R HADOOP INTEGRATION in Ubuntu 18.04(LTS Bionic)

Prerequisites

- Install Java (open default jdk java 11)
- Install Hadoop(our version 2.7.3)
- Installation of R and Rstudio

Installation of R and Rstudio in Ubuntu 18.04

In order to install RStudio on Ubuntu 18.04 we will first need to install the
r-base package. Open up terminal and enter:

```
$ sudo apt update

$ sudo apt -y install r-base
```

RStudio for Ubuntu system comes as the *.deb install-able package. The easiest way to install DEB file on Ubuntu Linux is by using the gdebi command. In case gdebi is not available on your system you can install it by executing the bellow command:

```
$ sudo apt install gdebi-core
```

Download RStudio

Next, navigate your browser to the official RStudio download page and download the latest Ubuntu/Debian RStudio *.deb package available. At the time of writing the Ubuntu 18.04 Bionic package is not available yet. If this is still the case download the Ubuntu 16.04 Xenial package instead.

```
$ cd Downloads/
$ ls
rstudio-xenial-1.1.442-amd64.deb
```

Install RStudio on Ubuntu

At this stage we are ready to install RStudio on our Ubuntu 18.04 system. Run the below `gdebi` command from the location of your downloaded RStudio package while replacing the package name where appropriate. When prompted, answer `y` to proceed with the installation:\

```
$ sudo gdebi rstudio-xenial-1.1.442-amd64.deb
```

Once the installation of RStudio on your Ubuntu system is completed you can start RStudio by executing the following linux command:

```
$ rstudio
```

Alternatively, search your start menu and start RStudio by clicking on its icon:
Start RStudio on Ubuntu 18.04
RStudio on Ubuntu 18.04

Install rJava in Rstudio

Write in Terminal: java -version
If it returns The program java can be found in the following packages, then Java hasn't been installed yet, so execute the following command:

```
sudo apt-get install default-jre
```

This will install the Java Runtime Environment (JRE).
Then install JDK

Write in Terminal: **sudo apt-get install default-jdk**

Then assotiate the JDK installed with R

Run in Terminal: **sudo R CMD javareconf**

Install RJava and Rgdal

Execute: **sudo apt-get install r-cran-rjava**
Then: **sudo apt-get install libgdal-dev libproj-dev**

Install package in RStudio

Run in RStudio: install.packages("rJava")

********Important********
sudo apt-get install libcurl4-openssl-dev libssl-dev libxml2-dev

R HADOOP Integration using rhdfs

Required Packages for Installing

We require several R packages to be installed for connecting R with Hadoop. The list of packages are as follows:

- rJava
- RJSONIO
- itertools
- digest
- Rcpp
- httr
- functional
- devtools
- plyr
- reshape2

Once the R and Rstudio is installed, install the above mentioned package in Rstudio console. There are 2 ways for the package installation

1st way :

```
install.packages( c('rJava','RJSONIO', 'itertools', 'digest','Rcpp
','httr','functional','devtools',
'plyr','reshape2'),dependencies=TRUE,repos='http://cran.rstudio.com/'
)
```

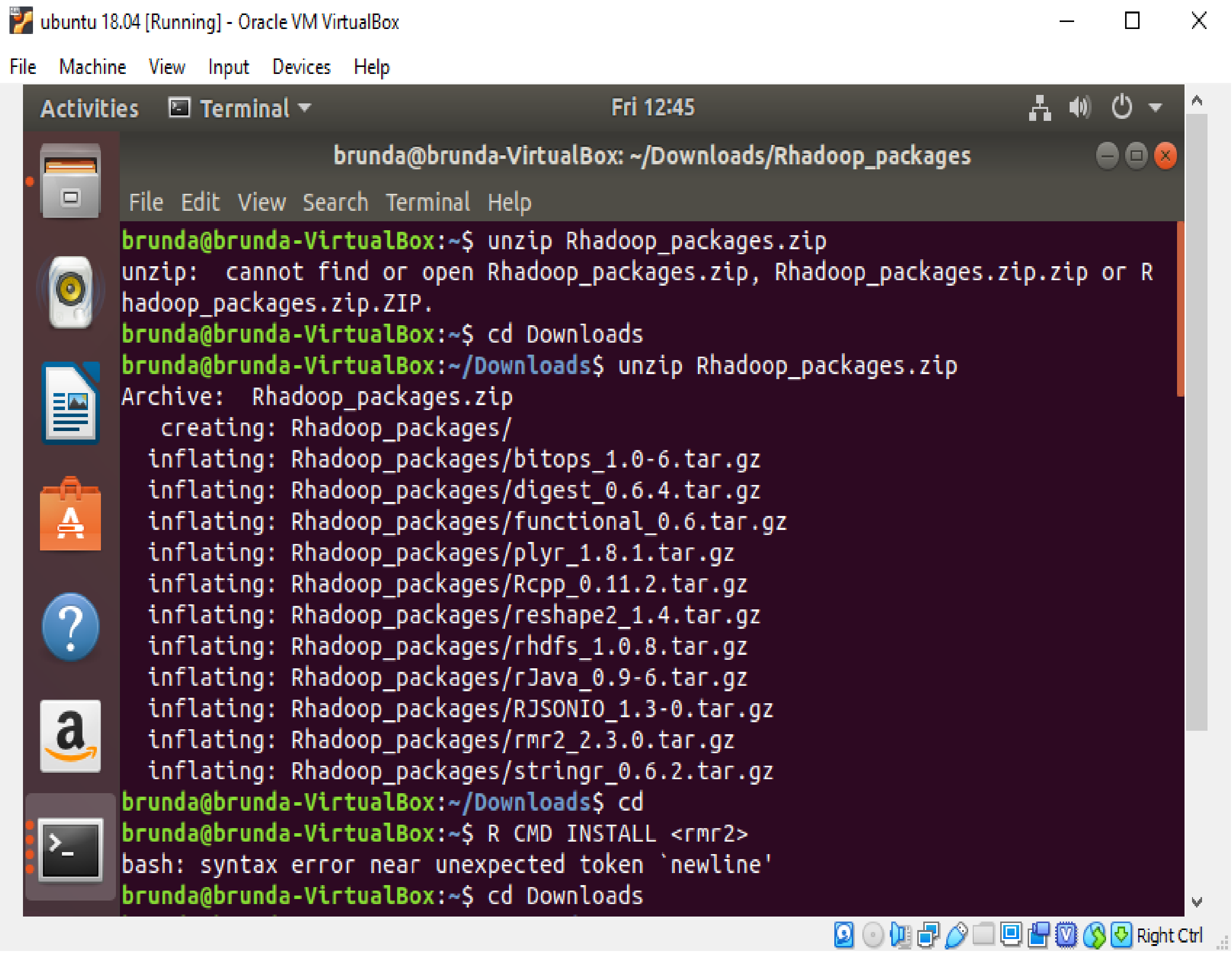
Note: Before installing rJava, we should set the JAVA_HOME path and should login to R with sudo privileges.

2nd way (we have chosen the 1st way for the installation of the packages)

Downloading Packages and installing through R cmd:

Download the required packages from the below link.
Link:
https://drive.google.com/open?id=0B5dejdhAYHztRkgzbGZOeUdXdVE
After downloading the packages, extract them and use the below command:

```
unzip Rhadoop_packages.zip
```


ubuntu 18.04 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Fri 12:45
brunda@brunda-VirtualBox: ~/Downloads/Rhadoop_packages
File Edit View Search Terminal Help
Right Ctrl

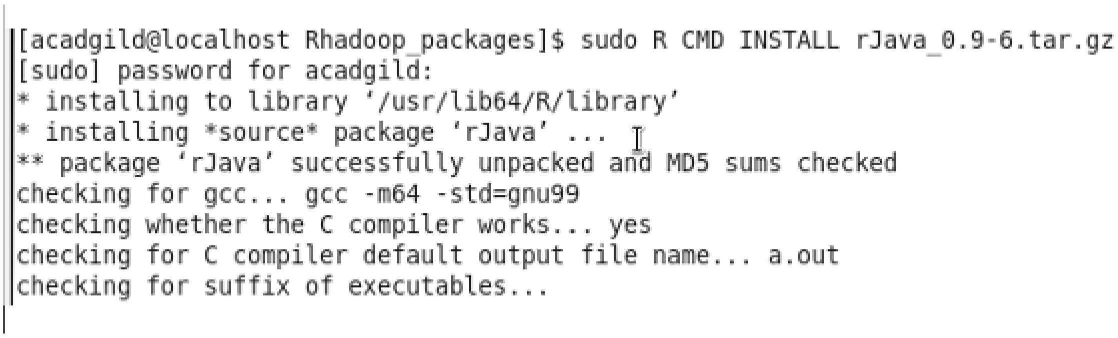
```
brunda@brunda-VirtualBox:~$ unzip Rhadoop_packages.zip
unzip:  cannot find or open Rhadoop_packages.zip, Rhadoop_packages.zip.zip or R
hadoop_packages.zip.ZIP.
brunda@brunda-VirtualBox:~$ cd Downloads
brunda@brunda-VirtualBox:~/Downloads$ unzip Rhadoop_packages.zip
Archive:  Rhadoop_packages.zip
   creating: Rhadoop_packages/
  inflating: Rhadoop_packages/bitops_1.0-6.tar.gz
  inflating: Rhadoop_packages/digest_0.6.4.tar.gz
  inflating: Rhadoop_packages/functional_0.6.tar.gz
  inflating: Rhadoop_packages/plyr_1.8.1.tar.gz
  inflating: Rhadoop_packages/Rcpp_0.11.2.tar.gz
  inflating: Rhadoop_packages/reshape2_1.4.tar.gz
  inflating: Rhadoop_packages/rhdfs_1.0.8.tar.gz
  inflating: Rhadoop_packages/rJava_0.9-6.tar.gz
  inflating: Rhadoop_packages/RJSONIO_1.3-0.tar.gz
  inflating: Rhadoop_packages/rmr2_2.3.0.tar.gz
  inflating: Rhadoop_packages/stringr_0.6.2.tar.gz
brunda@brunda-VirtualBox:~/Downloads$ cd
brunda@brunda-VirtualBox:~$ R CMD INSTALL <rmr2>
bash: syntax error near unexpected token `newline'
brunda@brunda-VirtualBox:~$ cd Downloads
```

To install these packages, we will be using R cmd.

```
R CMD INSTALL <package name>
```

Now we will be Installing rJava ,refer the below command for the same.

```
sudo R CMD INSTALL rJava_0.9-6.tar.gz
```

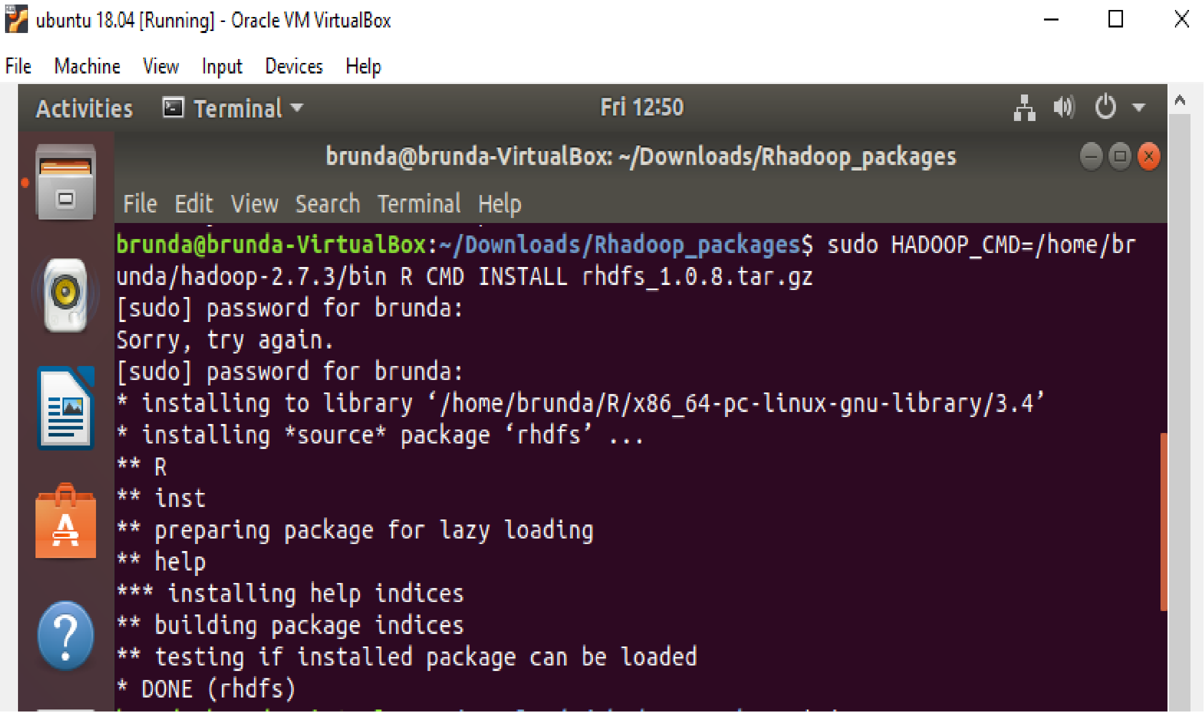


```
[acadgild@localhost Rhadoop_packages]$ sudo R CMD INSTALL rJava_0.9-6.tar.gz
[sudo] password for acadgild:
* installing to library ‘/usr/lib64/R/library’
* installing *source* package ‘rJava’ ...
** package ‘rJava’ successfully unpacked and MD5 sums checked
checking for gcc... gcc -m64 -std=gnu99
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
```

We need to follow the same command to install all the other required packages

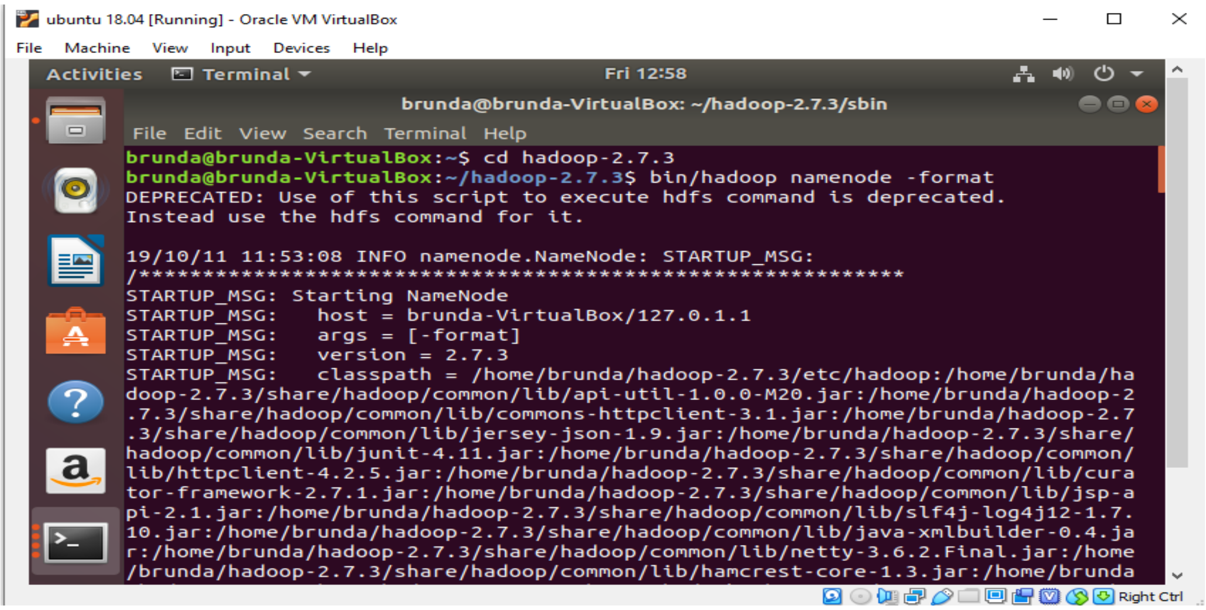
RHDFS INSTALLATION IN RSTUDIO

Before installing rhdfs we should set HADOOP_CMD environmental variable. You can refer to the below screen shot to follow the steps for **Installing Rhdfs.**


ubuntu 18.04 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Fri 12:50
brunda@brunda-VirtualBox: ~/Downloads/Rhadoop_packages
File Edit View Search Terminal Help

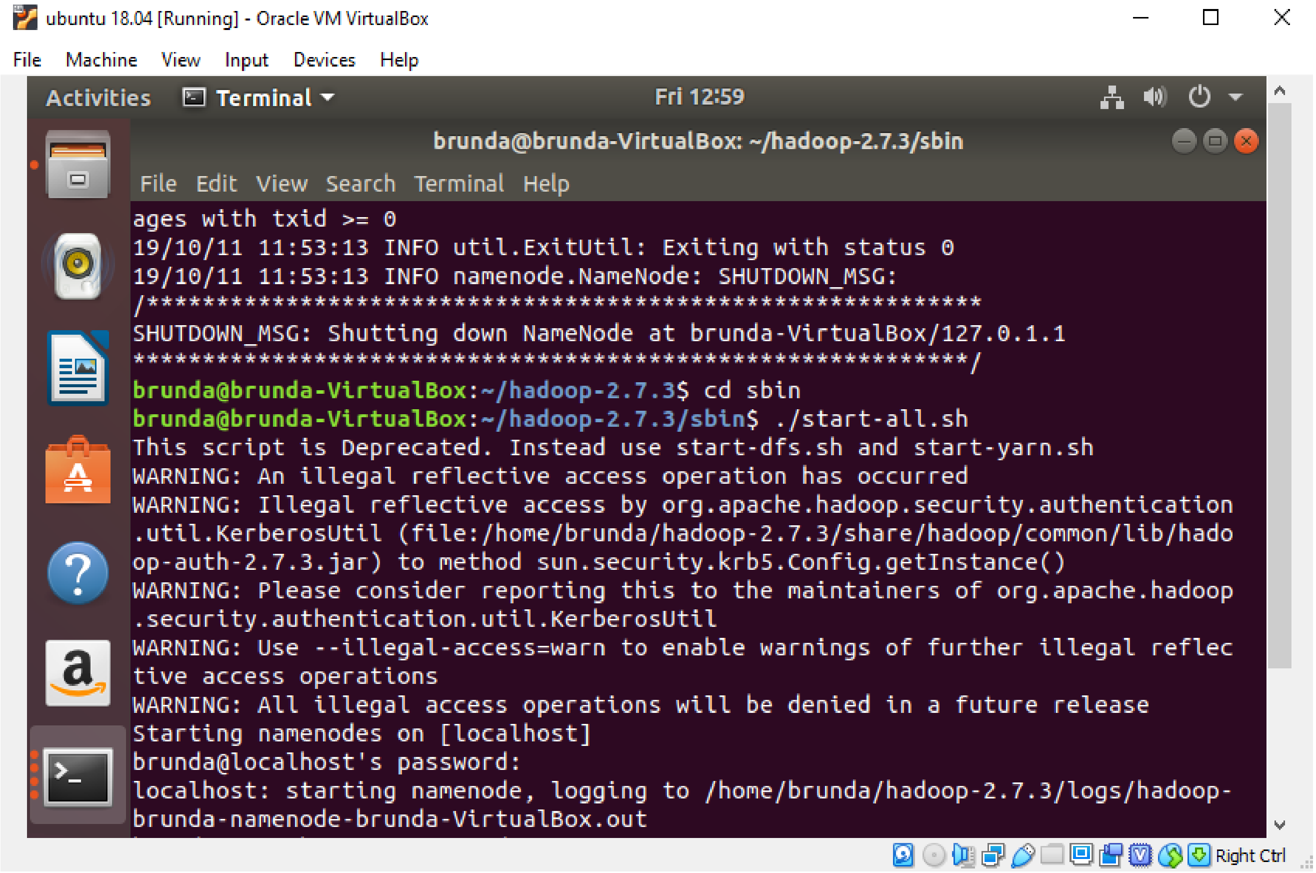
```
brunda@brunda-VirtualBox:~/Downloads/Rhadoop_packages$ sudo HADOOP_CMD=/home/br
unda/hadoop-2.7.3/bin R CMD INSTALL rhdfs_1.0.8.tar.gz
[sudo] password for brunda:
Sorry, try again.
[sudo] password for brunda:
* installing to library ‘/home/brunda/R/x86_64-pc-linux-gnu-library/3.4’
* installing *source* package ‘rhdfs’ ...
** R
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (rhdfs)
```

For accessing HDFS we should start hadoop demons, make sure that all your HDFS daemons are up.



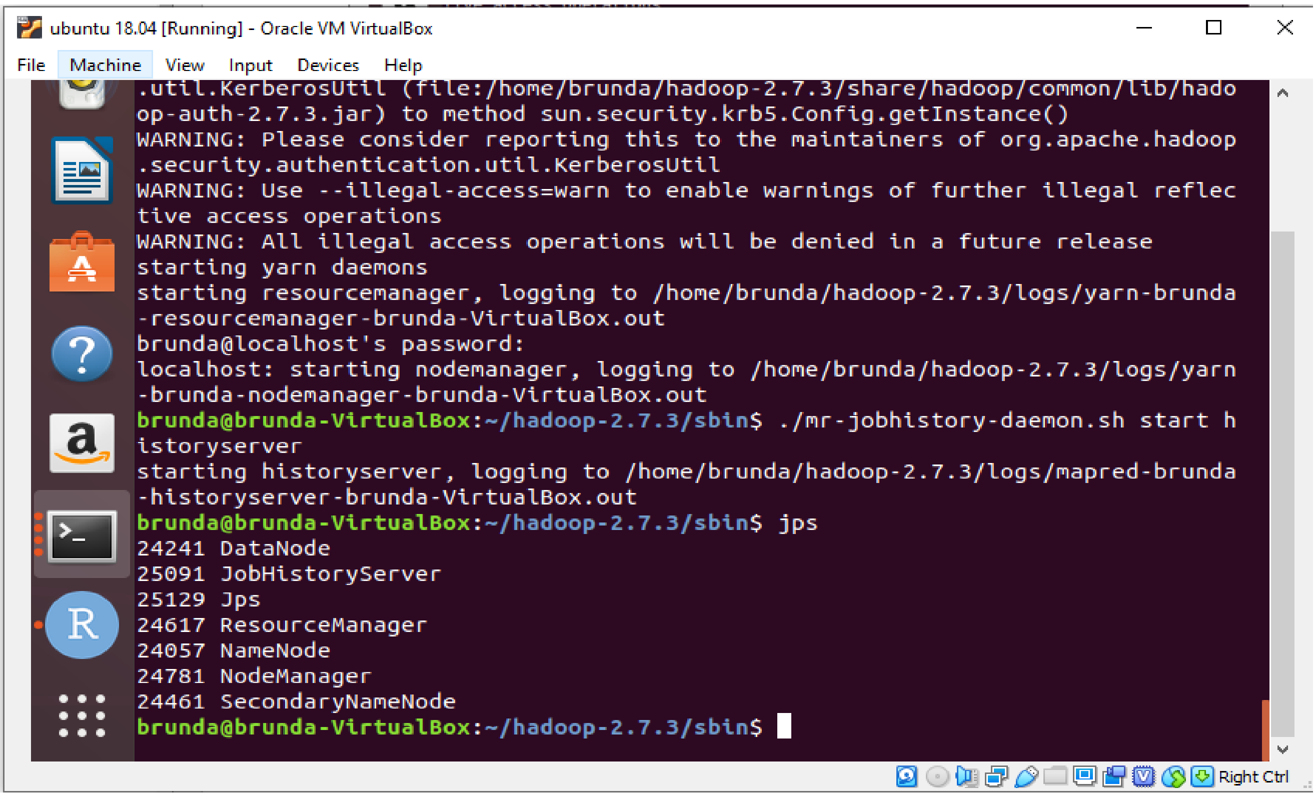
ubuntu 18.04 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Fri 12:58
brunda@brunda-VirtualBox: ~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
Right Ctrl

```
brunda@brunda-VirtualBox:~$ cd hadoop-2.7.3
brunda@brunda-VirtualBox:~/hadoop-2.7.3$ bin/hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/10/11 11:53:08 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = brunda-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /home/brunda/hadoop-2.7.3/etc/hadoop:/home/brunda/ha
doop-2.7.3/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/home/brunda/hadoop-2
.7.3/share/hadoop/common/lib/commons-httpclient-3.1.jar:/home/brunda/hadoop-2.7
.3/share/hadoop/common/lib/jersey-json-1.9.jar:/home/brunda/hadoop-2.7.3/share/
hadoop/common/lib/junit-4.11.jar:/home/brunda/hadoop-2.7.3/share/hadoop/common/
lib/httpclient-4.2.5.jar:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/cura
tor-framework-2.7.1.jar:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/jsp-a
pi-2.1.jar:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.
10.jar:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/java-xmlbuilder-0.4.ja
r:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/netty-3.6.2.Final.jar:/home
/brunda/hadoop-2.7.3/share/hadoop/common/lib/hamcrest-core-1.3.jar:/home/brunda
```



ubuntu 18.04 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Fri 12:59
brunda@brunda-VirtualBox: ~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
Right Ctrl

```
ages with txid >= 0
19/10/11 11:53:13 INFO util.ExitUtil: Exiting with status 0
19/10/11 11:53:13 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at brunda-VirtualBox/127.0.1.1
************************************************************/
brunda@brunda-VirtualBox:~/hadoop-2.7.3$ cd sbin
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/hado
op-auth-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
Starting namenodes on [localhost]
brunda@localhost's password:
localhost: starting namenode, logging to /home/brunda/hadoop-2.7.3/logs/hadoop-
brunda-namenode-brunda-VirtualBox.out
```



```
.util.KerberosUtil (file:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/hado
op-auth-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
starting yarn daemons
starting resourcemanager, logging to /home/brunda/hadoop-2.7.3/logs/yarn-brunda
-resourcemanager-brunda-VirtualBox.out
brunda@localhost's password:
localhost: starting nodemanager, logging to /home/brunda/hadoop-2.7.3/logs/yarn
-brunda-nodemanager-brunda-VirtualBox.out
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$ ./mr-jobhistory-daemon.sh start h
istoryserver
starting historyserver, logging to /home/brunda/hadoop-2.7.3/logs/mapred-brunda
-historyserver-brunda-VirtualBox.out
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$ jps
24241 DataNode
25091 JobHistoryServer
25129 Jps
24617 ResourceManager
24057 NameNode
24781 NodeManager
24461 SecondaryNameNode
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$
```

Steps to start the Hadoop demon

$ cd hadoop-2.7.3

$ bin/hadoop namenode –format

$ cd sbin

$ ./start-all.sh

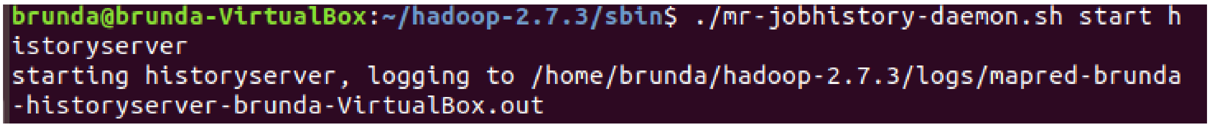
All the nodes will be started accordingly it will ask for the password

The demons started are namely:

- Namenode
- Datanode
- Resouremanager
- Nodemanager

Need to start the job history server

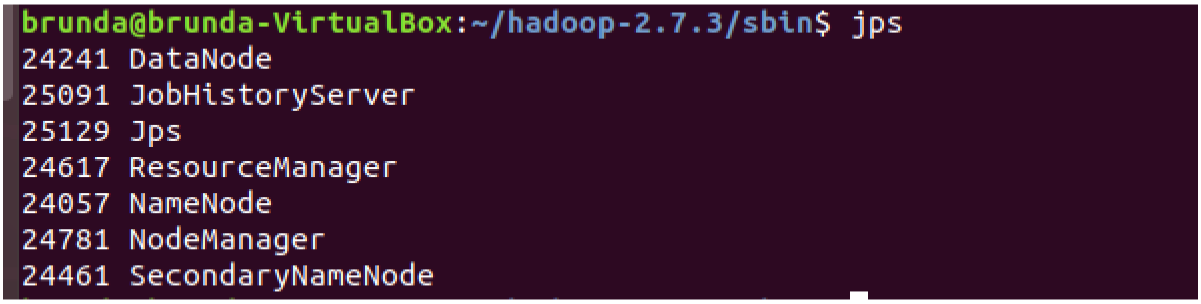
$ ./mr-jobhistory-daemon.sh start historyserver



```
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$ ./mr-jobhistory-daemon.sh start h
istoryserver
starting historyserver, logging to /home/brunda/hadoop-2.7.3/logs/mapred-brunda
-historyserver-brunda-VirtualBox.out
```

To check all the Hadoop services are up and running

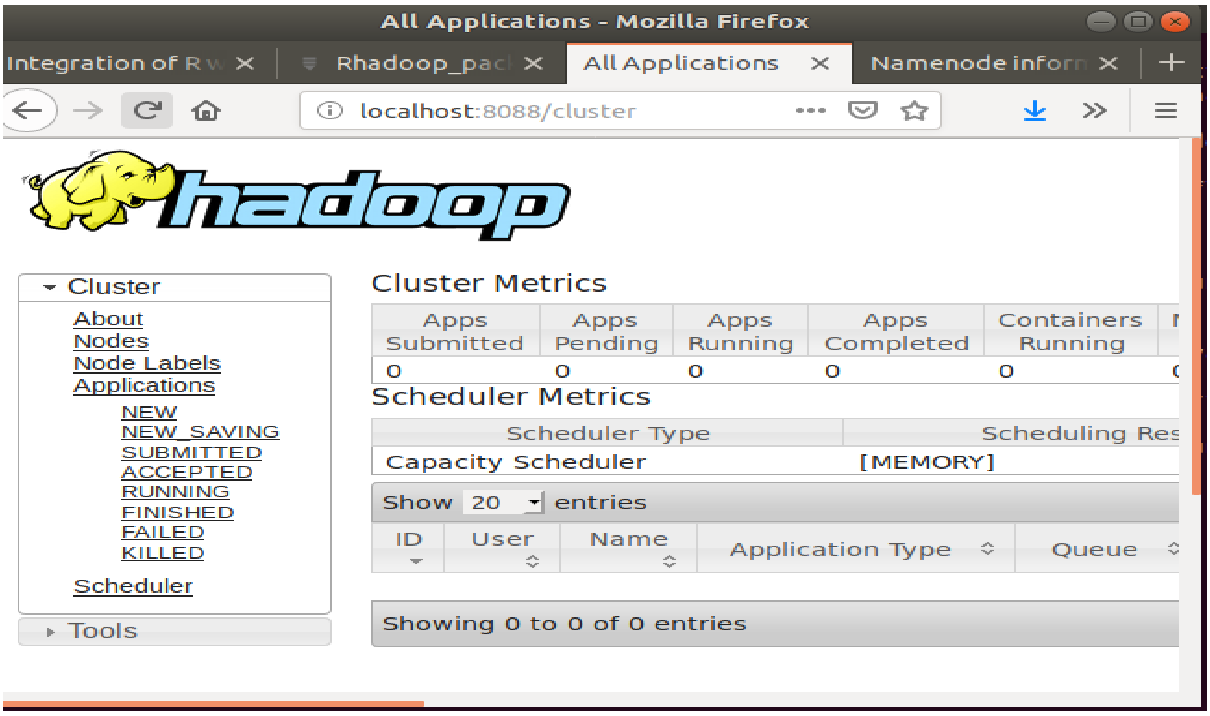
$ jps



```
brunda@brunda-VirtualBox:~/hadoop-2.7.3/sbin$ jps
24241 DataNode
25091 JobHistoryServer
25129 Jps
24617 ResourceManager
24057 NameNode
24781 NodeManager
24461 SecondaryNameNode
```

So Now successfully all the daemons are up. Now open browser and execute the

Localhost:8088   localhost:50070



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	M
0	0	0	0	0	0

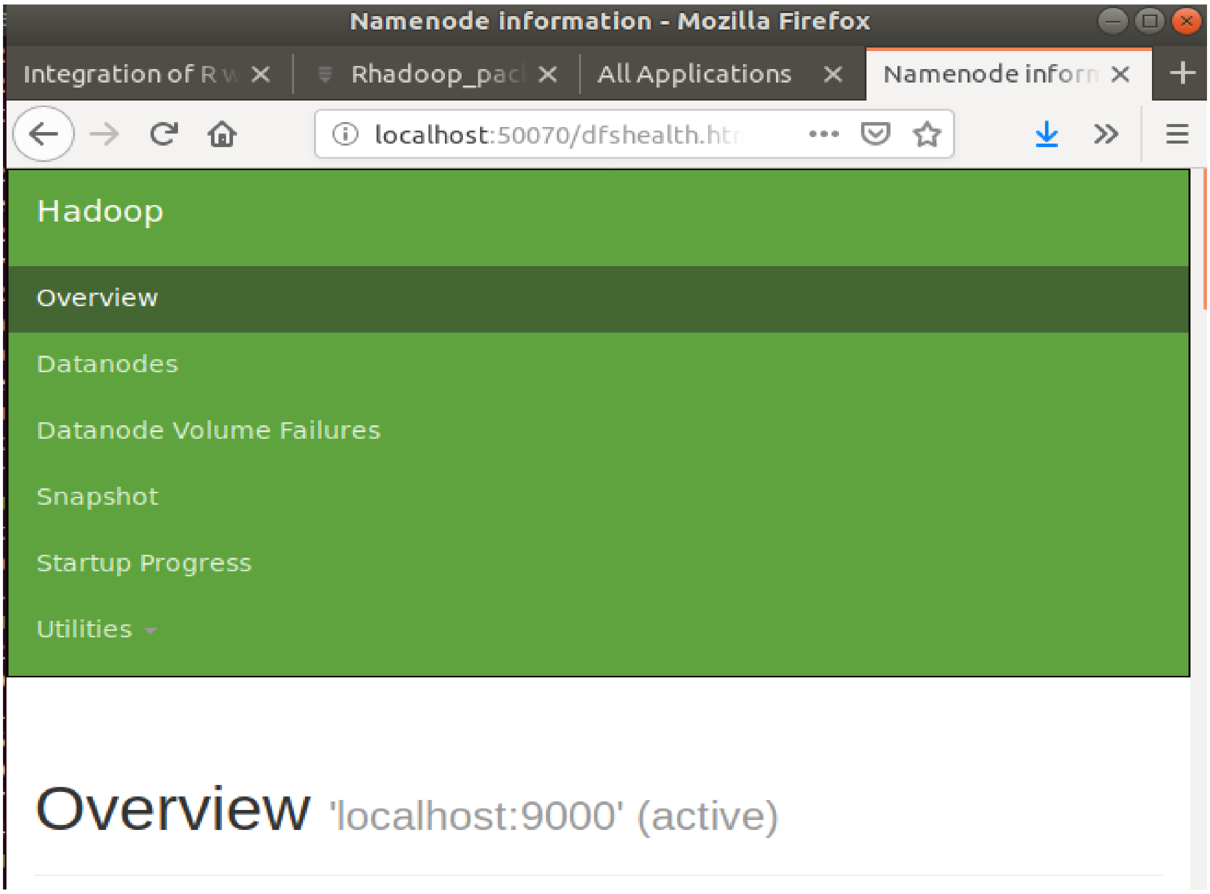
Scheduler Metrics

Scheduler Type	Scheduling Res
Capacity Scheduler	[MEMORY]

Show 20 entries

ID	User	Name	Application Type	Queue

Showing 0 to 0 of 0 entries



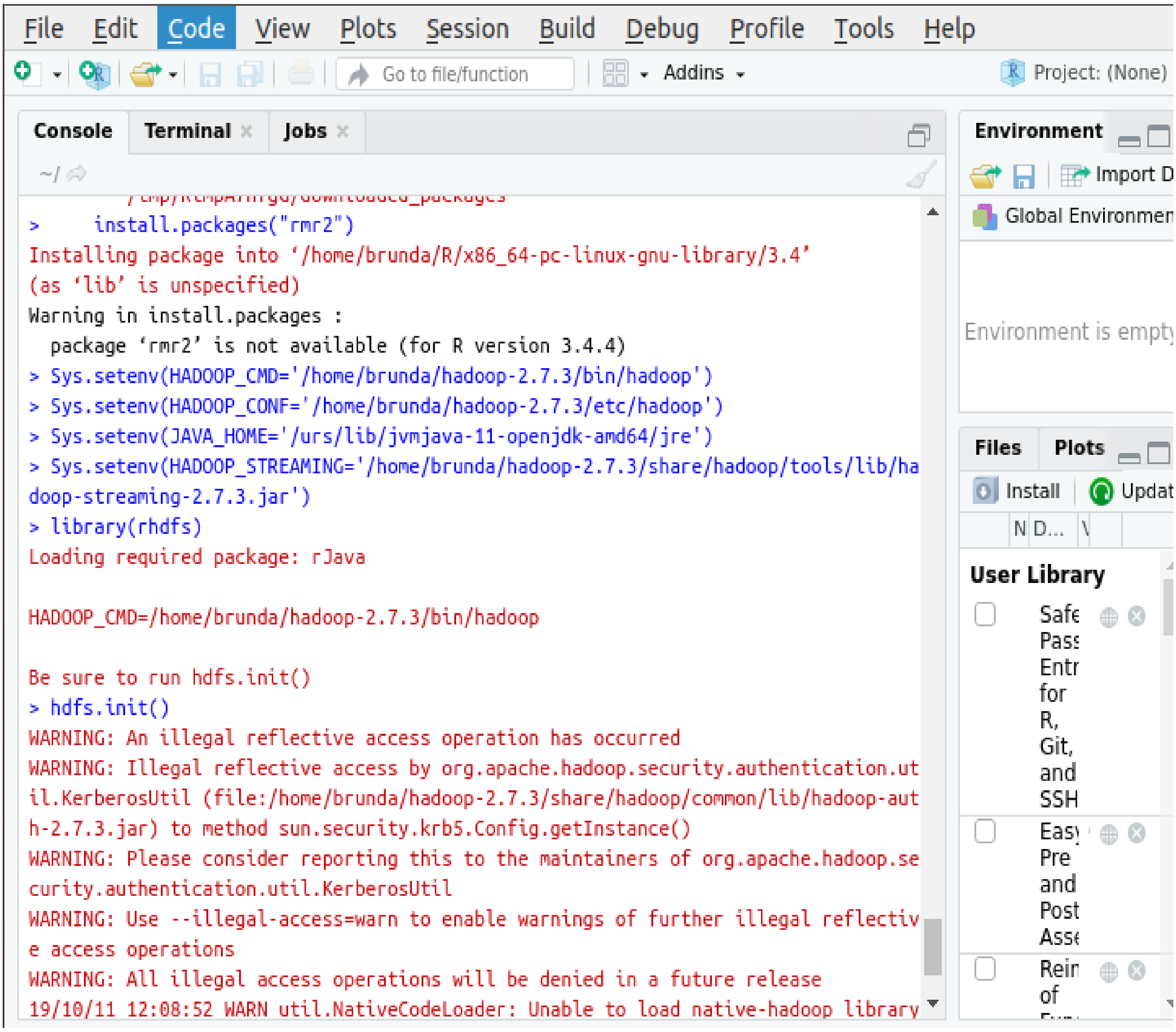
Overview 'localhost:9000' (active)

RHADOOP INTEGRATION

Now we will access HDFS from the R console

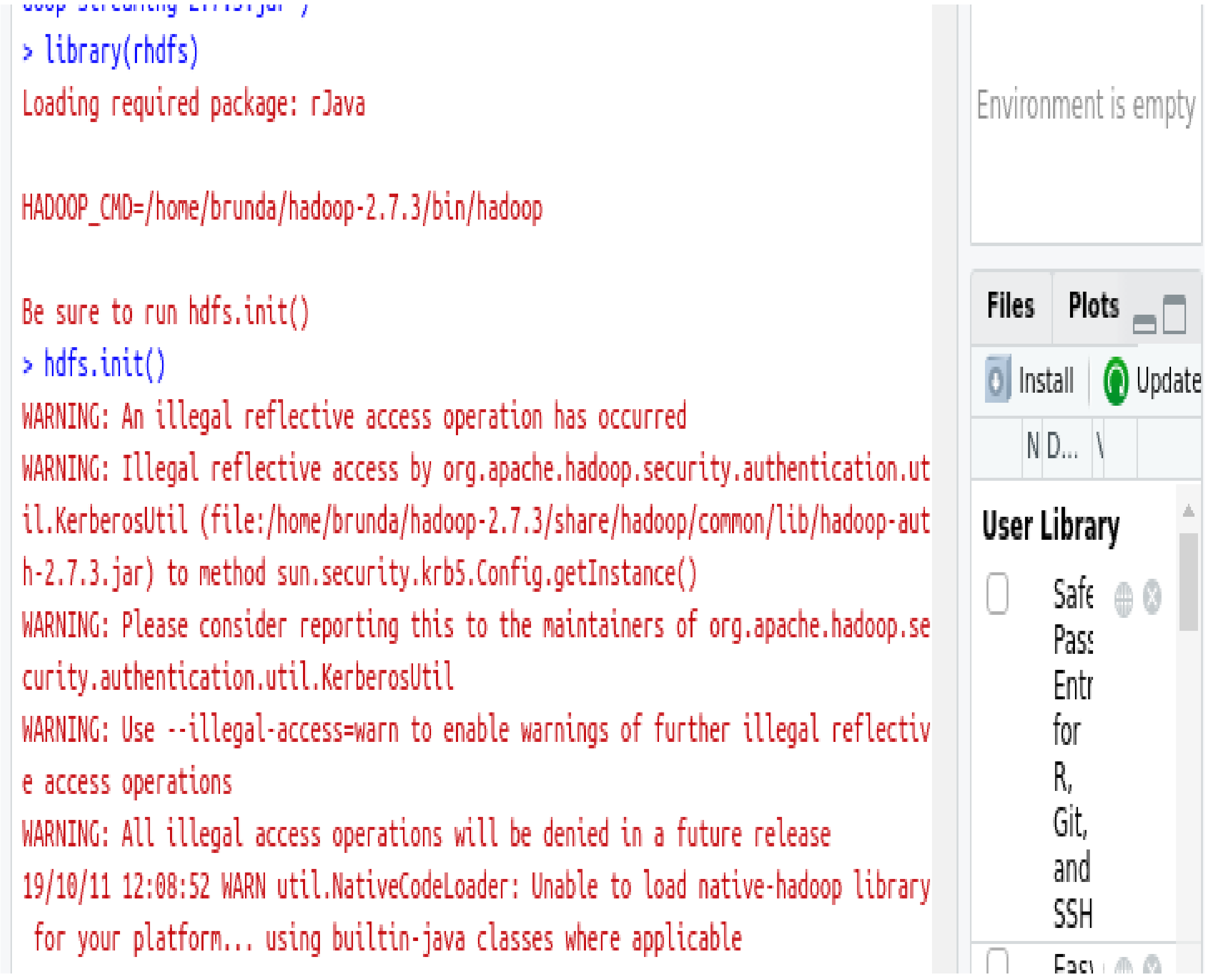
Login to R console

Set environment variables



```
>     install.packages("rmr2")
Installing package into ‘/home/brunda/R/x86_64-pc-linux-gnu-library/3.4’
(as ‘lib’ is unspecified)
Warning in install.packages :
  package ‘rmr2’ is not available (for R version 3.4.4)
> Sys.setenv(HADOOP_CMD='/home/brunda/hadoop-2.7.3/bin/hadoop')
> Sys.setenv(HADOOP_CONF='/home/brunda/hadoop-2.7.3/etc/hadoop')
> Sys.setenv(JAVA_HOME='/urs/lib/jvmjava-11-openjdk-amd64/jre')
> Sys.setenv(HADOOP_STREAMING='/home/brunda/hadoop-2.7.3/share/hadoop/tools/lib/ha
doop-streaming-2.7.3.jar')
> library(rhdfs)
Loading required package: rJava

HADOOP_CMD=/home/brunda/hadoop-2.7.3/bin/hadoop

Be sure to run hdfs.init()
> hdfs.init()
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.ut
il.KerberosUtil (file:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/hadoop-aut
h-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.se
curity.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflectiv
e access operations
WARNING: All illegal access operations will be denied in a future release
19/10/11 12:08:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library
```

Load the required packages rhdfs

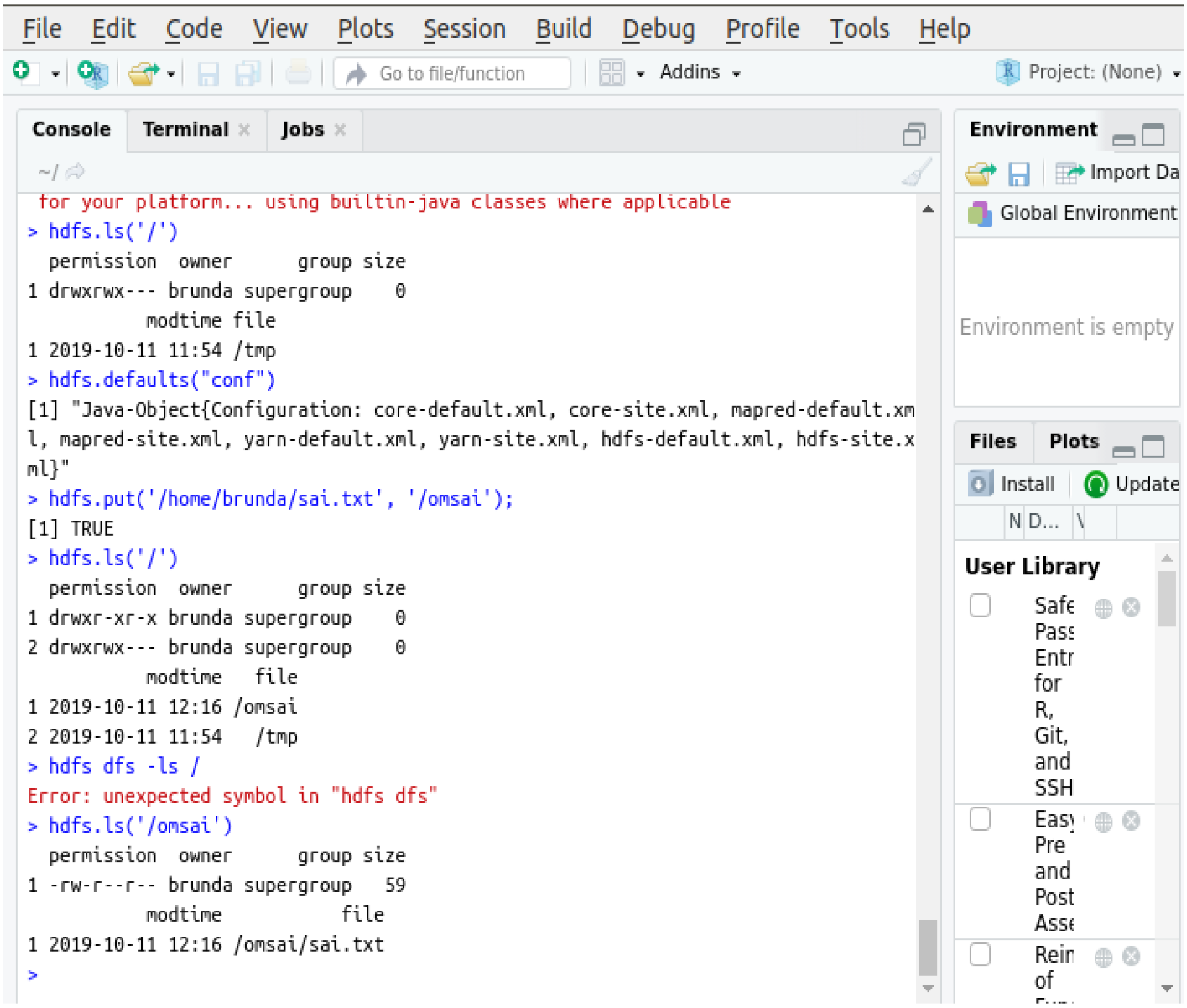


```
adoop-streaming-2.7.3.jar")
> library(rhdfs)
Loading required package: rJava

HADOOP_CMD=/home/brunda/hadoop-2.7.3/bin/hadoop

Be sure to run hdfs.init()
> hdfs.init()
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.ut
il.KerberosUtil (file:/home/brunda/hadoop-2.7.3/share/hadoop/common/lib/hadoop-aut
h-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.se
curity.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflectiv
e access operations
WARNING: All illegal access operations will be denied in a future release
19/10/11 12:08:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library
 for your platform... using builtin-java classes where applicable
```

After loading the rhdfs package we should initiate the connection using **hdfs.init()**
Accessing HDFS through R console

Listing the file in hdfs root directory

```
hdfs.ls('/')
```



```
 for your platform... using builtin-java classes where applicable
> hdfs.ls('/')
  permission  owner      group size
1 drwxrwx--- brunda supergroup    0
           modtime file
1 2019-10-11 11:54 /tmp
> hdfs.defaults("conf")
[1] "Java-Object{Configuration: core-default.xml, core-site.xml, mapred-default.xm
l, mapred-site.xml, yarn-default.xml, yarn-site.xml, hdfs-default.xml, hdfs-site.x
ml}"
> hdfs.put('/home/brunda/sai.txt', '/omsai');
[1] TRUE
> hdfs.ls('/')
  permission  owner      group size
1 drwxr-xr-x brunda supergroup    0
2 drwxrwx--- brunda supergroup    0
           modtime   file
1 2019-10-11 12:16 /omsai
2 2019-10-11 11:54   /tmp
> hdfs dfs -ls /
Error: unexpected symbol in "hdfs dfs"
> hdfs.ls('/omsai')
  permission  owner      group size
1 -rw-r--r-- brunda supergroup   59
           modtime          file
1 2019-10-11 12:16 /omsai/sai.txt
>
```

To get the HDFS default configurations used for this connection use

```
hdfs.defaults("conf")
```

File manipulation

° hdfs.put: This is used to copy files from the local filesystem to the HDFS filesystem.

```
hdfs.put('localfile source','hdfs destination')
```

hdfs.mkdir: used to create new directory in hdfs:

```
hdfs.mkdir('/new_dir')
```

hdfs.move: This is used to move a file from one HDFS directory to another HDFS directory.

```
hdfs.move('/test_file','/new_dir/')
```

hdfs.rename: This is used to rename the file stored at HDFS from R.

```
hdfs.rename('/new_dir/test_file','/new_dir/test_file1')
```

◦ hdfs.chmod: This is used to change permissions of some files.

```
hdfs.chmod('/Wc.txt', permissions= '777')
```

hdfs.delete: This is used to delete the HDFS file or directory from R.

```
hdfs.delete("/RHadoop")
```