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

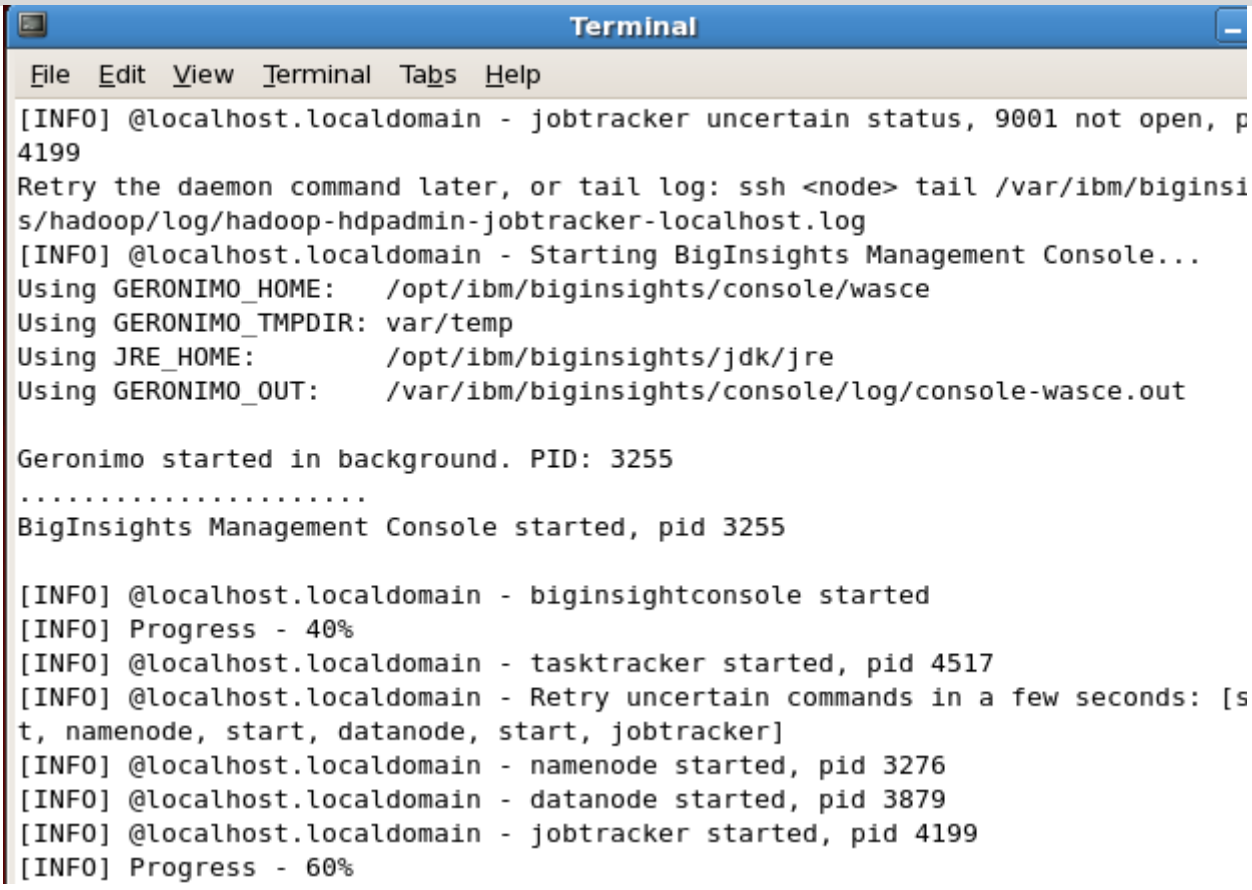
Для доступа к серверу IBM BigInsights необходимо выполнить инструкции из: «server_access.pdf». Если у Вас нет этого файла, необходимо написать мне на почту.

Стартуем Hadoop

Этот этап нужно делать только тем, кто работаете 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
```



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)

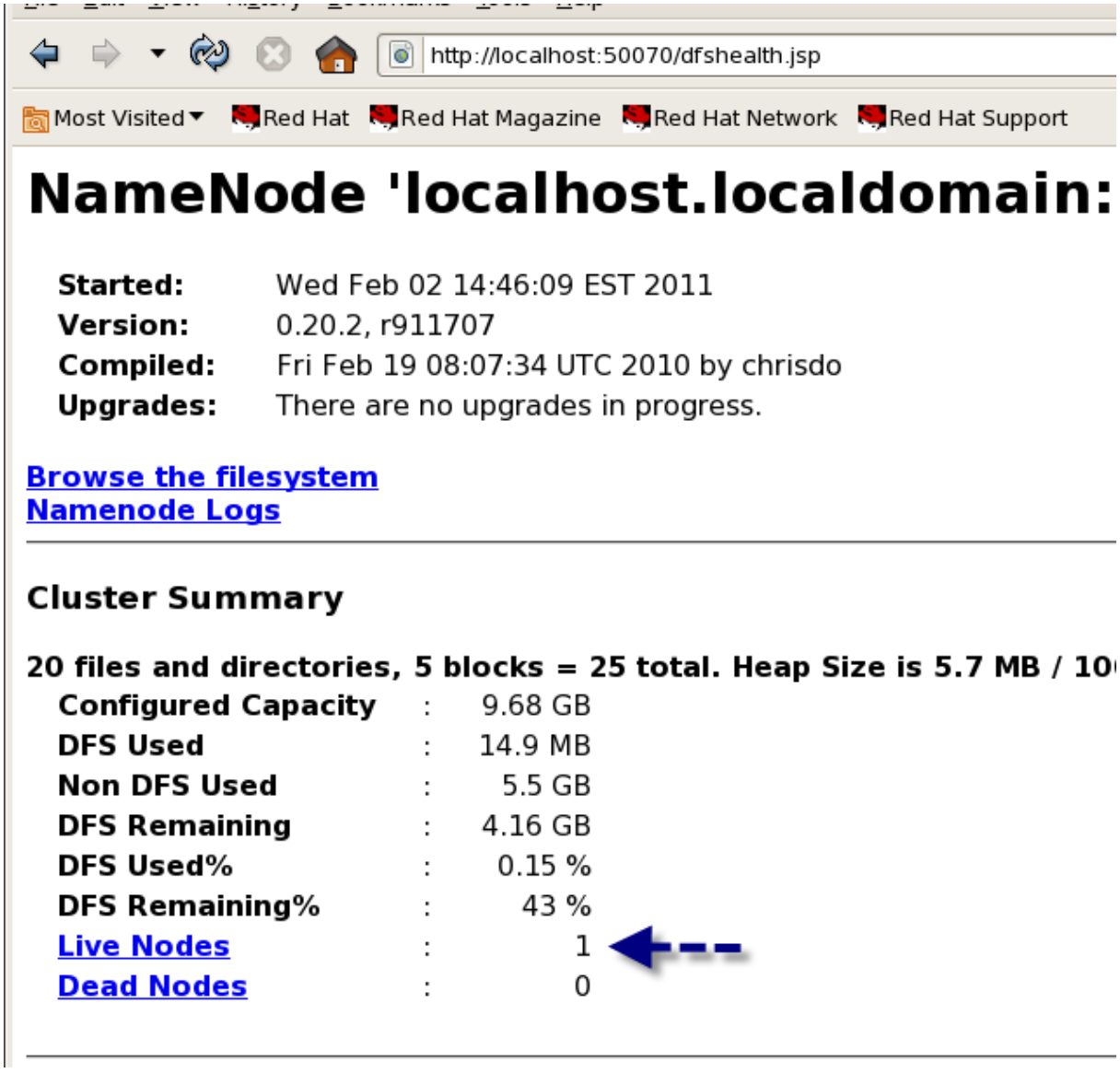
При работе на удаленном сервере

- 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)

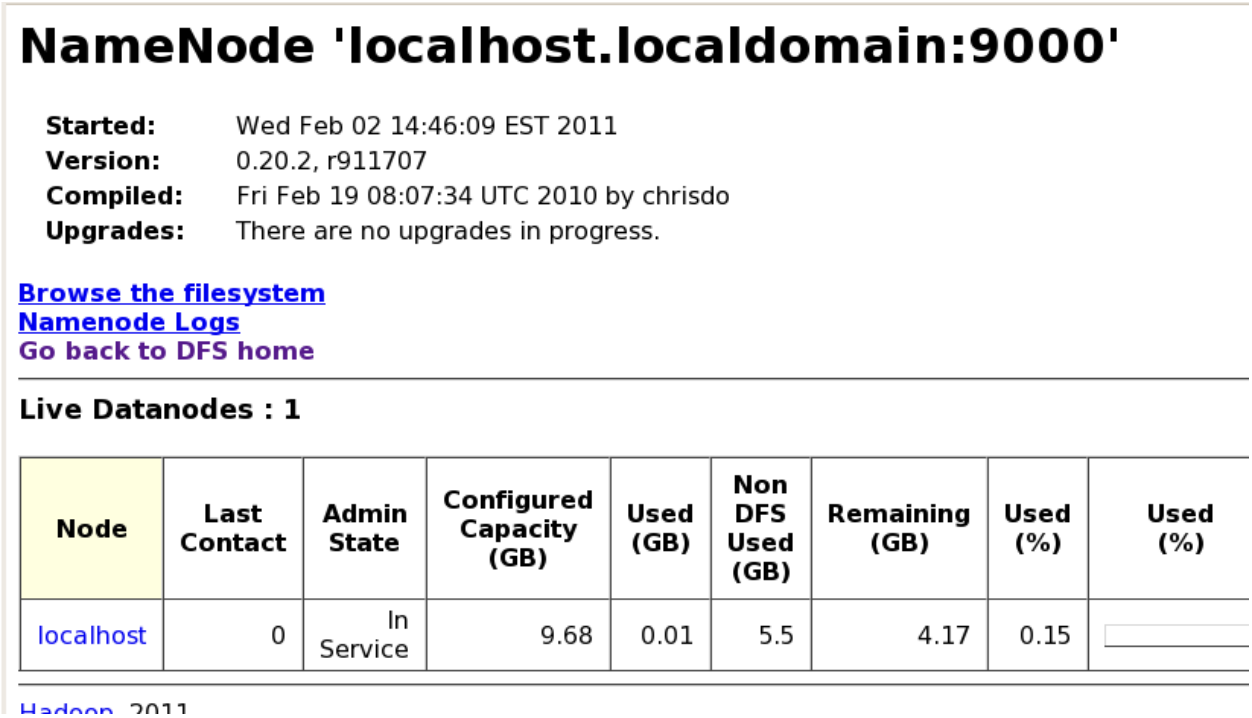
Рассмотрим web UI for HDFS name node(s)

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

На этой странице мы видим информацию о Hadoop кластере.



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	

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=/&nnaddr=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:

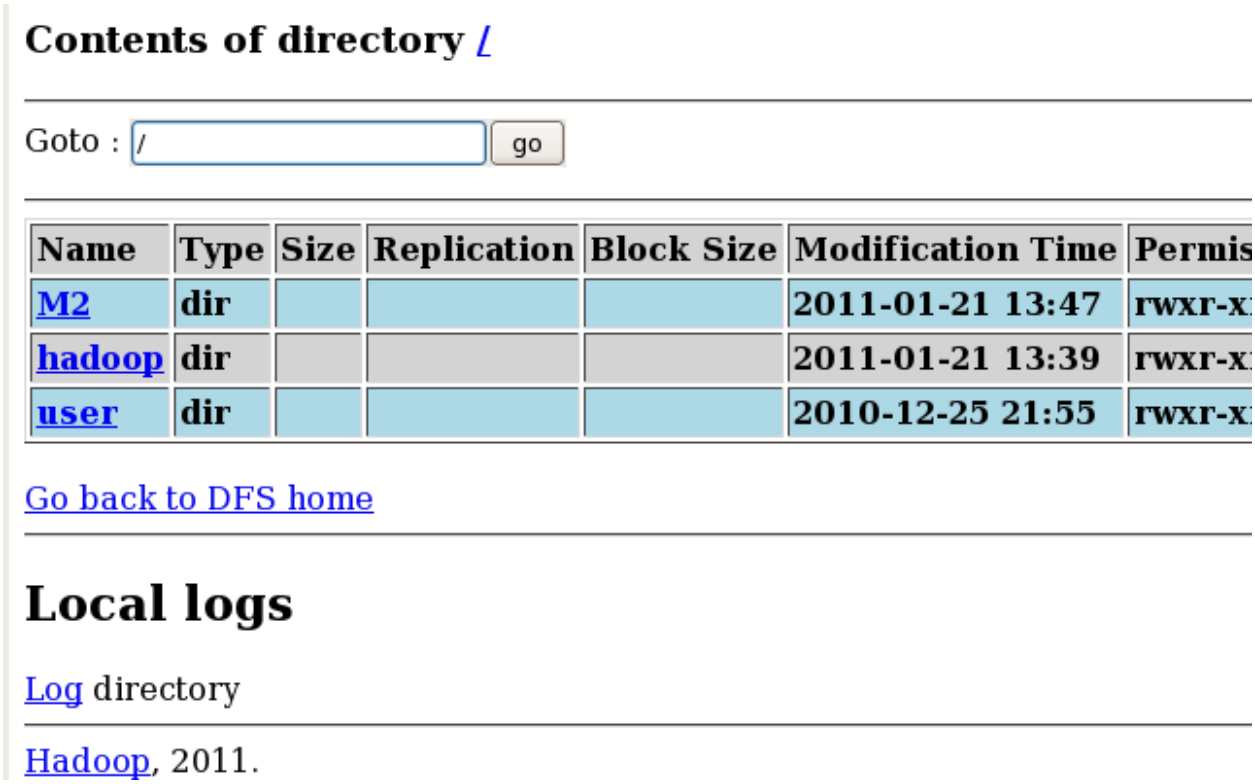
```
…
# For example:
#
#      102.54.94.97     rhino.acme.com          # source server
#       38.25.63.10     x.acme.com              # x client host

# 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.

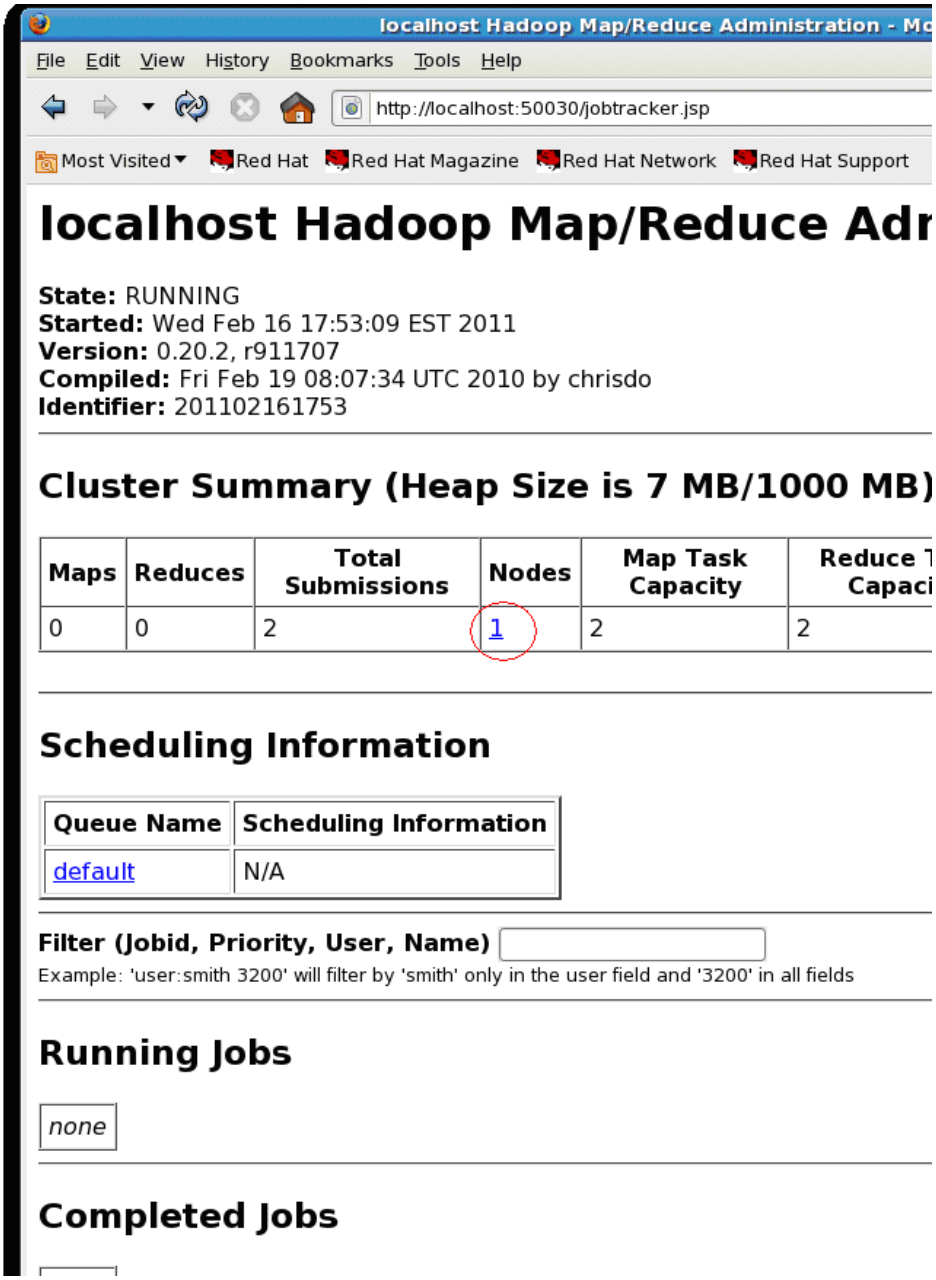


Здесь отображается древовидная структура файловой системы Hadoop Distributed File System (HDFS). Таким образом можно просматривать все данные в файловой системе.

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



Рассмотрим web UI for task tracker(s)

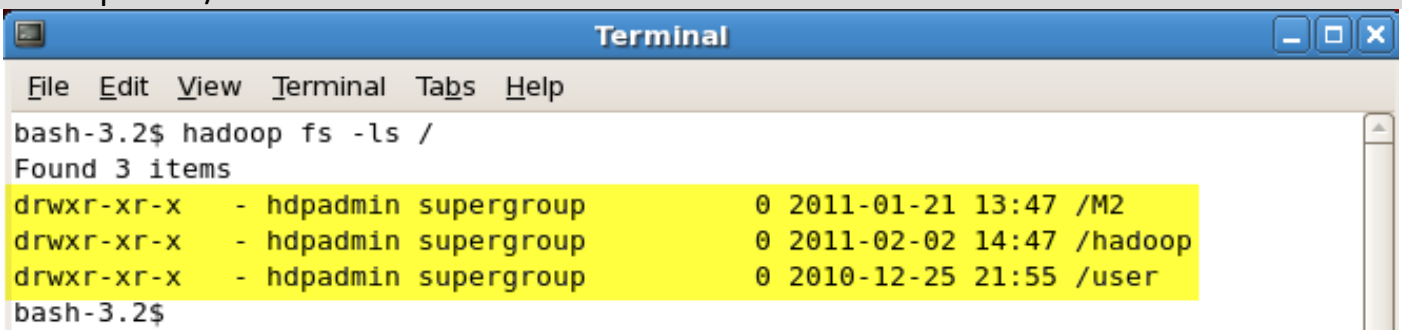
- http://bivm:50060/
- http://83.149.245.126:50060/
- Напрямую данная страница редко используется, однако на нее можно перейти из информации о Map-Reduce приложении
- Например, по этой ссылке, лог выполнения одной из Map задач
- http://83.149.245.126:50060/tasklog?attemptid=attempt_201410130843_0002_m_000000_0&all=true

Выполнение команд Hadoop

Просмотр файловой системы

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

```
hadoop fs -ls /
```



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.

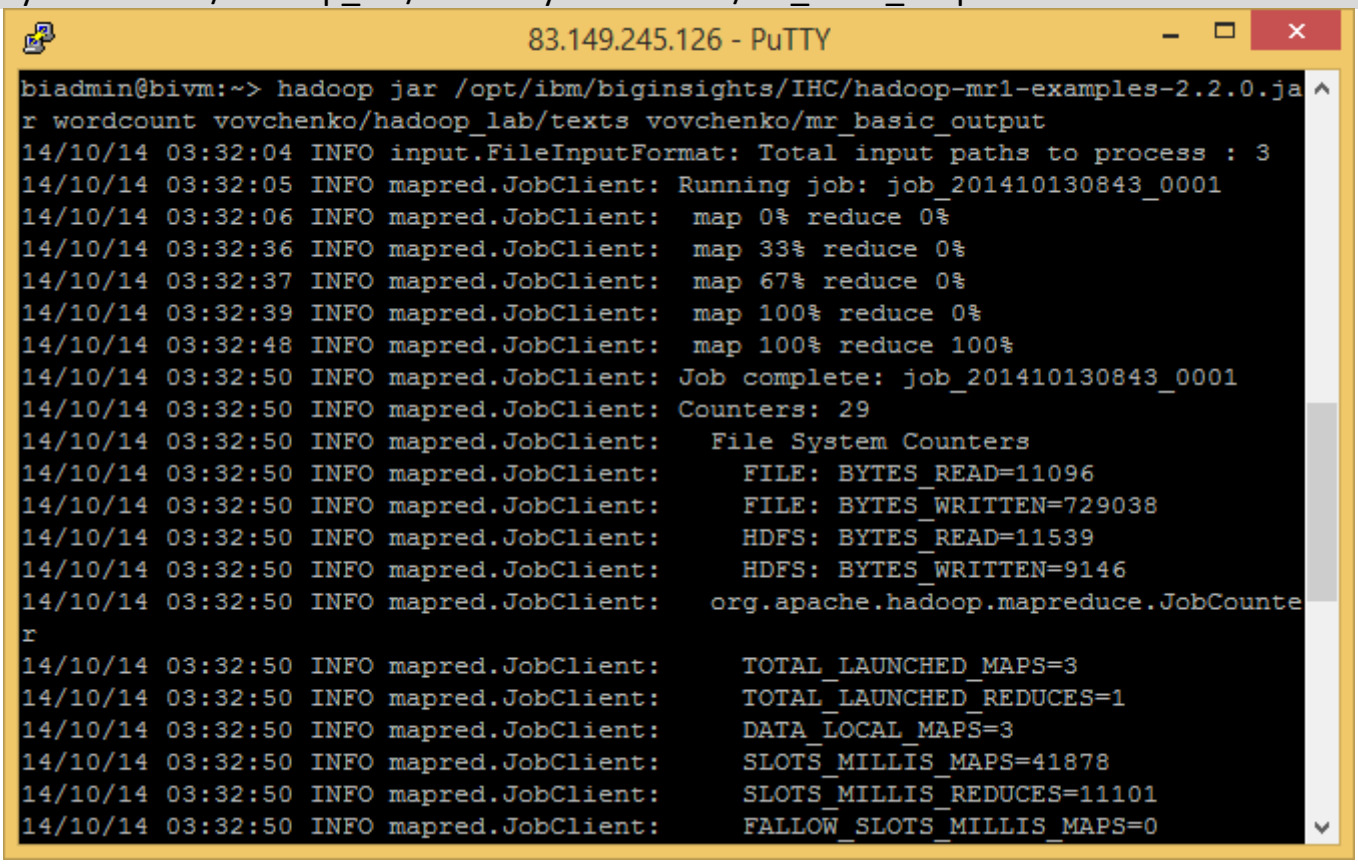
Работа с Map-Reduce приложениями

Запуск простого Map-Reduce приложения

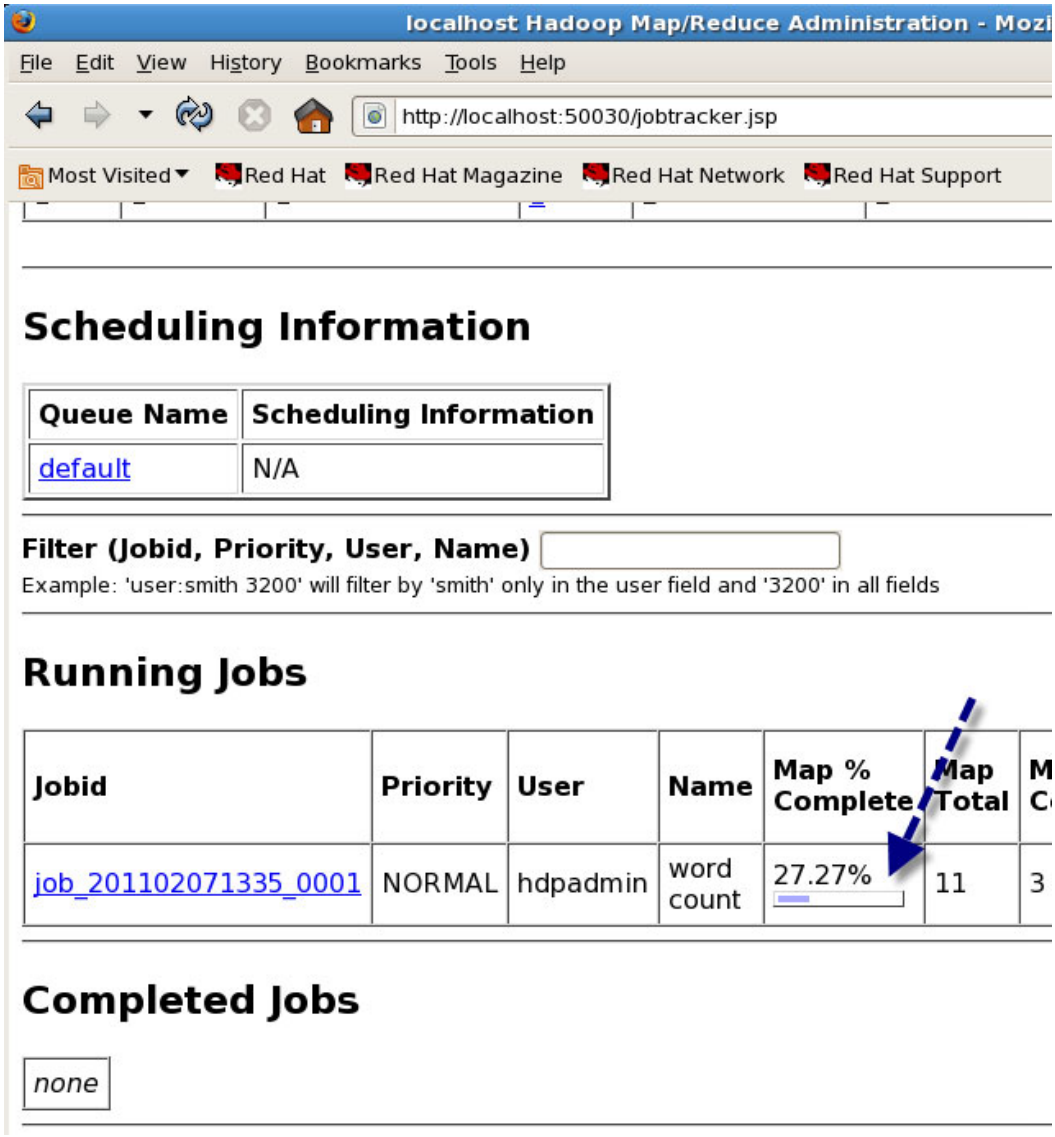
Собственное Map-Reduce приложение мы будет разрабатывать в следующей лабораторной работе. В этом задании мы будем запускать приложения, которые поставляются вместе с дистрибутивом Hadoop. Подробно можно прочитать здесь: http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Выполним команду:

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



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.



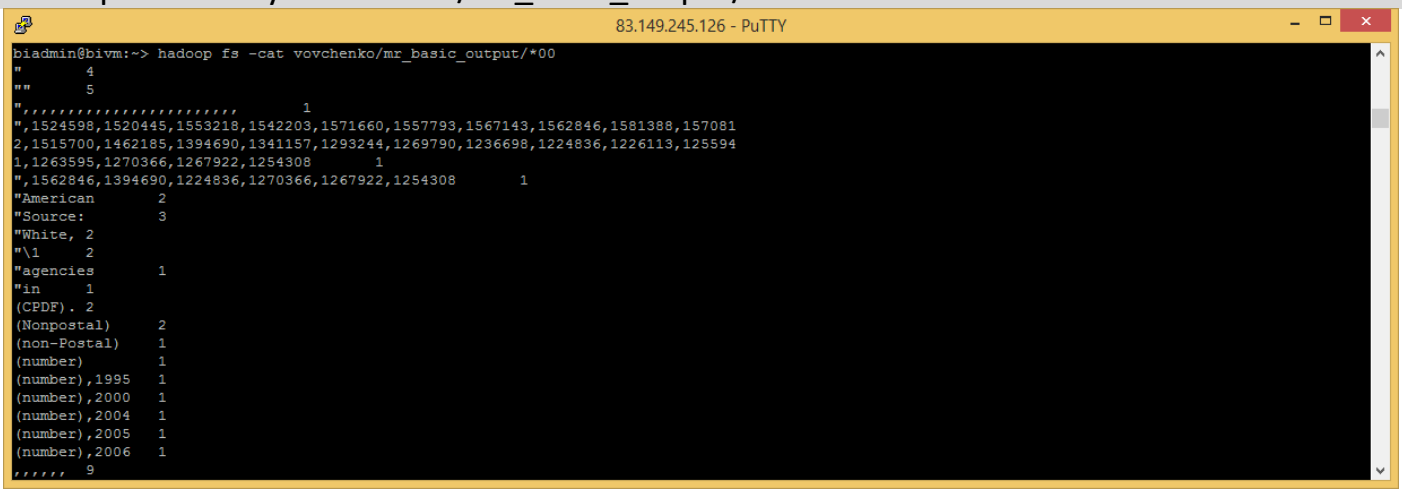
Просмотр результатов в консоли

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
```

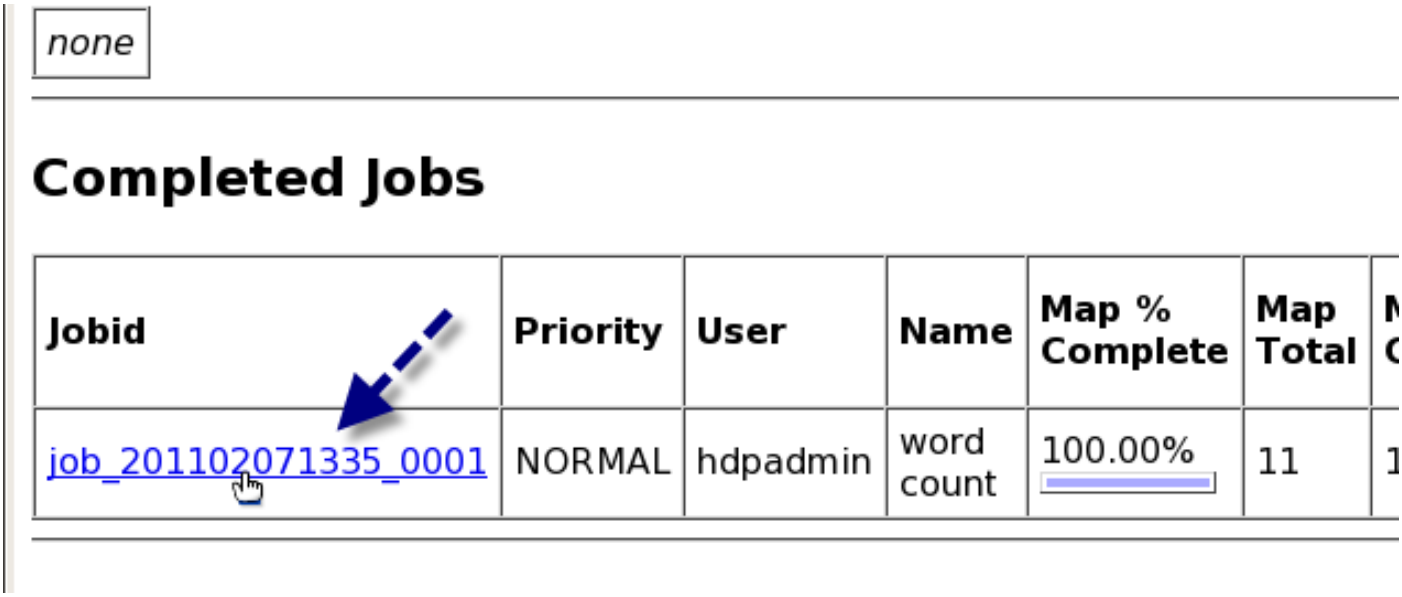
После этого можно посмотреть и сам результат:

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



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



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

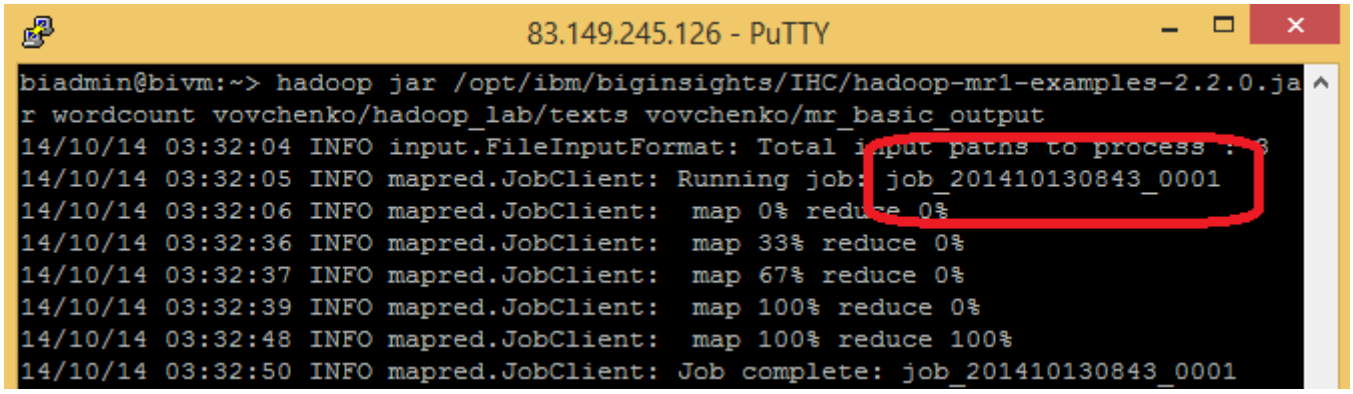
Останов Map-Reduce приложений

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:



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