The data that we want to get could be in different places and in different formats. We will provide some examples of how you can get data from different sources.

Get Data from SQL

It is very common for the data to be stored in an SQL database. We have provided an extensive example of how you can connect R with SQL.

Get csv/text Data from HTTP(s) URL

We can easily get structured data like `csv` or `txt` files that are under an HTTP(S) URL. I have created a public `S3` bucket where I stored some dummy data called `movie_metadata.csv`. Let's see how we can get them.

```
myURL<-"https://gpipisbucket.s3.amazonaws.com/movie_metadata.csv"

df<-read.csv(url(myURL))
```

Get/Download Data

If the data are of different formats, like `.jpg` , `png` , `pdf`, `xlsx` etc , usually, it's better to download them in a file. Let's see how we can do it. Note that we use the `download.file` command.

```
myURL<-"https://gpipisbucket.s3.amazonaws.com/movie_metadata.csv"
download.file(myURL, destfile = "movie_metadata.csv")
```

```
> download.file(myURL, destfile = "movie_metadata.csv")
trying URL 'https://gpipisbucket.s3.amazonaws.com/movie_metadata.csv'
Content type 'text/csv' length 1494688 bytes (1.4 MB)
downloaded 1.4 MB
```

Now, we have created a file called "movie_metadata.csv" in our working directory.

Get Data from JSON

On the web, most of the data are in a `json` format. Let's see how we can get them. We need the `httr` library.

```
library(httr)
# Get the url
url <- "http://www.omdbapi.com/?apikey=72bc447a&t=Annie+Hall&y=&plot=short&r=json"
resp <- GET(url)

# Store it to myresults
myresults<-content(resp)

myresults
```

Notice that in the `content` function you can define the type like `raw`, `application/json` etc.

Get Data from S3 to R

You can also get data from S3 provided that you know the `access_key_id` and the `secret_access_key`. You will need to work with the `aws.s3` library:

```
library(aws.s3)
Sys.setenv("AWS_ACCESS_KEY_ID" = "xxxxxxx",
           "AWS_SECRET_ACCESS_KEY" = "xxxxxxx")


# you need your path and your bucket
obj <- get_object("path", bucket = "my_bucket")


df=read.csv(text = rawToChar(obj), sep=",", header = FALSE)
```

Get Data from Hive to R

Assume that your data are stored in Hive under Hadoop. You need to download the `RJDBC` and `rJava` packages.

Then you can follow these steps:

```
library(RJDBC)
library(rJava)
#start VM
.jinit()

# set the maximum memory
options(java.parameters = "-Xmx8000m")

# add classpath
for(l in list.files('/opt/hivejdbc/')){ .jaddClassPath(paste("/opt/
hivejdbc/",l,sep=""))}

#load driver
drv <- JDBC("com.cloudera.hive.jdbc4.HS2Driver","/opt/hivejdbc/
HiveJDBC4.jar",
            identifier.quote="`")


conn <- dbConnect(drv, "jdbc:hive2://path/my_data_base", "username",
"password")

# show_databases <- dbGetQuery(conn, "show databases")
```

```
my_table <- dbGetQuery(conn, "select * from  my_data_base.my_table")…
```