碩士學位論文

데이터베이스에 연동한 통계패키지의 성능평가

濟州大學校 大學院

電算統計學科

姜 吉 南

2004年 6月

데이터베이스에 연동한 통계패키지의 성능평가

指導敎授 金 鎭 孝

姜 吉 南

이 論文을 理學 碩士學位 論文으로 提出함.

2004年 6月

姜吉南의 理學 碩士學位 論文을 認准함.

審査委員長 _____ (印)

委 員 _____ (印)

委 員 _____ (印)

濟州大學校 大學院

2004年 6月

Performance evaluation of the statistical Package which is connected with Database

Gil-Nam Kang

(Supervised by professor Jinhyo Kim)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE.

DEPARTMENT OF COMPUTER SCIENCE AND STATISTIC
GRADUATE SCHOOL
CHEJU NATIONAL UNIVERSITY

June 2000

목 차

List of Figures

List of Tables

Abstract

As performance of a computer improves, a large amount of information is collected with a form of data by large database. It was transformed into data to get useful information because of the development of related technology this information and it is continued analysis and a related researcher to deal with this data. In this way a statistical package is used with large data, and analysis and the frequency that must be dealt with are increasing gradually. Therefore, characteristics of database remote database is accessed to in a local computer with data accumulated by large database, and to change quickly must be reflected.

The SAS(ver8.0), SPSS(ver10.0), S-PLUS(S-PLUS 2000), EXCEL(EXCEL 2000), JMP(ver5.0.1), MATHEMATICA(ver4.0) which is a statistical package used generally is gone for, and remote database is accessed to through ODBC, and cross-check to divide which size of the largest data that it is possible becomes how does analysis and a process in this paper.

I. 서 론

컴퓨터의 성능이 향상됨에 따라 대량의 정보가 자료의 형태로 대용량 데이터베이스에 수집되어 관리되고 있다. 이러한 자료를 활용하여 유용한 정보를 얻기 위한 연구와 노력이 계속되고 있다. 이렇게 빠르게 변화하는 대용량 데이터베이스 자료의 분석 및 처리가 off-line으로 행하여지고 있기 때문에, 거의 실시간 형태로 저장되어지고 있는 데이터베이스의 특징을 반영하지 못하는 경우가 빈번하게 발생하고 있다. 그래서 빠르게 변화하는 대용량의 데이터베이스의 자료를 분석 및 처리하기 위해서 off-line이 아닌 on-line으로 데이터베이스에 접속하여 빠르게 변화하는 데이터베이스의 특징을 반영하여야 한다. on-line으로 데이터베이스에 접속하여 데이터를 분석 및 처리하기 위해서는 로컬(local) 컴퓨터에서 다양한 통계패키지들을 이용하여 원격(remote) 데이터베이스 테이블에 직접 액세스(access)하여야 한다. 그래서 휴대 가능한 크기의 데이터만을 개인용 컴퓨터에서 통계패키지를 활용하여 분석 및 처리하는 것이 아니라 원격의 데이터베이스에 저장된 대용량의 데이터를 액세스(access)하여 통계패키지를 사용해야한다. 로컬(local) 컴퓨터에서 원격(remote) 데이터베이스를 액세스(access)하기 위해서는 ODBC(Open DataBase Connectivity) 데이터 원본 관리자에 데이터베이스 ODBC와 관련한 DSN(Data Source Name)이 등록되어 있어야 한다.

본 논문에서는 일반적으로 많이 사용되고 있는 통계패키지들이 분석 및 처리 가능한 데이터 크기가 얼마나 되는지를 비교 검토하여, 데이터의 크기에 따라서 적절한 통계패키지를 선택하여 사용한다면 시간과 비용적인 측면에서 도움을 주고자한다. 통계패키지와 관련된 국내 논문으로는 개인용 컴퓨터에서의 통계패키지의 선택과 활용(김병천, 1987), EDA 기능을 중심으로 한 패키지의 비교연구(허명회, 정진환, 1990), 통계패키지에서의 시계열 분석 방법의 비교연구(김수화, 김승화, 조신섭, 1994), 통계적 공정관리를 위한 주요 통계패키지의 비교(조신섭, 신봉섭, 1997), 통계처리용 소프트웨어 패키지의 품질 비교에 관한 연구(이상석, 윤민석, 1999)와 수치해석을 이용하는 통계계산에 사용되는 패키지의 기능별 비교검토(김진효, 1998)가 있는데, 각각 특정분야의 초점을

맞추고 비교연구를 하였다. 위의 나열된 논문과 비교하여 본 논문에서는 데이터 크기를 기준으로 비교연구를 하였다. 물론 이들 통계패키지는 PC환경 뿐 아니라 여러 가지 운영체제나 플랫폼에서도 가능하지만 본 논문에서는 PC의 Windows version을 사용하였다. 본 논문에서 고려된 데이터베이스는 MS-SQL Sever와 MySQL Sever 이고 통계패키지는 SAS System(ver8.0), SPSS(ver10.0), S-PLUS(S-PLUS 2000), JMP(ver5.0.1), EXCEL(excel 2000), MATHEMATICA(ver 4.0)이다.

본 논문의 구성은 다음과 같다. 제2장에서는 대용량의 데이터를 분석 및 처리하기 위한 통계패키지들을 소개한다. 제3장에서는 데이터베이스 액세스를 위해 필요한 요소인 DSN(Data Source Name)등록에 대하여 기술하였고, 제4장에서는 성능평가를 위하여 MS-SQL과 MySQL 두 종류의 데이터베이스와 통계패키지별 연동 과정을 설명하였다. 그리고 제5장에서는 통계패키지별 처리 및 분석 가능한 최대 데이터 사이즈 결과를 보인 후 마지막 장에서 결론을 맺는다.

II. 이론적 배경

1. 통계패키지 소개

1) SAS

SAS시스템은 통계적 자료분석을 위하여 1976년 미국의 North Carolina 주립대학의 James H. Goodnight에 의하여 개발되었다. SAS라는 명칭은 Statistical Analysis System의 첫 문자를 따서 명명되었지만 그간 많은 발전과 확장을 거쳐 현재는 SAS 시스템으로 불릴만큼 활용범위가 넓어졌다. 이를 반영하기 위하여 SAS의 어원도 Strategic Applications Software으로 바꾸어 사용하고 있다. SAS는 Version 6.08 이후로부터 윈도우 환경으로 발표되기 시작한 이래 2004년 현재 Version 8.1까지 발표된 상황이다. 현재 SAS는 통계분석을 위한 소프트웨어일 뿐만 아니라 데이터마이닝(data mining), 데이터웨어하우스(data warehouse)등을 포함하여 응용프로그램 개발에 필요한 여러 가지 제품들을 포함하고 있다. SAS는 각종 데이터를 올바른 정보(Right Information)로 변환하여 정보를 필요로 하는 사람(Right Person)에게 가장 적절한 시기(Right Time)에 정보를 제공할 수 있는 통합 애플리케이션 소프트웨어라고 한다. (http://ksnam.toenter.net/nks/sas.htm)

2) SPSS

SPSS는 Statistical Package for Social Science의 약자로써 여러 학문분야, 특히 사회과학분야에서 얻어지는 각종 자료를 통계학적으로 분석하기 위하여 컴퓨터를 이용하여 복잡한 자료를 편리하고 쉽게 처리 분석할 수 있도록 만들어진 통계분석 전용 소프트웨어이다. SPSS를 이용하면 관심있는 변수의 각종 기술통계량을 구할 수 있을 뿐만 아니라 교차분석, 상관분석, 회귀분석, 분산분석, 더 나아가서 판별분석, 요인분석 등 복잡한 다변량분석을 할 수 있으며, 수집된 자료를 원하는 보고서 형태로 출력할 수도 있다. SPSS을 이용하면 연구자들은 원하는 통계결과를 신속하고 용이하게 얻어낼 수

있다. 이 통계 프로그램은 단순한 기술통계로부터 복잡한 다변량 통계분석까지 원하는 결과를 비교적 쉽게 얻어 낼 수 있게 해준다. 이뿐만 아니라, 일반 데이터베이스에서 작성된 자료를 이용할 수 있는 장점도 지니고 있다. SPSS는 크게 DOS용과 Windows 용으로 나눌 수 있으며, 예전에는 SPSS/PC+로 알려진 DOS용을 주로 사용하였으나, 하드웨어의 발전과 더불어 컴퓨터 운영체계가 Windows로 급격하게 변함에 따라 1993년에 MS Window용 SPSS for Windows 5.0이 처음 개발되었고, 2004년 현재 SPSS for Windows 12.0까지 개발되어 시판되고 있다.

3) S-PLUS

S-PLUS는 미국 AT&T사에서 자체개발하여 S라는 언어로 UNIX상에서 사용되던 것을 확장 및 보강하여 Statistical Sciences사에서 상업용으로 내놓은 언어로서 Windows 체제에서 사용할 수 있고, 탐색적인 자료분석, 통계적 자료분석, 그래픽방법 등에 대하여 다양한 패키지 기능을 제공하고 있다. 뿐만 아니라 고급언어인 C, FORTRAN, 또는 자체 언어로 작성된 절차를 첨가하는 것을 가능하게 하고 있다. S-PLUS의 장점중 하나는 객체지향적(Object oriented)인 언어로서 사용자가 원하는 일련의 과정을 보다 쉽게 구현할 수 있다는 것이다. 또한 1200개 이상의 내장함수(built-tin function)를 포함하고 있으므로 이것들을 수정(modify)하여 보다 효과적인 대화형식(interactive job)의 작업을 수행할 수 있다.

4) EXCEL

엑셀은 윈도우환경에서 표 계산, 수식작성, 데이터분석 등의 기능을 활용하는 스프레드시트(Spread Sheet)를 사용한다. 스프레드시트는 말 그대로 여러 장의 종이를 펼쳐 놓은 작업공간을 말한다. 엑셀이 나오기 이전 이미 1978년에 개발된 '비지칼크(VisiCalc)' 이후 개인용 컴퓨터의 운영체제로 사용된 DOS용 'Lotus 1-2-3'가 폭넓게 사용되다가 유사한 'QuatroPro'가 등장하였다. Windows 운영체제가 개발된 이후 MS 사에서 개발한 엑셀이 스프레드시트의 표준으로 인식되고 있다.

5) JMP

JMP는 1989년 SAS Institute에 의해 그 첫판이 발표된 통계분석용 패키지 프로그램이다. JMP는 SAS 시스템의 한 부분의 아니며, SAS의 간략한 버전 또한 아니다. 어떤 면에서는 SAS/INSIGHT라 불리는 SAS의 add-on제품과 관련이 있지만 그와 또 다른 독특성을 지니고 있다. JMP는 자료분석을 수행한 통계량과 탐색적 그래픽 인터페이스 화면을 상호간에 연계하여 수행한다. 따라서 사용자에게 매우 배우기 편리함과 분석의 편리성을 갖추고 있다. 현재 JMP는 2000년 후반부에 발표한 4.0버전에서 2004년 현재 5.0버전에 이르고 있다. JMP 4.0버전은 기존 버전과는 혁신적으로 다르게 매우 새롭고 막강한 기능으로 업그레이드되었다. 기존의 버전에서는 매킨토시용으로 제작된 것의 윈도우 버전으로의 변환으로 화면의 모습에서 매킨토시 프로그램의 모양을 상당부분 지니고 있었으나, JMP 4.0 버전에서는 완벽한 윈도우용 프로그램으로서의 면모를 보여주고 있다.

6) MATHEMATICA

매스매티카는 1989년 미국의 이론 물리학자 Stephen Wolfram에 의해서 버전 1이 매킨토시 컴퓨터용으로 최초로 개발되었고, 그 후 다양한 하드웨어 및 운영체제에 적용될 수 있도록 개발되어 사용되고 있다. 매스매티카는 자연과학이나 공학에서 발생하는 제반 문제가 일단 수학적 모델링이 된 후 다음과 같은 기능을 이용하여 해결을 해주는 소프트웨어이다. 첫째, IMSL 또는 MATLAB 등에서와 같이 수치적으로 계산하여 결과를 구할 수도 있다. 둘째, MACSYMA 또는 MAPLE 등에서와 같이 부호연산(symbilic manipulation)이 가능하여 결과를 기호로 얻을 수 있다. 셋째, 결과를 MATLAB 또는 MATHCAD에서처럼 그래픽으로 처리하여 이해를 용이하게 할 수가 있다. 넷째, 기존의 언어인 BASIC, C, FORTRAN처럼 일련의 계산과정을 매스매티가 언어를 이용하여 프로그램할 수 있다. 마지막으로 인터페이스(interface) 기능이 마련되어 있어 다른 프로그램과의 자료 교환이 가능하며 매스매티카에 내장된 함수(built-in function)를 외부에서 서브루틴으로 이용할 수가 있다.

2. 다양한 데이터베이스 연동 방식

서로 다른 데이터베이스간 연동 방법은 크게 직접연동(direct connectivity) 과 간접연동(indirect connectivity)으로 구분 될 수 있다. 직접연동은 연동하고자 하는 대상 데이터베이스에 직접 접근하여 질의(question) 및 조작(operation)을 하는 것을 의미하며, 간접연동은 연동 데이터베이스에 직접 접근하는 것이 아니라, 상대측에 데이터 요청 메시지를 보내면 상대측에서 이를 처리하여 요구받은 데이터를 회신해주는 방식을 의미한다. 데이터베이스 직접연동(direct connectivity)에 사용되는 기술로는 복잡한 서로 다른 환경에서 응용 프로그램과 운영체제 간에 원만한 통신을 이룰 수 있게 해주는 소프트웨어인 미들웨어(middleware) 기술이 있으며, 직접연동(indirect connectivity)방식의 일환이면서 데이터베이스간 데이터 구조와 제약 조건에 대한 명세(specification)를 기술한 스키마(schema)의 이질성을 극복하고자 하는 기술이 멀티데이터베이스 기능이다. 그리고 데이터베이스 간접연동을 위한 기술에는 데이터 변환 기술이 있다.

1) 서로 다른 데이터베이스 접속을 위한 미들웨어

과거에는 데이터베이스에 접속(connectivity)하기 위해서는, 개발자들이 데이터베이스에 접속을 위하여 각각의 데이터베이스 벤더(vendor)들이 내놓은 로우 레벨의 API(Application Programming Interface)를 알아야지만 가능하였다. 이 API를 통하여 질의를 보내고 처리해야 했는데 이러한 어려움으로 인하여 각각의 데이터베이스에 대한 API를 모두 알 필요 없이 공통의 API를 알기만 하면 모든 데이터베이스에 공통으로 접속할 수 있는 API가 필요하게 되었다.

CLI(Call Level Interface)는 인터페이스를 통하여 수많은 데이터베이스로의 접근을 제공함으로써, 다양한 데이터베이스에 걸치는 공통의 API이다. CLI는 비영리 그룹인 X/OPEN과 SQL Access Group에 의하여 SQL CLI가 1992년에 처음으로 표준화가 되어 1993년에 ISO 국제표준으로 제정되었다. CLI는SQL을 사용하는 것에서 알 수 있듯이 주로 관계형 데이터베이스의 연결에 사용된다.

SQL CLI를 기반으로 등장한 API들로는 ODBC(Open DataBase Connectivity), JDBC(Java DataBase Connectivity), OLE-DB 등이 있다.

(1) ODBC(Open DataBase Connectivity)

ODBC는 데이터베이스를 액세스하기 위한 C언어 응용 프로그램 인터페이스로서 프로그램 내에 ODBC문장을 사용하면 MS-Access, dBase, DB2, Excel, Text 등 여러 가지 종류의 데이터베이스를 액세스 할 수 있다.

ODBC는 X/OPEN과 ISO/IEC의 CLI(Call Level Interface)스팩을 기본으로 하고, 데이터베이스 접근 언어로서 SQL을 사용한다. ODBC는 SQL Access Group에 의해 만들어 졌으며, 1992년 9월에 처음 나왔다. 처음엔 마이크로소프트가 윈도우용 ODBC 제품을 공급했지만, 이제는 유닉스, OS/2 및 매킨토시 등을 위한 버전들 역시 생겨났다. 마이크로소프트가 ODBC의 제안자이자, 프로그램의 공급 지원을 맡고 있다.

ODBC는 프로그램들이 데이터베이스의 독점적인 인터페이스에 대해 알지 못하더라도, 데이터베이스 액세스를 위한 SQL 요청을 사용할 수 있게 한다. ODBC는 SQL 요청을 받아서, 그것을 개개의 데이터베이스 시스템들이 이해할 수 있는 요청으로 변환한다. 이를 위해서는 ODBC 소프트웨어 외에, 액세스할 각 데이터베이스마다 별도의 모듈이나 드라이버가 필요하다.

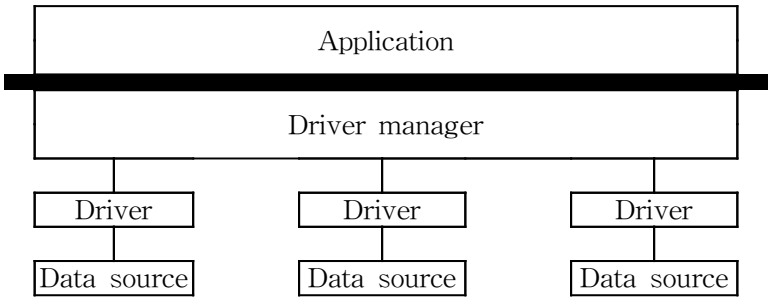


Figure 1. ODBC 구성도

Application	· ODBC function들을 call
Driver Manager	· 로드(load) 및 un-로드 드라이버 · ODBC function call을 처리 및 드라이버에 넘김
Driver	· ODBC function call 처리 · SQL 요청을 데이터 소스로 넘김 · 결과를 응용에게 되돌림 · DBMS 마다 ODBC 드라이버를 제공
Data source	· 사용자가 접근하고자 하는 데이터와 관련한 OS, DBMS, 네트워크 플랫폼들로 구성

Table 1. ODBC 구성별 역할

<Table 1>은 ODBC 구성요소들이 수행하는 역할이며, <Figure 1>과 같이 ODBC 애플리케이션(Application)은 사용자가 작성한 윈도우 애플리케이션이고, ODBC 드라이버 관리자(Driver Manger)는 드라이버를 연결하는 연결자이다. ODBC 인터페이스를 제공하는 애플리케이션들은 ODBC 드라이버가 존재할 데이터 소스를 접근할 수 있다. ODBC 데이터 소스 드라이버는 ODBC의 함수를 부르는 DLL(Dynamic Link Library)로서 부분적으로 데이터 소스에 접근하는데 사용된다. 즉 ODBC 인터페이스는 다음과 같이 정의 할 수 있다. 첫째, ODBC 함수의 라이브러리는 데이터 소스에 연결하고 SQL 문을 실행하고 그 결과를 반환한다. 둘째, 데이터 소스에 접근하는 표준화된 경로와, 데이터 타입에 대한 표준화된 표현, 에러에 대한 표준화된 형식을 제공한다.
그 다음으로 드라이버는 각 데이터베이스 회사에서 제공한 파일로서 바로 상위 계층인 ODBC 드라이버 관리자와만 통신한다. 마지막으로 데이터베이스는 각 회사의 여러 종류의 데이터베이스를 의미한다. 이런 구조로 인해 프로그래머는 각 데이터베이스에 대해 모두 알 필요가 없고 ODBC 드라이버 관리자의 구조만 잘 알고 있으면 되는 장점이 있다.

ODBC를 사용함에 있어서 데이터베이스의 특징에 따라 약간의 차이점이 있지만 그 차이가 그리 크지 않기 때문에 거의 모든 데이터베이스를 연결하는 애플리케이션 또한 ODBC를 지원하고 있다. 그러나 애플리케이션은 데이터베이스를 연결함에 있어서

ODBC뿐만 아니라 자체 연결자를 사용할 수도 있다. 다음 <Figure 2>처럼 데이터베이스를 연결함에 있어서 두 가지 통로를 다 가지고 있는 것이다. 표준화를 위한 여러 단계 처리시 발생되는 시간 지연으로 인하여 ODBC를 사용할 때 일부 데이터베이스에서 처리속도가 느려지는 문제점이 있다. 그러나 이때 자체 연결자(Native Connectivity)를 사용하면 눈에 띄게 속도 향상을 볼 수 있다.

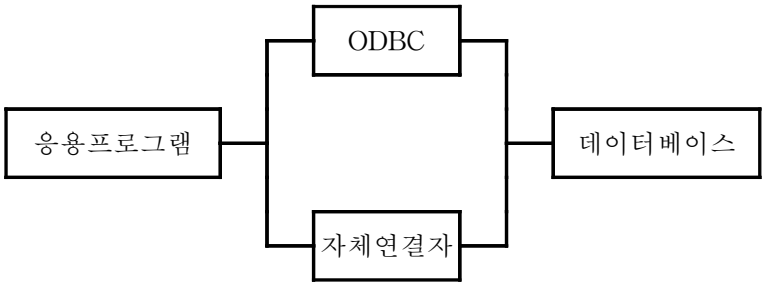


Figure 2. 자체연결자 사용

새로운 CORBA의 POS(Persistent Object Service)는 CLI와 ODBC 둘 모두를 포함한다. 자바 언어와 JDBC 응용 프로그램 인터페이스를 이용하여 프로그램을 작성할 때, ODBC로 액세스할 수 있는 데이터베이스에 접근하기 위해서는 일종의 브리지 프로그램인 JDBC-ODBC가 포함되어 있는 제품을 이용할 수 있다.

(2) JDBC(Java DataBase Connectivity)

JDBC는 자바로 작성된 프로그램을, 일반 데이터베이스에 연결하기 위한 응용프로그램 인터페이스 규격이다. 이 응용프로그램 인터페이스는 데이터베이스 관리시스템에 넘겨질 SQL 형태의 데이터베이스 접근요구 문장을, 각 시스템에 맞도록 바꾸어준다. 처리 결과도, 이와 비슷한 인터페이스를 통해 얻게 된다.

JDBC는 ODBC와 아주 유사해서, 조그만 연결 프로그램만 있으면, ODBC 인터페이스를 통해 데이터베이스에 연결하는 JDBC 인터페이스를 사용할 수 있다.

JDBC는 실제로도 두 계층의 인터페이스로 구성되어 있다. 주 인터페이스 외에도 JDBC “manager”에서 나온 API가 있는데, 이것의 역할은 개별 데이터베이스 제품의

드라이버들과 차례대로 통신을 하는 것이다. 이때, 만약 필요하다면 JDBC-ODBC bridge와, 그리고 자바 프로그램이 원격 데이터베이스를 액세스하기 위해 네트워크 환경에서 실행되고 있다면 JDBC 네트워크 드라이버 등과의 통신도 수행한다.
JDBC는 프로그래머가 SQL 요구를 만드는데 사용할 일련의 객체지향 프로그램의 클래스들을 정의하고 있으며, 별도의 추가 클래스 모음집에 JDBC 드라이버 API가 기술되어 있다. 자바 데이터 형식에 대응된 일반 SQL 데이터 형식들 대부분이 지원된다.

(3) OLE DB(Object Linking and Embedding)

ODBC가 관계형 데이터베이스에만 사용될 수 있는 한계로 인하여 OLE DB는 관계형/비관계형 모두 다 사용 할 수 있는 기술로, 마이크로소프트에서 제공하는 기술이다.

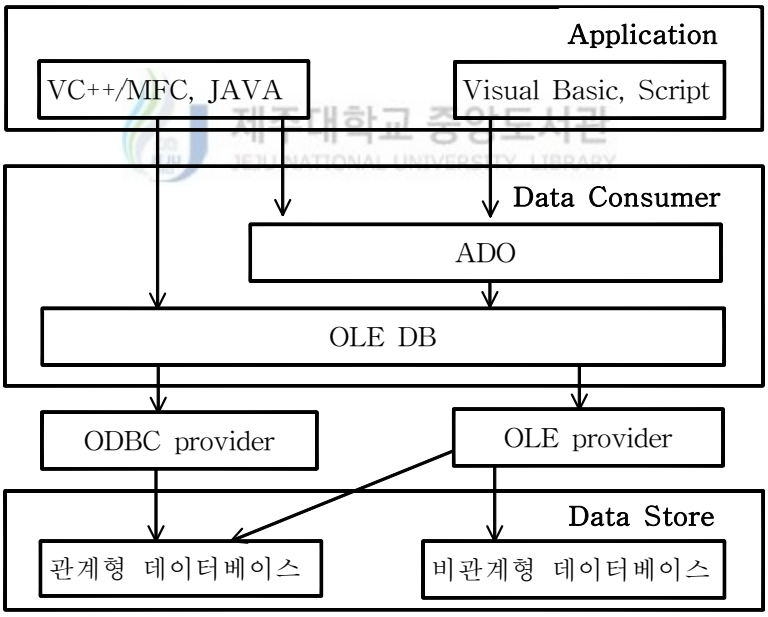


Figure 3. OLE DB 구조

<Figure 3> 과 같이 OLE DB의 구조를 보면 크게 데이터를 제공하는 Data Store와 데이터를 소비하는 Data Consumer로 나뉜다. ASP에서 Data Consumer 역할을 하는

것이 ADO이다. OLE DB를 사용하려며 API를 기존에 많이 사용되었던 DAO(Data Access Object), RDO(Remote Data Object)와 사용법이 비슷한 ActiveX 객체를 사용하였고, 이것이 바로 ADO(ActiveX Data Object)이다. 그리고 OLE Provider가 두 부분 사이에서 연결해 주는 역할을 하고 있다.

2) 멀티데이터베이스

멀티데이터베이스 시스템은 데이터베이스 시스템간 의미(semantic)의 차이를 처리하기 위한 방안으로, 다양한 다른 형태의 데이터베이스의 스키마를 통합한 전역 스키마(Global Schema)를 만들고 그 스키마에 대한 질의를 처리하는 구조를 갖는다.
<Figure 4>는 멀티데이터베이스의 일반적인 구조를 보여준다. 멀티데이터베이스는 통합하고자 하는 데이터베이스들에 대한 통합된 전체의 스키마(schema)만을 관리하고, 모든 사용자 데이터는 지역(local) 데이터베이스가 관리한다. 통합 스키마는 각각의 지역 데이터베이스의 스키마의 합병으로 구성되며 합병은 각각의 지역 데이터베이스들의 지역 스키마 충돌을 중재하면서 이루어진다.

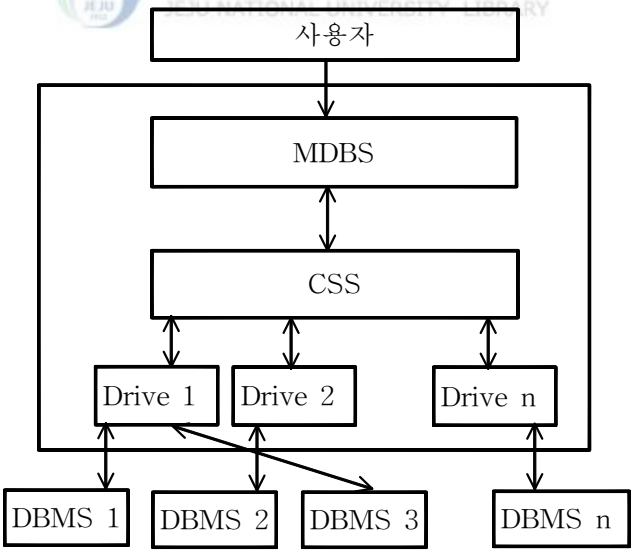


Figure 4. 멀티데이터베이스의 구조

이것은 다른 형태의 데이터베이스들을 하나의 시스템으로 통합하는데 있어서, 데이터베이스 시스템들 간의 스키마의 충돌을 어느 정도 해결할 수는 있으나, 초기 구축비용이 많이 소요되고, 대부분의 제품이 DBMS 회사에 의해 개발되었기 때문에 자사의 DBMS를 중심으로 타사의 DBMS를 통합하는 형태의 제품이 대부분이다.

3) 데이터 변환

데이터 변환 방법은 한 데이터베이스에서 사용되는 데이터를 다른 데이터베이스에서 사용할 수 있도록 변환하는 방법을 말한다. 손쉬운 방법은 데이터를 변환 프로그램을 이용하여 텍스트 형태로 전환한 후 이를 필요한 형태로 가공하여 필요한 데이터베이스로 적재(load)하는 것이다. DBMS와 DBMS, 파일 시스템과 DBMS간의 데이터 교환에 사용될 수 있다.

이 방법은 기술적으로 구현이 쉽고, 데이터의 교환 주기가 길거나 비정기적인 경우, 한번에 교환되는 데이터의 양이 많은 경우 효과적이다. 그러나, 데이터 변환 방법의 데이터의 포맷이 바뀔 때마다 변환 프로그램도 바꾸어 주어야 한다.

Ⅲ. 데이터베이스 연동을 위한 DSN 등록

원격의 데이터베이스를 액세스하기 위해서는 두 가지 요소가 필요하다. 하나는 로컬 컴퓨터의 ODBC 데이터 원본 관리자에 DSN을 등록하는 것이고, 다른 하나는 데이터베이스를 액세스하기 위한 데이터베이스 연동을 위해서는 SAS/ACCESS, SPSS Data Access Pack등의 소프트웨어가 필요하다.

원격 데이터베이스를 로컬 컴퓨터로 액세스하기 위해서는 ODBC(Open DataBase Connectivity)가 필요하다. ODBC는 데이터베이스를 액세스하기 위한 표준 개방형 응용 프로그램 인터페이스로서 ODBC를 이용하면 MS-SQL, MySQL 데이터베이스 외에 MS-ACCESS, DBASE, DB2, EXCEL, TEXT 등 여러 가지 종류의 데이터베이스를 액세스 할 수 있다. 따라서, 원격 데이터베이스를 액세스하기 위해서는 액세스할 각 데이터베이스에 해당하는 별도의 모듈이나 ODBC 드라이버가 필요하다. 즉, 로컬 컴퓨터에서 액세스하고자 하는 원격 데이터베이스의 모듈이나 ODBC 드라이버를 <Figure 5>와 같이 ODBC 데이터 원본 관리자의 DSN에 미리 등록해야 한다.

Figure 5. ODBC 데이터베이스 원본 관리자

DSN 등록은 사용자 DSN, 시스템 DSN, 파일 DSN 로 분류된다. 사용자 DSN은 다수

의 사용자가 컴퓨터를 사용하는 경우에 각 사용자에 맞게 DSN을 등록하는 것이고, 시스템 DSN은 사용자에 관계없이 DSN을 등록하며, 파일 DSN은 파일 형태로 DSN을 등록하는 것을 말한다.

1. MS-SQL 데이터베이스 액세스를 위한 DSN 등록

MS-SQL 데이터베베이스에 액세스하기 위해서는 우선 <Figure 5>와 같이 ODBC 관리자에서 추가버튼을 클릭하면 <Figure 6>과 같이 새 데이터 원본 만들기 창에서 데이터 원본을 설정할 드라이버를 선택한다.

Figure 6. 새 데이터 원본 만들기

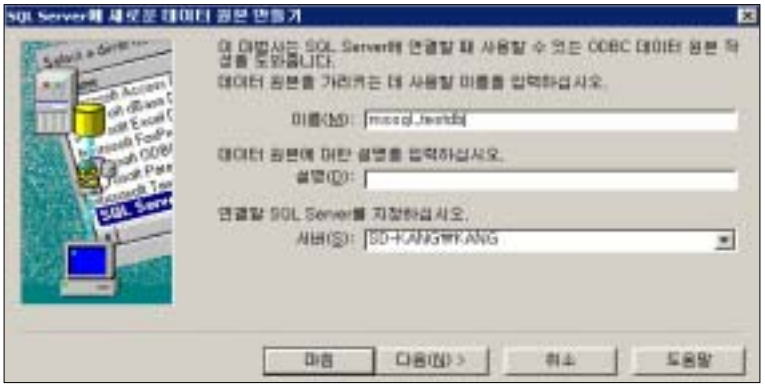


Figure 7. MS-SQL 데이터베이스 서버와 DSN 이름

<Figure 7>은 MS-SQL 데이터베이스를 액세스하기 위한 서버와 DSN 이름 등록화면으로, MS-SQL 데이터베이스를 사용하기 위한 DSN 이름인 mssql_testdb와 DSN 이름에 대한 설명, 데이터베이스 서버 이름과 사용할 데이터베이스 등에 대한 내용을 보여준다.

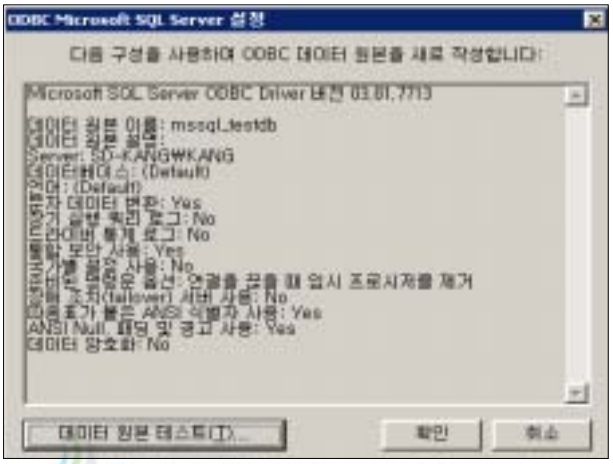


Figure 8. MS-SQL 데이터베이스 DSN 등록

<Figure 8>과 같이 DSN 등록이 제대로 이루어진 경우에 데이터 원본 테스트를 수행하면 <Figure 9>과 같은 결과를 볼 수 있다.

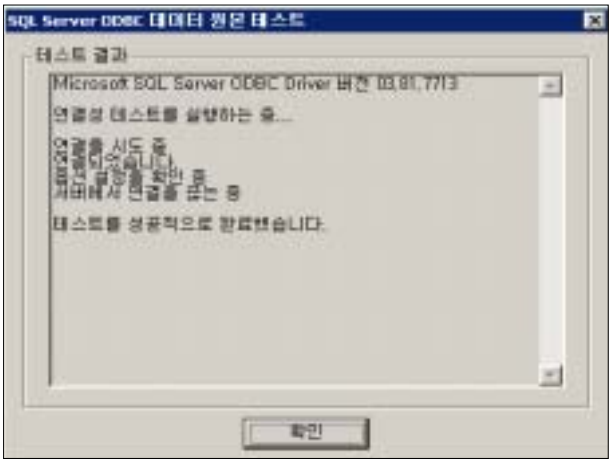


Figure 9. MS-SQL 데이터 원본 테스트 결과

2. MySQL 데이터베이스 액세스를 위한 DSN 등록

다음 <Figure10>은 MySQL 데이터베이스를 액세스하기 위한 서버와 DSN 이름을 등록하는 화면이다.

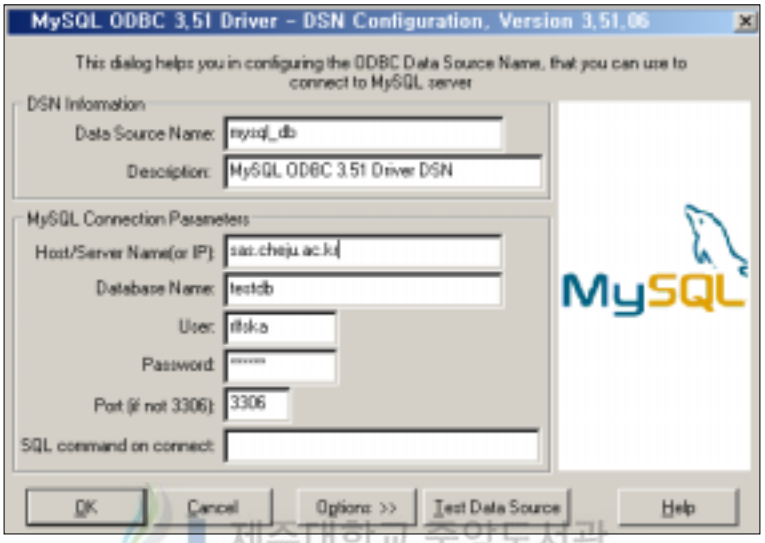


Figure 10. MySQL 데이터베이스 DSN 등록

DSN 정보에서 DSN 이름은 사용자가 임으로 정할 수 있으며, MySQL 연결 파라미터는 MySQL 데이터베이스가 존재하는 서버 이름 또는 IP를 입력하고, 액세스하고자 하는 데이터베이스 이름과 사용자, 패스워드를 입력한다. 입력이 완료된 후에 데이터 원본 테스트를 수행해보면 <Figure 11>과 같은 결과가 나타난다.

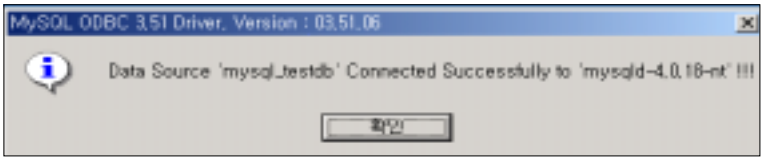


Figure 11. MySQL 데이터 원본 테스트 결과

Ⅳ. 통계패키지별 데이터베이스 연동

1. SAS에서 데이터베이스 연동

SAS/ACCESS ODBC 인터페이스로 ODBC에 등록되어 있는 DSN을 이용하여 데이터베이스 연동 과정은 다음과 같다.

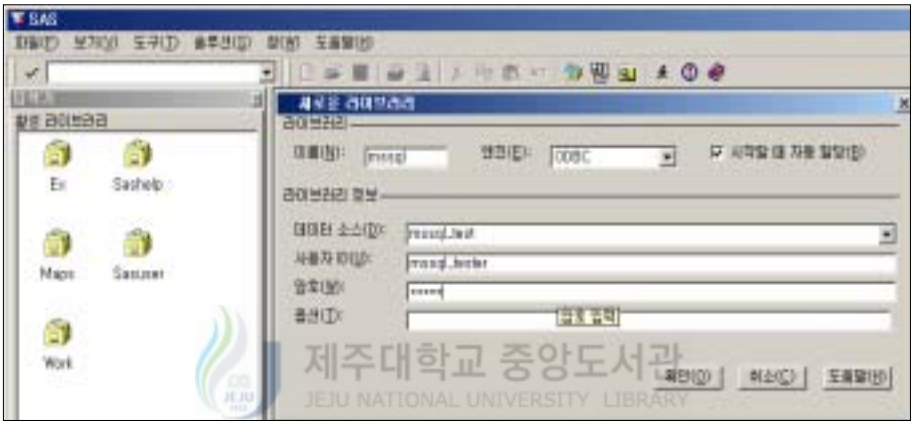


Figure 12. MS-SQL 데이터베이스 라이브러리 등록

데이터베이스를 라이브러리로 등록하는 방법은 일반적인 라이브러리 등록과 다른 차이를 보인다.

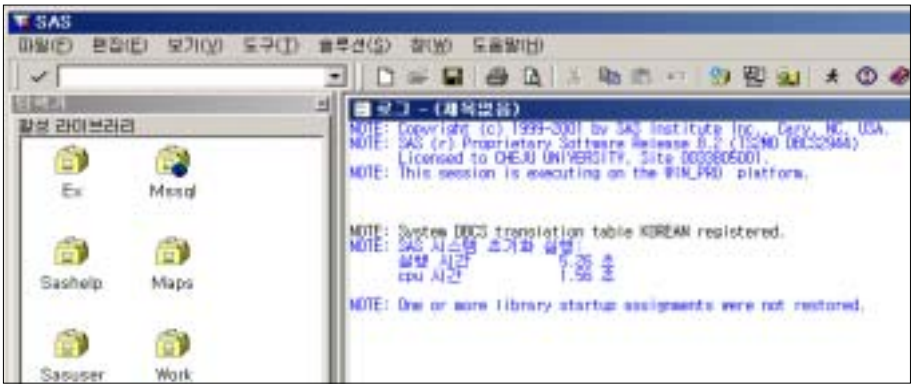


Figure 13. 데이터베이스 라이브러리 등록

<Figure 12>와 같이 우선 시작할 때 자동 할당되는 새로운 라이브러리를 생성하여, ODBC 엔진을 사용해 MS-SQL 데이터베이스에 원본 데이터를 데이터 소스에 작성한다. 그리고 DSN을 브라우저로 이용하여 입력하고, MS-SQL 서버와 연결하기 위하여 MS-SQL 사용자 DSN에 등록된 사용자 ID와 Password를 입력하면, <Figure 13>과 같은 라이브러리가 등록된다.

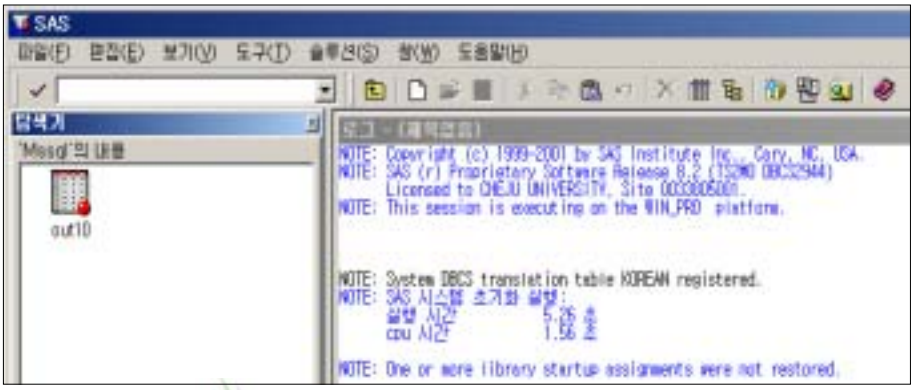


Figure 14. 등록된 라이브러리에 생성된 데이터

<Figure 14>은 Mssql이라는 라이브러리에 데이터베이스 사용자인 mssql_tester이 소유한 데이터를 보여주며, <Figure 15>은 Mssql 라이브러리에 저장된 데이터 out10을 실행시킨 후 생성된 데이터 테이블을 보여준다.

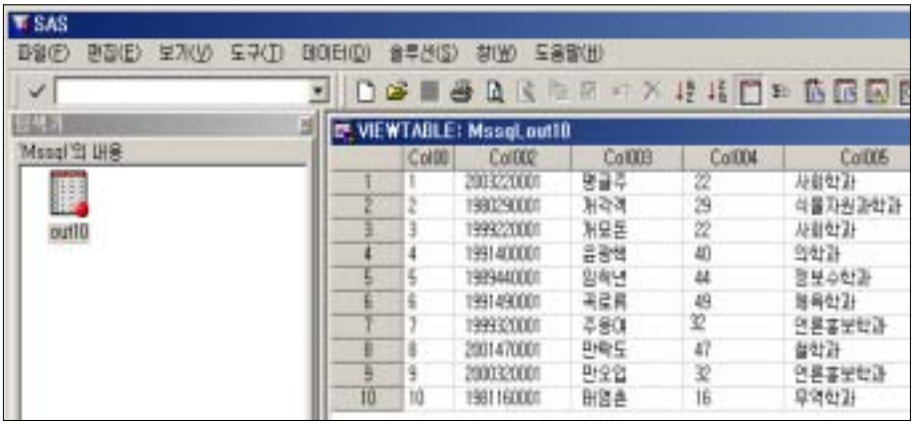


Figure 15. 생성된 데이터 테이블

2. SPSS에서 데이터베이스 연동

SPSS에서 원격 데이터베이스를 로컬 컴퓨터로 액세스하기 이해서는 ODBC를 이용하여 데이터를 Import 해야 한다. 우선 SPSS를 실행하면 SPSS for Windows 창이 생성된다. <Figure 16>과 같이 작업선택란에서 데이터베이스 마법사를 사용하여 새 쿼리 작성을 클릭한다.

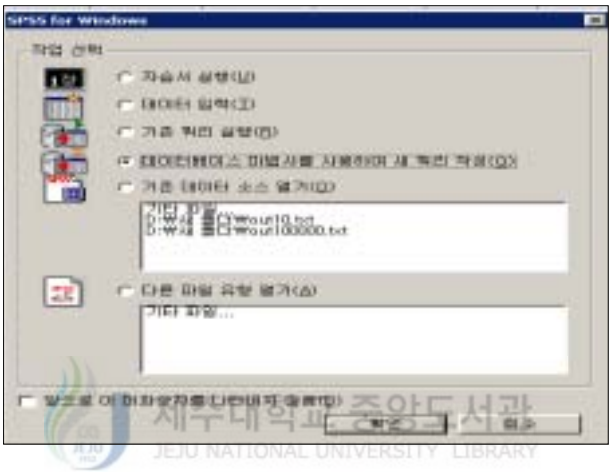


Figure 16. ODBC를 이용하여 외부데이터 가져오기(SPSS)

Figure 17. DSN에 등록된 데이터 소스 선택(SPSS)

먼저 <Figure 17>과 같이 ODBC 데이터 원본 관리자에서 등록된 DSN을 선택한다. DSN 등록은 사용자 DSN, 시스템 DSN, 파일 DSN으로 분류된다. 사용자 DSN은 다수의 사용자가 컴퓨터를 사용하는 경우에 각 사용자에 맞게 DSN을 등록하는 것으로, 사용자 DSN으로 DSN을 등록하면 다음과 같이 원격의 데이터베이스를 로컬 컴퓨터에서 액세스할 때 <Figure 18>과 같이 LoginID와 Password를 물어 허용된 사용자만이 데이터베이스에 접근하게 되고, 다른 사용자들을 제한한다. 그러면 데이터베이스 사용자 mssql_tester가 소유한 데이터를 <Figure 19>와 같이 보여주므로, '사용가능 표' 목록에서 작업할 데이터를 필드를 선택하여 필드 갱신 목록 위로 끌어가면 된다.

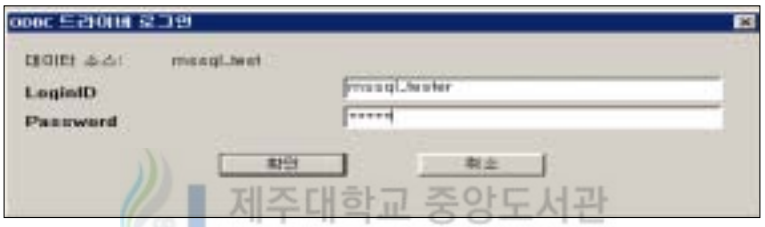


Figure 18. ODBC 드라이버 로그인

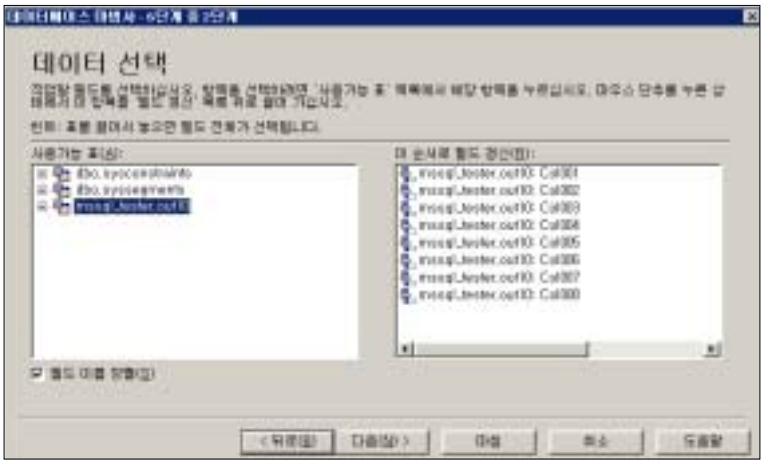


Figure 19. MS-SQL 데이터베이스에 저장된 데이터 선택

Figure 20. 변수정의

<Figure 20>은 선택한 데이터 소스로부터 갱신될 필드의 변수이름을 입력하거나 편집하는 과정으로 갱신될 필드의 변수이름을 입력하지 않으면 자동 변수이름(예:var001)으로 자동입력 된다. <Figure 21>은 사용자가 선택한 내용을 SQL 쿼리로 생성하여 보여준다.

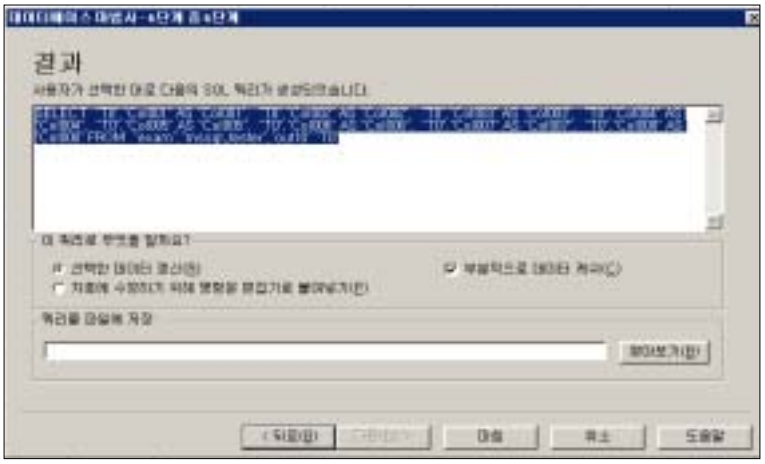
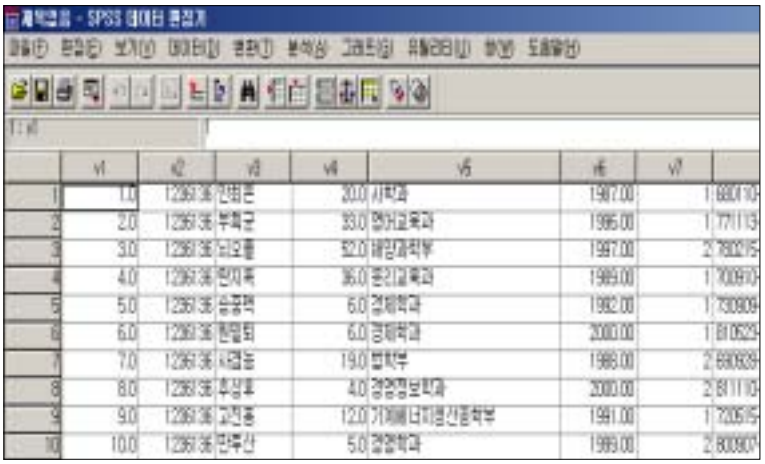


Figure 21. SQL 쿼리 생성 결과



	v1	v2	v3	v4	v5	v6	v7	
1	1.0	1226136	안태준	30.0	사학과	1987.00	1	680170-
2	2.0	1226136	부혁군	33.0	컴퓨교육과	1986.00	1	771113-
3	3.0	1226136	다오름	52.0	해양과학부	1987.00	2	7802015-
4	4.0	1226136	한지욱	36.0	중국어교육과	1989.00	1	700910-
5	5.0	1226136	송윤력	6.0	경제학과	1982.00	1	730909-
6	6.0	1226136	현일회	6.0	경제학과	2000.00	1	810523-
7	7.0	1226136	나랜능	19.0	법학부	1988.00	2	680609-
8	8.0	1226136	후성후	4.0	경영정보학과	2000.00	2	811110-
9	9.0	1226136	고건홍	12.0	기계에너지생산공학부	1991.00	1	720515-
10	10.0	1226136	안두산	5.0	경영학과	1989.00	2	800907-

Figure 22. 생성된 SPSS 데이터

<Figure 22>는 데이터베이스 마법사를 사용하여 새 쿼리 작성의 모든 과정을 마친 후 생성된 데이터 결과이다.

3. S-PLUS에서 데이터베이스 연동

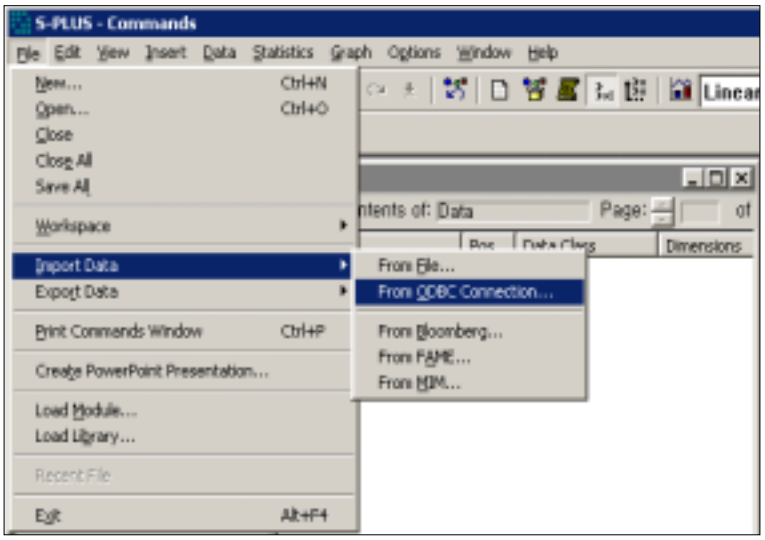


Figure 23. ODBC를 이용하여 외부데이터 가져오기(S-PLUS)

S-PLUS에서 ODBC를 이용하여 외부데이터를 가져오기 위해서는 <Figure 23>과 같이 파일메뉴에서 Import Data를 선택하고 그 서브메뉴인 From ODBC Connection을 선택한다. 그러면 <Figure 24>와 같이 Import ODBC창이 생성된다.

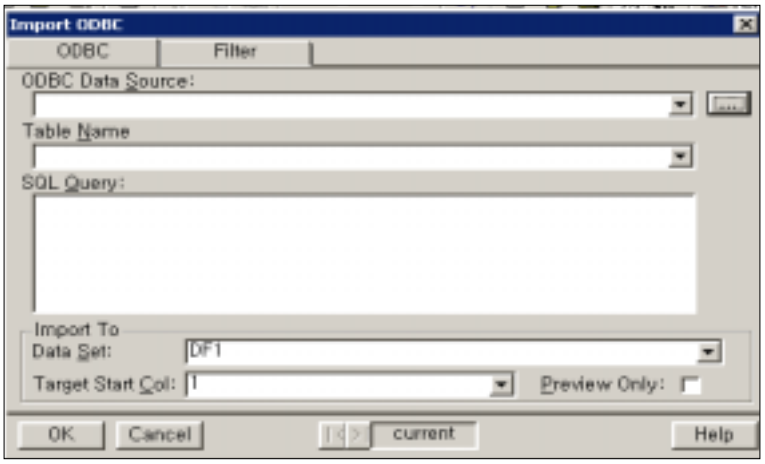


Figure 24. Import ODBC

ODBC Data Source에서 <Figure 25>와 같이 데이터 원본 선택 창에서 컴퓨터 데이터 원본 탭에서 사용할 데이터베이스를 선택한다.

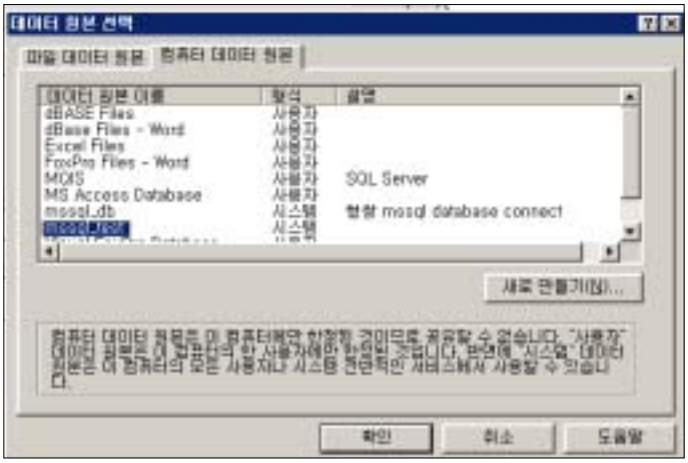


Figure 25. 데이터 원본 선택(S-PLUS)

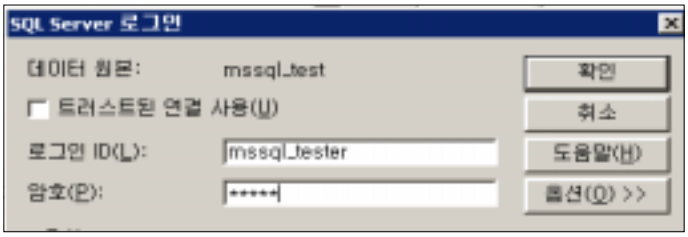


Figure 26. SQL Server 로그인

먼저 ODBC 데이터 원본 관리자에서 등록된 DSN을 선택한다. DSN 등록은 사용자 DSN, 시스템 DSN, 파일 DSN으로 분류된다. 사용자 DSN은 다수의 사용자가 컴퓨터를 사용하는 경우에 각 사용자에 맞게 DSN을 등록하는 것으로, 사용자 DSN으로 DSN을 등록하면 다음과 같이 원격의 데이터베이스를 로컬 컴퓨터에서 액세스할 때 <Figure 26>과 같이 LoginID와 Password를 물어 허용된 사용자만이 데이터베이스에

접근하게 되고, 다른 사용자들을 제한한다.

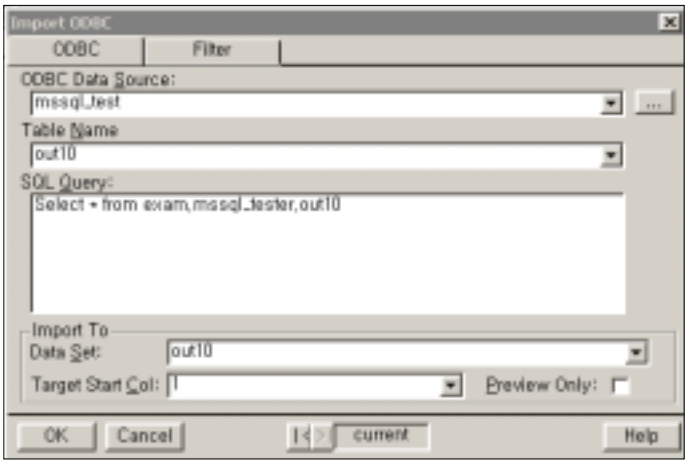


Figure 27. DSN에 등록된 데이터 소스 선택(S-PLUS)

Figure 28. 생성된 S-PLUS 데이터

<Figure 27>과 같이 ODBC Data Source에 DSN에 등록된 데이터 소스를 선택한 결과 생성된 데이터를 <Figure 28>에서 보여준다.

4. EXCEL에서 데이터베이스 연동

EXCEL에서 ODBC를 이용하여 외부데이터를 가져오기 위해서는 <Figure 29>와 같이 데이터 메뉴에서 외부데이터 가져오기를 선택하고 그 서브메뉴인 새 쿼리 만들기를 클릭한다.

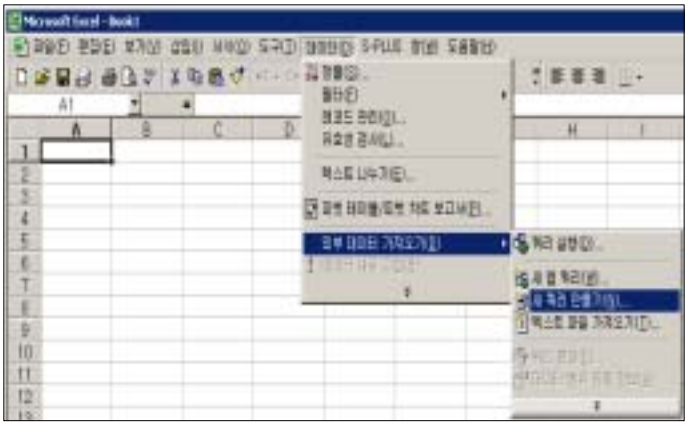


Figure 29. ODBC를 이용한 외부데이터 가져오기(EXCEL)

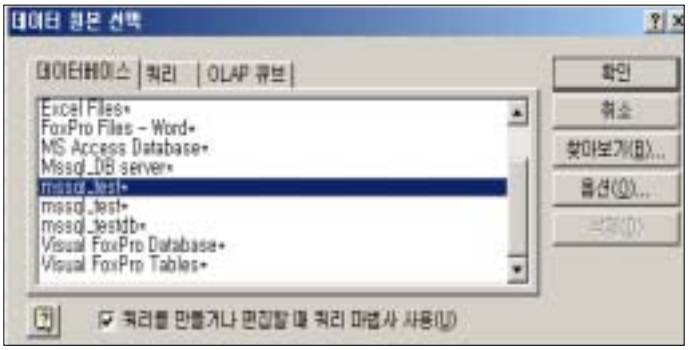


Figure 30. 데이터 원본 선택(EXCEL)

외부데이터를 가져오기 위해서는 <Figure 30>과 같이 ODBC 데이터 원본 관리자에 등록된 DSN인 데이터 원본을 선택한 후 <Figure 26>과 같이 LoginID와 Password를 물어 허용된 사용자만이 데이터베이스에 접근하게 되고, 다른 사용자들을 제한한다.

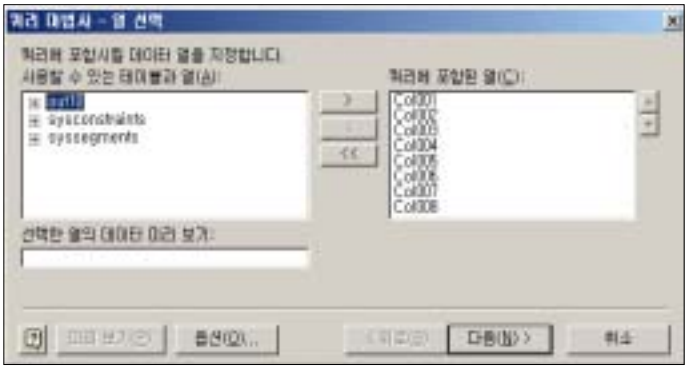


Figure 31. 쿼리 마법사 - 열 선택

<Figure 31>는 데이터베이스에 액세스한 다음 원본 데이터를 선택하여 쿼리에 포함시킬 데이터 열을 지정한다. 쿼리 마법사를 종료하면서 Microsoft Excel로 데이터 되돌리기를 선택하면 <Figure 32>과 같이 생성된 데이터를 보여준다.

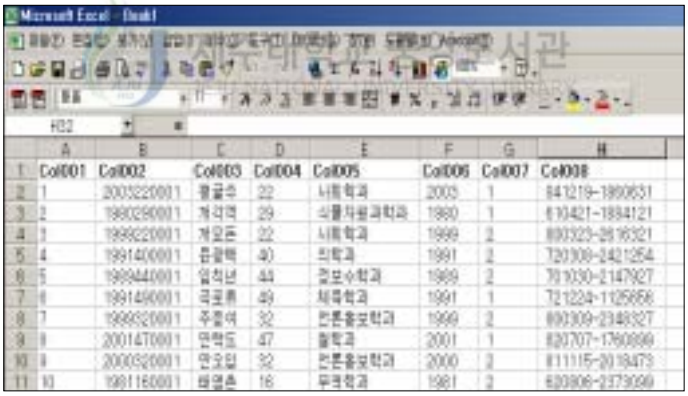


Figure 32. 생성된 EXCEL 데이터

5. JMP에서 데이터베이스 연동

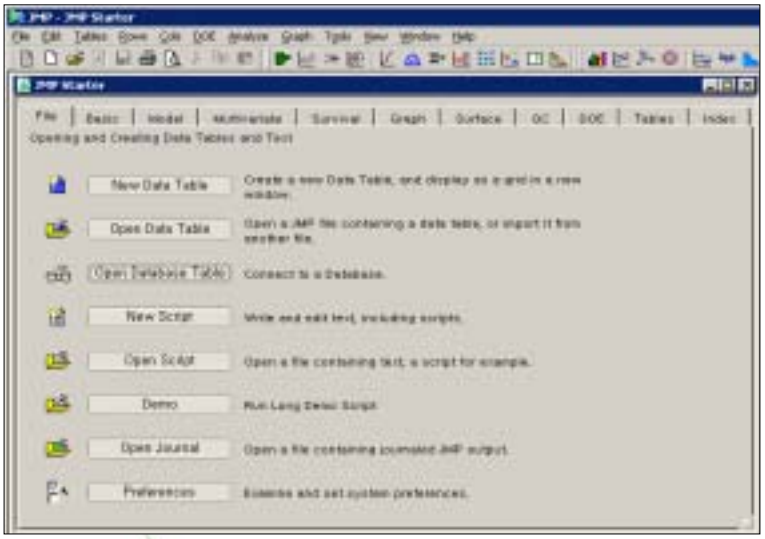


Figure 33. JMP Starter

JMP에서 ODBC를 이용하여 MS-SQL 서버에 액세스하기 위해서는 <Figure 33>과 같이 File 탭에서 Open Database Table을 선택하면 <Figure 34>과 같다.

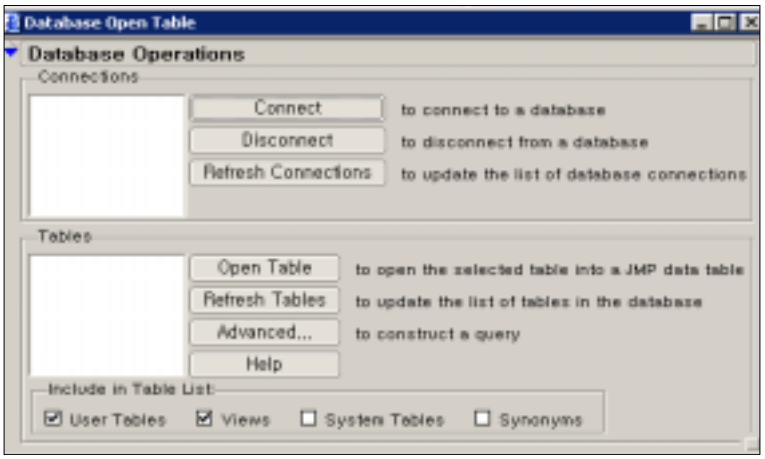


Figure 34. Database Open Table

Figure 35. DSN에 등록된 데이터 소스 선택(JMP)

<Figure 35>와 같이 Select Data Source 대화상자에 Machine Data source 탭에서 ODBC 데이터 원본 관리자에 등록된 DSN인 데이터 원본을 선택한다. <Figure 26>과 같이 LoginID와 Password를 물어 허용된 사용자만이 데이터베이스에 접근하게 되고, 다른 사용자들을 제한을 위하여 ODBC 등록시 사용자를 구분하여 데이터베이스에 접근하도록 사용자 DSN을 등록하였다.

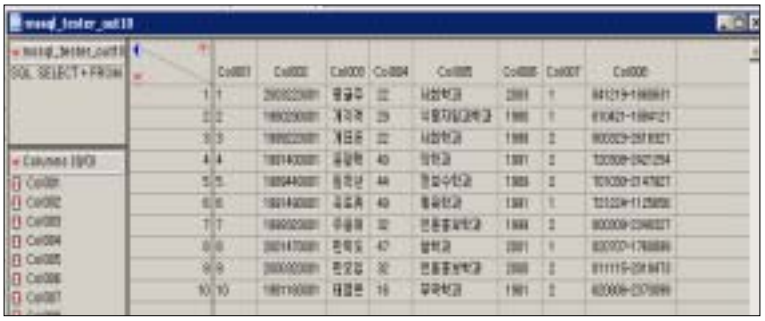


Figure 36. 생성된 JMP 데이터 테이블

ODBC를 사용하여 원격 데이터베이스에 접속 후 사용자 mssql_tester가 소유한 데이터베이스 테이블 중에서 원하는 테이블을 선택하면 <Figure 36>같이 테이블이 생성된다.

V. 성능평가

이 절에서는 성능평가를 위한 데이터 구축 방법과 생성된 데이터를 데이터베이스에 저장하여 통계패키지별 분석 및 처리 가능한 데이터의 레코드수와 컬럼수를 비교하여 그 결과를 살펴보자.

1. 데이터 생성을 위한 test 테이블 구축

생성 필드는 ID, 학번, 성명, 학과코드, 학과명, 입학년도, 성별, 주민번호 8가지이다. 이 test 테이블의 레코드를 100개에서 100만개 까지 다양하게 생성하였다. ID는 일련번호로 생성된 레코드 개수만큼 순서대로 자동 증가 시켰고, 학번은 입학년도 4자리와 학과코드 2자리를 결합하고 마지막 4자리 학과내의 번호는 입학년도와 학과코드별로 생성시킨 데이터의 순으로 일련번호를 부여하여 동일 입학년도의 동일 학과에서는 9999명까지는 학번의 중복이 없으며, 성명은 대한민국 성씨 154개가 저장된 파일과 이름으로 쓸 수 있는 글자 431개가 저장된 파일을 읽어들여 배열에 저장하여 성과 이름 두자를 랜덤하게 추출하여 결합하였고, 학과코드는 임의의 학과 55개를 가나다순으로 01~55의 2자리 코드를 주어 배열에 저장한 후 랜덤하게 추출하여 학과 코드와 해당 학과명을 출력하였다. 그리고 입학년도는 1980년에서 2003년 사이에서 가능한 24개의 입학년도를 랜덤하게 추출해냈으며, 성별은 남자는 1, 여자는 2로 구분하여 둘중 하나를 선택하였으며, 주민번호는 입학년도에서 19를 빼어 생년을 계산하고, 1과 12사이의 난수를 발생시켜 월을 구하고, 생성시킨 년이 월에 따라 최대일수를 계산하여 난수를 발생시켜 주민번호 뒷부분 중 첫 번째 코드는 발생시킨 성별 난수를 가져오고, 세자리 코드와 두자리 코드를 난수로 발생시고 난 뒤 마지막 체크 코드를 알고지름에 따라 규칙에 맞는 값을 생성해낸다. 생성된 파일을 MS-SQL 데이터베이스로 변환시키기 위해서는 먼저 Microsoft SQL Sever의 엔터프라이즈 관리자를 실행시킨 후 데이터 가져오기를 수행한 후, DTS(Data Transformation Services) 가져오기/내보내기에서 데이터 원본선택

란에서 변환을 원하는 소스타입(Access/Excel/Text)과 변환할 파일을 선택한 후 Microsoft SQL Sever의 데이터베이스를 대상으로 선택한다. 실행시기를 결정한 후 변환을 실행하면 진행률과 현재 상태가 표시되며 변환 완료 후에는 해당 데이터베이스 서버에 접속하여 정상적으로 테이블이 생성되었는지 확인해 볼 수 있다.

2. 통계패키지별 사용가능한 최대 데이터 사이즈 비교

CPU	Pentium Ⅳ 1.8GHz
RAM	512MB
HDD	7G 여유공간
OS	Windows 2000 Professional
DATABASE	MS-SQL Server, MySQL Server
PROGRAM	데이터생성(Visual C++6.0)

Table2. 실험환경

<Table 2>와 같은 실험환경에서 데이터를 1000개 단위를 나뉘어서 각각 저장하여 통계패키지별 off-line과 on-line상에서 각각의 데이터를 불러들였다. 이때 데이터를 불러들이지 못하거나 오류 메시지가 생성될 경우에는 데이터를 100개 혹은 10개 단위로 세분하여 다시 불러들여 가면서 통계패키지별 처리 가능한 데이터 크기를 비교하였다. 데이터의 레코드수는 SAS에서 원자료를 가지고 <Figure 37>와 같이 생성하였다.

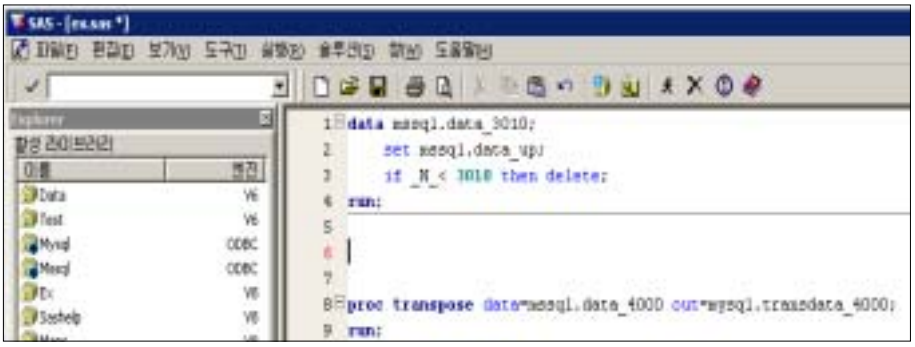


Figure 37. 데이터 크기별 생성

컬럼수						로컬 (local)	레코드수				
150	256	32767	35000	4만	6만		65,536	10만	30만	50만	100만
○	○	○	×	×	×	SAS	○	○	○	○	○
○	○	○	○	○	○	SPSS	○	○	○	○	○
○	○	○	○	×	×	S-PLUS	○	○	○	○	○
○	○	×	×	×	×	EXCEL	○	×	×	×	×
○	×	×	×	×	×	MATHEMATICA	○	○	○	○	○
○	○	○	×	×	×	JMP	○	○	○	○	○

Table 3. 로컬 컴퓨터에 저장된 데이터 불러오기

하지만 위 <Table 3>에서와 같이 엑셀은 레코드수가 현저하게 다른 통계패키지들과 차이를 보이는 것을 발견할 수 있었다.

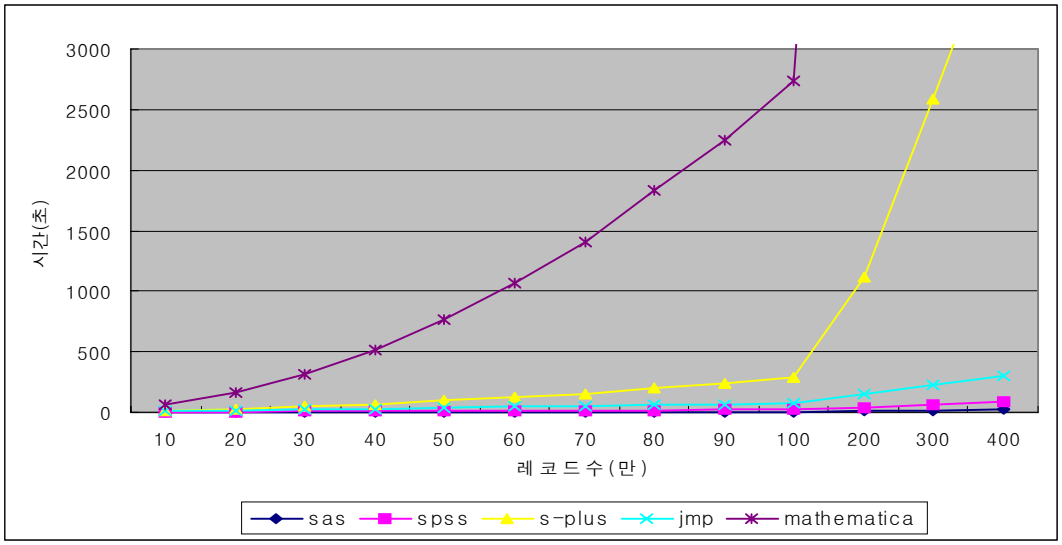


Figure 38. cpu시간 비교

<Figure 38>는 통계패키지별 로컬(local) 컴퓨터에 저장된 데이터를 불러들일 때 소요되는 cpu시간을 측정하여 비교하였다. 측정결과를 볼 때 100만개의 레코드를 가진 데이터를 불러들일 때 소요되는 cpu시간이 SAS가 4.75초로 가장 적었다. 그 다음으로 SPSS는 21초, JMP는 76초, S-PLUS는 294초, MATHEMATICA는 2739초가 소요됨을 볼 수 있었다. S-PLUS인 경우 100만개 레코드 이상인 경우 급격하게 cpu시간이 증가하였고, MATHEMATICA의 경우는 다른 패키지들과 현저히 차이가 났다. 100만개의 레코드를 가진 데이터를 불러들일 때 SAS는 SPSS보다 대략 4배, JMP보다 대략 16배, S-PLUS보다는 대략 60배가량 cpu시간의 차이를 보였다. 즉, 대용량 데이터를 로컬(local) 컴퓨터에서 불러들여 분석 및 처리할 때 SAS가 가장 적은 시간이 소유됨을 알 수 있다.

레코드수(만)	SAS	SPSS	S-PLUS	JMP	MATHEMATICA
10	0.47	3	12	8	57
20	0.98	5	26	16	163
30	1.41	7	45	23	319
40	1.87	9	66	30	517
50	2.43	11	95	38	767
60	2.82	13	129	45	1067
70	3.36	14	155	53	1405
80	3.81	17	197	60	1829
90	4.32	19	242	69	2243
100	4.75	21	294	76	2739

Table 4. cpu시간(초) 비교

어떤 자료를 가지고 분석을 할 때 레코드수가 65,536개 이상이 되었을 때 엑셀을 가지고서 통계처리를 한다면, 엑셀은 워크시트범위(65,536시트)를 초과하면 <Figure 39>과 같은 warning error가 발생한다. 즉 데이터의 레코드수가 65,536개를 초과하였을 경우는 EXCEL의 한계범위 내에서 데이터를 불러들이기 때문에 원 데이터에서 누락되는 데이터가 발생하므로 통계분석결과를 신뢰하기 힘들 것이다. 또한 로컬 컴퓨터에서 외부데이터를 불러올 때 걸리는 시간이 통계패키지별로 차이가 발생하는 것으로 보아서

통계패키지의 선택이 얼마나 중요한지를 다시 한번 보여준다.

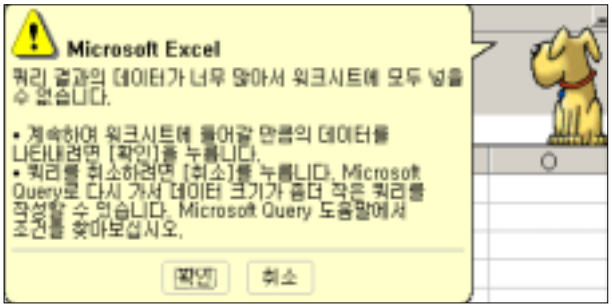


Figure 39. EXCEL 워크시트범위(65,536) 초과

이제 로컬 컴퓨터에서 원격의 데이터베이스에 액세스하여 데이터를 얼마나 처리가능한지를 보겠다. <Table 5>에서와 같이 로컬 컴퓨터에서 원격의 데이터베이스에 액세스한 데이터를 얼마나 처리하는가를 알아보기 위해서 레코드수와 컬럼수를 가지고 MS-SQL 데이터베이스와 MySQL 데이터베이스에 각각 같은 크기의 레코드수를 가진 데이터와 같은 크기의 컬럼수를 가진 데이터를 저장하였다.

여기서 MATHEMATICA는 ODBC를 사용하여 데이터를 import할 수 없으므로 비교에서 제외한다.

컬럼수(MS-SQL)			컬럼수(MySQL)				원격 (remote)	레코드수				
256	300	1024	256	300	2040	3080		65,536	10만	30만	50만	100만
○	○	○	○	○	○	○	SAS	○	○	○	○	○
○	○	○	○	○	○	○	SPSS	○	○	○	○	○
○	○	○	○	○	○	×	S-PLUS	○	○	○	○	○
○	×	×	○	×	×	×	EXCEL	○	×	×	×	×
○	○	○	○	○	○	○	JMP	○	○	○	○	○

Table 5. 원격 데이터베이스에 저장된 데이터 불러오기

처리 가능한 컬럼수를 알아보기 위해서 원자료를 레코드수를 1,300개부터 3,500개까지 50개씩 증가하여 원자료를 나누고, 이 나눠진 각각의 데이터를 전치(transpose)하였다. 이때 전치(transpose)시 오류가 발생하면 레코드수를 10개씩 증가 및 감소시켜 다시 데이터를 전치(transpose)하였다. 이 각각의 전치(transpose)된 데이터를 두개의 데이터베이스 MS-SQL Server와 MySQL Server에 각각 저장하였는데 데이터베이스의 종류에 따라 ODBC를 사용하여 자료를 up-load함에 차이가 발생하였다.

ODBC를 사용하여 데이터베이스에 저장된 데이터를 import할 때, MS-SQL Server는 컬럼수가 1024개 이상 up-load 할 수 없었지만, MySQL Server는 대략 3080개 정도의 컬럼을 가질 수 있다. 각각의 서버의 up-load 할 수 있는 데이터 사이즈가 다르다는 것을 알 수 있었다. 여기서 data를 MS-SQL과 MySQL에 up-load 시킬 때 발생하는 오류에 대해서 알아보면 다음과 같다.

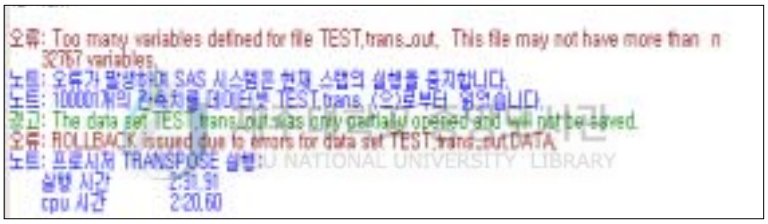


Figure 40. SAS의 변수 정의 최대 허용범위 초과

<Figure 40>은 변수 정의 최대 허용범위인 32767를 초과하였기 때문에 발생한 오류이다.

Figure 41. MS-SQL 데이터베이스 변수 정의 최대 허용범위 초과

<Figure 41>는 테이블의 'COL1023'열이 최대 열 개수인 1024를 초과하였기 때문에 테이블 생성에 실패하여 발생한 오류이다.

Figure 42. MS-SQL 데이터베이스 최대허용 테이블 행 크기 초과

<Figure 42>은 MS-SQL 행 크기의 최대값 8060 보다 큰 행을 만들 수 없기 때문에 발생한 오류이다.

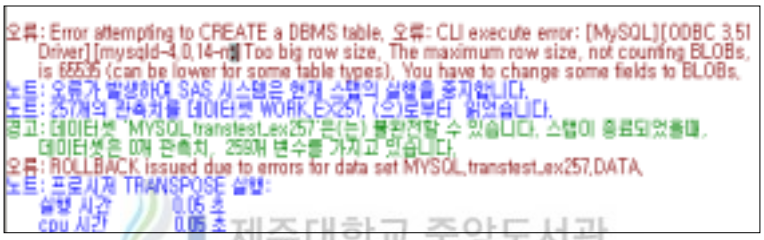


Figure 43. MySQL 데이터베이스 최대허용 테이블 행 크기 초과

따라서 MS-SQL 데이터베이스는 <Figure 42>과같이 데이터 최대허용 테이블 행 크기의 최대값 8060과 MySQL 데이터베이스는 <Figure 43>과 같이 데이터 최대허용 테이블 행 크기의 최대값 65535을 초과한 경우에 발생하는 오류이다. 두 데이터베이스가 가지는 변수 정의 최대 허용범위 1024(MS-SQL)와 32767(MySQL)보다 적은 변수만을 up-load한다. 로컬 컴퓨터에서 원격의 데이터베이스에 존재하는 데이터를 직접 액세스하여 데이터를 분석 및 처리할 경우 데이터의 성격을 잘 파악하여야 한다. 문자열로 된 데이터를 원격으로 액세스하는 경우에 변수에 정의된 문자열의 길이에 따라서 MS-SQL과 MySQL에 up-load할 수 있는 데이터의 변수의 개수가 달라지기 때문이다. 즉, MS-SQL과 MySQL은 고정적으로 변수 정의 최대 허용범위가 주어져 있으나, 각 데이터베이스에 주어진 최대허용 테이블 행 크기 때문에 두 데이터베이스의 변수 정의 최대 허용범위는 가변적이라고 할 수 있기 때문에 이 두 가지를 모두 고려하여야 한다.

VI. 결　　론

본 논문에서는 통계패키지별 ODBC를 이용하여 MS-SQL과 MySQL 데이터베이스 접속하는 방법과 분석 및 처리 가능한 최대 데이터 사이즈와 로컬(local)컴퓨터에 저장된 데이터를 불러올 때 cpu시간을 측정하여 비교하였다.

각각의 통계패키지들은 ODBC를 이용하여 원격 데이터베이스에 접속하는 방법이 달랐고, 그 중 Mathematica는 로컬 컴퓨터에서 원격 컴퓨터의 데이터베이스에 접속할 수 없었다. MS-SQL데이터베이스와 MySQL 데이터베이스는 처리 가능한 최대 허용범위가 서로 달라서 데이터베이스에 데이터를 up-load 할 때 데이터의 컬럼수에 차이를 보였다. 또한 통계패키지별 처리 가능한 데이터의 컬럼수와 레코드수에 차이를 보였는데, 특히 Excel인 경우 다른 통계패키지와 현저하나 레코드와 컬럼이 작은 데이터만을 처리할 수 있었다. 그리고 로컬(local) 컴퓨터에 저장된 데이터를 불러올 때 SAS를 활용하면 가장 적은 cpu시간에 불러올 수 있었다. S-PLUS는 100만개 레코드 이후부터 현저하게 다른 패키지들과 차이를 보임을 확인할 수 있었다. 그러므로 데이터베이스의 처리 가능한 최대 허용범위를 고려하여 데이터베이스 테이블을 생성함과 더불어 적절한 통계패키지의 선택을 데이터를 분석 및 처리하여야 한다.

이 후부터의 연구는 기존 실험에서 확장하여 ODBC를 이용하여 다양한 데이터베이스에 접속할 때 각각의 통계패키지의 액세스속도를 측정하고, ODBC가 아닌 다른 미들웨어인 JDBC를 활용하여 데이터베이스에 접속하는 방법과 그에 따른 최대 데이터 사이즈와 액세스 속도를 측정하여 서로 비교하여 볼 것이다.

VII. 참고 문헌

Anderson, E.E. Choice Models for The Evaluation and Selection of Software Packages. Journal of Management Information Systems, 1990, Vol. 6, No. 4

Bates, D. and Watts, D.(1998). Nonlinear Regression Analysis and its Applications. Wiley.

Burce, A. and Gao, H.(1996). Applied Wavelet Analysis with S-PLUS. Springer.

Hardle, W.(1990). Smoothing Techniques with implementation in S. Springer.

Huet, S., Bouvier, A., Gruet, M. and Jolivet, E.(1996). Statistical Tools for Nonlinear Regression. Springer.

Huber, P. J.. Languages for statistics and data analysis

Micah ALTMAN. A Review of JMP 4.03 With Special Attention to its Numerical Accuracy. The American Statistican, February 2002, Vol. 56, No. 1

ROBIN H. LOCK. A Comparison of Five Student Versions of Statistics Packages. The American Statistican, May 1993, Vol. 47, No. 2

Statsci. S User's Guide

Splus and Matlab: A comparison of two Programming Environments Find where it came from

강형창, 김철수(2002). 윈도우 SAS 시스템에서의 데이터베이스 연동, 제주대학교 기초과학연구소, 제15권 2호, 145-153

강형창, 김철수(2003). 윈도우 SAS 시스템에서 SAS/ACCESS 소프트웨어를 이용한 데이터베이스연결, 한국통계학회, 231-237

김병천(1987). 개인용 컴퓨터에서의 통계패키지의 선택과 활용, 응용통계연구, 제1권 1호, 75-90

김수화, 김승희, 조신섭(1994). 통계패키지에서의 시계열 분석방법의 비교연구. 한국통계학회 논문집, 제1권 제1호, 119-130

이진아, 허문열. R명령어들의 속도 평가. 2003 Proceedings of the Autumn Conference, Korean Statistical Society

이상석, 윤민석(1999). 통계처리용 소프트웨어 패키지의 품질 비교에 관한 연구. 품질경영학회지, 27권, 1호, 195-211

이정용(2003). 엑셀2002 통계함수의 번역 오류. Proceedings of the Spring Conference, Korean Statistical Society

조미순, 김순귀. 생존분석을 위한 통계패키지의 비교 연구-SAS, SPSS, STATA-. 2003 Proceedings of the Autumn Conference, Korean Statistical Society

조신섭, 신봉섭(1997). 통계적 공정관리를 위한 주요 통계패키지의 비교. 응용통계연구, 제10권 제1호, 29-36

조신섭, 송문섭, 이윤모, 성병찬, 윤영주, 이현부(1999). 기초통계교육을 위한 통계소프트웨어의 개발. 품질경영학회지, 27권, 2호, 277-14

최종후(1997). 그래픽 기법을 이용한 다변량자료의 구조탐색 -JMP IN을 중심으로-. Proceedings of the Spring Conference, Korean Statistical Society

차경준, 박영선(1999). 정적로딩 및 동적로딩을 통한 S-PLUS와 C언어간의 인터페이스 구현. 응용통계연구, 제12권 1호, 29-43

허명회, 정진환(1990). 탐색적 데이터분석(EDA) 기능에 관한 통계팩키지 프로그램의 비교검토. 응용통계연구, 제3권 제2호, 17-25

한국전산원(2002). XML이용한 데이터베이스 연동방안연구

이종원, 최현집. SAS를 이용한 통계분석. 박영사

최종후, 이성희. S-Plus를 이용한 통계그래픽스. 자유아카데미

원태연, 정성원. 한글 SPSS 통계자료분석. SPSS 아카데미

http://www.sas.com

http://www.spss.co.kr

http://user.spss.co.kr

http://www.jmp.com

http://www.s-plus.co.kr