

# 농업정책 실효성 증대를 위한 농업통계시스템 아키텍처 구축

이민수 · 최영찬 · 홍희연 · 최상호 · 김정섭

서울대학교 지역사회개발전공

## Construction of Integrated Agricultural Statistical System Architecture for Effective Policy

Lee, Min-Soo · Choe, YoungChan · Hong, Heeyeon · Choi, SangHo · Kim, Jeong-Seop

*Dept. of Agricultural Economics and Rural Development, Seoul National University*

**ABSTRACT** : This study designs an integrated data architecture to systematically manage the agricultural statistics database. Managing the agricultural statistics is important since it provides data for policies and decision making for agribusinesses. Ministry of Agriculture and the National Statistical Office collect the basic agricultural statistic data which provides the basis of logical decision making and agricultural policies. However, the agricultural statistic data has not well been used. The data has not been consistently collected nor managed. The raw data has not been organized nor processed to meet various demands. The needs has been arisen for a consistent agricultural statistics system to increase the relevance, accessibility, and efficiency of data for various users. There are massive amount of data accumulated over a long time period. Introducing the new system and reorganizing the data will bear large risks. A systematic method is required to reduce the risks in planing, building, and maintaining the database without hindering administration. This study provides a design of the agricultural statistics system architecture based on the user requirement analysis (URA) and similar systems abroad. We have also build a prototype to check the implementability of the system design.

**Key words** : Agricultural statistics, Data warehouse, ETL(Extraction, Transformation, Loading) tool

### I. 서론

농업정책 커뮤니케이션 과정에 있어 농업통계 자료는 매우 중요하다. 이 자료들로부터 정책 담당자가 의사결정을 내리도록 돕는 기본 지표를 확보할 수 있고, 정책효과를 파악할 수 있고, 정책의 장기적 목표를 설정하기 위한 기초 자료를 얻을 수 있다(USDA, 2005). 이와 함께 시장경쟁의 격화, 농식품 안전성 관련 사회문제의 대두 등, 최근에는 정책 환경이 급변하고 있어 관련 정보들을 통합적으로 관리하고 처리해야 할 필요성이 더욱 커지고 있다(Legg, 2005, OECD, 2003).

우리나라의 농업정책 커뮤니케이션 과정에서 활용되고 있는 정보나 지표는 주로 농림부와 통계청에서 작성하는 농업통계 자료로부터 산출된다. 해방 후 최초의 농업통계는 미군정에서부터 시작되었다. 미군정 시기부터 농업통계

는 농림부의 주도하에 이루어졌으나, 1998년 이후 통계청으로 5종의 통계 업무가 이관됨에 따라 통계청도 농업통계 수집에 중요한 역할을 하게 되었다. 이와 함께 현재 우리나라의 통계제도는 정보기관별로 각자 필요한 통계를 작성하는 분산형 통계제도를 채택하고 있기 때문에 농촌진흥청, 산림청, 농업기반공사, 농촌경제연구원 등 여러 정부 기관들이 서로 다른 성격의 공식적인 농업통계를 작성하고 있다. 분산형 통계제도는 각 기관이 필요한 통계를 효율적으로 수집하고 이용하는 데에는 장점을 가지지만, 통계자료를 통합적으로 이용할 필요가 있는 정책담당자, 연구자, 농업경영자 등이 농업통계 자료를 효율적으로 사용할 수 없도록 한다(Yost, 2000). 따라서 다양한 기관에서 생산되는 농업통계를 사용자들이 효과적으로 이용하기 위해서는, 농업통계들에 대한 관련성, 접근성, 효율성을 고려한 체계적이고 통합적인 농업통계시스템 구축이 필요하다.

농업통계 자료는 장기간에 걸친 조사를 통해 방대한 규모로 구축되는 경우가 일반적이기 때문에 새로운 시스템

Corresponding author : Cheo, YoungChan

Tel : 02-880-4743

E-mail : aggi@snu.ac.kr

템을 구축하여 재정비하는 데에는 많은 비용과 위험이 따를 수 있다. 그러므로 운영 업무에 지장을 주지 않으면서도 농업통계 통합 데이터베이스를 기획, 구축, 운영할 수 있는 방안이 필요하다. 이러한 현실적인 문제점을 해결하고 농업통계의 수집과 분석을 효과적으로 지원하고자 하는 의도에서 최근에 FAO의 FAOSTAT(<http://www.fao.org/faostat>), USDA의 FAS online 분석 시스템(<http://www.fas.usda.gov/ustrade>) 등 농관련 조직에서 많이 도입하고 있는 기술이 데이터 웨어하우스(Data Warehouse)이다. 데이터 웨어하우스는 조직의 의사결정 업무를 지원하기 위한 자료들의 집합 또는 의사결정 지원을 위해 기존의 데이터베이스로부터 독립된 별도의 데이터베이스이다. 데이터 웨어하우스는 기존의 상이한 플랫폼 및 포맷으로 존재하는 데이터를 재조명하여, 장단기 전략 계획 수립 및 의사결정에 필요한 분석적 정보처리의 적시적이고 정확한 정보 제공을 주요 목적으로 한다(Inmon, 2002, Griffin, 1995, McFadden, 1996).

본 연구에서는 다양한 기관에서 다양한 형식으로 장기간에 걸쳐 축적해 온 농업통계 자료를 통합하고, 이를 사용자가 효율적으로 활용할 수 있는 데이터 웨어하우스 기반의 농업통계시스템 아키텍처를 제시하고자 한다. 아키텍처를 구성하는 주요 시스템 구성요소의 특성을 규명하는 데 초점을 둔다. 또한 실제 사례에 대한 프로토타입 시스템 개발을 통해 본 구현 아키텍처에 대한 유용성도 파악한다.

## II. 농업통계 시스템 현황과 문제점

### 1. 농업통계 조사분석·배포 과정

농림부의 공식적인 농업통계 조사 업무는 국립 농산물품질관리원의 농업정보통계과가 담당하고 있다. 농산물품질관리원은 농업통계 자료를 조사하기 위한 지방 조직으로 경기, 강원, 충청, 호남, 영남, 제주의 6개 도 단위 지원과 80개의 시·군 단위 출장소를 두고 있다. 농업통계 자료를 수집하고 분석하고 배포하는 과정을 요약하면 다음 그림 1과 같다.

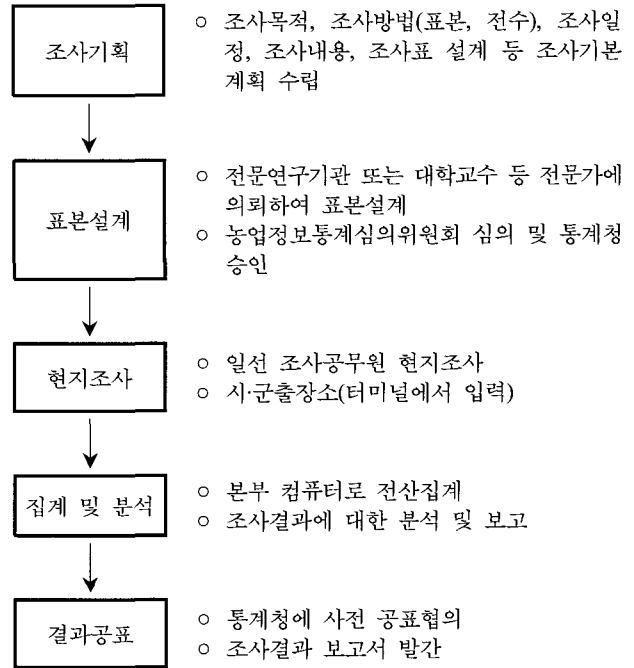


그림 1. 농업통계 조사 체계

조사기획 단계에서는 조사의 목적, 방법, 일정, 내용, 조사표 설계 등 기본적인 계획을 수립한다. 조사기본계획이 수립되면, 관련 연구기관이나 대학의 교수 또는 전문가가 표본을 설계하는 표본설계 단계가 진행된다. 이 단계에서 농업정보통계 심의위원회와 통계청의 승인을 받아야 한다. 대개의 경우 5년 간격으로 표본을 재설계하여 대표성을 유지하고 있다. 그 다음 단계는 현지조사 단계이다. 이 단계에서는 두 종류의 조사가 진행된다. 하나는 논이나 밭 등 농지 현장에 가서 자료를 수집하는 대지 조사이며, 다른 하나는 농업인 등 사람을 만나서 조사하는 면접 및 청취 조사이다. 농산물품질관리원 시·군 출장소 조사 공무원이 조사작업을 직접 수행하고, 기록한 조사표의 내용을 전산 입력하고 송신한다. 이렇게 조사가 끝나면, 농림부의 중앙 컴퓨터에서 자료를 집계하고 분석하고 가공하는 집계 및 분석 단계가 진행된다. 마지막으로 농림부는 통계청과 정부의 승인통계 기준에 맞추어 협의한 후 조사결과를 공표하고 보고서를 발간한다. 발

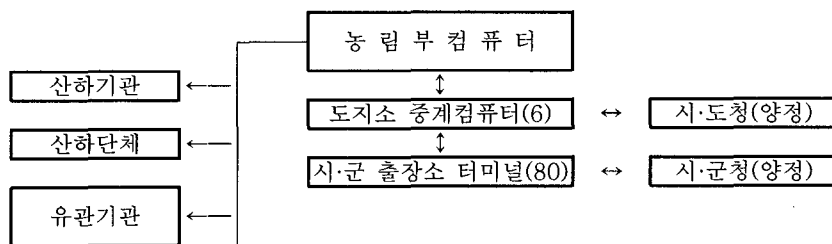


그림 2. 농업통계 전산 체계

표되는 농업통계 정보는 농산물품질관리원, 농림부 홈페이지, AFFIS 등을 통해 일반에게 전달된다.

## 2. 농업통계 담당 기관과 통계 작성 현황

우리나라는 기관별로 필요한 통계를 작성하는 분산형 통계제도를 채택하고 있다. 현재 103개 통계작성 기관들이 392종의 공식 통계를 생산하고 있다. 우리나라에서 통계청 승인 농업통계는 총 45종이 있는데, 이것들을 생산하는 기관은 농림부, 통계청, 농촌진흥청, 산림청, 농업기반공사, 농협중앙회이다<표 1>. 그 중 농업통계는 대부분 농림부가 작성하고 있다. 농림부에서는 경지면적, 작물생산량 등 20 종의 통계를 작성하고 있다. 통계청에서는 1998년 정

부조직 개편을 계기로 농림부로부터 이관 받은 5종의 통계(농업 총조사, 농업기본통계 조사, 농가경제 조사, 농산물 생산비 조사, 양곡소비량 및 재고량 조사)와 자체 5종의 통계를 작성하고 있다. 농촌진흥청은 2종, 산림청은 11종, 농업기반공사와 농협중앙회는 각각 1종의 농업통계를 작성하고 있다. 농림부 내에서 농업통계를 관장하는 조직으로는 국 단위로 농업정보통계관실이 있다. 그 소속 부서로 통계기획담당관실과 정보화담당관실을 설치하여 농업 통계의 기획 및 조정, 중장기발전계획 수립, 가공 및 이용, 국제 협력, 농업관측, 홍보 분산 및 정보화 등의 업무를 수행하고 있다.

## III. 농업통계 시스템 사용자 요구 분석

국내 농업통계 시스템의 문제점을 파악하기 위해 농림부, 농산물품질관리원, 통계청에서 관련 업무에 종사하는 공무원 11명을 대상으로 면접 조사를 실시했으며, 다양한 계층의 시스템 사용자 539명을 대상으로 설문조사를 실시했다. 설문조사를 진행하면서 농정조직, 농업 관련 연구조직, 농산업 현장 조직으로 구분하여 표집했다.

표 1. 주요 기관별 통계청 승인 농업통계

기관	승인 통계	
농림부	가축통계조사	경지면적조사
	과수실태조사	과실류가공현황조사보고
	농기구(보유)상황보고	농림어업인등에 대한 복지실태등 조사
	농림업생산지수	농업기계화율조사
	농지임대차조사	누에사육 및 양잠규모조사보고
	도축검사보고	배합사료생산실적 및 원료사용실적보고
	버섯생산통계	여성농업인실태조사
	우유및유제품 생산소비상황	작물통계조사
	채소류가공현황조사보고	축성재배시설현황 및 생산실적
	축산물생산비조사	화훼류재배현황보고
통계청	농가경제조사	임업총조사
	농어업법인사업체 통계조사	농산물생산비조사
	농업총조사	농업기본통계조사
	양곡소비량조사	어가경제조사
	어업기본통계조사	어업총조사
농촌진흥청	농산물소득조사	농촌생활지표조사
	관상수사업상황조사	목재수급통계
산림청	민유림조립실적보고	산림기본통계
	산림피해단속상황보고	산림형질변경허가 및 복구상황보고
	양묘사업실태보고	임산물생산통계
	임업경영실태조사	전국산림실태조사
	조림활착상황보고	
농업기반공사	농업생산기반정비 사업통계조사	
농협중앙회	농가판매및구입가격조사	

### 1. 면접 조사 결과

면접 조사에서는 농업통계 작성에 참여하고 있는 관계자들이 인식하고 있는 현재 우리나라 농업통계 시스템의 문제점들을 파악하는 데에 초점을 맞추었다. 면접 대상자들이 지적한 국내 농업통계 시스템의 주요 문제점은 다음과 같이 4가지로 요약할 수 있다. 여기에서는 해외 농업통계 사례와 비교하면서 그 내용을 설명한다.

#### 가. 농업통계 생산 및 관리를 총괄할 수 있는 전체적인 틀 부재

피면접자가 가장 많이 지적한 문제점은 농업통계 생산·관리의 틀이 부재하다는 점이었다. 현재 국내에는 농업통계 시스템의 목표, 비전(vision), 우선순위, 전략적 목표, 장단기 계획, 성과체계 등에 대한 전체적인 틀을 수립하고, 관리하는 조직이 없다. 농림부가 그러한 역할을 수행해야 함에도 불구하고 인력과 예산 부족으로 인해 공식 통계자료 발간 업무에만 치중하고 있으며, 1998년 일부 통계자료에 관한 업무가 통계청으로 이관됨에 따라 농업통계를 전체적으로 관리하기가 더욱 어려워졌다는 점이 많이 지적되었다.

외국사례를 보면, 미국 농무성의 국가농업통계국(National Agricultural Statistics Service, NASS)은 농업통계 생산의 목적, 비전, 핵심 성공요인, 우선순위, 전략적 목표

및 계획, 성과측정 등과 관련된 업무를 체계적으로 수행하고 있다. NASS는 미션과 목표에 근거하여 농업통계를 체계화하고, 성과측정 지표를 통해 장단기 전략계획과 성과를 평가하고 있다.

**나. 종합적인 농업통계 시스템 부재**

두 번째로 많이 지적된 사항은 농업통계의 수집·분석 시스템의 분산이다. 공식적인 농업통계의 경우도 농림부, 통계청, 농촌진흥청, 산림청, 농업기반공사, 농협중앙회 등이 각각 달리 조사하며, 조사한 자료 또한 기관들이 개별적으로 관리하고 배포한다. 공식 농업통계 외에도 농림부의 예산 지원을 받는 공공적 성격을 지니는 기관들인 농수산물유통공사, 농림수산물정보센터, 농촌경제연구원 등이 독자적으로 농업통계 자료를 생산하고 배포한다. 그리고 동일한 유형의 통계자료라 하더라도 이원화된 형태로 관리되는 경우가 있다. 예를 들면, 농업총조사 데이터의 경우는 현재 농림부와 통계청이 동시에 관리하고 있다. 업무가 통계청으로 이관된 1998년 이후의 원본 데이터는 통계청이 KOSIS 시스템을 통해 관리하고 있으나, 그 이전 자료는 농림부가 관리하고 있다. 또 보고서의 배포에서도 농림부는 한글 파일로, 통계청은 KOSIS를 통해 배포하고 있다.

미국 농무성의 경우, 코넬(Cornell) 대학과 협력하여 농업통계를 통합적으로 관리하고 배포하는 USDA Economics and Statistics System을 갖추고 있다. OECD는 통계 자료를 통합적으로 관리하는 StatsPortal을 운영하고 있으며, FAO도 Economics and Social Department Statistics Division을 두어 일원화된 형태로 농업통계 자료를 관리하고 있다.

**다. 통계자료의 사용성(usability) 미흡**

일부 피면접자는 통계자료가 주로 공식 보고서 작성에만 사용되고, 추가적인 분석에 거의 사용되지 않는다는 점을 문제점으로 지적하였다. 농림부가 웹을 통해 제공하고 있는 농업통계 자료는 다음 표 2와 같다. 농림부가 웹을 통해 통계자료를 배포하는 형태는 크게 두 가지로 나뉜다. 하나는 책자로 된 보고서를 한글파일이나 html 파일 형식으로 제공하는 것이며, 다른 하나는 쿼리 검색을 이용하도록 하는 것이다. 표 2에서 보듯이 농림부가 제공하는 농업통계 중 쿼리 검색은 매우 제한적으로 채택되고 있다. 사용자는 농림부가 사전에 설정한 보고서나 포맷으로만 정보를 획득할 수 있다. 대부분의 경우 의사결정에 필요한 정보를 웹 인터페이스를 통해 사전에 구조화하여 획득하는 것은 거의 불가능하다. 사용자들은 더욱 상세한 분석이나 통계자료 안의 콘텐츠들 간의 관계 분석, 회귀 분석이나 계량모형을 통한 고급분석을 원할 수도 있다. 그러나 현재의 농림부 통계 시스템은 보고서나 단순한 검

색 조건에 의한 보고서 출력 등의 기본적인 기능만 갖추고 있다. 이는 통계자료의 사용성 측면에서 매우 취약한 모습을 드러내고 있는 것이다. 이와는 대조적으로 미국 농무성의 Economics and Statistics System, OECD의 StatsPortal, FAO의 Economic and Social Department Statistic Division은 데이터 웨어하우스와 OLAP 기술을 활용하여 사용자들이 웹 인터페이스에서 통계 자료들을 실시간으로 분석하고 저장하며 원 자료를 손쉽게 획득할 수 있게 하고 있다.

**라. 체계적인 문서작업 및 배포일정 관리 미흡**

국내 농업통계 시스템의 자료는 많은 예산과 인력을 투입하여 수집한 것임에도 불구하고 체계적인 문서작업이 이루어지지 않고, 배포일정 관리가 정확히 이루어지지 않는다는 점도 주요 문제점으로 지적되었다. 현재 농업통계는 과거 조사 내용에 대한 이력(history)이 체계적으로 문서화되어 있지 않아 활용가능성이 적다. 과거에 조사한 통계자료들에 대한 의미를 알려주는 메타데이터(metadata)에 대한 문서작업이 체계적으로 이루어지지 않아, 과거의 농업통계자료를 분석하기가 쉽지 않다. 또 어떤 통계자료가 어느 시기에 공표될 것인지, 과거에 공표한 자료로는 어떤 것들이 있는지 등에 관한 정보를 제공하는 배포일정 관리 기능은 거의 없다. 그리고 통계자료를 발표한 후 이를 모니터링하고 평가하는 기능도 없다. 해외 사례를 살펴보면, OECD의 경우는 MetaStore를 통해 통계자료와 산출물에 대한 메타데이터를 효율적으로 관리하고 있으며, USDA의 경우는 보고서 발간일정을 캘린더 기능을 통해 관리하고 있다.

**표 2. 농림부가 웹을 통해 제공하는 농업통계 자료의 종류와 형태**

범주	통계 자료	배포 형태
농업 주요통계	* 농업 주요 통계	쿼리 방식, html 파일
	* 수출입 통계	쿼리 방식, html 파일
	* 농업총조사	한글 파일
	* 농림통계연보	한글 파일, PDF 파일
해외 농업통계	* 농림업 주요 통계	한글 파일, PDF 파일
	* 세계 속의 한국	PDF 파일
	* 중국 농업 통계	html 파일
	* 일본 농업 통계	html 파일
축산정보	* USDA 농업통계	PDF 파일
	* HACCP	한글 파일
유통정보	* 가축전염병 발생 정보	한글 파일, html 파일
	* 농산물 가격 정보	쿼리 방식
	* 유통 리포터	html 파일

표 3. 농업통계 전달 매체별 활용도

구분	보고서	CD-ROM	웹 보고서	웹 분석기능	원자료
연구조직 (N=146)	84.2%	27.4%	77.4%	34.2%	20.5%
농정조직 (N=186)	78.5%	21.0%	69.9%	30.1%	25.8%
현장조직 (N=205)	67.3%	20.0%	62.0%	35.1%	24.4%
전체 (N=537)	75.8%	22.3%	68.9%	33.1%	23.8%

2. 설문 조사 결과

가. 농업통계 전달 매체별 활용도

국내 농업통계 자료가 사용자에게 전달되는 주요 매체로는 ‘보고서’, ‘CD-ROM’, ‘웹 보고서’, ‘웹 분석기능’, ‘원자료’ 등의 다섯 가지를 들 수 있다. 설문 응답자들이 소속된 조직 이용별로 자주 활용하는 매체 형태를 살펴보면 다음 표 3과 같다. 전체적으로 보고서와 웹 보고서의 활용 빈도가 높게 나타나고 있으며, 웹 분석기능은 상대적으로 낮게 나타나고 있다. 농업통계 자료가 웹을 통해 전달될 수 있는 인프라가 이미 구축되어 있음에도 불구하고 웹 분석기능의 활용도는 30%대로 매우 낮은 것으로 나타났다.

나. 농업통계 배포 일정 정보 제공에 대한 응답자 평가

농림부 농업통계 시스템의 배포일정 정보 제공에 대한 평가에서는 개선이 ‘필요’하거나 ‘매우 필요’하다고 응답한 빈도가 전체의 86.0%였다. 특히, 현장 조직에 소속된 사용자들이 ‘매우 필요’하다고 응답한 비율이 20.6%로 상대적으로 높았다. 전체적으로 농업통계 배포일정 정보 제공에 대해 기능개선이 필요하다고 응답한 경우는 80%이 상으로 나타나 향후 시스템개발에 있어서 배포일정관리가 중요한 기능임을 보여준다.

다. 통계자료 검색 기능에 대한 응답자 평가

사용자들이 원하는 통계자료를 검색하는 기능에 대한 요구분석에서는 연구조직에 소속된 응답자들이 상대적으로 높은 만족도를 보이는 반면에, 현장조직에 소속된 응

표 4. 농업통계 배포일정 정보 제공에 대한 사용자 요구

구분	농림부 농업통계 시스템			통계청 농업통계 시스템				
	N	개선이 필요없음	개선이 필요	개선이 매우필요	N	개선이 필요없음	개선이 필요	개선이 매우필요
연구조직	146	12.3%	80.1%	7.5%	133	25.6%	66.9%	7.5%
농정조직	181	16.6%	70.7%	12.7%	166	18.7%	66.9%	14.5%
현장조직	194	12.9%	66.5%	20.6%	180	16.1%	60.6%	23.3%
전체	521	14.0%	71.8%	14.2%	479	19.6%	64.5%	15.9%

표 5. 통계자료 검색기능에 대한 사용자 요구

구분	농림부 농업통계 시스템			통계청 농업통계 시스템				
	N	개선이 필요없음	개선이 필요	개선이 매우필요	N	개선이 필요없음	개선이 필요	개선이 매우필요
연구조직	146	22.6%	59.6%	17.8%	133	28.6%	60.2%	11.3%
농정조직	177	19.2%	65.5%	15.3%	165	22.4%	63.6%	13.9%
현장조직	188	15.4%	62.8%	21.8%	180	17.8%	61.7%	20.6%
전체	511	18.8%	62.8%	18.4%	478	22.4%	61.9%	15.7%

답자들은 개선 요구가 높게 나타났다. 이는 농산업 현장에서 필요한 통계자료보다는 연구에 필요한 자료가 상대적으로 더 많이 제공되고 있음을 보여준다.

라. 웹 분석 기능에 대한 응답자 평가

웹 분석 기능에 대한 응답자 요구 분석에서는 농림부 통계시스템에 대해 농정조직에 소속된 사용자들의 요구가 다소 높게 나타났다<표 6>. 전체적으로 농업통계시스템을 통한 웹 분석기능에 대해 사용자들은 80%정도가 개선이 필요하다는 의견을 나타내, 향후 시스템개발에 있어서 웹 분석기능이 매우 중요함을 보여준다.

마. 원자료 제공 기능에 대한 응답자 평가

추가적인 분석을 원하는 사용자들에게 원자료를 제공하는 기능에 대한 요구는 농림부와 통계청 농업통계 시스템 각각 ‘필요하지 않다’는 의견이 11.4%와 14.8%에 그쳤다. 응답자 유형에 무관하게 개선 요구가 높게 나타나고 있어, 향후 시스템개발에 있어서 원 자료를 제공할 수 있는 기능이 매우 중요함을 보여준다.

표 6. 웹 분석 기능에 대한 사용자 요구

구분	농림부 농업통계 시스템			통계청 농업통계 시스템				
	N	개선이 필요없음	개선이 필요	개선이 매우필요	N	개선이 필요없음	개선이 필요	개선이 매우필요
연구조직	139	23.7%	58.3%	18.0%	128	25.8%	60.9%	13.3%
농정조직	163	19.0%	65.6%	15.3%	153	22.9%	65.4%	11.8%
현장조직	182	19.2%	56.0%	24.7%	170	18.8%	58.8%	22.4%
전체	484	20.5%	59.9%	19.6%	451	22.2%	61.6%	16.2%

표 7. 원자료 제공 기능에 대한 사용자 요구

구분	농림부 농업통계 시스템			통계청 농업통계 시스템				
	N	개선이 필요없음	개선이 필요	개선이 매우필요	N	개선이 필요없음	개선이 필요	개선이 매우필요
연구조직	144	9.0%	66.7%	24.3%	134	15.7%	63.4%	20.9%
농정조직	175	11.4%	66.9%	21.7%	165	12.1%	67.3%	20.6%
현장조직	188	13.3%	56.9%	29.8%	175	16.6%	57.7%	25.7%
전체	507	11.4%	63.1%	25.4%	474	14.8%	62.7%	22.6%

### 3. 요구분석 소결

면접조사 결과와 설문조사 결과를 토대로 농업통계 시스템을 구축할 때 가장 중요하게 고려해야 할 내용들을 정리하면 다음과 같다.

#### 가. 사용성

면접조사 때에 가장 많이 언급된 요구는 ‘사용하기 쉬운 환경’이었다. 농업통계 사용자, 특히 현장 농가 사용자들이 이해할 수 있고 직관적으로 사용할 수 있는 시스템이 필요하다고 피면접자 대부분이 언급했다. 그리고 자주 사용되는 웹 분석 기능에는 사용자가 손쉽게 보고서를 생성할 수 있게 하는 템플릿(templates)이 필요하다는 지적도 있었다. 일부 피면접자들은 특별한 형태의 보고서를 생성할 수 있는 고급의 기능을 요구했다. 그러나 가장 많이 요구된 웹 분석 기능은 계량 모형이나 고급 알고리즘을 이용한 분석이 아니라, 자료를 다양한 차원에서 평균이나 빈도를 분석할 수 있는 기능이었다.

#### 나. 접근성

사용성 다음으로 가장 많이 언급된 요구사항은 접근성이었다. 농업정책 관계자들이 분석을 위해 자료가 필요할 경우, 지금은 타 부서나 타 기관(농림부 관계자의 경우 통계청)에 요청해야 한다. 농업통계 자료들을 데이터 웨어하우스로 통합하고, 그 통합된 시스템을 통해 타 부서나 타 기관에 자료를 요청할 필요 없이 직접 접근할 수 있도록 할 필요가 있다. 한편, 경우에 따라서는 개인 정보 등이 자료에 포함되어 있기도 하므로 통계 정보를 생산하고 분석하는 권한에 따라 데이터 접근가능성과 분석가능성을 제한할 필요도 있다. 접근성을 제고시키려면 웹 기반 어플리케이션(application)이 필요하다는 지적이 많았다.

#### 다. 유연한 시스템

세 번째로 많이 언급된 요구사항은 사용자의 특별한 요구에도 응답할 수 있는 시스템의 유연성이었다. 사용자 대부분은 기존의 농업통계 시스템이 설정해 둔 보고서만 획득할 수 있는 것에 대해 불만을 갖고 있었다. 쿼리 검색을 가능하게 한 농림부의 일부 통계 자료들의 경우에도, 단순한 형태의 필터링이나 검색기능을 이용한 분석만을 지원하고 있다. 따라서 보다 심도 있는 분석을 원할 경우에는 사용자가 농업통계 담당자나 통계 전문가의 도움을 받아야만 하는 실정이다. OLAP나 DM 등의 고급 기능을 활용하여 사용자가 원하는 결과를 웹 인터페이스에서 바로 분석할 수 있는 시스템에 대한 요구가 많았다.

#### 라. 일관성 있는 정보

현재의 농업통계는 조사 시점에 따라 조사항목이나 기

준이 바뀌는 경우가 종종 있다. 그런데 사용자들은 해당 시점의 자료가 아니라 연도별 추세 등 시계열 분석에 대한 요구를 높게 갖고 있었다. 따라서 변동 정보를 일관되게 관리하는 것이 매우 중요하다. 요구분석에 응한 피면접자들은 자료의 일관성, 정확성 등에 대한 요구가 많았다.

## IV. 농업통계시스템 프로토타입 개발

### 1. 농업통계시스템(AGSTAT) 아키텍처

Umar(1993)는 아키텍처에 대해 “시스템의 구성 요소, 구성 요소가 제공하는 기능 및 시스템 구성 요소들간의 인터페이스 또는 상호작용을 묘사하는 것”이라고 정의내리고 있다. 또한 Poe(1995)는 “시스템 또는 제품의 전반적 설계를 위한 개념적 틀(Framework)을 제공하는 규칙 및 구조들의 집합”이라고 정의했다. 이상을 근거로 하여 본 연구에서는 데이터 웨어하우스를 이용한 농업통계시스템 아키텍처를 전반적인 시스템 설계를 하기위한 개념적 틀을 제공하는 구조들의 집합으로 정의하였다.

농업통계시스템(가칭: AGSTAT)는 데이터의 수집에서 사용자에게 대한 데이터의 배포의 전체과정을 담당하도록 설계되었다. 따라서 AGSTAT는 이 과정들을 포괄할 수 있는 아키텍처인 데이터웨어하우징(data warehousing)<sup>1)</sup>을 기반으로 설계되었다. 데이터웨어하우징(data warehousing)은 정보시스템 분야에서 가장 중요한 전략 중의 하나로(Eckerson, 1988, 1999, Eckerson과 Watson, 2000, Inmon, 2001), 최근 Palo Alto Management Group에 의하면 데이터 웨어하우징 시장은 2003년 현재 약 135억 달러에 이른다. 데이터웨어하우스는 의사결정을 지원하기위한 데이터를 준비하는 저장소로, 다양한 시스템으로부터 원본데이터를 통합한다(Gray와 Watson, 1998). 데이터웨어하우스를 통해 사용자는 더 좋은 데이터에 접근할 수 있고, 이로 인해 데이터 분석능력과 의사결정 능력이 향상된다.

AGSTAT의 아키텍처는 그림 3과 같다. AGSTAT는 크게 원본자료(Data Source), ETL 소프트웨어(ETL Software), 데이터웨어하우스(Data Warehouse), 어플리케이션(Applications)의 4계층으로 이루어져 있다. 이와 함께 시스템 정보와 분석데이터를 관리하는 메타데이터(metadata)도 포함하고 있다.

1) 데이터웨어하우스(data warehouse)와 데이터웨어하우징(data warehousing)은 명백히 구분되는 개념이지만, 많은 경우 이를 혼동하여 사용하는 경우가 많다. 데이터웨어하우스는 의사결정을 효율적으로 하기위한 하나의 데이터저장소이다. 이에 비해 데이터웨어하우징은 훨씬 포괄적인 개념으로 데이터웨어하우스를 포함한 전체적인 과정이다. 즉 데이터 웨어하우스는 원본데이터의 추출(extraction), 변형(transformation), 적재(loading), 사용자와 어플리케이션의 데이터에 대한 접근 등을 포함한다.

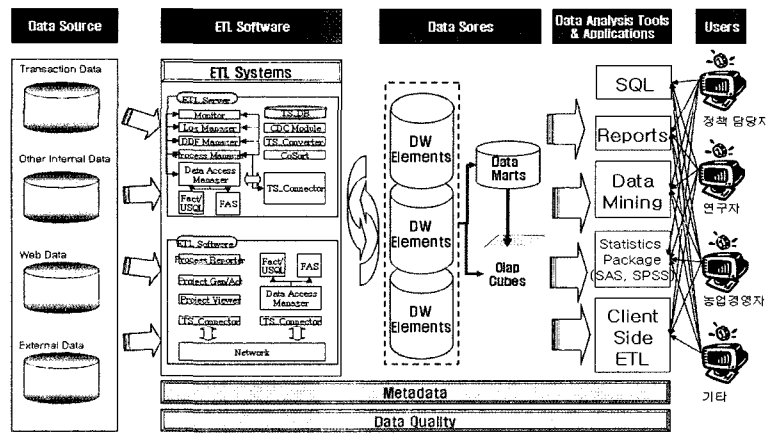


그림 3. AGSTAT 아키텍처

표 8. 원본자료(Data Source)의 종류

구분	설명
트랜잭션 자료 (Transaction Data)	- 생산, 회계, 마케팅 등의 트랜잭션 시스템 (운영시스템)으로부터 생산되는 자료 - 대부분의 자료는 이 범주에 속함
기타내부자료 (Other Internal Data)	- Transaction Data외의 다른 내부 자료 - 예) SAP, PeopleSoft와 같은 ERP로부터 생산되는 자료
웹자료 (Web Data)	- 웹 로그 등의 웹에서 생산되는 자료
외부자료 (External Data)	- 타 기관으로부터 생산되는 외부자료 - 예) 여론조사기관에 의해 생성되는 설문조사, 국가기관에서 생산되는 통계자료 등

## 2. 원본자료(data sources)

원본자료(Data Sources) 계층은 아키텍처의 제일 왼쪽에 위치하고 있다. 원본자료는 표 8과 같이 보통 4개의 범주로 구분할 수 있다(Watson, 2001, Werner와 Abramson, 2001).

AGSTAT에서 가장 중요한 데이터의 범주는 지역의 통계사무소 등에서 생산되는 내부 자료이다. 이 데이터들은 지역에서 입력되어 AGSTAT의 데이터 웨어하우스에 저장된다. AGSTAT가 일반적인 데이터 웨어하우스와 다른 점은 운영시스템으로부터 생산되는 데이터인 ‘트랜잭션 자료’를 다루지 않는다는 점이다. 이에 따라 AGSTAT는 다양한 원본데이터 시스템에서 생산되는 데이터를 실시간으로 통합하는 데이터 저장소인 ODS(operational data stores)가 존재하지 않는다. AGSTAT의 지속적인 업그레이드와 기능 향상을 위해서는 사용자들의 이용패턴 파악, 자주 이용하는 정보 파악, 분석형태 파악 등이 필요하다. 이를 위해서는 사용자들의 웹 사용 정보인 ‘click stream’ 데이터를 수집하여야 한다. 이는 웹 서버의 로그데이터를 통해

수집할 수 있다.

## 3. ETL Software

이 부분은 Data Source로부터 데이터를 추출하고(Extract), 변형하여(Transform), 목표 DB에 적재하는(load) 기능을 담당하는 계층(layer)이다. 원본데이터(Data Source)를 가지고 있는 원본시스템(source systems)으로부터 직접 프로그램을 이용하거나, 상업적인 ETL 도구를 이용하여 자료를 추출한다. 이후 추출된 데이터는 계획영역(Staging area)에서 변형(transformation)이 일어난다. 데이터 변형에서 가장 중요한 부분은 데이터를 정제(cleansing)하는 것이다. 데이터 정제를 편리하게 하는 소프트웨어들이 개발되어 있으며, 대부분 이 소프트웨어들을 이용하여 데이터 정제가 이루어진다. 데이터 변형과정이 완료되면 데이터웨어하우스에 데이터가 적재(loading)된다.

대부분의 농업통계 자료에는 수백 개의 항목이 포함되어 있었으며, 연도별 데이터포맷도 다양하고, 연도별로 항목변동과 코드설정이 변경되는 경우가 종종 있다. 따라서 현재 년도로 표준화하더라도 향후에 또 변경해야 하는 문제점과 사용자에게 따라 과거를 기준으로 분석해야 하는 경우도 생길 수 있다. 또 통계자료가 공표되고 나면, 더 이상 수정되거나 추가되지 않는 특성을 지니고 있다. 이를 고려하면 기존의 ETL 툴은 통계자료를 다루기에 적절하지 않다. 본 연구에서는 프로토타이핑 개발시 통계특성에 맞도록 기존 ETL툴을 커스텀마이징하였다.

## 4. 자료저장소(Data Stores)

원본시스템의 데이터는 ETL 과정을 거쳐 데이터 저장소로 적재된다. 일반적인 아키텍처에서의 데이터 저장소는 표 9와 같다.

표 9. 자료 저장소(Data Stores)

구분	설명
ODS(operational data stores)	· 다양한 원본데이터 시스템에서 생산되는 데이터를 실시간으로 통합하는 기능을 하는 데이터저장소
데이터 웨어하우스 (Data Warehouse)	· 의사결정을 지원하기 위해 사용되는 통합된 데이터 저장소
(데이터 마트) Data Marts	· Data Warehouse의 일부 데이터로 구성된 데이터 저장소 · 특정 사용자(예: 유통부서, 재무부서)의 의사결정을 지원하기위해 구축 · 대부분 다차원 데이터베이스로 구성되며, 매우 빠른 응답시간을 가짐

ODS는 운영계 어플리케이션을 지원하기 위한 목적을 가진 데이터 저장소로, ODS에 적재된 데이터는 일정 시간이 흐르면 데이터 웨어하우스에 적재된다. 이 경우 ODS는 데이터 웨어하우스의 원본데이터 역할을 한다. 그러나 앞서에서도 논의되었듯이 AGSTAT에는 현재 ODS가 존재하지 않는다. 이는 아직 실시간으로 처리해야할 운영계 데이터가 존재하지 않기 때문이다.

5. 메타데이터(Metadata)

메타데이터는 데이터를 기술하기 위한 2차적인 데이터이다. 크게 시스템 정보를 관리하기 위한 메타데이터와 분석하고자 하는 데이터를 기술하기 위한 메타데이터로 구분된다. 메타데이터 생성 능력을 향상시키고, 벤더(vendor)들 간의 데이터를 교환할 수 있도록 하는 두 가지 방법이 있다<표 10>.

현재 CWM와 같은 메타데이터 표준화를 따르는 것이 메타데이터 교환의 추세가 되고 있다(Markus, 2001, Vaduva 와 Vetterli, 2001, Watson, 2001, Stephens, 2004). AGSTAT는 벤더(vendor)들이 제공하는 메타데이터 틀을 이용하여 메타데이터를 처리한다.

표 10. 메타데이터 교환 방법

구분	설명
메타데이터 표준화	· 컴포넌트간의 상호운용을 지원하기 위하여 어떤 형태의 메타데이터가 저장되어야 하는지를 정의 · OIM(open information model) : Texas Instruments 와 Microsoft에 의해 제시된 첫 번째 표준 · CWM(command warehouse model) : Object Management Group이 OIM를 기준으로 발전시킨 표준
APIs (Application program interfaces)	· 해당 제품이 어떤 메타데이터를 저장하고 있고, 어떻게 접근이 가능한지에 대해 플랫폼에 독립적인 APIs를 통해 제공 · 대부분 XML(extensible markup language)이 APIs를 개발하는 표준

표 11. 데이터웨어하우스 접근 및 분석도구의 예

구분	설명
SQL queries	- SQL 쿼리를 이용해 자신이 필요한 보고서를 직접 산출 - 데이터웨어하우스내의 data model과 SQL 쿼리를 이해하는 고급사용자(power users)가 사용
Managed query environment	- SQL 쿼리를 손쉽게 작성할 수 있는 윈도우 인터페이스 제공하는 소프트웨어 - SQL을 잘 모르는 일반 사용자도 손쉽게 자신이 원하는 보고서 작성 가능 - Business Object, Cognos, Brio, ProClarity 등의 대표적인 제품
Management Reporting Systems	- 조직에서 정기적으로 필요한 보고서를 산출 - 과거에는 부서에 따라 다른 데이터베이스를 산출, 따라서 보고서의 내용이 불일치하는 문제 발생. Data warehouse기반에서는 동일한 데이터를 사용해 일관된 보고서 산출이 가능
Data Mining	- 데이터로부터 의미있는 패턴을 발견하도록 하는 고급 분석 도구 - Data Mining과 관련된 고급 지식을 가지고 있는 사용자가 데이터웨어하우스에 직접 접근하여 분석하거나 필요한 데이터를 자신의 컴퓨터로 다운로드하여 분석
DSS/EIS	- 다양한 DSS와 ESS 어플리케이션이 데이터웨어하우스의 데이터를 기반으로 구축 - MicroStrategy의 MicroStrategy Agent, Crystal Decisions의 Holog가 대표적인 제품

6. 어플리케이션(Data Analysis Tools and Applications)

사용자들은 매우 다양한 분석도구와 어플리케이션을 이용하여 데이터웨어하우스의 데이터에 접근한다. AGSTAT에서는 데이터 웨어하우스에 저장된 데이터를 표준화된 데이터 형식으로 제공함으로써 다양한 분석도구가 AGSTAT의 데이터를 사용할 수 있도록 설계되었다. 데이터 웨어하우스의 자료를 이용하여 분석하는 대표적인 도구를 살펴보면 표 11과 같다.

AGSTAT의 경우에는 위에서 제시된 일반적인 분석도구 외에 통계데이터의 특성을 고려하여 SAS, SPSS 등의 통계패키지를 이용해 추출한 데이터를 분석하고, 이를 웹에 게시할 수 있는 어플리케이션을 포함하였다. 이와 함께 국가의 공식 통계데이터는 원본 데이터에 대한 요구가 많다. 이를 고려하여 사용자들이 손쉽게 원본데이터를 가 공할 수 있는 클라이언트 ETL 도구를 포함하였다.

V. 농업통계시스템 프로토타입 개발

앞 장에서 설정한 농업통계시스템 아키텍처의 유용성을 확인하기위해서 농업통계시스템(AGSTAT) 프로토타입을 개발하였다. 전체 시스템은 크게 원본자료(Data Source)



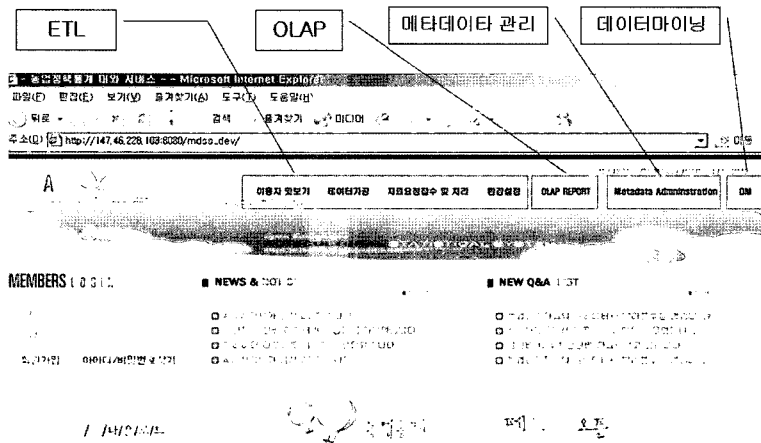


그림 4. AGSTAT 프로토타입 메인 화면

부분과 데이터 웨어하우스(Data Warehouse), 어플리케이션 부분으로 나뉘어진다. 원본자료로는 1984년부터 1987년까지의 수도작 생산비 자료를 사용하였다. 수도작 생산비는 노동력, 노동시간, 영농형태 등 약 400여개의 항목을 포함하고 있으며, 현재 텍스트 파일 형태로 저장되어 있다. 텍스트 파일이 데이터 웨어하우스에 적재되기 위해서는 변환과정이 필요하다. 변환을 위해 TeraStream ETL Software를 사용하였다. 데이터 웨어하우스를 이루는 주요 데이터베이스 서버는 MS-SQL 2000을 사용하였으며, MS-SQL 2000에 포함된 Analysis Services를 이용하여 OLAP Cubes와 Data Marts, DM Engine을 구성함으로써 데이터 웨어하우스를 완성했다. 이와 함께 SAS를 이용하여 MS SQL 2000의 데이터베이스에 접근하여 Data Mining 기능을 수행할 수 있도록 하였다. 데이터웨어하우스를 운영하는 서버로는 Windows 2003을 사용하였다. 어플리케이션 부분은 클라이언트 ETL 및 분석서버를 이용해 구성된 데이터 웨어하우스 결과물들을 사용자들에게 구현해주는 단계다. 구현 기능에는 ETL 기능, OLAP 기능, 데이터마이닝 기능, 메타데이터 관리(Metadata Administration) 기능이 있다.

사용자 어플리케이션을 중심으로 AGSTAT 프로토타입을 살펴보면 다음과 같다. 사용자 어플리케이션은 모두 웹 환경에서 동작하도록 개발되었으며, 전체 사용자 어플리케이션 화면은 그림 4와 같다. 그림 4에서 보듯이 AGSTAT가 보유하고 있는 데이터를 다양하게 추출·가공할 수 있는 ETL 기능, OLAP 기능, 데이터마이닝 기능, 메타데이터 관리 기능으로 구분된다. 사용자 어플리케이션에는 회원가입, 게시판, 자료실 등이 기본적으로 제공되고 있으며, 각각의 메뉴는 권한이 부여되어 회원이 아닌 사람은 이용할 수 없도록 서버구성 단계에서부터 통제하고 있다.

### 1. ETL 기능

농업통계시스템이 기존의 데이터 웨어하우스와 가장 구별되는 점은 데이터의 성격이다. 농업통계데이터는 일반적인 운영데이터(operational data)에 비해 하나의 테이블이 수백개의 항목을 포함하고 있으며, 연도별로 항목이 변경되는 경우가 많다. 따라서 원본데이터를 적재한 후에는 데이터에 대한 가공이나 변형이 일어날 필요가 거의 없는 일반 데이터 웨어하우스에 비해 분석목적에 따라 원본데이터가 새롭게 가공되어 적재되어야 하고, 사용자에게 원본데이터를 직접 제공해야 할 필요성이 많다. 따라서 농업통계시스템은 원본데이터를 다양하게 추출·가공·적재할 수 있는 ETL 툴의 기능이 강화되어야 한다. 따라서 본 연구에서는 AGSTAT 아키텍처의 유용성을 확인하기 위해서 웹을 통한 데이터 가공·추출에 대한 ETL 프로토타입을 개발하였다.

ETL 기능을 공용 웹 서버를 통해 사용자나 관리자가 데이터웨어하우스에 저장되어 있는 자료를 추출·가공할 수 있는 정보 시스템이다. 데이터 가공은 회원 가입이 끝난 사용자가 본격적으로 ETL 서비스를 요구하는 메뉴다. 하부 메뉴로는 데이터 추출 및 집계와 있으며, 이것으로 사용자는 웹을 통해 필요한 데이터를 추출, 집계할 수 있다. 다만, 정책 결정자, 행정적 관리자, 모듈 디자이너의 판단에 의해 사용자가 요구한 데이터가 보안과 정보 불균형에 어긋나지 않은 상태일 경우에만 자료요청 및 처리 메뉴를 통해서 자료를 보급할 수 있도록 시스템이 구성되어 있다. 필요한 자료를 추출하기 위해서는 우선 추출항목을 선택해야 한다<그림 5>.

이 후 선택된 자료의 항목이 나타나면 필요한 항목을 선택한다<그림 6>. 만약 추출조건이 필요할 경우에는 추출조건을 선택하고 실행할 수 있다<그림 7>.

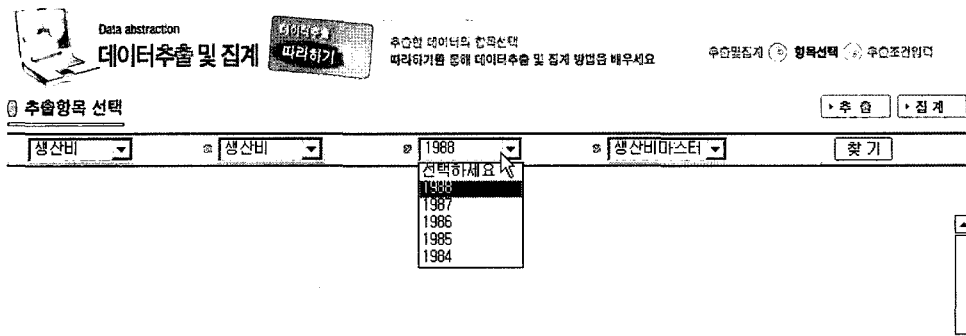


그림 5. 추출 그룹 및 파일(FDD) 선택 화면

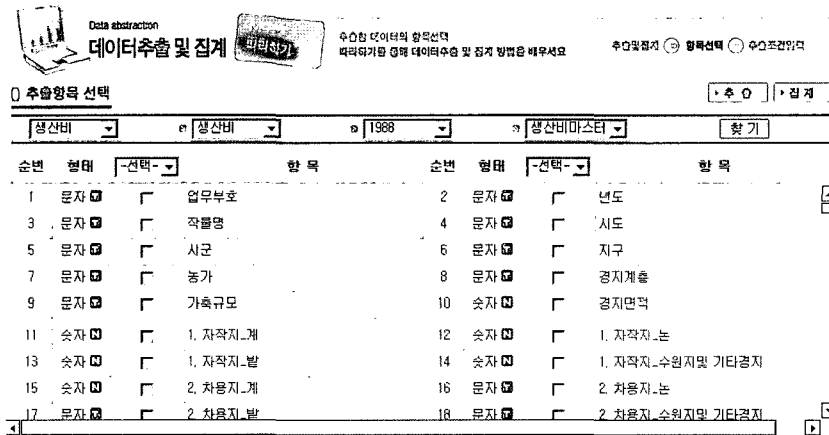


그림 6. 선택 항목에 따른 칼럼 목록

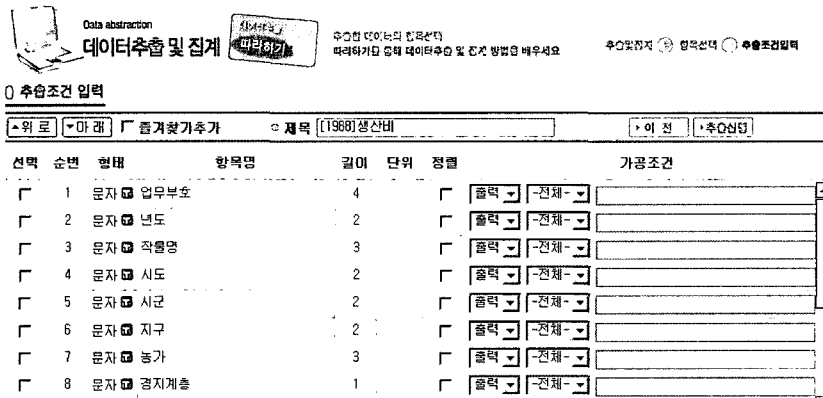


그림 7. 추출 조건 선택 및 추출 실행

필요한 데이터를 추출·가공하면 처리결과를 사용자는 텍스트파일, HTML 파일, 엑셀파일 등 다양한 형태로 받아볼 수 있다.

## 2. OLAP 기능

OLAP은 데이터 웨어하우스로부터 정보를 가공하고 전

달하는 가장 대표적인 방식이다. 본 연구에서도 웹을 통한 농업통계 온라인분석의 유용성을 확인하기 위해서 OLAP 프로토타입을 개발하였다. OLAP는 크게 MOLAP과 ROLAP으로 나누어 볼 수 있다. MOLAP는 다차원 분석을 목적으로 생긴 방식으로 응답속도가 빠르고 분석을 위한 다양한 고급 통계 기법을 지원하는 장점을 가진다. 따라서 본 프로토타입 개발에서도 MOLAP이 사용되었다.

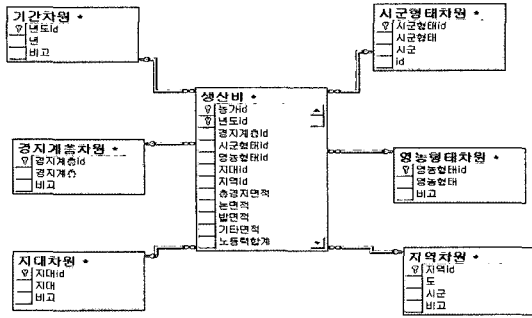


그림 8. 별 모양 스키마를 적용한 생산비 자료

본 연구에서는 농업통계 OLAP 시스템을 개발하기 위하여 수도작에 대한 생산비 원본데이터를 확보하였다. 데이터 보안문제 등을 고려하여 최근의 자료보다는 과거의 자료인 84~87년도의 데이터를 이용하여 시스템 개발을 하였다. 생산비의 경우 총 400여개의 변수를 가지고 있다. 그리고 연도마다 항목에서 차이가 있다. 따라서 본 연구에서 개발한 ETL 툴을 이용하여 원본 텍스트파일에서 필요한 항목만을 각 연도별로 추출하였다.

가. 데이터모델링 및 차원·큐브 설정

본 연구의 OLAP프로토타입을 위한 OLAP 서버로 마이크로소프트사의 Analysis services를 사용하였다. Analysis services는 데이터 모델링으로 두 개의 기법 - 별 모양 스키마(star schema), 눈송이 스키마(snowflake schema)-을 모두 지원하지만, 별 모양 스키마를 사용하도록 권고하고 있다. 따라서 본 연구에서는 별 모양 스키마(star schema)를 적용하였다. 그림 8은 본 OLAP 프로토타입을 위해 생성된 별 모양스키마이다.

그림 8에 나타난 모델을 기반으로 6개의 차원을 설정하였다. 설정된 차원은 경지계층, 기간, 시군형태, 영농형태, 지대, 지역이다. 그리고 팩터테이블(fact table)을 기준으로 큐브를 설정하였다. 설정된 큐브는 면적, 노동력, 생

산비, 무기질비료사용량, 무기질 비료사용금액, 유기질비료 사용량, 유기질비료 사용금액, 평가액이다.

나. 사용자용 웹 어플리케이션 개발

사용자용 OLAP분석 기능은 크게 두 가지로 구분된다. 첫 번째는 마이크로소프트 오피스 프로그램이 가지고 있는 OWC(Office XP Web Components) 컨트롤을 이용해 OLAP서버에 접속해 분석하는 방법이며, 두 번째는 MDX 쿼리를 이용해 분석하는 방법이다.

1) OWC 컨트롤 이용

OWC 컨트롤을 이용할 경우에는 MDX쿼리를 모르는 초보 사용자들도 엑셀의 피벗기능과 동일한 방식을 통해 분석하는 것이 가능하다. 로그인 후 Pivot을 이용한 OLAP 분석을 선택하면, 기존에 분석했던 피벗 자료가 리스트 되어 나타나는 것을 확인할 수 있다<그림 9>. 이때 관리자용 레포트 버튼을 선택하면 기존의 자료에 대한 수정, 삭제를 진행 할 수 있다.

레포트작성	관리자용레포트
<ul style="list-style-type: none"> <li>• 경지계층/영농형태별 면적</li> <li>• 영농형태/경지계층 별 면적</li> <li>• 경지계층별 면적</li> <li>• 영농형태별 면적</li> <li>• CC</li> <li>• 영농형태/시군/년도별 총면적</li> <li>• aa</li> <li>• aa</li> <li>• bbb</li> <li>• ff</li> <li>• 경지계층/지대별 자작지 경지면적</li> <li>• 년도별/지역형태별</li> </ul>	

그림 9. 초기 메뉴

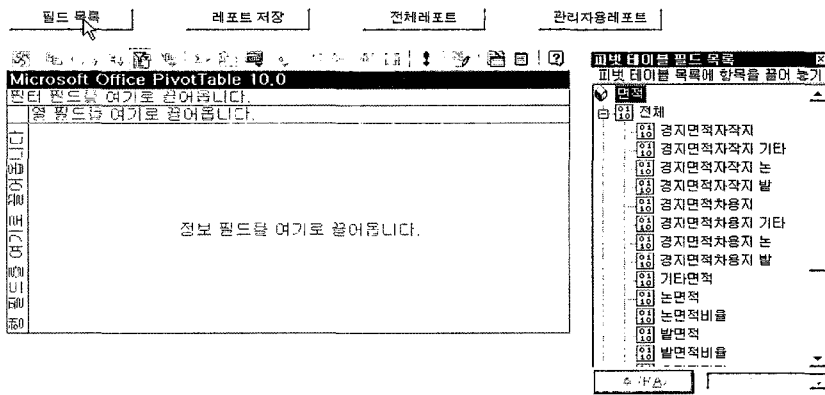


그림 10. 레포트 작성과 필드 목록 선택

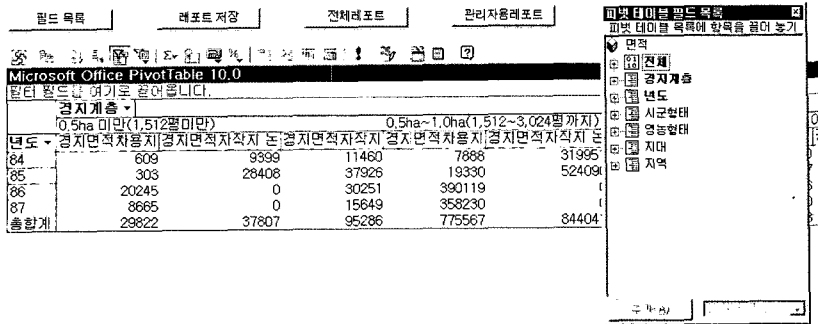


그림 11. 서버 컨트롤을 이용한 Pivot 틀 사용 결과

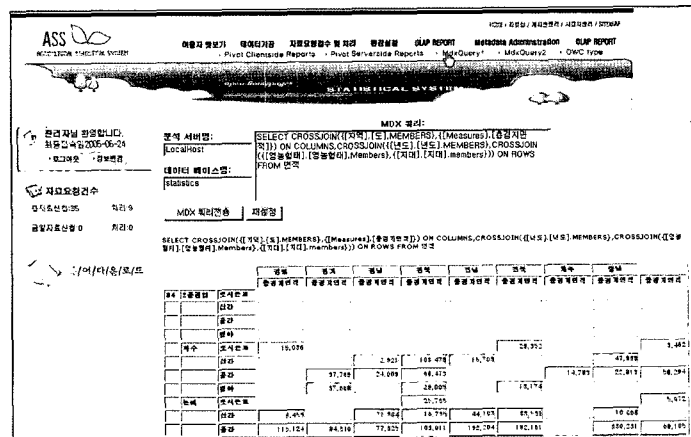


그림 12. MDX Query1

새롭게 피벗 기능을 이용해 데이터를 가공하고자 할 경우에는 레포트 작성 버튼을 클릭하면, 그림 10과 같은 화면이 나타난다.

레포트 작성 버튼과 함께 네 가지 버튼 메뉴가 나타난다. 필드목록을 선택하면 분석 서버의 큐브 및 SQL 데이터베이스 서버와 연동된 데이터 항목들이 자동으로 OWC 컨트롤을 통해 나타난다. 그림 11을 보면 각 차원과 설정된 큐브가 피벗 테이블 필드 목록에 나타나 있음을 볼 수 있다. 마우스로 드래그 & 드롭을 통해 데이터를 X축과

Y축에 옮기면 해당 데이터가 자동으로 분석 모듈을 수행해 데이터를 산출한다.

### 2) MDX를 이용한 어플리케이션

MDX 이용한 어플리케이션은 고급사용자용으로 사용자가 데이터베이스 및 큐브를 선택하여, 직접 쿼리를 수행할 수 있도록 한다. 그림 12는 사용자가 분석서버와 데이터베이스를 선택한 후 실행한 MDX 쿼리문과 결과를 보여준다.

#### OWC type-in ASS (agricultural statistical systems)

Server 선택: [AisLab(큐브전용)]  
 Database 선택: [Statistics]  
 측정값 선택: [종경지면적]

Cube 선택: [연적]  
 그래프 형태 선택: [직접선 그래프]

차원 선택: [경지계층] [영농형태] [지역]  
 레벨 선택: [경지계층] [영농형태] [지역]  
 빈 데이터 포함

Query:  
 SELECT NON EMPTY {[경지계층].[경지계층].Members} on columns, NON EMPTY {[영농형태].[All].Members} on rows FROM 연적 WHERE {(Measures).[종경지면적]}

[결과값 보기]

그림 13. MDX 쿼리 화면

MDX 쿼리를 작성하는 것은 고급사용자만 사용가능하다. 따라서 X축과 Y축, 차원, 큐브를 선택하면 MDX 쿼리가 자동으로 생성되고 결과를 보여주는 어플리케이션도 개발하였다<그림 13>. 그림 14를 보면 사용자는 차원, 큐브, 데이터베이스를 사용자가 직접 선택한 후, 선택한 계층에 대한 측정값을 직접 생성, 가공할 수 있다. 다양한 조건을 사용자가 선택하면 MDX 쿼리문이 자동으로 만들어진다.

‘결과값 보기’ 버튼을 클릭하면 해당 MDX 쿼리의 결과가 테이블과 그래프로 나타난다<그림 14, 15>.

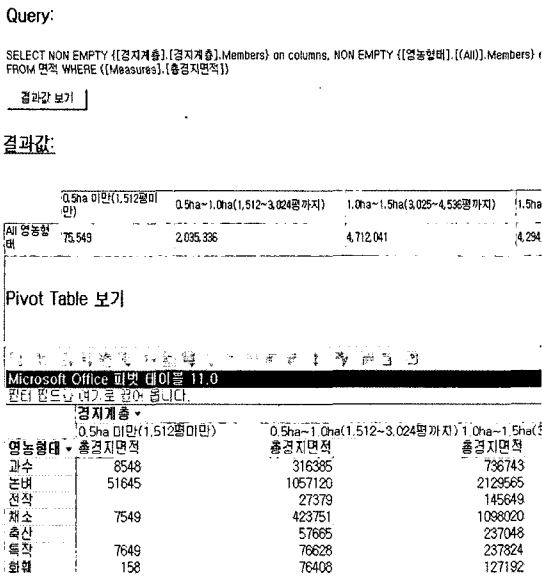


그림 14. MDX 결과 값

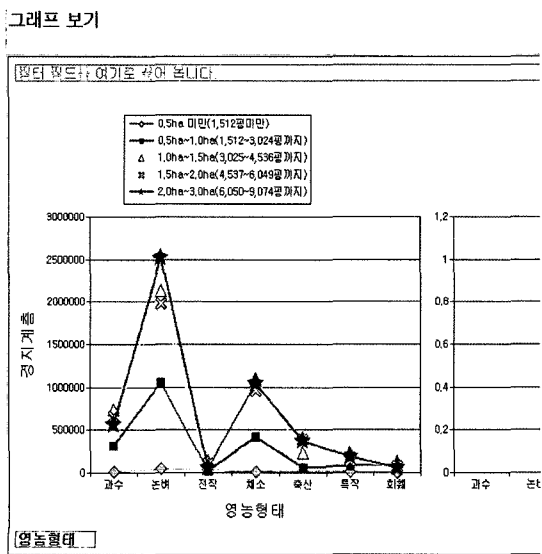


그림 15. MDX 결과값 2

### 3. 데이터마이닝 기능

데이터마이닝에 관한 여러 정의가 있지만, 간단히 요약하면, “대용량의 데이터에 존재하는 데이터간의 관계, 패턴, 규칙들을 찾아내고 모호화해서 유용한 정보로 변환하는 일련의 과정”이라고 할 수 있다(Cabena 등, 1998, Fayyad 등, 1996). 기존의 농업통계 분석은 대부분 빈도와 평균위주로 분석되어 왔다. 그러나 현재 수십년 간의 방대한 농업통계 데이터가 존재한다. 따라서 이 방대한 통계로부터 데이터간의 관계와 패턴을 분석하는 데이터마이닝이 기법의 도입이 필요하다. 데이터마이닝은 일반적으로 단일 데이터베이스와 같은 작은 용량의 데이터를 기반으로 하는 것이 아니라 통합된 대용량의 데이터를 기반으로 이루어진다(Chen 등, 1996, Han과 Kamber, 2001). 본 연구에서 설정한 AGSTAT 아키텍처의 유용성을 확인하기 위해서 데이터마이닝 프로토타입을 개발하였다.

#### 가. 원본데이터의 확보

데이터마이닝 프로토타입을 위해 사용된 데이터는 2001년 조사한 ‘전국 전업규모 양돈농가 경영실태 조사서’의 설문지 데이터이다. 이 자료는 농림부와 대한양돈협회가 공동으로 조사하는 설문지로, 2년 마다 조사하는 것으로 계획되어 있다.

본 데이터마이닝 어플리케이션 개발에 공식 농업통계를 사용하지 않은 이유는 다음과 같다. 공식 농업통계자료의 경우는 개별 농가에 대한 특성을 조사한 자료가 삭제되어 있다. 이는 개인 정보보호를 위해서이다. 따라서 데이터마이닝을 위해서는 각 개별농가의 특성에 대한 자료가 필수적이다. 왜냐하면 농업통계를 통해 파악하고자 하는 비즈니스 문제는 어떤 특성을 가진 농가가 더 많은 생산비를 사용하고 있는가? 특정 항목의 생산비를 많이 사용하는 농가의 특성은 무엇인가? 농가들을 어떻게 적절히 군집화 할 수 있는가? 등 농가와 밀접하게 관련되어 있기 때문이다.

본 연구에서는 ‘전국 전업규모 양돈농가 경영실태 조사서’의 설문내용 중 아래의 9개 변수를 추출하여 데이터마이닝 분석에 이용하였다<표 12>. 이 설문지는 총 4,976 농가의 데이터를 포함하고 있다.

#### 나. 모형 구축

데이터 마이닝을 위한 웹 어플리케이션 구성을 위해, 마이크로소프트사의 Analysis Services를 이용하였다. Analysis Services에서는 현재 데이터마이닝 기법으로 Decision Tree(의사결정나무)와 Clustering(군집분석)을 지원한다. 본 연

표 12. '양돈농가 경영실태 조사서'에서 추출한 문항

칼럼명	변수설명	변수값
Education	교육수준	초졸 이하, 중졸, 고졸, 전문대이상
JobLength	영농종사기간	3년 이내, ~5년, ~10년, ~20년, 20년 이상
FarmType	영농형태	일괄사육, 자돈생산비육, 자돈구입비육, 비육돈 위탁
Age	연령	(int)
Labors	노동력	가족, 1~3인 고용, 4인 이상
Standard	표준화정도	전축사표준화, 일부 표준화, 미표준화
Scale	사육규모	(int)
MIS	정보 시스템 수용	미사용, 정보수집, 경영S/W, 정보수집 + S/W
Enlarge	사육규모 확대 의향	사육감축, 현상유지, 1~25%확대, 26~50% 확대, 50% 이상 확대

구에서는 종속변수로 MIS, Enlarge를 설정하고 나머지 변수를 독립변수로 설정하였다. 따라서 Decision Tree(의사결정나무)에서는 두 개의 모형 - MIS를 종속변수로 하는 모형과, Enlarge를 모형으로 하는 모형- 이 생성되었다.

그림 16은 Analysis Services에서 MIS를 종속변수로 하여 생성한 의사결정나무이다. 이를 보면 왼쪽 제일 위쪽에는 모형을 생성하기 위해 사용한 변수들이 나타나 있다. 그리고 왼쪽 아래에는 모형을 설정하기 위한 속성이 있다. 여기서는 알고리즘으로 Microsoft Decision Tree를 사용했음을 보여준다. 그림의 중앙에는 의사결정나무의 결과를 보여주고 있다. 각 노드가 어떻게 결정되었는지를 이 의사결정 나무를 통해 알 수 있다. 여기서 나타나는 색깔은 진할수록 많은 케이스, 즉 농가를 포함하고 있음을 보여준다. 마지막으로 오른쪽 아래에는 특성이 나타난다. 이는 각 노드에 속한 농가의 MIS 수용상태를 보여준다. 현재는 '모두'에 클릭된 상태로 전체 농가들의 MIS 수용 빈도를 보여주고 있다.

다. 사용자용 웹 어플리케이션

Analysis Services에서 설정한 데이터와 모형은 Microsoft의 OLEDB for OLAP를 통해 연결할 수 있다. OLEDB for OLAP를 사용하여 웹에서 정보시스템 수용확률과 규모확대 의향에 대한 확률을 추출하는 데이터마이닝 어플리케이션을 구축하였다.

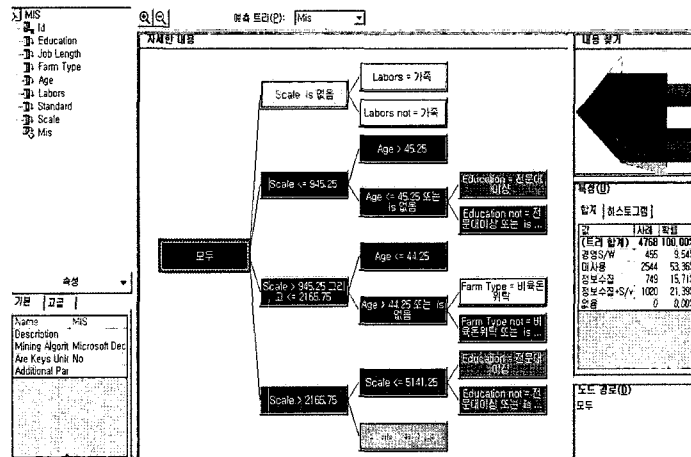


그림 16. Decision Tree 모형 구축

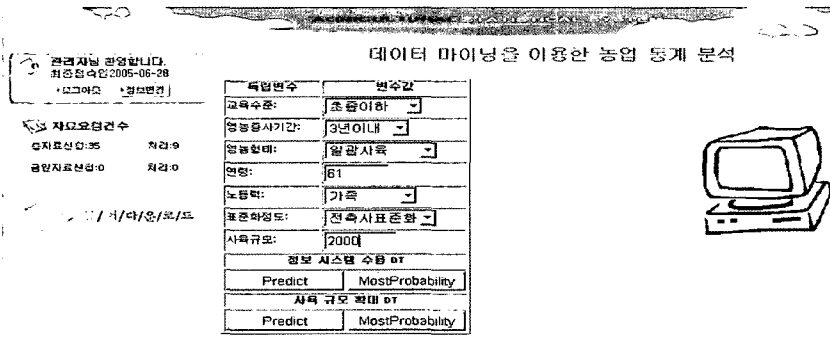


그림 17. 데이터마이닝 프로토타입 화면

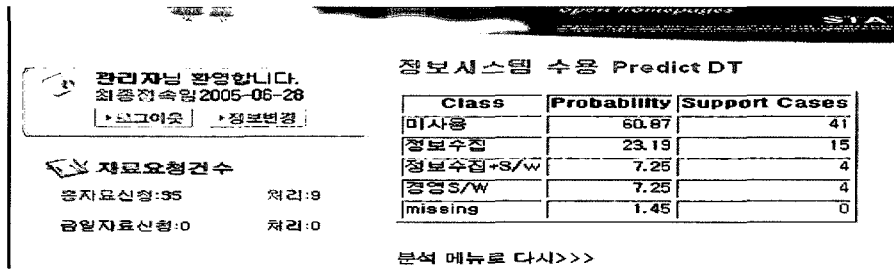


그림 18. 데이터마이닝 실행 결과

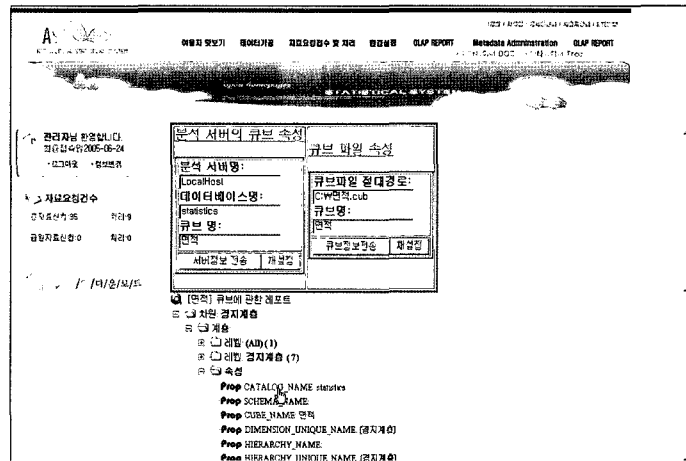


그림 19. 메타데이터 Tree 큐브 보고서 화면

메뉴중 'DM'을 클릭하면 그림 17과 같은 화면이 나타난다. 독립변수의 값을 설정한 후 정보 시스템 수용 DT의 'Predict' 버튼을 클릭하면 해당 농가의 정보시스템 확률이 산출된다. 'MostProbability' 버튼을 클릭하면 해당농가가 속할 확률이 가장 높은 정보시스템 수용 상태를 보여준다. 사육 규모 확대 DT의 'Predict' 버튼과 'MostProbability' 버튼은 향후 사육 규모 확대에 대한 확률을 예측한다.

예를 들면 위 그림에서 설정한 농가(초졸 이하, 3년 이내, 일괄사육, 61세, 가족농, 전축사표준화, 2000두 농가)의 경우에 정보시스템 수용확률을 추정하기 위해 정보시스템 수용 DT의 'Predict'를 실행하면 그림 18과 같은 결과가 나타난다. 즉 이 농가의 경우는 미사용 확률이 60.87% 정보수집으로 만 사용할 확률이 23.19%, 정보수집과 농가용 S/W를 함께 사용할 확률이 7.25%이다. 그리고 현재의 데이터 내에 위의 농가에 부합되는 농가에 대한 정보(support cases)도 함께 나타나 있다.

#### 4. 메타데이터정보 관리 기능

메타데이터 관리는 데이터베이스나 데이터 세트에 의해서 일시적으로 차원을 확장 축소하여 현재 서버에 구성되어

있는 메타데이터의 종류와 계층 구조를 분석할 수 있도록 한 메뉴다. 분석 서버명과 데이터베이스명, 그리고 큐브명을 선택한 후, '서버정보 전송' 버튼을 클릭하면 구성된 큐브의 차원, 레벨, 측정값 등의 속성에 대한 정보를 출력한다<그림 19>.

## VI. 결론

본 연구는 농업정책의사결정을 향상시키기 위해서 데이터 웨어하우스에 기반한 농업통계시스템(AGSTAT) 아키텍처를 개발하였다. 이를 위해 사용자 요구분석을 실시하였으며, 최근의 데이터 웨어하우스와 관련된 주요 개념을 토대로 아키텍처를 설계하였다. 이 후 아키텍처의 유용성을 평가하기 위해서 AGSTAT 아키텍처를 기반으로 응용 어플리케이션 프로토타입을 개발하였다.

일반적으로 데이터 웨어하우징 프로젝트에 의해 정보 분석환경을 재 구축한 경우는 다음과 같은 이점을 얻게 된다(Chen 등, 1996, EDS, 1995, Fayyad 등, 1996, Pendse, 2003). 첫째, 쉽게 다양한 각도로 업무를 분석할 수 있기 때문에 의사결정의 효율성과 효과성이 향상된다. 둘째, 내장되어 있는 분석기능과 리포팅 기능을 이용하여 최종

사용자 혼자서 처리할 수 있는 업무량이 증대될 수 있다. 셋째, 데이터를 새로운 관점에서 보게 됨으로써 새로운 아이디어나 대안을 마련할 수 있다. 넷째, OLAP이 제공하는 다차원 분석을 통해 사용자는 데이터를 여러 관점과 계층에서 분석함으로써 중요한 추세, 또는 문제점의 출처를 발견할 수 있도록 해준다. 다섯째, 부서 내에서 필요한 데이터를 업무별로 구조화함으로써 관리항목의 범위를 명확히 하고, 보고서 작성을 최소화할 수 있다.

본 연구에서도 위와 같은 장점을 고려하여, 데이터 웨어하우스를 기반으로 농업통계시스템 아키텍처를 설계하였다. 그러나 농업통계시스템의 경우는 주로 운용데이터(operational data)를 효과적으로 통합하는 일반적인 데이터 웨어하우스와는 차이가 존재한다. 기존 데이터 웨어하우스가 다루는 운영데이터에 비해 농업통계시스템이 다루는 데이터의 차이를 비교해보면 다음과 같다. 첫째, 통계 데이터는 일반적으로 수백 개의 항목을 포함하고 있어 일반 운용데이터에 비해 관리해야 할 항목이 많다. 둘째, 연도별로 항목이 분할·통합, 또는 삭제·생성되는 경우가 많다. 셋째, 매년 설문유형에 따라 데이터 레이아웃(Layout)이 상이하다.

위와 같은 농업통계데이터의 특성으로 인해 농업통계시스템은 원본데이터를 다양하게 추출·가공·적재할 수 있는 ETL 툴의 기능이 강화되어야 함을 보여준다. 즉 일반적인 데이터 웨어하우스의 경우에는 원본데이터를 적재한 후에는 데이터에 대한 가공이나 변형이 일어날 필요가 거의 없지만, 농업 통계데이터의 경우에는 분석목적에 따라 원본데이터가 새롭게 가공되거나 배포되어야 한다.

본 연구에서는 농업통계데이터의 특성을 고려하여 ETL 기능을 강화한 4개의 층(layer)으로 구성된 농업통계시스템 아키텍처를 설계하였다. 첫 번째 층은 다양한 원본데이터를 포함하는 원본데이터(data source) 계층이다. 두 번째 층은 상이한 원본데이터로부터 정보를 추출하고, 변환하고, 이를 목표 DB에 적재하는 ETL시스템이다. 세 번째 층(layer)은 SQL 서버의 DB, OLAP서버, DM서버로 구성된 DW 시스템이다. 네 번째 층(layer)은 사용자들이 이용하는 클라이언트 응용프로그램이다. 이 아키텍처를 기반으로 본 연구에서는 구체적인 물리적 아키텍처를 설계하였다. 설정된 물리적 아키텍처를 이용해 ETL, OLAP, 데이터마이닝, 메타데이터 관리 프로토타입을 구현하였다. 이를 통해 본 연구에서 설계한 농업통계시스템 아키텍처의 유용성이 검증되었다. 즉, 원본데이터를 손쉽게 데이터웨어하우스에 통합할 수 있을 뿐만 아니라, 원본데이터를 사용자들이 원하는 다양한 방식으로 가공·추출할 수 있었다. 또한 OLAP, 데이터마이닝 기능을 이용하여 사용자들은 좀 더 효율적으로 농업통계데이터를 분석할 수 있다. 따

라서 본 연구에서 구현한 논리적 아키텍처와 프로토타입은 향후 농업통계시스템을 구현하는 참조모델로 활용될 수 있다.

이 논문은 2004년도 농림부 연구비를 지원받아 수행되었습니다.

## 참고문헌

1. Cabena, P, P. hadjinian, R. Stadler, J. Verhees, and A. Zanasi, 1998, *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, New Jersey
2. Chen, M.S., J. Jan, and P.S. Yu, 1996, "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering* 8(6) : 866-883
3. Eckerson, W.W., 1988, "Post-Chasm Warehousing", *Journal of Data Warehousing* 3(3): 38-45
4. Eckerson, W.W., 1999, *Evolution of Data Warehousing: The Trend toward Analytical Applications*, Boston, MA: The Patricia Seybold Group, 1-8
5. Eckerson, W.W. and H.J. Watson, 2000, *Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions*. Seattle: The Data Warehousing Institute
6. EDS, 1995, "Data Warehouseing Primer", EDS, December
7. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communication of the ACM* 39(11) : 27-34
8. Gray, P. and H.J. Watson, 1998, *Decision Support in the Data Warehouse*, Upper Saddle River, Prentice-Hall, New Jersey
9. Griffin, J., 1995, "Customer Information Architecture", *DBMS* : 58-65
10. Han, J. and M. Kamber, 2001, *Data Mining: Concepts and Technique*, Morgan-Kaufmann Academic Press, San Francisco
11. Inmon, W.H., 2001, "Knowing Your Dss End User: Tourists, Explorers, Farmers", [www.billinmon.com/library/articles](http://www.billinmon.com/library/articles)
12. Inmon, W.H., 2002, *Building the Data Warehouse*. John Wiley and Sons, New York



13. Legg, W., 2005, "Agricultural Support Indicators: Measurement, Meaning and Use", In Statistics, Knowledge and Policy(pp. 635-644), OECD(Eds). <http://www.sourceoecd.org/9264009000/> (accessed 2005.12).
14. Markus, M.L., 2001, "Towards a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success," Journal of Management Information Systems 18(1) : 57-93
15. McFadden, F.R., 1996, "Data Warehouse for EIS: Some Issues and Impacts", Proceedings of 29th Annual Hawaii International Conference on System Science : 120-129
16. OECD, 2003, Methodology for the Measurement of Support and Use in Policy Evaluation, <http://www.oecd.org/agr.support> (accessed 2005.12).
17. Pendse, N., 2003, "What is OLAP, The OLAP Report", (<http://www.olapreport.com/FASMI.HTM>;White Paper).
18. Poe, V., 1995, "Data Warehouse: Architecture is not Infrastructure", Database Programming and Design
19. Stephens, T.R., 2004, "Knowledge: The Essence of Meta Data: The Meta Data Experience", DM Review 14(3)
20. Umar, 1993, Distributed Computing and Client/Server Systems. Prentice Hall
21. USDA, 2005, USDA Strategic Plan for FY 2002-2007, <http://www.usda.gov/ocfo/usdasp/usdasp.htm>(accessed 2005.12).
22. Vaduva, A. and T. Vetterli, 2001, "Metadata Management for Data Warehousing: An Overview", International Journal of Cooperative Information Systems 10(3) : 273-298
23. Watson, H.J., 2001, "Recent Developments in Data Warehousing", Communications of the Association for Information Systems 8 : 1-25
24. Werner, V. and C. Abramson, 2001, "Managing Clickstream Data", Journal of Data Warehousing 6(3) : 11-15
25. Yost, M., 2000, "Data warehousing and decision support at the National Agricultural Statistics Service," Social Science Computer Review 18(4) : 434-441