**Deployment Steps:**

One of the Goal is to make R to Run on Hadoop cluster and implement Regression analysis with R Studio Server. In this way I am going to analyze the Data and work on distributed system via parallel processing with Hadoop file system database.

Reason is make R code more scalable in multiple nodes in Cloud environment and available on many nodes across clusters. Here I will use R language instead of default Java to implement Map-Reduce jobs (Key-Value pairs) in Hadoop.

I will be using R functions to deploy Key Value pairs through R on Hadoop such as Mymap, Myreduce, Mapreduce.

Using Hadoop frame work, I will allow parallel processing of massive amount of data and will perform Horizontal Scalability with Statistical calculations

Usage of Packages and Technology used:

RHadoop Packages (Developed by Revolution Analytics)

* rmr2 : R Map-Reduce2 Interface
* rhbase: Interface Read/ Write of data from HBase tables
* rhdfs: Interface to work on hdfs using R
* I will be using R functions to deploy Key Value pairs through R on Hadoop such as

Mymap, Myreduce, Mapreduce job roles.

By this I can design API, flexible I/O subsystem, chains of mapreduce jobs and capable to store files in Local backend system for quick queries or prototyping.

**System Requirements:**

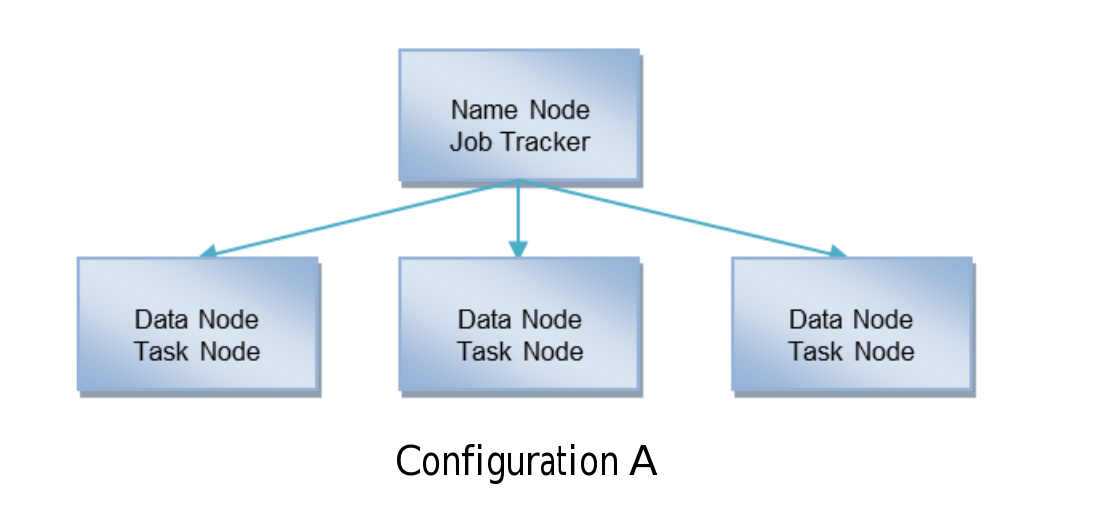
Operating System: Red Hat Enterprise Linux 5.4 and above with 64 bit processor.

Hadoop Cluster: Cloudera CDH3 or CDH4 with MapReduce Version2.

**R Hadoop Configuration requires three steps:**

The basic configuration can be done by any of the two architectures. I have used configuration A, however the reference of this configuration is taken from the following link:

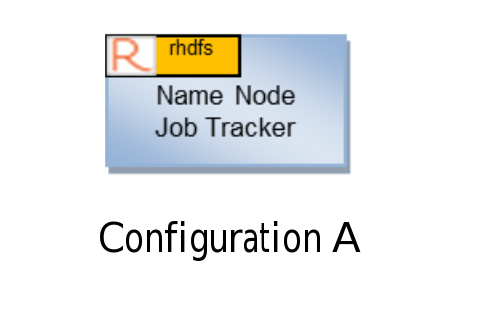
<https://s3.amazonaws.com/RHadoop/RHadoop2.0.2u2_Installation_Configuration_for_RedHat.pdf>



The above diagram shows the high level architecture of the Nodes which needs to be applied on Hadoop cluster while installation of RHadoop.

**Following are the steps to install rhdfs:**

Note: All Installations are done on Root directory.



First is the rhdfs configuration on the cluster. The following are steps and packages needs to deployed.

Package name: rJava

Command: R CMD INSTALL rJava\_0.9-4.tar.gz

Note: JDK has to be installed and make sure Java\_Home is pointing to the Java Jdk.

Command: R CMD javareconf (From root)

Step 2: Update the environment variable, which is needed by rhdfs to access.

Command: export HADOOP\_CMD =/usr/bin/Hadoop

Environment variable has to set on /etc/profile this will be available to all users.

Install rhfds

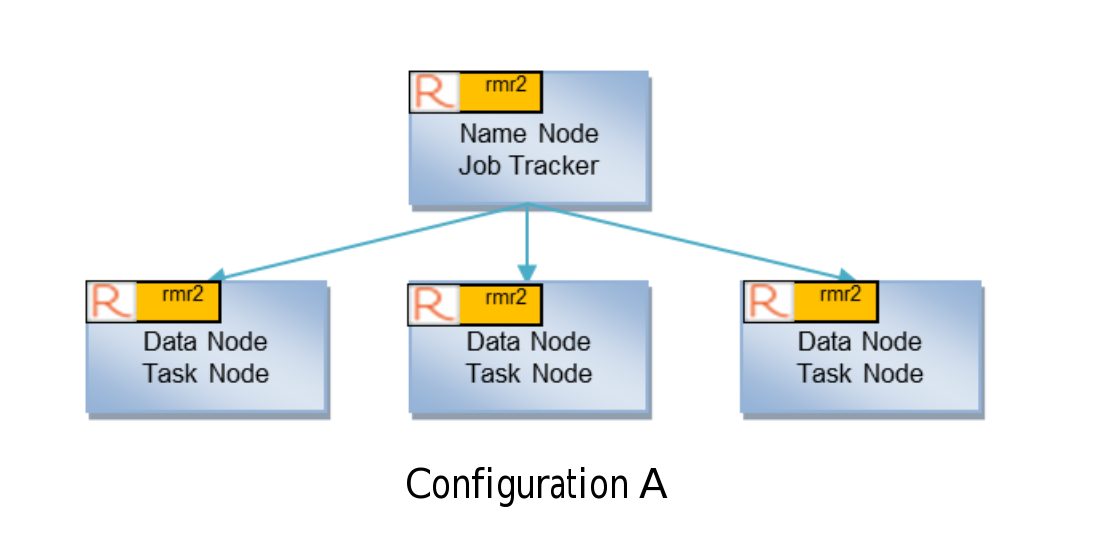
Command: R CMD INSTALL rhdfs\_1.2.1.tar.gz

**Following are the steps to install rmr2:**

Install the following files for rmr2 before Installation:

Revo-Ent-62.0-RHEL5.tar.gz or Revo-Ent-6.2.0-RHEL6.tar.gz

And R Hadoop-2.0.2u2.tar.gz



Above is the configuration of the Name Node and Data Node connection in Hadoop cluster.

Unpack the RHEL5 tar file

Command: tar –xzf Revo-Ent-6.2.0-RHEL.tar.gz

Change the directory to the versioned Revolution directory.

Command: cd RevolutionR\_6.2.0

Install Revolution R Enterprise

Command: ./install.py –no-ask -d

Go back to root by cd ..

Install R Hadoop bundle

Command: tar –xzf RHadoop-2.0.2u2.tar.gz

Install rmr2 and its denpendent R packages.

R CMD INSTALL digest\_0.6.3.tar.gz

R CMD INSTALL plyr\_1.8.tar.gz

R CMD INSTALL stringr\_0.6.2.tar.gz

R CMD INSTALL RJSONIO\_1.0-3.tar.gz

R CMD INSTALL Rcpp\_0.10.3.tar.gz

R CMD INSTALL functional\_0.1.tar.gz

R CMD INSTALL quickcheck\_1.0.tar.gz

Then last.

R CMD INSTALL rmr2\_2.0.2.tar.gz

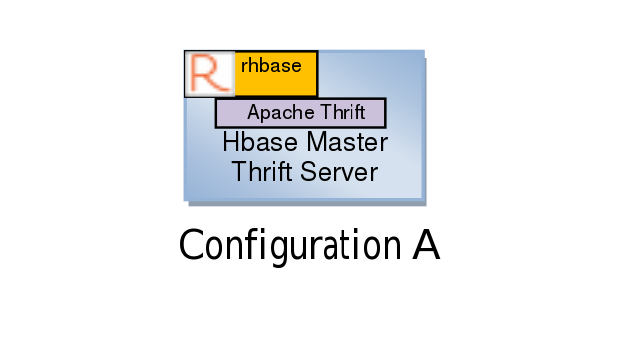
For Hadoop\_CMD and Hadoop\_Streaming I have exported the following files:

Command: export HADOOP\_CMD=/usr/bin/Hadoop

Next

Command: export HADOOP\_STRRAMING=/usr/lib/Hadoop/contrib/streaming/Hadoop-streaming-<version>.jar

Following are the steps to install rhbase (Which I got errors in this)



First install dependencies for Thrift:

Command: yum –y install automake libtool flex bison pkgconfig gcc-c++ boost-devel libevent-devel zlib-devel python-devel ruby-devel openssl-devel

Install the content of the Thrift

Command: tar –xzf thrift-0.8.0.tar.gz

Change the directory to the versioned thrift directory.

Command: cd thrift-0.8.0.tar.gz

Now Build the library. From the document he has used only C++ interface of Thrift, so he build with ruby or python.

Command: ./configure –without-ruby –without-python

Command: make

Install the Thrift library:

Command: make install

Command: ln –s /usr/local/lib/libthrift-0.8.0.so /usr/lib

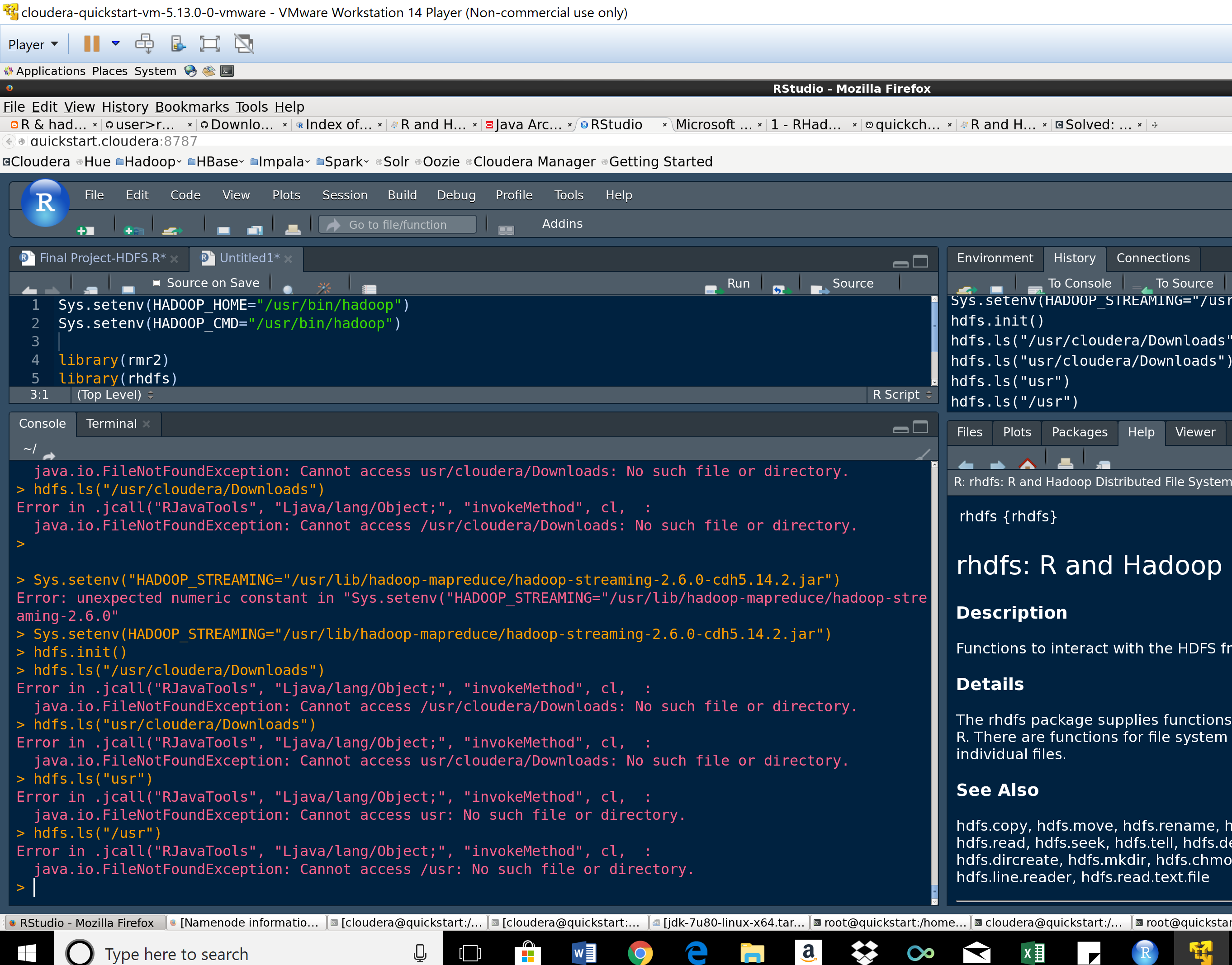
Setup the package PCK\_CONFIG\_PATH environmental variable.Command: export

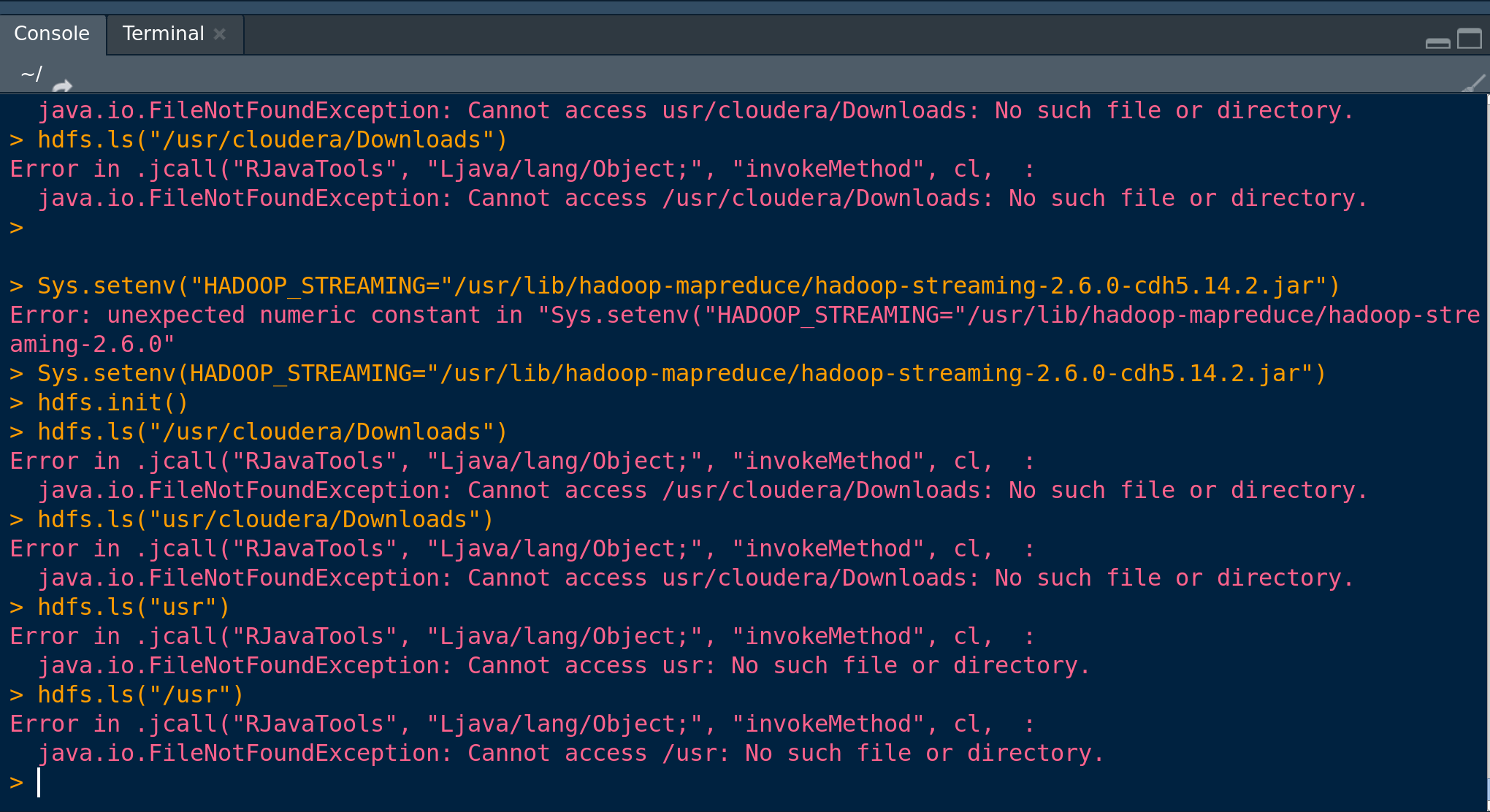
PKG\_CONFIG\_PATH=$PKG\_CONFIG\_PATH:/usr/local/lib/pkgconfig

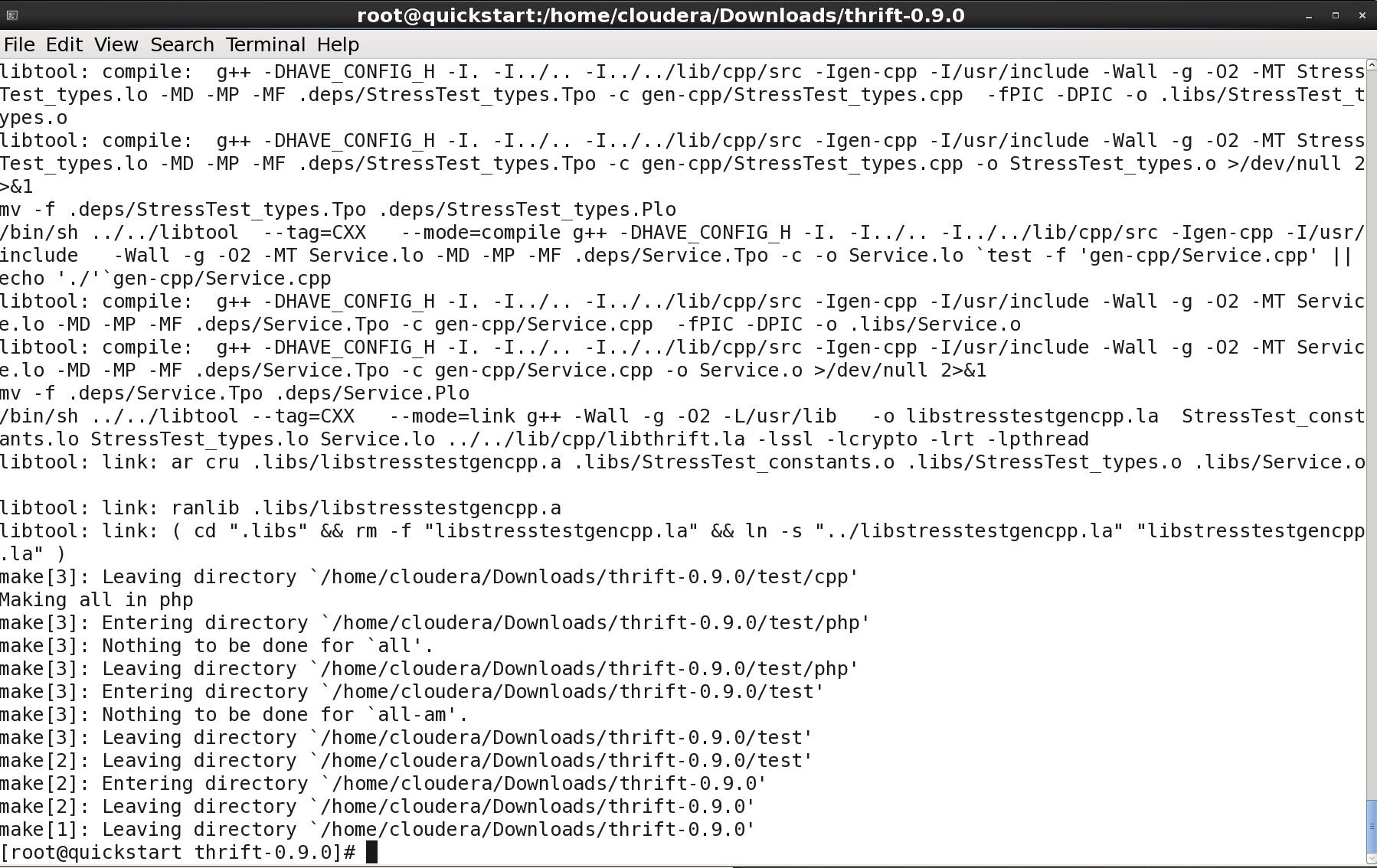
Last Install the rhbase package:

Command: R CMD INSTALL rhbase\_1.1.tar.gz

**The following are the screen shots of the error I got after installing the packages on R Studio Server while access the data on hdfs folder.**







To simplify the Storage and Management Huge amount of data, I have built Hadoop Cluster on a dedicated machine. This is the process I applied to make the to data access parallel across Hadoop.

References: Links and URLs

<https://s3.amazonaws.com/RHadoop/RHadoop2.0.2u2_Installation_Configuration_for_RedHat.pdf><https://github.com/kenahoo/RHadoop-Wiki>

<http://hsinay.blogspot.com/p/install-rhadoop-on-hortonworks-hdp-2.html>

<https://github.com/RevolutionAnalytics/RHadoop/wiki/user%3Ermr%3EHome>

<https://learnbigdataanalytics.blogspot.com/2016/06/how-to-install-rhbase-package-in-r.html>

<https://cwiki.apache.org/confluence/display/AMBARI/Installation+Guide+for+Ambari+2.6.0>

<https://stackoverflow.com/questions/29646482/how-to-install-rhadoop-packages-rmr-rhdfs-rhbase?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa>

<https://stat.ethz.ch/pipermail/r-help/2006-February/088003.html>

<https://coderwall.com/p/imxf6g/running-wordcount-on-hadoop-using-r-script>

<https://github.com/RevolutionAnalytics/RHadoop/wiki/user%3Erhdfs%3EHome>

<https://www.cyberciti.biz/tips/start-stop-vmware-virtualization-vms-command.html>

<https://stackoverflow.com/questions/29646482/how-to-install-rhadoop-packages-rmr-rhdfs-rhbase?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa>

Cloudera setup:

<http://community.cloudera.com/t5/CDH-Manual-Installation/Hadoop-streaming-jar-files-in-CDH5/td-p/15312>

Packages and Libraries:

<http://www.filewatcher.com/_/?q=rattle.tar.gz>

<https://archive.apache.org/dist/thrift/0.9.0/>

<https://github.com/RevolutionAnalytics/rhbase/blob/master/build/rhbase_1.2.0.tar.gz>

<https://github.com/RevolutionAnalytics/Rhadoop/wiki/user%3Erhbase%3EHome>

Cloudera:

<https://www.youtube.com/watch?v=4XBXJpYPkUk>

##################################### XXX ####################################