REVOLUTION ANALYTICS

RHadoop
Installation Guide for
Red Hat® Enterprise Linux®

Version 2.0.2 – Update 2

Contents

Overview

RHadoop is a collection of three R packages that allow users to manage and analyze data with Hadoop.

Package	Description
rhdfs	Connectivity to the Hadoop Distributed File System (HDFS). R programmers can browse, read, write, and modify files stored in HDFS
rhbase	Connectivity to HBase distributed database. R programmers can browse, read, write, and modify tables stored in HBase
rmr2	Statistical Analysis using R, via MapReduce on a Hadoop Cluster

System Requirements

Before installing, verify that the machine on which you will install has the following:

- **Operating System**. Red Hat® Enterprise Linux® 5.4, 5.5, 5.6, 5.7,5.8, 6.0, 6.1,6.2,6.3 (64-bit processors);
- **Hadoop Cluster**. Cloudera CDH3, CDH4 (MapReduce Version 1), Hortonworks HDP 1.2, IBM BigInsights 2.0, Intel Hadoop 2.2

Software Dependencies

In order to properly install and configure the RHadoop, a set of supported software dependencies must be installed first

Dependency and Version	All nodes in the Hadoop cluster	Single node in the Hadoop Cluster
Revolution R Enterprise 6.2 (and all of its dependencies)	Yes. Make sure that you have first installed the package prerequisites covered in the **Revolution R Enterprise Installation Guide for Red Hat Enterprise Linux**.	
HBase**	Per the instructions for the Hadoop Distribution you are using	
Apache Thrift **(and all dependencies)		Yes, recommended on the node containing the HBase Master

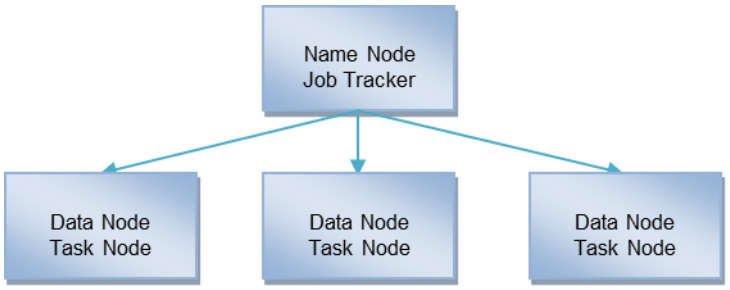
** Only needed if installing rhbase package

Typical Configurations for RHadoop

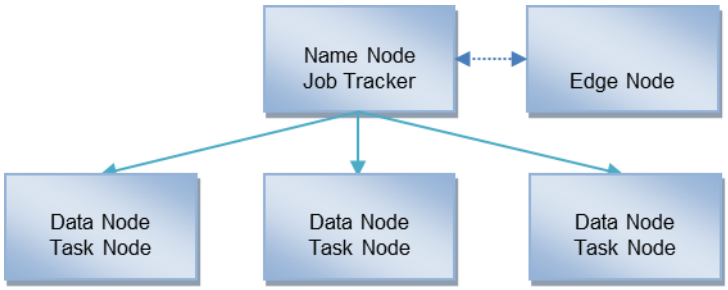
This section describes typical installations of RHadoop. Since Hadoop clusters can be configured in many different ways, this description should serve as a guide and not an absolute reference

Basic Hadoop Configuration

In the most basic Hadoop configuration, you'll have a topology that looks something one the following:



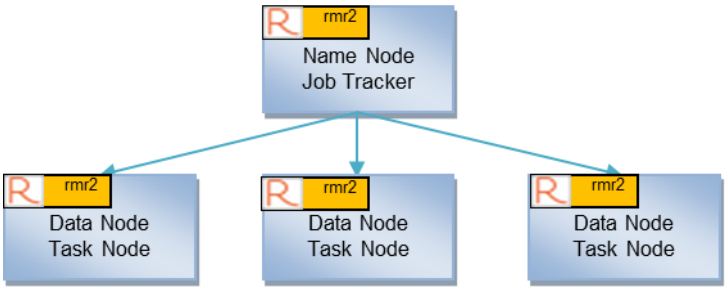
Configuration A



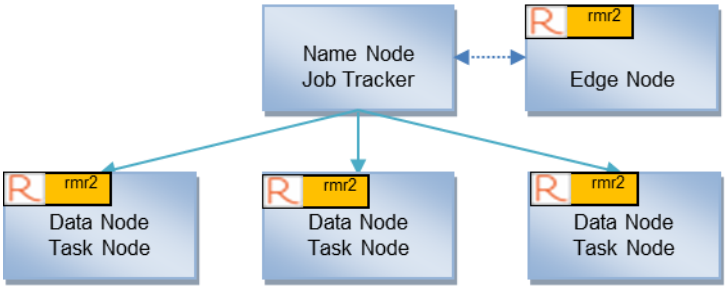
Configuration B

Basic Hadoop Configuration + Revolution R Enterprise + rmr2 package

The rmr2 package enables MapReduce jobs coded in R to be executed on the Hadoop cluster. In order for those MapReduce jobs to execute, Revolution R Enterprise and the rmr2 package (including its dependencies) must be installed on each Task node of the Hadoop cluster. If you are using an Edge node, Revolution R Enterprise and rmr2 can be installed on that node instead of the Name node (An Edge node has all of the Hadoop jar files, configuration files and network connectivity and access to the other nodes in the cluster).



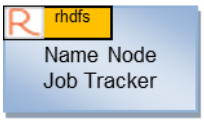
Configuration A



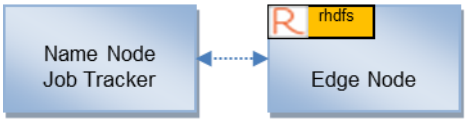
Configuration B

Basic Hadoop Configuration + Revolution R Enterprise + rhdfs package

The rhdfs package provides connectivity to HDFS. The package plus Revolution R Enterprise must be installed on the Name Node, a Data Node or an Edge Node in the Hadoop cluster.



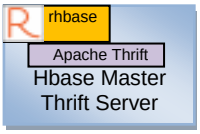
Configuration A



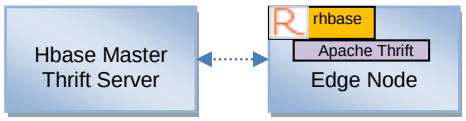
Configuration B

Basic Hadoop Configuration + Revolution R Enterprise + rhbase package

The rhbase package provides connectivity to HBase. The package plus Revolution R Enterprise must be installed on the node that has access to the HBase master. The package uses the Thrift API to communicate with HBase, so the Apache Thrift server must also be installed on the same node.



Configuration A



Configuration B

Installation

The RHadoop packages can be installed either manually or via a shell script. Both methods are described in this section. However, the commands listed in the shell script are to be used for **guidance only**, and should be adapted to standards of your IT department.

Installing as Root vs. Non-Root

Installing as root is the preferred method. There are a number of system dependencies that need to be installed for both Revolution R Enterprise and for Apache Thrift. If you install as non-root, you will need to have a system administrator available to install these dependencies for you. Also, because MapReduce jobs run as their own user, at least one symbolic link will need to be created for non-root installs, so Revolution R Enterprise can be invoked during the MapReduce jobs. Finally, there are some environment variables that need configured and you will have choices on how those are to be setup.

Installation Revolution R Enterprise and rmr2 on all nodes

Note: Make sure that you have installed the package prerequisites covered in the **Revolution R Enterprise Installation Guide for Red Hat Enterprise Linux** before installing Revolution R Enterprise.

1. Use the link(s) in your Revolution Analytics welcome letter to download the following installation files.

   - Revo-Ent-6.2.0-RHEL5.tar.gz or Revo-Ent-6.2.0-RHEL6.tar.gz
   - RHadoop-2.0.2u2.tar.gz

2. Unpack the contents of the Revolution R Enterprise installation bundle. At the prompt, type:

```
tar -xzf Revo-Ent-6.2.0-RHEL5.tar.gz
```

   Note: If installing on Red Hat Enterprise Linux 5.x, replace RHEL6 with RHEL5 in the previous tar command.

3. Change directory to the versioned Revolution directory. At the prompt, type:

```
cd RevolutionR_6.2.0
```

4. Install Revolution R Enterprise. At the prompt, type:

```
./install.py --no-ask -d
```

   Important! If installing as **NON-ROOT**, a list of missing system dependencies will be printed to the console. You must install any missing system dependencies before you continue.

   Important! If installing as **NON-ROOT**, you will need to do the following:

   - Add the location of the R executable to the PATH, or create a symbolic link from R to a location that is in the PATH. Example of symbolic link:

```
ln -s /home/users/<user>/local/bin/R /usr/bin
```

   - Because rmr2 uses the Rscript executable, and MapReduce jobs typically run as their own user, you'll need to create a symbolic link from Rscript to a location that is in the PATH. Example of symbolic link:

```
ln -s /home/users/<user>/local/bin/Rscript /usr/bin
```

5. Unpack the contents of the RHadoop installation bundle. At the prompt, type:

```
cd ..
```

```
tar -xzf RHadoop-2.0.2u2.tar.gz
```

6. Change directory to the versioned RHadoop directory. At the prompt, type:

```
cd RHadoop_2.0.2
```

7. Install rmr2 and its dependent R packages. At the prompt, type:

```
R CMD INSTALL digest_0.6.3.tar.gz plyr_1.8.tar.gz
  stringr_0.6.2.tar.gz RJSONIO_1.0-3.tar.gz Rcpp_0.10.3.tar.gz
  functional_0.1.tar.gz quickcheck_1.0.tar.gz rmr2_2.0.2.tar.gz
```

8. Update the environment variables needed by rmr2. The values for the environments will depend upon your Hadoop distribution.

   HADOOP_CMD – The complete path to the "hadoop" executable

   HADOOP_STREAMING – The complete path to the Hadoop Streaming jar file

   Examples of both of these environment variables are shown below:

```
export HADOOP_CMD=/usr/bin/hadoop
```

```
export HADOOP_STREAMING=/usr/lib/hadoop/contrib/streaming/hadoop-
  streaming-<version>.jar
```

   Important! These environment variables only need to be set on the nodes that are invoking the rmr2 MapReduce jobs (i.e. an Edge node as described earlier in this document). If you don't know which nodes will be used, then set these variables on each node. Also, it is recommended to add these environment variables to the file **/etc/profile** so that they will be available to all users.

Using a script to Install Revolution R Enterprise and rmr2 on all nodes

This shell script example should be used as **guidance** for installing Revolution R Enterprise and rmr2 on a node in the Hadoop Cluster. ***This script sample assumes ROOT privileges to run.***

```
#make sure we are root
sudo su

#unpack Revolution R Enterprise bundle
tar -xzf Revo-Ent-6.2.0-RHEL5.tar.gz
cd RevolutionR_6.2.0

#install Revolution R in silent mode
./install.py --no-ask -d
cd ..

#unpack RHadoop bundle
tar -xzf RHadoop-2.0.2u2.tar.gz
cd RHadoop_2.0.2

#Install dependencies for rmr2
R CMD INSTALL digest_0.6.3.tar.gz plyr_1.8.tar.gz
  stringr_0.6.2.tar.gz RJSONIO_1.0-3.tar.gz Rcpp_0.10.3.tar.gz
  functional_0.1.tar.gz quickcheck_1.0.tar.gz rmr2_2.0.2.tar.gz

#set the environment variables
su << EOF1
cat >> /etc/profile <<EOF

export HADOOP_CMD=/usr/bin/hadoop
export HADOOP_STREAMING=/usr/lib/hadoop/contrib/streaming/hadoop-
  streaming-<version>.jar

EOF
EOF1
```

Installation of rhdfs

Important! If you plan on using the RevoScaleR HDFS connector in addition to rhdfs, please read **Chapter 8** of the **Revolution R Enterprise Installation Guide for Red Hat Enterprise Linux** before beginning this installation.

1. Install the rJava packages. At the prompt, type:

```
R CMD INSTALL rJava_0.9-4.tar.gz
```

Important! If the installation of rJava fails, you may need to configure R to run properly with Java. First check to be sure you have the Java JDK installed, and the environment variable JAVA_HOME is pointing to the Java JDK. To configure R to run with Java, type the command:

```
R CMD javareconf
```

After running this command, try installing rJava again. Also,.

2. Update the environment variable needed by rhdfs. The value for the environment variable will depend upon your hadoop distribution.

HADOOP_CMD – The complete path to the "hadoop" executable

An **example** of the environment variable is shown below:

```
export HADOOP_CMD=/usr/bin/hadoop
```

Important! This environment variable only needs to be set on the nodes that are using the rhdfs package (i.e. an Edge node as described earlier in this document). Also, it is recommended to add this environment variable to the file **/etc/profile** so that it will be available to all users.

3. Install rhdfs. At the prompt, type:

```
R CMD INSTALL rhdfs_1.0.5.tar.gz
```

Installation of rhbase

1. Install Apache Thrift

   Important! rhbase requires Apache Thrift Server. If you do not have thrift already configured and installed, you will need to build and install Apache Thrift. Reference web site: http://thrift.apache.org/

2. Install the dependencies for Thrift. At the prompt, type

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

   Important! If installing as **NON-ROOT**, you will need a system administrator to help install these dependencies.

3. Unpack the contents of the Thrift archive. At the prompt, type.

```
tar -xzf thrift-0.8.0.tar.gz
```

4. Change directory to the versioned Thrift directory. At the prompt, type

```
cd thrift-0.8.0
```

5. Build the thrift library. We only need the C++ interface of Thrift, so we build without ruby or python . At the prompt type the following two commands

```
./configure --without-ruby --without-python
```

```
make
```

6. Install the thrift library. At the prompt, type:

```
make install
```

   Important! If installing as **NON-ROOT**, this command will most likely require root privileges, and will have to be executed by your system administrator:

7. Create a symbolic link to the thrift library so it can be loaded by the rhbase package. Example of symbolic link:

```
ln -s /usr/local/lib/libthrift-0.8.0.so /usr/lib
```

   Important! If installing as **NON-ROOT**, you may need a system administrator to execute this command for you.

8. Setup the PKG_CONFIG_PATH environment variable. At the prompt, type :

```
export PKG_CONFIG_PATH=$PKG_CONFIG_PATH:/usr/local/lib/pkgconfig
```

9. Install the rhdfs package. At the prompt, type:

```
cd ..
```

```
R CMD INSTALL rhbase_1.1.tar.gz
```

Using a script to Install rhdfs and rhbase

This shell script example should be used as **guidance** for installing R and rmr2 on a node in the Hadoop Cluster. ***This script sample assumes ROOT privileges to run.***

```
#This script installs both rhdfs and rhbase

#make sure we are root
sudo su

#---
#-rhdfs portion of the install
#---

#install rJava
R CMD INSTALL rJava_0.9-4.tar.gz

#set the environment variables
export HADOOP_CMD=/usr/bin/hadoop

#install rhdfs
R CMD INSTALL rhdfs_1.0.5.tar.gz

#---
#-rhbase portion of the install
#---

#install dependencies for thrift
yum -y install automake libtool flex bison pkgconfig gcc-c++
  boost-devel libevent-devel zlib-devel python-devel ruby-devel
  openssl-devel

#unpack thrift
tar -xzf thrift-0.8.0.tar.gz
cd thrift-0.8.0

#build thrift (We only need the C++ interface,
#so build without ruby and python)
./configure --without-ruby --without-python
make
make install

#set the PKG_CONFIG_PATH (needed for building rhbase)
export
  PKG_CONFIG_PATH=$PKG_CONFIG_PATH:/usr/local/lib/pkgconfig/

#make sure the thrift library
#is in a place where it can be loaded
ln -s /usr/local/lib/libthrift-0.8.0.so /usr/lib

#install rhbase
cd ..
R CMD INSTALL rhbase_1.1.tar.gz
```

Testing to be sure the packages are configured and working

There are two sets of tests you should do to verify that your configuration is working. The first set of test will check that the installed packages can be loaded and initialized

1. Invoke R. At the prompt, type:

```
R
```

2. Load and initialize the **rmr2** package, and execute some simple commands

   At the R prompt, type the following commands: (Note: the ">" symbol in the following code is the 'R' prompt and should not be typed.)

```
> library(rmr2)
> from.dfs(to.dfs(1:100))
> from.dfs(mapreduce(to.dfs(1:100)))
```

   - If any errors occur check the following:

     a. Revolution R Enterprise is installed on each node in the cluster.

     b. Check that rmr2, and its dependent packages are installed on each node in the cluster.

     c. Make sure that a link to Rscript executable is in the PATH on each node in the Hadoop cluster.

     d. The user that invoked 'R' has read and write permissions to HDFS

     e. HADOOP_CMD environment variable is set, exported and its value is the complete path of the "hadoop" executable.

     f. HADOOP_STREAMING environment variable is set, exported and its value is the complete path to the Hadoop Streaming jar file.

     g. If you encounter errors like the following (see below), check the 'stderr' log file for the job, and resolve any errors reported. The easiest way to find the log files is to use the tracking URL (i.e. `http://<my_ip_address>:50030/jobdetails.jsp?jobid=job_201208162037_0011`)

```
12/08/24 21:21:16 INFO streaming.StreamJob: Running job:
job_201208162037_0011
12/08/24 21:21:16 INFO streaming.StreamJob: To kill this job, run:
12/08/24 21:21:16 INFO streaming.StreamJob: /usr/lib/hadoop-
0.20/bin/hadoop job  -Dmapred.job.tracker=<my_ip_address>:8021 -kill
job_201208162037_0011
```

```
12/08/24 21:21:16 INFO streaming.StreamJob: Tracking URL:
http://<my_ip_address>:50030/jobdetails.jsp?jobid=job_201208162037_00
11
12/08/24 21:21:17 INFO streaming.StreamJob:  map 0%  reduce 0%
12/08/24 21:21:23 INFO streaming.StreamJob:  map 50%  reduce 0%
12/08/24 21:21:31 INFO streaming.StreamJob:  map 50%  reduce 17%
12/08/24 21:21:45 INFO streaming.StreamJob:  map 100%  reduce 100%
12/08/24 21:21:45 INFO streaming.StreamJob: To kill this job, run:
12/08/24 21:21:45 INFO streaming.StreamJob: /usr/lib/hadoop-
0.20/bin/hadoop job  -Dmapred.job.tracker=<my_ip_address>:8021 -kill
job_201208162037_0011
12/08/24 21:21:45 INFO streaming.StreamJob: Tracking URL:
http://<my_ip_address>:50030/jobdetails.jsp?jobid=job_201208162037_00
11
12/08/24 21:21:45 ERROR streaming.StreamJob: Job not successful.
Error: NA
12/08/24 21:21:45 INFO streaming.StreamJob: killJob...
Streaming Command Failed!
Error in mr(map = map, reduce = reduce, combine = combine, in.folder
= if (is.list(input)) { :
hadoop streaming failed with error code 1
```

3. Load and initialize the **rhdfs** package.

At the R prompt, type the following commands: (Note: the ">" symbol in the following code is the 'R' prompt and should not be typed.)

```
> library(rhdfs)
> hdfs.init()
> hdfs.ls("/")
```

- If any error occurs check the following:

  a. rJava package is installed, configured and loaded.

  b. HADOOP_CMD is set and its value is set to the complete path of the "hadoop" executable, and exported.

4. Load and initialize the **rhbase** package.

At the R prompt, type the following commands: (Note: the ">" symbol in the following code is the 'R' prompt and should not be typed.)

```
> library(rhbase)
> hb.init()
> hb.list.tables()
```

- If any error occurs check the following:

a. Thrift Server is running (refer to your Hadoop documentation for more details)

b. The default port for the Thrift Server is 9090. Be sure there is not a port conflict with other running processes

c. Check to be sure you are not running the Thrift Server in `hsha` or `nonblocking` mode. If necessary use the `threadpool` command line parameter to start the server

(i.e. `/usr/bin/hbase thrift –threadpool start`)

5. Using the standard R mechanism for checking packages, you can verify that your configuration is working properly.

Go to the directory where the R package source (rmr2, rhdfs, rhbase) exist. Type the following commands for each package.

Important!: Be aware that running the tests for the rmr2 package may take a significant time (hours) to complete

```
R CMD check rmr2_2.0.2.tar.gz
```

```
R CMD check rhdfs_1.0.5.tar.gz
```

```
R CMD check rhbase_1.1.tar.gz
```

- If any error occurs, refer to the trouble shooting information in the previous sections:
- Note: errors referring to missing package **pdflatex** can be **ignored**

```
Error in texi2dvi("Rd2.tex", pdf = (out_ext == "pdf"), quiet =
FALSE,  :
pdflatex is not available
Error in running tools::texi2dvi
```