

---

# 일선 기업 현장에서의 R의 활용

2011. 10. 28

(주) 베가스 김준기

## ‘분석(Analytics)’ : 정의

- 분석(analytics)은 사물을 이해하는데 필요한 광의의 분석(analysis)이나 데이터의 단순조회와 단순 리포팅의 생산의 과정이 아닌, 데이터에 근간한 통계분석, 예측과 트렌드 예측, 최적화가 여기에 해당함

### 분석 (Analytics)이란 ?

I

의사결정과 **action**에 활용하기 위한 데이터의 광범위한 활용, 통계적이며 정량적 측면의 분석, 탐색적 분석 및 예측모델링, 사실에 근거한 경영을 의미함

mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and action

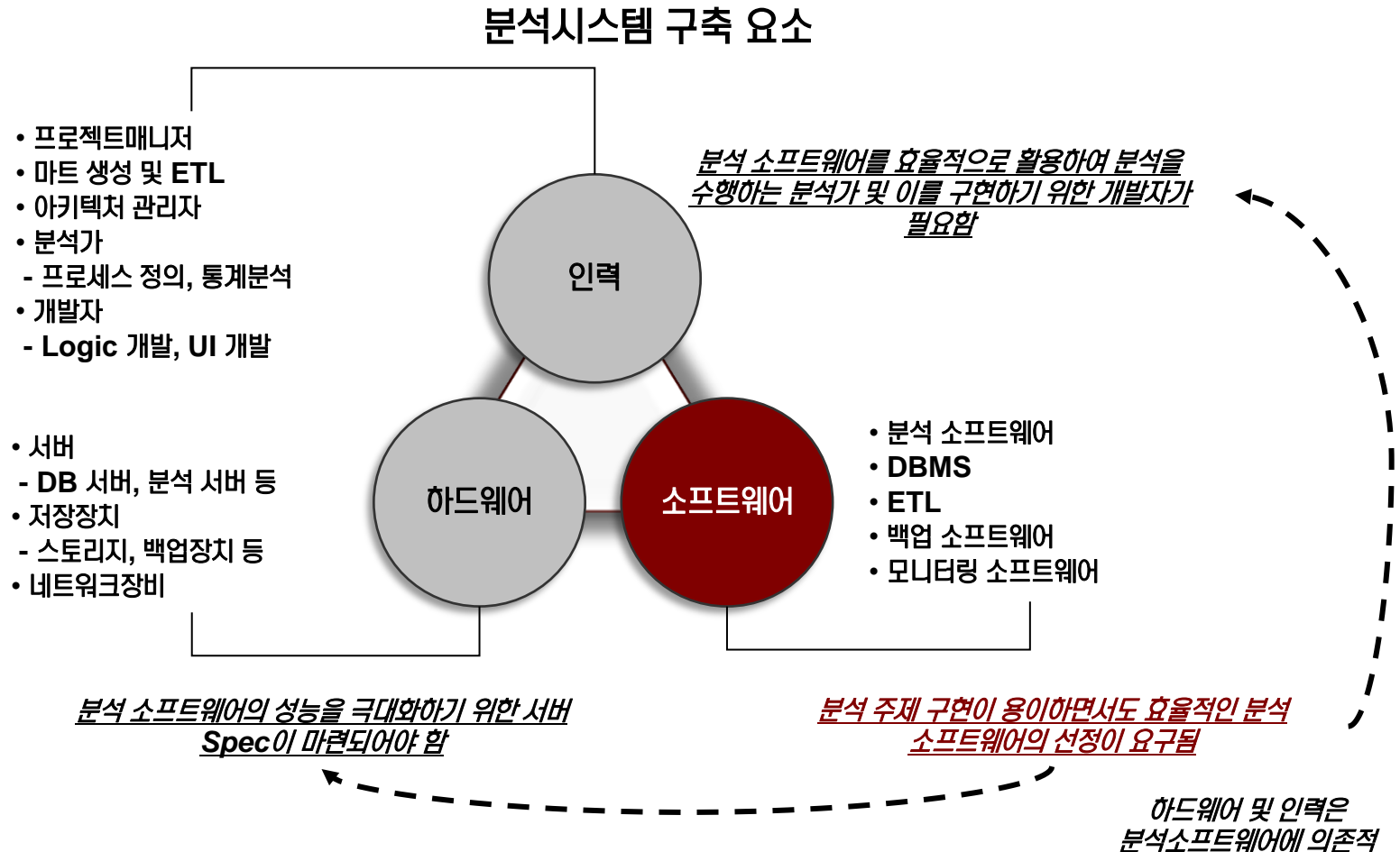
II

의사결정 혹은 완전 자동화된 의사결정의 입력이 될 수 있음

may be input for human decisions or may drive fully automated decisions

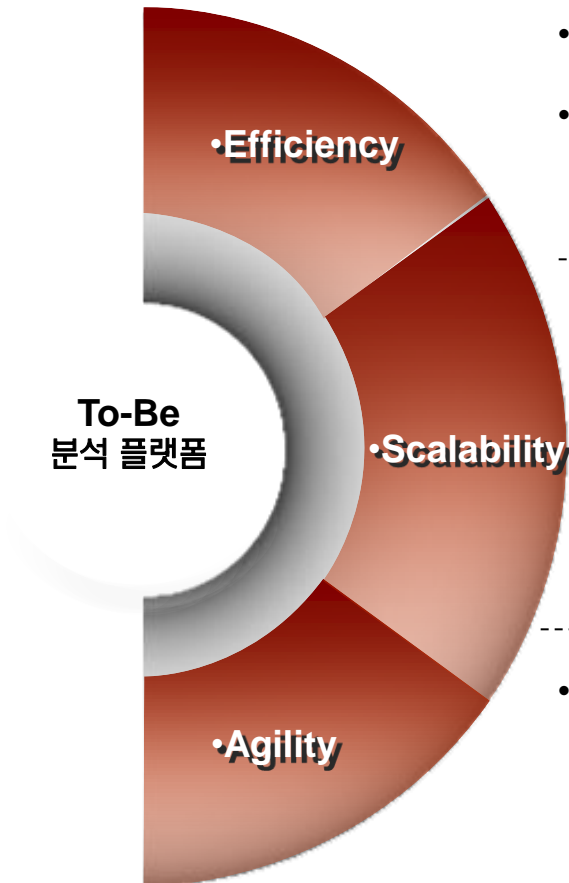
## ‘분석(Analytics)’ 시스템

- 분석시스템 구축에 있어서 소프트웨어(S/W), 하드웨어(H/W), 인력은 필수적인 요소임
- 이중 분석 소프트웨어의 제품 선정은 이에 따른 비용 지출 및 이를 구현하기 위한 인력 및 하드웨어에도 큰 영향을 미치는 중요한 요소임



## ‘분석(Analytics)’ 시스템

- 바람직한 분석시스템의 구축은 분석엔진을 중심으로 마련된 저비용(**Low Cost**)이지만 고성능이며 확장성이나 인터페이스가 뛰어난 (**Excellent**) **Analytic Platform**(분석 플랫폼)을 중심으로 이루어져야 함
- 소위 **LCBEx (Low Cost But Excellent) Analytic Platform**은 1) 효율적 (**efficiency**)이며 2) 확장성이 뛰어나고 (**scalability**) 3) 비즈니스 목표에 부합하는 시스템을 신속하게 구축 (**agility**)할 수 있어야 함



- 저비용
    - 오픈 소스 소프트웨어 기반으로 구축해 최대한 도입비용을 낮춰야 함
  - 고성능
    - 구현 사상을 고려하였을 때, 빠른 계산처리 및 새로운 알고리즘, 방법론이 제공되는 오픈소스 기반 분석엔진 필요함
- 
- 확장 및 통합 용이성
    - 독립된 형태의 분석 시스템 구축 없이 분산 처리를 통한 처리가 가능하여야 함
    - Hadoop과 같은 오픈소스 기반의 솔루션을 활용할 수 있음
- 
- 구현 신속성
    - 분석 방법이나 결과 등을 오브젝트로 관리하여 공유, 재활용이 가능하여야 함
    - 정형화된 분석 프로세스의 패키징이 용이하여 이관이나 재활용이 용이하여야 함

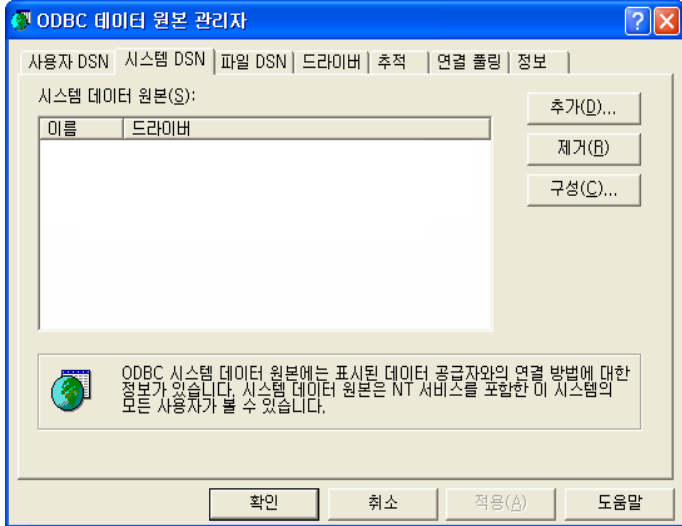
**Part I : DBMS 인터페이스**

**Part II : 데이터 Manipulation**

**Part III : Batch Program**

**Part III : Visualization Tool과의 연계**

## Part I : DBMS 인터페이스

R 관련 패키지	사전 설정 정보	ODBC 설정 화면
RODBC	DBMS Client 설치 Windows ODBC 설정	

## 관련 Tool 및 Package

DBMS Client : DBMS 접속을 위한 Tool  
 ODBC(Open DataBase Connectivity)  
 R : RODBC package

# RODBC 패키지

## RODBC: ODBC Database Access

An ODBC database interface

Version: 1.3-2  
Depends: R ( $\geq$  2.9.0), utils  
Imports: stats  
Published: 2010-07-26  
Author: Brian Ripley, and from 1999 to Oct 2002 Michael Lapsley  
Maintainer: Brian Ripley <ripley at stats.ox.ac.uk>  
License: [GPL-2](#) | [GPL-3](#)  
SystemRequirements: An ODBC3 driver manager and drivers.  
CRAN checks: [RODBC results](#)

### Downloads:

Package source: [RODBC 1.3-2.tar.gz](#)  
MacOS X binary: [RODBC 1.3-2.tgz](#)  
Windows binary: [RODBC 1.3-2.zip](#)  
Reference manual: [RODBC.pdf](#)  
Vignettes: [ODBC Connectivity](#)  
News/ChangeLog: [ChangeLog](#)  
Old sources: [RODBC archive](#)

- **RODBC Windows OS에서 DBMS와의 인터페이스를 위한 패키지임**
- **RODBC 패키지를 이용하여 데이터베이스를 통한 데이터를 R로 가져오기 위해서는 ODBC 설정에 대상 정보가 설정이 되어 있어야 함**
- **R에서 RODBC 패키지는 library(RODBC) 명령을 이용하면 내장된 Object를 사용할 수 있음**

# RODBC 패키지 내 Object 리스트

[close\\_RODBC](#)  
[getSqlTypeInfo](#)  
[odbcClearError](#)  
[odbcClose](#)  
[odbcCloseAll](#)  
[odbcConnect](#)  
[odbcConnectAccess](#)  
[odbcConnectAccess2007](#)  
[odbcConnectDbase](#)  
[odbcConnectExcel](#)  
[odbcConnectExcel2007](#)  
[odbcDataSources](#)  
[odbcDriverConnect](#)  
[odbcEndTran](#)  
[odbcFetchRows](#)  
[odbcGetErrMsg](#)  
[odbcGetInfo](#)  
[odbcQuery](#)  
[odbcReConnect](#)  
[odbcSetAutoCommit](#)  
[odbcTables](#)  
[RODBC](#)  
[setSqlTypeInfo](#)  
[sqlClear](#)  
[sqlColumns](#)  
[sqlCopy](#)  
[sqlCopyTable](#)  
[sqlDrop](#)  
[sqlFetch](#)  
[sqlFetchMore](#)  
[sqlGetResults](#)  
[sqlPrimaryKeys](#)  
[sqlQuery](#)  
[sqlSave](#)  
[sqlTables](#)  
[sqlTypeInfo](#)  
[sqlUpdate](#)

ODBC Close Connections  
 Specify or Query a Mapping of R Types to DBMS Types  
 Low-level ODBC functions  
 ODBC Close Connections  
 ODBC Close Connections  
 ODBC Open Connections  
 ODBC Open Connections  
 ODBC Open Connections  
 ODBC Open Connections  
 ODBC Open Connections  
 ODBC Open Connections  
 List ODBC Data Sources  
 ODBC Open Connections  
 ODBC Set Auto-Commit Mode  
 Low-level ODBC functions  
 Low-level ODBC functions  
 Request Information on an ODBC Connection  
 Low-level ODBC functions  
 ODBC Open Connections  
 ODBC Set Auto-Commit Mode  
 Low-level ODBC functions  
 ODBC Database Connectivity  
 Specify or Query a Mapping of R Types to DBMS Types  
 Deletion Operations on Tables in ODBC databases  
 Query Column Structure in ODBC Tables  
 ODBC Copy  
 ODBC Copy  
 Deletion Operations on Tables in ODBC databases  
 Reading Tables from ODBC Databases  
 Reading Tables from ODBC Databases  
 Query an ODBC Database  
 Query Column Structure in ODBC Tables  
 Query an ODBC Database  
 Write a Data Frame to a Table in an ODBC Database  
 List Tables on an ODBC Connection  
 Request Information about Data Types in an ODBC Database  
 Write a Data Frame to a Table in an ODBC Database

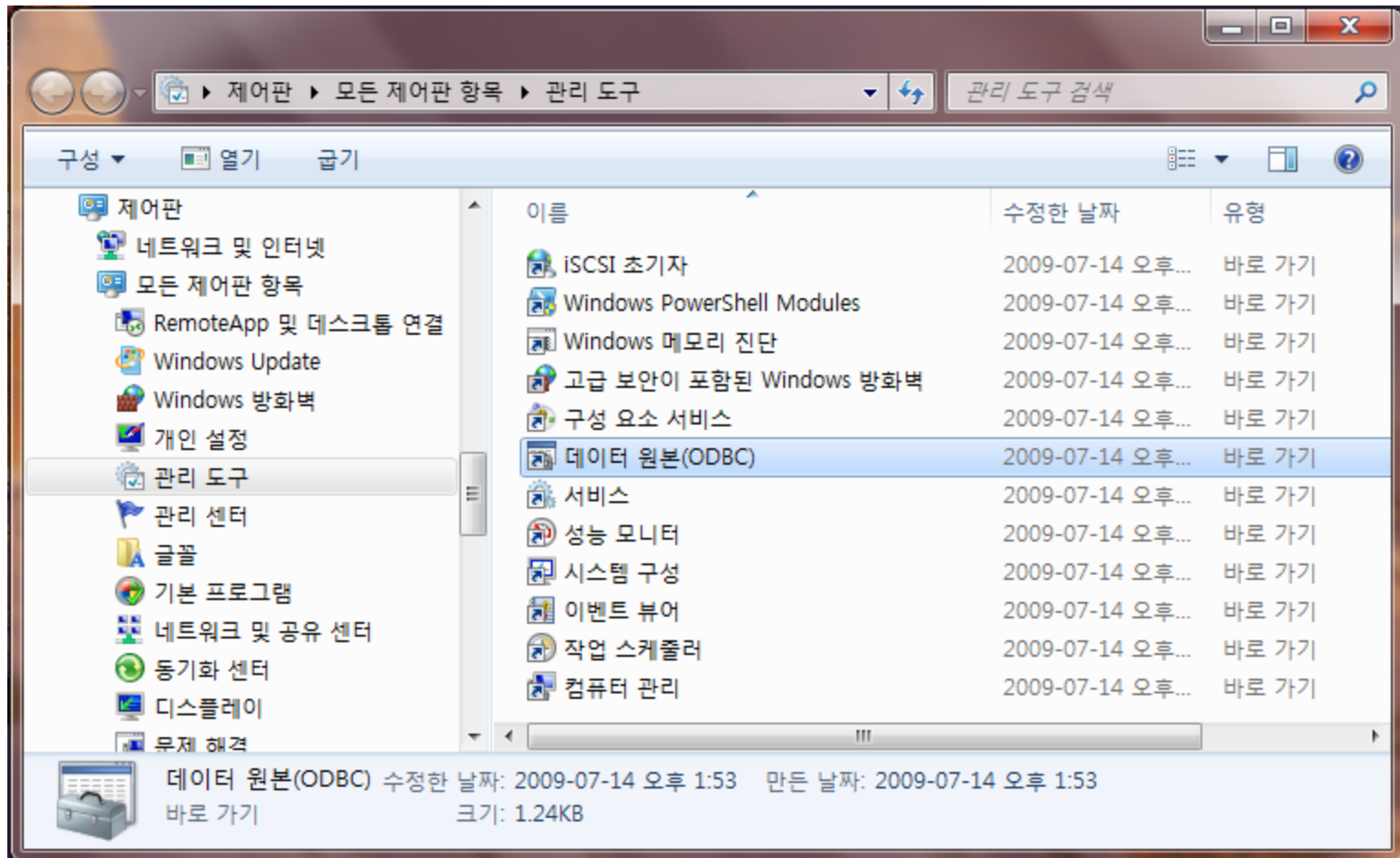
## RODBC 패키지 내 오브젝트

오브젝트 명	설명	사용방법
odbcConnect	ODBC 접속	odbcConnect(dsn, uid = "", pwd = "", ...)
odbcClose	ODBC 접속 해제	odbcClose(channel)
sqlQuery	SQL을 이용하여 데이터를 R의 Object로 생성	sqlQuery(channel, paste("select State, Murder from USArrests", "where Rape > 30 order by Murder"))
sqlUpdate	R의 data frame Object를 DB Table에 쓰기	sqlUpdate(channel, dat, tablename = NULL, index = NULL, verbose = FALSE, test = FALSE, nastring = NULL, fast = TRUE)

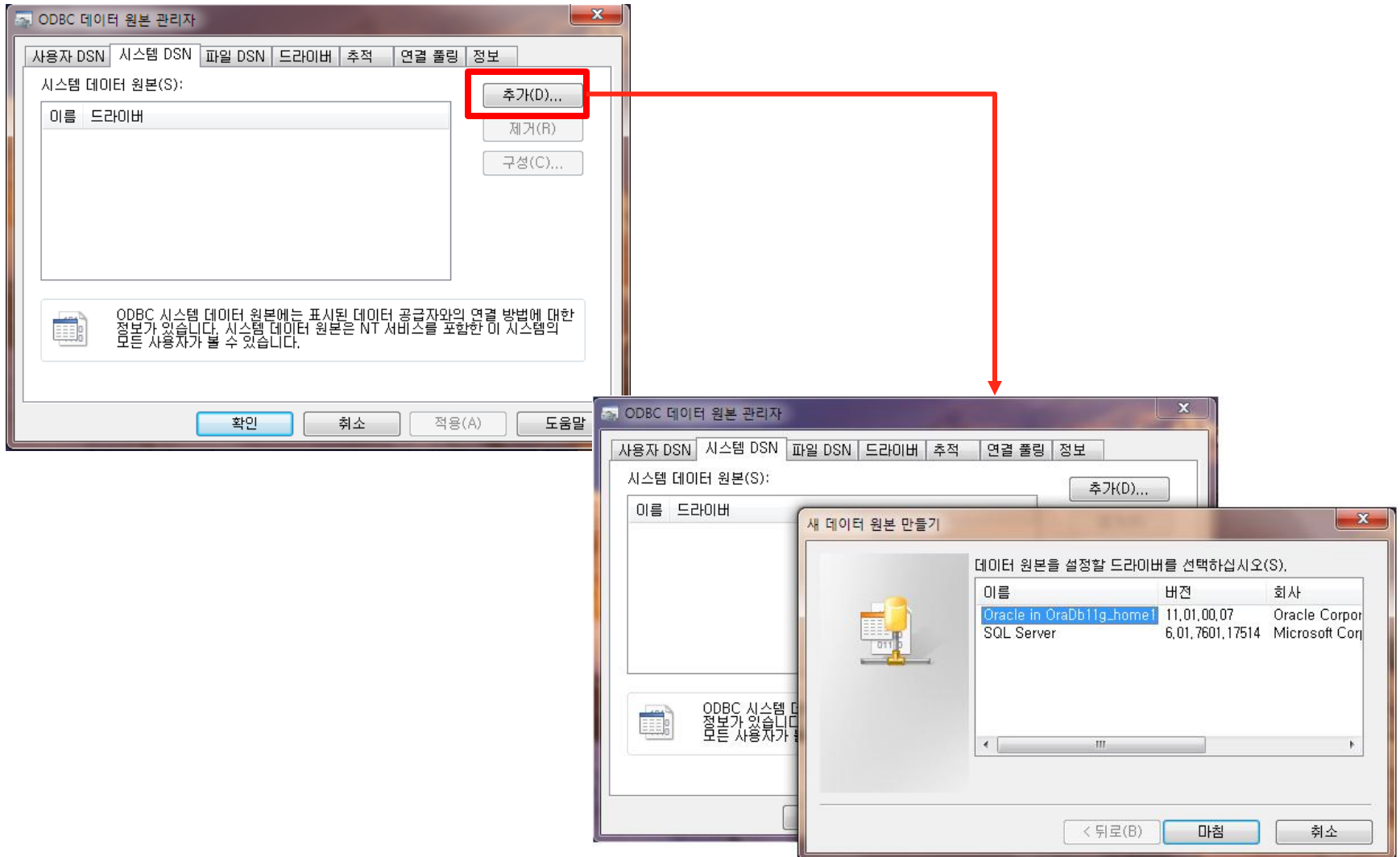


## ODBC 설정 하기 : Oracle DBMS 예

- 관리도구 – 데이터 원본(ODBC)

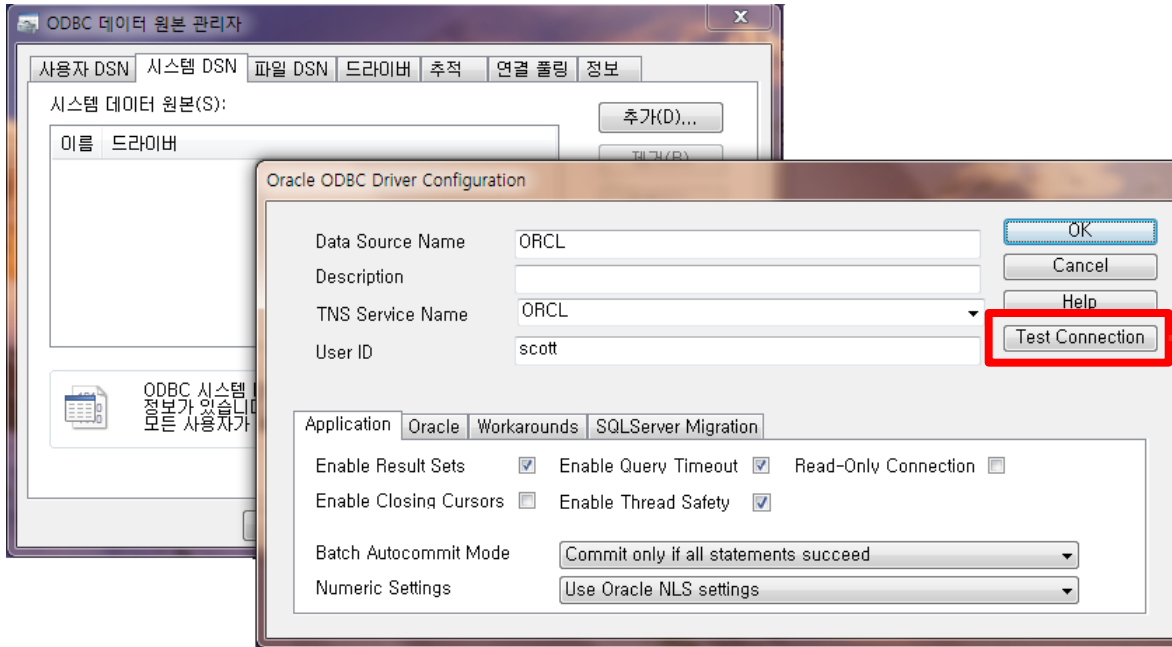


## • ODBC 데이터 원본 관리자



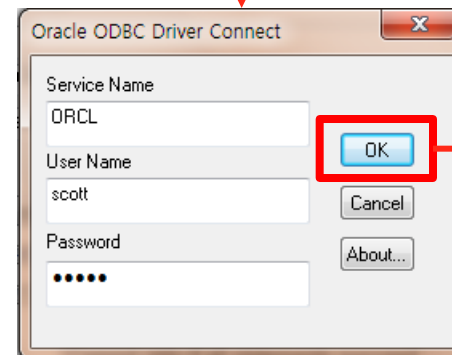
## • 새 데이터 원본 만들기

## • Oracle ODBC Driver Configuration

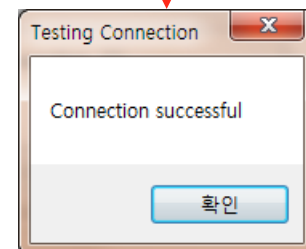


### 데이터베이스 정보 입력

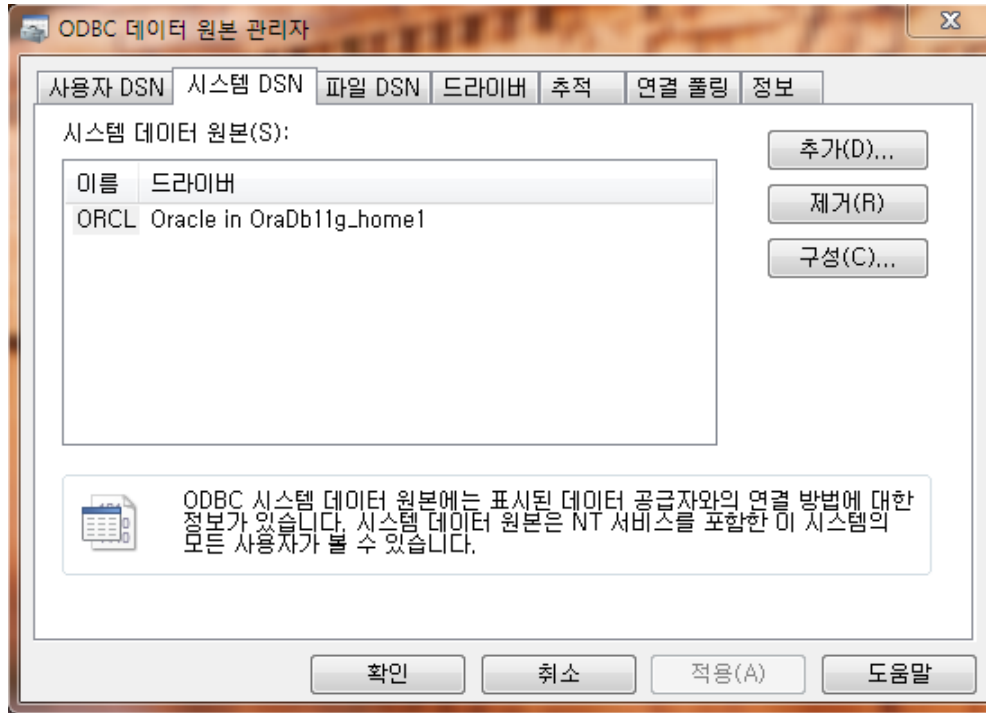
- **Data Source Name : ORCL**
- **TNS Service Name : ORCL**
- **User ID : scott**



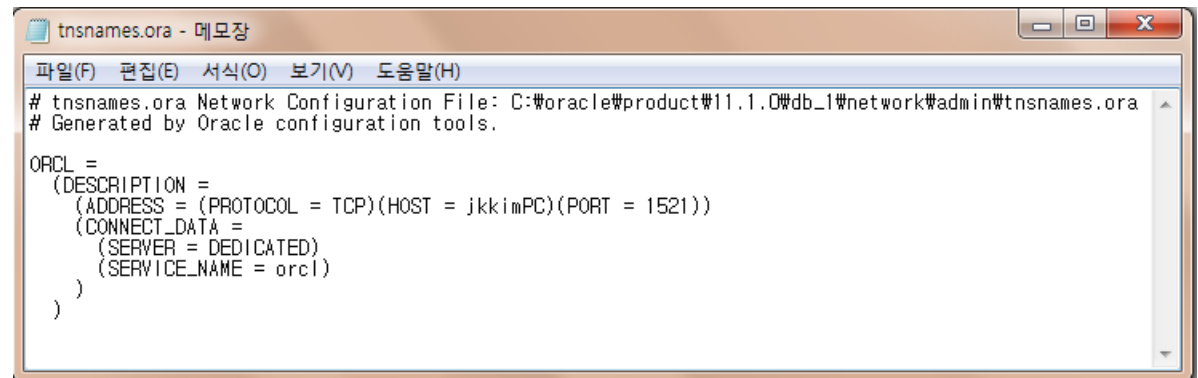
### • Test Connection



- ODBC 데이터 원본 관리자 : ODBC 설정 완료



- TNS Service Name : ORCL



## RODBC 패키지를 이용한 DB 접속 및 데이터 가져오기

```
# 패키지 불러오기
library(RODBC)

# Database 접속 정보
odbc.dsn <- "ORCL"
odbc.uid <- "scott"
odbc.pwd <- "tiger"

# =====
# 1. 테이블 리스트 가져오기
# =====

# SQL 작성
select.x <- "SELECT TNAME"
from.x <- "FROM tab"
sql.x <- paste(select.x, from.x, sep = " ")

# DB 접속 : DB 접속 및 ODBC 설정이 사전에 필요함
db.connect <- odbcConnect(dsn = odbc.dsn, uid = odbc.uid, pwd = odbc.pwd)
# 데이터 가져오기
import.tb.list <- sqlQuery(db.connect, sql.x, stringsAsFactors = FALSE)
# 접속정보 해제
odbcClose(db.connect)
```

## RODBC 패키지를 이용한 DB 접속 및 데이터 가져오기

```
# =====  
# 2. EMP 테이블 가져오기  
# =====  
# SQL 작성  
select.x <- "SELECT *"  
from.x <- "FROM EMP"  
sql.x <- paste(select.x, from.x, sep = " ")  
  
# DB 접속 : DB 접속 및 ODBC 설정이 사전에 필요함  
db.connect <- odbcConnect(dsn = odbc.dsn, uid = odbc.uid, pwd = odbc.pwd)  
# 데이터 가져오기  
import.emp.tb.data <- sqlQuery(db.connect, sql.x, stringsAsFactors = FALSE)  
# 접속정보 해제  
odbcClose(db.connect)  
  
# 가져온 데이터 보기  
import.emp.tb.data  
  
< ===== R Console 결과 화면 ===== >  
EMPNO ENAME      JOB MGR  HIREDATE  SAL COMM DEPTNO  
1  7369 SMITH    CLERK 7902 1980-12-17 800  NA   20  
2  7499 ALLEN  SALESMAN 7698 1981-02-20 1600 300   30  
...
```

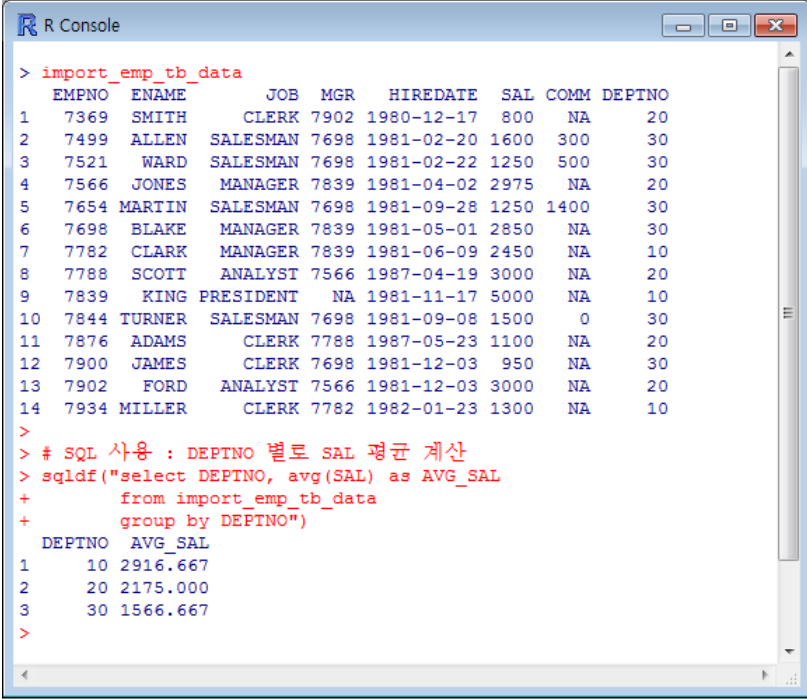
**Part I : DBMS 인터페이스**

**Part II : 데이터 Manipulation**

**Part III : Batch Program**

**Part III : Visualization Tool과의 연계**

## Part II : 데이터 Manipulation

R 관련 패키지	활용	예제
sqldf	SQL 문을 이용하여 데이터 Manipulation이 가능	 <pre> R Console &gt; import_emp_tb_data   EMPNO  ENAME      JOB   MGR  HIREDATE   SAL  COMM  DEPTNO 1   7369   SMITH      CLERK 7902 1980-12-17   800    NA     20 2   7499   ALLEN    SALESMAN 7698 1981-02-20  1600   300     30 3   7521   WARD    SALESMAN 7698 1981-02-22  1250   500     30 4   7566   JONES    MANAGER 7839 1981-04-02  2975    NA     20 5   7654   MARTIN  SALESMAN 7698 1981-09-28  1250  1400     30 6   7698   BLAKE    MANAGER 7839 1981-05-01  2850    NA     30 7   7782   CLARK    MANAGER 7839 1981-06-09  2450    NA     10 8   7788   SCOTT   ANALYST 7566 1987-04-19  3000    NA     20 9   7839   KING  PRESIDENT   NA 1981-11-17  5000    NA     10 10  7844   TURNER  SALESMAN 7698 1981-09-08  1500     0     30 11  7876   ADAMS    CLERK 7788 1987-05-23  1100    NA     20 12  7900   JAMES    CLERK 7698 1981-12-03   950    NA     30 13  7902   FORD    ANALYST 7566 1981-12-03  3000    NA     20 14  7934   MILLER   CLERK 7782 1982-01-23  1300    NA     10 &gt; &gt; # SQL 사용 : DEPTNO 별로 SAL 평균 계산 &gt; sqldf("select DEPTNO, avg(SAL) as AVG_SAL +       from import_emp_tb_data +       group by DEPTNO")   DEPTNO  AVG_SAL 1      10 2916.667 2      20 2175.000 3      30 1566.667 &gt; </pre>

sqldf 함수를 이용한 데이터 Manipulation에서는 오브젝트 이름에 '.'을 사용할 수 없음



## sqldf 패키지를 이용한 데이터 Manipulation

```
# 패키지 불러오기
library(sqldf)

# =====
# sqldf 함수를 이용한 Data Manipulation
# =====

# 데이터프레임 이름 변경
import_dept_tb_data <- import.dept.tb.data
import_emp_tb_data <- import.emp.tb.data

# SQL 사용 : DEPTNO 별로 SAL 평균 계산
sqldf("select DEPTNO, avg(SAL) as AVG_SAL
      from import_emp_tb_data
      group by DEPTNO")

# SQL 사용 : MGR이 NULL이 아닌 DEPTNO 별로 SAL 평균 계산
sqldf("select DEPTNO, avg(SAL) as AVG_SAL
      from import_emp_tb_data
      where MGR is not NULL
      group by DEPTNO")
```

## sqldf 패키지를 이용한 데이터 Manipulation

# Left Join 예제를 보여주기 위하여 DEPTNO 10을 50으로 변경

```
import_dept_tb_data[import_dept_tb_data$DEPTNO == 10, ]$DEPTNO <- 50
```

# SQL 사용 : Inner Join

```
# merge(import_emp_tb_data, import_dept_tb_data, by = c("DEPTNO"))
```

```
sqldf("select a.*, b.DNAME, LOC  
      from import_emp_tb_data a, import_dept_tb_data b  
      where a.DEPTNO = b.DEPTNO")
```

# SQL 사용 : Left Join

```
# merge(import_emp_tb_data, import_dept_tb_data, by = c("DEPTNO"), all.x = T)
```

```
sqldf("select a.*, b.DNAME, LOC  
      from import_emp_tb_data a left join import_dept_tb_data b  
      using(DEPTNO)")
```

**Part I : DBMS 인터페이스**

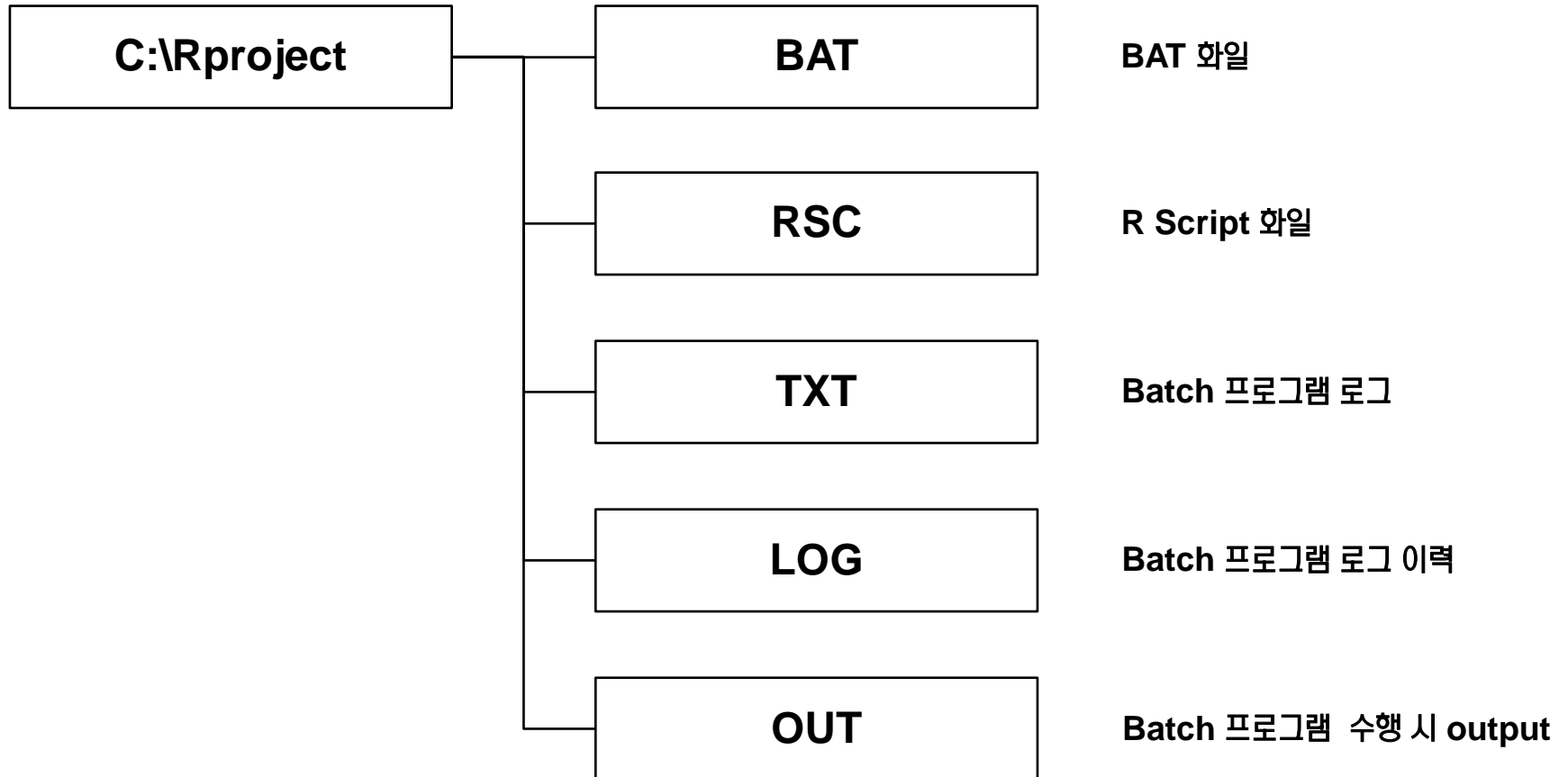
**Part II : 데이터 Manipulation**

**Part III : Batch Program**

**Part III : Visualization Tool과의 연계**

## Part III : Batch Program

### [Batch Program 관련 디렉토리]



**\*\* Batch Program 관련 디렉토리는 발표자가 설정한 구성임 (반드시 이렇게 구성해야 한다는 의무사항이 아님)**

## Batch 작업을 위한 R Script (C:\Rproject\RSC\RBatchSample.R)

```
# =====  
#                               프로그램 시작  
# =====  
cat("Batch Program Start!!!", as.character(Sys.time()), "\n")  
# -----  
  
# 오늘 날짜  
today.date <- as.character(Sys.Date())  
  
# 데이터프레임 생성  
sample.df <- data.frame(year = substr(today.date, 1, 4),  
                        month = substr(today.date, 6, 7),  
                        day = substr(today.date, 9, 10),  
                        sample.norm = rnorm(1000000), stringsAsFactors = TRUE)  
  
# 데이터프레임을 csv 파일로 저장  
write.csv(sample.df, paste("C:\\RProject\\OUT\\", today.date, ".", "sample.df.csv", sep = ""), row.names = FALSE)  
  
# =====  
#                               프로그램 종료  
# =====  
cat("Batch Program End!!!", as.character(Sys.time()), "\n")  
# -----
```

## Batch 작업을 위한 bat 파일 (C:\Rproject\BAT\BatchSample.bat)

```
CD C:\Program Files\R\R-2.13.1\bin\x64
```

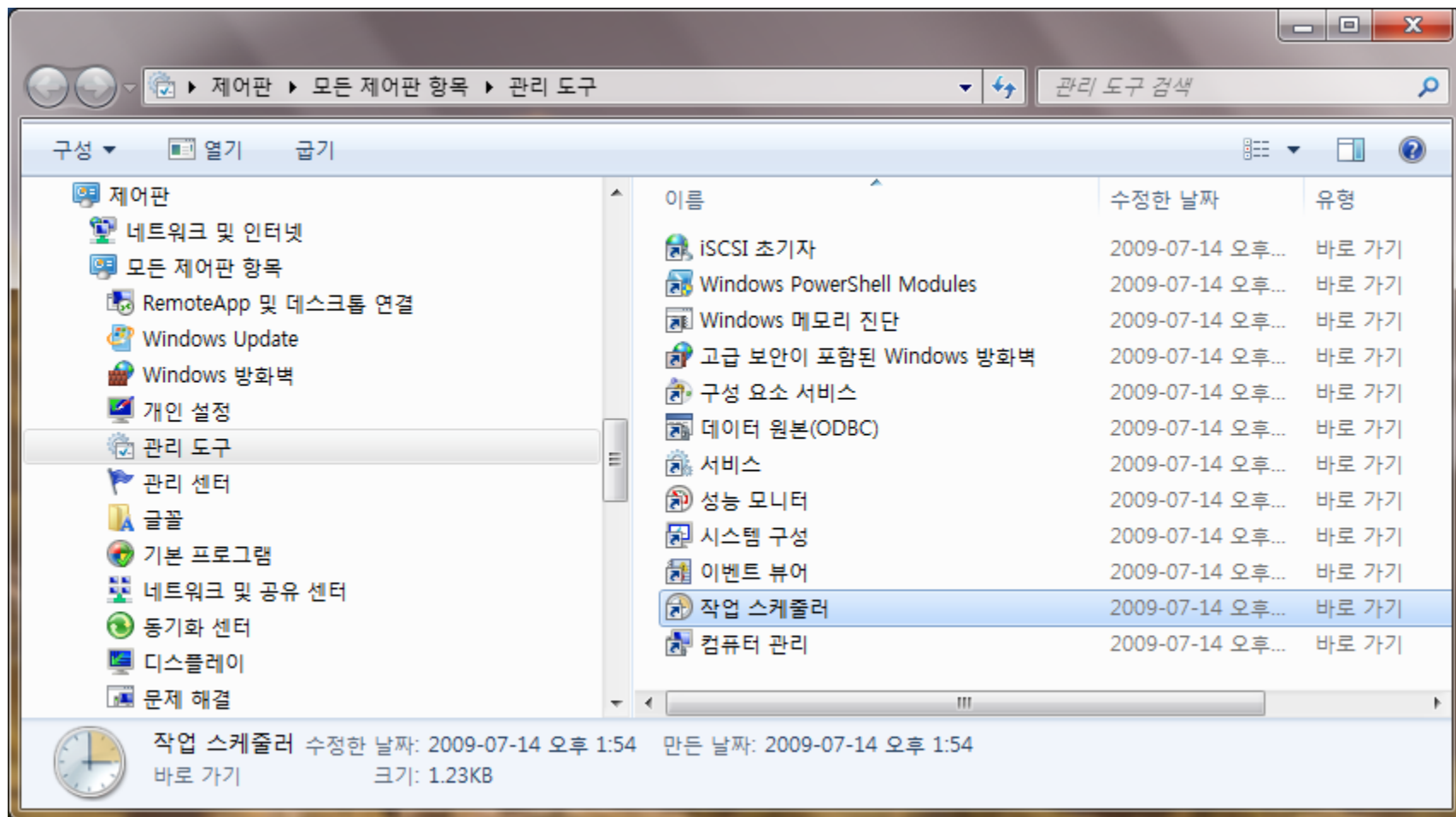
```
Rcmd BATCH C:\Rproject\RSC\RBatchSample.R C:\RProject\TXT\RBatchSample.txt
```

```
set filename=RBatchSample_~2%_~2%_~0,2%_~3,2%_~6,2%.log
```

```
copy C:\RProject\TXT\RBatchSample.txt C:\RProject\LOG\%filename%
```

- bat 파일
  - bat 파일 내 명령어는 dos 명령어 임
- Rcmd
  - R Batch를 수행하기 위해서는 Rcmd.exe 파일을 이용함
- Rcmd BATCH
  - BATCH 명령어는 대문자 임
  - Usage: Rcmd BATCH [options] infile [outfile]
  - outfile을 지정하지 않으면 .Rout 파일로 outfile이 생성됨

- 작업 스케줄러



## • 작업 스케줄러 - 작업 만들기



작업 스케줄러 개요

작업 스케줄러를 사용하면 지정된 시간에 컴퓨터에서 자동으로 수행되는 일반 작업을 만들고 관리할 수 있습니다. 시작하려면 [작업] 창에서 명령을 클릭하십시오.

작업은 작업 스케줄러 라이브러리 내의 폴더에 저장됩니다. 각 작업을 보거나 실행하려면 작업 스케줄러 라이브러리에서 작업을 선택하고 [작업] 메뉴에서 명령을 클릭하십시오.

작업 상태

다음 기간에 시작된 작업 상태: 지난 24시간

요약: 총 0개 - 0개 실행 중, 0개 성공, 0개 중지, 0개 실패

작업 이름	실행 결과	실행 시작	실행 끝	트리거 주제
-------	-------	-------	------	--------

실행 중인 작업

실행 중인 작업은 현재 사용하고 만료되지 않은 작업입니다.

요약: 총 51

마지막 새로 고침 시간 2011-10-27 오전 10:36:09 새로 고침

작업

작업 스케줄러 (로컬)

다른 컴퓨터에 연결...

작업 만들기...

작업 가져오기...

실행 중인 모든 작업 표시

모든 작업 기록 사용

AT 서비스 계정 구성

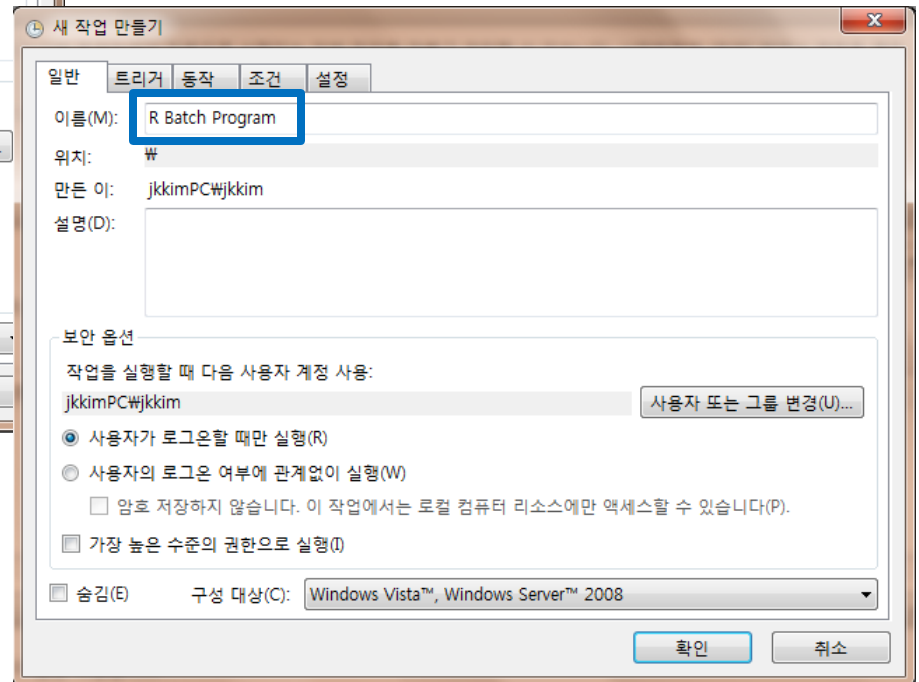
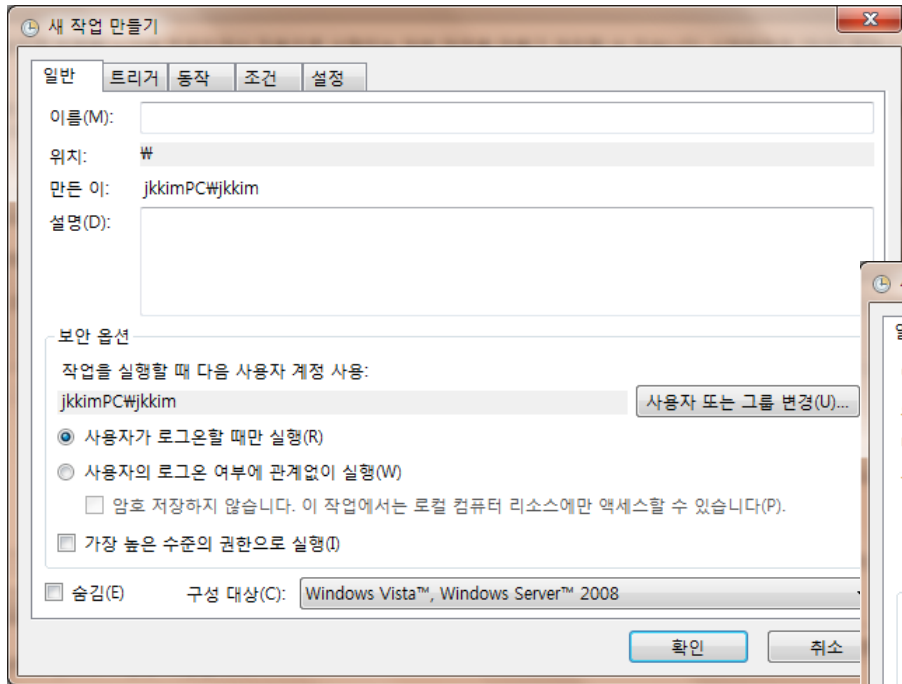
보기

새로 고침

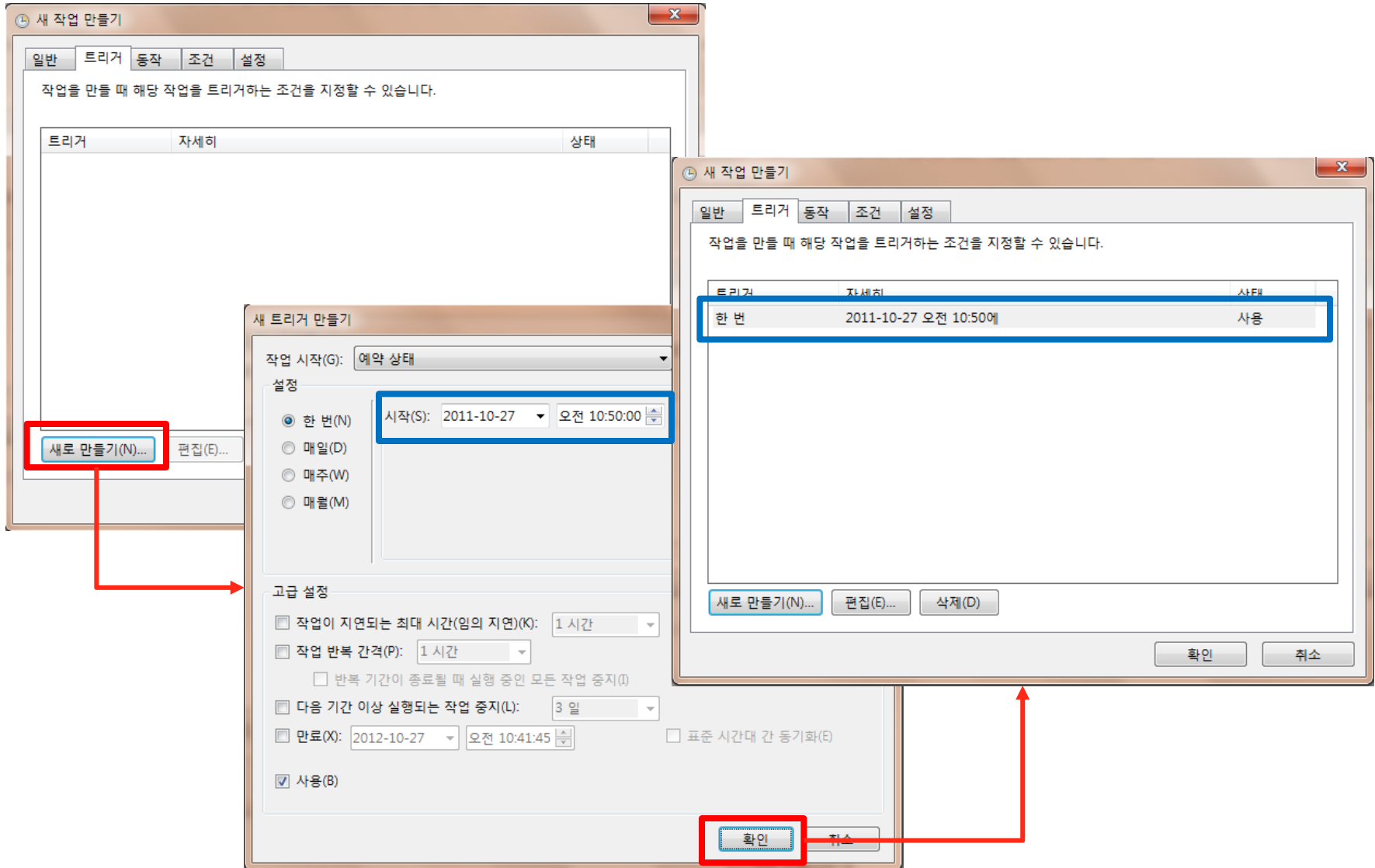
도움말



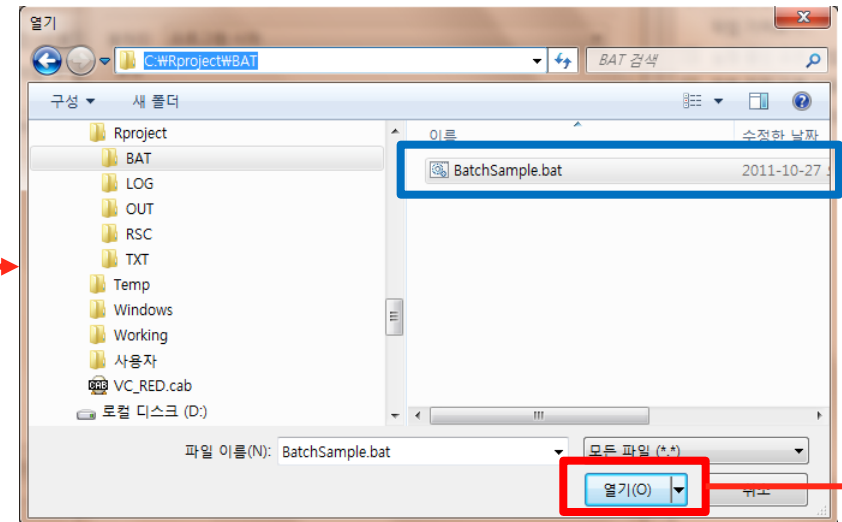
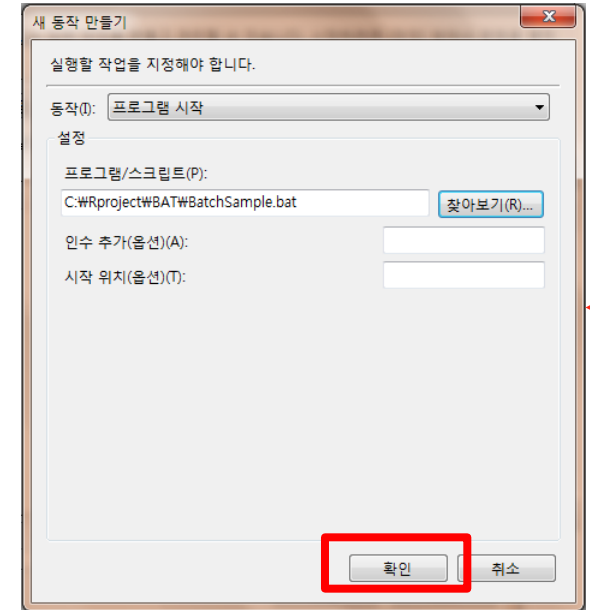
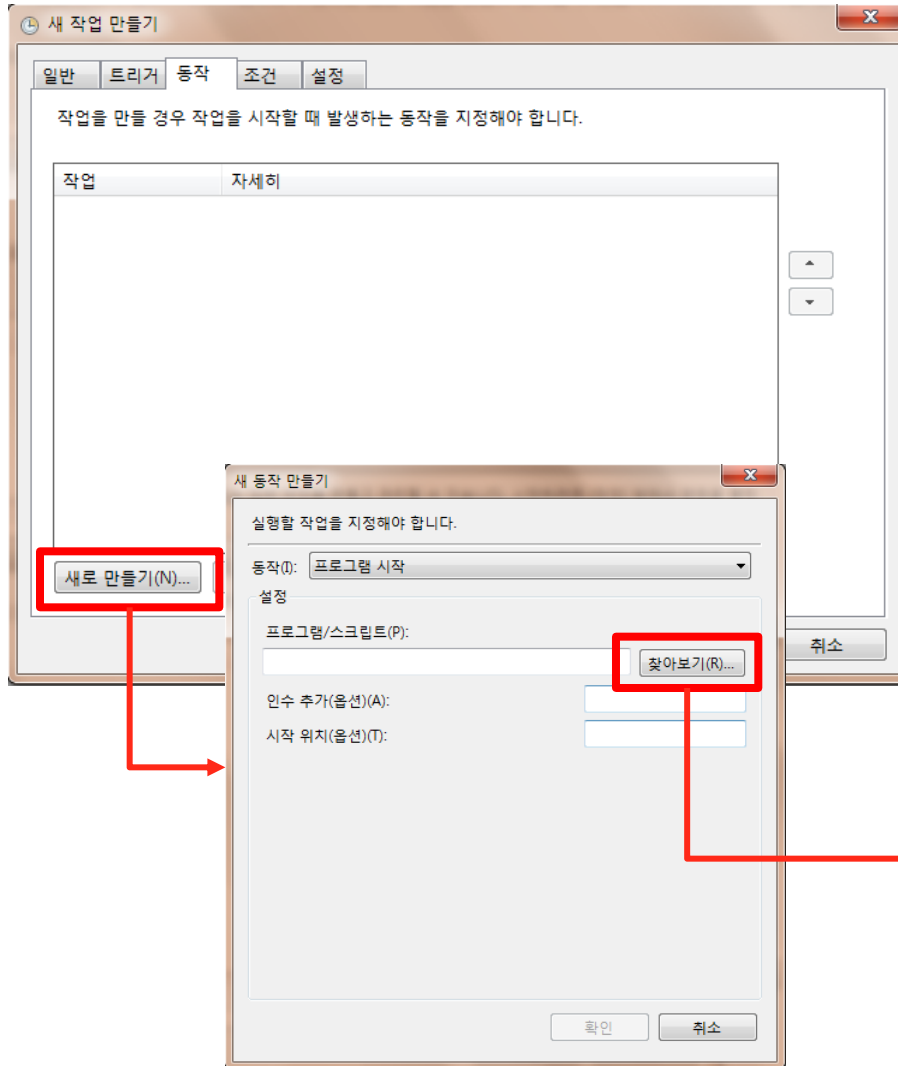
## • 새 작업 만들기 - 일반



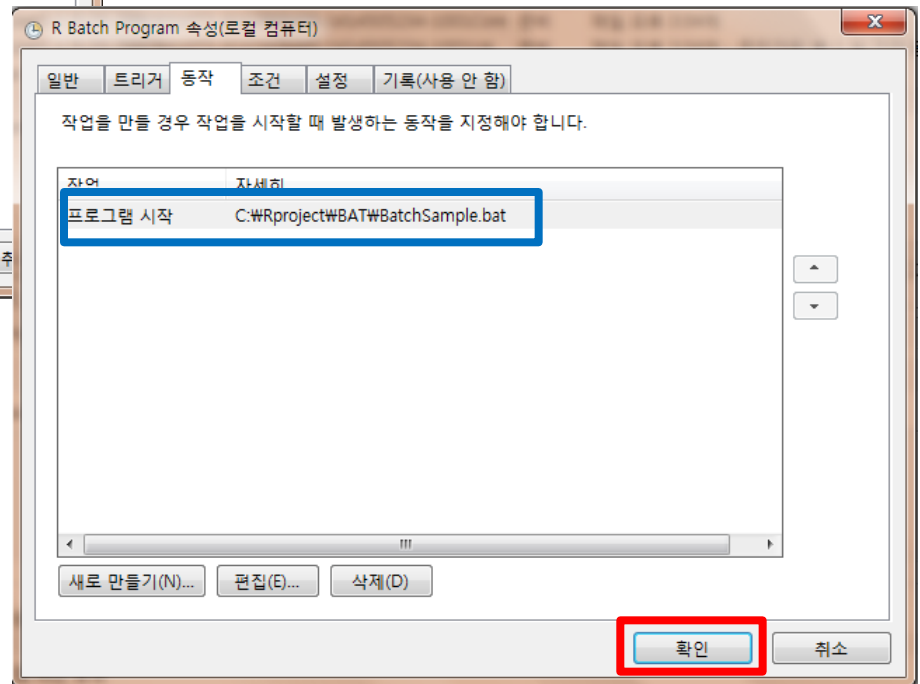
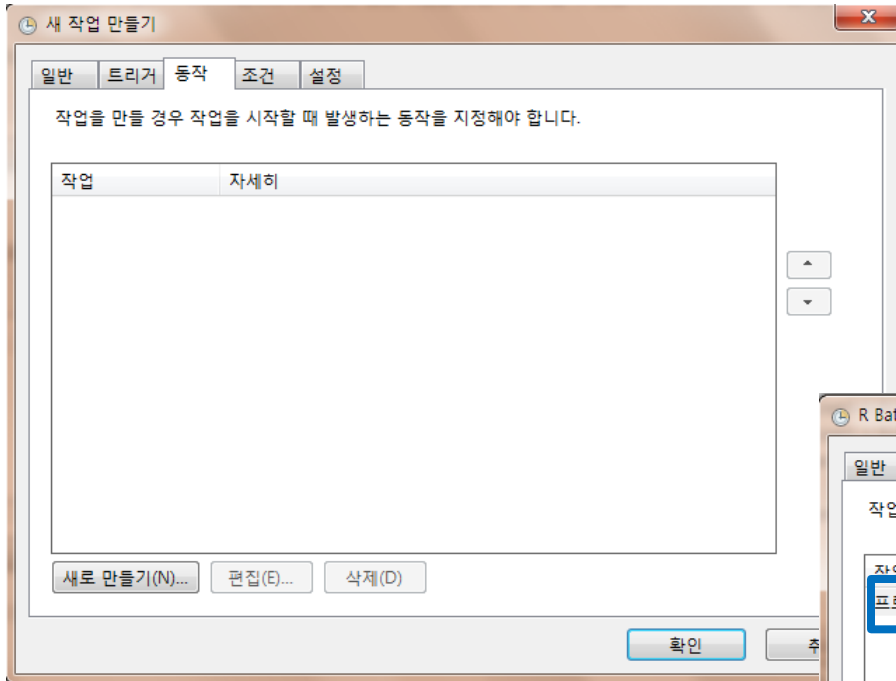
## • 새 작업 만들기 - 트리거



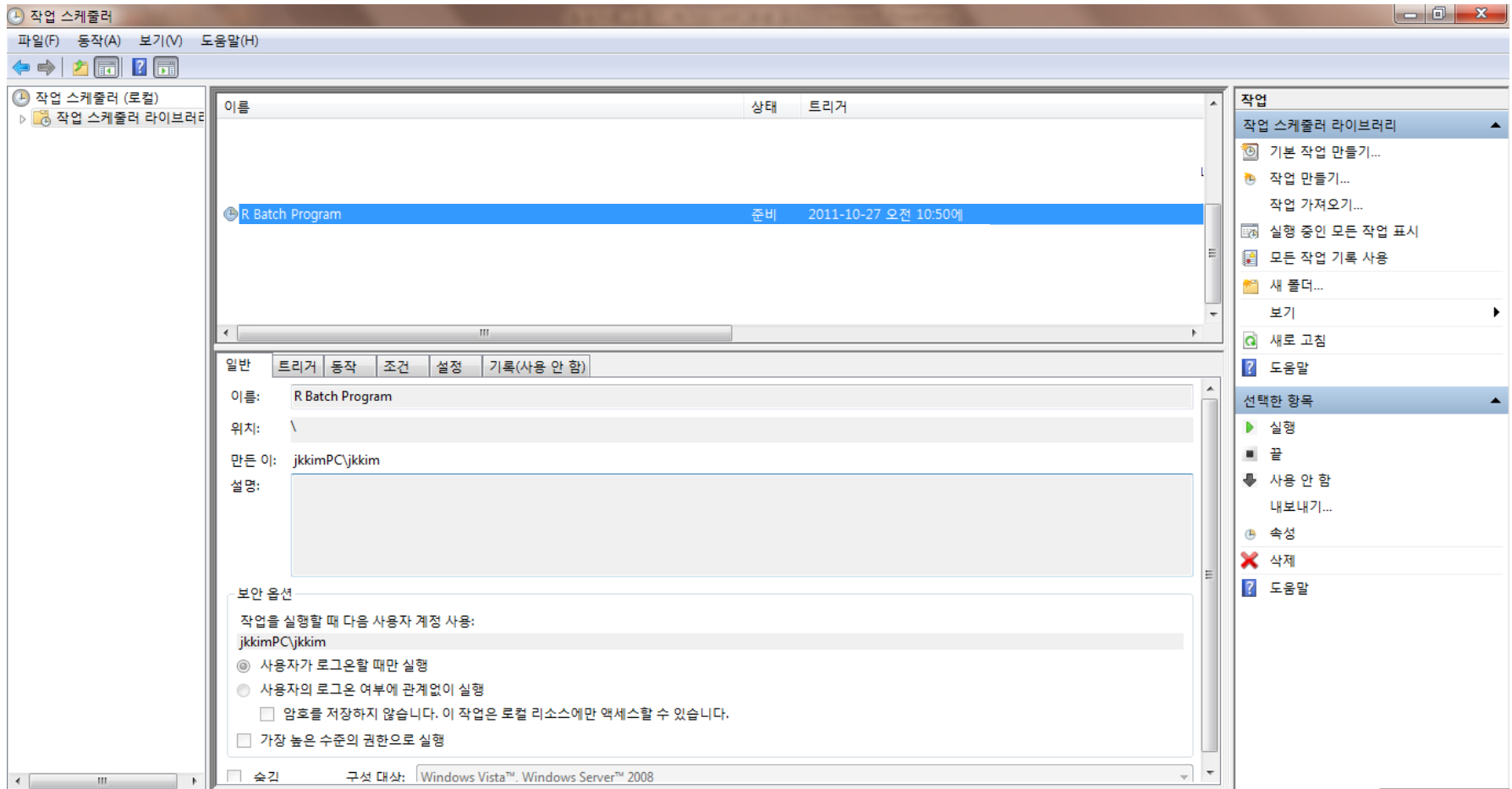
## • 새 작업 만들기 - 동작



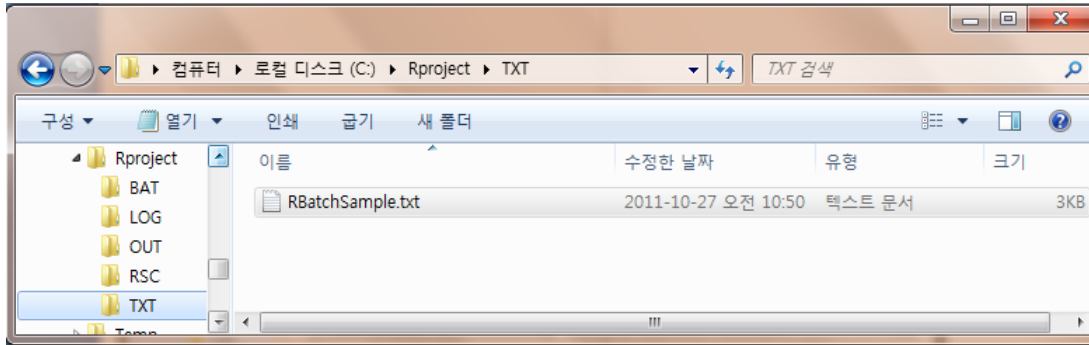
- 새 작업 만들기 - 동작



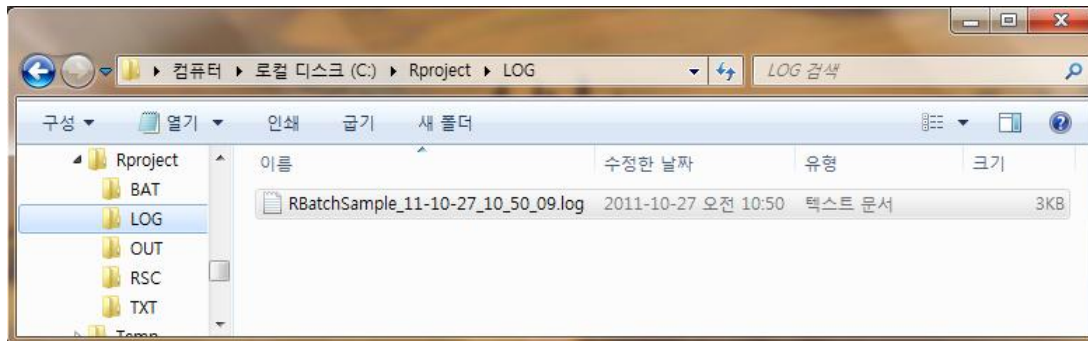
## • 작업 스케줄러 – 완료



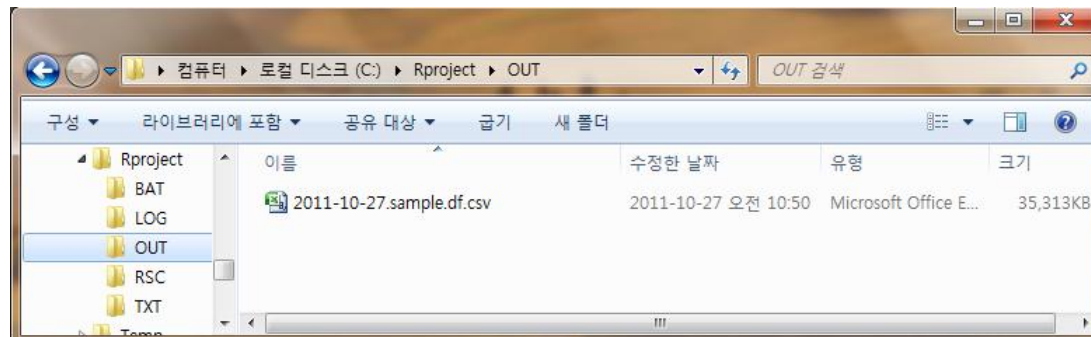
## Batch 작업 결과



- TXT 디렉토리



- LOG 디렉토리



- OUT 디렉토리

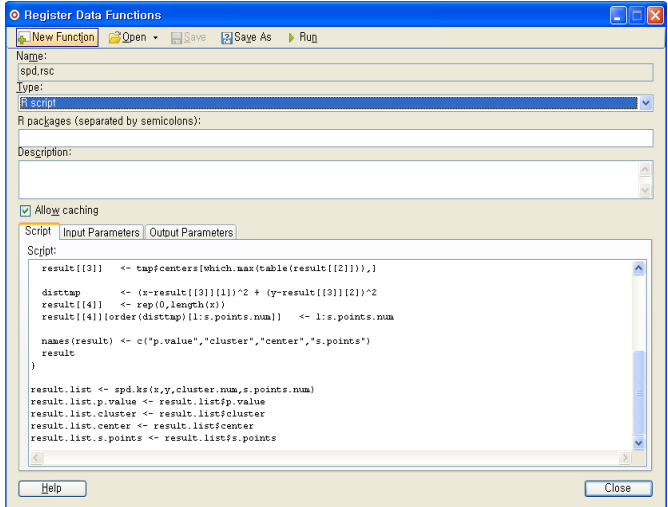
**Part I : DBMS 인터페이스**

**Part II : 데이터 Manipulation**

**Part III : Batch Program**

**Part III : Visualization Tool과의 연계**

## Part IV : Visualization Tool과의 연계

대상	인터페이스 방식	R 관련 패키지	Visualization Tool 관련 모듈	Interface 정의
Spotfire	Spotfire 모듈 이용	rJava	TIBCO Spotfire Statistics Services Local Adapter 3.3.0	

### 관련 Tool 및 Package

**Spotfire Server** : 분석함수의 저장 및 제공

**Spotfire Enterprise Developer** : 분석함수의 작성 및 실행 결과를 표출

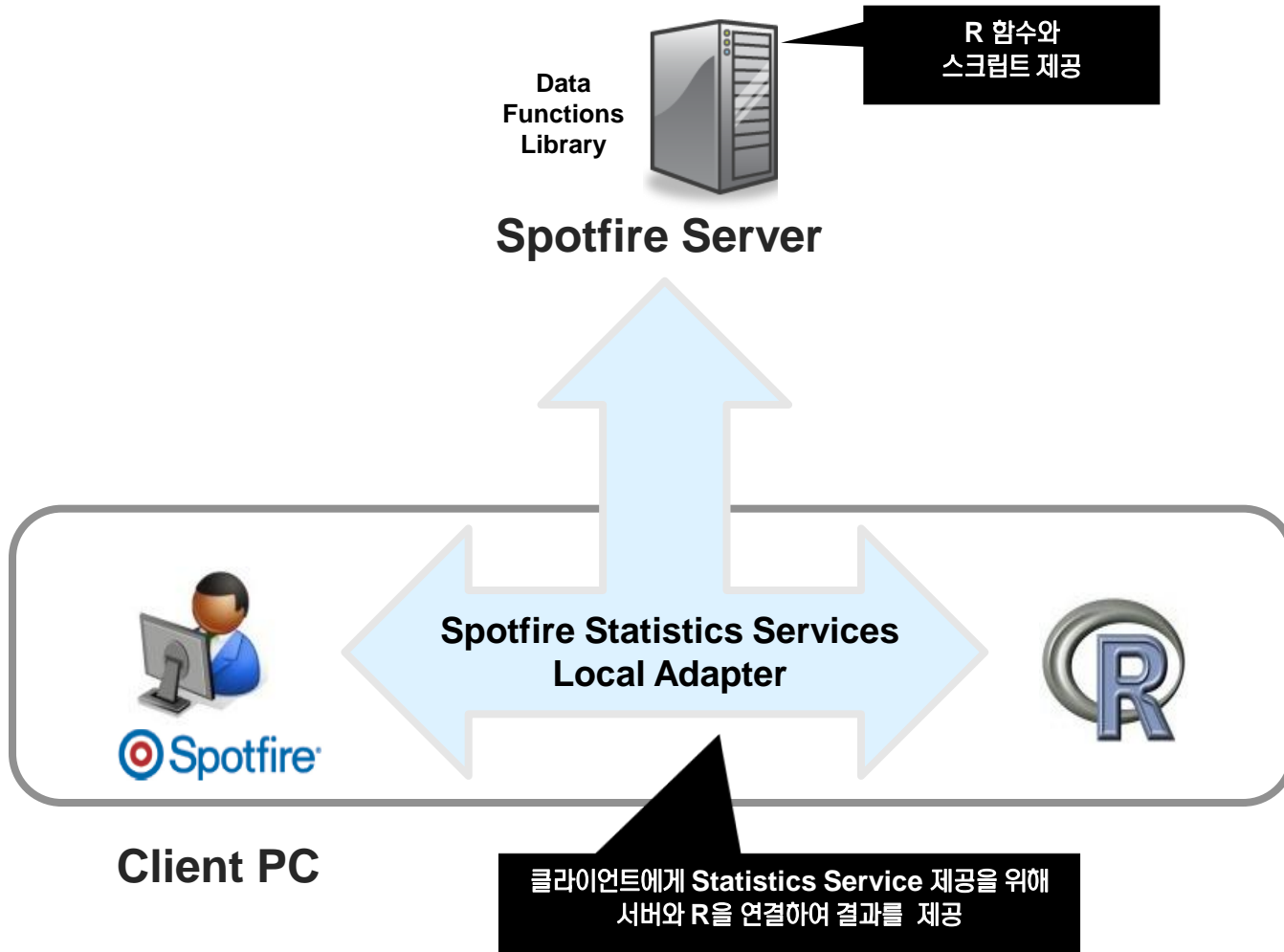
**Spotfire Statistics Services Local Adapter** : Spotfire 서버와 클라이언트 PC와의 분석 함수 연동

**R** : 클라이언트 PC에서 통계분석을 실행

- **rJAVA package가 설치되어야 함** : `install.packages("rJava")`

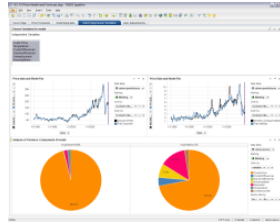


## Spotfire와 R 연동 Architecture



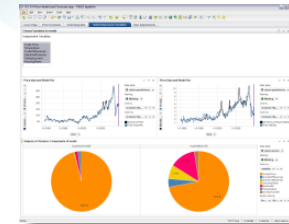
## Spotfire와 R 연동 작업 흐름

Managers, Consumers,  
Executives



One-click 전개로  
web에서 적용

Centrally-managed  
application based  
on best practices

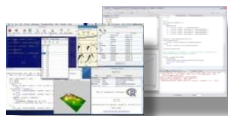


코딩지식 필요 없음

Statisticians



R을 이용하여 분석  
개발



Spotfire Statistics Services

Data Functions 사용을  
위해 구성된 화면을 설정

Data Function 을  
서버에 저장

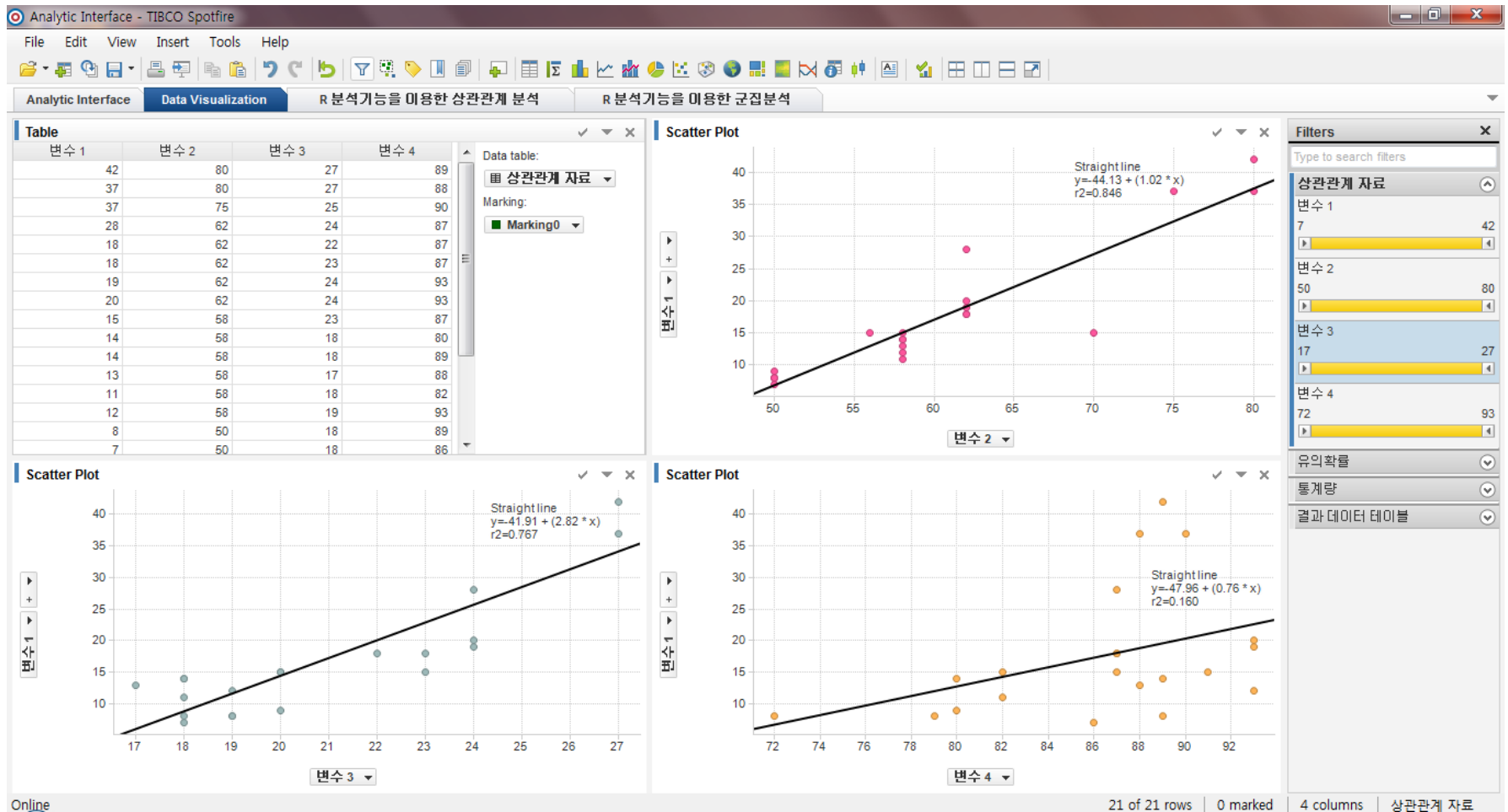


application  
authors를 위한 화면

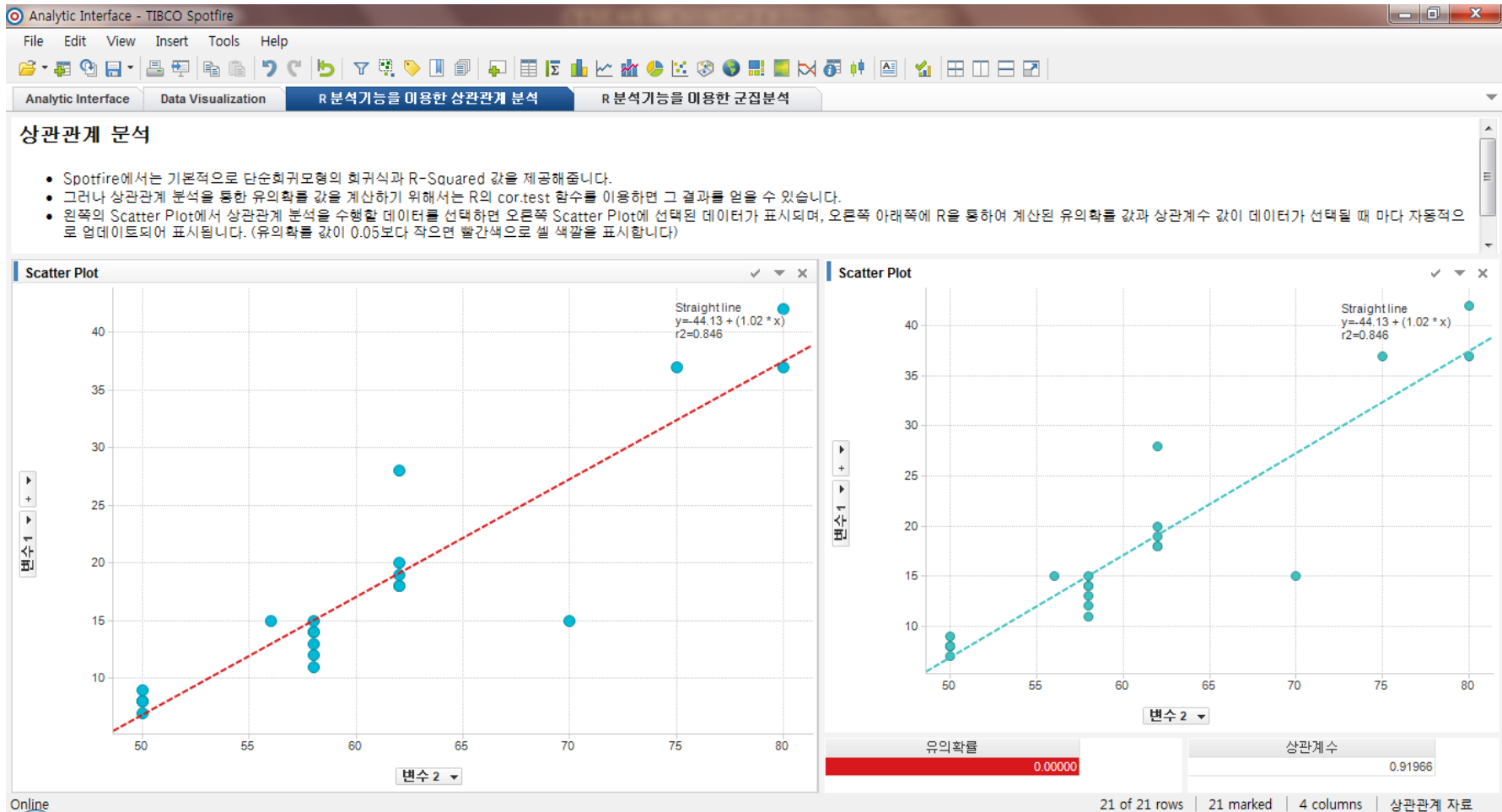
Analysts,  
Application Authors



# Data Visualization



# 상관관계 분석





감사합니다

김준기

**Tel : 010-7109-7291**

**email : [jkkim@begas.co.kr](mailto:jkkim@begas.co.kr)**