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](https://linuxconfig.org/how-to-open-a-terminal-on-ubuntu-bionic-beaver-18-04-linux) 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](https://linuxconfig.org/install-deb-file-on-ubuntu-18-04-bionic-beaver-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](https://www.rstudio.com/products/rstudio/download/" \l "download" \t "_blank) 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](https://linuxconfig.org/linux-commands" \t "_blank):

$ 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/')

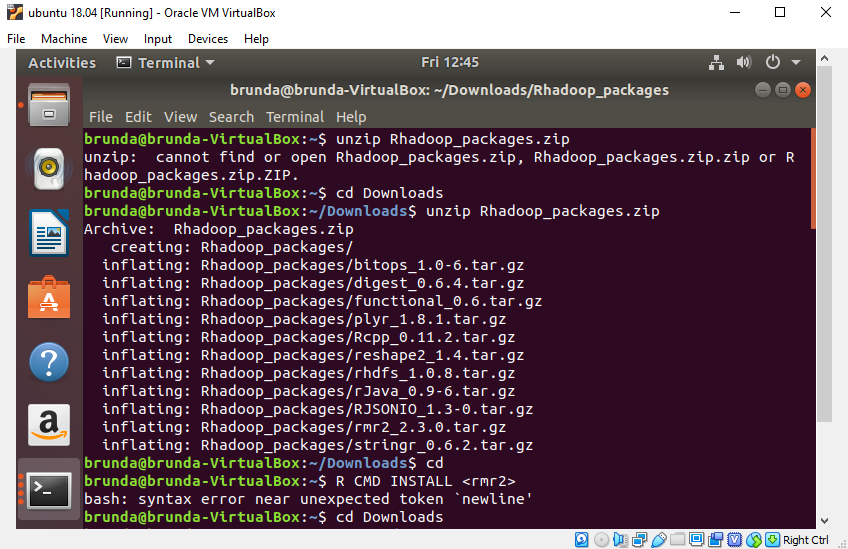
**Note:** Before installing rJava, we should set the JAVA\_HOME path and should login to R with sudo privileges.

**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

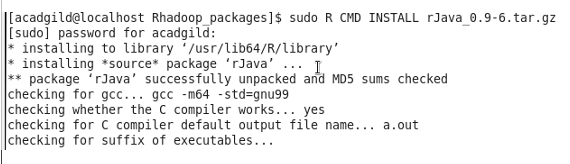


To install these packages, we will be using R cmd.

R CMD INSTALL <package name>

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

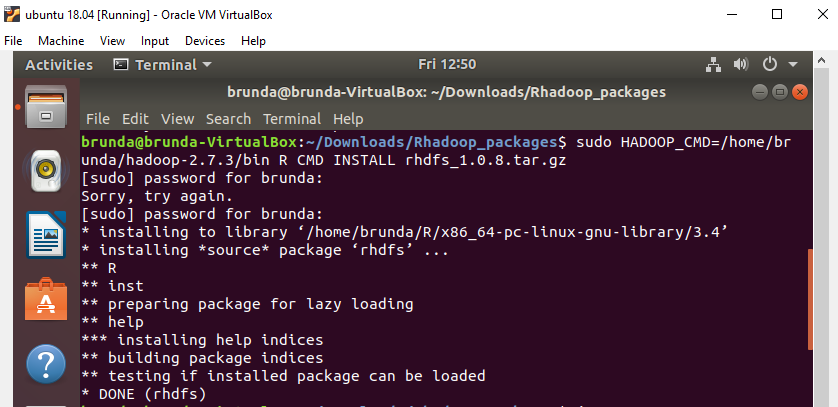
sudo R CMD INSTALL rJava\_0.9-6.tar.gz



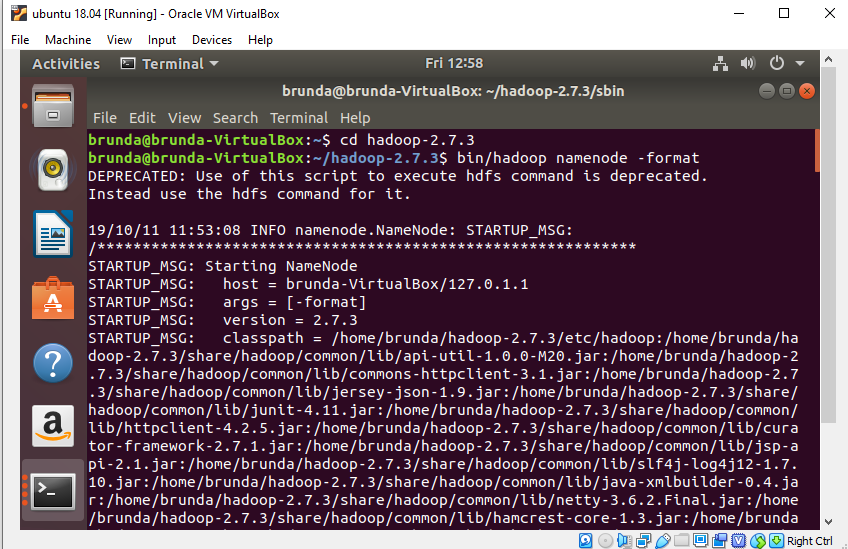
We need to follow the same command to install all the other required packages

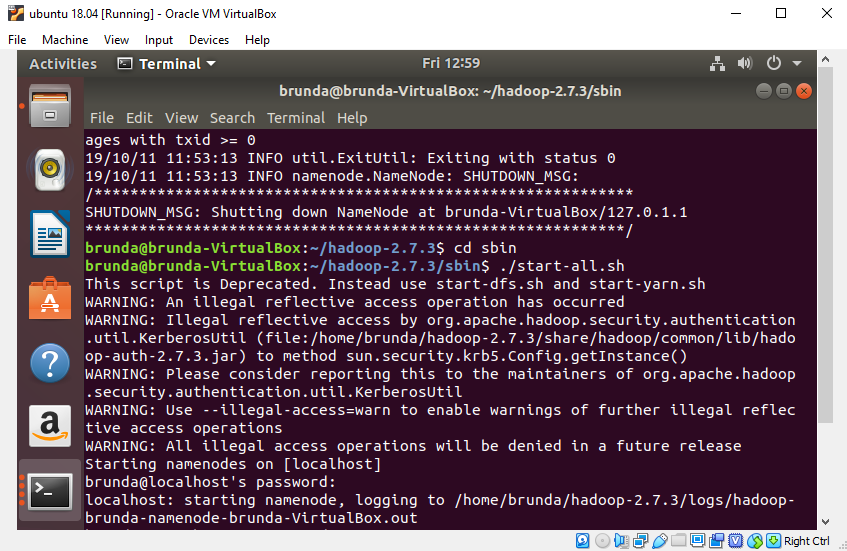
**RHDFS INSTALLATION IN RSTUDIO**

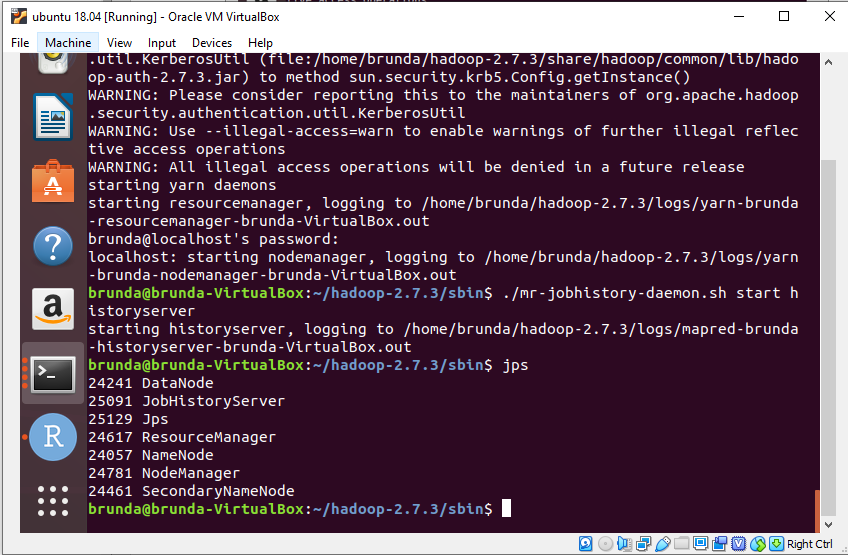
Before installing rhdfs we should set HADOOP\_CMD environmental variable. You can refer to the below screen shot to follow the steps for **Installing Rhdfs.**



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







Steps to start the Hadoop demon

$ cd hadoop-2.7.3

$ bin/hadoop namenode –format

$ cd sbin

$ ./start-all.sh

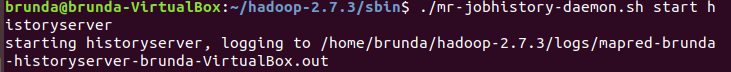
All the nodes will be started accordingly it will ask for the password

The demons started are namely:

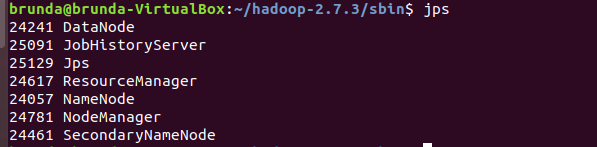
* Namenode
* Datanode
* Resouremanager
* Nodemanager

Need to start the job history server

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

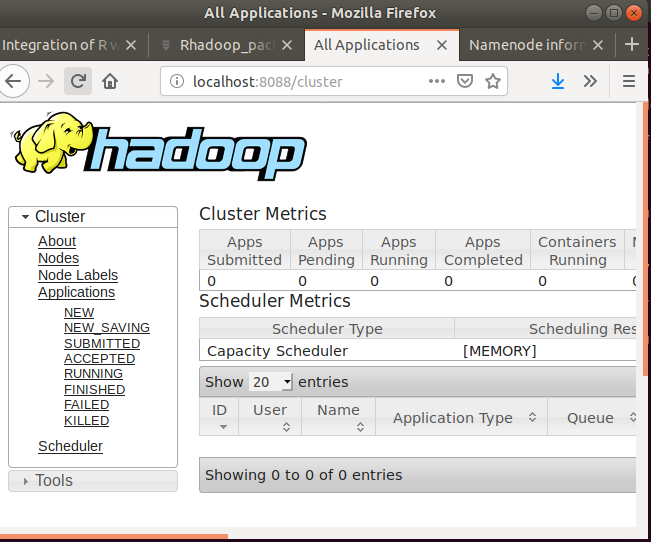


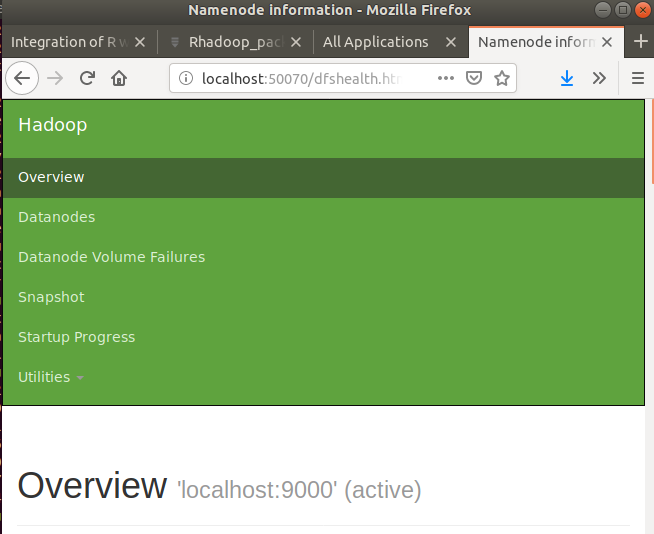
To check all the Hadoop services are up and running

$ jps 

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

Localhost:8088 localhost:50070



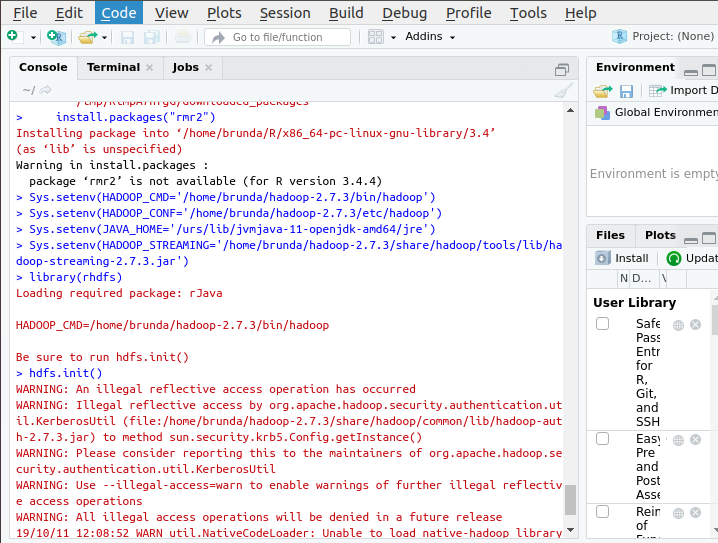


**RHADOOP INTEGRATION**

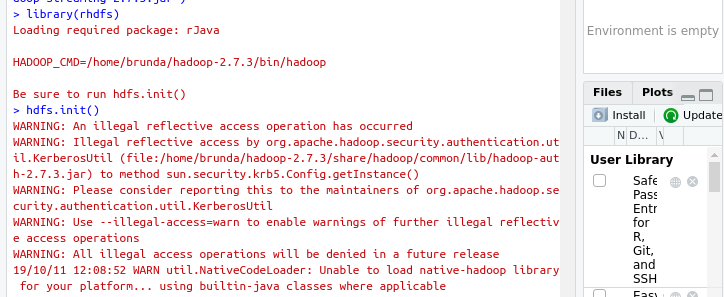
**Now we will access HDFS from the R console**

Login to R console

Set environment variables



Load the required packages rhdfs

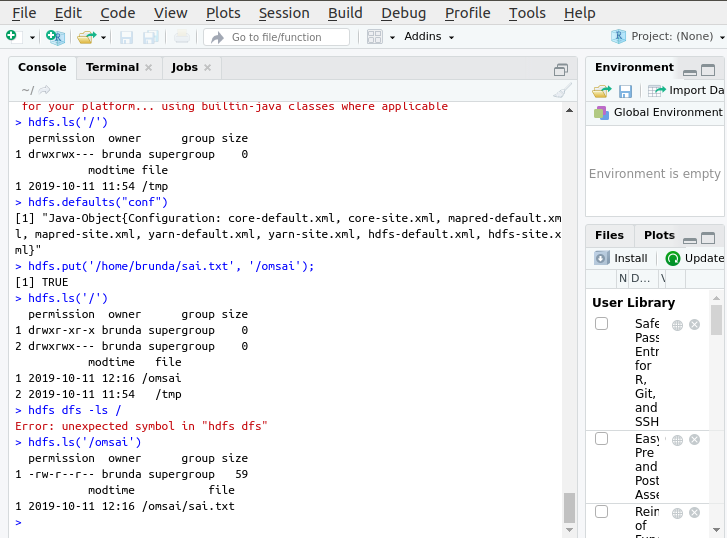


After loading the rhdfs package we should initiate the connection using**hdfs.init()**

Accessing HDFS through R console

Listing the file in hdfs root directory

hdfs.ls('/')



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")