Hands-on Lab: Creating and Querying Database Objects from R

Welcome!

In this hands-on lab, we will create and query database objects from an R notebook in Jupyter, and use ggplot2 to plot the data using R libraries.

Tasks

a. Pre-requisites & Dataset
b. Load RODBC
c. Create a database connection
d. Create a connection string and connect to the database
e. View database and driver information
f. Create the tables
g. Load data into the database
h. Fetch data from the database
i. Plot the data (using ggplot2)
j. Dis-connect

Estimated Time Needed: **30 min**

a. Pre-requisites & Dataset

Pre-requisite: In this lab we will use Jupyter Notebooks within SN Labs to access data in a Db2 on Cloud database using RODBC. Information about Jupyter notebooks, SN Labs, and Db2 services is provided in the previous labs.

Dataset used in this lab: For this lab we will utilize the Ontario public schools enrollment dataset. This data set is available under the Open Government License – Ontario and sourced from: https://www.ontario.ca/data/ontario-public-schools-enrolment

For simplicity we have already split it into two separate files: board.csv and school.csv.

Prior to starting the lab, ensure the data set files are present in the "/resources/data/samples/osb/" folder under My Data.

b. Load RODBC

The RODBC package and the ODBC driver for Db2 are pre-installed on your workbench. Let's load the RODBC package by clicking on the following cell and executing it (Shift+Enter):

[ ]:
```
library(RODBC);
```

c. Create a database connection

[ ]:
```
dsn_driver <- "{IBM DB2 ODBC Driver}"
dsn_database <- "bludb"           # e.g. "bludb"
dsn_hostname <- "<Enter Hostname>" # e.g "54a2f15b-5c0f-46df-8954-.databases.appdomain.cloud"
dsn_port <- "<Enter port number"   # e.g. "32733"
dsn_protocol <- "TCPIP"            # i.e. "TCPIP"
dsn_uid <- "<Enter UserID>"        # e.g. "zjh17769"
dsn_pwd <- "<Enter Password>"      # e.g. "zcwd4+8gbq9bm5k4"
dsn_security <- "ssl"
```

▼ Click here to view/hide hint

```
# Fill in the <...>
dsn_driver <- "{...}"
dsn_database <- "..."
dsn_hostname <- "<Enter Hostname>"
dsn_port <- "..."
dsn_protocol <- "..."
dsn_uid <- "<Enter UserID>"
dsn_pwd <- "<Enter Password>"
dsn_security <- "ssl"
```

▼ Click here to view/hide solution

```
dsn_driver <- "{IBM DB2 ODBC Driver}"
dsn_database <- "bludb"           # e.g. "bludb"
dsn_hostname <- "<Enter Hostname>" # e.g "54a2f15b-5c0f-46df-8954-.databases.appdomain.cloud"
dsn_port <- "<Enter port number"   # e.g. "32733"
dsn_protocol <- "TCPIP"            # i.e. "TCPIP"
dsn_uid <- "<Enter UserID>"        # e.g. "zjh17769"
dsn_pwd <- "<Enter Password>"      # e.g. "zcwd4+8gbq9bm5k4"
```

```
dsn_security <- "ssl"
```

d. Create a connection string and connect to the database

[ ]:
```
conn_path <- paste("DRIVER=",dsn_driver,
                   ";DATABASE=",dsn_database,
                   ";HOSTNAME=",dsn_hostname,
                   ";PORT=",dsn_port,
                   ";PROTOCOL=",dsn_protocol,
                   ";UID=",dsn_uid,
                   ";PWD=",dsn_pwd,
                   ";SECURITY=",dsn_security,
                    sep="")
conn <- odbcDriverConnect(conn_path)
conn
```

▼Click here to view/hide hint

```
# Fill in the ...
conn_path <- paste("DRIVER=",...
                   ";DATABASE=",...
                   ";HOSTNAME=",...
                   ";PORT=",...
                   ";PROTOCOL=",...
                   ";UID=",...
                   ";PWD=",..."
                   ";SECURITY=",...")
conn <- odbcDriverConnect(conn_path)
conn
```

▼Click here to view/hide solution

```
conn_path <- paste("DRIVER=",dsn_driver,
                   ";DATABASE=",dsn_database,
                   ";HOSTNAME=",dsn_hostname,
                   ";PORT=",dsn_port,
                   ";PROTOCOL=",dsn_protocol,
                   ";UID=",dsn_uid,
                   ";PWD=",dsn_pwd,
                   ";SECURITY=",dsn_security,
                    sep="")
conn <- odbcDriverConnect(conn_path)
conn
```

e. View database and driver information

[ ]:
```
#View database and driver information
sql.info <- sqlTypeInfo(conn)
conn.info <- odbcGetInfo(conn)
conn.info["DBMS_Name"]
conn.info["DBMS_Ver"]
conn.info["Driver_ODBC_Ver"]
```

▼Click here to view/hide hint

```
# Fill in the ...
sql.... <- sql...(conn)
conn.... <- odbc...(conn)
conn....["..._Name"]
conn....["..._Ver"]
conn....["Driver_..._Ver"]
conn
```

▼Click here to view/hide solution

```
#View database and driver information
sql.info <- sqlTypeInfo(conn)
conn.info <- odbcGetInfo(conn)
conn.info["DBMS_Name"]
conn.info["DBMS_Ver"]
conn.info["Driver_ODBC_Ver"]
```

f. Create the tables

You will need to *remove* the BOARD and SCHOOL tables in case they already exist.
Note: Your Db2 non-system Schema name is your userID/username in uppercase used in creating database connection.

[ ]:
```
myschema <- "tjamesbu@gmail.com" # e.g. "ZJH17769"
tables <- c("BOARD", "SCHOOL")

    for (table in tables){
      # Drop School table if it already exists
      out <- sqlTables(conn, tableType = "TABLE", schema = myschema, tableName =table)
      if (nrow(out)>0) {
        err <- sqlDrop (conn, paste(myschema,".",table,sep=""), errors=FALSE)
        if (err==-1){
          cat("An error has occurred.\n")
          err.msg <- odbcGetErrMsg(conn)
          for (error in err.msg) {
            cat(error,"\n")
          }
        } else {
          cat ("Table: ",  myschema,".",table," was dropped\n")
        }
      } else {
          cat ("Table: ",  myschema,".",table," does not exist\n")
      }
    }
```

▼Click here to view/hide hint

```
myschema <- "..."
tables <- c("...", "...L")

    for (table in ...){
      # Drop ... table if it already exists
      out <- sql...(conn, table... = "...", schema = my..., ...Name =table)
      if (nrow(...)>0) {
        err <- sql... (conn, paste(my...,".",...,...=""), errors=...)
        if (err==-1){
```

```
          ...("An error has occurred.\n")
          err.... <- odbc...Msg(conn)
          for (error in ....msg) {
            cat(...,"\n")
          }
        } else {
          cat ("...: ",  my...,".",table," was ...\n")
        }
      } else {
          cat ("...: ",  my...,".",table," does not ...\n")
      }
    }
```

▼Click here to view/hide solution

```
myschema <- "<Enter Schema>" # e.g. "ZJH17769"
tables <- c("BOARD", "SCHOOL")

    for (table in tables){
      # Drop School table if it already exists
      out <- sqlTables(conn, tableType = "TABLE", schema = myschema, tableName =table)
      if (nrow(out)>0) {
        err <- sqlDrop (conn, paste(myschema,".",table,sep=""), errors=FALSE)
        if (err==-1){
          cat("An error has occurred.\n")
          err.msg <- odbcGetErrMsg(conn)
          for (error in err.msg) {
            cat(error,"\n")
          }
        } else {
          cat ("Table: ",  myschema,".",table," was dropped\n")
        }
      } else {
          cat ("Table: ",  myschema,".",table," does not exist\n")
      }
    }
```

Let's create the BOARD table in the database.

[ ]:
```
df1 <- sqlQuery(conn, "CREATE TABLE BOARD (
                            B_ID CHAR(6) NOT NULL,
                            B_NAME VARCHAR(75) NOT NULL,
                            TYPE VARCHAR(50) NOT NULL,
                            LANGUAGE VARCHAR(50),
                            PRIMARY KEY (B_ID))",
                errors=FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
df1 <- sql...(conn, "CREATE ... BOARD (
                            B_ID ...(6) NOT ...,
                            B_NAME ...(75) NOT ...,
                            TYPE ...(50) NOT ...,
                            LANGUAGE ...(50),
                            ... KEY (B_ID))",
                ...=FALSE)
```

▼Click here to view/hide solution

```
df1 <- sqlQuery(conn, "CREATE TABLE BOARD (
                            B_ID CHAR(6) NOT NULL,
                            B_NAME VARCHAR(75) NOT NULL,
                            TYPE VARCHAR(50) NOT NULL,
                            LANGUAGE VARCHAR(50),
                            PRIMARY KEY (B_ID))",
                errors=FALSE)
```

Check if successful

[ ]:
```
if (df1 == -1){
  cat ("An error has occurred.\n")
  msg <- odbcGetErrMsg(conn)
  print (msg)
} else {
  cat ("Table was created successfully.\n")
}
```

▼Click here to view/hide hint

```
# Fill in the ...
if (... == -1){
  cat ("An ... has occurred.\n")
  msg <- odbc...Msg(conn)
  print (...)
} else {
  cat ("Table was ... ....\n")
}
```

▼Click here to view/hide solution

```
if (df1 == -1){
  cat ("An error has occurred.\n")
  msg <- odbcGetErrMsg(conn)
  print (msg)
} else {
  cat ("Table was created successfully.\n")
}
```

Now let's create the SCHOOL table.

[ ]:
```
df2 <- sqlQuery(conn, "CREATE TABLE SCHOOL (
                  B_ID CHAR(6) NOT NULL,
                  S_ID CHAR(6) NOT NULL,
                  S_NAME VARCHAR(100),
                  LEVEL VARCHAR(70),
                  ENROLLMENT INTEGER WITH DEFAULT 10,
                  PRIMARY KEY (B_ID, S_ID))", errors=FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
df2 <- sql...(conn, "CREATE ... SCHOOL (
```

```
                    B_ID ...(6) NOT NULL,
                    S_ID ...(6) NOT NULL,
                    S_NAME ...(100),
                    LEVEL ...(70),
                    ENROLLMENT ... WITH ... 10,
                    PRIMARY ... (B_ID, S_ID))", ...=FALSE)
```

▼Click here to view/hide solution

```
df2 <- sqlQuery(conn, "CREATE TABLE SCHOOL (
                    B_ID CHAR(6) NOT NULL,
                    S_ID CHAR(6) NOT NULL,
                    S_NAME VARCHAR(100),
                    LEVEL VARCHAR(70),
                    ENROLLMENT INTEGER WITH DEFAULT 10,
                    PRIMARY KEY (B_ID, S_ID))", errors=FALSE)
```

Check if successful.

[ ]:
```
if (df2 == -1){
  cat ("An error has occurred.\n")
  msg <- odbcGetErrMsg(conn)
  print (msg)
} else {
  cat ("Table was created successfully.\n")
}
```

▼Click here to view/hide hint

```
# Fill in the ...
if (... == -1){
  cat ("An ... has occurred.\n")
  msg <- odbc...Msg(conn)
  print (...)
} else {
  cat ("Table was ... ....\n")
}
```

▼Click here to view/hide solution

```
if (df2 == -1){
  cat ("An error has occurred.\n")
  msg <- odbcGetErrMsg(conn)
  print (msg)
} else {
  cat ("Table was created successfully.\n")
}
```

g. Load the data into the database

Fetch the tables present in the current database schema.

[ ]:
```
tab.frame <- sqlTables(conn, schema=myschema)
nrow(tab.frame)
tab.frame$TABLE_NAME
```

▼Click here to view/hide hint

```
# Fill in the ...
tab.... <- sql...(conn, ...=myschema)
n...(tab....)
tab....$..._NAME
```

▼Click here to view/hide solution

```
tab.frame <- sqlTables(conn, schema=myschema)
nrow(tab.frame)
tab.frame$TABLE_NAME
```

Print column 4, 6, 7, 18 details for the tables BOARD and SCHOOL

[ ]:
```
for (table in tables){
        cat ("\nColumn info for table", table, ":\n")
        col.detail <- sqlColumns(conn, table)
        print(col.detail[c(4,6,7,18)], row.names=FALSE)
}
```

▼Click here to view/hide hint

```
# Fill in the ...
for (table in ...){
        cat ("\n... info for table", ..., ":\n")
        col.detail <- sqlColumns(conn, ...)
        print(col....[c(4,6,7,18)], row....=FALSE)
}
```

▼Click here to view/hide solution

```
for (table in tables){
        cat ("\nColumn info for table", table, ":\n")
        col.detail <- sqlColumns(conn, table)
        print(col.detail[c(4,6,7,18)], row.names=FALSE)
}
```

Load the data from the board.csv into the BOARD dataframe.

[ ]:
```
boarddf <- read.csv("/resources/data/samples/osb/board.csv", header = FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
board... <- ....csv("/resources/.../samples/.../board.csv", header = ...)
```

▼Click here to view/hide solution

```
boarddf <- read.csv("/resources/data/samples/osb/board.csv", header = FALSE)
```

Display initial data from the BOARD dataframe.

```
[ ]: head(boarddf)
```

▼Click here to view/hide hint

```
# Fill in the ...
...(board...)
```

▼Click here to view/hide solution

```
head(boarddf)
```

Save the dataframe to the database table BOARD.

```
[ ]: sqlSave(conn, boarddf, "BOARD", append=TRUE, fast=FALSE, rownames=FALSE, colnames=FALSE, verbose=FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
sql...(conn, ...df, "BOARD", ...=TRUE, ...=FALSE, row...=FALSE, col...=FALSE)
```

▼Click here to view/hide solution

```
sqlSave(conn, boarddf, "BOARD", append=TRUE, fast=FALSE, rownames=FALSE, colnames=FALSE, verbose=FALSE)
```

Load the data from the school.csv into the SCHOOL dataframe

```
[ ]: schooldf <- read.csv("/resources/data/samples/osb/school.csv", header = FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
school... <- ....csv("/resources/.../samples/osb/....csv", ... = FALSE)
```

▼Click here to view/hide solution

```
schooldf <- read.csv("/resources/data/samples/osb/school.csv", header = FALSE)
```

Display some records from the beginning of the SCHOOL dataframe.

```
[ ]: head(schooldf)
```

▼Click here to view/hide hint

```
# Fill in the ...
...(school...)
```

▼Click here to view/hide solution

```
head(schooldf)
```

Change the encoding of the 3rd column character vector from latin1 to ASCII//TRANSLIT

```
[ ]: schooldf$V3 <- iconv(schooldf$V3, "latin1", "ASCII//TRANSLIT")
```

▼Click here to view/hide hint

```
# Fill in the ...
...df$V3 <- i...(...df$V3, "latin1", "ASCII//TRANSLIT")
```

▼Click here to view/hide solution

```
schooldf$V3 <- iconv(schooldf$V3, "latin1", "ASCII//TRANSLIT")
```

Save the dataframe to the database table SCHOOL.

```
[ ]: sqlSave(conn, schooldf, "SCHOOL", append=TRUE, fast=FALSE, rownames=FALSE, colnames=FALSE, verbose=FALSE)
```

▼Click here to view/hide hint

```
# Fill in the ...
sql...(conn, school..., "SCHOOL", append=..., fast=..., row...=FALSE, col...=FALSE)
```

▼Click here to view/hide solution

```
#NOTE: This may take a long time because of the large size of the database. When there is lot of data to insert in the database it may be faster to use the native LOAD
utility of the database instead of through R.

sqlSave(conn, schooldf, "SCHOOL", append=TRUE, fast=FALSE, rownames=FALSE, colnames=FALSE, verbose=FALSE)
```

h. Fetch data from the database

Fetch the data from the database table BOARD and display some rows from the end of the data.

```
[ ]: boarddb <- sqlFetch(conn, "BOARD")
tail(boarddb)
```

▼Click here to view/hide hint

```
# Fill in the ...
...db <- sql...(conn, "...")
...(board...)
```

▼Click here to view/hide solution

```
boarddb <- sqlFetch(conn, "BOARD")
tail(boarddb)
```

Fetch the data from the database table SCHOOL and and display some rows from the end of the data.

```
[ ]: schooldb <- sqlFetch(conn, "SCHOOL")
tail(schooldb)
```

▼Click here to view/hide hint

```
# Fill in the ...
...db <- sql...(conn, "...")
tail(...db)
```

▼Click here to view/hide solution

```
schooldb <- sqlFetch(conn, "SCHOOL")
tail(schooldb)
```

i. Plot the data (using ggplot2)

[ ]:
```
library(ggplot2);
```

▼Click here to view/hide hint

```
# Fill in the ...
...(gg...2);
```

▼Click here to view/hide solution

```
library(ggplot2);
```

Get the elementary school data from the database from both tables in descending sequence.

[ ]:
```
elequery <- query <- paste("select s.enrollment as ENROLLMENT
from school s, board b
where b.b_name = 'Toronto DSB' and b.b_id=s.b_id
and s.level = 'Elementary'
order by enrollment desc")
```

▼Click here to view/hide hint

```
# Fill in the ...
...query <- ... <- paste("... s.enrollment as ...
from ... s, ... b
where b.b_... = 'Toronto DSB' and b.b_id=s....
and ....level = 'Elementary'
order by ... desc")
```

▼Click here to view/hide solution

```
elequery <- query <- paste("select s.enrollment as ENROLLMENT
from school s, board b
where b.b_name = 'Toronto DSB' and b.b_id=s.b_id
and s.level = 'Elementary'
order by enrollment desc")
```

create the elementary school dataframe.

[ ]:
```
eledf <- sqlQuery(conn, elequery)
dim(eledf)
```

▼Click here to view/hide hint

```
# Fill in the ...
ele... <- sql...(conn, ...query)
dim(...df)
```

▼Click here to view/hide solution

```
eledf <- sqlQuery(conn, elequery)
dim(eledf)
```

Create a density plot of elementary school enrollments.

[ ]:
```
qplot(ENROLLMENT, data=eledf, geom="density",  main="TDSB School Size - Elementary")
```

▼Click here to view/hide hint

```
# Fill in the ...
q...(ENROLLMENT, ...=eledf, ...="density",  ...="TDSB School Size - ...")
```

▼Click here to view/hide solution

```
qplot(ENROLLMENT, data=eledf, geom="density",  main="TDSB School Size - Elementary")
```

Create the secondary school enrollments query in descending sequence.

[ ]:
```
secquery <- paste("select s.enrollment as ENROLLMENT
from school s, board b
where b.b_name = 'Toronto DSB' and b.b_id=s.b_id
and s.level = 'Secondary'
order by enrollment desc")
```

▼Click here to view/hide hint

```
# Fill in the ...
sec... <- paste("... s.enrollment as ...
from ... s, board b
where b.b_... = 'Toronto ...' and b.b_id=s....
and s.... = 'Secondary'
order by ... desc")
```

▼Click here to view/hide solution

```
secquery <- paste("select s.enrollment as ENROLLMENT
from school s, board b
where b.b_name = 'Toronto DSB' and b.b_id=s.b_id
and s.level = 'Secondary'
order by enrollment desc")
```

Create the dataframe using the data in the database.

[ ]:
```
secdf <- sqlQuery(conn, secquery)
```

▼Click here to view/hide hint

```
# Fill in the ...
secdf <- sql...(conn, sec...)
```

▼Click here to view/hide solution

```
secdf <- sqlQuery(conn, secquery)
```

Create a density plot of secondary school enrollments.

[ ]:
```
qplot(ENROLLMENT, data=secdf, geom="density", main="TDSB School Size - Secondary")
```

▼Click here to view/hide hint

```
# Fill in the ...
q...(ENROLLMENT, ...=secdf, ...="density", ...="TDSB School Size - ...")
```

▼Click here to view/hide solution

```
qplot(ENROLLMENT, data=secdf, geom="density", main="TDSB School Size - Secondary")
```

Query the BOARD database for enrollments.

[ ]:
```
denquery <- paste("select b.b_name, s.s_name, level as LEVEL, enrollment
 from board b, school s where b.b_id = s.b_id and b.b_name = 'Toronto DSB'")
```

▼Click here to view/hide hint

```
# Fill in the ...
den... <- paste("select b.b_..., s.s_..., level as ..., ...
 from board b, ... s where b.... = s.b_id and b.b_... = 'Toronto DSB'")
```

▼Click here to view/hide solution

```
denquery <- paste("select b.b_name, s.s_name, level as LEVEL, enrollment
 from board b, school s where b.b_id = s.b_id and b.b_name = 'Toronto DSB'")
```

Query the database.

[ ]:
```
dendf <- sqlQuery(conn, denquery)
```

▼Click here to view/hide hint

```
# Fill in the ...
d...f <- sql...(conn, den...)
```

▼Click here to view/hide solution

```
dendf <- sqlQuery(conn, denquery)
```

Create a box plot of enrollements in elementary and secondary schools in Toronto.

[ ]:
```
dendf$LEVEL <- as.factor(dendf$LEVEL)
boxplot(ENROLLMENT ~ LEVEL, dendf, names =c("Secondary","Elementary"), main="Toronto DSB")
```

▼Click here to view/hide hint

```
# Fill in the ...
d...f$LEVEL <- as....(d...f$LEVEL)
box...(ENROLLMENT ~ ..., d...f, names =c("...","..."), ...="Toronto DSB")
```

▼Click here to view/hide solution

```
dendf$LEVEL <- as.factor(dendf$LEVEL)
boxplot(ENROLLMENT ~ LEVEL, dendf, names =c("Secondary","Elementary"), main="Toronto DSB")
```

j. Dis-connect

Finally, as a best practice we should close the database connection once we're done with it.

[ ]:
```
close(conn)
```

▼Click here to view/hide hint

```
# Fill in the ...
...(conn)
```

▼Click here to view/hide solution

```
close(conn)
```

Summary

In this lab you created and queried database objects from an R notebook in Jupyter, and you used ggplot2 to plot the data using R libraries.

Thank you for completing this module on creating and querying database objects from R.

Authors

- Rav Ahuja
- Agatha Colangelo
- Sandip Saha Joy

Changelog

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-07-14	2.1	Lakshmi Holla	Added ssl changes
2021-01-22	2.0	Sandip Saha Joy	Created revised version of the lab
2017	1.0	Rav Ahuja & Agatha Colangelo	Created initial version of the lab

---