

Installation of R, RStudio, and Packages for RHadoop

1- Install R and RStudio

Software	R and RStudio
Version*	R 3.3.1 RStudio Server 1.0.44
Download link(s)	Use the provided command in the tutorial
File size	R: 65 MB RStudio Server: 55 MB
Install size	Variable
Requirements	• Free disk space. 250 MB • RAM. 1 GB required, 2 GB recommended • Ubuntu version 12.04 or higher
* This version is used in this tutorial	

This section describes the installation process for R and RStudio on Ubuntu. To develop this guideline, I used installation codes on Cran R for Rbase from this [link](#). Also, I used the codes on Cran Rstudio for Rstudio Server from this [link](#).

To install the latest version of R package, CRAN repository should be added to the system. Use the following code for this purpose:

```
~$ sudo sh -c 'echo "deb http://cran.cnr.berkeley.edu /bin/linux/ubuntu xenial/" >>
/etc/apt/sources.list'
```

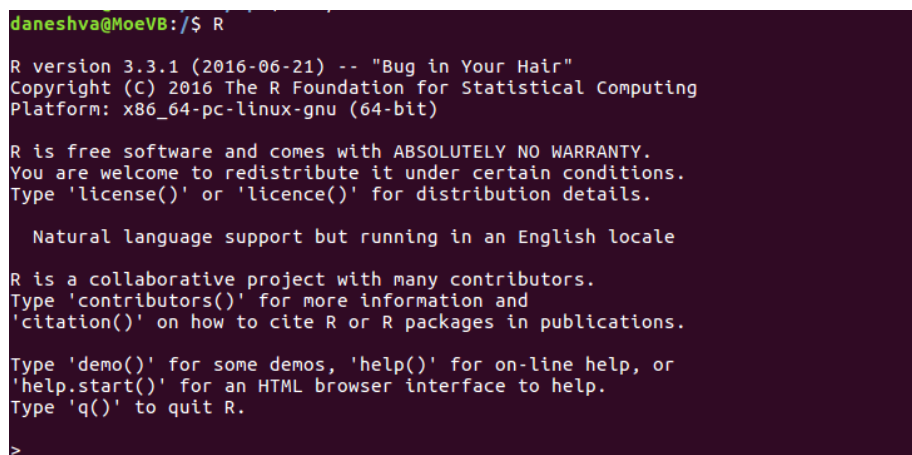
You can replace the mirror address with an address from [this list](#) that is closer to your country.

Please make sure that you install R with your hduser (or any other user that you identified for Hadoop). Use the following commands to install the complete R system and compile R packages from the source.

```
~$ sudo apt-get update
~$ sudo apt-get install r-base
~$ sudo apt-get install r-base-dev
```

At this point the installation is complete and you can run R with the following command. The output of this command is shown in Fig 1. To quit R, type q() and select to save/or not as prompted.

```
~$ R
```



```
daneshva@MoeVB:/$ R
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Fig 1. R environment in Ubuntu

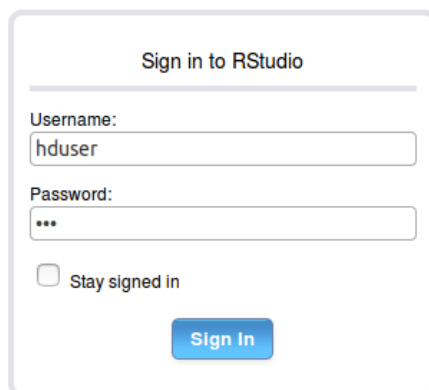
To create a better user interface for R, install RStudio Server. The prerequisite for installation of RStudio Server is installation of R. To install RStudio Server, use the following commands (for 64 bits version). The commands required for installation of 32 bits version is presented in this [link](#).

```
~$ sudo apt-get install gdebi-core
~$ wget https://download2.rstudio.org/rstudio-server-1.0.44-amd64.deb
~$ sudo gdebi rstudio-server-1.0.44-amd64.deb
```

The RStudio server is ready to use. To execute RStudio server, I used the <http://10.0.2.15:8787>. You need to replace the IP address (10.0.2.15) with your VM IP address before running the RStudio server. The following command is useful for identification of your VM IP address:

```
~$ ifconfig
```

Once you run the program on browser, you need to login with your hduser (Fig 2).

A login form titled "Sign in to RStudio". It contains two input fields: "Username:" with the text "hduser" entered, and "Password:" with three dots "..." entered. Below these fields is a checkbox labeled "Stay signed in" which is unchecked. At the bottom is a blue button labeled "Sign In".

Sign in to RStudio

Username:
hduser

Password:
...

☐ Stay signed in

Sign In

Fig 2. Log in with your credentials

After log in, you will see the interface shown in Fig 3.

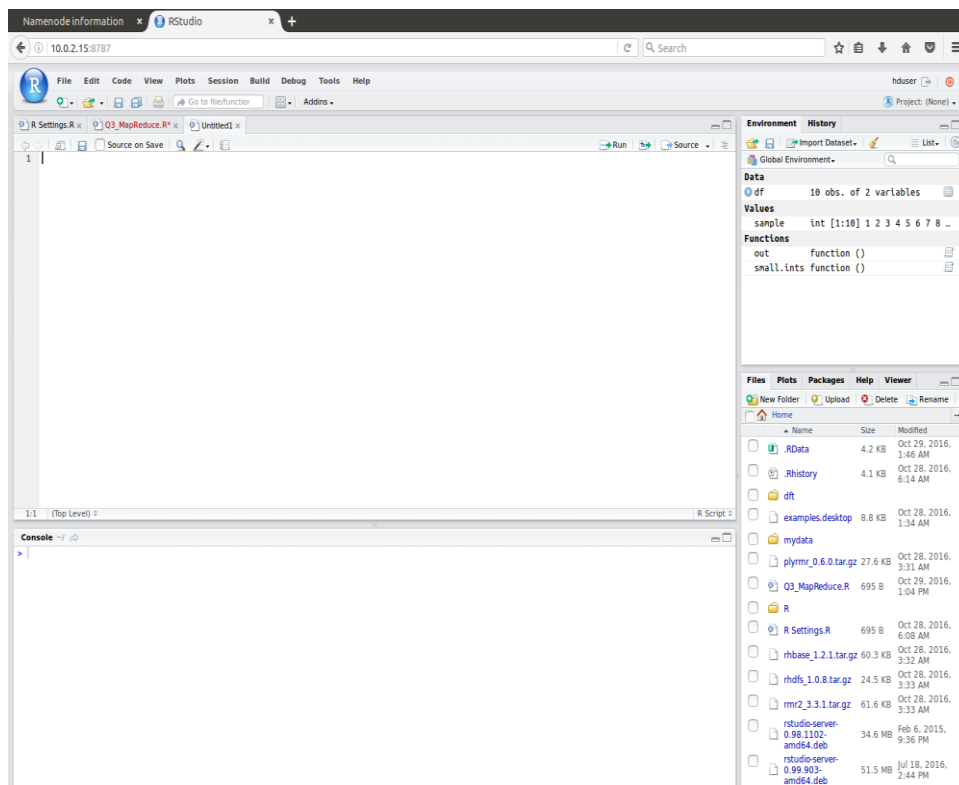


Fig 3. The interface of RStudio Server

You can write and execute any R script in both R and Rstudio Server.

2- Install Packages for RHadoop

Software	RHadoop Packages
Version*	plyrmr_0.6.0 rmr2_3.3.1 rhdfs_1.0.8 rhbase_1.2.1
Download link(s)	https://github.com/RevolutionAnalytics/RHadoop/wiki
File size	178 KB
Requirements	• Install with root privileges • Need other R packages
* This version is used in this tutorial	

RHadoop helps in an integrated interaction of R with Hadoop. RHadoop is a collection of R packages that enable R to use Hadoop data management. These packages are:

- 1- plyrmr
- 2- rmr
- 3- rhdfs
- 4- rhbase

The plyrmr package is a data processing tool. The MapReduce service is provided by rmr package. rhdfs enables working with the Hadoop file management system. Finally, rhbase is the means of database management.

Therefore, to enable R to work with data files that are being manage by Hadoop, we need to install these packages. To proceed to installation, please consider to use root user. Since packages should not be installed on private directories, root user is the right user with enough privileges to install these packages in R. Therefore, to start the process, we need to use sudo to install packages. Be aware that installing the packages from R environment might cause permission issues. Therefore, we install packages in Terminal. Prior to installation of main packages, install the following packages in R environment.

```
> install.packages(c("Rcpp", "RJSONIO", "bitops", "digest", "functional", "stringr", "plyr",  
  "reshape2", "dplyr", "R.methodsS3", "caTools", "Hmisc"))
```

The next step is to install rJava through following commands in Terminal.

```
~$ sudo apt-get install oracle-java9-installer  
~$ sudo apt-get install openjdk-9-jdk  
~$ sudo apt-get install liblzma-dev  
~$ sudo apt-get install r-cran-rjava
```

Then, download the rhdfs, rhbase, rmr2, and plyrmr from [here](#) and install them. To install the downloaded packages, move to the directory that the packages are downloaded in. In the following commands, replace daneshva (my username) with yours.

```
~$ cd /  
~$ cd home/daneshva/Downloads  
~$ HADOOP_CMD="/usr/local/hadoop/bin/hadoop"  
~$ sudo R CMD INSTALL plyrmr_0.6.0.tar.gz  
~$ sudo R CMD INSTALL rmr2_3.3.1.tar.gz  
~$ sudo R CMD INSTALL rhdfs_1.0.8.tar.gz  
~$ sudo R CMD INSTALL rhbase_1.2.1.tar.gz
```

Depending on the version of the packages, you might get error messages indicating that you need to install a specific package.

To run any program in RHadoop, one must ensure that appropriate user with enough permissions initiated the Hadoop (on terminal through commands stard-dfs.sh and start-yarn.sh).

Now, the RHadoop is ready to use. I run a simple MapReduce code in the RStudio Server to test the installation of the program. The program is as follows. The results are shown in Fig 4.

```

> Sys.setenv(HADOOP_HOME="/usr/local/hadoop")
> Sys.setenv(HADOOP_CMD="/usr/local/hadoop/bin/hadoop")
> Sys.setenv(HADOOP_STREAMING="/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-
  2.7.1.jar")
> library(rJava)
> library(rhdfs)
> hdfs.init()
> library(rmr2)

> sample<-1:10
> small.ints<-to.dfs(sample)
> out<-mapreduce(input = small.ints, map=function(k,v) keyval(v,v^2))
> from.dfs(out)
> df<-as.data.frame(from.dfs(out))
> print(df)

```

```

Console ~/
mapreduce: Spilled Bytes=0
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=312
CPU time spent (ms)=1660
Physical memory (bytes) snapshot=220708864
Virtual memory (bytes) snapshot=4189093888
Total committed heap usage (bytes)=125960192

File Input Format Counters
  Bytes Read=793
File Output Format Counters
  Bytes Written=2008
16/11/04 02:57:58 INFO streaming.StreamJob: Output directory: /tmp/file1b3132d958df
> from.dfs(out)
$key
[1] 1 2 3 4 5 6 7 8 9 10

$val
[1] 1 4 9 16 25 36 49 64 81 100

> df<-as.data.frame(from.dfs(out))
> print(df)
  key val
1    1   1
2    2   4
3    3   9
4    4  16
5    5  25
6    6  36
7    7  49
8    8  64
9    9  81
10  10 100
> |

```

Fig 4. The results of MapReduce Code