

## Experiment NO. 11

**Aim: To install HADOOP** (Advance Topic)

**Theory :-**

### What is Hadoop

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

### Modules of Hadoop

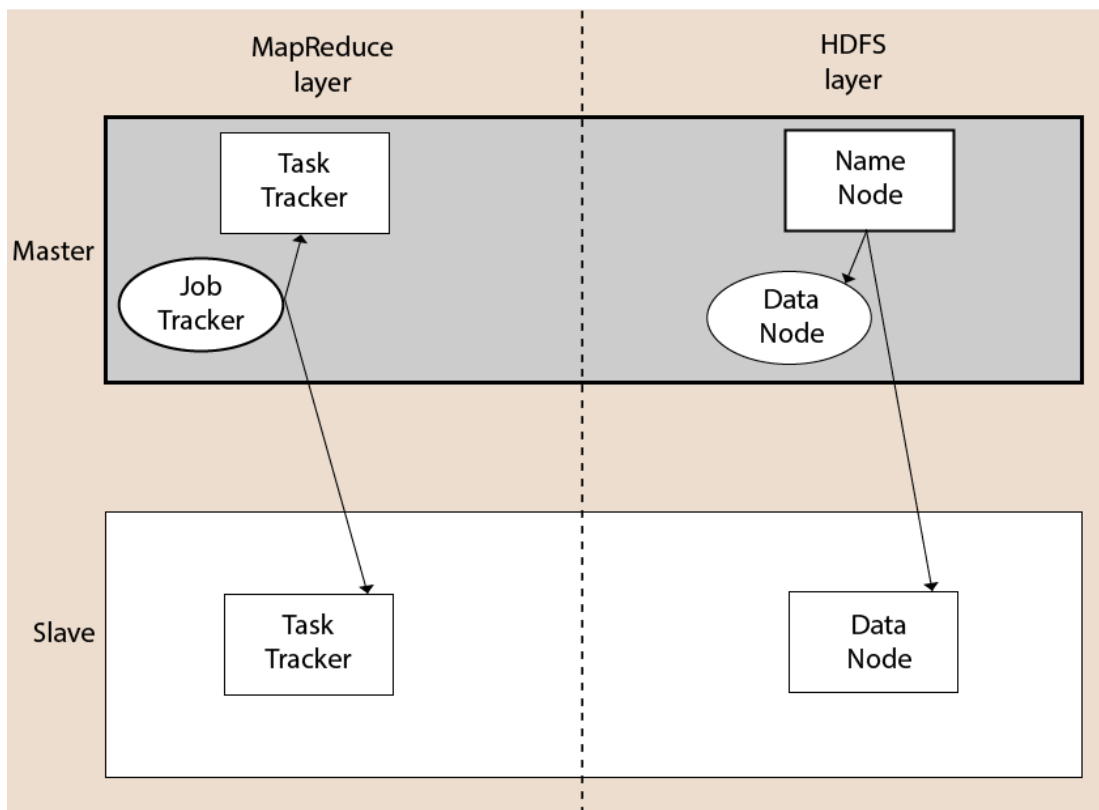
1. **HDFS:** Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
2. **Yarn:** Yet another Resource Negotiator is used for job scheduling and manage the cluster.
3. **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
4. **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

### Hadoop Architecture

The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.





#### Advantages of Hadoop

- **Fast:** In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
- **Scalable:** Hadoop cluster can be extended by just adding nodes in the cluster.
- **Cost Effective:** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
- **Resilient to failure:** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

#### Introduction to R and r studio

**R** is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accumulation of very specific and inflexible tools, as is frequently the case with other data analysis software. Many users think of R as a statistics system. It is preferable to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are several packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R is an implementation of the S programming language. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New

Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S. R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.

**RStudio** is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux). RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. JJ Allaire, creator of the programming language ColdFusion, founded RStudio. Hadley Wickham is the Chief Scientist at RStudio. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Windows, OS X, and Linux. RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface. Work on RStudio started at around December 2010, and the first public beta version (v0.92) was officially announced in February 2011.

## R Hadoop

The R Hadoop methods are the collection of packages. It contains three packages i.e., rmr, rhbase, and rhdfs.

### The rmr package

For the Hadoop framework, the `rmr` package provides MapReduce functionality by executing the Mapping and Reducing codes in R.

### The `rhbase` package

This package provides R database management capability with integration with HBASE.

### The `rhdfs` package

This package provides file management capabilities by integrating with HDFS.

## 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	<ul style="list-style-type: none"><li>• Free disk space. 250 MB</li><li>• RAM. 1 GB required, 2 GB recommended</li><li>• Ubuntu version 12.04 or higher</li></ul>
* 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.

>

```

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

```

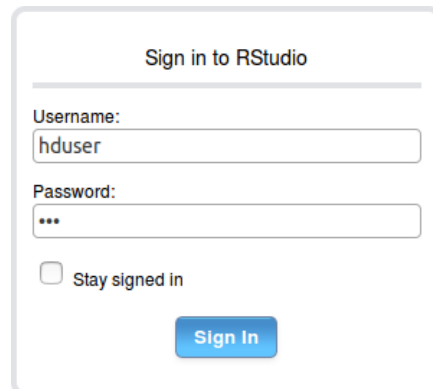
The RStudio server is ready to use. To execute RStudio server, I used the <http://10.0.2.15:8787>. You need to replace the IP address (10.0.2.15) with your VM IP address before running the RStudio server. The following command is useful for identification of your VM IP address:

```

~$ ifconfig

```

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



The image shows a web form titled "Sign in to RStudio". It contains two input fields: "Username:" with the text "hduser" entered, and "Password:" with three dots indicating a masked password. Below these fields is a checkbox labeled "Stay signed in" which is currently unchecked. At the bottom of the form is a blue button labeled "Sign In".

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

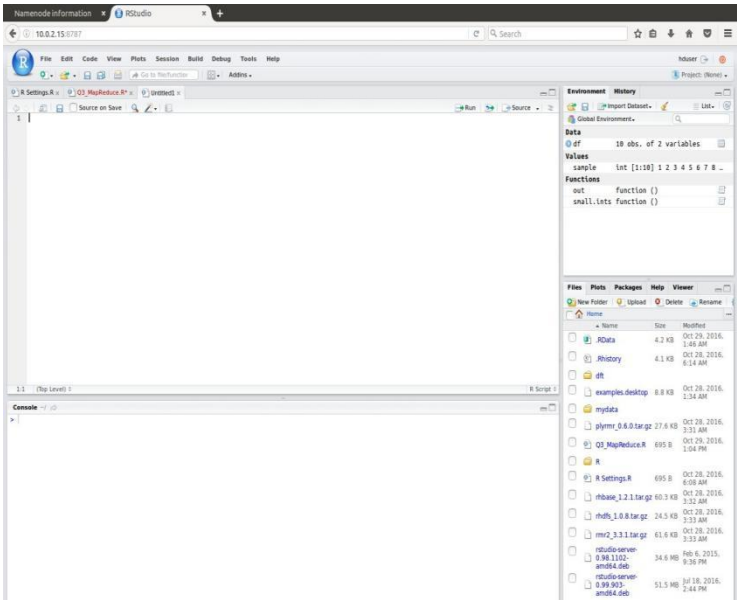


Fig 3. The interface of RStudio Server

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

## Install Packages for RHadoop

Software	<b>RHadoop Packages</b>
Version*	plyrmr_0.6.0 rmr2_3.3.1 rhdfs_1.0.8 rhbase_1.2.1
Download link(s)	<a href="https://github.com/RevolutionAnalytics/RHadoop/wiki">https://github.com/RevolutionAnalytics/RHadoop/wiki</a>
File size	178 KB
Requirements	<ul style="list-style-type: none"><li>• Install with root privileges</li><li>• Need other R packages</li></ul>
* 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 managed 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 `start-dfs.sh` and `start-yarn.sh`).  
Now, the RHadoop is ready to use.

Conclusion: Thus Hadoop is installed successfully.

#### VIVA QUESTIONS.

- 1) Explain big data and list its characteristics.
- 2). Explain Hadoop. List the core components of Hadoop
- 3) What is R-studio?
- 4) What is the need of R-studio?