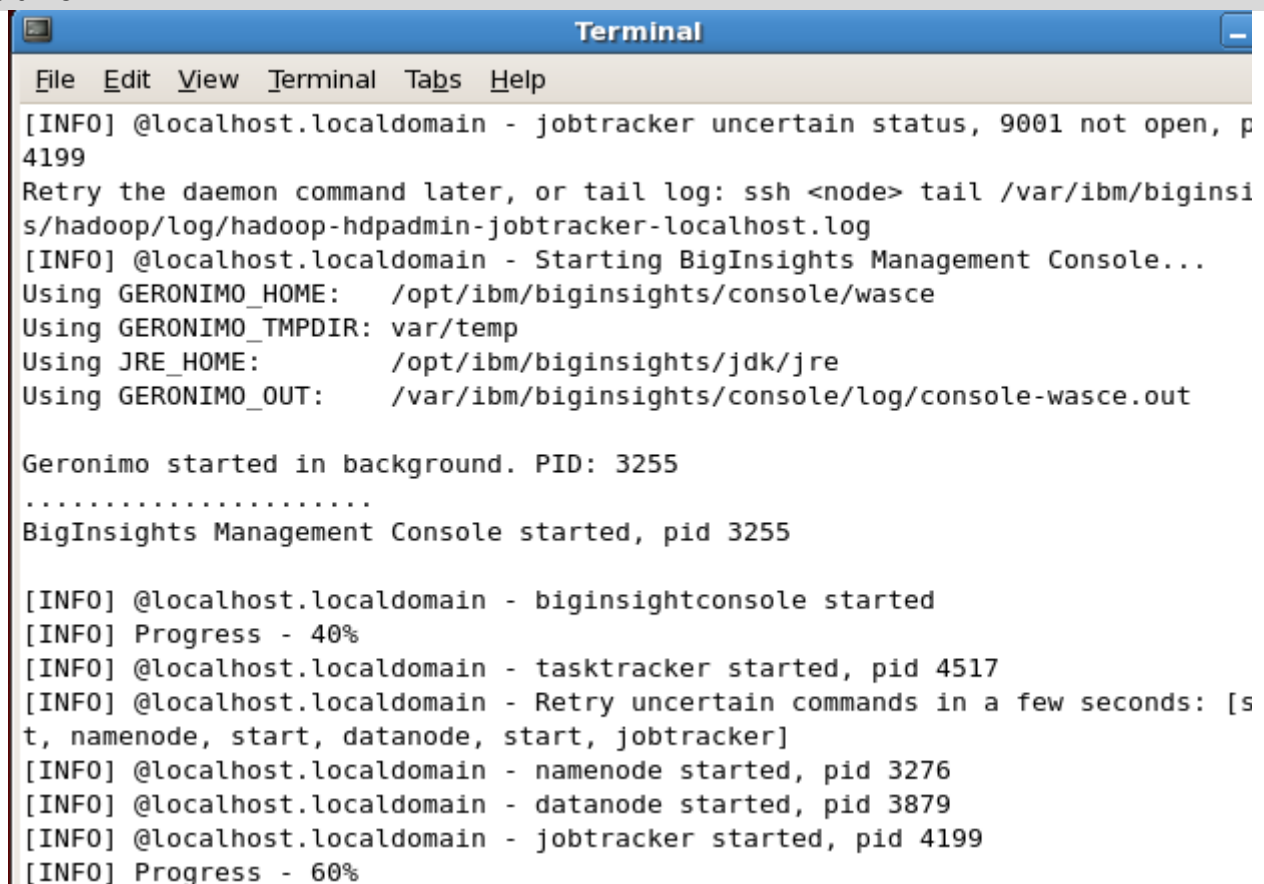
Для доступа к серверу IBM BigInsights необходимо выполнить инструкции из: «server_access.pdf». Если у Вас нет этого файла, необходимо написать мне на почту.

Стартуем Hadoop

Этот этап нужно делать только тем, кто работает на собственных виртуальных машинах. Тем, кто работает на удаленной виртуальной машине это делать не нужно, т.к. там Hadoop уже запущен.

Иконка на рабочем столе, или из shell:

start-all.sh



```
Terminal
File Edit View Terminal Tabs Help
[INFO] @localhost.localdomain - jobtracker uncertain status, 9001 not open, p
4199
Retry the daemon command later, or tail log: ssh <node> tail /var/ibm/biginsi
s/hadoop/log/hadoop-hdpadmin-jobtracker-localhost.log
[INFO] @localhost.localdomain - Starting BigInsights Management Console...
Using GERONIMO_HOME: /opt/ibm/biginsights/console/wasce
Using GERONIMO_TMPDIR: var/temp
Using JRE_HOME: /opt/ibm/biginsights/jdk/jre
Using GERONIMO_OUT: /var/ibm/biginsights/console/log/console-wasce.out

Geronimo started in background. PID: 3255
.....
BigInsights Management Console started, pid 3255

[INFO] @localhost.localdomain - biginsightconsole started
[INFO] Progress - 40%
[INFO] @localhost.localdomain - tasktracker started, pid 4517
[INFO] @localhost.localdomain - Retry uncertain commands in a few seconds: [s
t, namenode, start, datanode, start, jobtracker]
[INFO] @localhost.localdomain - namenode started, pid 3276
[INFO] @localhost.localdomain - datanode started, pid 3879
[INFO] @localhost.localdomain - jobtracker started, pid 4199
[INFO] Progress - 60%
```

Веб интерфейс Hadoop

При работе на собственной виртуальной машине

- <http://bivm:50070/> – web UI for HDFS name node(s)
- <http://bivm:50030/> – web UI for MapReduce job tracker(s)
- <http://bivm:50060/> – web UI for task tracker(s)

При работе на удаленном сервере

- <http://83.149.245.126:50070/> – web UI for HDFS name node(s)
- <http://83.149.245.126:50030/> – web UI for MapReduce job tracker(s)
- <http://83.149.245.126:50060/> – web UI for task tracker(s)

Рассмотрим web UI for HDFS name node(s)

- <http://bivm:50070/> или <http://83.149.245.126:50070/>

На этой странице мы видим информацию о Hadoop кластере.

NameNode 'localhost.localdomain:

Started: Wed Feb 02 14:46:09 EST 2011
Version: 0.20.2, r911707
Compiled: Fri Feb 19 08:07:34 UTC 2010 by chrisdo
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)
[Namenode Logs](#)

Cluster Summary

20 files and directories, 5 blocks = 25 total. Heap Size is 5.7 MB / 10

Configured Capacity	:	9.68 GB
DFS Used	:	14.9 MB
Non DFS Used	:	5.5 GB
DFS Remaining	:	4.16 GB
DFS Used%	:	0.15 %
DFS Remaining%	:	43 %
Live Nodes	:	1
Dead Nodes	:	0

Перейдем по ссылке «Live Nodes» - функционирующие узлы кластера.

NameNode 'localhost.localdomain:9000'

Started: Wed Feb 02 14:46:09 EST 2011
Version: 0.20.2, r911707
Compiled: Fri Feb 19 08:07:34 UTC 2010 by chrisdo
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)

[Namenode Logs](#)

[Go back to DFS home](#)

Live Datanodes : 1

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)
localhost	0	In Service	9.68	0.01	5.5	4.17	0.15	

Hadoop 2011

Здесь мы видим узлы, из которых состоит кластер, перейдет теперь по ссылке «Browse the filesystem». Если вы работаете на удаленном сервере, то может возникнуть ошибка.

Для ее исправления можно либо переписать URL руками (заменив bivm.ibm.com на 83.149.245.126)

<http://83.149.245.126:50075/browseDirectory.jsp?namenodeInfoPort=50070&dir=/&nnadr=83.149.245.126:9000>

Либо для Windows можно прописать данное имя в список известных имен. Для этого надо открыть файл:

```
"C:\Windows\System32\drivers\etc\hosts"
```

И добавить туда строку вида:

```
83.149.245.126      bivm.ibm.com
```

Пример содержимого файла:

```
...  
# For example:  
#  
# 102.54.94.97  rhino.acme.com    # source server  
# 38.25.63.10  x.acme.com          # x client host  
  
# localhost name resolution is handled within DNS itself.  
# 127.0.0.1    localhost  
# ::1         localhost  
83.149.245.126      bivm.ibm.com
```

Для добавления данного адреса можно также использовать скрипт `add_hadoop_uri.cmd`, запустив его от имени администратора

```
..\labsData\add_hadoop_uri.cmd
```

После того как ошибка устранена, мы должны увидеть следующее изображение.

Contents of directory /

Goto :

Name	Type	Size	Replication	Block Size	Modification Time	Permis
M2	dir				2011-01-21 13:47	rwxr-x
hadoop	dir				2011-01-21 13:39	rwxr-x
user	dir				2010-12-25 21:55	rwxr-x

[Go back to DFS home](#)

Local logs

[Log directory](#)

[Hadoop](#), 2011.

Здесь отображается древовидная структура файловой системы Hadoop Distributed File System (HDFS). Таким образом можно просматривать все данные в файловой системе.

Рассмотрим [web UI for MapReduce job tracker\(s\)](#)

- <http://bivm:50030/> или <http://83.149.245.126:50030/>

Здесь будут отображаться запускаемые Map-Reduce приложения

localhost Hadoop Map/Reduce Administration - Mo

File Edit View History Bookmarks Tools Help

http://localhost:50030/jobtracker.jsp

Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support

localhost Hadoop Map/Reduce Adr

State: RUNNING
Started: Wed Feb 16 17:53:09 EST 2011
Version: 0.20.2, r911707
Compiled: Fri Feb 19 08:07:34 UTC 2010 by chrisdo
Identifier: 201102161753

Cluster Summary (Heap Size is 7 MB/1000 MB)

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity
0	0	2	<u>1</u>	2	2

Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Completed Jobs

Рассмотрим web UI for task tracker(s)

- <http://bivm:50060/>
- <http://83.149.245.126:50060/>
- Напрямую данная страница редко используется, однако на нее можно перейти из информации о Map-Reduce приложении
- Например, по этой ссылке, лог выполнения одной из Map задач
- http://83.149.245.126:50060/tasklog?attemptid=attempt_201410130843_0002_m_000000_0&all=true

tracker_localhost.localdomain:local Task Tracker Status



Version: 0.20.2, r911707
Compiled: Fri Feb 19 08:07:34 UTC 2010 by chrisdo

Running tasks

Non-Running Tasks

Tasks from Running Jobs

Local Logs

[Log](#) directory

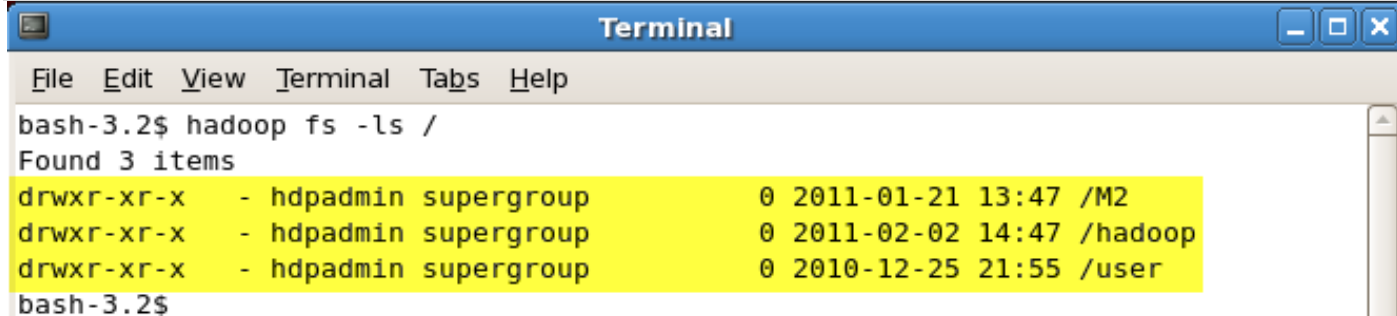
[Hadoop](#), 2011.

Выполнение команд Hadoop

Просмотр файловой системы

Для просмотра файловой системы HDFS можно также использовать консольные команды. Общий вид всех команд – «hadoop fs <args>». В качестве аргументов передается сама команда. Например, список файлов в HDFS можно посмотреть следующим образом:

```
hadoop fs -ls /
```



```
bash-3.2$ hadoop fs -ls /
Found 3 items
drwxr-xr-x  - hdpadmin supergroup    0 2011-01-21 13:47 /M2
drwxr-xr-x  - hdpadmin supergroup    0 2011-02-02 14:47 /hadoop
drwxr-xr-x  - hdpadmin supergroup    0 2010-12-25 21:55 /user
bash-3.2$
```

Копируем данные

Для тех, кто работает на удаленном сервере, везде где используется `##yourname##` нужно использовать свою фамилию. Например, я использую «vovchenko». Для тех кто работает на своей виртуальной машине можно использовать любое имя.

Создадим в HDFS каталог с нашим именем:

```
hadoop fs -mkdir ##yourname##
```

Можно посмотреть содержимое созданного каталога, которое будет пустым:

```
hadoop fs -ls ##yourname##
```

Затем скопируем туда входные данные из локальной файловой системы

```
hadoop fs -put /home/biadmin/labsData/hadoop_lab/ ##yourname##
```

Если заново выполнить команду, для просмотра содержимого своего каталога, то можно увидеть не пустой результат.

```
hadoop fs -ls ##yourname##
```

Если зайти через веб-интерфейс web UI for HDFS name node(s), и перейдя по ссылке Browse the Filesystem, можно также обнаружить свои данные, которые были загружены в Hadoop.

Работа с Map-Reduce приложениями

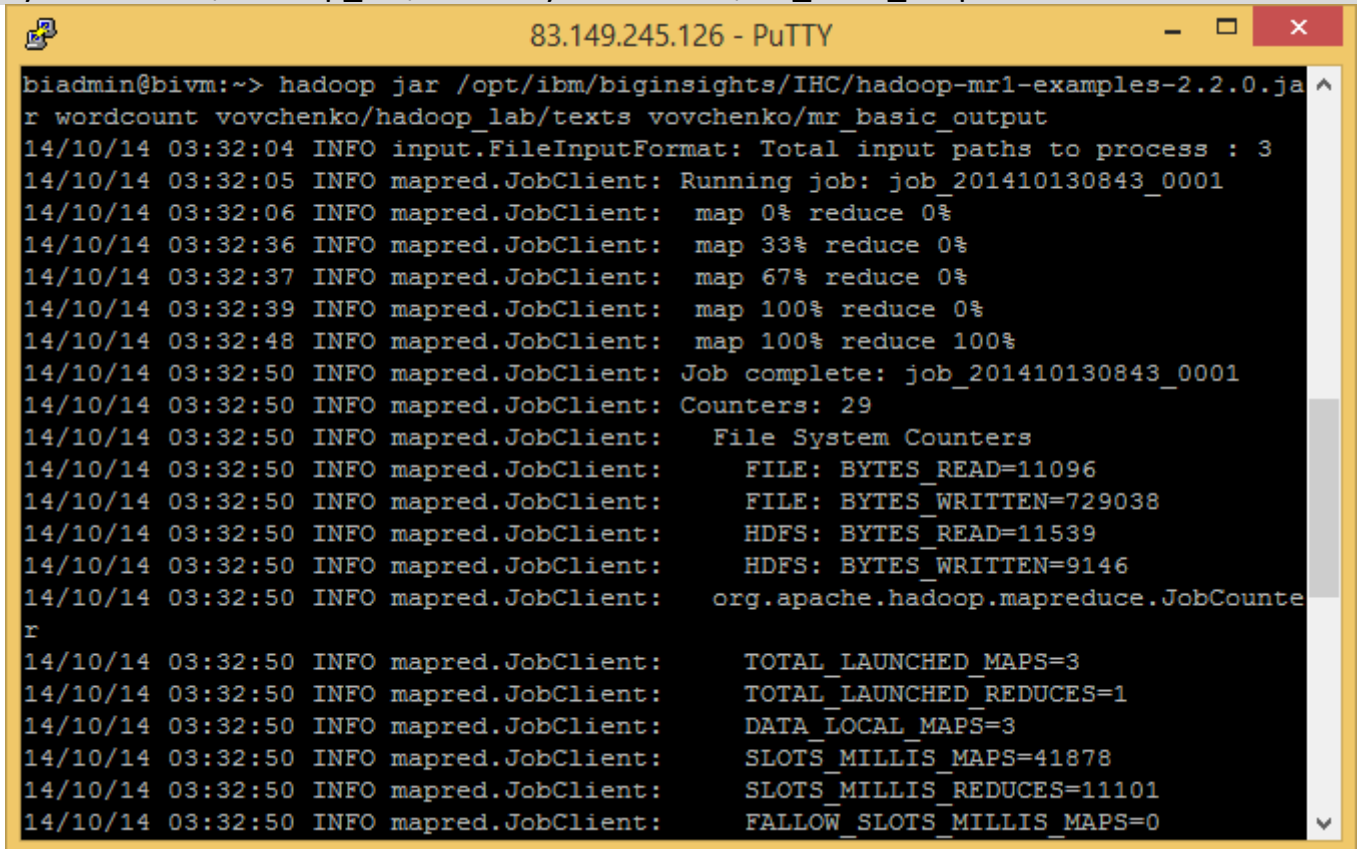
Запуск простого Map-Reduce приложения

Собственное Map-Reduce приложение мы будем разрабатывать в следующей лабораторной работе. В этом задании мы будем запускать приложения, которые поставляются вместе с дистрибутивом Hadoop. Подробно можно прочитать здесь:

http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Выполним команду:

```
hadoop jar /opt/ibm/biginsights/IHC/hadoop-mr1-examples-2.2.0.jar wordcount  
##yourname##/hadoop_lab/texts ##yourname##/mr_basic_output
```



```
83.149.245.126 - PuTTY  
biadmin@bivm:~> hadoop jar /opt/ibm/biginsights/IHC/hadoop-mr1-examples-2.2.0.jar  
wordcount vovchenko/hadoop_lab/texts vovchenko/mr_basic_output  
14/10/14 03:32:04 INFO input.FileInputFormat: Total input paths to process : 3  
14/10/14 03:32:05 INFO mapred.JobClient: Running job: job_201410130843_0001  
14/10/14 03:32:06 INFO mapred.JobClient: map 0% reduce 0%  
14/10/14 03:32:36 INFO mapred.JobClient: map 33% reduce 0%  
14/10/14 03:32:37 INFO mapred.JobClient: map 67% reduce 0%  
14/10/14 03:32:39 INFO mapred.JobClient: map 100% reduce 0%  
14/10/14 03:32:48 INFO mapred.JobClient: map 100% reduce 100%  
14/10/14 03:32:50 INFO mapred.JobClient: Job complete: job_201410130843_0001  
14/10/14 03:32:50 INFO mapred.JobClient: Counters: 29  
14/10/14 03:32:50 INFO mapred.JobClient: File System Counters  
14/10/14 03:32:50 INFO mapred.JobClient: FILE: BYTES_READ=11096  
14/10/14 03:32:50 INFO mapred.JobClient: FILE: BYTES_WRITTEN=729038  
14/10/14 03:32:50 INFO mapred.JobClient: HDFS: BYTES_READ=11539  
14/10/14 03:32:50 INFO mapred.JobClient: HDFS: BYTES_WRITTEN=9146  
14/10/14 03:32:50 INFO mapred.JobClient: org.apache.hadoop.mapreduce.JobCounte  
r  
14/10/14 03:32:50 INFO mapred.JobClient: TOTAL_LAUNCHED_MAPS=3  
14/10/14 03:32:50 INFO mapred.JobClient: TOTAL_LAUNCHED_REDUCE=1  
14/10/14 03:32:50 INFO mapred.JobClient: DATA_LOCAL_MAPS=3  
14/10/14 03:32:50 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=41878  
14/10/14 03:32:50 INFO mapred.JobClient: SLOTS_MILLIS_REDUCE=11101  
14/10/14 03:32:50 INFO mapred.JobClient: FALLOW_SLOTS_MILLIS_MAPS=0
```

Открываем <http://bivm:50030/> или <http://83.149.245.126:50030/> – web UI for MapReduce, и можно увидеть свое запущенное задание, либо как выполняющиеся, либо как уже выполненное, если прошло достаточно времени.

localhost Hadoop Map/Reduce Administration - Mozilla

File Edit View History Bookmarks Tools Help

http://localhost:50030/jobtracker.jsp

Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support

Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	M C
job_201102071335_0001	NORMAL	hdppadmin	word count	27.27%	11	3

Completed Jobs

[Просмотр результатов в консоли](#)

После того как задача выполнена, и ваше задание попало в раздел «Completed Jobs», можно посмотреть результаты. Для этого выполним команду:

```
hadoop fs -ls ##yourname##/mr_basic_output
```

После этого можно посмотреть и сам результат:

```
hadoop fs -cat ##yourname##/mr_basic_output/*00
```

```
83.149.245.126 - PuTTY
biadmin@bivm:~> hadoop fs -cat vovchenko/mr_basic_output/*00
"
  4
""
  5
"
  1
",1524598,1520445,1553218,1542203,1571660,1557793,1567143,1562846,1581388,157081
2,1515700,1462185,1394690,1341157,1293244,1269790,1236698,1224836,1226113,125594
1,1263595,1270366,1267922,1254308
",1562846,1394690,1224836,1270366,1267922,1254308
"American
"Source:
"White,
"l
"agencies
"in
(CPDF).
(Nonpostal)
(non-Postal)
(number)
(number),1995
(number),2000
(number),2004
(number),2005
(number),2006
"
  9
"
  9
```


Просмотр результата в веб-интерфейсе

Перейдем по ссылке «по имени задачи в веб интерфейсе», например:
http://83.149.245.126:50030/jobdetails.jsp?jobid=job_201410130843_0001&refresh=0

none

Completed Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Reduce Total
job_201102071335_0001	NORMAL	hdpadmin	word count	100.00%	11	1

Увидим страницу администрирования, на которой можно посмотреть всю интересующую нас информацию по задаче. Сколько было Map, сколько Reduce, сколько времени работало и т.д.

Hadoop job_201102071335_0001

User: hdpadmin
Job Name: word count
Job File: hdfs://localhost.localdomain:9000/hadoop/mapred/system/job_201102071335_0001
Job Setup: Successful
Status: Succeeded
Started at: Mon Feb 07 13:51:41 EST 2011
Finished at: Mon Feb 07 13:53:44 EST 2011
Finished in: 2mins, 2sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed
map	100.00%	11	0	0	11	0
reduce	100.00%	1	0	0	1	0

	Counter	Map	Reduce
Job Counters	Data-local map tasks	0	0
	Launched map tasks	0	0
	Launched reduce tasks	0	0
FileSystemCounters	HDFS_BYTES_READ	7,242,877	0
	FILE_BYTES_WRITTEN	2,789,783	2,789,371
	FILE_BYTES_READ	0	2,789,371
	HDFS_BYTES_WRITTEN	0	1,071,320

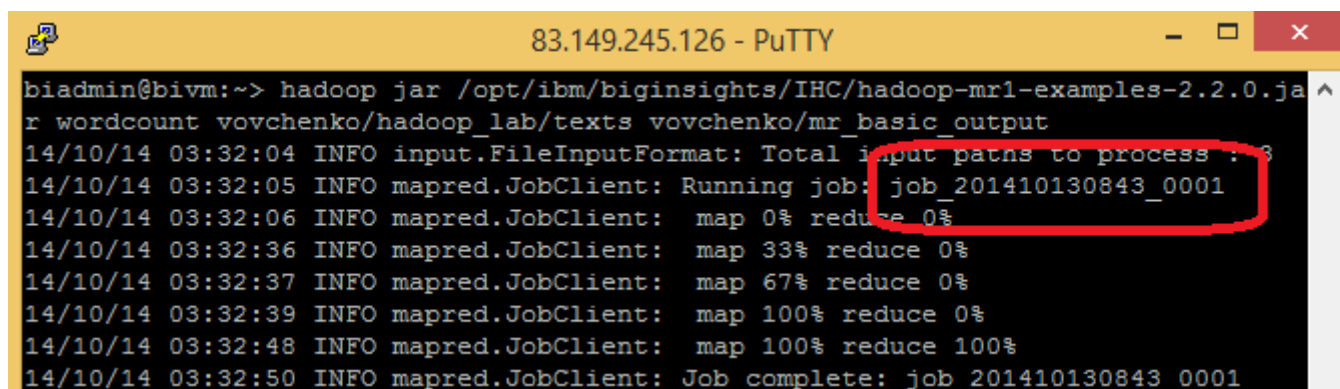
Останов Map-Reduce приложений

Иногда нам может понадобиться остановить приложение. Например, если возникает ошибка при выполнении, или приложение зависло. Для этого откроем два окна консоли.

В одном запустим Map-Reduce приложение.

```
hadoop jar /opt/ibm/biginsights/IHC/hadoop-mr1-examples-2.2.0.jar wordcount
##yourname##/hadoop_lab/texts ##yourname## /mr_basic_output2
```

После запуска можно посмотреть имя задания:



```
83.149.245.126 - PuTTY
biadmin@bivm:~> hadoop jar /opt/ibm/biginsights/IHC/hadoop-mr1-examples-2.2.0.jar
wordcount vovchenko/hadoop_lab/texts vovchenko/mr_basic_output
14/10/14 03:32:04 INFO input.FileInputFormat: Total input paths to process : 3
14/10/14 03:32:05 INFO mapred.JobClient: Running job: job_201410130843_0001
14/10/14 03:32:06 INFO mapred.JobClient: map 0% reduce 0%
14/10/14 03:32:36 INFO mapred.JobClient: map 33% reduce 0%
14/10/14 03:32:37 INFO mapred.JobClient: map 67% reduce 0%
14/10/14 03:32:39 INFO mapred.JobClient: map 100% reduce 0%
14/10/14 03:32:48 INFO mapred.JobClient: map 100% reduce 100%
14/10/14 03:32:50 INFO mapred.JobClient: Job complete: job_201410130843_0001
```

И затем остановить приложение в другой консоли:

```
hadoop job -kill job_201410130843_0002
```

Вместо `job_201410130843_0002`, нужно использовать идентификатор вашего приложения.

Зайдя в статус приложения можно увидеть:

Hadoop job_201410130843_0002 on [bivm](#)

User: biadmin

Job Name: word count

Job File: hdfs://bivm.ibm.com:9000/user/biadmin/.staging/job_201410130843_0002/job.xml

Submit Host: bivm.ibm.com

Submit Host Address: 83.149.245.126

Job-ACLs:

mapreduce.job.acl-view-job: No users are allowed

mapreduce.job.acl-modify-job: No users are allowed

Job Setup: [Successful](#)

Status: Killed

Failure Info: NA

Started at: Tue Oct 14 03:58:56 EDT 2014

Killed at: Tue Oct 14 03:59:13 EDT 2014

Killed in: 16sec

Job Cleanup: [Successful](#)