International Journal of Computer Science and Business Informatics

IJCSBI.ORG

An Efficient Connection between Statistical Software and Database Management System

Sunghae Jun
Department of Statistics, Cheongju University
Chungbuk 360-764 Korea

ABSTRACT
In big data era, we need to manipulate and analyze the big data. For the first step of big data manipulation, we can consider traditional database management system. To discover novel knowledge from the big data environment, we should analyze the big data. Many statistical methods have been applied to big data analysis, and most works of statistical analysis are dependent on diverse statistical software such as SAS, SPSS, or R project. In addition, a considerable portion of big data is stored in diverse database systems. But, the data types of general statistical software are different from the database systems such as Oracle, or MySQL. So, many approaches to connect statistical software to database management system (DBMS) were introduced. In this paper, we study on an efficient connection between the statistical software and DBMS. To show our performance, we carry out a case study using real application.

Keywords
Statistical software, Database management system, Big data analysis, Database connection, MySQL, R project.

1. INTRODUCTION
Every day, huge data are created from diverse fields, and stored in computer systems. These big data are extremely large and complex [1]. So, it is very difficult to manage and analyze them. But, big data analysis is important issue in many fields such as marketing, finance, technology, or medicine. Big data analysis is based on statistics and machine learning algorithms. In addition, data analysis is depended on statistical software, and the data are stored in database systems. So, for big data analysis, we should manage statistical software and database system effectively. In this paper, we consider R project system as statistical software. R is an environment for statistical computing including statistical analysis and graphical display of data [2]. This program provides most of statistical and machine learning methods for big data analysis. We use MySQL for connecting database system from R project. The MySQL is a database management system (DBMS) product that is the most popular open source database in the world, in addition, this is a free software like R system [3]. So, in our research, we use R and MySQL for an efficient connection between statistical software and DBMS. There was a work about DB access through R [4]. This covered

the DB access problems of R, and showed the ODBC (open database connectivity) drivers for connecting R and DBMS such as MySQL, PostgreSQL, and Oracle. Also, the authors of this paper introduced the installation and technological environment for the DB access. But, they did not illustrate detailed approaches for real applications. That is, their work was about a conceptual suggestion for the access of R to MySQL. So, in this paper, we perform more specific study for connection between statistical software, R to DBMS, MySQL. In our case study, we will show detailed and efficient connection of R to MySQL using specific data set from the University of California, Irvine (UCI) machine learning repository [5]. We will cover our research background in next section. In section 3, our proposed methodology will be shown. We also introduce an efficient connection between statistical database and DBMS in section 4. Lastly we conclude our study and offer our future works for statistical database system.

2. RESEARCH BACKGROUND

2.1 Statistical Software

To analyze data, we can consider diverse approaches using statistical software. These days, there are so many products for statistical software. SAS (statistical analysis system) is the most popular software for statistical analysis [6]. But, this is expensive, so there are not many companies using SAS except large size companies. SPSS (statistical analysis in social science) is another representative software [7], but this is also expensive. Minitab [8] and S-Plus [9] are well used statistics packages and these are all not free. Recently, R has been used in many works for statistical data analysis, and this is free. In addition, R also provides most of statistical functions included in SAS, or SPSS. R is open source program, so we can modify R functions for our statistical computing. This is very useful advantage of R. Therefore, we consider R for connection to database system in this research.

2.2 Database Management System

Database is a collection of data, and database management system (DBMS) is a software for managing database using structured query language (SQL) [10],[11]. Oracle is one of popular DBMS products [12], but it is expensive. MySQL is another DBMS, which is widely used open source software in the world [3]. Also, most functions of MySQL are similar to Oracle [3]. So, in this paper, we use MySQL for DBMS connecting to statistical software, R. Using MySQL DBMS efficiently, we use RODBC package supported by R CRAN in our research [13].

3. STATISTICAL DATABASE SYSTEM

The main goal of our study is to solve the cost problem for constructing statistical database system, because we should buy additional product to connect statistical software to DBMS. For example, for the connection

between SAS and DBMS, we need 'SAS/Access' product as supplementary software. In general, this is expensive. So, we tried to make the connection between statistical software and DBMS without cost. The 'efficient' of our paper was about 'cost'. There are many approaches to connect statistical software and DBMS. To use most of them, we should buy additional products. But, there are few free approaches. So, we find an approach to connect statistical software and DBMS without cost. In this paper, we study an efficient connection between DBMS and statistical software. We select the MySQL as a DBMS for our research, and use R project as statistical software because not only they are free but also they have good functions. In addition, the R and MySQL have strong performance in statistical computing and DBMS respectively for constructing statistical database system [14],[15],[16],[17]. In general, big data are transformed to structured data type for statistical analysis as follow;

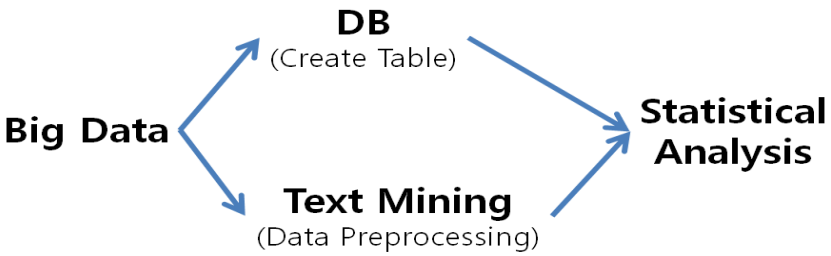

Figure 1. From big data to statistical analysis

First, big data are stored in DB by creating table. Second, big data are changed to structured data by preprocessing based on text mining. All data by DB and text mining are analyzed by statistical analysis. We find that text mining process is hard work for data preprocessing [18]. So, we know that table creation is more effective approach for big data analysis. To construct MySQL DB, we use console or graphic user interface (GUI) environments as follow;

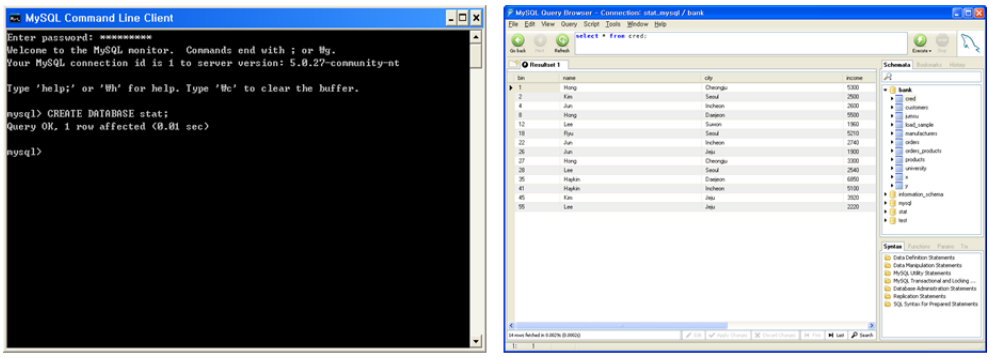

Figure 2. User interface of MySQL

In this paper, we use SQL codes in the MySQL console. Also, we use RODBC as an ODBC database interface between R and MySQL [13]. In general R system, package is a set of additional R functions. R packages are not installed in basic R system. If we need to use a package, we have to add the package to the R system. Also we can search all packages from the R CRAN, and install them from the CRAN [19]. The RODBC package provides efficient functions for ODBC database access. So, our research is based on RODBC package to connect R to MySQL. To install RODBC in R system, we should select R CRAN mirror site. After RODBC installation, we load this package on R system as follow;

```
>library (RODBC)
```

The R system uses 'library' function for loading a package. By this R code, we can use all functions provided by RODBC package such as odbcConnect, sqlFetch, and sqlQuery. They are used in our research for DB accessing and connecting. To connect MySQL DB, we use 'odbcConnect' function of RODBC package as follow;

```
>db_con =odbcConnect("stat_MySQL")

User = , Password = , Database =
```

The DSN is 'stat_MySQL' and the 'db_con' object of R system includes the connecting result. Also, in this connecting process, we decide user name, password, and determined database. If R and MySQL are connected each other, we can show the tables of MySQL DB using 'sqlTables' function as follow;

```
>sqlTables(con)

TABLE_CAT TABLE_SCHEM TABLE_NAME TABLE_TYPE REMARKS
```

The result of this function is the information of connected DB and its tables.

3.1 Structure of DB Connection Software

In general, for connecting DBMS to application software, we should use ODBC connector [20]. R as a statistical software is also needed to ODBC driver to access MySQL DBMS. In this paper, we consider RODBC package for efficient connection between R and MySQL. Figure 3 shows the ODBC connection between DBMS and statistical software, and their specific products.

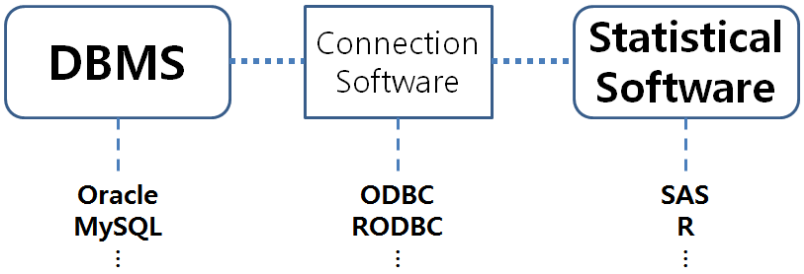

Figure 3. Connection between DBMS and statistical software

Oracle and MySQL are representative DBMS products, and SAS and R system are popular software for statistical analysis. General ODBC program is used for connecting application software to DBMS. So, there are so many ODBC drivers for diverse DBMS and application products. Our work is focused on the connection R and MySQL, and we select RODBC as an ODBC driver. The RODBC is a package of many R packages for DB accessing. RMySQL is another R package for R and MySQL [21]. This package is also R interface to access the MySQL DBMS. In addition to RODBC and RMySQL, there are some packages for connecting R to MySQL. In this paper, we use RODBC for MySQL accessing. This is an ODBC driver like SAS connection to DBMS as follow.

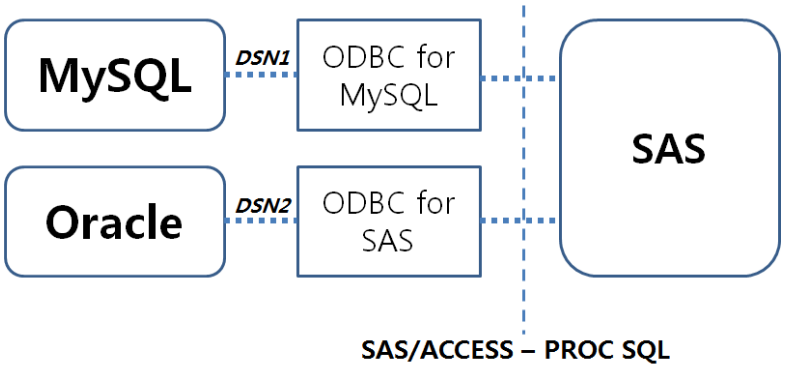

Figure 4. Connection between MySQL/Oracle and SAS

SAS uses some ODBC drivers for diverse DBMS such as MySQL and Oracle. Also, the drivers use their data source name (DSN). In this research, we also use DSN for RODBC package. Next, we show more detailed connection between R and MySQL.

3.2 Efficient Connection between R and MySQL

The RODBC package of R system is an efficient ODBC connector. This includes diverse functions to access DBMS as follow;

· odbcConnect: function for open connections to ODBC

· sqlFetch: function for fetching tables from DB

· sqlQuery: function for SQL query

· sqlSave: function for writing data frame to table in DB

Also, we can use more functions for accessing and manipulating MySQL DB by RODBC packages. The process of connection between R and MySQL is as follow;

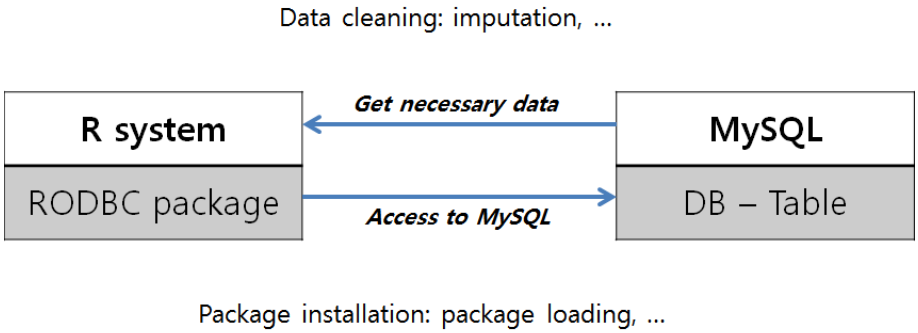

Figure 5. Connecting process between R and MySQL

Using RODBC package, R system get necessary data from MySQL DB, and we analyze the connected data. Also, R system accesses to MySQL by sqlQuery function of RODBC, and create a table for storing analysis result using R system. Our process of connection between R and MySQL is shown as follow;

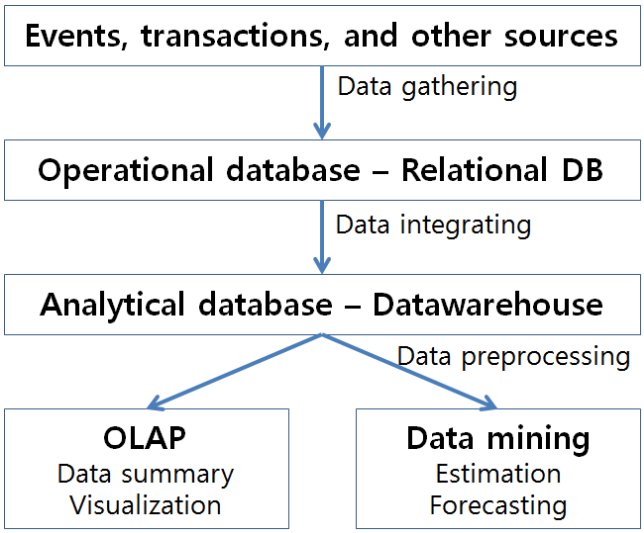

Figure 6. Connecting process between R and MySQL

A table of MySQL DB is transformed to an object in R by RODBC connector. So, we are able to analyze the object data from the DB table. We also perform online transaction processing (OLAP) for data summarization and visualization. Next, we carry out a case study for verifying our work.

4. CASE STUDY

To illustrate a case study in real problem, we used 'RODBC' package from R-project [13]. This is the software for ODBC database connection between R and DBMS such as MySQL. Also, we made experiment using an example data set from the UCI machine learning repository [5].

4.1 UCI Machine Learning Repository

For our case study, we used "Abalone" data set from the UCI machine learning repository [5]. This data set consisted of 8 variables (columns) and 4,177 observations (rows). The main goal of the data is to predict the age of abalone from the physical measurements. Next table shows the variables and their values [5].

Table 1. Table captions should be placed above the table

Variable	Data type	Description
Sex	Nominal	M(male), F(female), I(infant)
Length	Continuous	Longest shell measurement
Diameter	Continuous	perpendicular to length
Height	Continuous	with meat in shell
Whole_weight	Continuous	whole abalone
Shucked_weight	Continuous	weight of meat
Viscera_weight	Continuous	gut weight (after bleeding)
Shell_weight	Continuous	after being dried
Rings	Discrete	+1.5 gives the age in years

The last variable (rings) is target variable, and others are all input variables. We constructed MySQL DB using this data set. The original data from UCI machine learning repository was text file separated by 'comma', but the MySQL needed data file separated by 'tab key' for DB loading file. So, we transformed the data type using Excel as follow.

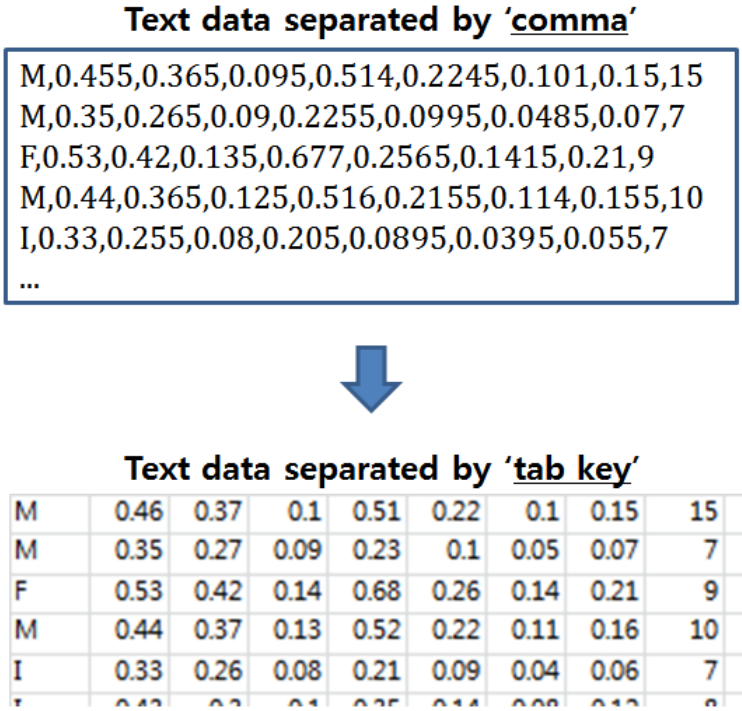

Figure 7. Data transformation for MySQL loading

To load text data file on MySQL, we should make a table to save these data. So, we create the table in next step.

4.2 DB Creation

We used SQL to create table for loading Abalone data set on MySQL DBMS as follow;

· CREATE DATABASE case_study;

· USE case_study;

· CREATE TABLE abalone(Sex CHAR(3), Length FLOAT(10), Diameter FLOAT(10), Height FLOAT(10), Whole_weight FLOAT(10), Shucked_weight FLOAT(10), Viscera_weight FLOAT(10), Shell_weight FLOAT(10), Rings INT(5));

· LOAD DATA INFILE 'd:/data/abalone.txt' INTO TABLE abalone;

· SELECT * FROM abalone;

Using above SQL codes, we constructed a table of Abalone data in MySQL DB(case_study). Next, we connected the table of abalone in case_study DB to R system.

4.3 Connecting R to MySQL

We used RODBC package for connecting R to MySQL as follow;


```
>library(RODBC)
>abalone_con=odbcConnect("abalone_ODBC")
>sqlTables(abalone_con)
TABLE_SCHEM     TABLE_NAME      TABLE_TYPE
case_study      abalone         TABLE
>vars=sqlQuery(abalone_con, "SELECT sex, diameter, rings FROM abalone")
Sex     Diameter        Rings
1       M       0.365           15
2       M       0.265           7
3       F       0.420           9
4       M       0.365           10
5       I       0.255           7
…
```

Using above R codes, we saved three variables of abalone data set to 'vars' R object. We found the abalone table was created well from the SQL query result by sqlQuery function. This function enabled the usage of SQL in R system. So, we analyzed abalone data using analytical functions of R system. Next, the result of data analysis is shown.

4.4 Data Analysis

First, we performed data summarization of three variables using 'summary' function of R system as follow;

```
>summary(vars)
sex             diameter        rings
F:1307 Min.   :0.0550 Min.   : 1.000
I:1342 1st Qu.:0.3500 1st Qu.: 8.000
M:1528 Median :0.4250 Median : 9.000
       Mean   :0.4079 Mean   : 9.934
       3rd Qu.:0.4800 3rd Qu.:11.000
       Max.   :0.6500 Max.   :29.000
```

This function provided frequency or descriptive statistic according to data type (continuous or nominal). For example diameter is continuous variable, so we got minimum, 25 percentile, median, mean, 75 percentile, and maximum values. Next we carried out data visualization as follow;

```
>boxplot(vars$diameter)
```

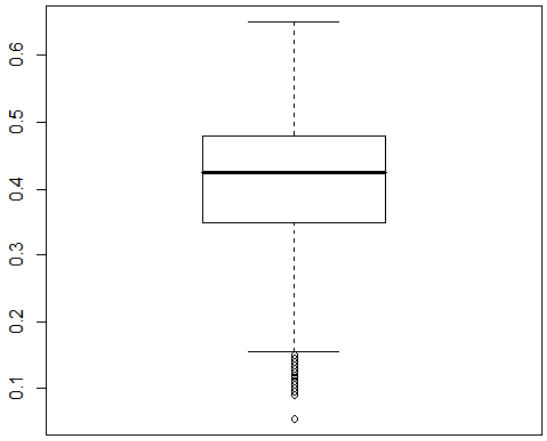

Figure 8. Boxplot: data visualization of MySQL table

This shows boxplot of diameter variable of abalone table. Using graphical functions supported by R system, we can also get diverse visualization results such as histogram, plot, and so on. Lastly we constructed regression model using 'reg' function as follow;

```
>regression_result=lm(rings~diameter, data=vars)
>sunnary(regression_result)
Residuals:
Min          1Q      Median      3Q     Max
-5.19        -1.69   -0.72       0.91        16.00
Coefficients:
Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.3186     0.1727      13.42    <2e-16 ***
diameter     18.6699   0.4115      45.37    <2e-16 ***
R-squared: 0.3302,  Adj. R-squared: 0.3301
```

The regression is popular model in statistical analysis. The dependent and independent variables are 'rings' and 'diameter' respectively. So, we got the following regression equation;

Rings=2.3186+18.6699diameter. Therefore, in our case study, we illustrated a case study of connection between R and MySQL.

5. CONCLUSION

In this paper, we studied on the efficient connection between DBMS and statistical software. We used R system and MySQL as statistical software and DBMS respectively. The RODBC package was used for DB connection in our study. After connecting between R and MySQL, we analyzed the data of MySQL table. So, this can be expanded to the big data analysis. In our

case study, we illustrated how our approach could be applied in real application. We selected Abalone data set from the UCI machine learning repository for our case study. Our result contributes to the works related to big data analysis. In addition, we can analyze the data in DBMS directly by statistical methods. In our future works, we will expand the scope of the connection between DBMS and statistical software to more products.

6. DISCUSSION

The biggest problem of statistical database system is the cost of connecting between statistical software and DBMS. For example, we should buy 'SAS/Access' product additionally and install it to SAS base system for connecting SAS and DBMS. Generally this supplementary product is expensive, so most users have had difficulty to use statistical databases system. In this paper, we selected R system as statistical software instead of SAS, and we used RODBC as ODBC connector instead of SAS/Access, because R and RODBC are all free. But, their performance is similar to SAS. Also, in new analytical functions such as statistical leaning theory and machine learning algorithm, they surpass SAS.

REFERENCES

[1] Sathi, A. *Big Data Analytics*. An Article from IBM Corporation, 2012.

[2] Heiberger, R. M., and Neuwirth, E.*R through Excel – A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer, 2009.

[3] MySQL, *The World's most popular open source database*. http://www.mysql.com, accessed on October 2013.

[4] Sim, S., Kang, H., and Lee, Y. Access to Database through the R-Language. *The Korean Communications in Statistics, 15, 1* (2008), 51-64.

[5] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, accessed on October 2013.

[6] SAS, http://www.sas.com,accessed on October 2013.

[7] SPSS, http://www-01.ibm.com/software/analytics/spss/, accessed on October 2013.

[8] Minitab, http://www.minitab.com, accessed on October 2013.

[9] S-Plus, http://solutionmetrics.com.au/products/splus/, accessed on October 2013.

[10] Wikipedia, *the free encyclopedia*. http://en.wikipedia.org, accessed on October 2013.

[11] Date, C. J.*An Introduction to Database Systems*. 7th edition, Addition-Wesley, 2000.

[12] Oracle, http://www.oracle.com, accessed on October 2013.

[13] Ripley, B.*Package RODBC*. CRAN R-Project, 2013.

[14] R-bloggers, *On R versus SAS.* http://www.r-bloggers.com/on-r-versus-sas/, accessed on December, 2013.

[15] Linkin,*Advanced Business Analytics, Data Mining and Predictive Modeling*. http://www.linkedin.com/groups/SAS-versus-R-35222.S.65098787, accessed on December, 2013.

[16] Clever Logic, *MySQL vs. Oracle Security*, http://cleverlogic.net/articles/mysql-vs-oracle, accessed on December, 2013.

[17] Find The Best, *Oracle vs MySQL*, http://database-management-systems.findthebest.com/saved_compare/Oracle-vs-MySQL, accessed on December, 2013.

[18] Han, J., and Kamber, M. *Data Mining Concepts and Techniques*. Morgan Kaufmann, 2001.

[19] R system, *The R Project for Statistical Computing*. http://www.r-project.org, accessed on October 2013.

[20] Spector, P. *Data Manipulation with R*, Springer, 2008.

[21] James, D. A., and DebRoy, S.*Package RMySQL*. CRAN R-Project, 2013.