Project 3

Adejare Windokun

Wednesday, October 08, 2014

Project 3: Loading Data into a Database

1. Will need RPostgreSQL which contains the driver and connectivity functions to interact with the PostgreSQL Server

```
if (!require(RPostgreSQL)) install.packages('RPostgreSQL')
```

```
## Loading required package: RPostgreSQL
## Loading required package: DBI
```

```
library(RPostgreSQL)
```

Now will build the connection string passing the username and password parameters. In addition will create the Database, 650G to store the imported text file

Will delete the database if it already exists, and then recreate it. Drop Database does not allow transaction processing therefore have to do it in two statements

```
drv = dbDriver("PostgreSQL")
con = dbConnect(drv, user= "postgres", password="cuny2014")
sql = 'DROP DATABASE IF EXISTS "650G" ;'
rs = dbSendQuery(con, sql)
sql = 'CREATE DATABASE "650G";'
rs = dbSendQuery(con, sql)

dbDisconnect(con)
```

```
## [1] TRUE
```

Will now create the table to store our data. Will name the table with tbl followed by the data for easy indentification. The column names are from those given at Amazon's introduction to the 650G schema

```
con = dbConnect(drv, user= "postgres", password="cuny2014", dbname = "650G")

sql = paste('DROP TABLE IF EXISTS tbl20141001140000;',

'CREATE TABLE tbl20141001140000
(
    record_id serial PRIMARY KEY,
    language VARCHAR (50) NULL,
    page_title TEXT NULL,
    num_requests BIGINT NULL,
    size_content BIGINT NULL
```

```
);')

rs = dbSendQuery(con, sql)


dbDisconnect(con)
```

```
## [1] TRUE
```

Show the structure of the table(s) that you created. (A single table is acceptable; you'll need to create the database and table(s)).

Code from: http://www.alberton.info/postgresql_meta_info.html, Extracting META information from PostgreSQL (INFORMATION_SCHEMA), Lorenzo Alberton

```
con = dbConnect(drv, user= "postgres", password="cuny2014", dbname = "650G")

sql = "SELECT ordinal_position, column_name, data_type, column_default, is_nullable, character_maximum_l
information_schema.columns WHERE table_name = 'tbl20141001140000' ORDER BY ordinal_position;"

dbGetQuery(con, sql)
```

```
##   ordinal_position  column_name         data_type
## 1                1    record_id           integer
## 2                2     language character varying
## 3                3   page_title              text
## 4                4 num_requests            bigint
## 5                5 size_content            bigint
##                                      column_default is_nullable
## 1 nextval('tbl20141001140000_record_id_seq'::regclass)          NO
## 2                                                 <NA>         YES
## 3                                                 <NA>         YES
## 4                                                 <NA>         YES
## 5                                                 <NA>         YES
##   character_maximum_length numeric_precision
## 1                       NA                32
## 2                       50                NA
## 3                       NA                NA
## 4                       NA                64
## 5                       NA                64
```

```
dbDisconnect(con)
```

```
## [1] TRUE
```

Will now import the data from the text file into a dataframe and then export it into our newly created PostgreSQL table

```
mfile = "C:/Users/jare/SkyDrive/WorkDocs/CUNY/607/Week 7/Project 3/20141001140000.txt"
mydata = read.table(mfile, stringsAsFactors = FALSE)
head(mydata)
```

```
##   V1                                    V2  V3      V4
## 1 en                                  Data 203 8573050
## 2 en Data%20Structures%20and%20Algorithms   2      40
## 3 en                      Data%20retention   1      20
## 4 en       Data,_Context,_and_Interaction   1   20740
## 5 en        Data,_context_and_interaction   1       0
## 6 en                  Data-Link_Switching   2   39328
```

```
con = dbConnect(drv, user= "postgres", password="cuny2014", dbname = "650G")
s = dbWriteTable(con, "tbl20141001140000", value = mydata, append = T)

print (s)
```

```
## [1] TRUE
```

. Provide the SQL script to show the five most often visited Wikipedia pages.

```
sql =   "SELECT page_title, num_requests FROM tbl20141001140000 ORDER BY num_requests DESC LIMIT 5;"

dbGetQuery(con, sql)
```

```
##
## 1
## 2 Data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAEAAAAuCAIAAABmjeQ9AAAARElEQVR42mVO2wrAUAhy/f8fzQ2wrAUAhy/f8fz%2Bf
## 3
## 4
## 5
##   num_requests
## 1          257
## 2          220
## 3          203
## 4          196
## 5          174
```

```
dbDisconnect(con)
```

```
## [1] TRUE
```

Clean up

```
dbDisconnect(con)
```

```
## [1] TRUE
```

```
dbUnloadDriver(drv)
```

```
## [1] TRUE
```

```
drv = NULL
con = NULL
```